Data Analysis and Graphing

Objective(s):

1. Demonstrate the ability to determine independent and dependent variables in an experiment.
2. Understand the concepts of graphing and when to use a specific type of graph: bar/column graphs, line graphs, pie charts, and scatterplots.
3. Demonstrate the ability to graph experimentally generated data.
4. Know the concepts of experimental design.

Background:

You have probably heard that a picture is worth a thousand words. Likewise, a graph could be said to be worth a thousand data points. A **graph** or a chart is simply a pictorial form of the data. Graphs are used to display the relationship between variables or show the amount of scatter of data about a given variable or mathematical equation. Graphs are often more informative and more readily understood by the reader than lists or tables of data because graphs communicate the information visually.

The simplest form of a graph is a 1-dimensional number line (see Figure 1) where a single variable can be shown along a horizontal straight line. A number line is frequently used as a teaching aid for addition and subtraction problems involving negative numbers.

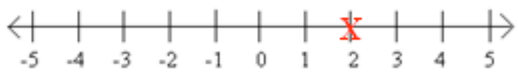


Figure 1: A **"Number Line"** is an example of a 1-dimensional graph. On this graph, it can be seen that the variable "X" has a value of 2.

Number lines are marked with evenly spaced vertical lines to represent the progression of numbers. Traditionally, zero is located at the center of the line with positive numbers being listed to the right of zero and negative numbers to the left of zero. The number line in Figure 1 only shows the integers from -5 to 5; however, the line could be extended to any length desired. Also, the divisions do not have to be integers, they can be any set value. For example, the same size line could be marked with intervals of 0.5 with the resulting labels: -2.5, -2.0, -1.5, -1.0, -0.5, 0.0, 0.5, 1.0, 1.5, 2.0, 2.5.

In Figure 1, the arrowheads on either end of the line indicate that the line continues indefinitely in the given direction (right = positive or left = negative). If the arrowhead is replaced by a closed circle over the final value (in our example 5 or -5), then the limit of the values for the variable is equal to the last value given. (See Figure 2a.) If the arrowhead on the right is replaced by an open circle (at 5) the limit is less than the last value given (<5); likewise, if the arrowhead on the left is replaced by an open circle the limit is greater than the last given value (> -5). (See Figure 2b.)

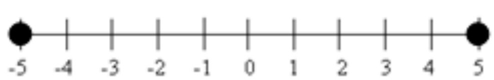


Figure 2a. A number line showing $-5 \le x \le 5$.

Figure 2b. A number line showing $-5 < x < 5$.

While number lines are useful for understanding mathematical concepts (e.g., counting numbers, addition and subtraction problems), they are generally inadequate to aid in the understanding of scientific concepts where one is looking for the relationship between two variables. When a graph is a visual representation of a relationship between two variables (x, y), it takes the form of a 2-dimensional figure where the axes for each variable intersect. The two axes called the x-axis (horizontal) and y-axis (vertical) correspond to the values of the variables. The place where the two axes intersect is called the **origin**. The origin is also identified as the point (0,0). [**Note: There are also 3-dimensional graphs for 3 variables (x, y, z). These are often used to discover the relationship of 2 variables over time (the third variable)*.]

Sometimes, graphs are used to portray a **trend**. An example of a trend would be how well the data fits a straight line. With many chemistry experiments, the results require the use of a graph for proper identification of trends. Constructing a decent graph then can make data analysis more efficient and easier to perform. The principles that go into constructing a decent graph are given in the following sections.

In order to construct a graph, one must first identify the variables and determine their relationship. A **variable** is simply anything that can be varied. Examples of variables would be an object, an event, an idea, a time period or any type of category that can be measured. There are two types of variables: independent and dependent.

The **independent variable "x"** is easy to identify in an experiment because it is the variable that is being deliberately changed. The values of the independent variables are plotted along the x-axis (*horizontal axis*). Examples of independent variables include time passing, geographic locations, or individual items.

A second variable, the **dependent variable "y",** will be measured to see how, or even if, it varies in response to changes in the independent variable. It is thus "*dependent"* on the changes made to the independent variable. The values of the dependent variables are plotted along the y-axis (*vertical axis*). Examples of dependent variables (*y values based on x values*) would be the temperature of water with ice melting in it (based on time passing), the amount of rainfall (based on geographic location), or the number of individuals who prefer a particular type of ice cream (based on flavor of ice cream).

When doing experiments, some people have difficulty remembering which is the independent and which is the dependent variable. A simple way to figure it out is to insert the variables into a "cause and effect" statement.

A General Cause & Effect Statement:

When blank was done to the independent variable, it changed the dependent variable. It is not possible / probable that the dependent variable could cause the change in the independent variable.

A Specific Cause & Effect Statement:

When heat was applied to the ice cubes over a 5 minute period, the temperature of the water decreased 18 ^{o}C. It is not possible / probable that decreasing the temperature of the water by 18 ^{o}C would cause 5 minutes to pass. So time (min) is the independent variable, x, and temperature is the dependent variable, y.

Types of Graphs:

Bar and Column Graphs – Bar graphs and column graphs are essentially the same type of chart, except that in the bar graph the data is displayed horizontally as bars (See Fig. 3a.) and in a column graph the data is displayed vertically as columns (See Fig. 3b.). (*Note: We will be using column graphs or "vertical bars", rather than the traditional horizontal bar graphs.*) In each case though the length of the bar (or column) is proportional to the data that it represents.

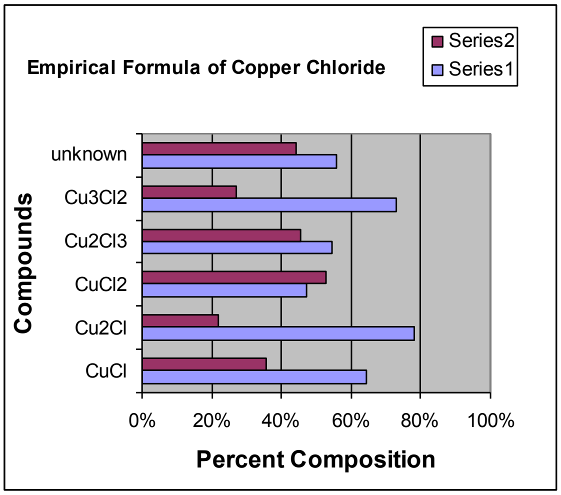


Figure 3a. Bar Graph – data is displayed horizontally as bars.

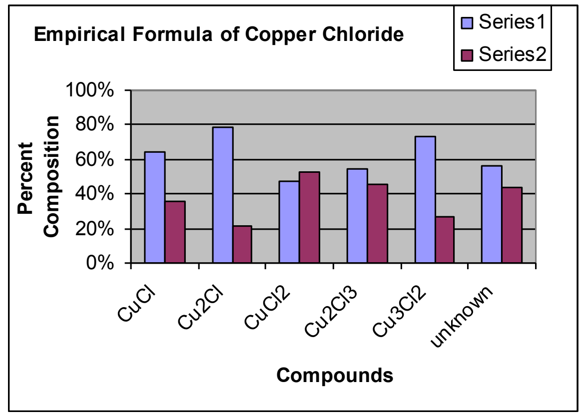


Figure 3b. Column Graph – data is displayed vertically as columns.

Bar and column graphs are both primarily used to show **comparisons of data**. The main advantage of a horizontal bar graph is that it uses the y-axis (vertical line) for labeling; and, there is more room on the y-axis to fit text labels for categorical variables. Despite this, column graphs are more commonly used than bar graphs because they more closely correspond with the reading process in Western cultures; that is, looking at the data from left to right as opposed to from top to bottom. We more readily identify this left to right trend in linear plots. For example, in a linear plot, if the data being displayed is dependent on time, then the overall trend of the data with time progressing from the left to the right is more easily understood by the reader.

Column graphs are often used to compare things between different groups or to track changes over time. Column graphs are best used when comparing static data or a qualitative independent variable. For example if you wanted to compare sales for 5 separate years, then a column graph would be a good choice.

While you can extract trends from looking at the heights of the columns (e.g., *are they getting taller or shorter?*), you should not calculate a slope from those heights. If you want to extrapolate the data to show future events (e.g., *sales predictions for future years*), then you should use a line graph instead.

Constructing a Column Graph

When creating a column graph, draw a vertical column for each group or value. The height of the column is proportional to the data which it represents. The width of each column should be the same for a given dataset. The exact width usually depends on the number of groups represented and the ease of viewing.

Stacked Column Graphs – The stacked column graph uses segments or blocks to form the columns. Despite the fact that stacked column graphs can convey a lot of information, they are rarely used. This is because the stacked column graph can be very difficult to analyze if too many items are in each stack. It is a convenient way to contrast values, but it is not necessarily the simplest manner to understand.

Multiple Column Graphs – The double (or group) column graph is another effective way of comparing sets of data about the same places or items. This type of column graph gives two or more pieces of information for each item on the x-axis instead of just one. This allows one to make direct comparisons on the same graph by time or anything else that can be compared. However, if a group column graph has too many sets of data, the graph becomes cluttered and it can be very confusing to read.

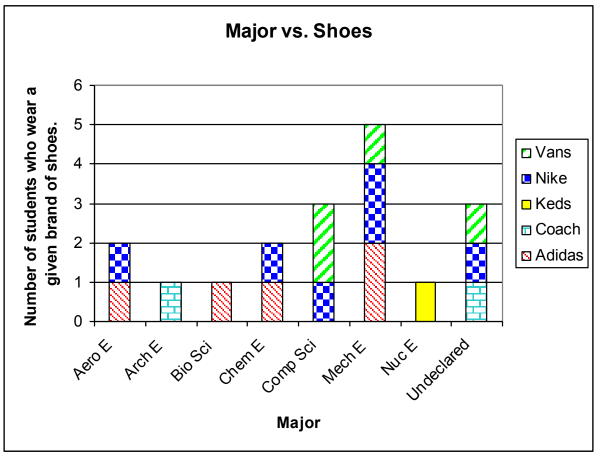


Figure 4a. Stacked Column Graph.

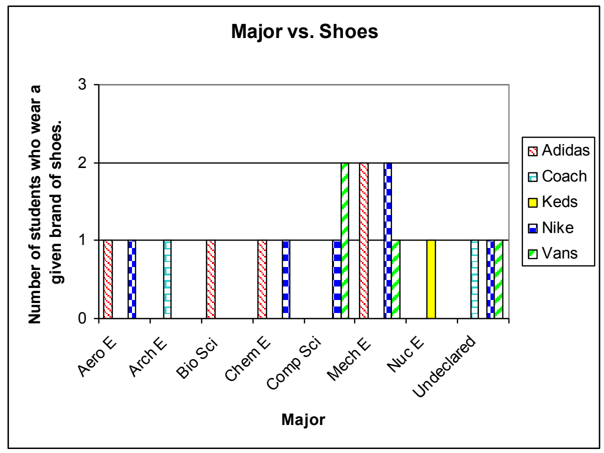


Figure 4b. Group Column Graph.

Line Graphs – A line graph can be thought of as a column graph where only the tops of the columns are represented by discrete points and the rest of the bar is omitted. Often in a line graph, the data points are connected by line segments. It is easier to see the changes over time with the line graph because it is less cluttered than the column graph.

It is important to note that line graphs should be used only when the x - y data are sets of ordered pairs. If one or the other is a qualitative variable, then it is better to use a column graph and display the qualitative variable on the x-axis. The problem with showing qualitative data on a line graph is that it misleads the reader into thinking that there is some natural chronology to the data when there clearly is not.

Line graphs provide an excellent way to determine trends between independent and dependent variables when both are quantitative values. When both variables are quantitative, then the line segment between the two data points is representative of the **slope, m:**

$$\mathbf{m} = \frac{\mathbf{y_2 - y_1}}{\mathbf{x_2 - x_1}}$$

where the slope is the change in the y-values over the change in the x values or *rise over run*.

You are probably familiar with the equation for a line:

$$\mathbf{y = mx + b}$$

where **x** is the independent variable, **y** is the dependent variable, **m** is the slope of the line and **b** is the y-intercept. (The y-intercept is where the data crosses the y-axis at x = 0).

Line graphs are generally used to determine changes over a time period. When smaller changes occur, it is better to use a line graph than a column graph. Line graphs can also be used to compare changes over the same time period for multiple datasets. While column graphs can also be used to display multiple datasets, line graphs are generally less cluttered and thus less complicated for the reader to understand.

Pie Charts – A pie chart is simply a circle that has been divided into a series of segments where each segment represents a different piece of information. The area of the segment is proportional to the data it represents. A pie chart then is a way of displaying the different values of a given variable (e.g., percentage distribution). So each segment of a pie chart represents a part of a whole.

Pie charts are best used when trying to compare percentages. (See Figure 5.) By themselves, they do not show changes over times. Percentages can also be shown using column graphs, but their relative amounts are often easier to visualize using a pie chart.

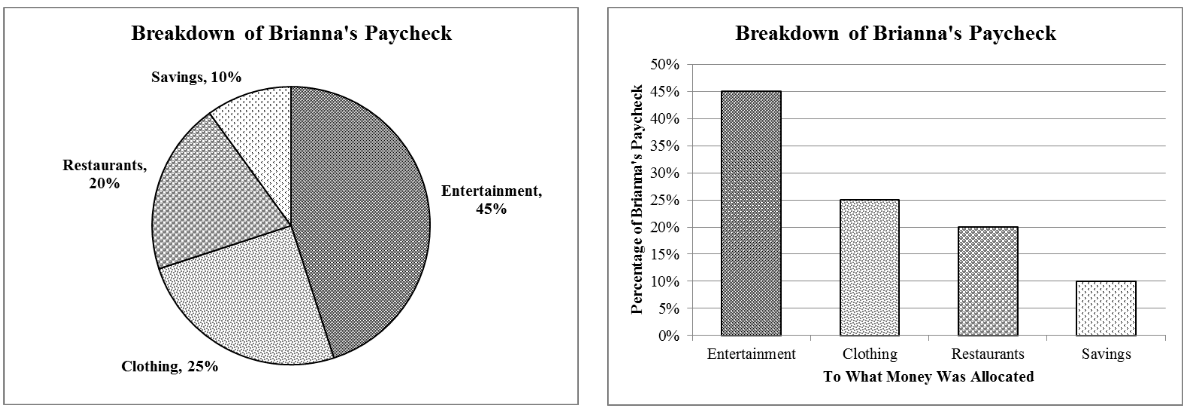


Figure 5: Comparison of Pie Chart (*left*) and Column Graph (*right*).

Constructing a Pie Chart

(***Note:*** *The easiest way to construct a pie chart is to follow this step-by-step approach*.)

1. Draw a circle with your protractor.
2. Convert each component into a percentage of 360 degrees.
3. Starting from the 12 o'clock position on the circle, measure an angle corresponding to the largest percentage with your protractor. Mark this radius off with your ruler. (***Note****: When drawing a pie chart, it is customary to order the segments by size (largest to smallest) and to insert them into the chart in a clockwise direction*.)
4. Repeat the process for each of the remaining categories, drawing in the radius according to its percentage of 360 degrees. The final category need not be measured as its radius is already in position.
5. Labeling the segments with percentage values often makes it easier to tell quickly which segment is larger. Whenever possible, the percentage and the category label should be indicated beside their corresponding segments. This way, the readers do not have to constantly look back at the legend in order to identify how each category has been represented.

Scatterplots (a.k.a., X-Y Plots) – Scatterplots are similar to line graphs in that they are used to determine relationships between quantitative data points. The difference is that the data in a line graph is directly connected with a line, whereas in a scatterplot the trend in the data is represented by a regression line. By analyzing the regression line then, you can determine if there is any correlation* between the x and y values.

Depending on how tightly the data points cluster together, the trend may be obvious prior to the addition of the regression line. Since the data points represent actual experimental data rather than theoretically calculated values, they will represent all of the error inherent in the experimental collection process. When there is a large amount of error, the data points will be widely scattered about the regression line. When there is a small amount, the data points will be close to the line.

(***Correlations** – If both variables increase at the same time, they have a ***positive correlation***. If one variable decreases while the other one increases, then they have a ***negative correlation***. Sometimes the variables do not show any relationship between them. They are said to have ***no correlation***.)

Proper Labeling of the Graph

Titles – The graph must have an overall title that appears at the top of the graph and describes the data presented. For titles that indicate a comparison, the quantity plotted on the x-axis is always listed first and then compared to the y-axis (e.g., Time vs. Temperature for Melting Ice Cubes). Sometimes the title should include the origin of the data (i.e., the experimenter's name and date of experiment).

Axes Labels – The labels that appear along the x-axis and the y-axis need to include the name of the variable which is being measured along with its corresponding units (e.g., x-axis label "Time (min)" and y-axis label "Temperature (^oC)").

Keys – Keys are included if there are multiple datasets being shown on a single graph. Multiple datasets can be shown on one graph, when the datasets have comparable scales. Each dataset would have its own symbol. (An example of a key then would be ■ = Trial 1, ◆ = Trial 2, ★ = Trial 3.)

Specifically then, the key shows the symbol for a given dataset and identifies the dataset by name. If only one dataset is shown, then a key is not necessary because the information is provided in the axes labels.

Scales – The scales are essentially a range slightly greater than the range of the data. The scale is recorded and evenly distributed along an axis. The scale of the axes should be adjusted so that the data fills the graph as much as possible. If appropriate the origin (0, 0) needs to be included.

Interpreting Graphical Data

Trendlines (Regression Lines) – Trendline is another name for regression line. A regression line then is a statistical tool used to mathematically express a *trend* in the distribution of the data points. The equation for the regression line can be used to extrapolate or interpolate information not expressly represented by the data.

Regression lines do *not* have to be straight lines. They can be exponential, linear, logarithmic, polynomial or power based. In any case, an equation can be generated for the regression line that shows the relationship of the x and y values.

Regression Coefficient (R^2) – A regression coefficient indicates the percentage of data that falls along the trendline. If $R^2 = 0.956$, then 95.6% of the data fits the trend.

Slope (m) – The slope is the ratio of the change in the dependent variables to the change in the independent variables. Or, $m = (y_2-y_1) / (x_2-x_1)$

Calculating Unknowns from Graphs – Using the regression line equation, you can insert values of x and calculated values of y or vice versa.

Extrapolation – Predicting unknown values from a regression line, where the unknown value falls outside the original dataset.

Interpolation – Predicting unknown values from a regression line, where the unknown value falls inside the original dataset.

Appendix

Variable – A variable is simply anything that can be varied. Examples of variables would be an object, an event, an idea, a time period or any type of category that can be measured. There are two types of variables: independent and dependent.

Independent Variable "x" – The independent variable is easy to identify in an experiment because it is the variable that is being deliberately changed. The values of the independent variables are plotted along the x-axis (*horizontal axis*). Examples of independent variables include time passing, geographic locations, or individual items.

Dependent Variable "y" – A second variable will be measured to see how, or even if, it varies in response to changes in the independent variable. It is thus "*dependent"* on the changes made to the independent variable. The values of the dependent variables are plotted along the y-axis (*vertical axis*).

Examples of dependent variables (*y values based on x values*) would be temperature of water with ice melting in it (based on time passing), amount of rainfall (based on geographic location), or number of individuals who prefer a particular type of ice cream (based on flavor of ice cream).

When doing experiments, some people have difficulty remembering which is the independent and which is the dependent variable. A simple way to figure it out is to insert the variables into a "cause and effect" statement.

A General Cause & Effect Statement:

When blank was done to the independent variable, it changed the dependent variable. It is not possible/probable that the dependent variable could cause the change in the independent variable.

A Specific Cause & Effect Statement:

When heat was applied to the ice cubes over a 5 minute period, the temperature of the water decreased 18 ^{o}C. It is not possible/probable that decreasing the temperature of the water by 18 ^{o}C would cause 5 minutes to pass. So time (min) is the independent variable, x, and temperature is the dependent variable, y.

Data Sets

x values – the values that correspond to the independent variables. They are measured along the x-axis (*horizontal axis*).

y values – the values that correspond to the dependent variables. They are measured along the y-axis (*vertical axis*).

Data points – This is the ordered pairs of (*x, y*) values that correspond to a given point.

Datasets – The series of ordered pairs ($x_1, y_1 : x_2, y_2: \ldots x_n, y_n$) that correspond to the experimental data.

Other Useful Terms:

Mean (Average) – The mean (x_{bar}) is the average value of a distribution. It can be calculated by adding up all of the observations and then dividing by the number of observations: $x_{bar} = (x_1 + x_2 + \ldots + x_n) / n$. (***Note:*** The mean and the median will agree for symmetric distributions.)

Median – The median is the midpoint of the distribution. That is the point in the list where half of observations fall above it and half fall below it. To determine the median it is always necessary to first reorder the observations based on their size, from smallest to largest. If the number of observations (*n*) is *odd*, the median is found by counting (*n*+1)/2 observations up from the bottom of the list. If the number of observations (*n*) is even, the median is found in the same way except that after finding the center of the list, the two numbers on either side of the center are averaged.

Mode – If a number appears more than once in a distribution, it is considered a mode.

Qualitative Observation – Making a determination based on a characteristic or attribute of an item (e.g., color, moral characteristic, or level of excellence).

Quantitative Observation – Making a determination based on a measured attribute of an item (e.g., length, weight, or volume).

Scales – The scales are essentially a range slightly greater than the range of the data. The scale is recorded and evenly distributed along an axis. The scale of the axes should be adjusted so that the data fills the graph as much as possible. If appropriate the origin (0, 0) needs to be included.

http://www.brighthub.com/computing/windows-platform/articles/17857.aspx
Microsoft Excel 2007: Bar Charts vs. Column Charts
Article by Michele McDonough (86,918 pts)
Edited & published by Neil Henry (18,753 pts) on May 26, 2009

http://cnx.org/content/m10933/latest/
Module by: David Lane

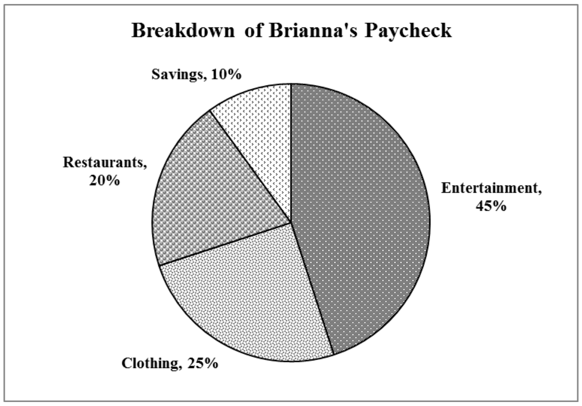
http://www.statcan.gc.ca/edu/power-pouvoir/ch9/pie-secteurs/5214826-eng.htm
Statistics Canada

Name: ________________________________

Section: __________ **Date:** ______________

Graphing Worksheets

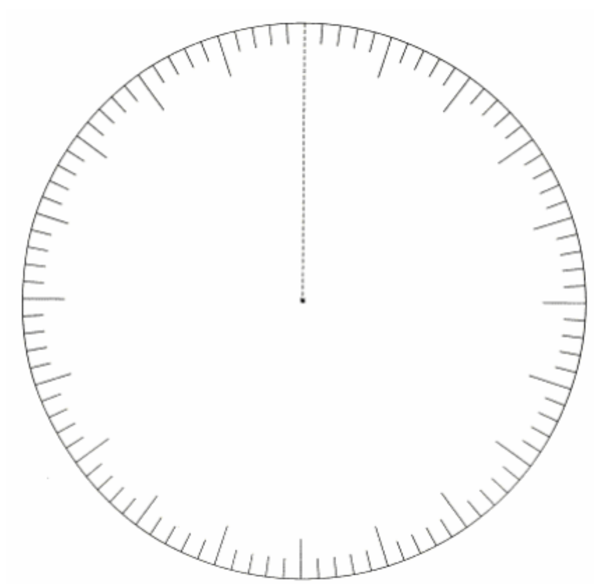
<u>Pie Chart Questions</u>:



1. Brianna, a junior in high school, is working at the mall. She receives a paycheck of $525 per month from her job. According to the information in the pie chart, calculate how much money does Brianna spend on clothing in one month. (Round to the nearest penny or $0.01.)

2. The table below lists 4 different pizza toppings and the percentage of the number of students in Mrs. Brown's biology class who prefer those toppings. Following the directions on page 5 of this handout, create a pie chart that represents the information from the table. Be sure to give your graph a title, and label each section with the correct pizza topping & percentage. You are welcome to use a fill pattern or color to represent the toppings as well.

Topping	%
Cheese	25
Mushroom	5
Pepperoni	40
Sausage	30



Dataset Questions:

3. Ms. Davis recorded her students' test scores alphabetically. In order to look for trends in the data, it would be useful to have the data listed in order of test scores. Rearrange the test scores from lowest to highest and post them in the table provided. Then answer the questions below regarding the data.

History Test Scores

Student's Name	Test Scores
Albert	48
Beth	50
Karl	35
Linda	46
Mark	30
Pamela	46
Ryan	18
Stephanie	39
William	26

Test Scores	Student's Name

a. What is the range for this data set? (lowest value to highest value)

b. Who received the highest score?

c. Who received the lowest score?

d. Are there any modes for this data set? If so, where?

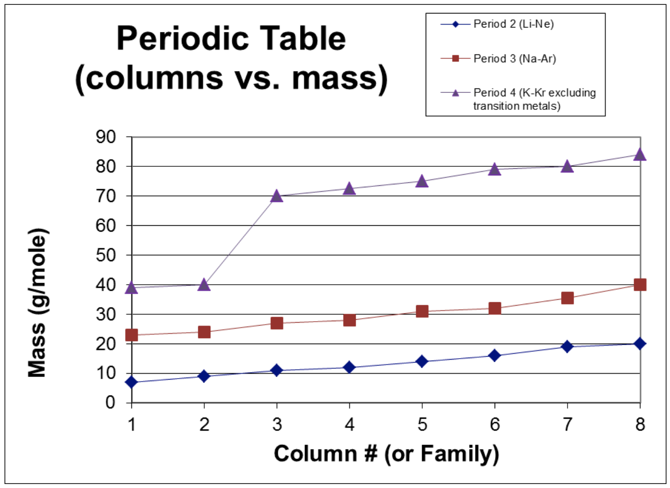
e. What is the median for this data set?

f. What is the mean (*average*) for this data set? Show your calculation.

Name: ______________________________

Section: __________ **Date:** ______________

4. Scatterplots/Line Graphs



A plot of the masses of the atoms Lithium to Krypton (*excluding the transition metals Sc-Zn*) based on their column number (*i.e., group or family where 1 corresponds to alkali metals through 8 as noble gases*).

Directions: Answer the following questions based on the **Graph of the Periodic Table** shown above where the masses of the atoms Lithium to Krypton (excluding the transition metals Sc-Zn) were plotted versus their column number (i.e., group / family):

a. Why were the families plotted along the x axis?

b. Would it have been okay to plot the masses along the x-axis and the families along the y-axis? Explain using the concept of dependence.

c. When looking at the graph, do you see any trends in the masses of the elements going from left to right across a period? Explain. (***Note:*** *The question is about the graph above not the periodic table itself.*)

d. When looking at the graph, do you see any trends in the masses going from top to bottom down a column (comparing groups / families)? Explain.
(***Note:*** *The question is about the graph above not the periodic table itself.*)

e. List another properties of atoms could be used to demonstrate periodic trends.

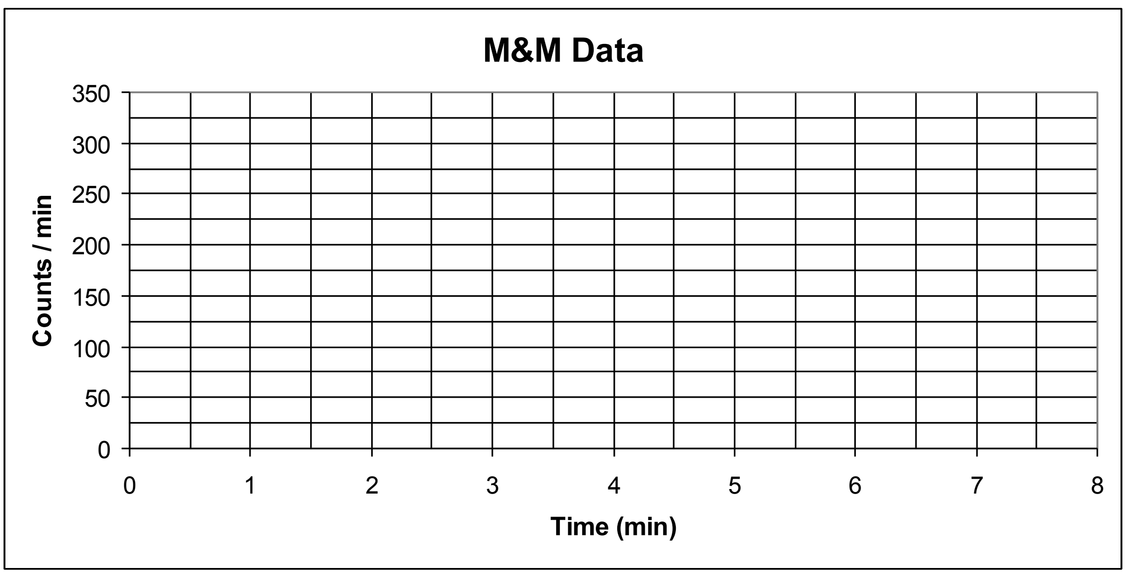
5. Scatterplot – Showing Rate of Decay

Joey's mom bought a bag of M&Ms for a party. She told Joey not to eat all of the M&Ms. Joey figured out that if he ate half of the M&Ms, then there would still be half left and he would not have eaten all of them. After eating half of them, Joey used the same logic and ate half of what was remaining. He continued to use this logic and ate half of the remaining M&Ms at a regular pace.

Unfortunately for Joey, his sister Allison caught on to what he was doing and started keeping track of the number of M&Ms. Allison reported Joey's antics to their mom. When their mom asked Joey what happened to the M&Ms, he claimed their dog Zeno had eaten them.

Time Passed (minutes)	**Number of M&Ms remaining**
2	321
3	215
4	144
5	97
6	65

a. In order to see how fast Joey inhaled those M&Ms, graph Allison's observations. On the graph provided, graph the data and determine the time and rate at which Joey was eating the M&Ms. Once the data is plotted, connect the data points with a smooth line.



b. **Estimate Your Half-life.** On your M&M Data graph draw a horizontal line at 300 M&Ms (crossing the y-axis at 300). Draw another horizontal line at half that value. Where your horizontal lines cross your smooth line for the data, drop a vertical line to intersect the x-axis. Subtract the one time from the other, this will be your half-life. Show your calculation here.

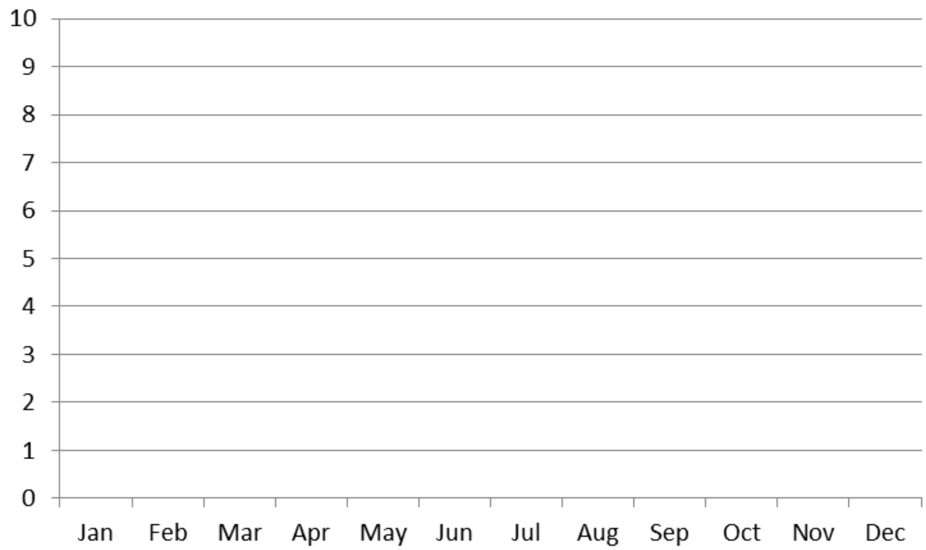
c. **Solve for the Rate (k).** The **rate (k) of a half-life reaction** is given by the equation $t_{½} = 0.693 / k$ where $t_{½}$ is the time it takes for half of the amount to disappear. Using the half-life value you calculated in part b. above, determine the rate. Show your calculation here.

(*Note: Answer must have units to receive full credit.*)

Given : $t_{½} = 0.693 / k$

6.) Graphs of Personal Data *(Use the data from the datasheet that you got in class.)*

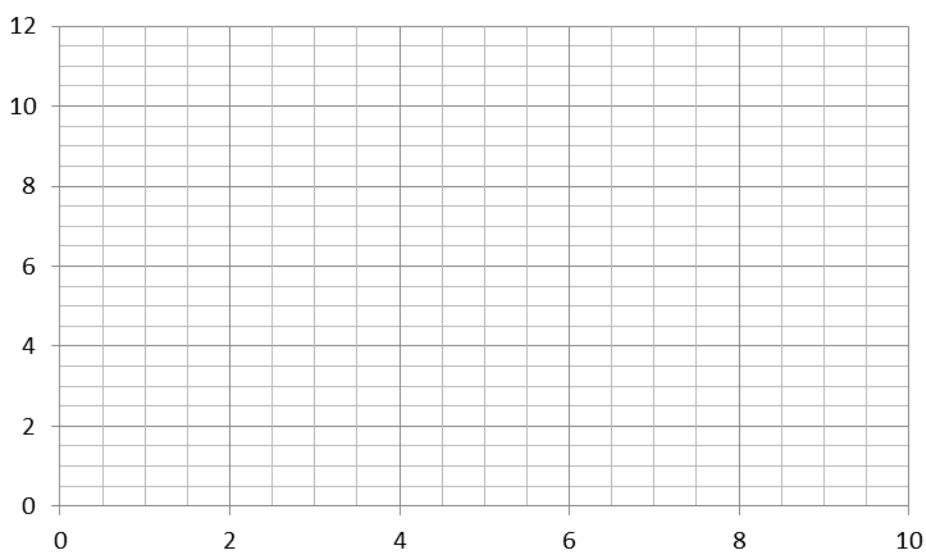
a. On the graph provided, make a stacked column graph (*See **Fig. 4a** page 4 of this handout*) using the birth months of your class as the x values and the number of students with a given major on the y-axis. Make a key in the right margin that indicates what color or fill pattern was used to indicate the majors. (*Don't forget to title the graphs, label the axes and make a key*.)



b. Record any trends you see here. (*If you do not see any trends, record that instead*.)

c. On the graph provided, make an X-Y scatterplot using the number of meals on the x-axis and the number of hours of sleep on the y-axis. For your key, you may use different symbols or different size points to indicate when there is more than one person who gave the exact same answer.

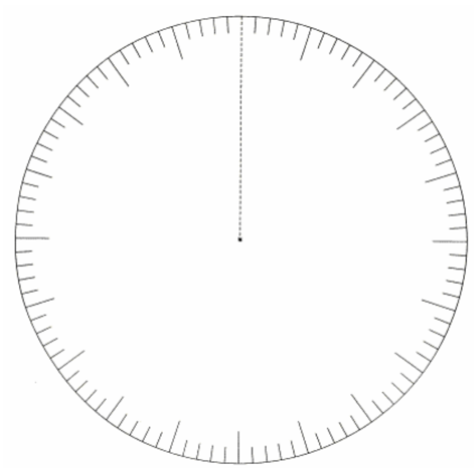
(*Don't forget to title the graphs, label the axes and make a key*.)

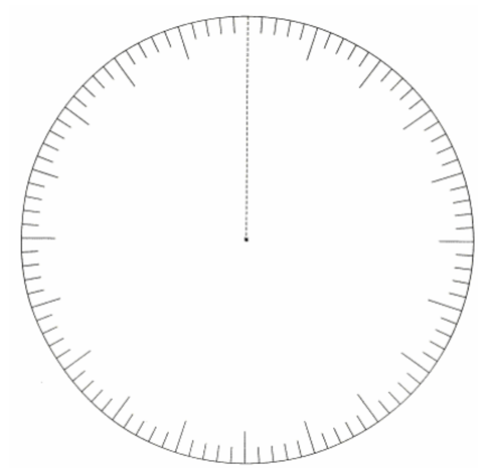


d. Does the number of meals a person eats directly affect the number of hours that s/he sleeps? (*i.e., Is there a correlation between the x and y values? If so, is it positive or negative?)*

7.) Smarties (Using the data from the datasheet that you got in class):

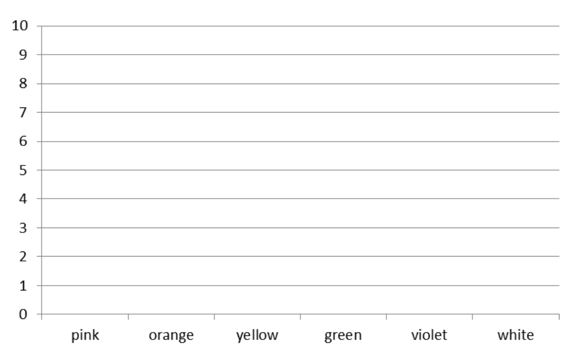
a. Write the number of Smarties that you have of each color in the corresponding column.
b. Get the data for the rest of your Chem 2 section and record it in your table.
c. Sum each of the rows. (They should equal 15.)
Sum each of the columns, including the total column.
d. Record your data and the total your section data again in the second chart.
e. Calculate the percent of each color that you have, using the equation:
% = [number / (total=15)] x 100
For example, if you have 5 red ones you would take (5 / 15) x 100 = 33%
f. Calculate the percent of each color for your Chem 2 section data.
g. Plot your percentages in the pie chart on the left and your section data on the right. Label the percentages on the graphs. If you do not use actual colors to indicate the different colors, please make a key. (*Directions for constructing a pie chart are on page 5 of this handout. Don't forget to title each of the graphs.*)

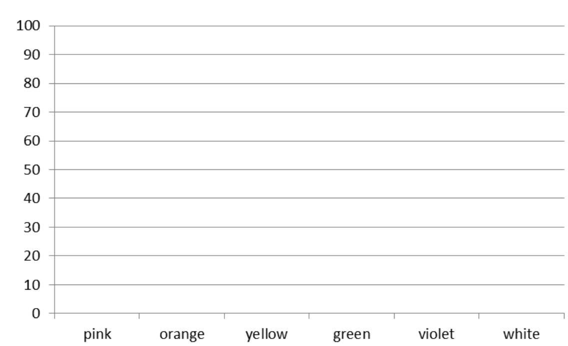




h. Which data do you feel more closely resembles the data for the entire class (where the entire class is <u>all 12 sections</u> of Chem 2)? Explain.

i. In the space below, plot your data and your section data in column graphs with your data on the left and your section data on the right. (*Don't forget to label each of the axes & give overall titles to both of the graphs.*)





j. In your opinion, from which type of graph (the pie chart or the column graph) is it easier to determine the ratio of a given color of Smarties? Explain your answer.

