TECHNICAL WRITING AND SIMPLE STATISTICS

For laboratory classes

By: Stephanie G. Wettstein
with chapter contributions from:
J. T. Browne, K.L. Morrissey, P. Russell

Licensing and Citations:

Abstract

This upper division resource focuses on how to communicate results through technical writing, use Excel to perform simple statistics, and create professional charts/documents. Excel tutorials are provided for performing descriptive statistics, t-tests, and linear regression as well as using text boxes, formatting figures and captions, and using Equation Editor to insert equations. Additionally, guidance and examples of different communication components are provided along with team writing strategies and guidelines on how to hold efficient meetings.

This book was made possible through grant support from the TRAILS OER program, funded by the Office of the Commissioner for Higher Education, Montana University System.

Table of Contents

Statistics

1. Descriptive Statistics

KATHRYN L. MORRISSEY AND STEPHANIE G. WETTSTEIN

Once you have collected data, what will you do with it? Data can be described and presented in many different formats and starting with the basics of analysis is important. The Descriptive Statistics tool in Excel can analyze multiple data points at a given experimental setting; that is, when no variables are changing. An example of this would be collecting multiple data points at set conditions like collecting 10 samples over time from a distillation column that is running at steady state. Descriptive Statistics provides numerical descriptions that include, but are not limited to, mean, median, and mode (central tendency) as well as range, standard deviation, and variance (dispersion or variation).

Mean, Median, and Mode
The mean, median, and mode of the data set tell us what data value represents or is typical in a data set. The mean is the average value of the data set and for a population (μ), is calculated using

$$\mu = \frac{\sum_{i=1}^{n} x_i}{N} \tag{1.1}$$

where x is the value of the data point and N is the number of data points in the data set. If the data set is a sample, or a subset of the population, the mean is represented by $\bar{x}$, but is calculated the same way as for a population (μ). Most data sets are samples of a larger data set known as the population. For example, if we wanted to know the average height of women and men aged 20-80 in Montana, the population would be every single person aged 20-80 in Montana. Since that would be nearly impossible to determine and time restrictive, we would randomly select a number of women and men aged 20-80 in Montana, which makes the data set a sample.

The median is the central value of the data set. The data are arranged from smallest to largest and the central value is the median. In the case that the data set has an even number of data points, the two central values are averaged. The mode of the data set is the value that occurs most often. In the case that no values occur more than once, the data set has no mode.

Outliers
At this point it is worth mentioning that in order to make solid conclusions from data, the data has to be good. This means that the experiment must be done well and, as a result, no outliers exist that could skew the data. An outlier is a data point that is abnormally far away from the other values. There are multiple ways to determine outliers, but a common way is to use quartiles that divide the data into quarters and percentiles that divide the data into hundredths. Quartiles are just special percentiles where the first quartile, Q_1, is the same as the 25th percentile and the third quartile, Q_3, is the same as the 75th percentile. The median is called both the second quartile and the 50th percentile. To score in the 90th percentile of an exam does not mean that you received 90% on a test. It means that 90% of test scores are the same or less than your score and 10% of the test scores are the same or greater than your test score. However, if scores of two students who did not complete the test were in the data set, those scores are going to skew the data lower, leading to a poor analysis. These outliers need to be identified and removed if there is a reason to do so prior to analysis.

In order to calculate quartiles and percentiles, the data must be ordered from smallest to largest and then interquartile ranges (IQR) can be used. The IQR is a number that indicates the spread of the middle 50% of the data and is calculated using

$$IQR = Q_3 - Q_1 \qquad (1.2)$$

where Q_3 is the third quartile and Q_1 is the first quartile. A value is identified to be a potential outlier if it is outside of the range defined by the upper and lower fences, which are calculated with

$$Upper\ fence = Q_1 - (1.5)(IQR) \qquad (1.3)$$

$$Lower\ fence = Q_3 + (1.5)(IQR) \qquad (1.4)$$

Potential outliers always require further investigation or mention in your report findings; therefore, it is important to be able to determine true data outliers. For example, if a data point lies just outside of the fence and there is no reason to remove it, it may be better to leave it in the data set. However, if it is determined that a data point may be bad due to a specific reason, e.g., incorrect scribing, experiment upset, etc., then it should be removed and noted in the findings.

Standard Deviation
An important characteristic of any set of data is the variation in the data. In some data sets, the data values are concentrated closely near the mean; in other data sets, the data values are more widely spread out. The most common measure of variation, or spread, is the standard deviation which measures how far data values are from their mean.

The standard deviation is always positive or zero and is calculated for a sample using

$$s = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}} \qquad (1.5)$$

where $\bar{x}$ is the mean of the sample population. In reports and other written documents, the standard deviation should always be reported with a mean and is commonly reported as $\bar{x} \pm s$. The lower-case letter s represents the sample standard deviation whereas the Greek letter σ (sigma, lower case) represents the *population* standard deviation determined by

$$\sigma = \sqrt{\frac{\sum(x_i - \mu)^2}{n}} \qquad (1.6)$$

where μ is the mean of the population. If the sample has the same characteristics as the population, then s should be a good estimate of σ.

The standard deviation is small when the data are all concentrated close to the mean, exhibiting little variation or spread. The standard deviation is larger when the data values are more spread out from the mean, exhibiting more variation. A data value that is two standard deviations from the average is just on the borderline for what many statisticians would typically consider to be far from the average. More will be discussed on this in the Process Control Chart section.

Sample Variance
The sample variance is useful for determine what type of t-test we be run on data sets, which will be discussed in Chapter 6. The sample standard deviation, s, is the square root of the sample variance, s^2, which is calculated using

$$s^2 = \frac{\sum(x - \bar{x})^2}{n-1} \qquad (1.7)$$

If the data are from a **sample** than a population, we divide by $n - 1$, one less than the number of items in the sample, just as we did for standard deviation as shown in equations 1.5 but if the data come from the entire **population** and not a sample, the denominator is just n, the number of items in the population.

Standard Error
Standard error and standard deviation are both reported in the Excel's Descriptive Statistics; however, standard deviation is more commonly used as a measure of data variation. This is because the standard error is always less than the standard deviation. The standard error of a sample is defined as

$$SE_{\bar{x}} = \frac{s}{\sqrt{n}}$$

(1.8)

which means the standard deviation divided by the square root of the number of data points is equivalent to the standard error. The standard error provides information on how far the sample mean is from the population mean while the standard deviation provides information on the variability of individual data points to the mean. When reporting on experimental data, the standard deviation is a more important value and therefore, the standard error is not typically reported.

Confidence Interval
The confidence interval (CI), expressed as a percentage of sureness, is a measure of how confident you are that your data will fall within a specific range of values and is calculated using

$$CI = t\frac{s}{\sqrt{n}}$$

(1.9)

where t is the t-value that can be found in a statistics table. It is common to use a 95% confidence interval, which uses a t-value of 1.960, and indicates that you are willing to be wrong 5 times out of 100, or 1 time out of 20. Sometimes much higher levels of confidence are appropriate. If you choose a very high level of confidence, you are indicating that you hardly ever want to be wrong. If you never want to be wrong, simply use an infinitely wide confidence interval – that is, choosing a high level of confidence produces a wide confidence interval. If you set a 99% confidence interval, you are willing to be wrong 1 time out of 100, which will result in a larger confidence interval than 95%.

Once the confidence interval has been determined, the CI value is subtracted from the mean, which is the lower confidence interval bound, and then the value is added to the mean, which is the upper confidence interval bound. For example, say you measured the circumference and diameter of 20 circles to determine if π equals the circumference divided by the diameter. If the 95% confidence interval of the data was calculated to be 0.21 and the mean of the data was 3.25, the confidence interval would be 3.04 – 3.46. That is, we are 95% confident that the true mean of our data lies between 3.04 and 3.46. Since the value of π (3.14) does lie within our confidence interval, our data, although having high variability, is representative of the population. The confidence interval can also be reported as $\bar{x} \pm CI$, but be sure to differentiate between the confidence interval and the standard deviation if you choose to report both in this manner.

Percent Error and Percent Difference
Percent error and percent difference allow you to compare your data to another data point or set and are especially useful for when you only have singular data points as t-tests cannot be done on single data points. This could involve experimental and/or theoretical data. The data source will determine whether you use percent error or difference. If you do not know which value is considered "correct," you use percent difference. If you know one of the values is the true value, for example, a theoretical value like π, a value from a textbook, or perhaps a value from a research paper, you use percent error.

The percent difference is calculated using

$$\% \, difference = \frac{|E_1 - E_2|}{0.5(E_1 + E_2)} * 100 \tag{1.10}$$

where E_1 is the value from the first experiment and E_2 is the value from the second experiment. These can be single data points or averages of several data points.

For the percent error, the calculation is made using

$$\% \, error = \left| \frac{T - E}{T} \right| * 100 \tag{1.11}$$

where E is the experimental value and T is the true or theoretical value. In most cases, T is a single value, but E could be a single data point or an average of several data points.

Excel Tutorial: Performing Descriptive Statistics Using Excel

Excel's *Data Analysis Add-In* is a powerful set of statistical utilities that will provide most of the statistical tools necessary to analyze the data from experiments. Check that the *Data Analysis Add-In* is available by clicking on the "***Data***" tab and looking for the "***Data Analysis***" link in the "***Analyze***" section of the navigation ribbon as shown in Figure 1.1.
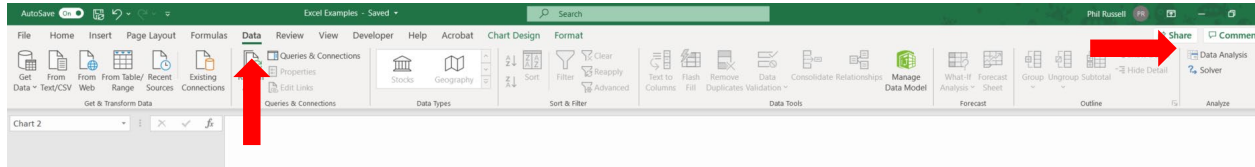
Figure 1.1: Location of Data Analysis link in the Data tool ribbon.

(If "***Data Analysis***" is not showing, directions to load it can be found using Microsoft online help.)

Listed in Table 1.1 are 100 data points that will be analyzed with ***Descriptive Statistics*** and were generated with Excel's "=NORM.INV(Probability, Mean, Standard_Dev)" function. This function returns the inverse of the normal cumulative distribution for the specified mean and standard deviation. To generate the data, the mean was set to 0 and the standard deviation was set to 1. We will compare this to the mean and standard deviation calculated for the generated data using the ***Descriptive Statistics*** tool.

Table 1.1: One hundred random numbers selected from normally distributed data ranging from -1 to 1.

-0.01	-0.93	-0.05	-0.52	2.56	0.45	0.26	-0.54	0.20	1.75
0.24	0.17	-0.25	-0.78	-0.23	-1.19	0.89	-0.84	-0.80	2.10
0.46	-0.40	1.00	-0.80	0.91	0.90	-1.61	1.23	-1.52	0.77
0.30	-0.87	1.00	-0.37	-1.45	-0.21	-1.48	1.06	-1.22	-0.63
0.95	0.53	1.24	-1.59	0.37	-0.40	0.50	1.96	0.74	0.60
-0.29	1.45	-1.05	-0.25	0.69	-0.49	-0.27	-1.81	1.28	0.84
1.68	0.45	0.69	-0.92	1.29	-1.66	-0.29	-1.32	1.25	1.07
-0.26	-1.31	-0.43	-1.11	0.97	-0.02	-0.75	-0.96	-0.55	1.03
-1.35	0.87	1.38	0.60	0.82	-2.16	0.74	-1.47	0.76	0.91
-0.34	0.05	-1.22	0.39	1.43	0.30	-0.06	-1.73	1.15	-1.97

Click on the ***Data Analysis*** link shown in Figure 1.1. This brings up the ***Data Analysis*** *tool selection* dialogue box (Figure 1.2). Scroll to and select ***Descriptive Statistics*** then click ***OK*** as shown in Figure 1.2.
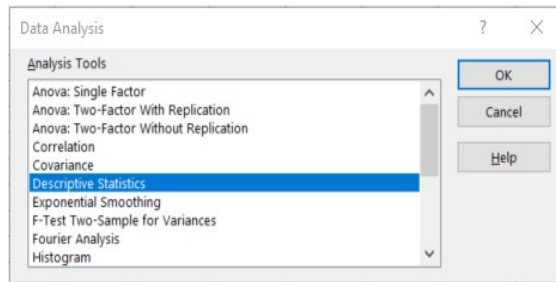
Figure 1.2: Data Analysis dialogue used to select Descriptive Statistics tool.

The **Descriptive Statistics** tool dialogue box appears and is shown in Figure 1.3. The data shown in Table 1.1 were contained in the range **E9:E108**, which is entered into the **Input Range** box by highlighting the data. In this example, the data was contained in one column so the **Columns** radio button was selected. (Even though there is only one column of data, the **Columns** radio button still must be selected).

The location where *Excel* will write the **Descriptive Statistics** results can be selected from a:
- **Output Range**, which could be a location on the same worksheet,
- a *New Worksheet Ply* (creates a new Excel worksheet tab)
- or a *New Workbook* (creates a new Excel workbook/file).

In addition, it is common to check the boxes for **Summary statistics** and **Confidence Level for the Mean** (95% confidence level requested in this case). Clicking on **OK**, gives the results shown in Table 1.2.
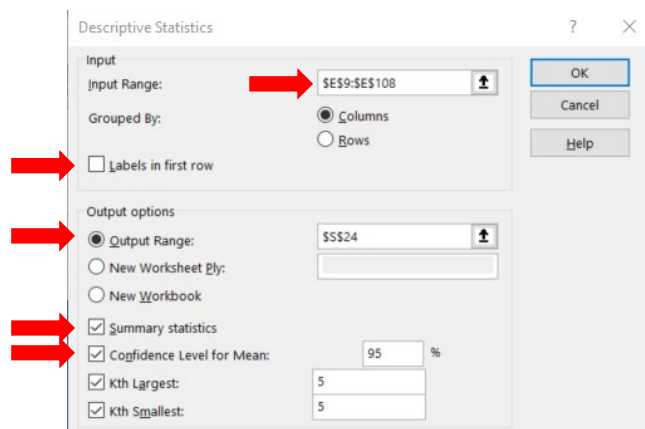
Figure 1.3: Input parameters provided to Descriptive Statistics tool to produce summary statistics and the confidence interval around the mean.

You should not just copy and paste the generated table into report narratives; you only need the data that will help support your analysis and discussion. Commonly reported values from the summary table are the (*sample*) *mean* (0.0249), (*sample*) *standard deviation* (1.0395), and the *confidence level* (0.2063). NOTE that the "Confidence Level" is not the level of confidence (which is typically 95%), it is the confidence interval given by eqn. 1.9. In this case, the confidence interval is between

$$\bar{x} - t \cdot \frac{s}{\sqrt{N}} \tag{1.12}$$

and

$$\bar{x} + t \cdot \frac{s}{\sqrt{N}} \tag{1.13}$$

as described in the Confidence Interval section; therefore, the confidence interval ranges from -0.18 and 0.23 for this data set.

Table 1.2: *Example descriptive statistics output for the data shown in Table 1.1.*

Column1	
Mean	0.025
Standard Error	0.104
Median	0.02
Mode	0.3
Standard Deviation	1.04
Sample Variance	1.08
Kurtosis	-0.75
Skewness	-0.05
Range	4.72
Minimum	-2.16
Maximum	2.56
Sum	2.5
Count	100
Confidence Level(95.0%)	0.206

In terms of reporting the data, you should rely on the number of significant figures in the data set. Typically, the number of significant figures plus one is acceptable to report. In this data (Table 1.1), there are three data points that only have one significant figure (0.05, -0.05, and -0.01), many that have two significant figures (0.24, 0.46, 0.30, etc.), and some that have three (1.68, -1.35, 1.00, etc.). Therefore, we should report two significant figures since one is the lowest in our data set. Our data would then be reported as: 0.025 ± 1.04. Notice that the standard deviation has three significant figures since we should report the standard deviation to the same number of decimal places as the mean.

Compare the mean and standard deviation 0.025 ± 1.04 to the data set we generated the value from (0.0 ± 1.0), it shows that, although close, our data had some variability in it. Increasing the data set would help this, but is it a concern? In order to find out, a t-test (Chapter 6) should be run using our data set testing against the "theoretical value" 0. If we do a *t-test: Two-Sample Assuming Unequal Variances*, it results in a p-value of 0.81, meaning that the generated data is not significantly different than 0. (More to come in Chapter 6!)

2. Using Graphs to Display Data

KATHRYN L. MORRISSEY AND STEPHANIE G. WETTSTEIN

A graph can be a more effective way of presenting data than a table because we can see where data clusters, where the data is sparse, as well as trends in the data. A graph is a tool that helps you learn about the shape or distribution of a sample or a population and the type of graph you choose is dependent on the data you collect.

Line Graphs

A type of graph that is useful for specific data values is a line graph, which typically show the data values on the x-axis (horizontal axis) and the frequency on the y-axis (vertical axis). When the data are related, the frequency points can be connected using line segments. For example, data such as temperature, product specifications, revenue, sales, and safety incidents, are often plotted using line graphs to show trends with time.

Let's say a chemical and biological engineering department is interested in seeing how the enrollment in the senior laboratory class has changed with time. The data in Table 2.1 was collected, which one can use to view the data. However, Figure 2.1 makes the data easier to analyze as one can look at the figure, quickly see enrollment peaked in 2018, and decreased to around 90 students in later years, which is below average. Additional information, such as the average enrollment over the period, can be added to make analysis even easier. The goal of figures is to provide the reader with an easy way to view important data and trends in the data.

Table 2.1: *Number of students enrolled in a senior lab course per year.*

Year	Number of students
2020	89
2019	93
2018	120
2017	97
2016	88
2015	105
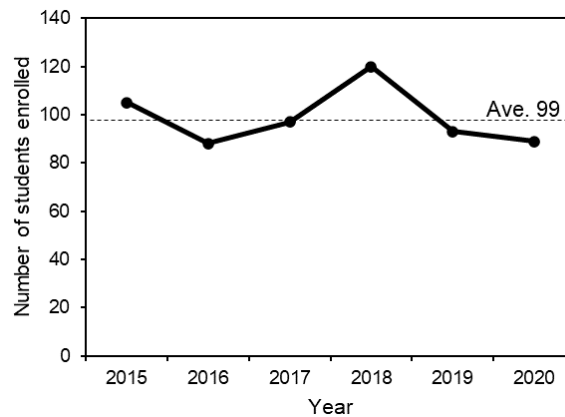
Figure 2.1: *Line plot of senior lab enrollment with time.*

Bar Graphs

Bar graphs consist of bars that are separated from each other. The bars are usually rectangles and they can be vertical or horizontal to show comparisons among categories. One axis of the chart shows the specific categories being compared, and the other axis represents a discrete value. Bar graphs are especially useful when categorical data is being used.

If we plot the same data in Table 2.1 on bar graphs, there is a visual difference compared to Figure 2.1. Figure 2.2 shows the two main types of bar graphs, vertical and horizontal. What you select depends on the number of categories, the space you have, and what you would like the reader to take away from the figure. In this case, the vertical figure better coveys the data as when comparison data over time, since time being on the x-axis is more intuitive for the reader. Sometimes making multiple figures for comparison is useful to see which style displays your data with the most clarity.
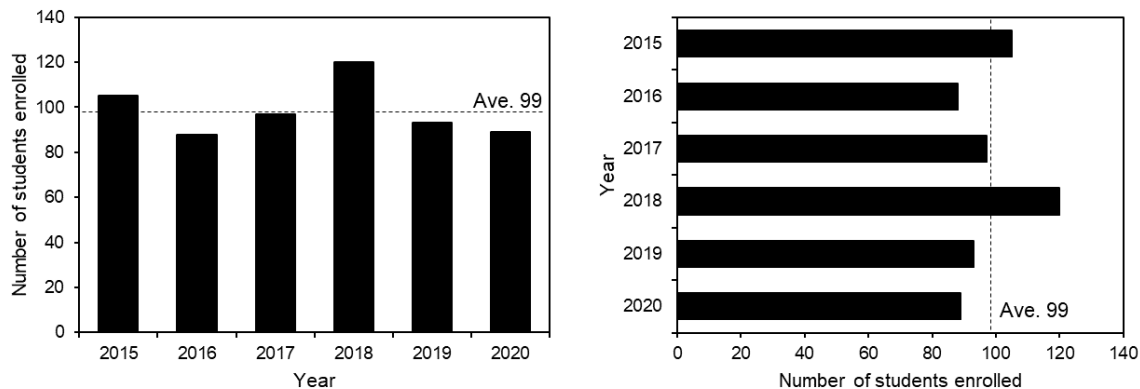
Figure 2.2: *Vertical (left) and horizontal (right) bar graphs of senior lab enrollment with time.*

Box Plots

Box plots are a type of graph that can help visually organize data in order to identify outliers and are good ways to display data that you want to present the mean for. To graph a box plot, the following data points must be calculated: the minimum value, the first quartile, the median, the third quartile, and the maximum value (see section on "Outliers" in Chapter 1). Once the box plot is graphed, you can display and compare distributions of data.

For example, in Figure 2.3, the data is comparing the percentage of problems that students answered correctly on a quiz in week 5 and week 13 of a course. The two asterisks identify two outliers that were removed from the analysis, the square markers represent the means, the boxes represent the 2nd and 3rd quartiles, while the "whisker" extending below the box, represents the lower, or 4th, quartile. In this case, the maximum value achieved is 100%, which was directly above the 2nd quartile and therefore, no

Figure 2.3: *Comparison of the percent correct of attempted questions for the quiz problems. The filled squares represent the mean scores. n=21 in week 5 and n=23 in week 13. Outliers are indicated by an asterisk.*

whisker extends above the box. Figure 2.3 shows that there is likely no significant difference in the percentage of correct answers for the quiz in week 5 compared to week 13 indicating that the problems were of similar difficulty for the students.

Error Bars

All experimental data contains error in the measurement. For example, if you collect multiple data points at the same conditions, an average and standard deviation can be calculated due to variability in the data. The standard deviation can be useful to put on a graph of the means so that readers understand the amount of error associated with the data. Including error bars allows the reader to visually see how precise the data is and to provide insight as to if there is a significant difference between data.

> Note that just because the standard deviations between two data points do not overlap, does not mean the points are significantly different. A statistical test, like a t-test, is necessary (that takes into the number of samples) in order to conclude that.

Typically, error bars are the standard deviation in the data, but other options include a fixed value, a percentage, or standard error. The average value is the data point (i.e., marker) and then the error bar extends from that point (Figure 2.4). Typically end caps are include to indicate the end of the error bar. In order to not overload the reader and make the graph difficult to understand, sometimes it is best to use error bars sparingly. For example, in Figures 2.4, data is plotted that consists of the average circumference and dimeters of five circles that were measured by multiple people. As the diameter of the circle gets larger, the circumference increases as well. Since multiple people measured the circles, there is error in the data. Showing that error on a graph can be done in multiple ways.

In Figure 2.4a, all the error bars are the same value in the x-direction as well as in the y-direction. That is, the error is the same for all circumference measurements and the same for all diameter measurements. In this case, it is best to display the error bars on a single point like is shown in Figure 2.4b in order to present a cleaner graph. If the error is the same, but smaller, as shown in Figure 2.4c, the error bars do not clutter the graph substantially and could be presented. If the error changes for each data point, as it does in Figure 2.4d, all error bars should be shown so that the reader can draw meaningful conclusions. In the case of Figure 2.4d, the error (standard deviation) got larger as the circles measured got larger. Since the error bars represent the uncertainty in the data, it can be seen that the data points in Figure 2.4c has a greater chance of being statistically different than the data points in Figure 2.4a due to the error bar size. However, in order to confirm this, a statistical test must be done.
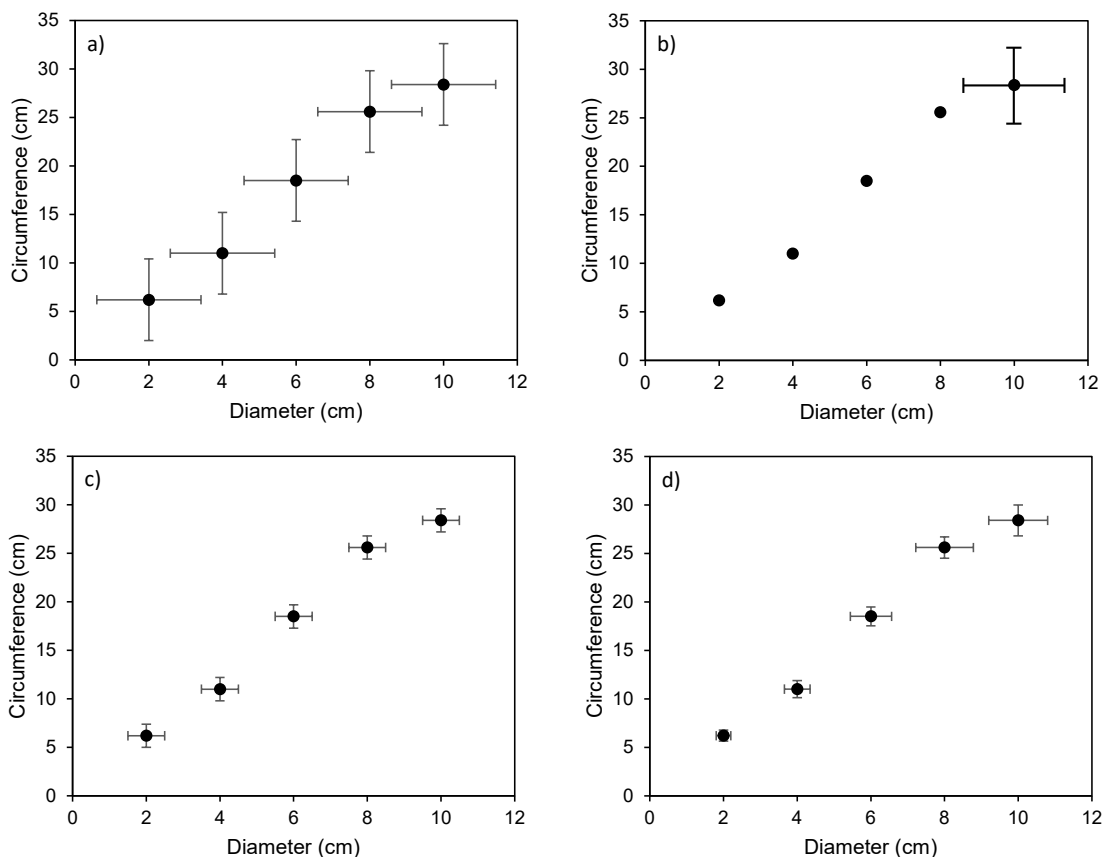
Figure 2.4: *Comparison of the diameter and circumference of different circles that were measured by multiple people. Figures a-d have different displays of error bars on the data where a) is a standard amount of uniform error, b) displays the same amount of error as a), but more cleanly, c) has a smaller uniform amount of error, and d) has different amounts of error for each point as given by the standard deviation of each data point.*

Charts – What Not to Do

Anyone who has made figures in Excel knows that there are many different templates that can be selected. Remembering that the goal of figures is to present your data in the simplest, clearest way to help the reader understand your results should help you design your chart. The figures presented previously in this chapter are typical of what you would see in publications: simple, clear, to the point (Figure 2.2), and how to make your figures look like these is in the appendix. However, let's take a look at some of the templates and options that Excel has to offer and why using them is not the best choice:
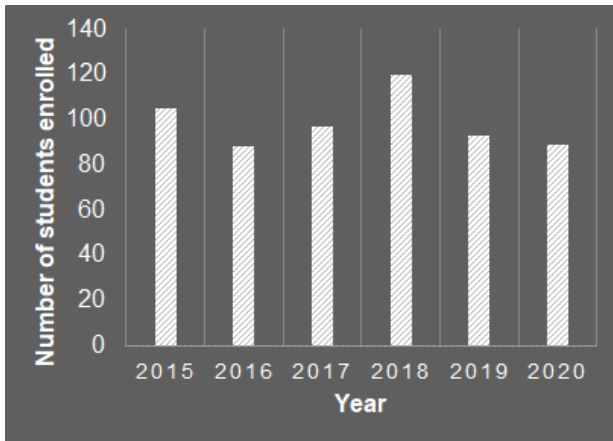
This figure has a harsh change in the background color. For printed documents, dark backgrounds with white/light text should be avoided due to potential printer resolution issues. Additionally, the striped bars are unnecessary.
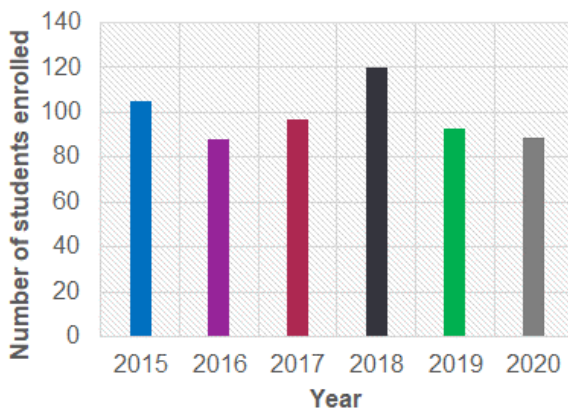
Although the major horizontal and vertical gridlines may be okay (still not recommended), this figure has a distracting diagonal background that does not add anything to the interpretation of the figure. Also, different colors are typically used to indicate different data sets. In this case, colors were used inappropriately as the reader will expect a different category (e.g., different course enrollment) for each color. Also, the gray font and lines are harder to see than just using black.

This figure has a distracting faded background along with "shadows" on the data bars that add no value. Again, the gray font is harder to see than black and there should not be a border around the entire figure. The addition of the labels (i.e., the numbers on each bar) may be of use depending on the narrative. If you are just conveying a trend and exact numbers are not critical, labels just clutter the figure. If exact numbers are important, labels should be considered or potential display the data as a table.
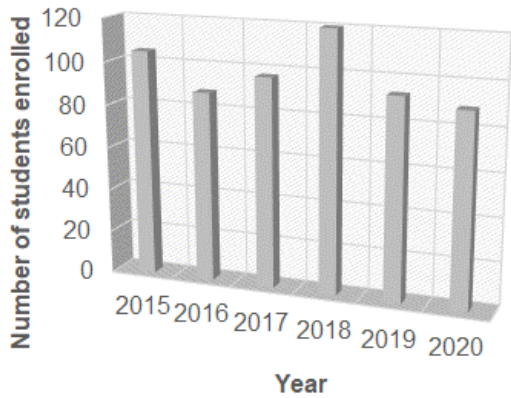
Speaking of distracting from the data, using "3D" figures when only presenting two dimensions of data is useless. In this case, the x-axis is angled, making it difficult to orient and the trend in data is not as clear. Again, gray font and lines are harder to see than black.
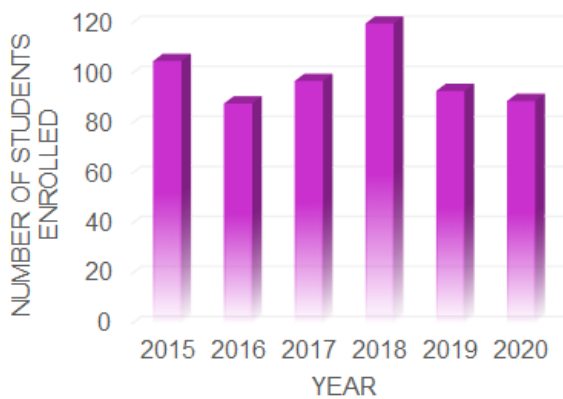
This "3D" figure is not angled, which is a better than the previous figure, but the capitalization of the axes reads aggressively and sentence capitalization (or capitalize the first letter of each word) should be used. Also, the unnecessary fade of the bars detracts from the data, while gray lines and fonts should be black. Another issue with this figure is that the y-axis line is missing.

So, what makes a good figure? See "Excel Tutorial: Formatting Excel Charts for a Professional Look" in the appendix for details and Figure 2.5 for an example, but key features are:

- Black axes font and lines around the entire chart/data area portion of the graph
- No outside border that encompasses the axes and chart area.
- Properly labeled axes and, if appropriate, labels directly on the chart area near the data (Figure 2.5).
- No excess decimal places on axes. (For example, the figures above do not have 20.0, 40.0, etc.)
- A good caption that describes the figure but does not contain discussion/results.
- No legends. (Markers should be identified in the caption as is shown in Figure 2.5.)
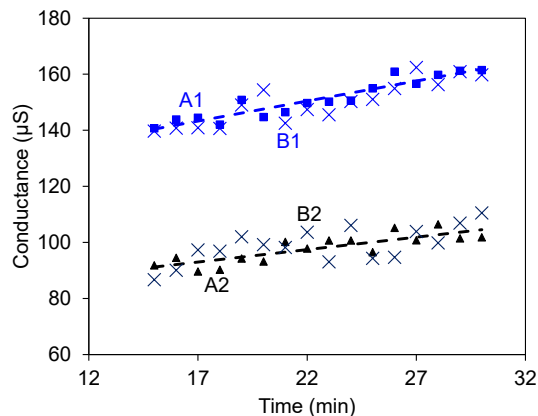
Figure 2.5: *Data for four liquid diffusion experimental runs (data sets A and B) and the linear regression fits for: run A1 (■), run A2 (▲), run B1 (×) and run B2 (×).*

3. Statistical Process Control & Process Control Charts

J. TRENT BROWNE AND STEPHANIE G. WETTSTEIN

A process control chart is a special type of statistical chart that is used to monitor process performance over time in order to identify sources of variation that could possibly be reduced to improve a process. For example, if there is product specification that needs to be met, say an alcoholic beverage proof of 60, the company is committing to the consumer that the product contains at least 30% ethanol by weight. If the beverage does not meet this specification, that could lead to consumer complaints, lawsuits, and fines. Now, no process can produce 30% ethanol exactly, all day, every day, because there is process variability. Therefore, the company needs to be able to oversee their process to ensure the amount of alcohol in their beverage is at least 30% and meets specifications. This is done by monitoring the variable of interest using a tool called Statistical Process Control, or SPC.

SPC is a statistical method that was developed to better monitor and control a process with the goal to produce more on-specification product and subsequently less off-grade and rework product. In short, SPC helps achieve process stability while improving the ability of the process to meet specifications by reducing variability due to unnecessary process changes. The key tool used to control variation is the process control chart (PCC). One type of PCC is a 6σ chart that assumes if you collect enough data, the distribution of the data would follow a bell-shaped curve with normal distribution. In a normal distribution, approximately 68.26% of the data falls within the area of two sigma (2σ), meaning one standard deviation above the mean and one standard deviation below the mean. Likewise, 95.44% of the data falls within four sigma (4σ) and 99.74% of the data falls within six sigma (6σ), the latter being three standard deviations above and three standard deviations below the mean ($3\sigma + 3\sigma = 6\sigma$). Figure 3.1 shows the regions highlighted on a chart that allow for users to quickly determine if a process is experiencing too much variation. These sigma regions will also help the operator decide when the process is statistically "in control" or "out of control." By knowing that the process is out of control the operator can preemptively make changes to a process control parameter(s) and bring the process back into control before a considerable amount of off-specification product is produced.
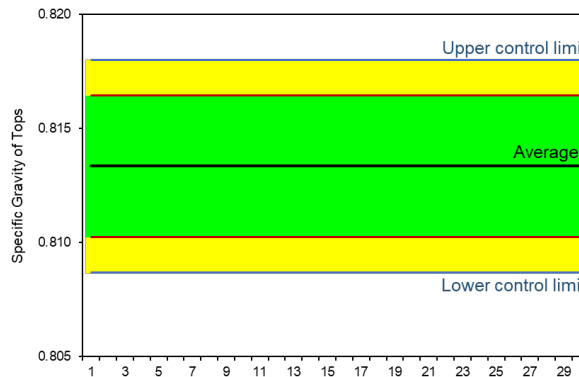
Figure 3.1: Control chart with 6σ regions shaded green (2σ and 4 σ) and yellow (6 σ).

Before any process control charts can be made and used, there needs to be a basic understanding of the manufacturing processes so that sources of variation can be identified. In the case of the alcoholic beverage, once the amount of variability is known, the process control limits can be determined and monitored to ensure a product containing at least 30% ethanol. If there is a process upset, there needs to be a determination as to what is wrong and subsequently how to fix it.

Process Variation

There are two types of variation: 1) random and 2) assignable cause variation. Random variation, sometimes called natural or normal variation is all the variation taking place that is inherent and unexplainable in any process at some small scale. This variation can be thought of as "background noise," with no need to implement controls.

Assignable cause, also called abnormal condition, non-random variation, or special cause is a type of variation that makes the process unstable and incapable of producing products that meet specifications. This type of variation must be controlled as much as possible. Companies that successfully control the assignable cause variations are the ones that have a better chance of being successful in the marketplace over the long run. Consistent product quality reduces the costs associated with returning and replacing defective product, handling and reprocessing off-specification product, and reducing the risk of losing customers. By using a PCC, product quality can be tracked and changes made to the process only when necessary. If changes are made to the process unnecessarily, *higher* product variability will result.

One of the first items to identify is where variation can enter in the system (Figure 3.2). Process variations can come from many sources including:

- Raw material inputs like where the raw materials were sourced from, how old they are, the moisture content, etc.
- Manufacturing process variations like machine speed, operator variability, mixing ratios, aging times, etc.
- Product testing and quality inspections like analytical testing methods, standards data, etc.
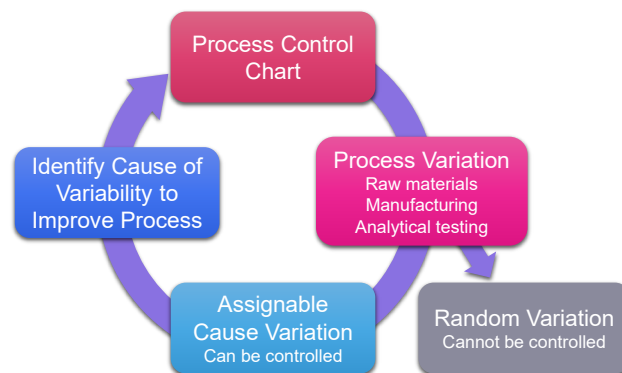
Figure 3.2: *Value in using process control charts.*

Say you were hired to be the new quality improvement manager for a company that produces distilled spirits such as vodka, whiskey, and bourbon. The owner is concerned because the company is getting customer complaints that the alcohol content (proof) of its high-end spirits seems to widely fluctuate from bottle to bottle, even though the actual *average* proof is consistent with the labeled proof on each bottle. Since the level of proof can affect taste, the owner is concerned that the company may start losing business to its competitors and this is not a path to good profitability and long-term growth. The owner has asked you to implement a quality improvement program to help correct this problem by reducing product variability.

In the case of the distillation example, there are many variables to consider that could affect the final alcohol content such as:

Raw Materials Inputs - The first step at any distillery is to purchase grains such as corn, rye and malt. These grains are then milled and mixed in batches with fresh water at percentages specified by the company's spirit recipes in preparation for the downstream fermentation step. What are some common raw material variations here?

- The first thing could be the grains themselves. Every farm has differences in soil make-up and moisture, sunlight availability, levels of irrigation and possibly even the source of the original seeds. Maybe the natural sugars that develop in one farmer's crop is different from that of another farmer. Since the distillation process involves fermenting these natural sugars to ethanol, the difference in initial sugar levels have a significant impact on the amount of alcohol that is produced.
- Secondly, the water used to prepare each batch can have a dramatic effect on the final taste. Is the water rich in minerals? Has it been properly filtered? What kind of filtration was used? Water

quality is so important that bourbon manufacturers love to put their distilleries near limestone water sources because they say it is needed for good taste.

Manufacturing Process Variations - Consider how many producers of spirits there are in the world and while each manufacturer uses similar operations, equipment, and procedures, every brand has dedicated customers that will only buy their one favorite, special brand. Each manufacturer has its own special processing steps to set themselves apart from their competitors and are fully aware that there are variations from batch to batch, even when making the same product. While changes to the incoming raw materials have already been mentioned, variations between batches can also occur because of such things as:

- Individual operators doing things slightly differently,
- Environmental variations such as barometric pressure changes that affect the boiling points of the alcohol solutions feeding the stills, and
- Variations due to process control issues. For example, the instruments used to measure process variables such as flow, temperatures and pressures can change due to shifting calibration, electrical/mechanical breakdowns or process issues such as material plugging or scaling.

Product Testing and Quality Inspections - Even after the products have been produced, product variations can take place during final quality control testing. Hydrometers, the equipment used to measure the alcohol content of the final distilled product, must be interpreted by quality control personnel, and individual interpretations can be a source of variation. Likewise, final product testing usually includes taste testing by a team of distillery employees. Taste testing can vary greatly from person to person and depending on who is on the taste-test team, this can introduce variations in the final product.

In summary, while the spirits industry is used as an example here, it should be stated that <u>all</u> manufacturing processes need to be aware of varying raw materials, variations in manufacturing processes and variations in quality control. Variation is basically a fact of manufacturing life and must be understood and appropriately controlled to maintain good product quality.

For the sake of this discussion, we will assume that after reviewing the product specifications associated with your spirits process, you determined that the most important variable to track is the alcohol content in the top product leaving the main still. If the distillery can lessen the variation of the alcohol content in the still top product, the alcohol level in the final product will be less variable and a more consistent product can be delivered to the customer. In order to determine the variability in the product and track when the product is not meeting specifications, you need <u>real data</u> about your still operation. You decide to have the still operators record specific gravity readings, which can be converted to weight percent ethanol, of the still top product every hour, around the clock, for 30 days when the column is at steady state. You average the twenty-four hourly specific gravity readings for each day, average the 30 daily averages, and then plot them on a chart (Figure 3.3).
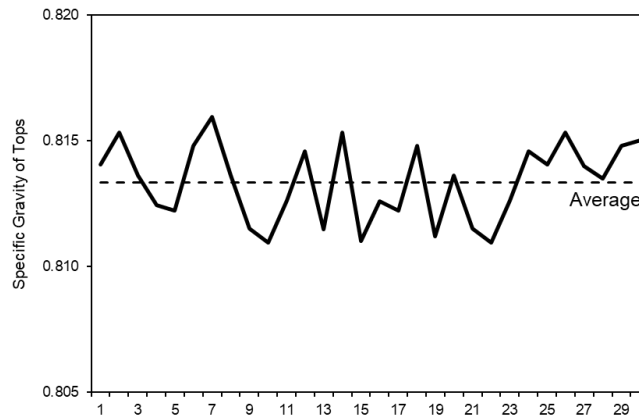
Figure 3.3: *Daily average still tops specific gravities for one month (monthly average indicated by a dashed line).*

14

In reviewing the chart, the first thing you see is that there appears to be a lot of variability from day to day. The second thing you see is that the **target** specific gravity of 0.8136 (specific gravity of 60 proof alcohol) is slightly higher than the **average** still tops specific gravity (0.8133), which is not much, but you would prefer the average value to be equal to the specification target. After seeing this data, you decide that this process would be a good application for SPC with the ultimate goal to reduce the variation of the alcohol content in the still tops from day to day and to make sure the daily average is equal to the target specific gravity. You also decide that because of the continuous data available, the best choice in control charts for this operation is a 6σ process control chart.

Making the Process Control Chart

The first thing to do in constructing a process control chart is to estimate the variability of the data using the standard deviation and the average value **for data when the process is at steady state**. As mentioned in a previous chapter, the standard deviation is a measure of how far away the overall set of data is from the mean of the same data. It cannot be emphasized enough that the mean and standard deviation to set the sigma values for the initial process control chart needs to be collected (or estimated) when the system is at steady state. That means no samples can be collected during that time, no changes are taking place, and the system is not in flux, say for example, heating or cooling. The goal of SPC programs is to maintain control such that the smallest standard deviation possible is achieved.

NOTE: There are a couple exceptions to setting the mean and standard deviation using steady-state data and that is when 1) there is a minimum product threshold to meet and a maximum about of acceptable variability and 2) when there is enough process data to create a chart.

Scenario 1 is common in many industries and the alcohol proof example we are discussing is one case of this. If a manufacturer commits to a certain level, like 60 proof, that product needs to contain a minimum of that. Other examples would be 12 oz of soda in a can, 50% cotton, 500 sheets of paper in a pack, as well as the amount of medicine in prescriptions. In these cases, the average would need to be set slightly higher than the stated value since lower variations need to still meet the stated requirements. That is, one could not sell Amoxicillin (an antibiotic) that claims to have 250 mg per tablet in it, but really only contains 200 mg because that may not be effective in killing the bacteria. The pharmaceutical industry is one that has usually has tight sigmas (i.e., process control) as product variability can have significant impacts compared to a soda manufacturer that if cans are produced with an extra milliliter or two, will have little impact.

Scenario 2 happens when you have collected enough steady state data while the process is running that you process control chart may be modified. For example, initially you may determine that the standard deviation for a process is 1 mg/tablet, but with process improvements, the standard deviation decreased to 0.75 mg/tablet. The process control chart could then be modified for the improved process.

By applying the 2σ, 4σ, 6σ areas to the data in Figure 3.3, Figure 3.4 is generated.
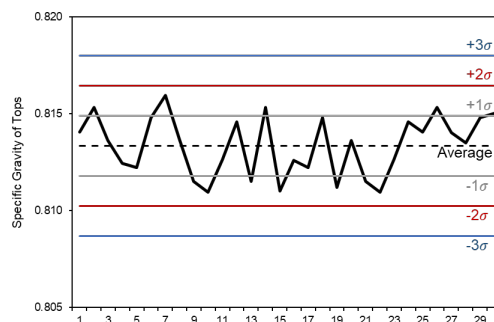
Figure 3.4: *Daily average still tops specific gravities for one month with standard deviations indicated.*

To simplify Figure 3.4, our process control chart can be colored by 4σ (green areas) and 6σ areas (green and yellow areas combined), which is shown in Figure 3.5.
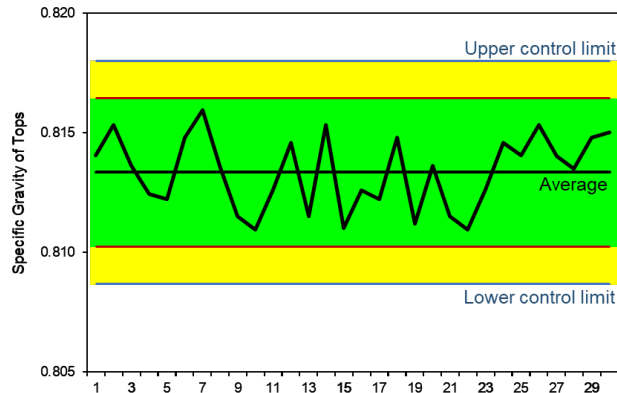
Figure 3.5: *Control chart with 6σ regions shaded for daily average specific gravity data.*

This process control chart is now ready for use in statistical process control; however, in this case, the sample should be at least 0.8316 specific gravity. A discussion should take place as to how to improve the process in order to meet this specification and have all variability above this limit. For other processes, as time goes on and more data is collected, the standard deviation ranges might be modified.

Using a process control chart with accurately defined limits (that is, σs) helps with process control because changes should only be made to the process when it goes out of control; else you risk adding more variability into the system making unnecessary process changes.

Since the purpose of the process control chart is to help the operator decide when the process is statistically "in control" or statistically "out of control," there needs to be criteria to trigger making a process change. The terms "in control" or "out of control" are usually defined by the locations of individual data points or trends of a few or several data points. A process is typically considered out of control if:

- One point lies outside LCL and UCL (6σ area)
- Two out of three consecutive points lie outside the 4σ area
- Four out of five consecutive points lie beyond 1σ on the same side of the center line
- Eight consecutive points lie on one side of the center line
- Five consecutive points trending in any one direction
- Other patterns that trend with time

If any of these scenarios occur, process changes are considered and typically made in order to get the process back within control. However, considerations do need to be made as to how the process control chart was made. For example, perhaps the process is new and the data was not "fine-tuned" enough when producing the limits (typically a minimum of 20 sequential points are used to create the chart) or the product specification is not achievable given the process.

Example #1: Bullet Speeds All Data
At a bullet manufacturing facility, bullets are fired from a gun with the velocity measured at 10 ft. from the muzzle to determine if the bullet lot meets quality control. In order to determine when the process should be changed, a process control chart needs to be developed. Using the data in Table 3.1, we will create a process control chart that can be used for future quality control testing.

Table 3.1: *The velocity of bullets 10 ft from the muzzle for 56 bullet lots.*

Test #	Velocity (ft/s)	Test #	Velocity (ft/s)	Test #	Velocity (ft/s)	Test #	Velocity (ft/s)
1	2958	15	3016	29	3015	43	3042
2	2970	16	3071	30	3022	44	3048
3	2979	17	3003	31	3019	45	3028
4	2992	18	3060	32	3101	46	3041
5	2979	19	2997	33	3137	47	3041
6	2966	20	2907	34	3118	48	3002
7	3037	21	2941	35	3110	49	3053
8	3070	22	2972	36	3133	50	3078
9	3008	23	2968	37	3135	51	3033
10	3013	24	3094	38	2984	52	3009
11	3004	25	3014	39	2998	53	3003
12	3001	26	3079	40	2989	54	3056
13	3000	27	3056	41	3008	55	2993
14	3022	28	3022	42	2980	56	3030

When the data is plotted on a scatter plot, a decent scatter in the data can be seen (Figure 3.6). From Table 3.1, the mean is 3025 ft/s with a standard deviation of 49.3 ft/s.
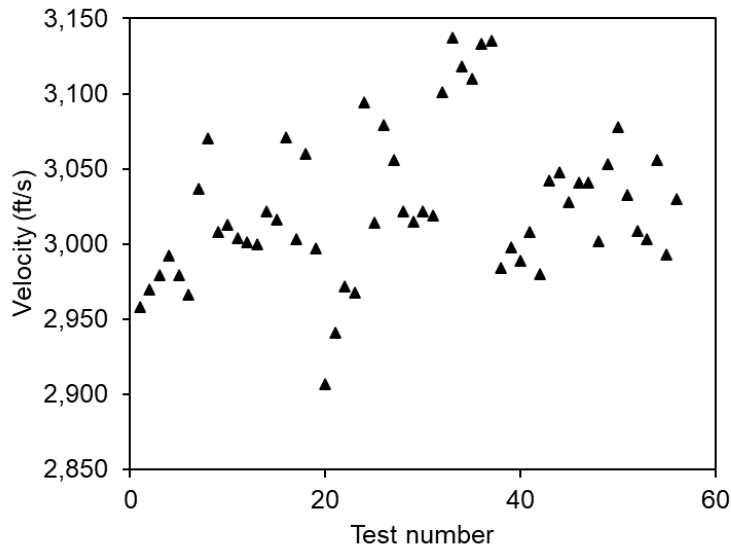
Figure 3.6: *The velocity of bullets 10 ft from the muzzle for 56 bullet lots.*

Using the average and the standard deviation, we can create the 4σ and 6σ data regions for a process control chart. In this case, Figure 3.7 is generated with the 4σ being +/-98.5 ft/s and 6σ data regions being +/-147.8 ft/s. Notice that there are no outliers past 6σ and that is because all of the data was used in determining the average and standard deviation of the data. When creating a process control chart, it is particularly important to have good baseline data that represents the process when it is in control which means it is important to exclude non-steady-state data when determining the average and standard deviation for the control chart.
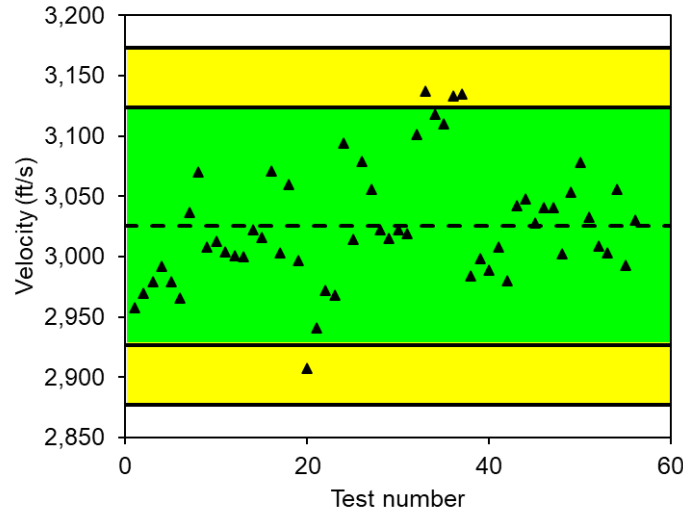
Figure 3.7: Process control chart based on the data in Table 3.1 to produce bullets.

Example #2: Bullet Speeds with Narrower Specifications

For this example, we assume the data in Table 3.1 included bullets produced during process upsets, while the process was starting up, and during shutdown. Due to this, only the middle 28 data points from Table 3.1 should be used for the process control chart as those were the bullets produced while the process was at steady state. Those 28 points have a new average and standard deviation, which resulted in an average of 3019 ft/s with a standard deviation of 16.5 ft/s, which is much narrower than the 49.3 ft/s from when all of the data was considered. In this case, the 4σ region is +/-32.9 ft/s and 6σ data regions are +/-49.4 ft/s that results in Figure 3.8 after all the data is plotted on the chart.
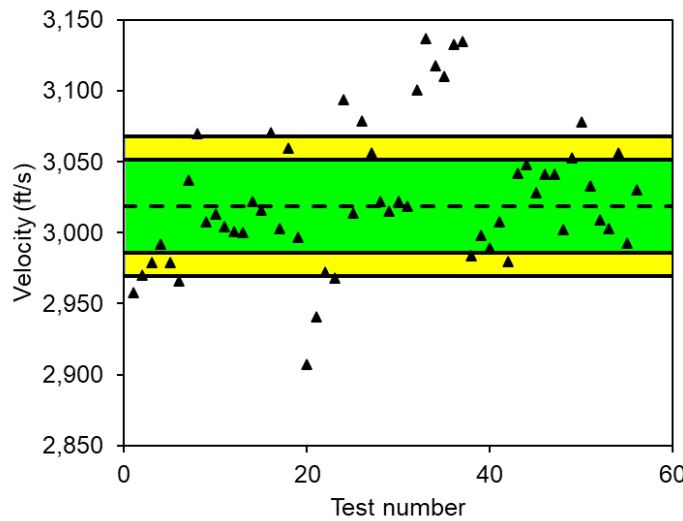
Figure 3.8: Process control chart based on the data in Table 3.1 to produce bullets with only the middle 28 data points used for the average and standard deviation.

In Figure 3.8, we can see that several points are out of control as the upper and lower control limits are narrower than in Figure 3.7. Using the out-of-control guidelines given previously, several issues are seen:
- One point lies outside LCL and UCL (6σ area)
  - This happens 12 times with several other points right on the control limit lines including, but not limited to, test numbers 20, 21, 24, 26, 32-37. These are the points that fall in the white regions above and below the 6σ lines.

18

- Two out of three consecutive points lie outside the 4σ area
  - These would be 2 out of 3 points within the yellow region, which happens for test numbers 2-4, 15-17, and several other times. Note that this could be one point in the lower 6σ region, a point within the 4σ region, and another point in the upper 6σ region. This potentially indicates that a process change was made that overcorrected the issue.
- Four out of five consecutive points lie beyond 1σ on the same side of the center line
  - Points 2-6 show this trend.
- Eight consecutive points lie on one side of the center line
  - This does not occur on this control chart.
- Five consecutive points trending in any one direction
  - The most obvious case of this, is for tests 32-37, which are also above the upper control limit line.

If Figure 3.8 was the process control chart, as an engineer you should be concerned with how often the process is out of control. Either the quality control specifications need to be reconsidered because the process cannot make the product within the current specifications, or the process needs to be improved to reduce the amount of variation in the product.

Process control charts are an extremely useful tool for production processes as they can provide a lot of insight as to the manufacturing process and how changes to the process affect product specifications.

4. Linear Regression

PHILLIP RUSSELL AND STEPHANIE G. WETTSTEIN

The goal of linear regression is to determine if one variable can predict another. The independent (also called the predictor) variable is the variable we are basing our prediction on (x) while the dependent (also called the criterion) variable is the one we are predicting (y). When there is one independent variable, one dependent variable, and a straight line is the best fitting line for the data, the prediction method is called linear regression.

Performing Linear Regression

The data shown in Table 4.1 is the diameter of a circle (x) and the measured circumference of the circle (y) in centimeters. When plotted against each other and a linear trendline is added, the black diagonal line shown in Figure 4.1a is known as the regression line.

Table 4.1: *Diameter (x) and circumference (y) data in centimeters for regression plot.*

Point	x (Predictor variable)	y (Criterion variable)
1	2	6.2
2	4	11
3	6	18.5
4	8	25.6
5	10	28.4

The regression equation of the line consists of the predicted score on y for each possible value of x that is determined in the form of

$$y = mx + b \tag{4.1}$$

where m is the slope of the trendline and b is the intercept. The best fit equation is formed by minimizing the sum of the squared errors of prediction. In the example shown in Figure 4.1 on the left, the linear relationship is in the form of

$$y = mx \tag{4.2}$$

because since the line was forced through (0, 0), there is no intercept since at a diameter of zero, the circumference of a circle is zero.
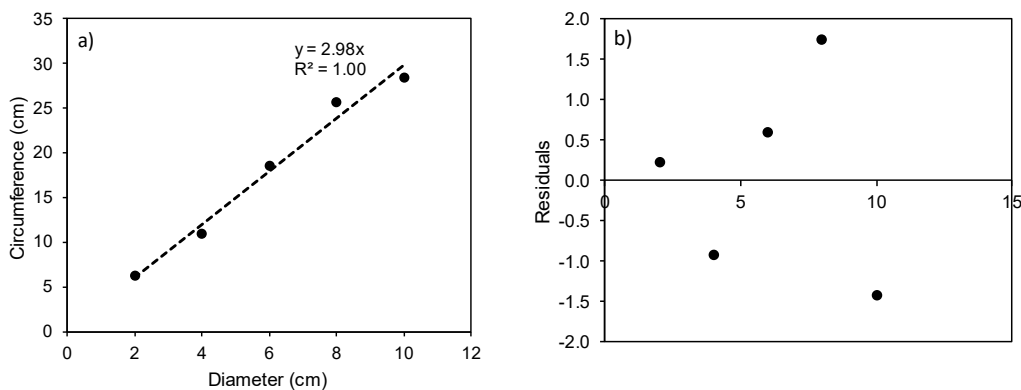
Figure 4.1: *Using the example data in Table 4.1, a) scatter plot and the b) residuals.*

Note that forcing the regression line through (0, 0) should only be done if when the independent value is 0, the dependent variable will always be 0.

Therefore, the relationship between diameter (d) and circumference (C) for the data in Table 4.1 was determined to be

$$C = 2.98d \tag{4.3}$$

This means that multiplying a given x (the diameter of a circle) by 2.98, will result in the circumference of the circle based on the collected data. However, the actual relationship between diameter and circumference is:

$$C = \pi d \tag{4.4}$$

There is clearly error in the data and we have not yet statistically determined if diameter can predict circumference; however, using Excel's Data Analysis tools for regression, we can test if the dependence has statistical significance, which will be discussed later in this chapter. But before we do that, we need to check to ensure that the residuals from the prediction are random and not systematic. If there is systematic error, patterns in the residuals will be apparent indicating a different fit (e.g., polynomial, logarithmic, etc.) may be better and that linear regression should not be used.

To check for systematic error, one can plot the residuals of the experimental data as compared to the predicted data. The residuals represent the error of prediction, which can be visualized by vertical lines drawn from the experimental points to the regression line and is the difference between the actual y value and the predicted y value. The closer the point is to the regression line, the smaller the error of prediction. For example, point 1 in Figure 4.1a at a diameter of 2 cm is near the regression line and therefore, its error of prediction is small compared to the 4th point for a diameter of 8 cm. We can see this on the residual plots (Figure 4.1b), where points 4 and 5 have higher residuals than points 1-3. Even so, Figure 4.1b looks like random error. If the residuals follow a trend (e.g., slowly increasing or decreasing through a range), a different regression line (e.g., perhaps a polynomial fit) may be a better option. In the case of the data in Table 4.1, the linear regression is a good fit for the data set.

Excel Tutorials: Performing Regression Analysis Using Excel

Consider an experiment where conductance of an aqueous salt solution is measured over time. Salt is continuously diffusing into the solution in a controlled manner. This increases the conductance, and the rate of increase is essentially proportional to the rate of diffusion of the salt into the solution. The data for an experimental run is shown in Figure 4.2 and listed in Figure 4.3. (Note that it would be best to include this data as a table, but a screenshot was taken to include the formulas, so it is a figure.) It is desired to determine the slope of the linear regression line for this data set.

The slope of the trendline can be found in multiple ways below but note that method 3 is the preferred method. That is because the Regression Analysis discussed in method 3, like the other two methods, calculates the regression coefficients, but additionally provides
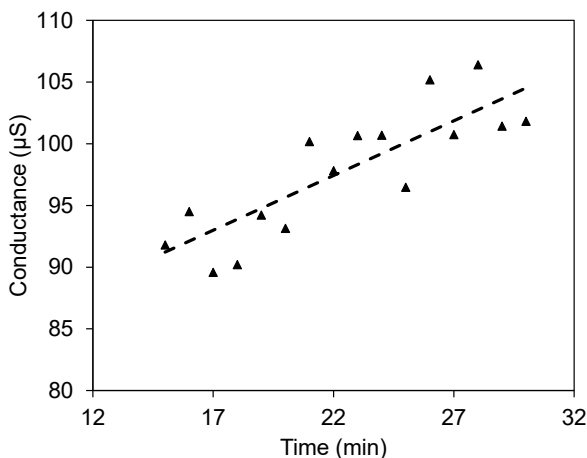
Figure 4.2: Data for a liquid diffusion experimental run.

information on the uncertainty of the regression. That is, method 3 can answer the question: "is the dependent variable truly dependent on x?"

Method 1: Using Excel Formulas

You can use the built-in Excel functions of "=SLOPE(y-range, x-range)" and "=INTERCEPT(y-range, x-range)" as seen in Figure 4.3. The second row from the bottom has the value of the slope (0.909) on the left and the equation used to obtain it on the right. The bottom row contains the information for the intercept (77.3). NOTE: You cannot set the intercept to zero using this method. Therefore, the equation for the regression line would be

$$y = 0.909x + 77.3 \tag{4.5}$$

Time (min)	Conductance (microS)
x	y
15	91.76
16	94.51
17	89.59
18	90.21
19	94.23
20	93.16
21	100.19
22	97.83
23	100.67
24	100.7
25	96.48
26	105.19
27	100.75
28	106.41
29	101.43
30	101.82
0.909455882352941	=SLOPE(R60:R75,Q60:Q75)
77.3453676470589	=INTERCEPT(R60:R75,Q60:Q75)

Figure 4.3: Data for a liquid diffusion experimental run.

Method 2: Using Excel Trendlines

Another option that most engineering undergraduates are familiar with is to plot the independent variable against the dependent variable and adding a trendline. This is done in Excel by right clicking the data set in a figure that and click "Add trendline" (Figure 4.4).
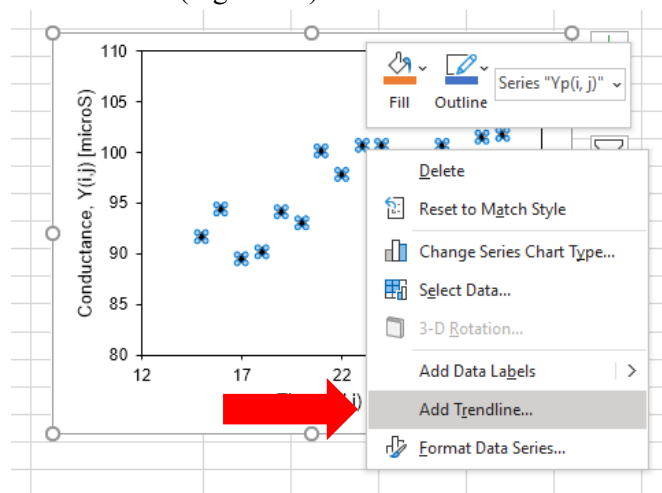
Figure 4.4: Adding a trendline using the chart in Excel.

A sidebar appears (Figure 4.5) that allow you to select options for the trendline. This includes the type of trendline (in this case, linear), if the y-intercept should be set to a value (if this is chosen, it is usually set to 0; red arrow), and whether to display of the equation and R^2 value of the line on the figure (blue arrow).
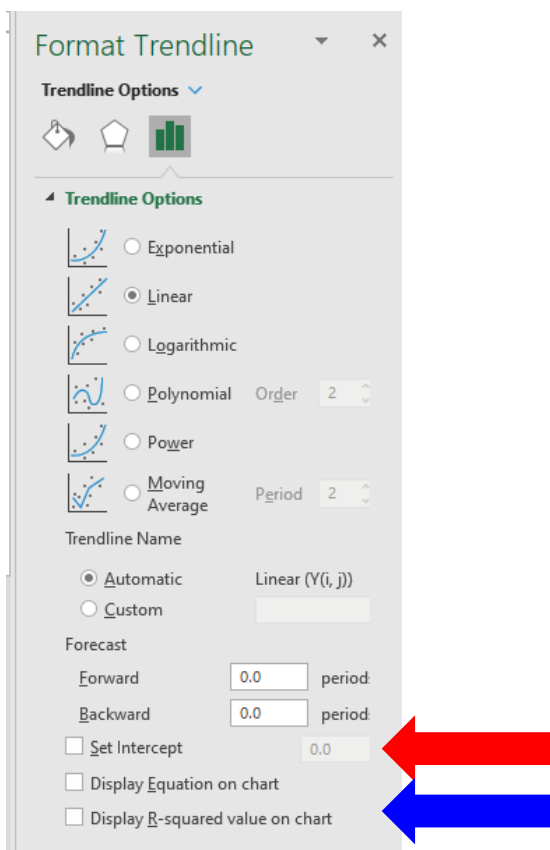
Figure 4.5: Options for a trendline.

For Figure 4.6, "Display Equation on chart" was selected and linear trendlines were added to each data set. In this case, the slope of the line is 0.909 (with a y-intercept of 77.3). These are the same results that we got in method 1, which is expected.
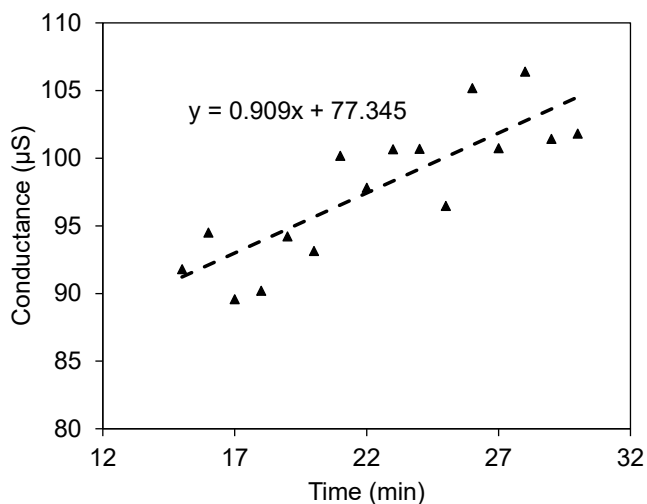
Figure 4.6: Data for an experimental liquid diffusion run and the linear trendline and trendline equation.

To change the significant figures in the trendline equation, you can double-click inside the formula box and manually change it, or you can click on the box and in the sidebar (Figure 4.7), select "Number" as the Category and then select the number of decimal places to display.
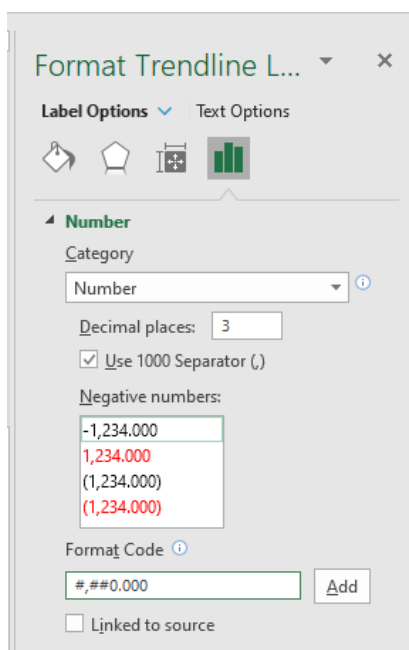
Figure 4.7: *Sidebar for formatting the trendline equation.*

NOTE: The number of decimal places is not the same as the number of significant figures. In this example, Figure 4.6 shows that the slope has three significant figures, but the intercept has five due to values in the ones and tens positions (i.e., 77). *Be sure the significant figures are correct for both values if presenting the equation in a report.* This means you may have to add a textbox and manually enter the equation onto the figure which was done for Figure 4.8 and displays the correct number of significant figures.
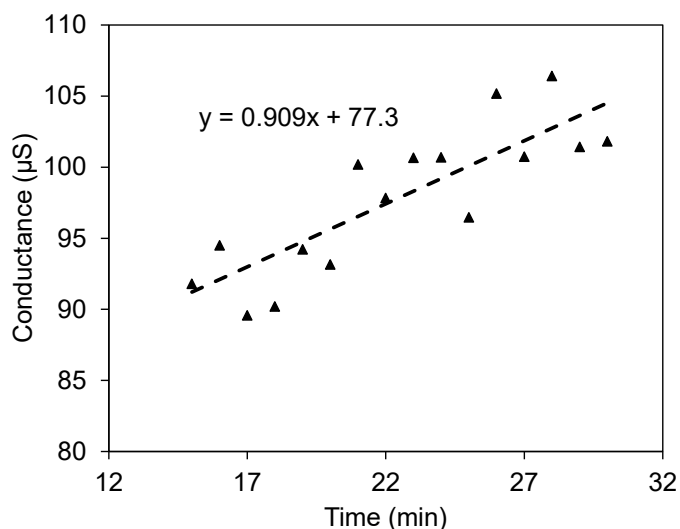
Figure 4.8: *Data for an experimental liquid diffusion run and the linear trendlines and trendline equation with a textbox for the equation adder to have the correct number of significant figures.*

Method 3: Using Excel Data Analysis
Select the "Data Analysis" toolpak add-in listed under the "Data" tab of Excel (Figure 4.9).
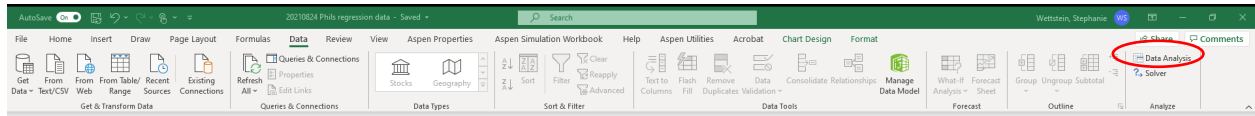
Figure 4.9: *Data Analysis location for regression analysis in Excel.*

From the "Data Analysis" menu (Figure 4.9), select "Regression" and click "OK" (Figure 4.10).
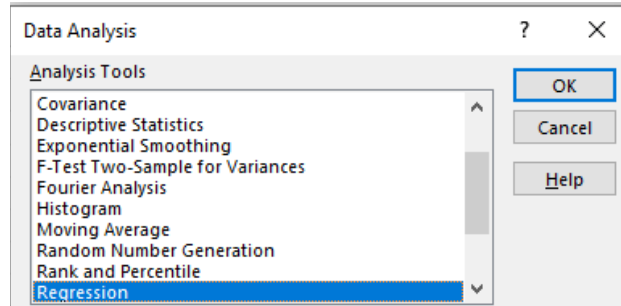
Figure 4.10: *Regression selection.*

Then, your data needs to be selected for Excel to perform the regression. The dialogue box requires the Y input range, the X input range, confidence level box to be checked, the output location, and a checkmark in the "Residual Plots" box so that the residuals can be checked to ensure they are random. Figure 4.11 shows the selections for the liquid diffusion data example.
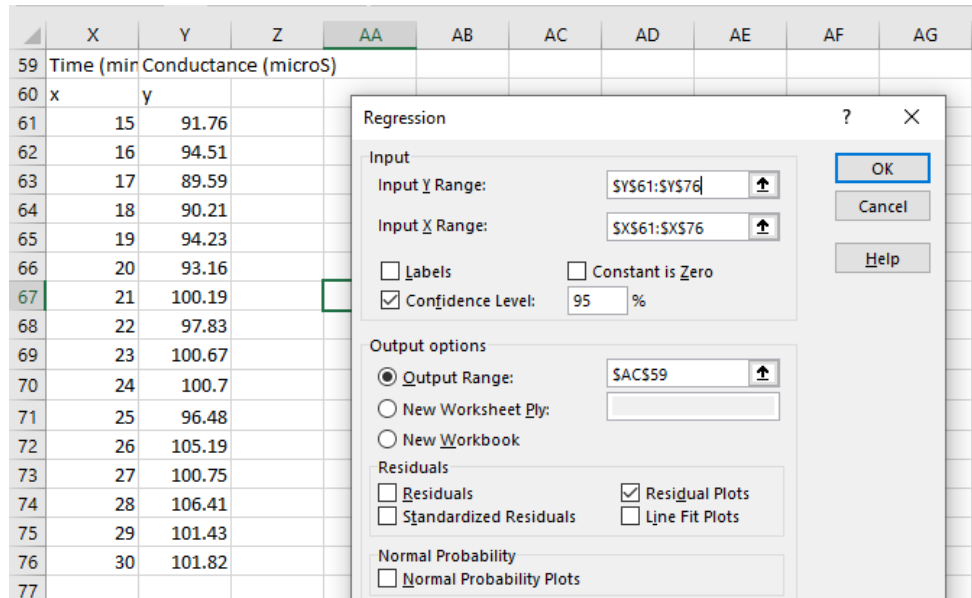
Figure 4.11: *Regression dialogue box selections for the example.*

To use the regression dialogue box (Figure 4.11):
1) The **Input Y Range** is for the dependent variable, in this example, the conductance.
2) The **Input X Range** is for the independent variable, in this example, time. The column headings were not included in the input ranges so the **Labels** box was not checked. If you have column headers and include them in the highlighted data, the **Labels** box should be checked.

25

3) For relationships where you expect the y-intercept to be zero, that is, the curve must go through the point (0, 0), the **Constant is Zero** box should be checked.

   a) For example, say you started with five plants at the same height and want to know if how much you water them affects the plant growing height. Since the plants do not start at 0 height, you would not click that box. Same if you want to correlate the number of white blood cells present after a drug treatment. Since the body already contains white blood cells even with no drug present, you would not check the **Constant is Zero** box.

   b) If you are seeing how tall seeds grow with watering and all are 0" high on the first day, then you would check the **Constant is Zero** box. Additionally, if you are measuring the amount of a drug not normally found in a person's blood, you would set the constant to zero since with no drug, the concentration should be zero.

   c) Even if you expect a particular value, you should look at your data and see if it makes sense. For example, you may not expect a concentration of chemical A in chemical B, but due to a low-quality distillation, chemical A is present without adding it separately.

4) Checking the **Confidence Level** box is not required unless you want to change the confidence level from the default value of 95%; however, checking the box is fine as seen in Figure 4.11.

5) Indicate where the **Output** table should be placed. Typically, it is good to have all the analysis on the same worksheet ply (*i.e.*, worksheet space), so choosing "Output Range:" and the cell it should start is preferred. However, ensure there is enough space for the analysis tables so they do not overwrite data. Selecting "New Worksheet Ply" eliminates the need to make space and you could always copy and paste the data after into the existing worksheet ply.

6) The **residual plots** box should be checked to see if the residuals of the data (experimental data compared to the data predicted by the trendline) are random. If the residuals are **not** randomly scattered that means there are trends in the data indicating a different trendline fit (e.g., exponential, polynomials, etc.) may be a better option than a linear fit.

Several output tables are produced from the analysis under the heading "Summary Output," which are shown in Figure 4.12. The cells described below are based on the grid shown in Figure 4.12.

1) The first table, "Regression Statistics," provides information about the R-value and error.

   a) Multiple R: This is the square root of R squared,

   b) R Square: This is what most people think of when they think of the fits of trendlines (R^2) and is also known as the coefficient of determination; be aware that there are many reasons that you may get a high R Square (1 or close to 1), but the fit is not good,

   c) Adjusted R Square: This value is used if you have more than one independent (x) variable

   d) Standard Error: This is the standard error of *the regression* and is *not* the standard error in descriptive statistics (also note that this is *not* the standard deviation),

   e) Observations: The number of samples.

For this example, the R^2 is 0.71 (cell AI6), which certainly indicates some variability in the data points. This table also provides the number of observations (in this case, 16; cell AI9), which you should compare to the number of data points you wanted to analyze to ensure you selected the data properly.

2) The second and third tables fall under the header "ANOVA." The second table will not be used, but the third one will.

   a) Coefficients: Excel uses a least squares estimate, which minimizes the vertical distance from the data points to the regression line and provides the slope (m) and intercept (b) of the regression line

   b) Standard Error: Least squares estimate of the standard error. Note that the standard error is NOT the same as the standard deviation (eqn. 1.5). The standard error is calculated by equation 1.8.

   c) t-Stat: Provides the t-statistic for testing the hypotheses; the greater the t, the more likely your value is significantly different from the comparison data.

d) P-value: P-value for the hypothesis test of the data; this tells you the odds that your results could have happened by chance.
e) Lower 95%: Lower bound of the confidence interval.
f) Upper 95%: Upper bound of the confidence interval.

The "Coefficients" column lists the y-intercept value (in this case, 77.35; cell AI18) and the "x variable 1," which is the slope of the regression line. In this case, the slope is 0.909 (cell AI19). As expected, these values are the same as we calculated with methods 1 and 2. (If they do not match, you likely have an error in the analysis or forced the intercept through 0 for an analysis.) In cells AJ18 and AJ19, the standard error of the coefficients are given. The p-values listed in AL18 and AL19 are the important statistical values needed to determine if there is a significant dependence of y on x. In this case the p-values are both much smaller than 0.05, which means that the conductance is dependent on time and that the intercept is necessary to explain the relationship. The 95% confidence intervals for the slope and intercept are given in columns AM and AN. For this example, you are 95% confident that the true value of the slope is between 0.57 and 1.24.

3) The final data table is the "Residual Output" and contains the predicted y value based on the trendline equation and the residual between the predicted and actual values.
4) Figure 4.12 shows that the residuals listed in column AJ are random and that the linear trendline is a good trendline option for this data. If there are trends in the residuals (e.g., the first half of the data set has positive residuals while the second half has negative residuals), this indicates that the trendline may need to be different (e.g., exponential, polynomial, etc.) or that there may be systematic error in the data (e.g., baseline drift of the signal with time).
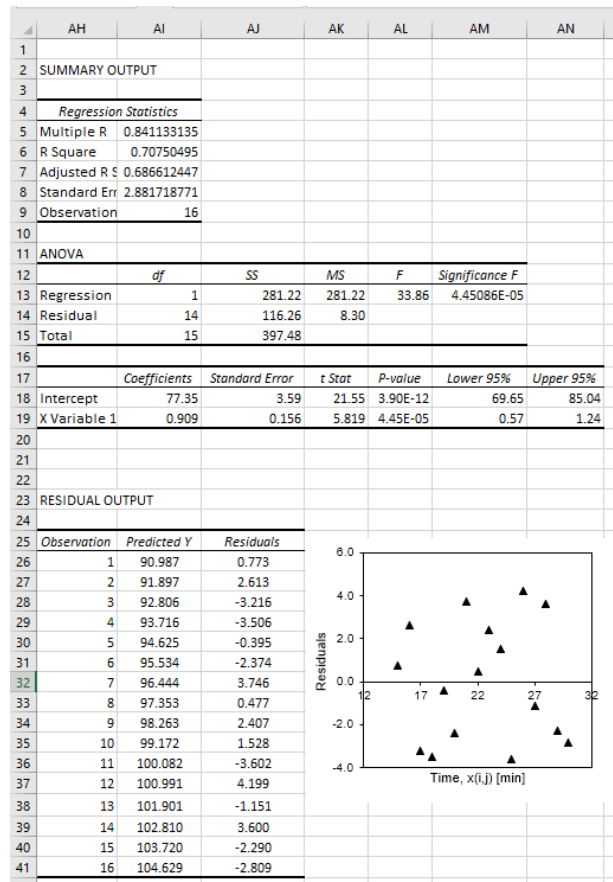
Figure 4.12: *Regression outputs for the example.*

Example: Regression Analysis of Diameter Data

Going back to the diameter/circumference example, if the Regression analysis is run using Excel on the data in Table 4.1, Tables 4.2 and 4.3 result.

Table 4.2: Regression statistics that result from Excel's Regression analysis tool for the data in Table 4.1.

Regression Statistics	
Multiple R	0.998
R Square	0.997
Adjusted R Square	0.747
Standard Error	1.259
Observations	5

Table 4.3: Regression statistical analysis that result from Excel's Regression analysis tool.

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	0	#N/A	#N/A	#N/A	#N/A	#N/A
Variable 1 (Diameter)	2.983	0.085	35.129	3.92E-06	2.747	3.218

The coefficient for the intercept is zero since for diameter versus circumference, if the diameter is zero, the circumference would be zero. For the x-variable, the coefficient is 2.98, which is what resulted in Figure 4.1a. Again, this means that multiplying x (the diameter) by 2.98 will result in y, which is the circumference in this case. The additional data provided in Table 4.3 regarding the p-value and confidence intervals determine the if y really depends on x. In this case, the p-value is 3.92×10^{-6}, which indicates the chances of the circumference depending on the diameter due to random chance is small since it is much lower than 0.05. Therefore, the circumference is dependent on diameter. Also, the confidence interval (Lower 95% to Upper 95% from Table 4.3) means that we are 95% confident that the true *population* mean (not the sample mean) falls between 2.747-3.218. We know that the true value of π is 3.14. (That is, if we were to measure the diameter and circumference of every circle in the World, the value of π would be 3.14.) Since 3.14 falls within the 95% confidence interval, the data set is good given the number of samples. If the confidence interval was, say 2.74-3.08 and did not encapsulate the population mean of 3.14, we would want to do further analysis to see if there is a point (or several points) that are outliers or collect more data to improve our confidence interval.

5. Statistical Comparison of Two Regressed Slopes

PHILLIP RUSSELL AND STEPHANIE G. WETTSTEIN

In certain situations, it would likely be beneficial to compare one regressed slope to another to see if they are statistically different or not. One example would be in determining the diffusion coefficient of a salt, which is proportional to the slope of a conductance versus time plot for each run. It may be desired to compare the resulting slopes to see if they are statistically different or not to determine equipment repeatability using same salt, determine if two different pieces of equipment give the same results using the same salt, or to compare the slopes of different salts on the same piece of equipment. Another example would be if you want to compare the performance of multiple fuel cells, the ohmic slope for each cell could be statistically compared to see if there is a significant difference between the cells.

> NOTE: Only one variable can change to properly compare regressed slopes! If the equipment changes, the salt has to be the same; if the salt changes, the equipment has to be the same. Also, most regression comparisons have time on the x-axis.

In situations where you get linear data, you cannot run a two-sampled t-test as you have multiple variables changing at once and therefore, need to perform Regression analysis. In the case of Figure 5.1, we ran two experiments (A1 and A2) that consisted of taking conductance measurements at different times. Here we are evaluating repeatability since we want to compare two runs that have multiple time points, a t-test (see Chapter 6) will not work due to two variables changing and therefore, we need to compare the two regressed slopes.
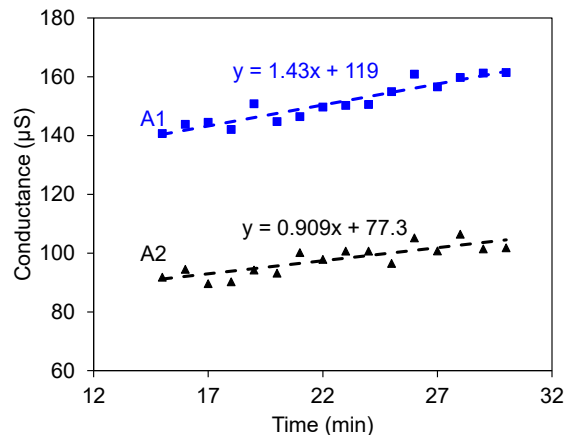
Figure 5.1: *Conductance data for two experimental runs and the linear regression fits for run A1 (■) and run A2 (▲) with the linear trendlines and equations for each data set.*

General Approach for Comparing Two Linearly Regressed Slopes
The null hypothesis is that the two regressed slopes are equal, which is what we are going to test. (More details about developing hypotheses is presented in the 'Hypotheses" section in Chapter 6.) In order to test the hypothesis and compare the two slopes, we need to calculate the key statistical variable t_{stat} using:

$$t_{stat} = \frac{\left|b_{1,1} - b_{1,2}\right| \sqrt{DOF}}{\sqrt{SSE \left[\frac{1}{\sum_{j=1}^{N_1}(x_{1,j} - \bar{x}_1)^2} + \frac{1}{\sum_{j=1}^{N_2}(x_{2,j} - \bar{x}_2)^2}\right]}} \tag{5.1}$$

where $b_{1,1}$ is the slope of the first line, $b_{1,2}$ is the slope of the second line, DOF are the degrees of freedom, N is the number of data points, SSE is the summed squared errors, x is the independent variable, and $\bar{x}$ is the average of the independent variable. The larger t_{stat} is, the more likely it is that the two regressed slopes are statistically different since the larger the difference in the two slopes (the numerator) the bigger t_{stat} will be. More details on the different terms are given here:

$\left|b_{1,1} - b_{1,2}\right|$ – absolute difference in the value of the two regressed slopes. The more different these slopes are, the more likely it is that the slopes will be statistically different. If the two slopes are identical (i.e., the lines are parallel), then it will always be determined that the slopes are statistically not different.

$\sqrt{DOF}$ – degrees of freedom correspond to the number of data points available for analysis. More data points will increase t_{stat} and increase the ability to differentiate the two slopes as statistically different. For two data sets, the equation for DOF is

$$DOF = N_1 + N_2 - 4 \tag{5.2}$$

where N_1 is the number of data points in data set 1 and N_2 is the number of data points in data set 2.

$\sqrt{SSE}$ – the sum of the squared errors corresponds to the amount of "*noise*" in the data. Noisy data reduces confidence in the true slope of the data being accurately represented by the regressed value.

$\sqrt{\left[\dfrac{1}{\sum_{j=1}^{N_1}(x_{1,j}-\bar{x}_1)^2} + \dfrac{1}{\sum_{j=1}^{N_2}(x_{2,j}-\bar{x}_2)^2}\right]}$ – This term represents the "dispersion" along the independent variable.

The sums in the denominator get larger as more data points that are "far" from the average value of x are included. This means that if you are taking data over time, the longer the time interval, the larger the denominator sum. To visualize this, imagine receiving the task of accurately drawing the slope of a line passing through two points if one of the points is located at $x = 10$ and the other point is located at $x = 12$ (the red data points and red lines showing the maximum and minimum estimated slopes shown in Figure 5.2). The diameter of the markers represents the uncertainty in the measured value, which are equal. The minimum slope is determined by drawing a line that is tangent to the upper portion of the first data point and is also tangent to the lower part of the second data point. The maximum slope is determined by drawing a line that is tangent to the lower portion of the first data point and is also tangent to the upper portion of the second data point. Quantitatively, $\bar{x} = 11$ and

$$\sum_{j=1}^{N_1}(x_{1,j} - \bar{x}_1)^2 = (10 - 11)^2 + (12 - 11)^2 = 2 \tag{5.3}$$
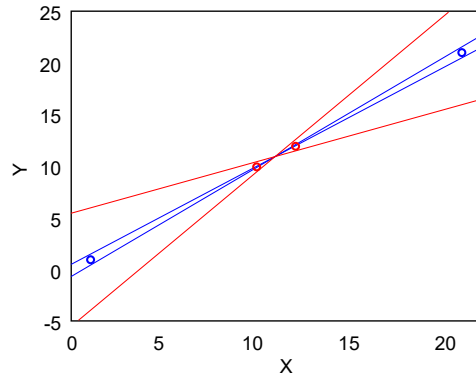
Figure 5.2: *Comparison of uncertainty in estimated slopes for data with different dispersions in the independent variable. Red data points are located at x=10 and x=12. Blue data points are located at x=1 and x=21. Colored lines are the maximum and minimum slopes around the data.*

Compare this with drawing a line through two points (all data points are assumed to have the same uncertainty in y values) where one of the points is located at $x = 1.0$ and the other point is located at $x = 21$ (see the blue data points and blue lines in Figure 5.2). The slope for the widely dispersed data can be much more precisely estimated than the slope for the (red) data which are narrowly dispersed. Quantitatively, $\bar{x} = 11$ and

$$\sum_{j=1}^{N_1} (x_{1,j} - \bar{x}_1)^2 = (1 - 11)^2 + (21 - 11)^2 = 200 \tag{5.4}$$

Both the blue and red points have the same "uncertainty" (diameter of the marker), but because the blue data (wide dispersion) are more widely dispersed, the slope can be estimated with far less uncertainty than for the red (narrow dispersion) case.

This is something to consider when gathering data as *collecting data at longer time intervals, if possible, will lead to better statistical results*. That is, if an experiment is at steady-state for 1 h, collecting 10 samples for 10 min, once a minute will result in less dispersion than collecting samples for the entire hour, once every 6 min. The latter would result in better statistical results.

$t_{critical}$ – The value of t_{crit} "draws the line" for determining if the slopes are statistically "different" or "not different." If $t_{stat} > t_{crit}$, the null hypothesis is rejected, and **the two slopes are statistically different**. If $t_{stat} < t_{crit}$, the null hypothesis is accepted, and **the two slopes are statistically not different**.

Note that t_{crit} is a function of α, which is related to the level of certainty, and DOF, which is the number of degrees of freedom related to the number of data points available for use in the regression analysis. A value of 0.05 for α is commonly used in statistical analyses and implies that the null hypothesis is rejected 5 % of the time when it is in fact true. If rejecting the null hypothesis 20% of the time when it is if fact true is acceptable, α would be set to 0.20. The impact on quantitative results for t_{crit} at different values of α and DOF is shown in Table 5.1. The more data that is available (high values of DOF), the smaller t_{crit} becomes and this makes "passing" the test for statistical similarity less likely, which is better for statistical analysis.

Table 5.1: *Values of t_{crit} at different values of α and DOF.*

α	DOF	t_{crit} $(= TINV(\alpha, DOF)$
0.05 (95% confidence level)	6	2.45
0.05 (95% confidence level)	28	2.05
0.20 (80% confidence level)	6	1.44
0.20 (80% confidence level)	28	1.31

Analysis Steps

The key question is "are these slopes statistically different or not different?" To compare two regressed slopes, the following calculations need to be made:

1. Perform a linear regression on each data set to determine the slopes.
2. Calculate the degrees of freedom with
$$DOF = N_1 + N_2 - 4 \tag{5.2}$$
3. Calculate the SSE, which is the *sum of squared estimate of errors*, using

$$SSE = \sum_{i=1}^{2} \sum_{j=1}^{N_j} \left(y_{i,j} - y_{p_{i,j}}\right)^2 \tag{5.6}$$

where $y_{i,j}$ is the experimental data and $y_{p_{i,j}}$ is the predicted value from the regression line.

4. Determine $\bar{x}_i$ for the data sets with

$$\bar{x}_i = \frac{\left(\sum_{j=1}^{N_1} x_{i,j}\right)}{N_i} \tag{5.7}$$

5. Then calculate

$$\sum_{j=1}^{N_1}(x_{i,j} - \bar{x}_i)^2 \tag{5.3}$$

6. With this information, calculate t_{stat} using eqn 5.1, which is shown here again

$$t_{stat} = \frac{|b_{1,1} - b_{1,2}| \sqrt{DOF}}{\sqrt{SSE\left[\frac{1}{\sum_{j=1}^{N_1}(x_{1,j} - \bar{x}_1)^2} + \frac{1}{\sum_{j=1}^{N_2}(x_{2,j} - \bar{x}_2)^2}\right]}} \tag{5.1}$$

7. Determine t_{crit} from Excel ("=TINV(α, DOF)") or a statistics table.
8. Complete the hypothesis test noting:
   a. If $t_{stat} > t_{crit}$ – Reject the null hypothesis that the slopes are equal ($b_{1,1} = b_{2,1}$) (or, "the slopes are statistically different")
   b. If $t_{stat} \leq t_{crit}$ – Fail to reject the null hypothesis (or "the slopes are not statistically different")

Example: Two Regressed Slopes that are Statistically Different
Consider the data presented in Figure 5.3, which was from an experiment where conductance of an aqueous salt solution is measured over time. (This is the same data presented in Figure 5.1 without the regression lines.) Salt is continuously diffusing into the solution in a controlled manner. This increases the conductance, and the rate of increase is essentially proportional to the rate of diffusion of the salt into the solution. The data for two separate experimental runs are shown in Figure 5.3 and listed in Table 5.2 (at the end of the section). It is desired to statistically compare the slopes of the regression lines for these data sets to see if the experiment is repeatable.
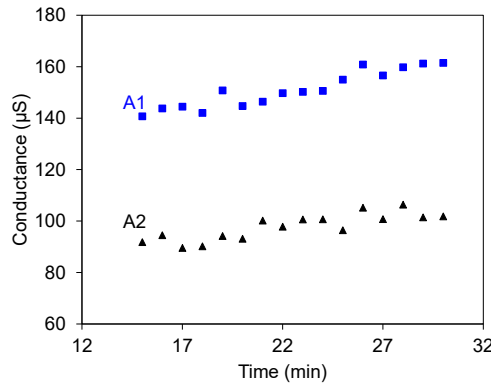
Figure 5.3: Conductance data for two experimental runs (data set A) run A1 (■) and run A2 (▲)

Excel's linear regression tool can be used to determine the best slope and intercept for each run, which is given by:

$$y_p = b_{1,i}\,x + b_{0,i} \tag{5.5}$$

where y_p is the predicted value (i.e., dependent variable), $b_{1,i}$ is the slope of the regression curve for the ith run, x is the independent variable, and $b_{0,i}$ is the y-intercept of the regression curve for the ith run. Note

that this is the same equation as 4.1, but with different variable notation. In this regression example, y_p is the conductance in μS and x is time in minutes.

As mentioned in Chapter 4, when adding a trendline to data, the residuals should always be checked to make sure that they are random, which indicates a good fit of the data. Trends in residuals (e.g., the first half of the data set has positive residuals while the second half has negative residuals) indicate that the trendline may need to be different (e.g., exponential, polynomial, etc.). In this case, Figure 5.4 shows that the residuals are random for both data sets, meaning a linear trendline is the correct shape.
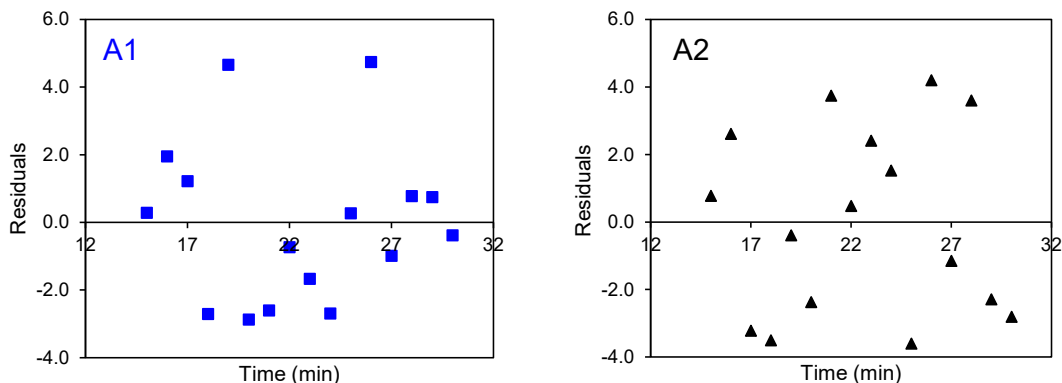
Figure 5.4: Residuals for a linear trendline for run 1 (■) on left and run 2 (▲) on right.

NOTES on chart uniformity: When two graphs are plotting the same data and next to each other, keeping the scales of the axes the same helps the reader compare the data. In this case, Figure 5.4 on the left defaulted to a max y-axis value of 6.0. Since the left figure had a max of 6.0, the y-axis on the left was changed from a max of 5.0 to 6.0 so that an easier comparison can be made.

Also note that in order to make it easier for the reader to understand which residuals go with each data set, the marker colors and shapes are consistent with the ones in Figure 5.3 and other related figures, such as Figure 5.5. Additionally, inserting the marker shapes in the caption (use "Insert-> Symbol) and then using the same color coding (highlight the symbol and change the font color) makes the data much easier to interpret and follow. Step-by-step directions on how to do this can be found in the appendix.

Once the trendlines are added to the data in Figure 5.3, Figure 5.5 results.
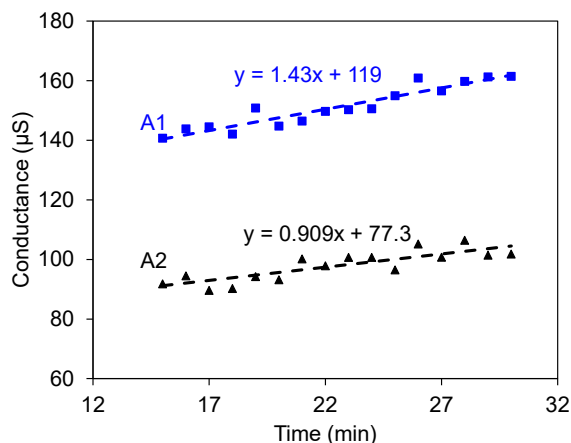
Figure 5.5: Conductance data for two experimental runs and the linear regression fits for run A1 (■) and run A2 (▲) and the linear trendlines and trendline equations for each data set are shown.

33

NOTE on significant figures: As previously mentioned, be aware that the default number of significant figures for regression lines may be more than required for the data. A general guideline is that you should present the same number of significant figures as the data allows or plus one. Since Table 5.2 has only two significant figures for time, data should be presented with a maximum of three significant figures as shown in Figure 5.5. The trendline equations had to be manually added using a text box in order to make this work due to the decimal places. (That is, using Excel's default of "3 decimal places" resulted in six significant figures for A1's intercept and five for A2's, which is incorrect.)

Example Analysis: Regression Slope Comparison

Applying the analysis steps to the data found in Table 5.2 results in:

1) Runs 1 and 2 have slopes of 1.43 ($b_{1,1}$) and 0.909 ($b_{1,2}$), respectively, as seen in Figure 5.5.
2) Both runs 1 and 2 have 28 degrees of freedom as calculated by:
$$DOF = 16 + 16 - 4 = 28$$
3) The SSE equals 200.4 (summing values from the sixth column in Table 5.2).
4) $\bar{x}_i$ equals 22.5 for run 1 and run 2 (averaging values from the third column in Table 5.2).
5) The data for Equation 5.3 are in the last column and equal to 340 for both run 1 and run 2 (from the seventh column in Table 5.2).
6) The t_{stat} then equals:
$$t_{stat} = \frac{|1.430 - 0.909| \sqrt{28}}{\sqrt{(200.4)\left[\frac{1}{340} + \frac{1}{340}\right]}} = 2.54$$
7) For this data set, $\alpha = 0.05$ and $DOF = 28$ so to determine t_{crit} in Excel, "=TINV(0.05, 28)" was typed in to a cell and 2.05 results.

Result

$t_{stat} > t_{crit}$ (2.54 is greater than 2.05) so the null hypothesis is rejected, and the *two regressed slopes* are found to be *statistically different at an alpha value of 0.05*. The experimental data have little noise, and the regressed slopes fit the data well. The excellent "*goodness of fit*" reduces the uncertainties associated with the slopes and this makes it easier to statistically differentiate between the two.

Example Two Regressed Slopes that are Not Statistically Different

In the previous example, data set A came from an experiment that had little noise in the data. A second data set with more noise is shown overlapping data set A in Figure 5.6 and is denoted with X's. "More noise" can be visually seen by the markers being further away from the linear regression line. Data with more noise will increase the uncertainties in the two regressed slopes and makes it more difficult to determine that the regressed slopes are statistically different. The statistical analysis will be repeated to determine if the regressed slopes from the two experiments in data set B regressed slopes.
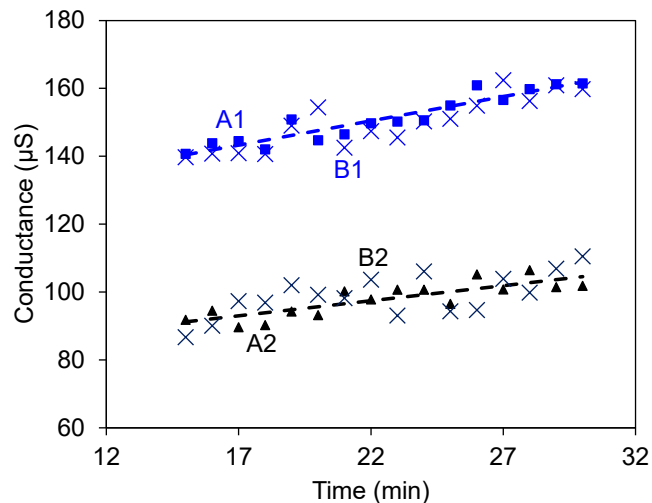
Figure 5.6: Data for four liquid diffusion experimental runs (data set A and B) and the linear regression fits for data set A: run A1 (■) and run A2 (▲) and data set B: run B1 (×) and run B2 (×). The regression lines are nearly identical for run 1 and run 2 from both data sets and therefore, overlap.

<u>Analysis</u>

Using equation 5.4, the slope and the intercept of each run can be determined for data set B (Table 5.3; end of chapter). Data set B has significantly more noise than data set A as seen by comparing the values in the sixth column of squared errors in Tables 5.2 and 5.3. To answer the question of whether the regressed slopes for Run 1 and Run 2 for the "B" data are statistically different or not different, the same steps outlined in the previous section will be repeated:

1) The linear regression slopes from each run are $b_{1,B,1} = 1.433$ and $b_{1,B,2} = 0.9067$. (Note that both data sets have nearly identical values of the slopes and intercepts.)
2) Degrees of freedom are the same for both data sets (Equation 5.2), which is 28.
3) Using equation 5.6, the SSE is 544.5 for data set B. This is the term where noise or lack of quality of fit occurs. For data set A, the sum of squared estimates of errors was only 200.4. The "noise" in data set B is more than double data set A.
4) Since the number of runs and $x_{i,j}$ are the same for data set A and B, $\bar{x}_{B1}$ and $\bar{x}_{B2}$ both equal 22.5 (see the third column in Table 5.3).
5) Since $x_{i,j}$ and $\bar{x}_i$ are the same for data set A and B, these sums are also 340 (see the seventh columns in Table 5.3).
6) Using equation 5.1, the t_{stat} is 1.605.
7) Determine t_{crit} with $\alpha = 0.05$ and $DOF = 28$ from *Excel:* =TINV(0.05, 28) = 2.05
8) Complete the hypothesis test: In this example, $t_{stat} \leq t_{crit}$ (1.605 is less than 2.05), so we fail to reject the null hypothesis, or, said another way, *the slopes are not statistically different*.

<u>Results</u>

This is the opposite result from what we obtained for data set A, even though the regressed slopes and intercepts for the regression fits were essentially identical. The SSE from data set B had sufficient noise to force $t_{stat} < t_{crit}$. This meant that the two regressed slopes from data set B were statistically not different while in data set A, the two slopes could be differentiated and it was determined that they were statistically different because their data had significantly less noise. Even a relatively modest decrease in the quality of the data reduces the ability to see differences in calculated quantities like regressed slopes. This reinforces the need to do experiments carefully to minimize the errors in the experimental data we collect.

Table 5.2: Conductance and time data from two runs (data set A).

Expt ID i	Run # j	Time $x_{i,j}$	Conductance $y_{i,j}$	Predicted conductance $y_{p_{i,j}}$	Squared error $(y - y_p)^2$	$(x_{i,j} - \bar{x}_i)^2$
A1	1	15	140.7	140.4	0.09	56.25
A1	2	16	143.8	141.9	3.61	42.25
A1	3	17	144.5	143.3	1.44	30.25
A1	4	18	142	144.7	7.29	30.25
A1	5	19	150.8	146.2	21.16	12.25
A1	6	20	144.7	147.6	8.41	6.25
A1	7	21	146.4	149	6.76	2.25
A1	8	22	149.7	150.4	0.49	0.25
A1	9	23	150.2	151.9	2.89	0.25
A1	10	24	150.6	153.3	7.29	2.25
A1	11	25	155	154.7	0.09	6.25
A1	12	26	160.9	156.2	22.09	12.25
A1	13	27	156.6	157.6	1.00	20.25
A1	14	28	159.8	159	0.64	30.25
A1	15	29	161.2	160.25	0.90	42.25
A1	16	30	161.5	161.9	0.16	56.25
		AVE: 22.5				SUM: 340
A2	1	15	91.76	91.21	0.30	56.25
A2	2	16	94.51	92.1	5.81	42.25
A2	3	17	89.59	92.99	11.56	30.25
A2	4	18	90.21	93.88	13.47	20.25
A2	5	19	94.23	94.76	0.28	12.25
A2	6	20	93.16	95.65	6.20	6.25
A2	7	21	100.19	96.54	13.32	2.25
A2	8	22	97.83	97.43	0.16	0.25
A2	9	23	100.67	98.31	5.57	0.25
A2	10	24	100.7	99.2	2.25	2.25
A2	11	25	96.48	100.09	13.03	6.25
A2	12	26	105.19	100.98	17.72	12.25
A2	13	27	100.75	101.86	1.23	20.25
A2	14	28	106.41	102.75	13.40	30.25
A2	15	29	101.43	103.64	4.88	42.25
A2	16	30	101.82	104.53	7.34	56.25
		AVE: 22.5		SUM OF ALL	200.4	SUM: 340

$x_{i,j}$ is the time (min) for a given experiment ID (i) and run number (j), $y_{i,j}$ is the experimental value of the conductance (μS), $y_{p_{i,j}}$ is the predicted value of the conductance (μS), $\bar{x}_{i,j}$ is the average time for all runs for a set of experiments.

Expt ID i	Run # j	Time $x_{i,j}$	Conductance $y_{i,j}$	Predicted conductance $y_{p_{i,j}}$	Squared error $(y - y_p)^2$	$(x_{i,j} - \bar{x}_i)^2$
B1	1	15	139.7	139	0.4404	56.25
B1	2	16	140.8	140.5	0.09978	42.25
B1	3	17	140.9	141.9	1.0645	30.25
B1	4	18	140.6	143.3	7.2952	30.25
B1	5	19	149	144.8	18.2254	12.25
B1	6	20	154.4	146.2	66.7305	6.25
B1	7	21	142.5	147.6	25.8827	2.25
B1	8	22	147.4	149.1	2.7239	0.25
B1	9	23	145.6	150.5	24.4766	0.25
B1	10	24	150.3	151.9	2.6693	2.25
B1	11	25	151	153.4	5.5877	6.25
B1	12	26	154.9	154.8	0.0135	12.25
B1	13	27	162.4	156.2	38.0109	20.25
B1	14	28	156.3	157.7	1.8544	30.25
B1	15	29	160.9	159.1	3.2599	42.25
B1	16	30	159.8	160.5	0.5288	56.25
		AVE: 22.5				SUM: 340
B2	1	15	86.7	92.2	30.1459	56.25
B2	2	16	90	93.1	9.4134	42.25
B2	3	17	97.3	94	11.0748	30.25
B2	4	18	96.9	94.9	3.8277	20.25
B2	5	19	102	95.8	38.2267	12.25
B2	6	20	99.2	96.7	6.0288	6.25
B2	7	21	98.2	97.6	0.2937	2.25
B2	8	22	103.6	98.5	25.6018	0.25
B2	9	23	93.1	99.4	40.6133	0.25
B2	10	24	106	100.4	32.4812	2.25
B2	11	25	94.3	101.3	47.9631	6.25
B2	12	26	93.7	102.2	71.6382	12.25
B2	13	27	104.9	103.1	3.3467	20.25
B2	14	28	99.5	104	20.0461	30.25
B2	15	29	107.9	104.9	9.0965	42.25
B2	16	30	110.5	105.8	22.367	56.25
		AVE: 22.5		SUM OF ALL	544.5	SUM: 340

$x_{i,j}$ is the time (min) for a given experiment ID (i) and run number (j), $y_{i,j}$ is the experimental value of the conductance (μS), $y_{p_{i,j}}$ is the predicted value of the conductance (μS), $\bar{x}_{i,j}$ is the average time for all runs for a set of experiments.

6. t-Tests

There are many applications of t-tests, which determine whether data sets/theoretical values are significantly different or not, in research and in industry. Is your product within quality specifications? Did a manufacturing change really result in a "New and Improved" product claim? Was my test score significantly different than the rest of the class? A t-test can be used to answer these questions and many more. In general terms, the questions that can be answered using a t-test are:

- Is one experimental data set significantly different than another?
- Is my experimental result consistent with a published value?

In order to answer these questions, it is imperative that the data be collected in the correct manner. For t-tests, only one variable can be changed between data sets. For example, if I want to compare the pressure drop of two different packing materials in a packed bed, the material is the variable that changed; therefore, the temperature and pressure of the surroundings and the air flow rate need to be the same for each of the packed beds. If I want to compare the pressure drop at two different flow rates, I can only do that if the temperature and pressure of the surroundings and the packing material is the same. Only one variable can change and there can only be two levels. That is, I cannot determine if data collected at 50, 60, and 70 sccm are equal with a t-test, I can only compare two of the flow rates to each other. It is critical that you set up your hypotheses and data collection properly in order to get analyses that are correct.

There are several different tests that can be performed:

Difference between a t-test and z-test: You use a t-test when the sample size is small (less than 50) and the population standard deviation is unknown. You use a z-test with large sample sizes (greater than 50) or the population variance is known.

Difference between paired and unpaired t-tests: This text will not discuss paired t-tests, which occur when a data point in one data set is associated with a data point in another set. For example, say we wanted to test the null hypothesis that the front left paw and the front right paw of a bear are of equal size. This is trying to test that the variation in the data set is because there is a difference between left and right paws; however, we also need to consider that bears are different sizes and that paw size is likely correlated to bear size. Therefore, we need to "pair" the left and right paws of the same bear since they would be associated, which is why we would need to use a paired t-test. An unpaired t-test means that data points are assumed to be independent from other data points.

Difference between a one-tail and two-tail t-test: A *two-tailed t-test* tests in both directions of the mean to see if the data are significantly difference. For example, was the mean test score 80%? The two-tailed test will determine if the mean test score was equal to 80%, and if not, the mean test score could have been above or below 80%. You use a *one-tailed test* if you only want to test in one direction of the mean. For example, was the mean test score greater or equal to 80%? You could also test if the mean test score was less than or equal to 80%, but you cannot test both with a one-tailed t-test; that would require a two-tailed test. Setting up the hypothesis is where the difference between two-tailed and one-tailed t-tests become clear (Table 6.1).

Difference between an equal and unequal variance t-test: The choice between an equal and unequal variance t-test is based on the variance of the data you are comparing with the t-test. *If you are comparing a set of experimental data to a single data point, such as a published or theoretical value*, the variance on the published/theoretical value will be 0 (as it is a single data point). Therefore, an unequal t-test should always be selected in this case. *In the case of comparing two sets of experimental data*, each set will have a variance associated with it that can be determined using *Descriptive Statistics.* Then,

- If the variances are within one order of magnitude, assume equal variances and use the "t-Test: Two-Sample Assuming Equal Variances" analysis tool in Excel.

- If the variances differ by more than an order of magnitude (10 times different from each other), than assume unequal variances and use the "t-Test: Two-Sample Assuming Unequal Variances" analysis tool in Excel.

If the variances are close to being an order of magnitude different, run both an unequal and equal variance t-test to see if it makes a difference on whether you accept or reject the null hypotheses. Chances are you will get the same results regardless of the t-test in this case.

Selecting which t-test to run is important as well as several other steps. From Introductory Statistics (Holmes, 2022), which is an excellent open-source resource you can access, to perform a hypothesis test a statistician will:

1) Set up two contradictory hypotheses and plan the data collection
2) Collect sample data
3) Analyze sample data by performing the correct t-test
4) Interpret the data
5) Write a meaningful results and discussion section (More to come starting in Chapter 7)

1) Hypotheses

First, you need to develop the null and test hypotheses. The null hypothesis states that there is no difference between the data sets while the test hypothesis (also known as the alternative hypothesis) states the opposite of the null hypothesis. As seen in Table 6.1, it depends which type of t-test you perform as to what your null and test hypotheses are. In general, the null hypothesis always has a symbol with an equal in it (=, ≤, or ≥) and the test hypothesis never has a symbol with an equal in it.

Table 6.1: Operators for the hypotheses in t-tests.

	Null (H_o)	Test (H_1)
Two-tailed	Equal to*	Not Equal to
One-tailed	Greater than or equal to*	Less than
One-tailed	Less than or equal to*	Greater than

*NOTE: "Equal to" really means "not significantly different"

Be cautious when selecting between the two-tailed and one-tailed tests. Initially the one-tailed test may seem to be the correct test, when in reality, the two-tailed test is correct. For example, if the null hypothesis is that the new distillation column produces an ethanol purity that is equal to the ethanol purity of the old column, it may be tempting to want to use a one-tailed test to see if the ethanol purity of the new column is greater than or equal to the old column. However, using the one-sided test fails to test for the possibility that the new column produces an ethanol purity less than the old column. In this case, a two-tailed test is more appropriate since we just want to determine if the ethanol purity is equal to the old column regardless of if it is higher or lower.

2) Sample Collection and Determining Alpha

Once you determine your hypotheses, you will then need to collect data samples. It is important to note that in order to complete a t-test, some assumptions about the data need to be made:

- The data should be a random sample – Consider this when planning your sample collections
- The data is assumed to be from a normally distributed population – That is, if you took enough samples, the results would be normally distributed
- The sample standard deviation can be used to approximate the population standard deviation
- The samples are independent and that each data point in one sample set is not associated with another data point in the other data set. If they are associated, then you need to use a paired t-test.

When considering sample collection, keep in mind that the more samples you have, the better your test results will be. You need a minimum of three data points *at a given condition* (ideally, at least five) to complete a t-test and that the data should also be examined for outliers prior to running the t-tests.

Now that you have gathered the data, in order to run a t-test you need to select the level of significance (α), which is the probability of rejecting the null hypothesis when it is true. In other words, an alpha of 0.05 would mean there is a 5% risk of concluding that a difference exists when there is no actual difference. Therefore, decreasing alpha will result in a greater chance of accepting the null hypothesis (i.e., the data sets are not significantly different). For a two-sided test, alpha will be divided equally on both ends of data (Figure 6.1a) while for a one-side test, alpha will only lie on one side of the data (Figure 6.1b).
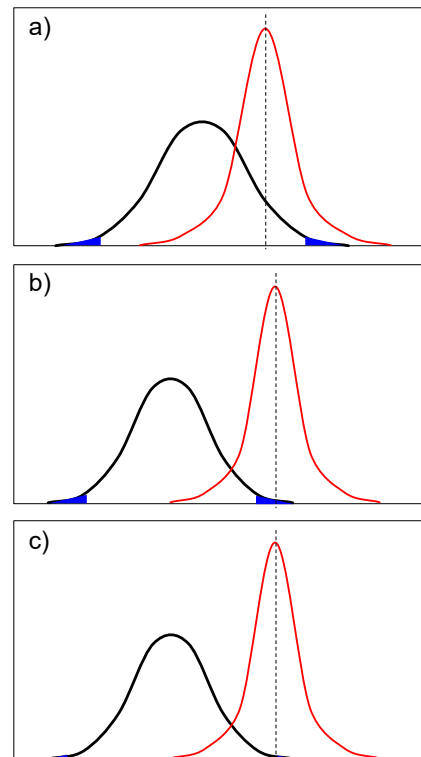
The alpha is an important variable that can be seen visually in Figure 6.2. Each graph has two, normally distributed data sets plotted on it with the mean of data set 2 represented by a dashed line and the area encompassed by the alpha region shaded in blue on data set 1. The alpha for Figure 6.2a and 6.2b are 0.05 while for Figure 6.2c, it is 0.01. In Figure 6.2a the data sets are not significantly different at an alpha of 0.05. In Figure 6.2b, the data sets would be determined to be significantly different at an alpha of 0.05 even though they are not (this is a Type I error, which will not be discussed in this text). However, using the same data sets, but reducing the alpha to 0.01 (Figure 6.2c) results in the data sets being not significantly different. If no alpha is given, it is common to use 0.05 and accept that there is a 5% chance of concluding that a different exists when there is no actual difference. Once the alpha value is chosen, a t-test can be run.

3) Performing a t-test using Excel
Consider the two randomly generated data sets listed in Table 6.2. Due to their means, we can see that samples 1a and 1b differ by 0.4 in set 1, and in set 2, the means differ by 2.2. Both of these differences may seem significant, but when the high standard deviation is considered, the means now look to not be statistically different. We can use Excel to run t-tests that can determine if the data sets are significantly different or not.

Figure 6.1: Normally distributed data with the alpha shaded in blue where a) represents a two-sided t-test and b) a one-sided t-test in which you are testing if the data set is greater than the other.

Figure 6.2: Normally distributed data with the p-value shaded in blue where a) has an alpha of 0.05 and not significantly different (accept the null hypothesis), b) shows data sets that are significantly different (reject the null hypothesis) at an alpha of 0.05, but c) are not significantly different (accept the null hypothesis) at an alpha of 0.01.

Table 6.2: *Data Set 1 contains two distributions with close means and Set 2 contains two distributions with means that are more different. (Shading is done for referencing for Figure 6.5.)*

|	Set 1: Similar			Set 2: Dissimilar	
Sample	1a	1b		2a	2b
	5.3	4.7		3.6	4.2
	5.5	6.2		3.4	6.2
	6.1	6.1		3.6	6.1
	6.7	5.7		3.5	6
	4.9	5.1		3.1	5.2
	4.5	6.1		3.2	6.1
	3.8	5.8		4.1	6.3
mean:	5.3	5.7		3.5	5.7
std. dev.:	0.97	0.57		0.33	0.77

Excel's Data Analysis tool is used to perform t-tests and we first need to determine what type of *t-test* we should run: *t-Test: Two-Sample Assuming Equal Variances* or a *t-Test: Two-Sample Assuming Unequal Variances*. A general guideline is to use an unequal variance t-test when the variance of one data set is an order of magnitude or more different than the other data set. If you are unsure, you could run the t-test both ways to see if the p-value (and therefore, the result) changes.

NOTE: Performing a t-test in Excel uses exactly the same steps for both equal and unequal variances and therefore, only the equal variance case is shown. The steps are the same if *t-test: Two-Sample Assuming Unequal Variances* is selected other than the selection from the Data Analysis pop-up.

Using the **Descriptive Statistics** method described in chapter 1, the variances are:
- 1a: 0.946
- 1b: 0.322
- 2a: 0.107
- 2b: 0.586

Since the variances are **within an order of magnitude of each other** (10 times), a *t-test: Two-Sample Assuming Equal Variances* should be run. From the *Data Analysis* dialog box (Figure 6.3) select "t-Test: Two-Sample Assuming Equal Variance" to open the t-test dialog box shown in Figure 6.4.
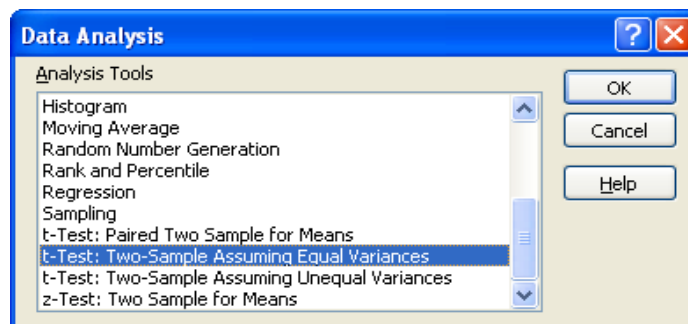
Figure 6.3: *Select "t-Test: Two-Sample Assuming Equal Variances" from the Data Analysis dialogue box.*

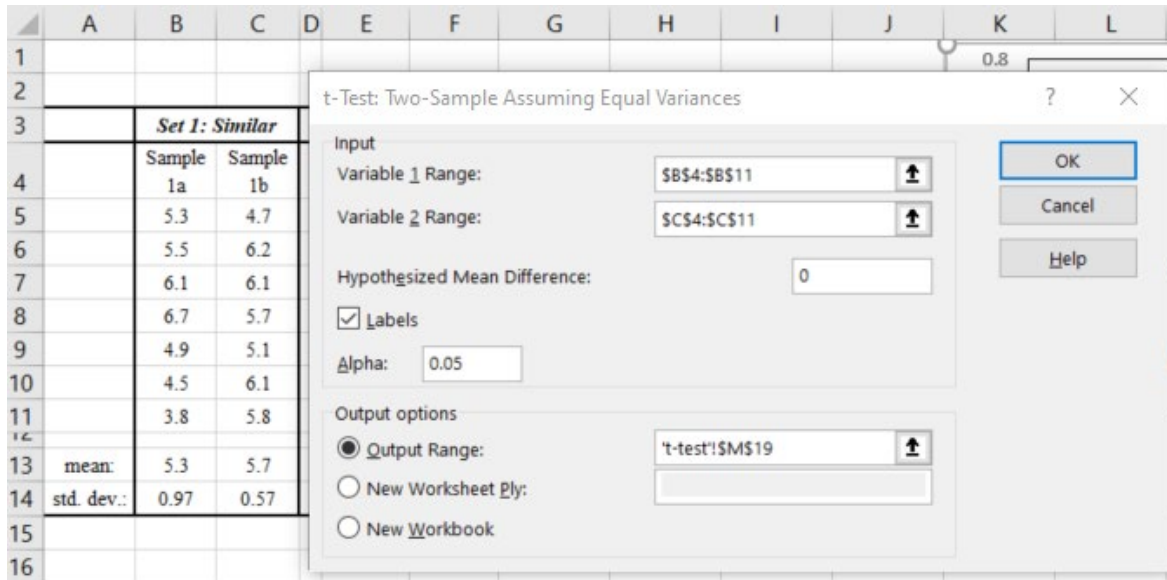|	A	B	C	D E	F	G	H	I	J	K	L
1										0.8	
2				t-Test: Two-Sample Assuming Equal Variances						? ✕	
3		*Set 1: Similar*		Input							
4		Sample 1a	Sample 1b	Variable 1 Range:			$B$4:$B$11	⬆		OK	
5		5.3	4.7	Variable 2 Range:			$C$4:$C$11	⬆		Cancel	
6		5.5	6.2							Help	
7		6.1	6.1	Hypothesized Mean Difference:				0			
8		6.7	5.7	☑ Labels							
9		4.9	5.1	Alpha: 0.05							
10		4.5	6.1								
11		3.8	5.8	Output options							
12				⊙ Output Range:			't-test'!$M$19	⬆			
13	mean:	5.3	5.7	○ New Worksheet Ply:							
14	std. dev.:	0.97	0.57								
15				○ New Workbook							
16											

Figure 6.4: t-Test: Two-Sample Assuming Equal Variance dialogue box.

To use the t-test dialogue box (Figure 6.4)
- The **Variable 1 Range** is the first set of data. Do not highlight the mean, standard deviation, or anything other than the data set and the column header (optional). The column headings were included in the input ranges so the **Labels** box was checked. If you do not have column headers, the **Labels** box should not be checked.
- The **Variable 2 Range** is the second set of data.
- Set the **Hypothesized Mean Difference** to zero to *test* if the samples are "*equivalent*".
- Choose the level of significance (**alpha**). The default is α = 0.05.
- Indicate where the **Output** table should be placed. Typically, it is good to have all the analysis on the same worksheet ply (*i.e.*, worksheet space), so choosing "Output Range:" and the cell it should start is preferred. However, ensure there is enough space for the analysis tables so they do not overwrite data. Selecting "New Worksheet Ply" eliminates the need to make space and you could always copy and paste the data after into the existing worksheet ply.

NOTE: If you test your data against a single data point, like a theoretical value, you will need to perform a *t-test: Two-Sample Assuming Unequal Variances*. Excel needs to calculate the variance, so the single data point needs to be entered into two cells. (That way the variance in the data is 0.

The analysis produces a table with the heading "t-test: Two-Sample Assuming Equal Variances," with the data shown in columns 1-3 in Table 6.3.
- The mean and variance are just the mean and variance of each sample, which should be the same values as given by Descriptive Statistics.
- The p-value is listed as **P(T<=t) two-tail** and is the important statistical values needed to determine if there is a significant difference between the data sets.
  - In the case of the example, the p-value is 0.35, which is larger than 0.05, which tells us that the means of the two sets of data are not significantly different from each other at the 95% confidence level.
- Note that these tables as-is should not be incorporated into technical reports unless you are going to describe and use all of the rows. Only use the data from the tables that are important to support and explain your results in a concise manner.

Table 6.3: *Two t-Test results for similar data (data Set 1) and the dissimilar data set (data Set 2).*

	Sample 1a	Sample 1b	Sample 2a	Sample 2b
Mean	5.26	5.67	3.5	5.73
Variance	0.95	0.32	0.11	0.59
Observations	7	7	7	7
Pooled Variance	0.63		0.346	
Hypothesized Mean Difference	0		0	
df	12		12	
t Stat	-0.97		-7.09	
P(T<=t) one-tail	0.175		6.36E-06	
t Critical one-tail	1.78		1.78	
P(T<=t) two-tail	0.350		1.27E-05	
t Critical two-tail	2.18		2.18	

4) Interpret the Data

Now that you have the analyses, a decision needs to be made. The level of significance (alpha) is critical because that is the value that the t-test results are compared to. After you perform the Excel analysis and generate a p-value (probability), you compare that to the alpha.

If the probability is less than or equal to the level of significance ($p \leq \alpha$):
If the t-test returns a probability less than or equal to alpha, you reject the null hypothesis and conclude that:
- the population means behind the two samples (the new data and the constant) are significantly different or that
- the new data is not consistent with the published value

If the probability is greater than the level of significance ($p > \alpha$):
If the t-test returns a probability greater than alpha you accept the null hypothesis and conclude that:
- the population means are not significantly different or that
- the new data is consistent with the published value

The greater the difference between the p-value and alpha, the more confidence you can have in the decision. For example, if the p-value is 0.6 for one data set, but 0.06 for another, you can have more confidence in not rejecting the null hypotheses for the first data set. The same goes for p-values less than alpha. If the p-value is 0.003 versus 0.045, you can be more confident that the first set is significantly different than you could be with the second data set. When the p-value is close to α, it is likely worth suggesting collection of more data to increase the certainty.

In the case of the example, a t-test samples 1a and 2b resulted in a p-value is 0.35 (columns 2 and 3 of Table 6.3), which is larger than 0.05, which tells us that the means of the two sets of data are *not significantly different from each other at the 95% confidence level*. Using the second set of data (samples 2a and 2b) results in columns 4 and 5 in Table 6.3 and had a p-value of 1.27E-5. Since 1.27E-5 is much less than 0.05, the data sets are indeed *significantly different at an alpha of 0.05*.

It may be helpful to consider a graphical view of these results. Figure 6.5 compares the two distributions in Set 1 with the two distributions in Set 2. The difference in means must be viewed relative to the amount of variability in the distributions. In Set 1 (left side of Figure 6.5), the difference in means is "small" compared the variabilities (width of the distributions) seen in the two curves. In Set 2 (right side of Figure 6.5), the

difference in means is "large" compared to the variability seen in the two curves. This "gut-level" analysis is consistent with the rigorous quantitative results obtained above.
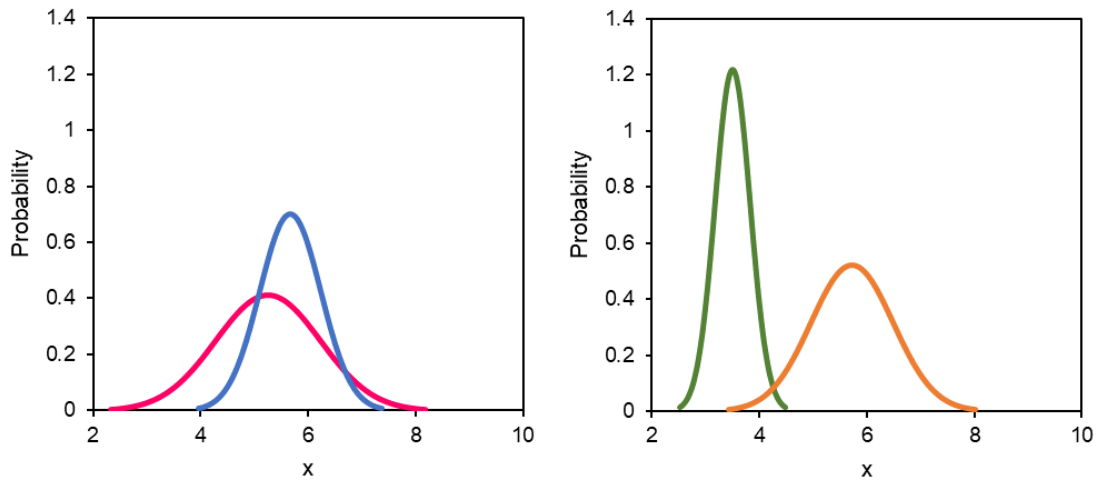
Figure 6.5: *The left figure shows the two distributions in Set 1 that are not statistically different. The right figure shows the two distributions from Set 2 where the means are statistically different at an alpha level of 0.05.*

References

Holmes, A., *Introductory Statistics*, OER Commons, 7 July 2022, https://www.oercommons.org/courses/introductory-statistics-6.
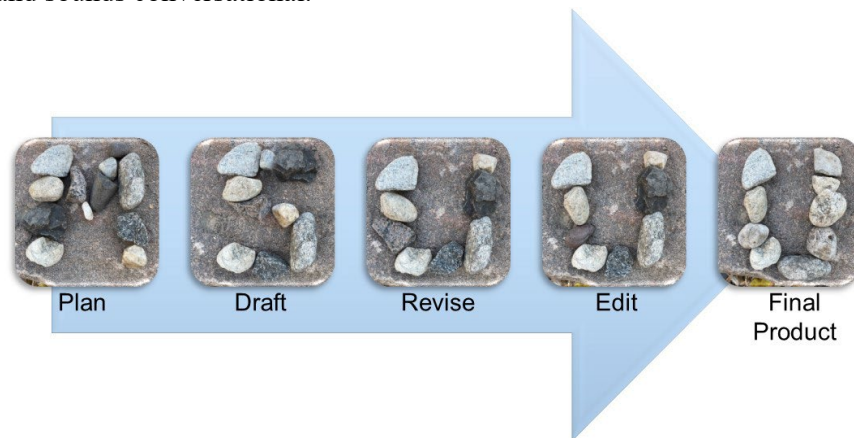
Writing

7. The Writing Process

Writing, especially when compiling a larger document, is not something you sit down, complete in one session, and quickly submit. This is especially true when writing for the workplace where accuracy and clarity are critical. In fact, writing should be seen as an iterative process where the writer moves in and out of various stages of writing and often times revisits some of the stages. The writing process should consist of the following (Figure 7.1):

Planning/Organizing: This is the planning done before writing a document. It may be defining the purpose of the task, determining the audience of the document (i.e., who will read it), sketching/outlining the document and what will go in each section, and/or gathering research.

Drafting: This is writing and compiling a first draft of the document. The writer should focus more on getting ideas down than on punctuation, concision, and flow. Often, writers go back to the planning and organizing step in this phase in order to improve clarity.

Revising: When a writer revises, a writer revisits the draft and makes substantial changes to it. This is more than editing. It is adding, deleting, and moving entire sections of the document around to enhance the flow and accuracy to prepare it as a final, comprehensive document. In fact, it is here that many writers ask others for feedback before revising to ensure that another, unbiased set of eyes have looked over the document and easily understand it. Revising is also a time to ensure the reader has all the information they need to understand the results and that the results are logical and discussed in enough detail.

Editing: This is the final part of the process. It is reading through the document several times while looking for clarity, concision, and grammar. In fact, consider reading the document aloud and listening to it. Most individuals communicate mostly through talking and listening, and therefore, when read aloud, if something in the document does not sound right, it can be corrected. The document should be read in a way that it is understandable and sounds conversational.

Figure 7.1: Steps of a writing process.

8. Team Writing Strategies

Writing is inherently an individual activity. That is, "divide and concur" may be the best plan for the first draft, but teamwork needs to happen for communicating changes, editing the report as a single document (versus individual sections), and making the writing cohesive. Here are some tips to help:

1) Start early. This is critical to have enough time to edit the document.
2) Have regular meetings and prepare for meetings by reading necessary documents, inspecting equipment, and review videos/manuals provided to you. (See Chapter 9.)
3) One person needs to take the lead and set up a meeting. It is best to send out a meeting agenda in order to keep everyone on task.
4) Meet as a group to determine what your objective(s) is/are going to be (see lab objective statement) and discuss who is responsible for what.
   o In the case of writing lab reports, some sections are highly related and will need to be rewritten once the corresponding section is completed to ensure all important information is covered. These sections include:
      ▪ Objective/Results
      ▪ Theory/Results
      ▪ Results/Conclusions
      ▪ Methods of Analysis/Results
   o Other sections, such as Background, Equipment, and Method, are more independent and are easier to work on individually.
   o Discuss any questions you may want to ask the instructor/your supervisor.
   o Set up the next meeting and determine who will facilitate it.
   o The leader of the meeting should send out meeting notes that clearly state member responsibilities and required completion date.
5) Contact the lab instructor if you have questions to set up a meeting (preferred) or email. It is recommended that the group come to the meeting together, if possible. Some instructors may require a meeting.
   o When emailing, it is good practice to copy your entire team on the correspondence so that everyone is on the same page.
6) Set up an electronic file exchange location, such as Microsoft Teams, Google docs, Dropbox, OneDrive, etc., for everyone to work from.

Once the document is put together, **everyone should individually read the document** and provide feedback. Perhaps a team member is not as strong as another in the theory section and in this step, the other people can provide feedback or insert necessary information.

NOTE: Just because someone worked on a section, does NOT mean that no one else should work on it or provide feedback.

9. Meeting agendas and notes

As mentioned, team communication is a critical piece to creating a cohesive document. Meetings help facilitate communication and are important to make sure that everyone involved with the document are on the same page. Meetings are valuable for making decisions, a plan, deciding who does what, and for many other purposes. Everyone has been to a meeting that was not worth their time. In order to increase the chance that a meeting is worthwhile, meeting agendas can help keep everyone on the same page. For project related meetings, an agenda will typically list the time frame, topics, and outcomes, which therefore, increases efficiency. That is, the agenda answers the questions of:

- why am I coming to the meeting?
- what do I need to prepare for? and
- what will be decided at the meeting?

In order to maximize efficiency, post-meeting notes should be sent out so that participants know their roles in project and the time frame for completion.

Agenda

A good agenda contains all the information that participants will need to make progress on the project at the meeting. It will contain:

- Date, time, and location of the meeting
- Expected attendees
- Things to be done before the meeting
  - Examples (but not all encompassing): share contact information, schedules of availability, read certain documents, search for a few references, look at equipment in lab
- Agenda of what will be discussed during the meeting
  - Likely similar to what was mentioned in the previous bullet, but also consider: brainstorm ideas, statistics discussion, objective goals, data that needs to be collected, who is writing what, document management, next meeting leader and date/time to meet

It is important that agenda items should be specific and results oriented. For example, an agenda item of "Discuss assignment" is not specific and may lead to a discussion of how the assignment fits into the entire curriculum. A more specific and results oriented goal would be "Discuss and understand lab objectives" as it is focused on getting all participants on the same page of what needs to be done to complete the lab.

For example, a meeting agenda may look like:

July 18, 2022
3-4 pm
Sorenson Conference Room
Invitees: Nate, Emily, Mike
Led by: Stephanie

Agenda Items:
- Discuss and understand lab objectives
- Determine what the deliverables are (that is, what needs to be completed/turned in)
- Individuals select sections to begin working on based on deliverables
- Brainstorm how to do final edits
- Determine what time works for Writing Center appointment and schedule it
- Set next meeting and leader

Notes

After the meeting it is important to follow up with the meeting attendees (and those who could not attend) to ensure that everyone knows what the plan is going forward within a short time of the meeting (typically within a few hours of the meeting end). Starting with the meeting agenda, details are added to provide attendees with what to accomplish next. These details should include:

- Date, time, and location of the meeting
- Meeting attendees
- What was discussed at the meeting – including brief summaries of each item
- Action items: - a "to do" list with people assigned to tasks with deadlines

It is important to have enough detail so that people who were absent (or who will not look at them for a day or two) understand what was discussed and the next steps, but not include minor details that do not progress the project. For example, if Mike commented on how good the sandwich was that they were eating, that does not impact the project and should not make the meeting notes. Additionally, having good formatting so that the topics and summary are clear is valuable. Using bullet points make the flow of topics and summaries easy to follow. Mirroring the flow of the agenda is also beneficial so that meeting attendees are not having to search for information.

For the example agenda above, the meeting notes may look like:

July 18; 3-4 pm; Sorenson Conference Room
Attendees: Stephanie, Emily, Nate Absent: Mike

Agenda Items:
- Discuss and understand lab objectives
    - Lab objectives were discussed and the objective of changing the CO_2 flow was clear
    - Need to follow up with manager on how we could change the temperature for objective 2 – *Nate will email manager*
- Determine what are the deliverables (that is, what needs to be completed/turned in)
    - Need to complete technical report by 8/15/22
- Individuals select sections to begin working on based on deliverables
    - Intro/background – *Emily will complete by 7/20*
    - Methods – *Mike will complete by 7/20*
    - *Stephanie will find references to add*
- Brainstorm how to do final edits
    - Easiest to work in Microsoft Teams so that everyone has access to the documents and can edit together – *Nate will set up a Microsoft Teams and invite everyone*
- Determine what time works for Writing Center appointment and schedule it
    - Determined 7/21 works for everyone so intro/background (*Emily*) and methods (*Mike*) sections will be completed by 7/20
    - *Nate* scheduled it at 1 pm during the meeting
- Set next meeting and leader
    - *Nate* will lead next meeting after Writing Center appointment on 7/23
    - Everyone needs go to lab and check out equipment for safety assignment due next week


Sending out the meeting notes within a few hours (or 24 h at the latest) is important so that everyone can be clear on what their tasks are. Having the meeting notes in an editable document (e.g., Google Docs, Microsoft Teams, etc.) can be beneficial since people can update their progress prior to the next meeting. However, caution should be exercised that "to do" items do not get deleted or revised without the consent of the team.

10. Communication Components

No matter what sort of professional work you do, you are likely to do lots of writing—and much of it technical in nature. The more you know about some basic technical-writing skills, the better job of writing you are likely to do, and that will be good for the projects you work on, for the organizations you work in, and—most of all—good for you and your career.

Generally speaking, these steps should be followed when communicating technically (Figure 10.1):
1. Understand the purpose for writing
2. Determine and analyze your audience for what background they will need to understand the results and discussion. This could be:
   o General public
   o Management
   o Supervisor
   o Colleague outside your field of expertise
   o Colleague inside your field of expertise
   o Subordinate (e.g., intern, employees, etc.)
3. Select the appropriate communication
   o Email
   o Business Memo
   o Proposal/Experimental Plan
   o Technical Report
4. Perform the writing process (Chapter 7)
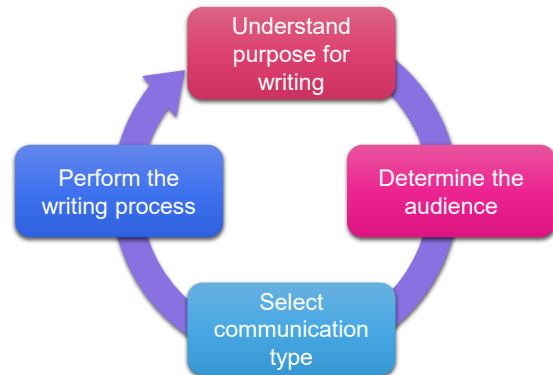
Figure 10.1: *The iterative cycle of developing a communication.*

Once you have determined the purpose of the document based upon your audience, writing can begin. The components you need to include in the document depend on what communication you selected and the targeted audience. As general guidelines the following communications typically contain:

- Email – Brief and to the point. The email will usually be similar to the executive summary of the business memo. Emails will contain a brief background, key results and discussion, as well as recommendations. If anything changed significantly in the methods and it affected the results, that should be mentioned as well.
- Business Memo – Typically no more than a few pages long and it focuses on the key results, discussion, and recommendations. The bare minimum introduction and background give the reader some guidance, and the methods might only be mentioned if they are complex or something significant changed/happened.
- Proposal/Experimental Plan – The theory and methods are the key components of a proposal and many details are needed in these sections. The reader needs to know that you understand what you are doing and how you are going to do it. The background and introduction are important as well but are secondary to the theory and methods. In a typical proposal, you do not have results, discussion, or conclusions section, but instead are answering the questions:
  1) Why are you going the experiment (intro/background)?
  2) What are you doing (methods)?
  3) How are you going to analyze your data (methods of analysis)?
  4) What is the reason you are doing the methods the way you are (theory)?
- Technical Report –Technical reports are usually thorough documents that can be filed away as references for future work. These are common in research and development as well as graduate studies (for example, a peer-reviewed journal article). They typically contain an abstract, introduction, background, theory, results and discussion, and a conclusion section.

Each of the report components are described in detail on the following pages.

Abstract (or Summary)

The abstract of a report is one of the most important pieces of technical communications as it provides the reader with concise information of what is contained within the report. Other than maybe figures and tables, which some readers look at first, the abstract will be the first thing read and needs to deliver the importance of the document. It is typically best to write the narrative of the document, then write the abstract as concise summary of the most important pieces. The abstract can range from 6-15 sentences and is a condensed version of the narrative that can stand on its own.

Key features of your work need to be included in the abstract so that *the people who read your abstract* can decide what to do with your report. If you are writing an abstract for the general public, it will look much different than if you are writing an abstract for the CEO of a company. It is important to keep the audience in mind so that you can tailor it (and the highlights of the report that they would want to know) directly for the reader.

Most abstracts for experimental results will include one or two sentences each on the:
1. background/introduction
2. experimental objectives
3. equipment and methods used

Then, the remaining 3-12 sentences consist of the important results (including numerical values), conclusions, and recommendations. Many example abstracts are available because the vast majority of research articles require one. Using Google Scholar (https://scholar.google.com/) you can search most anything and have research papers that you could view to get samples. Note that you may not have access to some journals due to subscription fees, but the abstracts are typically available regardless. Finding a paper that you can skim over and then see the important pieces the authors pulled for the abstract is of value and can help identify the important pieces for your work.

There is an exception to the typical abstract. The business memo is a much shorter document than a technical report and is typically written for management and upper-level employees. In the memo, the goal is to communicate the most important aspects of the project, so a typical abstract contains more information than necessary. For the business memo, the abstract should contain the purpose of the report, key results and discussion, conclusions, and recommendations and are typically half the length of a standard abstract.

Background/Introduction

The background or introduction section is a paragraph or two that describes the motivation for doing the work. This is more "big picture" and may discuss things like:
- Why is this topic/unit operation important?
- What are the key characteristics of the topic/unit operation?
- Where is this unit operation or system commonly used (e.g., what industries use it)?
- Why is this lab/data/recommendation important? (Think big picture.)

It does not get into details about the experimental set-up or method, but may include, for example, that you are separating x from y, which is done in ABC industry. The *background* section discusses why the experiment has broad appeal and importance, which provides a lead-in for the objective.

It is common to use the first paragraph of the *background* to explain what is known generally in the area of your study. Cite key references, but do not write an extensive review of literature; instead, possibly direct the reader to a recent review. Then, focus in on the problem that your study addresses. **References are expected in this section** and should strengthen your reasoning for performing the experiment. (Again, Google Scholar is a great place to find references as is Web of Science.) For example, if you say that "XXX BTUs of energy come from fossil fuels and fuel cells have the potential to replace YY% of this," this statement should have a reference as to where you got the information.

Objective Statement

Depending on the situation, you may be provided an objective (common for consulting companies) or you may have to develop your own (common as an engineer). Objectives need to be "SMART": (For more information: https://ctb.ku.edu/en/table-of-contents/structure/strategic-planning/create-objectives/main)

- *Specific*. That is, they tell *how much* of *what* is to be achieved *by when*.
- *Measurable*. Information concerning the objective can be collected, detected, or obtained.
- *Achievable*. It is feasible to pull them off.
- *Relevant:* Makes sense based on the experiment.
- *Timed*. Your organization has developed a timeline by which they will be achieved.

In the case of experimental objectives, specific and measurable are probably the most important considerations. That is, what will you measure in order to come to the conclusion you want to know about? If the task is to determine if experimental data matches a computer model prediction, the objective may be "Compare the experimental cooking distance of unpeeled potatoes to the provided Python model at 2 and 8 min." For experiments, objectives commonly begin with the verbs:

- Measure…
- Analyze…
- Test…
- Evaluate…
- Determine…
- Compare…

The experimental objectives are what you want to accomplish during the laboratory procedure *and resulting statistical analysis*. For example, in a report to a supervisor your objective may be to "Perform an experiment that evaluates the effect microwave power level and heating time has on the temperature of water." But it should also include "Then, compare the experimental value of 50% power to 100% power using a t-test at both 8 and 10 min to determine if the power level significantly impacts water temperature." Now, that is not the best wording, but the point is that the experimental objectives are to: A) run the lab to collect data and B) analyze the data. Even though an experiment is run, the goal is likely to compare the data you get to a theoretical value, compare data between two different pieces of equipment, two different variable settings, etc. This needs to be stated in the objective as well. *All major objectives* you are tasked with for experiments should be stated in this section, not just some of them.

The experimental and statistical objectives should allow the reader to have a clear understanding of what you plan on evaluating, measuring, and analyzing. Later, conclusions will be discussed which should be related back to the objectives. That is, the objectives set the stage for the entire narrative.

Theory

The Theory section of a report or publication is to give the reader a thorough background on the information they need to understand the results and conclusions you present. This may include an overview of the topic and equations needed to calculate technical results (typically statistics equations are not presented like mean, t-test, etc.). Understanding your audience and their background is important for this section as their needs will determine how much information you present. For example, if you are writing a business memo for the CEO of a company, the amount of theory required will be far less than for a technical report that will be indexed in the company's files for reference at a future date.

In the end, the reader needs to know that you understand what you are writing/talking about and be able to understand your results *at whatever level they care about*. For example, if you are writing a/an:

- email – likely little theory is necessary, and equations may not need to be presented since emails should be brief and to the point.

- business memo – similar to an email, a business memo is top level with a focus on results, recommendations, and conclusions; perhaps a sentence or two of theory is appropriate and in rare cases, an equation, but not much more.
- technical report – the technical report will likely have less theory than a proposal but will contain equations important to understanding the results.
- proposal – this document likely has the most theory as why you are doing what you are doing and equations that will be needed to make calculations need to be presented.

You also need to consider the audience. A presentation or document to the public will need to have a theory section that is straight-forward and able to be understood by someone without a technical background. A document for a colleague/supervisor in the same field can be much more technical and higher level, but a communication to the CEO may fall somewhere in-between there depending on the presentation time or document length.

The theory section can also contain equations if they are presented and should appear in the order they appear in the document. Then, when presenting the results, the equation number can be referenced. Equations should be consecutively numbered and follow standard guidelines, which can be seen in most technical textbooks. More can be found about inserting equations using Microsoft Word in the Appendix, but key takeaways are that:
- equations need to be their own line in a document, but part of a sentence,
- the variables are defined immediately after the equation is presented, and
- all equations *and variables* used in the document are inserted with Equation Editor, which has a specific Cambria Math font.

Be sure to properly reference equations that are from other sources. For example, the Ergun equation or Fick's Law would not be considered common knowledge and if presented, a reliable resource (e.g., textbook, *Perry's Chemical Handbook*, etc.) should be cited.

An equation should appear as follows:

"After measuring the radius and the circumference (C) of each of the circles, the value of π was calculated using:

$$C = \pi d \tag{1}$$

where d is diameter of the circle. Using this information…"

Notice that the equation:
- is part of the sentence but is on its own line.
- is numbered so that it can be referenced in other sections. (For example: "using equation 1, it was found…")
- has the variables defined directly after the equations is presented. In this case, the circumference was defined prior to the equation, so C did not have to be defined again. Variables only need to be defined once.
- the variables are all in the equation editor font (e.g., d).

More information on how to insert equations using Microsoft Word, including step-by-step directions) can be found in the appendix.

Methods
The Methods section involves several major components that can appear in the narrative (i.e., body) of a communication or the appendix as supporting information. These include:

- Apparatus (i.e., equipment; typically a block flow diagram),
- Safety,
- Experimental Protocol (step-by-step directions),
- Experimental Design (experimental method in paragraph form),
- Methods of Analysis, and
- Data Collection

which are all dependent on each other in one way or another. For example, you cannot develop the protocol without knowing the apparatus and you cannot design the experiment without knowing how you are going to analyze the data. Therefore, although the components are presented in a certain order, know that this section will be iterative since certain pieces may have been given to you, are easier to obtain, or may have to be revised as the experiment develops.

Assuming the objectives are known, in order to determine what factors can be changed, the best place to start is with the equipment itself and then build the methods section by adding the necessary information, like the *experimental design* and *methods of analysis*.

Apparatus (or Equipment)

Details are needed about the apparatus and equipment so that you can create an experimental design that will allow you to accomplish the objectives. It should include details of both the equipment and analytical instruments, but not read as a list and should be in paragraph form. Examining the apparatus allows you to learn what the equipment does, what variables (also known as "factors") can be changed on the equipment, the possible ranges of the variables (e.g., flow rates from 10-100 sccm), and the volume/size you are working with. This allows you to create a better experimental design as you will know the limits of the equipment.

When examining equipment, look for instruments, control elements, indicators, or recorders that can give you information about key variables such as flow, levels, pressure, and temperature. Controllers, as expected, are devices that control one of the process variables, like a flow controller, back flow pressure regulator, or temperature controller. Indicators just provide an output, like a temperature gauge or pressure read-out. This might be a visual display, a signal to be sent to a computer, or a graphical printout (e.g., strip chart recorder or computer screen). Examine the equipment and read any manuals associated with the equipment to determine what can be varied and what data you can collect from the equipment. Just as important are the things that should be recorded but cannot be controlled. For example, atmospheric pressure and temperature may affect the experiment and need to be noted but are likely not able to be controlled.

As an example, say the goal of an experiment was to perform an experiment that evaluates the effect microwave power level and heating time has on the temperature of water. Bulleted lists like:
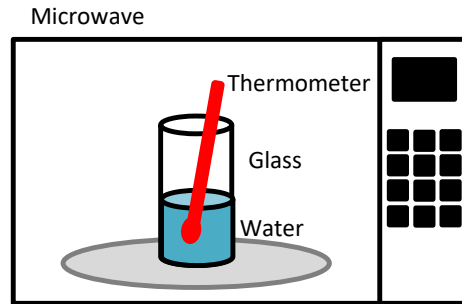- Microwave
- 4 16 oz glasses
- Thermometer
- Water

should not be used for describing the equipment. Instead, describing the apparatus and equipment should be *built into the experimental design section* and not a standalone section (will be discussed later, but again, all the methods components are all tied to each other), such as:

"A 400 mL glass beaker will be filled using a 1 cup measuring cup with 8 oz of water that has been refrigerated overnight. The beaker will be placed in the microwave (GE, 1000 W) that has the ability to have specific times and power levels set. The temperature of the water will be measured before and after

heating using a standard alcohol thermometer. After heating for a given amount of time, the microwave will be opened, and a thermometer inserted into the cup (Figure 10.2). After stirring the water briskly for 15 s to allow the temperature to equilibrate, the temperature will be read and recorded in Excel."

It is **not** every single detail since some can be inferred. This next example contains too many details and could be more concise:

"A 400 mL glass beaker obtained from the lab will be filled with 8 oz of water from the faucet that was measured with a 1 cup measuring cup. The beaker will then be placed into the refrigerator. The next day, the refrigerator will be opened, and the glass will be taken out and the microwave will be plugged into the wall. The microwave door will be opened and the glass containing water placed into the microwave. The microwave will be closed and the time entered on the keypad. The start button will be pushed and after the stated amount of time, the microwave will be opened, and a thermometer inserted into the cup (Figure 10.1). After stirring the water briskly for 15 s and allowing the temperature to equilibrate, the temperature will be read and recorded on the data sheet in Excel on tab "Experimental Data." The cup will then be removed from the microwave, and the water in the cup will be disposed of down the drain."

Figure 10.2: *Schematic of microwave set-up after the water has been heated.*

The schematic (i.e., block flow diagram (BFD)) should be used to show the readers the apparatus and can be drawn in software programs such as Microsoft PowerPoint, Microsoft Visio, AutoCAD, SolidWorks, etc. For simple diagrams, a diagram such as Figure 10.2 is adequate, but for more complex systems, gauges, flow direction indicated by arrows, and major pieces of the equipment should be labeled (Figure 10.3). After completing the drawing, be sure the following are shown:
- Major equipment and instruments with labels for key pieces of the system
- Flows of materials are indicated using lines **with 90° angles** and arrows pointing in the direction of flow (Figure 10.3)
- Important controls, instruments, and sampling points (as appropriate) are noted.
  Do not use photographs or scans of diagrams; you must draw the schematic yourself

As an example, a BFD of a packed bed experiment is shown in Figure 10.3. This schematic was drawn using Microsoft PowerPoint and then pasted into Microsoft Word as a picture.

Figures **should not** have borders around the image (Figure 10.4) or the figure and caption (i.e., text box, if used) as seen Figure 10.5. It is highly recommended to insert a text box and then paste the figure and write the caption within the text box as shown in Figure 10.3. That way, the figure is easy to move and you can wrap text (if desired) as shown around Figures 10.2 and 10.3. How to do use text boxes is explained in detail in the appendix.

Drawings and schematics are critical to understanding in academia, research and development and industry. Drawings are often the main piece of information engineers use when running experiments or are in the field, therefore, these drawings must be easy to understand, complete, correct and up to date in order to accurately convey engineering design and operations.
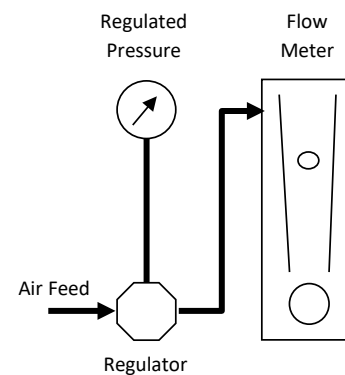
Figure 10.3: *This is how images/schematics should be presented.*
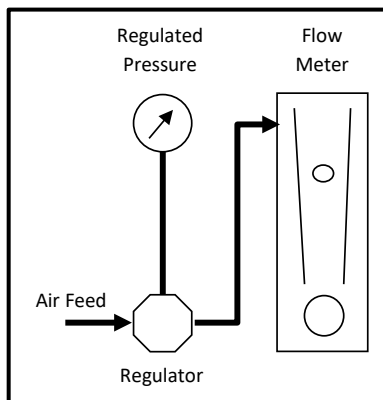
Figure 10.4: Do not have a border around image.
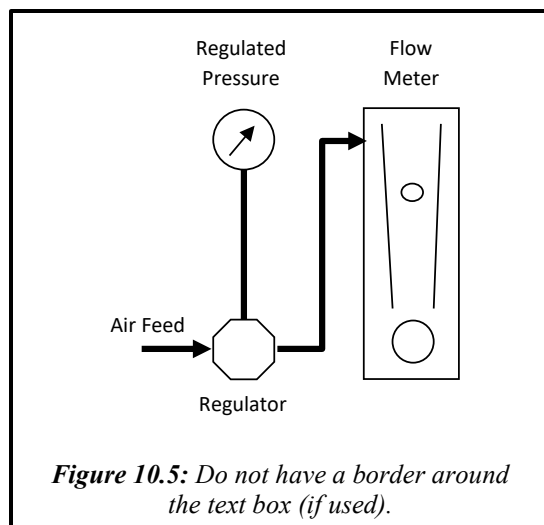
Figure 10.5: Do not have a border around the text box (if used).

NOTE: It is important that no sections begin with a figure (or table). Figures need to be discussed BEFORE they appear (or beside the reference as shown in the example above). The first few sentences of the methods section will often summarize the equipment being used and reference the equipment diagram as written above. Report sections should **always begin with text** that leads the reader to the figure(s).

Safety

While examining the apparatus, it is important to consider safety, as safety is a **critical** piece of running experiments and working with equipment. Ask yourself:

- What hazards are involved?
- What could go wrong?
- What would I do if something went wrong?

A good place to start is using safety considerations like those that can be found on the Dow Chemical Workplace Safety Website (https://corporate.dow.com/en-us/science-and-sustainability/innovation/safety-at-dow.html). These considerations include:

- EMERGENCY SHUTDOWN PROCEDURES: Include location of plugs and devices
- CHEMICAL EXPOSURE HAZARDS & RESPONSE: Use material safety data sheets (MSDS) to get more information on specific chemicals
  - EYE CONTACT:
  - SKIN CONTACT:
  - INHALATION:
- PERSONAL PROTECTIVE EQUIPMENT
  - EYE PROTECTION:
  - BODY PROTECTION:
  - REPSIRATORY PROTECTION:
  - SPILL CLEANUP/EQUIPMENT
  - DECONTAMINATION RESPONSE: Are clean-up or decontamination supplies compatible and readily available?
- CHEMICAL, PHYSICAL OR MECHANICAL HAZARDS - Consider chemical compatibility, reactivity, flammability, explosion potential, toxicity. Can your equipment handle rapid evolution of volumes of gas? - Consider pressure, temperature, corrosivity, electrical hazards, cylinder storage/use

along with many other considerations that found can be found at: https://corporate.dow.com/content/dam/corp/documents/science-sustainability/920-00006-01-safety-operation-card-template.pdf. Once you develop an outline, then write a concise paragraph or two on what the hazards are, the potential injuries, and how the situation would be handled if someone was injured for the most likely incidents for the narrative. You do not need to include low probability situations for simple experiments. That is, although you are using a microwave to perform the experiment, the chances of the microwave exploding are slim and would not need to be addressed.

For example, for the microwave experiment, the form may look like this:

EMERGENCY SHUTDOWN PROCEDURES: Turn off water at faucet and if faucet doesn't work, turn off water shut off below sink. Unplug microwave from wall if "stop" button is not working.
CHEMICAL EXPOSURE HAZARDS & RESPONSE: Water is a low-risk chemical with low safety precautions
EYECONTACT: Hot water may splash into eyes. Safety glasses will be worn at all times, but if hot water gets into eyes, flush eyes at eye wash and get medical attention if symptoms occur.
SKIN CONTACT: If the hot water or container creates a burn, flush skin with plenty of cool water. Get medical attention if symptoms occur.
INHALATION: N/A
PERSONAL PROTECTIVE EQUIPMENT: Safety glasses, long pants, and closed toed shoes are required. Heat gloves will be worn when removing glass from microwave.
EYE PROTECTION: Safety glasses throughout experiment.
BODY PROTECTION: Heat gloves for handling hot containers.
REPSIRATORY PROTECTION: None needed
SPILL CLEANUP/EQUIPMENT: Any water spills will be cleaned up with sponges and/or paper towels that are available in the lab and disposed of in the trash cans.
DECONTAMINATION RESPONSE: N/A
CHEMICAL, PHYSICAL OR MECHANICAL HAZARDS – If glass breaks, it will be cleaned with a broom and dustpan and disposed of in the broken glass containers located within the lab.

The paragraph that appears in the narrative would then look like:

"Due to the potential high temperature of the heated water, there is a risk of skin burns. In order to reduce the risk, gloves that can withstand high temperatures will be used to remove the heated cup from the microwave. If a burn occurs, cool water will be run over the affected area for at least 20 min. and assessed. In the case of a mild burn, the burn will be bandaged and the instructor notified. In the case a of severe burn, treatment will be sought at a medical facility and the instructor notified."

These statements cover all of the important aspects:
- The **hazard** is the hot water,
- the **potential injury** is a skin burn, and
- the **situation will be handled** based on the severity of the burn.

Besides hazards to people, there are also hazards to the equipment should be considered. These types of hazards can be different or can be linked. For example, running a pump dry can cause the pump seal to overheat and break. If this happens, you have equipment damage and the potential for water leaking from the seal and causing a slipping hazard to people. It is critical that safety is not overlooked and is seriously considered. Safety is crucial in industry, research, and life in general. It is important to do due diligence and be prepared for things that may go wrong during experiments and how to handle them.

Experimental Protocol, Design, and Analysis

When you think of a methods section, the experimental design and protocol are typically what come to mind. The *experimental protocol* consists of the step-by-step directions to run the experiment while the *experimental design* consists of the variables you are going to change to what levels and how you are going to do this in paragraph form. Often forgotten in the *experimental design* is the data analysis piece, which is how data will be analyzed and part of the *methods of analysis*. All three of these sections are highly intertwined as they all depend on each other (Figure 10.6). For example, if you decide to run a t-test, you will need at least three data points at the same condition, which means you will have to modify the protocol to make sure you are getting enough samples and the design will have to be clear on what conditions the samples are being collected at.
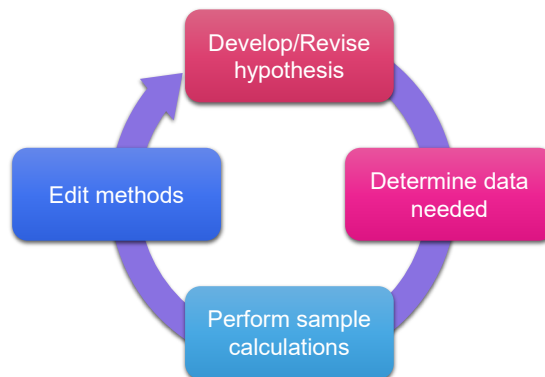
Figure 10.6: The iterative cycle of developing a data collection plan and methods section.

Designing the experiment begins with two pieces of information:
1. the goal of the experiment based on the experimental objective and
2. the limitations of what can be changed and to what level using the equipment.

Using these two pieces of information, hypotheses can be developed as to what will be compared/determined and the variables and levels chosen to test the hypothesis. An important piece to this is deciding how you are going to test each hypothesis, which will typically require a statistical test of some sort. For example, if you want to determine if the experimental results you got were significantly different than a software model prediction, you may want to do a t-test comparing the experimental data at specific conditions to the model prediction using those same conditions. That means you have to know that the model can predict data at the experimental conditions you chose (and if not, modify the experiment) and you need to have at least three (ideally five) experimental repeats at those specific reaction conditions in order to do the t-test. The methods section needs to discuss the conditions of the experiment, how you are going to run the experiment, the number of repeats, and how you are going to compare the data. For example, if you only want to calculate a percent difference between the experiment and model, you may only need one data point. The *methods of analysis* are critical for the experimental plan since if you do not collect the correct data, or the correct amount of data, you will need to go back and collect the proper amount of data. A general rule of statistics is that the more data you collect, the better your results will be and if you had infinite time, you could have amazing statistical results. However, time is limited so you will need to consider how many changes you can realistically make (that is, the number of variable and level changes) in to order to meet the objectives and provide solid statistical conclusions. This ties back to examining the equipment/apparatus as hopefully you can get an idea of how long data collection will take.

Experimental Protocol

The *experimental protocol* is **a list of steps** (e.g., 1. do this, 2. do that, 3. then this) used to perform the experiment and not written in paragraph form, but rather a list. Think of it as "the recipe" to run the experiment. If you gave the protocol to someone else, they should be able to run the experiment exactly how you did. This should be **a thorough step-by-step list** of the steps required to:
1. Start the equipment
2. Perform the experiment
3. Shut down the equipment

Sometimes a protocol will be given to you and can be used as a starting point. However, it is important to incorporate the factors, levels, parameters, and repetitions you decided you will be running into your protocol **and importantly** revise the protocol after the experiment. Something will be missed that you realize during the experiment, or sometimes steps will change; these should be noted during the experiment and the protocol revised.

Here is an example *experimental protocol* sample for the microwave lab discussed above:
1. Add 8 oz of distilled, refrigerated water to a 400 mL glass beaker using a 1 cup measuring cup
2. Place beaker in microwave and heat at 100% power for 2 min.
3. Open the microwave door and immediately place thermometer in water
4. Stir water with thermometer and read temperature after it reaches steady state (~15 s)
5. Repeat steps 1-4 for 4, 6, 8 (4 repeats), and 10 min (4 repeats)
6. Repeat step 1-5 at 50% power

Notice how each time was not a separate line, which would look like this:
1. Add 8 oz of distilled, refrigerated water to a 400 mL glass beaker using a 1 cup measuring cup
2. Place beaker in microwave and heat at 100% power for 2 min.
3. Open the microwave door and immediately place thermometer in water
4. Stir water with thermometer and read temperature after it reaches steady state (~15 s)
5. Add 8 oz of distilled, refrigerated water to a 400 mL glass beaker using a 1 cup measuring cup
6. Place beaker in microwave and heat at 100% power for 4 min.
7. Open the microwave door and immediately place thermometer in water
8. Stir water with thermometer and read temperature after it reaches steady state (~15 s)
9. Etc, etc for 6, 8, and 10 min.

For steps that repeat, instead of having a long list of repeating steps, the first protocol had number 5: "Repeat steps 1-4 for 4, 6, 8 (4 repeats), and 10 min (4 repeats)." This is a much easier way to shorten the protocol and eliminate repeated directions. It also makes clear that you are **repeating** certain steps at different levels, which makes it easier for the reader to determine if you are collecting enough data for the analyses you plan on doing. Remember, the protocol should allow anyone to repeat the experiment exactly so details are important.

Experimental Design
The *experimental design* is an extremely significant part of the methods section. The experimental design incorporates the important pieces of the *experimental protocol*, but in a paragraph form. It contains a description of the scope of the experiments that were performed and identifies the:
- Factors – the variables that were adjusted
- Levels – the values of the varied variables
- Parameters – variables that were treated as constants in the experiment
- Frequency – Repetition and randomization

Randomization is important as you can minimize trends in data due to time and other effects. In the case of the microwave experiment, the data collection should be set up to randomize the order of both the time and microwave power. For some experiments the data collection cannot be randomized either due to the process or because changes are too difficult to make (think a manufacturing process where the raw materials are being changed), but the order in which the samples are analyzed can be. For example, if you are analyzing samples on a gas chromatograph, the baseline, which should be zero, can drift with time. If you run the samples in a particular order, like all the 2 min samples first, then 4 min, then 8, then 10, and the baseline drifts, the samples run later will be artificially high or low due to the drift. Randomizing the samples (and

in some cases, including an internal standard) can minimize this effect. Some experiments neither the data collection or analysis can be randomized.

TIP: Use the random number generator ("=RANDBETWEEN(x, y)" or "=RAND") in Excel to randomize the order of your runs by generating a column of random numbers for each run that you sort the table by.

For example, to incorporate the *experimental design* into the previously written *apparatus* section for the microwave lab method, it could look something like this (Note that the added text is in italics so it is easier to see; it would not be in italics in the methods section):

"A 400 mL glass beaker will be filled using a 1 cup measuring cup with 8 oz of water that has been refrigerated overnight. *By using a 1 cup measuring cup, it is assumed that each glass will contain the same amount of water.* The beaker will be placed in the microwave (GE, 1000 W) *and heated for 2, 4, 6, 8, and 10 min at a power rating of and 100%. At 8 and 10 min, four repeat measurements will be taken. These steps will be repeated at the 50% power setting. Excel will be used to determine the order of the experiments using the random number function.* The temperature of the water will be measured before and after heating using a standard alcohol thermometer. After heating for a given amount of time, the microwave will be opened, and a thermometer inserted into the cup (Figure 10.2). After stirring the water briskly for 15 s to allow the temperature to equilibrate, the temperature will be read and recorded in Excel. *Briskly stirring the water after heating will allow for the assumption that all liquid in the cup is at the same temperature. This temperature difference between the hot and cold measurements will be used for the analysis.*"
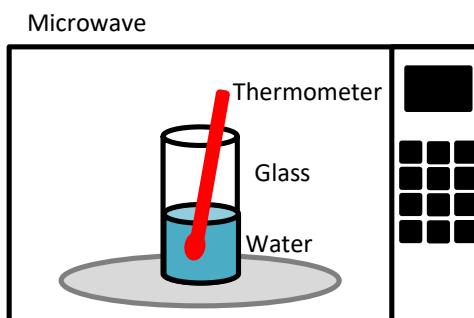
Figure 10.2: *Schematic of microwave set-up after the water has been heated.*

In this paragraph, everything is identified:
- Factors: power and time
- Levels: 50%/100% and 2, 4, 6, 8, 10 min
- Parameters: water amount and temperature gradient
- Repetition and randomization – at 8 and 10 min four sets of data will be collected and the order of the experiments will be randomized using Excel
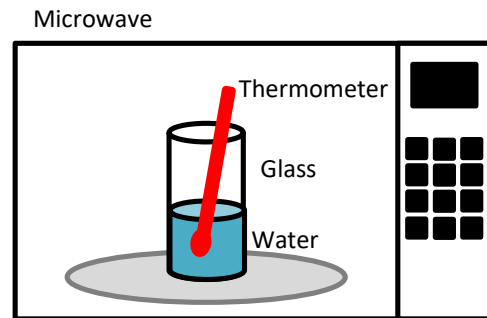
Methods of Analysis

When writing a proposal, it is important that you convey to the reader that the data you collect (*experimental design*) is enough to provide conclusions regarding your experimental hypothesis (statistics). To do this, the *methods of analysis* section is where you describe the statistical tests and figures that you will be using to support your recommendations. The methods of analysis may consist of t-tests, regression analysis, or could simply be percent differences or averages and standard deviations amongst others. Thinking about the figures that you will want to include also helps inform the data collection. For example, if you will be incorporating a process control chart, then it is necessary to collect steady-state data to determine the averages and standard deviations for the chart, as well as collect the data that will be displayed as markers.

For example, to incorporate the *methods of analysis* into the previously written *experimental design/apparatus* section for the microwave lab method, it could look something like this (Note that the added text is in italics so it is easier to see; it would not be in italics in the methods section):

*This paragraph represents an example of a **COMPLETE** *methods* section for this microwave lab as it integrates the: *apparatus*, *experimental design*, and *methods of analysis*. For a proposal, your methods will likely be more detailed and require several paragraphs. The different parts of the *methods* can be separate or integrated (as shown below) depending on your preference.*

"A 400 mL glass beaker will be filled using a 1 cup measuring cup with 8 oz of water that has been refrigerated overnight. By using a 1 cup measuring cup, it is assumed that each glass will contain the same amount of water. The beaker will be placed in the microwave (GE, 1000 W) and heated for 2, 4, 6, 8, and 10 min at a power rating of and 100%. At 8 and 10 min, four repeat measurements will be taken *in order to have data points to perform an equal variance t-test between 8 and 10 min to determine if the heating time makes a significant difference on the temperature*. These steps will be repeated at the 50% power setting *and equal variance t-tests will be done on the 8-min samples at 50% and 100% and separately for the 10-min samples as well to see if the power level makes a significant difference on the temperature*. Excel will be used to determine the order of the experiments using the random number function *and for the t-tests*. The temperature of the water will be measured before and after heating using a standard alcohol thermometer. After heating for a given amount of time, the microwave will be opened, and a thermometer inserted into the cup (Figure 10.2). After stirring the water briskly for 15 s to allow the temperature to equilibrate, the temperature will be read and recorded in Excel. Briskly stirring the water after heating will allow for the assumption that all liquid in the cup is at the same temperature. This temperature difference between the hot and cold measurements will be used for the analysis. *Additionally, a figure will be made plotting the time versus water temperature of both the 50% and 100% power data.*"

Figure 10.2: *Schematic of microwave set-up after the water has been heated.*

Or, the *methods of analysis* could be a separate paragraph stating:
"At 8 and 10 min, an equal variance t-test will be performed to determine if the heating time makes a significant difference on the temperature at both the 50% and 100% power levels. Additionally, equal variance t-tests will be done on the 8-min samples at 50% and 100% and separately for the 10-min samples as well to see if the power level makes a significant difference on the temperature. Additionally, a figure will be made plotting the time versus water temperature of both the 50% and 100% power data."

Data Collection

As the engineer on a project, you need to decide what data is important to use and what other data you will need to collect to confirm/refute a hypothesis. Knowing what and how much data you need to draw conclusions for each objective will help determine what data you need to collect. After you come up with a list of data you think is sufficient, it is beneficial to test your calculations using estimated data of what you plan on collecting. This makes it much easier to see if you are missing a critical piece of data. For example, if I was evaluating pressure drop and know that the pressure gauge reads gauge pressure, in the sample calculations I might realize that I need to record the atmospheric pressure as well in order to calculate the absolute pressure needed for the calculations. By creating a data collection sheet and doing sample analyses/calculations prior to running the experiment, you can ensure you are collecting the necessary data in order to not have to rerun the experiment.

It is useful to include a table that will be used in the lab to collect your data for proposals. This allows your supervisor to ensure you are collecting all the data you need and therefore, you have a much lower chance of needing to rerun the experiment due to lack of data or not collecting the correct data. For example, in the microwave lab introduced earlier, Table A1 would be the associated data collection table and is entitled "A1" since it is being presented as the first table that appeared in an appendix. Notice that the "Planned heating time" column is not what you would expect. One would likely expect to see the times increase consecutively like: run 1 be 2 min, run 2 be 4 min, run 3 be 6 min, etc, which is how the table was initially made (see Table A2 in three pages). However, in order to randomize the runs, the table was made in Excel and in a ninth column, the formula "=RANDBETWEEN(1, 100)" was entered and copied down for all the runs. The entire table was highlighted and sorted by this ninth row, which resulted in the order show in Table A1. Using the random number generator in Excel is an easy way to randomize your runs.

Table A1: Excerpt of raw data collection sheet for microwave experiment.

Run #	Time	Planned heating time (min)	Actual heating time (min)	Microwave power	Temp before	Rest time before T reading (s)	Temp after
1		8		100			
2		4		50			
3		10		100			
4		10		50			
5		2		100			
6		4		100			
7		10		50			
8		8		50			
9		10		50			
10		10		50			
11		8		50			
12		10		100			
13		10		100			
14		6		50			
15		6		100			
16		8		100			
17		10		100			
18		8		50			
19		8		100			
20		2		50			
21		8		50			
22		8		100			

Obviously the variables you change are important to record, but also consider if any other data is important to collect. For example, if you are boiling water, the atmospheric temperature and pressure would be good to record as that will affect the boiling point. Are different people going to be collecting samples? Perhaps a column for initials is necessary since someone might have done something different, which affected the results.

Results and Discussion

The most critical piece of running a lab or experiment are the results and how you interpret and discuss them. There are two standard ways most reports present the results and discussion: 1) as separate sections and 2) as a combined section. Presenting the results and discussion as separate sections is required for some reports/publications but tends to be more difficult to do as you need to *only* present data with no interpretation in the results section. Additionally, when you discuss the data, you tend to refer the reader back to the results section, which can lead to confusion and repeating information. A combined "Results and Discussion" section is recommended as you can interpret the data as it is presented and give the readers the "what does this mean?" immediately.

When reporting what data was collected and the results that were produced from your analysis, it is critical to be concise and clear. You do not need to present every piece of data in the narrative. You want to focus on the data that help confirm/refute your hypotheses regarding the objective and the data that show significant trends. (The appendix is the place to put all of the data regardless of if it is presented in the narrative or not.) One of the best ways to present data is to choose an appropriate visual, like a figure or table, to display key results pertaining to your objectives and present that visual with corresponding text in the narrative. Ideally, while you developed the experimental design, you thought through how your data would be presented in the final report. Anticipating what you need to present with a solid *methods of analysis* makes the results section much easier to complete after you collect the data.

Keep in mind these questions when developing the discussion of your results, which can be found at https://lsa.umich.edu/sweetland/undergraduates/writing-guides/how-do-i-present-findings-from-my-experiment-in-a-report-.html):
- Is your hypothesis supported?
- Was there any data that surprised you?
- Are the results useful?
- What are the implications of your work?
- What are the shortcomings of your work?

Here is a brief, and should in no way be considered thorough, example using microwave heating data:

"The data from both the 50% and 100% power trials follow similar linear trends (Figure 10.7). Both power levels resulted in temperature changes of around 44°F and 75°F after 2 and 10 min of heating, respectively. When looking at the statistical analysis (Tables A2-A5), this is confirmed since there is no significant difference at the 95% confidence level between both the 8 min and 10 min heating times when comparing 50% power to 100% power. This was determined as the p-value for the unequal variance test was greater than the alpha value of 0.05. At 8 min, the temperature change was 69 ± 2°F for 100% power and 70.2 ± 0.7°F at 50% power. At 10 min and 100% power, the temperature change was 75 ± 1°F while at 50% power the temperature change was 76.0 ± 0.9°F (Table 10.1). Therefore, in this microwave, heating at 50% power or 100% power provides a similar amount of energy to the water indicating that the microwave is not working properly since at 50% power, the temperature change of the water should be less than at 100% power.

Where the microwave oven does seem to be working properly is that additional time increases the temperature change (Figure 10.7). At both powers there was a significant difference in the temperature change comparing 8 to 10 min of heating times at a 95% confidence level. This was determined as the p-value of the equal variance test was less than 0.05. At 50% power, the temperature change increase from 70.2 ± 0.7°F to 76.0 ± 0.9°F and at 100% power, the temperature change increased from 69 ± 2°F to 75 ± 1°F as seen in Table 10.1.

It is important to note that at the 2 min time, there is more separation between the 50% and 100% power temperature (4.5% difference). This may mean that at shorter times, there is a difference due to power

level which should be evaluated since most people likely use their microwaves for short durations. Also, gathering more data may reduce variability leading to more significance in the results. Additionally, since the expected water temperature increase could be calculated based on the microwave wattage, these calculations should be made and compared to the experimental results."

(Sample of the data that would appear in the appendix of a technical report for the microwave experiment follow on the next two pages.)

Notice that the example addresses:
- Is your hypothesis supported?
    - The hypothesis supported regarding heating time,
- Was there any data that surprised you?
    - The power level did not make a significant difference at 8 and 10 min.,
- Are the results useful?
    - The data was good (no outliers) and the statistics were ran correctly.
- What are the implications of your work?
    - This microwave works well for increased time, but the power level did not effect the water temperature significantly. Perhaps the microwave is broke.
- What are the shortcomings of your work?
    - Most people do not use microwaves for 8 and 10 min, so short times should be evaluated. Additionally, a comparison to calculations of how much the temperature should increase should be made.
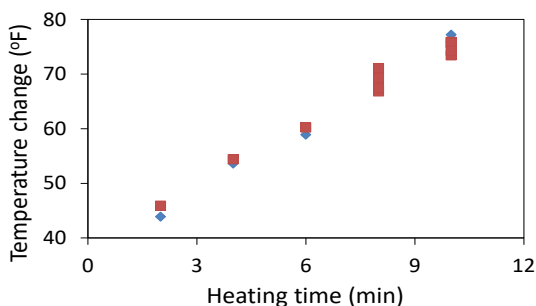
Figure 10.7: Change in water temperature after heating for a given time in a 1000 W microwave. ■ *represent 100% power and* ♦ *represent 50% power.*

Table 10.1: *Data from repeat measurements for the microwave experiment.*

Power (%)	Heating time (min)	Average temperature change (°F)
50	8	70.2±0.7
50	10	76.0±0.9
100	8	69±2
100	10	75±1

When required, include the full data (Sample Table A2 on next page) and analysis tables (Sample Tables A3-A6 following) from Excel in the Appendix as shown on the following pages.

NOTE: Significant figures matter within the narrative and for the statistics tables pasted into the Appendix, as significant figures indicate precision of the equipment and data.

Table A2: Collected data for microwave experiment.

Run #	Heating time (min)	Microwave power (%)	Temp before (°F)	Temp after (°F)
5	2	100	36.5	82.4
6	4	100	36.6	91.0
15	6	100	36.5	96.7
1	8	100	36.5	105.8
3	10	100	36.6	112.5
20	2	50	36.5	80.4
2	4	50	36.4	90.0
14	6	50	36.6	95.5
8	8	50	36.5	106.5
4	10	50	36.5	112.5
16	8	100	36.6	103.4
19	8	100	36.4	107.5
22	8	100	36.4	103.8
12	10	100	36.6	110.0
13	10	100	36.6	112.3
17	10	100	36.5	110.7
11	8	50	36.5	106.0
18	8	50	36.4	107.5
21	8	50	36.5	106.8
7	10	50	36.5	111.4
9	10	50	36.6	112.5
10	10	50	36.6	113.8

Note that this table has the data organize for the figures and tables that were made. The run number corresponds to the number in Table A1 where the data collection was randomized, but was sorted using Excel to be presented in a more logical manner.

Table A3: *Comparing 8 and 10 min at 100% power.*

	50 8	50 10
Mean	70.2	76
Variance	0.449	0.887
Observations	4	4
Hypothesized Mean Difference	0	
df	6	
t Stat	-9.99	
P(T<=t) one-tail	2.91E-05	
t Critical one-tail	1.94	
P(T<=t) two-tail	5.81E-05	
t Critical two-tail	2.45	

Table A4: *Comparing 50% and 100% power at 8 min.*

	50 8	100 8
Mean	70.2	68.6
Variance	0.449	3.80
Observations	4	4
Hypothesized Mean Difference	0	
df	4	
t Stat	1.53	
P(T<=t) one-tail	0.101	
t Critical one-tail	2.13	
P(T<=t) two-tail	0.201	
t Critical two-tail	2.45	

Table A5: *Comparing 8 and 10 min at 100% power.*

	100 8	100 10
Mean	68.6	74.8
Variance	3.80	1.45
Observations	4	4
Hypothesized Mean Difference	0	
df	6	
t Stat	-5.37	
P(T<=t) one-tail	0.000857	
t Critical one-tail	1.94	
P(T<=t) two-tail	0.00172	
t Critical two-tail	2.45	

Table A6: *Comparing 50% and 100% power at 10 min.*

	50 10	100 10
Mean	76	74.8
Variance	0.887	1.47
Observations	4	4
Hypothesized Mean Difference	0	
df	6	
t Stat	1.57	
P(T<=t) one-tail	0.0836	
t Critical one-tail	1.94	
P(T<=t) two-tail	0.167	
t Critical two-tail	2.45	

Note that the notation used in the spreadsheet was "Power time," so 100 8 indicates heating at 100% power for 8 min. In these analyses, the header was selected and the "Labels" box was checked.

.

Conclusions and Recommendations

Conclusions and recommendations could be thought of a more detailed abstract, minus the methods, but still concise. Everything that appears in conclusions should be mentioned elsewhere in the report; **there should be no new results presented in the Conclusions section**. For example:

- DISCUSSION: At 8 min, the temperature change was $69 \pm 2^{\circ}F$ for 100% power and $70.2 \pm 0.7^{\circ}F$ at 50% power. At 10 min and 100% power, the temperature change was $75 \pm 1^{\circ}F$ while at 50% power the temperature change was $76.0 \pm 0.9^{\circ}F$.
- CONCLUSION: There was no significant difference when 50% and 100% power was used to heat the water at 8 or 10 min when, but there was a significant difference comparing 8 and 10 min of heating for both powers. Therefore, the microwave power function may not be working correctly, but the addition of time leads to increased heating.

An excellent reference for writing scientific conclusions can be found at the following website: https://www.dvusd.org/cms/lib011/AZ01901092/Centricity/Domain/3054/conclusionwriting.pdf and a summary of what is contained in the document is found in the numerical list and paragraph below.

A conclusion paragraph contains a description of the purpose of the experiment, a discussion of your major findings, a brief explanation of your findings, and recommendations for further study. Address the following points in paragraph form:

1. Restate the overall purpose of the experiment.
   a. One format: The purpose of the experiment was to investigate the effect of (THIS) on (THAT)
   b. Example: The purpose of the experiment was to investigate the effect of stress on the growth of bean plants by comparing the growth of bean plants subjected to stress for 15 days with a control of non-stressed plants.
2. What were the major findings? Summarize your data and graph results, but do not refer to tables and figures.
   a. Example: No significant difference existed between the height of stressed plants (10.2±0.5 cm) and non-stressed plants (10.5±0.3 cm).
3. Was the hypothesis supported by the data?
   a. One format: The hypothesis that (insert your hypothesis) was (supported, partially supported, or not supported.) Do not use the word "prove" – we do NOT prove hypotheses true in science.
   b. Example: The hypothesis that stressed plants would have a significantly lower mean height was not supported.
4. How could this experiment be improved?
   a. Example: This experiment relied on an artificial source of stress – just digging out the plants at one time and replanting them. Perhaps this experiment could be improved by simulating real-life stressors, including drought and lack of nutrients in soil.
   b. NOT acceptable: This experiment would have been better if we had done it correctly – we did sloppy work and made careless measurements.
5. What could be studied next after this experiment? What new experiment could continue study of this topic?
   a. Example: Additional investigations using various sources of stress at more frequent intervals would be a good additional experiment. Also, other crops could be subjected to the same experiment, such as corn and squash. Perhaps scientists could find a chemical that the plants release during stress.

Sample Conclusion Paragraph
Experiment: Dr. Carlson is trying to create the best environment for his fish in his aquarium. He wants to figure out the relationship between the water's temperature and the amount of oxygen dissolved in it (fish need enough oxygen to breathe through their gills in the water). He sets up an experiment where he uses ice and a hot plate to make different temperatures of water, then measures the amount of dissolved oxygen present in the sample.

Hypothesis: Oxygen levels will depend on the temperature water.

Conclusion paragraph (where the numbers reference the question the statement answers above. Do NOT include the numbers in your narrative): The purpose of this experiment was to see the effect of changing water temperature on the amount of dissolved oxygen that it carries (1). The coldest temperature water had the most oxygen in it – about 6.50 mg/L at 5°C, while the warmest temperature water had the least oxygen in it and the trend was linear; as the temperature increased, the amount of available oxygen decreased (2). This data confirms that the oxygen levels change based on the water temperature (3) and was due to the solubility of oxygen decreasing as the temperature increased [Hibbeler, 2001]. Although the data trended properly, the oxygen testing procedure was too long causing the water to warm which increased the variability and made accurate temperature measurements difficult. Perhaps future tests could use an alternate method to measure oxygen levels to minimize temperature drift (4). Additionally, future experiments could test for other factors that affect oxygen levels in water – perhaps adding plants to the aquarium could affect oxygen levels due to the oxygen they make in photosynthesis (5).

This contains more details than an abstract would and focuses on results and recommendations. However, you do not want to state undeliverable conclusions. For example, the fuel cells may offer the promise to replace propane-driven forklifts, but you do not want to imply that they will save the World from greenhouse gases or make the manufacturing facility LEED (Leadership in Energy and Environmental Design) certified.

The conclusions should always end with the main takeaway. When you have objectives, the recommendations should be tied to those. However, when running an experiment, there may be something that should be changed or extended. This may be collecting more data to have more samples for better statistics, ensuring steady-state was reached, a different method, etc. It is important to identify the shortcomings of your results if any exist.

References: Citing the work of others
Giving credit to others for their ideas (and not plagiarizing) is critical. Using citations and a reference list is how you give others credit in technical writing. As a writer, you have a responsibility to:
1. consult and analyze sources that are relevant to the topic of inquiry,
2. clearly acknowledge when you draw from the ideas or the phrasing of those sources in your own writing, and
3. learn and use appropriate citation conventions.

When you fail to adhere to these responsibilities, you may intentionally or unintentionally "use someone else's language, ideas, or other original (not common-knowledge) material without properly acknowledging its source"(http://www.wpacouncil.org/aws/CWPA/pt/sd/news_article/272555/_self/layout_details/false) and that is plagiarism. (https://www.montana.edu/writingcenter/faculty/faculty_resources/syllabus-and-plagiarism.html)

Citation Location
When finding resources for backing up your claims and for information, it is best to reference peer-reviewed sources such as textbooks and journal articles. It is highly recommended that website references are kept to a minimum and only are used if a peer-reviewed source cannot be found. That is, you may be able to find

an equation or data on Wikipedia but referencing a textbook or journal article is a much better and reliable option.

When you are citing a resource, the generally accepted method is to do an in-line citation, where the citation is placed next to the information you referenced. This can be in the middle of a sentence if it is one thing or might be at the end of the sentence containing the reference material. For example, these made-up statements show how you want to reference material as close to mentioning it as possible:

- "Previous work has shown that cylindrical,[1] spherical, and powder[2] catalysts have the least amount of mass transfer diffusion in these cases."
    - Note that reference 1 is for the cylindrical shape, but reference 2 contains the information for both spherical and powder catalysts.
- "In previous work, Pt, Pd, and Ru catalysts all had similar activation energies for the reaction.[1]"
    - In this case, reference 1 contained information on all three metals and their activation energies.

It is important to not reference statements that do not come from the resource and to give credit for all the material from a reference. For example, if the statement "have the least amount of mass transfer diffusion" came from a source, that should be referenced too. In the case of the second statement, if the source does not mention "similar activation energies for the reaction" you do not want to attribute this information to the source and would move the citation behind the word "catalysts."

Numerical Formatting
For a numerical bibliography, references are numbered in the order they appear in the text and can use other notations as well, such as:

- The oxygen content decreases with temperature due to decreased solubility.[1]

- The oxygen content decreases with temperature due to decreased solubility [1].

- The oxygen content decreases with temperature due to decreased solubility.[1]

The reference list would then list the references in numerical order, which is the order they appear in the text, using standard APA formatting like:

1. Hibbeler, R. C., *Engineering Mechanics. Statics & Dynamics*, 9th ed., Prentice-Hall, Inc., Upper Saddle River, New Jersey (2001), p. 82.

In those samples, numerical citations are used, which is easy to do using citation software like EndNote, Zotero (free), Paperpile, and Mendeley (free). If you have written a journal article before, you may have experience with citation software, a program that makes inserting references easy. You are welcome to use these programs, but this section does not go through how to use them although online tutorials are available to walk you through using the software. Numerical citations are used most often because if the same author has multiple papers from the same year, more information is needed using the alphabetical format as discussed below. Additionally, if you are referencing multiple sources, using numerical citations look like [1-4] versus (Hibbeler, 2001a; Hibbeler, 2001b; Wettstein, 2018; Sorenson, 2010).

Alphabetical Formatting
Alternatively, and what works well for team documents, is using the first author's last name and the year of the publication in the format of (Last name of author, year) such as:

- The oxygen content decreases with temperature due to decreased solubility (Hibbeler, 2001).

The reference list would then list the references in alphabetical order, regardless of the order they appear in the narrative using APA formatting like shown above (but without the numbers):

Hibbeler, R. C., *Engineering Mechanics. Statics & Dynamics*, 9th ed., Prentice-Hall, Inc., Upper Saddle River, New Jersey (2001), p. 82.

If there are multiple authors, the first author's last name is used followed by "*et al.*" For example, (Wettstein *et al.*, 2018). To distinguish between references with the same last name and year, letters are often used after the year. For example, (Sorenson *et al.*, 2015a) and (Sorenson *et al.*, 2015b). One final difference with last name citations compared to numerical citations is that if you mention the author's last name in the narrative, only the year needs to be cited. For example, "In the paper by Wettstein *et al.* (2018), the catalyst deactivated after 26 h on stream." If last name notation is used, the references should appear in alphabetical order in the reference section at the end of the paper, which makes ordering references in the bibliography easier, especially in group work, since you do not have to renumber your references if you add a new reference to the paper.

Reference list/Bibliography
These two reference styles are generalized versions of the styles presented in Turabian, K. L., *A Manual for Writers of Term Papers, Theses, and Dissertations*, that can be found here: http://jcs.edu.au/wp-content/uploads/2016/09/A-manual-for-writers-of-research-papers-theses-and-dissertations.pdf in Chapter 15.3. Also presented in that same book starting in Chapter 16.1 is how the references should be cited in the bibliography. **Many citation styles are acceptable as long as they contain all the necessary information AND remain consistent in your reference list.**

Appendices
An appendix acts as a supporting document to the main technical communication. It is important to not only provide proof of your work, experimental data, calculations, and other details, but it is imperative to communicate to a reader how to use the appendix appropriately as they navigate your technical communication. Appendices should be broken down into appropriate sections that are numbered. Refer to said section within the body of your communication to direct your reader to a specific area of your appendix for additional information or clarification, such as "*refer to Appendix section A3, Figure A2 for…*" However, do not put critical figures and tables in the appendix. Figures and tables that explain the data and conclusions should appear in the narrative. For example, perhaps you made three runs on each of three packed beds. A figure in the narrative could show one run of each packed bed that is representative and then the other two runs for each packed bed could appear in the appendix.

An appendix/supplemental information document for a final technical communication typically contains the below items:
- Data collection sheets and raw data
- If the calculations are complex, sample calculations using experimental data are valuable for the reader
- Formatted spreadsheets or tables of calculations and statistics, and Excel analysis tables
- Supporting figures and/or tables that are not in the body of the communication
- Appropriate supporting methods sections (step-by-step protocols, safety, etc.)

The appendix should contain extra information that is not necessary to understand your results and conclusions (that information should be in the narrative). Keep in mind that many people will not even look at the appendix, but it is still important to keep it organized and well formatted for those that do reference it.

Appendix

Excel Tutorial: Using Equation Editor in Microsoft Word
Using the equation editor in *Microsoft Word* will help create professional looking documents. Equation editor inserts equations using fonts and symbols used in textbooks and academic research journals, which facilitates interpretation and understanding of the equation.

For example, say you want to insert the following equation into a *Word* document

$$\frac{B - D_i}{C + (E - \pi)} + C = \varphi^3 \tag{1.1}$$

Step 1: Insert an Equation Region
Equations are contained in an *equation region* as shown in Figure 1.1. An equation region is analogous to a text box and the equation region is only displayed when the equation is being edited. Clicking on an equation region activates equation edit mode and the **Equation Tools tab** on the Ribbon shown in Figure 1.2 becomes available. To insert an empty equation region into your Word document, position the cursor at the desired location.
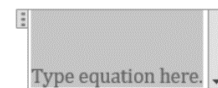
Figure 1.1: Example of an equation region.
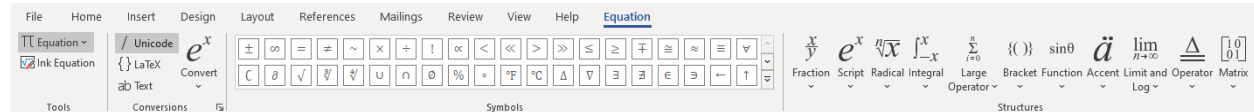
Figure 1.2: Equation Tools / Design Tab located on the Ribbon.

Click on the "**Insert**" tab, then click on the "**Equation**" icon as shown in Figure 1.3.
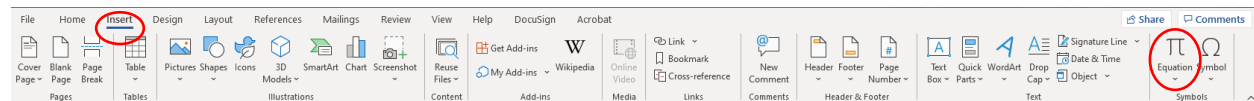
Figure 1.3: Insert tab is used to find Equation icon.

An empty equation region (Figure 1.1) will be created at the cursor location.

A handy shortcut is to insert an equation using "**Alt**" + "**=**" (press the "**Alt**" key and the "**=**" key simultaneously.

Step 2: Use Equation Tools Structures to Create Basic Equation Layout
If the equation just uses simple (*i.e.*, non-subscripted and non-superscripted) variables, simply start typing into the equation region. Most engineering equations are sufficiently complex that an equation structure should be created initially. The ribbon's **"Equations Tools/Design tab"** (Figure 1.2) allows the structure of the equation to be created. The Design Tab is displayed whenever an equation region is inserted into a Word document.

The Design tab's **"Structures Group"** offers a series of drop-down menus that are used to create the basic structure of the equation: an example of an equation structure with placeholders is shown in Figure 1.4. To create this structure the **"Fraction"** drop-down was used to create the numerator and denominator placeholders, and the **Script** drop-down was used to create the subscripted symbol placeholders. The order

to insert placeholders does take some getting used to. If you are not getting the results you want, it may be best to just restart the equation insertion process.

$$\frac{B - }{} + C =$$

$$\frac{B - D_i}{C + (E - \pi)} + C = \varphi^3$$

Figure 1.4: *Top: An equation structure showing a subscript, a superscript, and fraction placeholders. Bottom: Completed equation with placeholders populated to match Equation 1.1.*

Step 3: Fill in the placeholders to complete the equation

Select a placeholder with the mouse and enter the symbols either from the keyboard, or from the palettes available in the Design tab's **"Symbols Group".**

By default, the **Symbols Group** shows some commonly used math symbols, called the **"Basic Math Palette"**. There is a scroll bar at the right side of the palette to access additional symbols. You can also click the **"More"** button at the bottom of the scroll bar to show the entire **Basic Math palette**, or select one of the other available symbol palettes, including:

- Basic Math
- Greek Letters (can also be entered by name, *e.g.,* '\pi ' gives π, insert a space after the name)
- Letter-like Symbols
- Operators (can also be entered using a slash, '\' before then name: '\sum')
- Arrows
- Negated Relations
- Scripts
- Geometry

For most reports and papers, labeling the equations in order is necessary so that you can refer to the equation later in the document. For example, in the theory or background section of the document you might put multiple equations that will help the reader understand your results. To easily refer to the equations, labels should be added on the right-hand side of the equation. One of the most useful short-cuts is typing a hashtag ("#") followed by the equation number (in the equation box) like

$$E = mc^2 \#(1)$$

and then after hitting "**Enter**"

$$E = mc^2 \hspace{10cm} (1)$$

whatever follows the hashtag becomes right justified automatically. Typically, numbers are used for labeling. In the results section, you could now say "The mass of the object directly affects the energy contained in the system as described in Equation 1." This helps the reader analyze and better understand your data in relation to the theory.

There are keyboard short-cuts that reduce the number of clicks required compared to using the *Excel* **Toolbar**.

Cliff Notes/Shortcuts

To enter an equation box: Press "**Alt**" + "**=**" keys

After you type the following, press the spacebar:

- Subscript: "**_X**" for $_X$
- Superscript: "**^X**" for ^{X}
- Greek letters: "**\greekletter**" so "**\pi**" is π
- Operators: "**\sum**" for $\sum$

You can also use "**/**" followed by enter for a fraction like $-$

One of the most useful short-cuts is typing a hashtag ("**#**") followed by the equation number (in the equation box), then pressing "**Enter**" to center the equation center and right-justify the equation number.

$$E = mc^2 \#(Eq.A)$$

and then after hitting "**Enter**":

$$E = mc^2 \hspace{6cm} (Eq.A)$$

Excel Tutorial: Using Text Boxes

Inserting *text boxes* may make reports look more professional and usually take up less space since the text can be wrapped around the text box.

NOTE: Text boxes sometimes move with the addition of another text box or text. A way to avoid this is to complete the text and then insert the text boxes at the end of the process.

To insert a *text box* (see Figure 2.1 below):

1) Go to the "**Insert**" tab of *Word*

2) Select the "**Text Box**" drop-down menu and select "**Simple Text Box**"
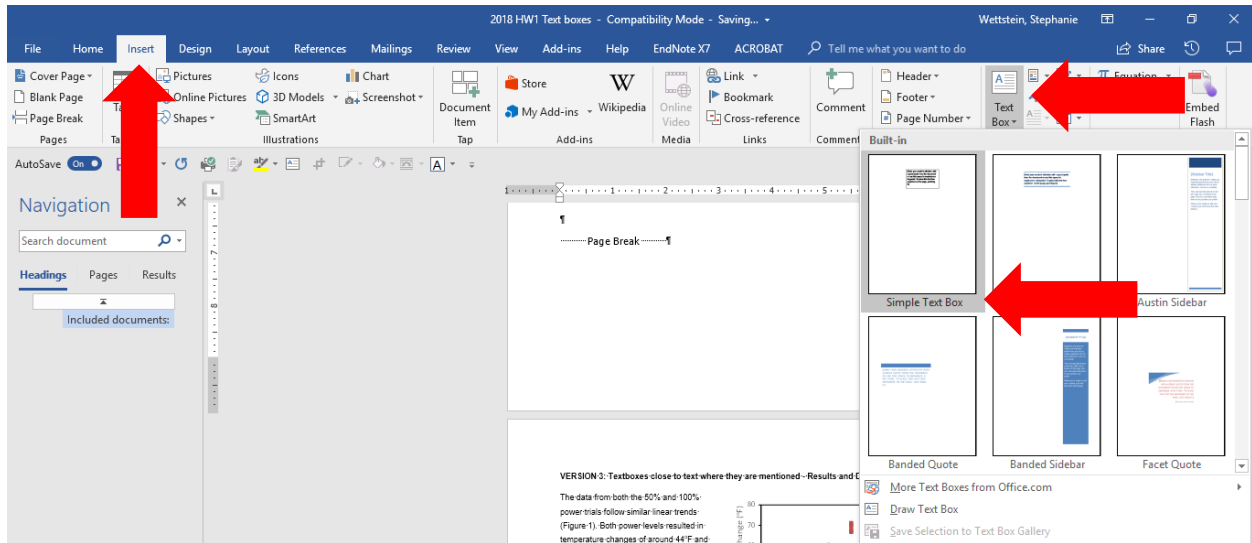
Figure 2.1: Steps to insert a text box in a document.

3) Word inserts a text box at the cursor. The default when inserting a text box in this way is to have text wrapping around the text box, which is shown in Figure 2.2.
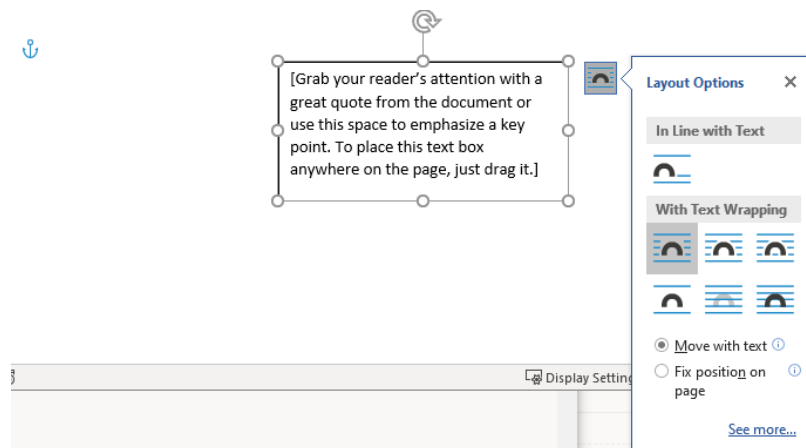
Figure 2.2: Accessing text wrapping options around a text box.

Change how the text is displayed around the text box by clicking on the border of the text box selecting the "*Layout Options*" icon that should appear near the upper-right corner of the text box. NOTE: The text box MUST be "**active**," which means you clicked on the border to see the "*Layout Options*" dialogue box.

An alternative way to access text wrapping controls is shown in Figure 2.3. Click on the *Layout* link and then the *Wrap Text* icon to bring up the dialogue box to control text wrapping.
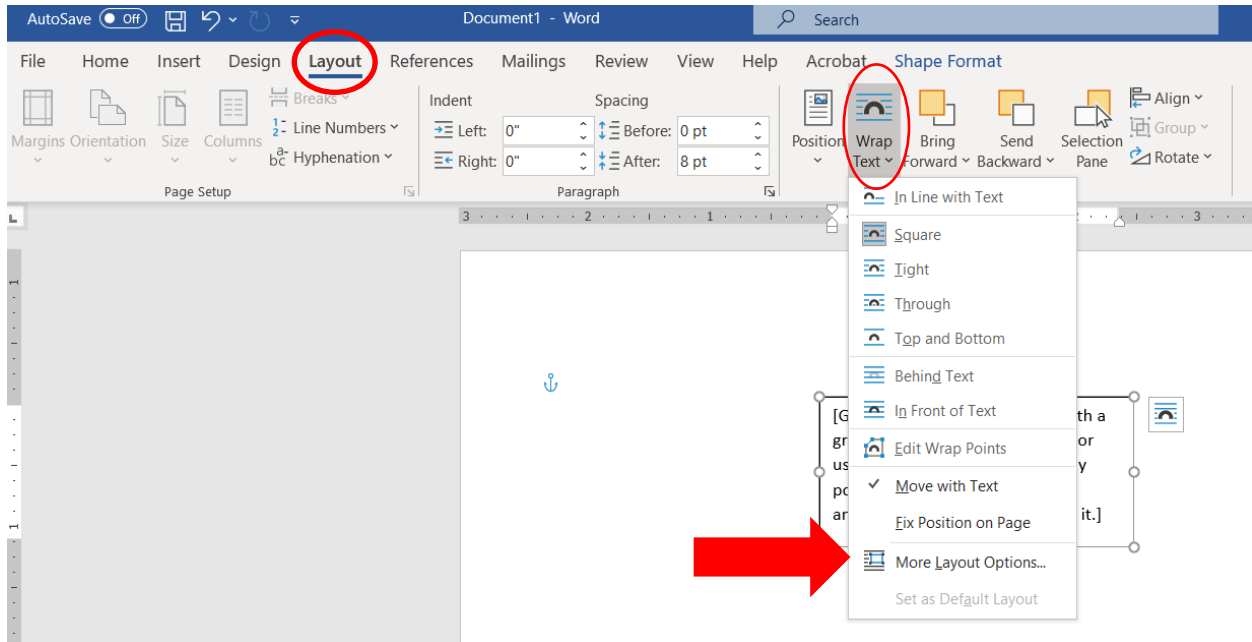
Figure 2.3: Alternative method to access text wrapping options around text boxes.

For example, if you have a large figure and want the text only above and below the text box, select "**Top and Bottom**." Figure 2.4, below, has no text to the left or right of the text box.
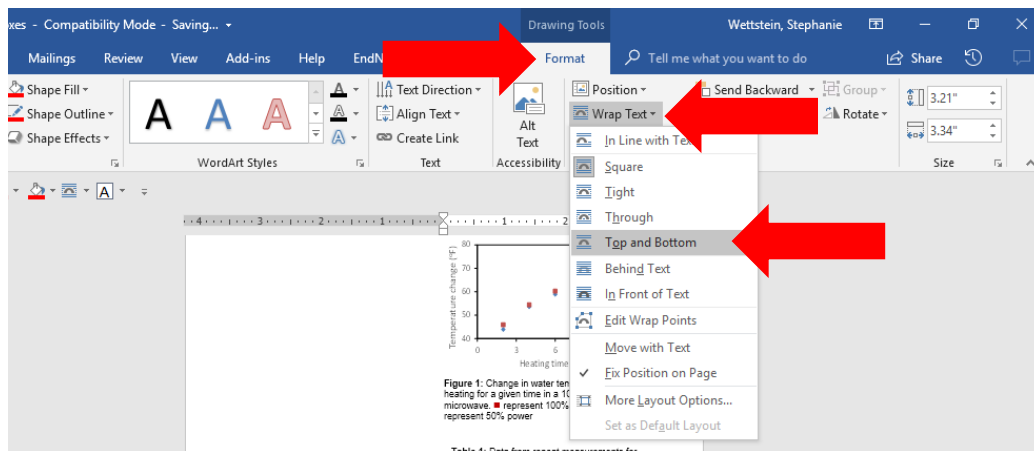
Figure 2.4: Example of "Top and Bottom" wrapping that prevents text from being placed to the sides of the text box.

4) Next, paste the figure/table and caption into the text box so that you can properly resize the text box using the white circles along the border of the text box.

5) There are two options for positioning an active text box. The first option is to move it manually to the desired location or use the *Position* tools in *Word*. The second option is accessed by returning to Figure 2.3

and clicking on "***More Layout Options…***". This brings up the full "***Layout***" dialogue box shown in Figure 2.5. Note the two additional tabs that allow control of the **Position** and the **Size** of the text box. Figure 2.6 shows the options available to automatically position the text box in a desired location.
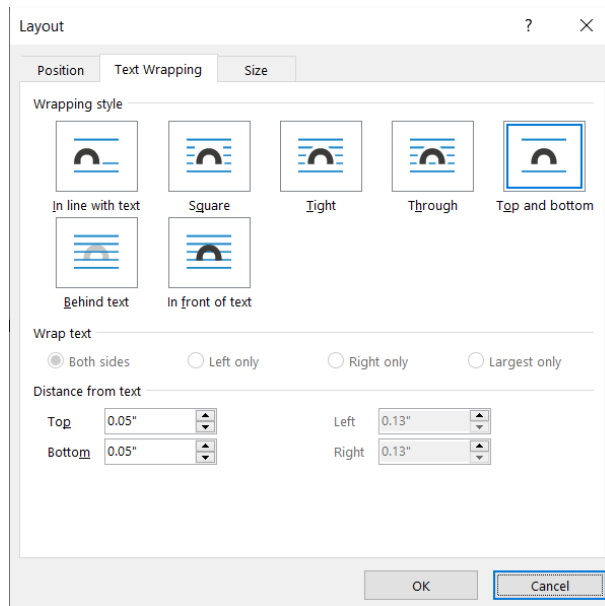
Figure 2.5: *Layout dialogue box. Note tabs for "Position," "Text Wrapping" (shown here), and "Size."*
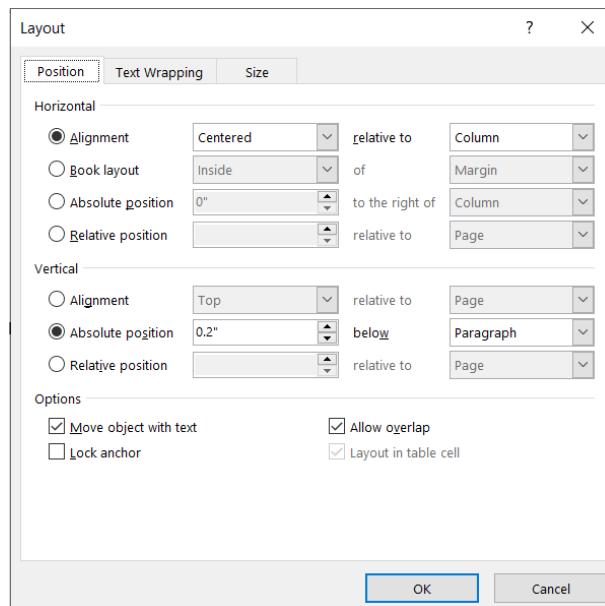
Figure 2.6: *Position tools for text box in Layout dialogue box.*

Using the "**With Text Wrapping**" options create a professional look. The text box can still be manually moved even if it has been initially placed by one of the **With Text Wrapping** options.

6) If desired, the text box border can be removed. *Right-click* on the text box to access the menu shown in Figure 2.7. Select the **Format Shape** at the bottom of this drop-down menu. The **Format Shapes, Shape**

Options provides *Fill* and *Line* options. Select the *No line* radio button in the *Line* options. This removes the border of the text box as shown in Figure 2.8.
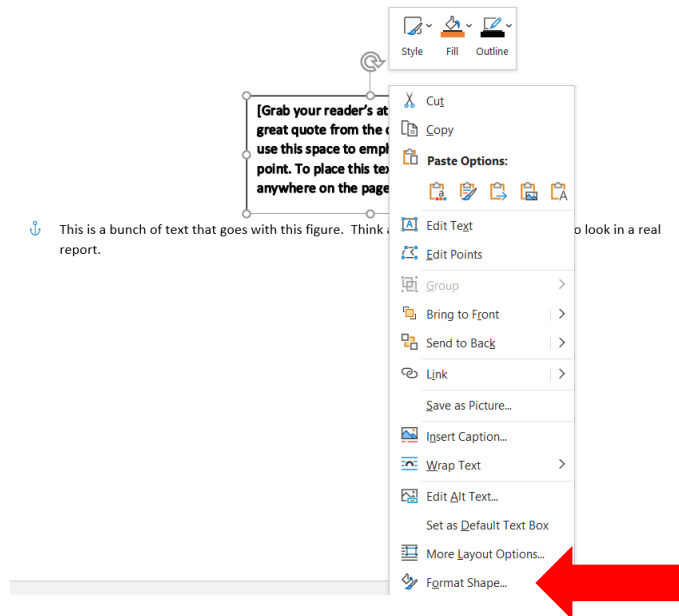
Figure 2.7: *Menu accessed by right-clicking on the text box.*
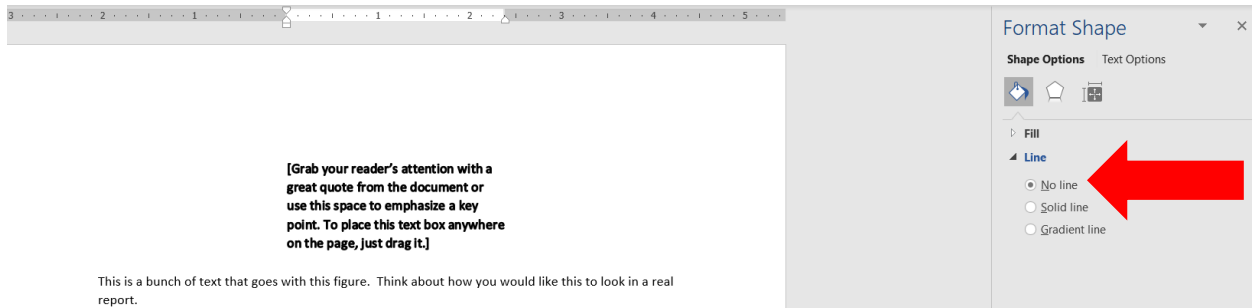
Figure 2.8: *Selecting the "No line" option removes the border from the text box.*

Text Box Management: Three Examples

The next section contains three versions of the same material illustrating the impact that different text box formats and locations have on the appearance of the document.

Version 1: The text boxes are inserted, then the position function is used to put Figure V1-1 in the top right corner and Table V1-1 in the middle right of the page. Then each box was activated and manually moved to the desired location and the text boxes did not overlap.

Version 2: The position function is used to put Figure V2-1 in the lower-left corner and Table V2-1 in the lower-right corner. Some manual adjusting had to be done so the figure and table text boxes did not overlap. Also, some resizing was done so both would fit.

Version 3: No text boxes were used. The figure and table are both inserted with the text wrapped.

In your option, which version gives the best results?

VERSION 1: Textboxes Near Text Description

Results and Discussion

The data from both the 50% and 100% power trials followed similar linear trends (Figure V1-1). Both power levels resulted in temperature changes of around 44°F and 75°F after 2 minutes and 10 minutes of heating, respectively. Statistical analysis confirms there is no significant difference at the 95% confidence level between the 50% and 100% power levels at both 8 minutes and 10 minutes. This was determined since the p-value for the equal variance t-test was greater than the alpha value of 0.05. This validates the null hypothesis that the power of the microwave at 50% was equivalent to the power at 100%. At 8 minutes, the temperature change was 69.0 ± 2°F for 100% power and 70.2 ± 0.7°F at 50% power. The temperature change at 10 minutes and 100% power was 75 ± 1°F, while at 50% power, the temperature change was 76.0 ± 0.9°F (Table V1-1). Therefore, in this microwave, heating at 50% power or 100% power provided an equivalent amount of energy to the water indicating that the microwave was not working properly. Normally at 50% power, the temperature change of the water should be approximately 50% of the change observed at 100% power.
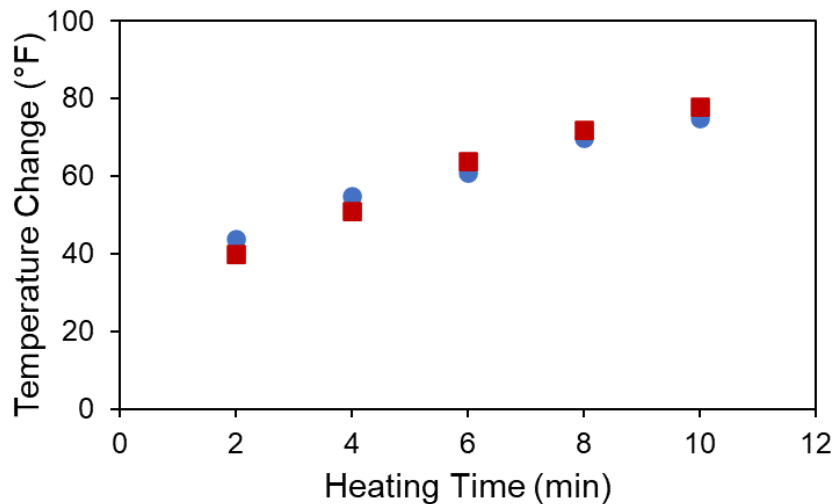

Figure V1-1: Change in water temperature after heating for a given time in a 1000 W microwave. ■ represents 100% power and ♦ represents 50% power

Table V1-1: Data from repeat measurements for the microwave experiment.

Power (%)	Heating time (min)	Average temperature change (°F)
50	8	70.2±0.7
50	10	76.0±0.9
100	8	69.0±2
100	10	75.0±1

Additional time did increase the temperature change. At both the 50% and 100% power levels, there was a significant difference in the temperature change (at the 95% confidence level) comparing 8 minutes to 10 minutes of heating time. The p-value of the equal variance t-test was less than 0.05, indicating a statistically significant temperature difference between 8 and 10 minutes. At 50% power, the temperature change increased from 70.2 ± 0.7°F to 76.0 ± 0.9°F and at 100% power, the temperature change increased from 69.0 ± 2°F to 75.0 ± 1°F.

VERSION 2: Textboxes at Bottom of Page

Results and Discussion

The data from both the 50% and 100% power trials followed similar linear trends (Figure V2-1). Both power levels resulted in temperature changes of around 44°F and 75°F after 2 minutes and 10 minutes of heating, respectively. Statistical analysis confirms there is no significant difference at the 95% confidence level between the 50% and 100% power levels at both 8 minutes and 10 minutes. This was determined since the p-value for the equal variance t-test was greater than the alpha value of 0.05. This validates the null hypothesis that the power of the microwave at 50% was equivalent to the power at 100%. At 8 minutes, the temperature change was 69.0 ± 2°F for 100% power and 70.2 ± 0.7°F at 50% power. The temperature change at 10 minutes and 100% power was 75 ± 1°F, while at 50% power, the temperature change was 76.0 ± 0.9°F (Table V2-1). Therefore, in this microwave, heating at 50% power or 100% power provided an equivalent amount of energy to the water indicating that the microwave was not working properly. Normally at 50% power, the temperature change of the water should be approximately 50% of the change observed at 100% power.

Additional time did increase the temperature change. At both the 50% and 100% power levels, there was a significant difference in the temperature change (at the 95% confidence level) comparing 8 minutes to 10 minutes of heating time. The p-value of the equal variance t-test was less than 0.05, indicating a statistically significant temperature difference between 8 and 10 minutes. At 50% power, the temperature change increased from 70.2 ± 0.7°F to 76.0 ± 0.9°F and at 100% power, the temperature change increased from 69.0 ± 2°F to 75.0 ± 1°F.
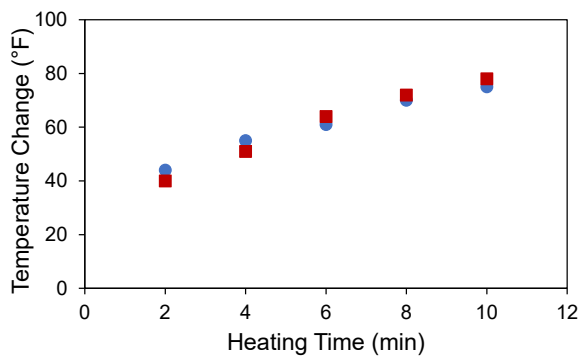
Figure V2-1: Change in water temperature after heating for a given time in a 1000 W microwave. ■ represents 100% power and ◆ represents 50% power

Table V2-1: Data from repeat measurements for the microwave experiment.

Power (%)	Heating time (min)	Average temperature change (°F)
50	8	70.2±0.7
50	10	76.0±0.9
100	8	69.0±2
100	10	75.0±1

Results and Discussion

The data from both the 50% and 100% power trials followed similar linear trends (Figure V3-1). Both power levels resulted in temperature changes of around 44°F and 75°F after 2 minutes and 10 minutes of heating, respectively. Statistical analysis confirms there is no significant difference at the 95% confidence level between the 50% and 100% power levels at both 8 minutes and 10 minutes. This was determined since the p-value for the equal variance t-test was greater than the alpha value of 0.05. This validates the null hypothesis that the power of the microwave at 50% was equivalent to the power at 100%. At 8 minutes, the temperature change was 69.0 ± 2°F for 100% power and 70.2 ± 0.7°F at 50% power. The temperature change at 10 minutes and 100% power was 75 ± 1°F, while at 50% power, the temperature change was 76.0 ± 0.9°F (Table V3-1). Therefore, in this microwave, heating at 50% power or 100% power provided an equivalent amount of energy to the water indicating that the microwave was not working properly. Normally at 50% power, the temperature change of the water should be approximately 50% of the change observed at 100% power.

Figure V3-1: Change in water temperature after heating for a given time in a 1000 W microwave. ■ represents 100% power and ◆ represents 50% power

Additional time did increase the temperature change. At both the 50% and 100% power levels, there was a significant difference in the temperature change (at the 95% confidence level) comparing 8 minutes to 10 minutes of heating time. The p-value of the equal variance t-test was less than 0.05, indicating a statistically significant temperature difference between 8 and 10 minutes. At 50% power, the temperature change increased from 70.2 ± 0.7°F to 76.0 ± 0.9°F and at 100% power, the temperature change increased from 69.0 ± 2°F to 75.0 ± 1°F.

Table V3-1: Data from repeat measurements for the microwave experiment.

Power (%)	Heating time (min)	Average temperature change (°F)
50	8	70.2±0.7
50	10	76.0±0.9
100	8	69.0±2
100	10	75.0±1

NOTE: Do NOT have tables "break" (*i.e.*, display) across two pages. If Table V3-1 would have been inserted on the previous page, the top half might have been the previous page and the bottom half on this page. It is better to insert it at the end of the next paragraph as in this case, even though it was mentioned in the previous paragraph.

If you have a table so large that it will not fit on a page, even after column and font sizes are adjusted, create a second table on the next page with the same columns and same table title with (Continued) after the title. Be sure that the column titles (i.e., column headers) are shown on the second page as well.

Excel Tutorial: Formatting Excel Charts for a Professional Look

The default *Excel* formatting of charts can be improved with a few minor changes. The default *Excel* chart format is shown in Figure 3.1 so it can be directly compared to the preferred format shown in Figure 3.2.

The modified format (Figure 3.2) is much easier to read and understand. The following steps will transform the chart from *Excel's* default to the desired format.

1) Place time and temperature change data into two columns in *Excel* (Figure 3.3). Title each column with the desired axes titles.

Axes Titles: The temperature axis label requires a "degree" symbol. One choice is to use a superscript "o". There is a *symbol* for degrees which looks more professional.

As shown on Figure 3.3, go to the **Insert** tab, click **Symbol**, then in the **Subset** drop-down, select **Basic Latin** (the drop-down title will change to **Latin 1 Supplement** when the **Degree Sign** is selected). Click on the **Degree Sign** and then hit **Insert**. Close the "**Symbol**" box using the "**X**" in the upper-right corner. This allows multiple symbols to be inserted without reopening the **Symbol** selection box.

Figure 3.1: Default Excel chart format.
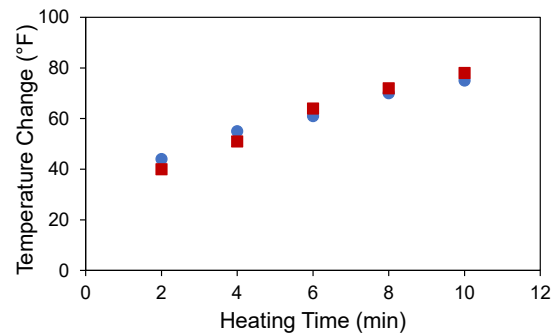
Figure 3.2: Improved chart format.

The short-cut for inserting a **Degree Sign** (for both *Word* and *Excel*) is to hold down the **Alt** key and type 0176 (**Alt+0176**). The degree symbol should appear when the **Alt** key is released.
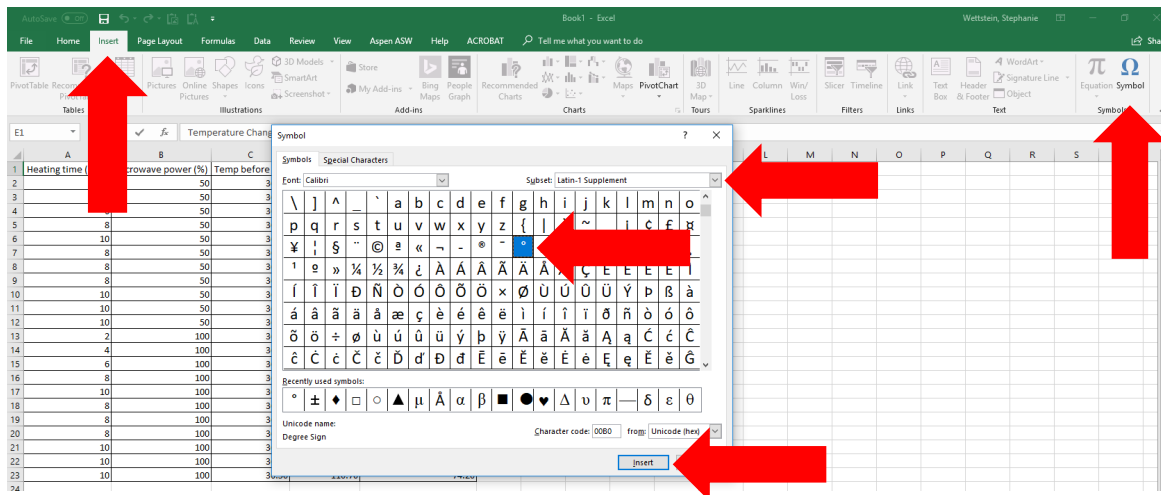
Figure 3.3: Insert a Degree Sign from symbol library.

NOTE: When documents are converted to *pdf* format, the degree symbol occasionally changes to some undesired character. If this occurs in a figure, paste the figure as a *Bitmap* into *Word*. (Select **Bitmap** (or
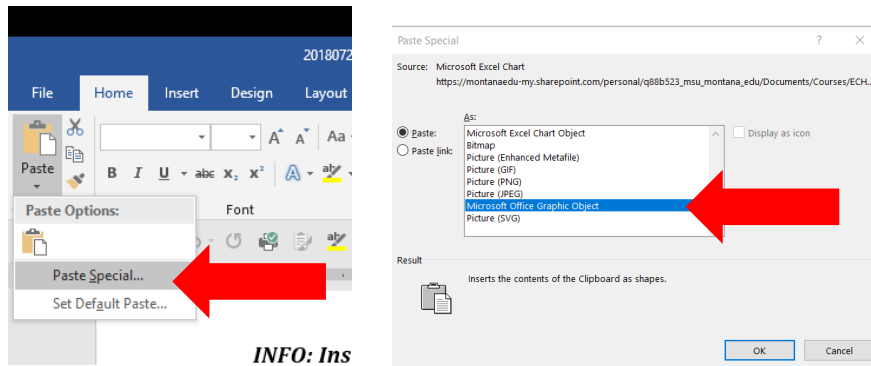
Figure 3.4: *Formatting options when pasting a picture.*

If it is not a figure, the alternate way is to insert it as a superscript "o." In *Excel*, highlight the text you want to superscript, then right-click and select "***Format Cells…***" From the menu that pops up, click the box next to "***Superscript***", then click "***OK***" As shown in Figure 3.5.
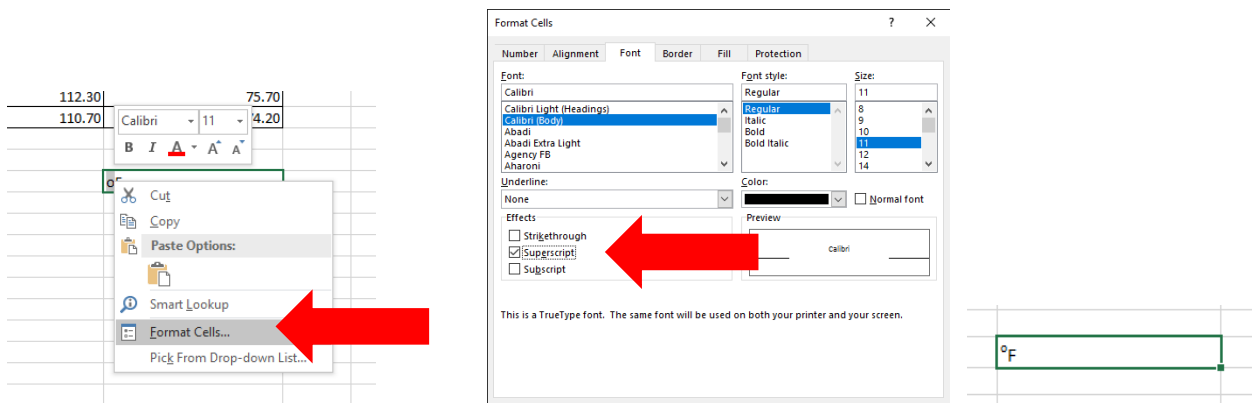
Figure 3.5: *Creating a superscripted character in Excel.*

2) Highlight the data and make a chart. Excel will use its default formatting, which should be edited to make it easier to read.

3) **Removing the border:** Make the chart active by clicking near the edge in the white space as shown in the left image in Figure 3.6 below. Then, under the "***Format***" tab, click the "***Shape Outline***" drop-down and select "***No Outline***."
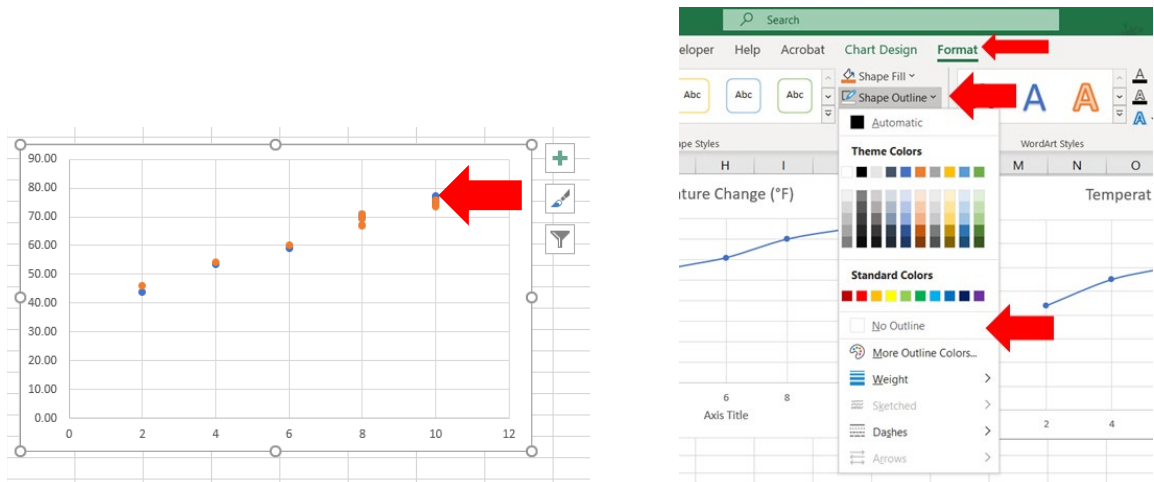
Figure 3.6: Removing the outline (border) from a figure.

4) **Remove the gridlines:** "Activate" the gridlines by clicking on them and then press the delete key on the keyboard. This will have to be done separately for both the horizontal and vertical gridlines (see Figure 3.7).
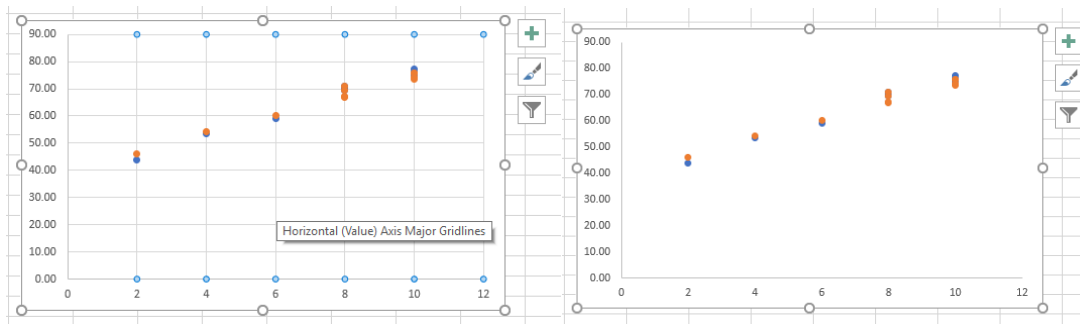
Figure 3.7: Click on gridlines to activate them. Then hit the delete key on the keyboard to remove them.

5) **Add a border to the chart area (not the outside):** Click in the area containing the data points to make it active, then from the "***Format***" tab, open the "***Shape Outline***" drop-down (see Figure 3.8) and select black.
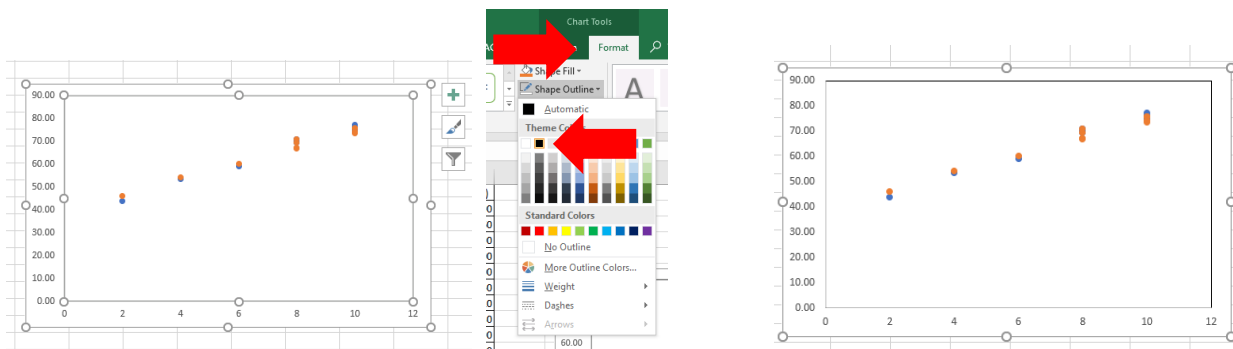
Figure 3.8: Adding an outline to the chart area.

The axes lines did not change color. Activate the axes by clicking on the axis values. (Same dropdown as above in Figure 3.8.) From the "*Format*" tab, open the "*Shape Outline*" drop-down and select black. Do this for both the *x* and *y* axis with the result shown in Figure 3.9.
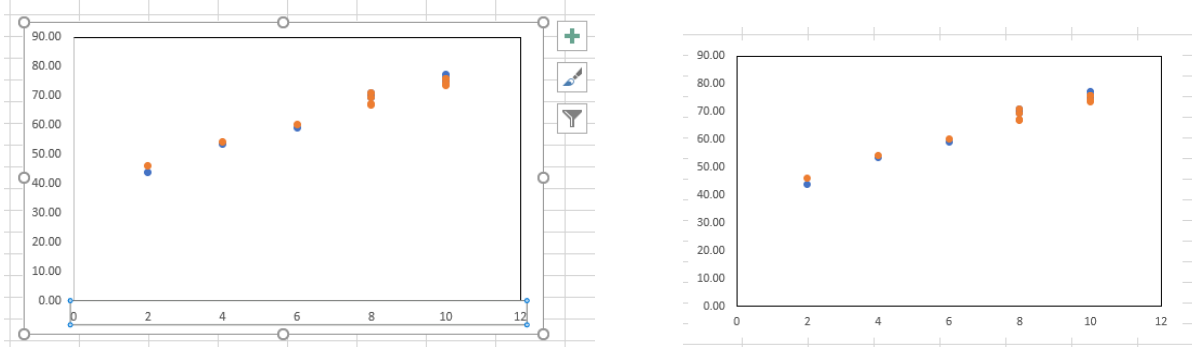
Figure 3.9: *Change axis color to black.*

6) **Adding tick marks:** Double-click an axis. The "*Format Axis*" side bar should pop up. Expand the "*Tick Marks*" menu, then select "*Major*" and select where you want them to appear from among "*Inside*", "*Outside*", or "*Cross*". The "*Outside*" option has been selected for the example shown in Figure 3.10. Do this for both axes.
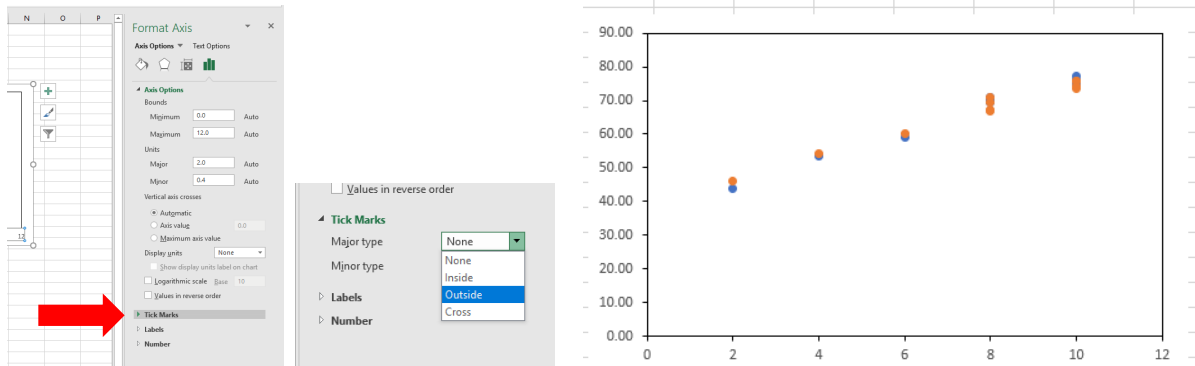
Figure 3.10: *Inserting tick marks on both axes using the "Outside" option for tick placement.*

7) **Formatting the font:** By default, the font is also in gray. Activate the entire chart by clicking just outside of the chart area, but inside the border and then from the "*Home*" tab, select the font color drop-down and pick black as shown in Figure 3.11.
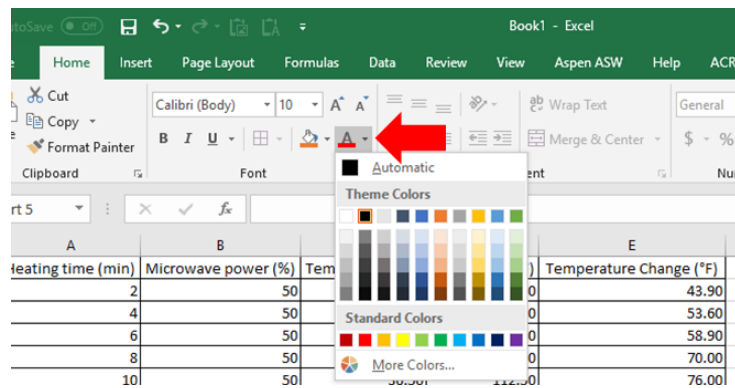
Figure 3.11: *Change font color of chart to black.*

The *y-axis* on Figure 3.11 has two decimal places. For the range of values from 0 to 90 it is appropriate to drop the decimal zeros. Activate the axis you want to edit and then in the "***Format Axis***" side bar, expand the "***Number***" menu and change the value in the "***Decimal places***" box. The result of changing it from 2 to 0 is shown in Figure 3.12.
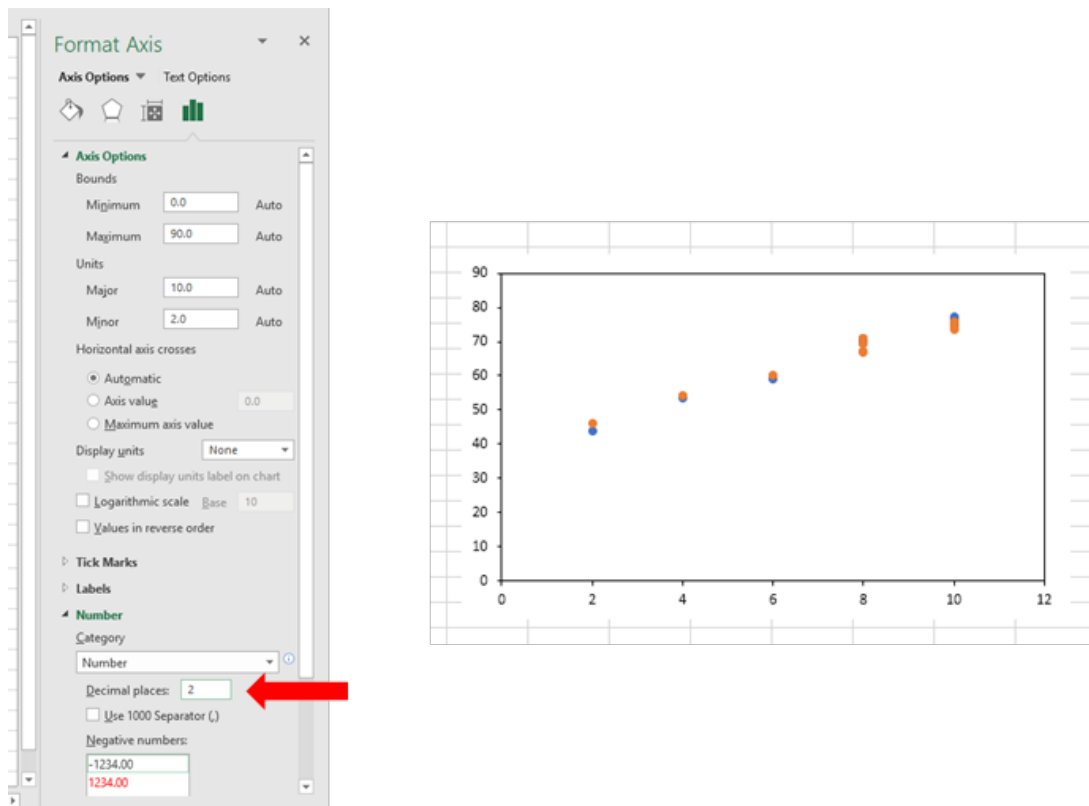
Figure 3.12: *Reducing the number of decimal places on axis values (compare with Figure 3.11).*

8) **Add axes labels:** Activate the chart and then click the green "**+**" symbol in the box just outside of the upper-right border of the chart. Click the "***Axis Titles***" check box in the "***Chart Elements***" drop down menu. Click in the text box that appears and type the axis title as shown in Figure 3.13.
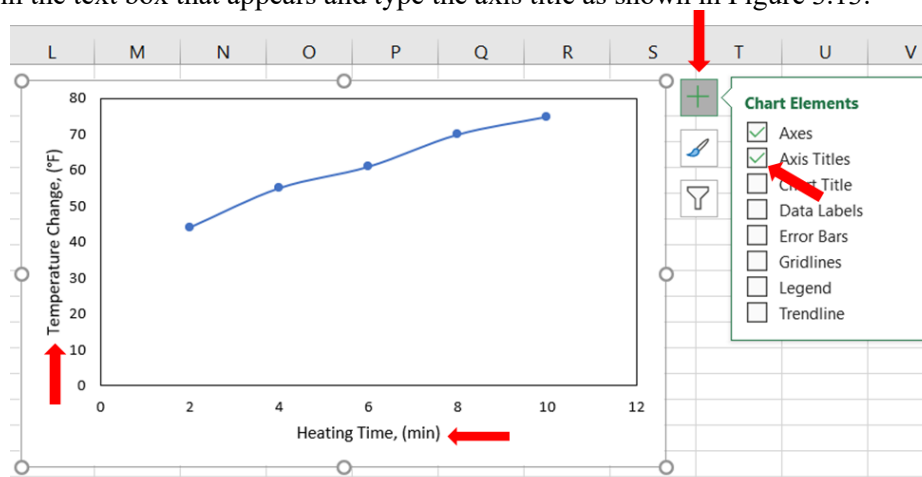
Figure 3.13: *Adding axis titles to a chart.*

It is often necessary to increase the font size ("***Home***" tab) to make it legible in *Word*. Activate the whole chart and then increase the *font size to 14* on the *y-axis title* as shown in Figure 3.14. The *x-axis* bounds

have been changed from 0-12 to 0-10 and the "***Major Units***" have been changed from every 2 units to every 1 unit using the "***Axis Options***" menu in the "***Format Axis***" sidebar as shown in Figure 3.14.
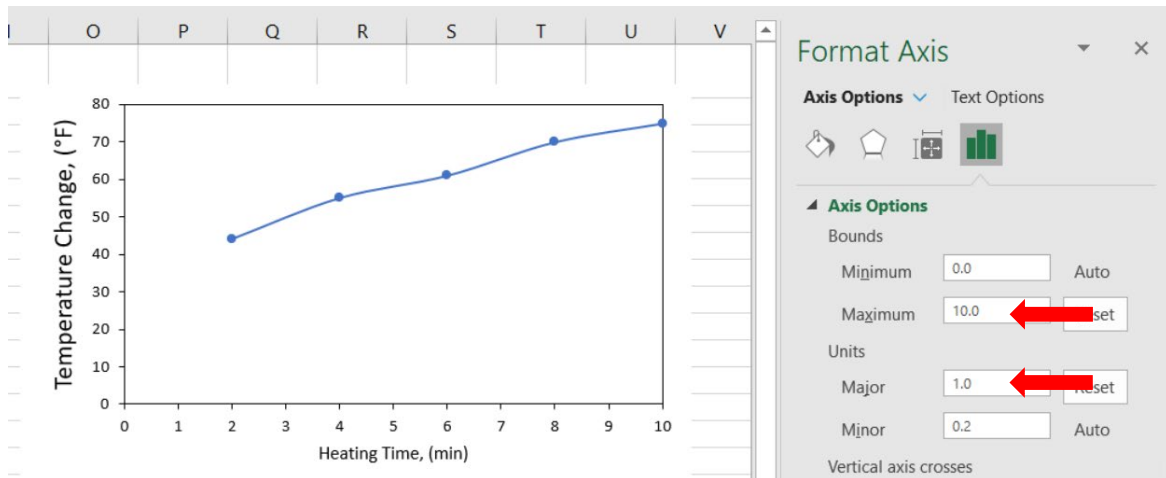
Figure 3.14: *Increasing font size on the y-axis title using the "Home" menu and changing the Maximum Bound on the x-axis from 12 to 10. Also changed the Major Units on the x-axis from 2.0 to 1.0 using the "Axis Options" drop-down menu.*

9) **Adjust chart markers if needed:** Data from a second run have been added and note that *Excel* made the markers for both data sets circles as shown in the left panel of Figure 3.15. Different marker shapes help clearly identify data sets, especially if the document is to be printed in black and white. Activate the "***Format Data Series***" sidebar menu by double-clicking the data series needing a change in marker style. Click the "***pouring paint can***" icon, then click "***Marker***" and expand the "***Marker Options***" menu. Click the radio button for the "***Built-in***" option and select the desired marker from the drop-down menu. Note that a square marker style has been selected. The size of the marker can also be changed. The color is selected under the "***Fill***" and "***Border***" menus from the lower section of the same side bar. (NOTE: The border color must also be changed if the fill color is changed. If this is not done, the marker will have an orange center with a blue border.)
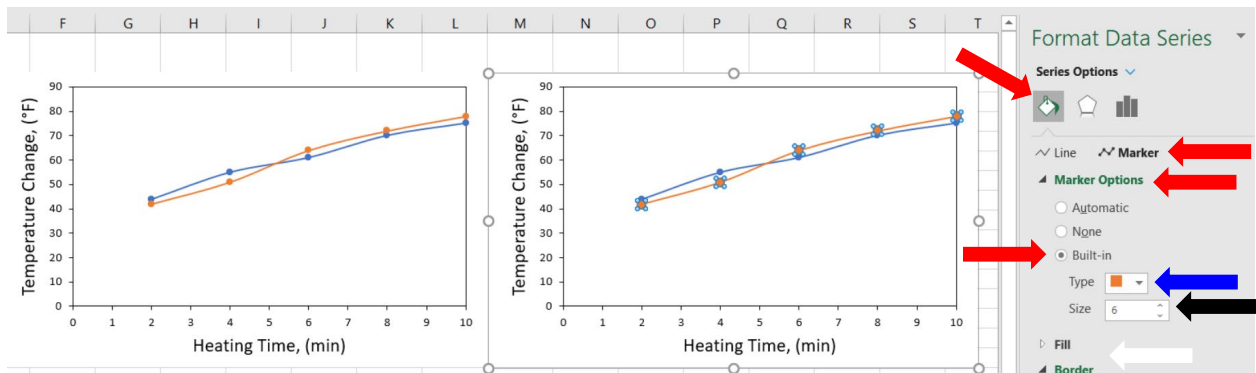
Figure 3.15: *Changing marker type (blue arrow), size (black arrow), and color (white arrow).*
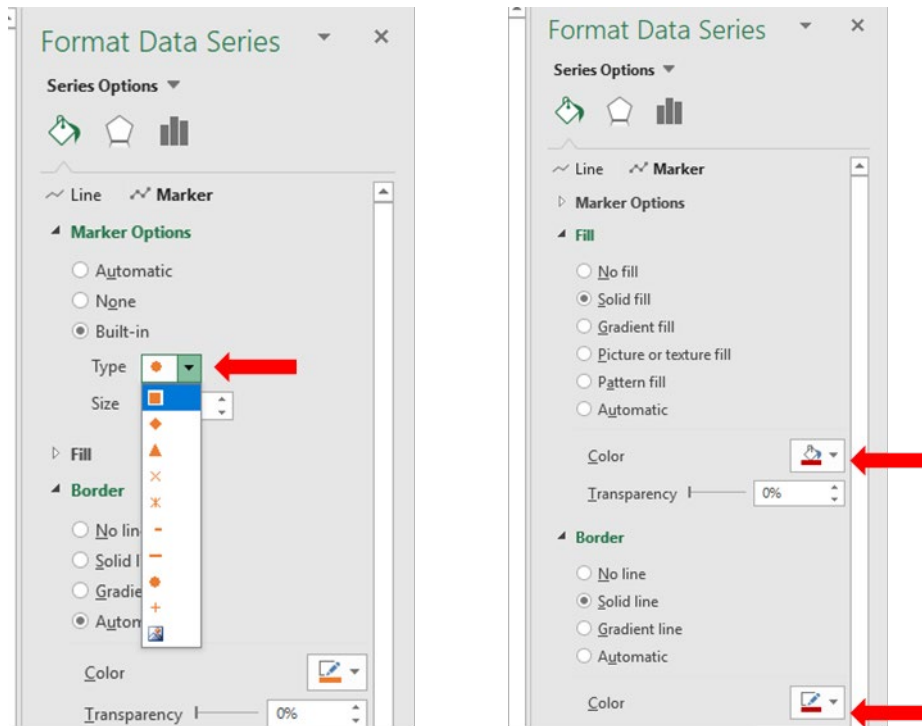
Figure 3.15 (expanded): *Changing marker type (options in lower left) and color (lower right).*

If it is desired to only show the data points, the line can be eliminated by activating the appropriate data series. Again, click on the "***pouring paint can***" icon, expand the "***Line***" option and click on the radio button for the "No line" option. Figure 3.16 shows the steps to remove the line from the data series with square orange markers.
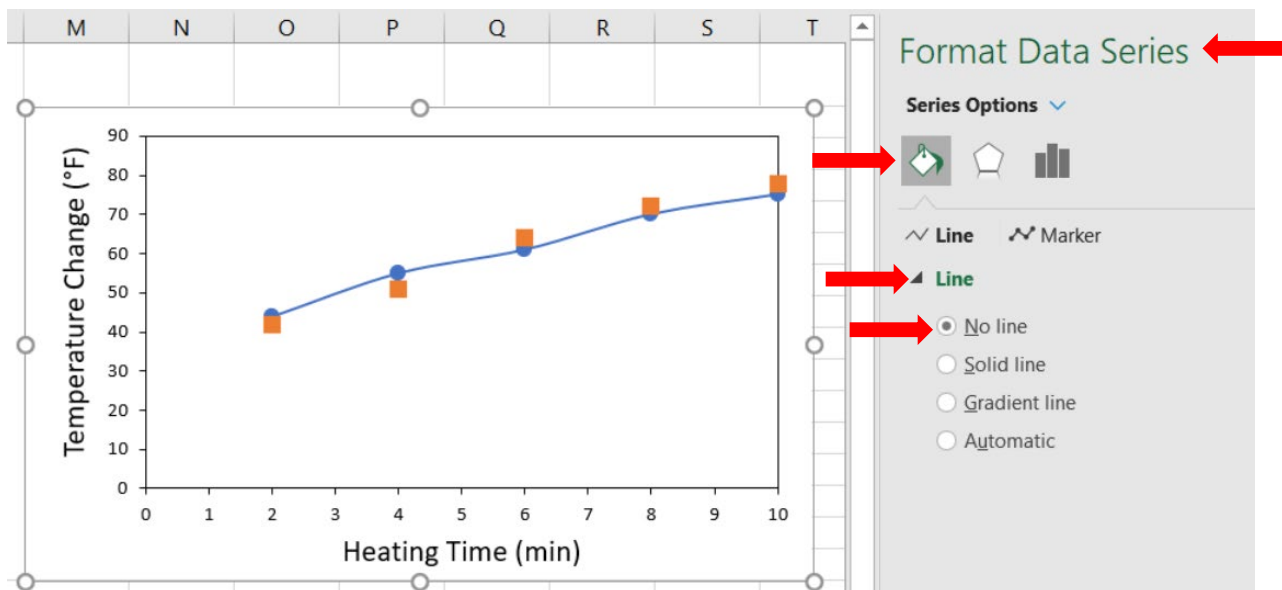
Figure 3.16: *Removing the line connecting discrete data points for the data series marked with orange squares.*

An analogous series of steps is used to remove the line from the date series with blue circle markers. The finished product is shown in Figure 3.17. NOTE: The data should not end on the border of the figure,

therefore, in Figure 3.17 the x-axis was set to range from 0-12. To allow more white space on the axis, the "Units" were selected to be 2.0 versus 1.0. Figure 3.17 was pasted directly from Excel and looks good! But wait… there's more…
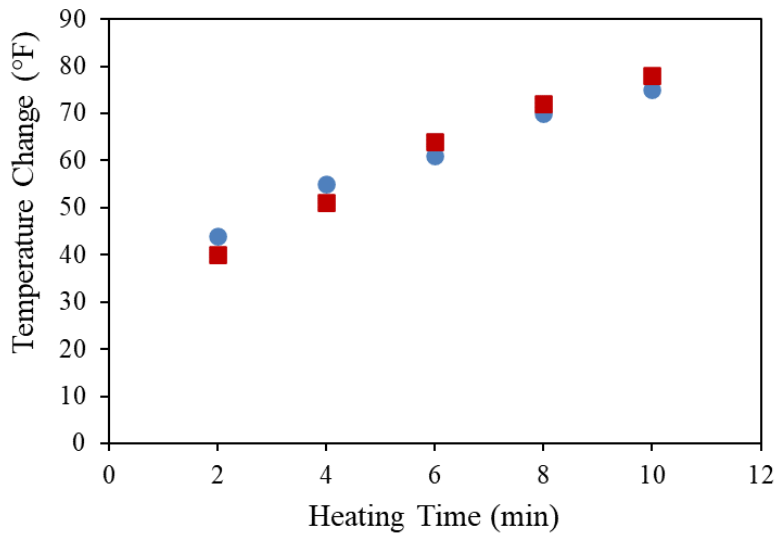
Figure 3.17: Modified chart pasted as is from Excel.

Resizing the chart when pasted directly from Excel produces a very *unsatisfactory* (Figure 3.18) result.
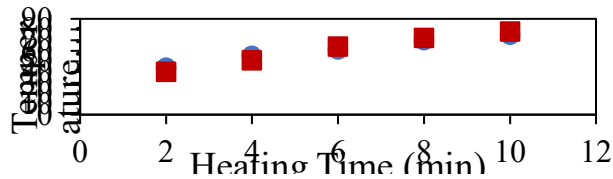
Figure 3.18: Example of undesirable behavior when a pasted Excel chart is resized to a much smaller size in Word.

There is a work-around. *Copy* the figure in *Excel* and *paste* it as a *picture* (or select "Paste Special…" and then a picture option like JPEG) and try resizing it again (see Figure 3.19):
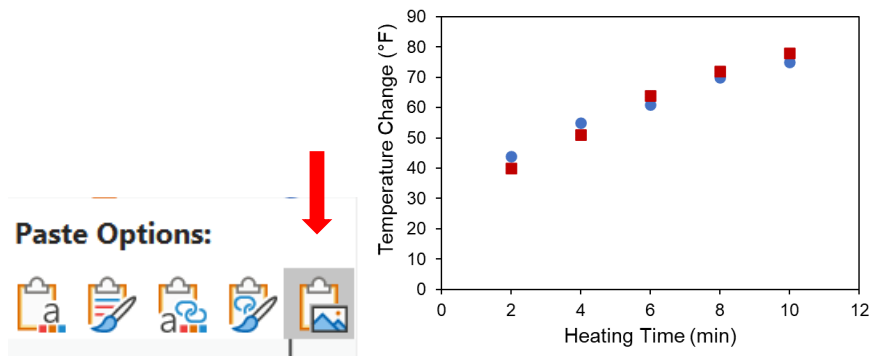
Figure 3.19: Paste the chart as a "Picture" and resizing will produce a better result. The "paste as picture" icon is shown below the red arrow.

88

Templates

Once you get a figure formatting set, create a template by right clicking on the figure and selecting "Save as Template…" (Figure 3.20). Name the template and click "OK."
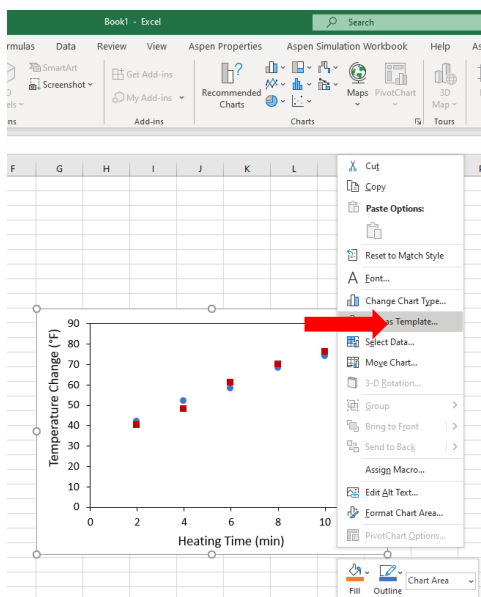

Figure 3.20: *Right clicking on figure pulls up options to create a template.*

Then, when you have another set of data, highlight the data, and click on the arrow found in the lower corner of the "Charts" section of the ribbon as seen in Figure 3.21. (There are multiple ways to apply the template, this is just one of them.)
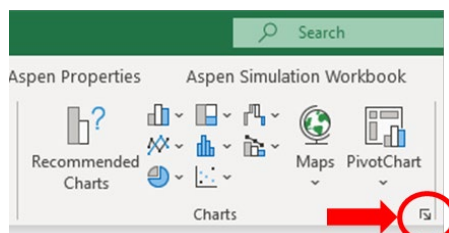
Figure 3.21: *Click on encircled area to apply template.*

Click on "All Charts" at the top of the "Insert Chart" box, and then "Templates" on the left-hand column. Your template should appear on the right as mine does entitled "Heating" (Figure 3.22). Note that I have several other templates as well; one entitled "A" and another entitled "Bar." The figures shown are previews of what your final figure will look like. Some modification may need to be done (especially to the axes). That is, be sure that after you apply the template, the figure is what you want and that the scales are okay.
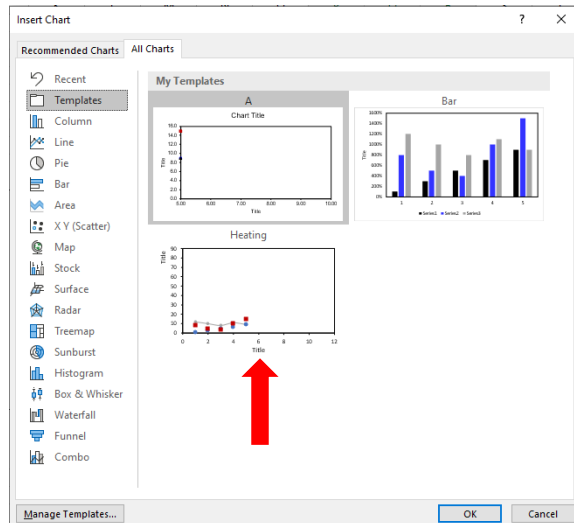
*Figure 3.22: Select which template you want to apply.*

After making some minor formatting changes, Figure 3.23 was created in fewer steps than the original figure. The templates can save a lot of time if you are creating multiple figures of the same type.
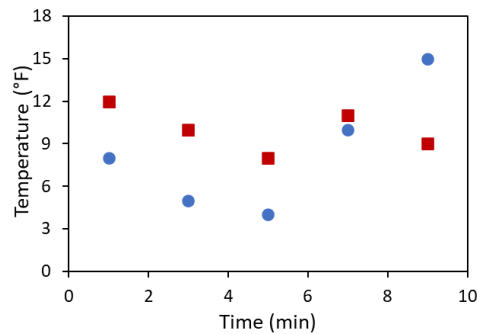
Figure 3.23: A figure that had previous template applied to it.

Captions

Captions are an important piece of the figure. You want to explain what the figure shows, but not discuss results. For example, for the Figure:
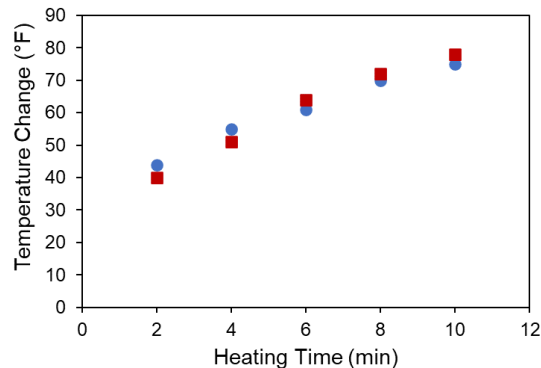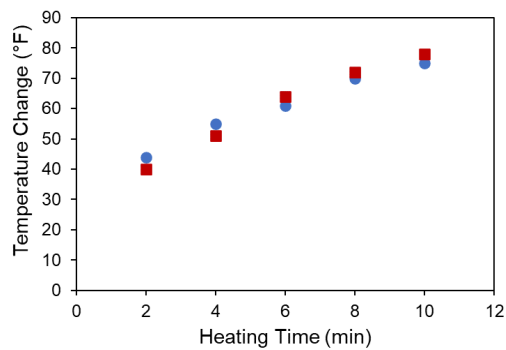

a caption of: "**Figure V3-1:** Change in water temperature after heating for a given time in a 1000 W microwave where ■ represents 100% power and ● represents 50% power" is great!

You *do not* want to have "**Figure V3-1:** Change in water temperature after heating for a given time in a 1000 W microwave that shows the error in heating for a given time is low and that increasing heating time increases the temperature change. ■ represents 100% power and ¨ represents 50% power" in the caption as the discussion should go into the narrative and refer to the figure.

When referring to figures (and tables), always capitalize "Figure" and "Table." For example, *"The increasing trend is shown in Figure 2."* and *"There was a significant error in the measurements as listed in Table 1."* Still capitalize "Figure" and "Table" if referencing them as in: *"The increasing trend was exponential above a time of 1 h (Figure 2)."*

Inserting Symbols in Captions

Figures containing *multiple data series* should **not have legends**. The *marker symbols* for the different *data series* should be defined in the caption of the figure (see Figure 3.24). Inserting the symbols can be done in *PowerPoint* (using **shapes**) or *Word* as described below.

Figure 3.24: Change in water temperature after heating for a given time in a 1000 W microwave oven where ■ represents 100% power and ● represents 50% power.

The above example *caption* contains the *marker symbols* for each *data set* in the figure. **No legend appears in or around the figure**. *Labels* can also be placed directly in the figure as shown in Figure 3.25 (but the *caption* should still include the symbols as shown in the caption for Figure 3.24).
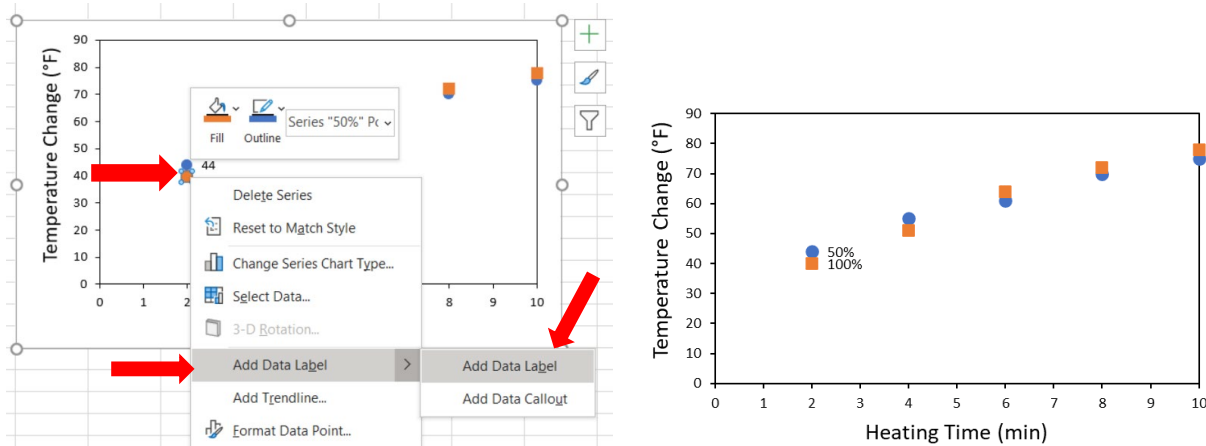
Figure 3.25: Adding a label to a data point where ■ represents 100% power and ● represents 50% power.

Step 1: Find the symbol feature

To insert a *symbol* into your *Word* document, position the cursor at the desired location, then use the following ribbon options:

Insert tab / Symbols group / Symbol

Then click "***More symbols***" if the symbol you need is not in the dropdown box (Figure 3.26). The dropdown box will populate with symbols you have previously inserted making future figure captions easier to format.

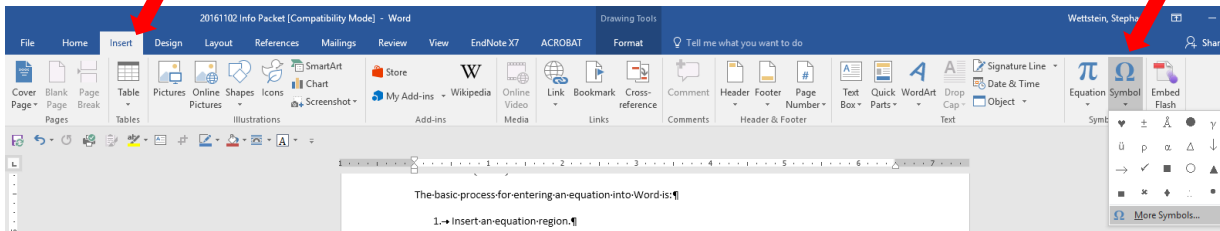After clicking "more symbols" the pop-up in Figure 3.27 will appear.

Figure 3.26: The location on the ribbon for inserting symbols.

Step 2: Find the symbol

The default that will pop-up is likely to be "***Symbol***" font, which does not contain the most commonly used symbols. **NOTE:** For example, do **NOT** use "**$**" as a symbol just because it was the easiest one to find!
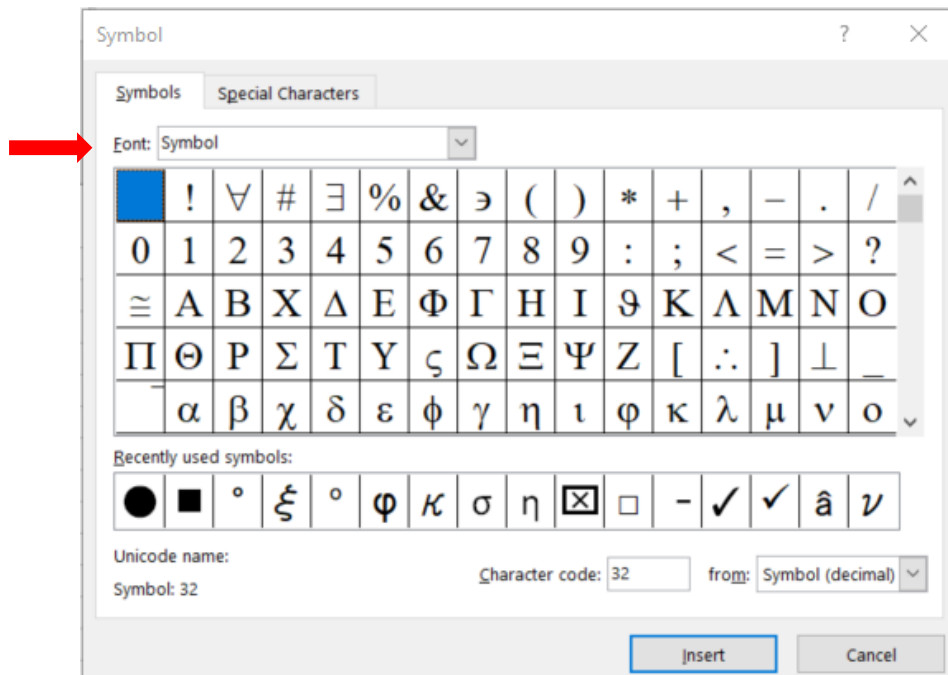

Figure 3.27: The default pop-up menu for "More Symbols."

To access more *symbols*, select "***Arial***" from the *font* dropdown (red box in Figure 3.28) and "***Geometric Shapes***" from the *subset dropdown menu* (you may have to scroll up; see the blue box in Figure 3.28). This *subset* contains the most used *symbols*, such as *square* (■), *triangle* (▲), and *circle* (●).
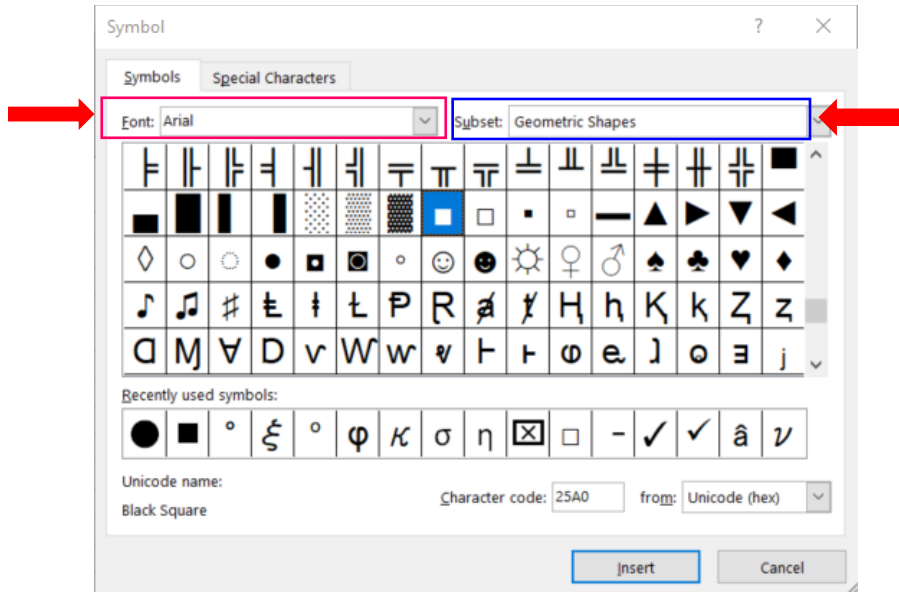
Figure 3.28: *Inserting Geometric Shapes from Symbol dialogue box.*

Step 3: Insert the symbol

Click on the *symbol* you want to *insert* to *highlight* it, and then click "***Insert***" in the lower right of the pop-up box. Even though nothing appears after you click "***Insert***," the symbols are being inserted at the cursor location. Click "***Cancel***" to close the "***Symbol***" dialogue box. Move the cursor and repeat the process to insert another symbol.

TIP: Add all symbols to one location in the document, then cut and paste them to other desired locations.

Step 4: Change the symbol color

After all the *symbols* are *inserted*, *highlight* the *symbol,* and change the color by changing ***the font color*** (**not** the *shape fill* color; see Figure 3.29).
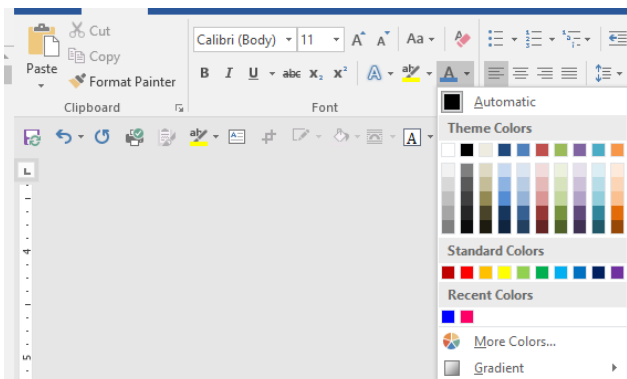

Figure 3.29: *Changing the color of an inserted symbol.*