Week Zero:
Introduction to Types of Data and Graphs

Week Zero Goals

- Learn some statistical vocabulary.
- Distinguish between numerical (quantitative) data and categorical (qualitative) data.
- Distinguish between continuous numerical data and discrete numerical data.
- Identify different types of graphs (bar chart, pie chart, dot plot, histogram). Provide a sentence of interpretation for a graph.
- Identify the objects in the population, sample and type of data collected.

Vocabulary You Should Know

Population

The objects in the population are all the people or things that you want to know about.

NOTE: Often we do not have access to every object in the population. This is why we rely on sampling techniques.

Sample

A sample is a subset of the population to be studied; the people or things selected from the population. It is best to select randomly.

Random Sampling

Random sampling is the process of using chance to select individuals from a population.

Simple random sampling

Simple random sampling (SRS) is when a sample of size n from a population of size N is obtained and every possible sample of size n has an equally likely chance of occurring.

We will assume that data in this course are collected using Simple Random Sampling techniques. Be aware that there are terms to describe other sampling techniques: Stratified, Systematic, Cluster, Convenience. You are not responsible for knowing them for this course.

Variable

A variable is a particular measurement or characteristic for an object being studied. If you conduct a survey and ask for the responded to indicate their age, then AGE is a variable. If you ask for respondents to indicate their gender, then GENDER is the variable. Some variables are numerical and some are categorical.

Categorical (Qualitative) Variable

The classification of individuals based on some attribute or characteristic. (e.g. Do you have a driver's license? Yes/No) BEWARE: Some data look like numbers but are really categorical, e.g. zip code.

Numerical (Quantitative) Variable

A numerical measure of individuals. Mathematical operations can be performed on such values to provide meaningful results. Numerical variables can be further described as either **discrete** (counts) or **continuous** (measured).

Discrete Numerical Variable

A discrete numerical variable has a countable number of possible values. The data values are numbers such as 0, 1, 2, 3…, e.g. number of apples on a tree; you won't get results such as 2½ apples.

Continuous Numerical Variable

A continuous numerical variable has an infinite number of possible values. The data values can be any value in an interval, e.g. weight of individual apples on a tree; results such as 5, 5.1, 5.33, 6 ounces are possible.

Examples of Identifying the Population and the Sample

Example 1: We are interested in studying customer satisfaction with the new food trucks at the Rochester Public Market. We survey every 5 people that leave the food truck area and end up with 65 responses. Describe the population in a sentence. Describe the sample in a sentence.

The population is all food truck customers at the Rochester Public Market. The sample is made up of those 65 food truck customers that responded to the survey. We will only have data on the sample of respondents.

Example 2: Wegmans is interested in the amount of time it takes to fill one of their Meals2GO orders. All Calkins Road Wegmans March 2020 meals2GO order times are gathered and will be studied. Describe the population in a sentence. Describe the sample in a sentence.

The population is All Meals2GO orders for March 2020 at Calkins Road Wegmans. Since they have all the data in their population, there is no sample. Our data represents the population data.

Examples of Identifying the Type of Data

Example 3: Identify each variable as Categorical, Numerical/Discrete or Numerical/Continuous.

A. A person's waist measurement (inches)

Numerical/Continuous since it is measured and a fractional measurement is possible (e.g. 32.2 inches).

B. The different egg sizes (when purchasing a dozen)

Categorical since eggs are measured in Small, Medium, Large, Extra Large.

C. Social Security Number

Categorical since a social security number represents an individual. It would not make sense to average the SS number or rank SS numbers, etc.

D. The total cost of plane fare to fly back home (dollars)

Numerical/Continuous since fractional dollar amount is possible (e.g. $355.58).

E. Classes of the Dewey Decimal System

Categorical since the system categories books by subject matter.

F. The number of attendees at a concert in the Gordon Field House

Numerical/Discrete since the attendees are counted and there will not be fractional responses.

Graphical Displays

Frequency Distribution

When the variable that you are studying is categorical (qualitative), then you can organize and visualize the data in a **Frequency Distribution/Table** – a table with two columns. One column lists the categories, and another lists the frequencies with which the items in the categories occur (how many items fit into each category). Each entry in the table contains the frequency or count of the occurrences of values within a particular group or interval, and in this way, the table summarizes the distribution of values in the sample. The table is a way of organizing your data.

A survey was conducted and children were asked: ***"Which emotion do you connect with the color red?"*** ("Color Association of Male and Female Fourth-Grade School Children," J. Psych., 1988).
The results have been summarized in the table below. This table is called a **FREQUENCY DISTRIBUTION.**

Emotion	Frequency
Anger	61
Happiness	31
Love	77
Pain	45

A frequency distribution organizes raw data. Additional columns of data can be calculated to help better understand these data. For example, the relative frequency could be calculated. To do this, first find the total of the sample size. We call this n and, for this example, n = 61+31+77+45. n = 214.

Emotion	Frequency	Relative Frequency
Anger	61	61/214 = .2850
Happiness	31	31/214 = .1449
Love	77	77/214 = .3598
Pain	45	45/214 = .2103

The benefit of studying relative frequencies is that you are looking at the count relative to the overall sample/population size. These values can be interpreted as percentages and are used to help build pie charts and bar charts as a way of visualizing data.

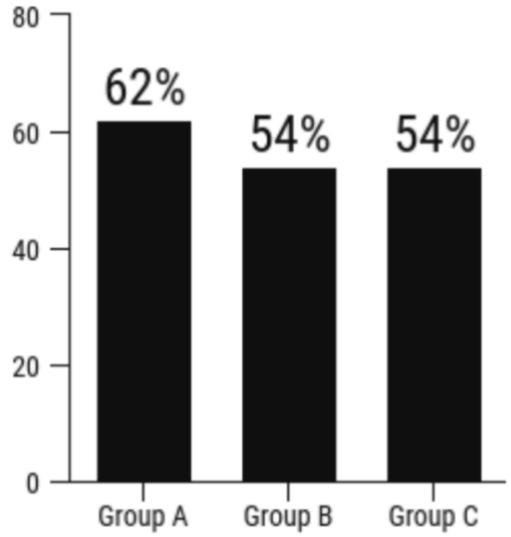
Bar Charts and Pie Charts

Once categorical data is organized into a frequency distribution, graphs like bar charts and pie charts are helpful visualization tools. Technology is helpful in producing graphical displays. A quick Google search reveals many websites that can build the graphs you need. We also have numerous free statistical software packages available at RIT including Minitab 19, Minitab Express and JMP Pro. We will not be building any graphs this week but rather looking over graphs and forming a statement of interpretation.

Pie Chart – a graph that shows proportions and percentages between categories by dividing a circle into proportional segments. Pie Charts are ideal for giving the reader a quick idea of the proportional distribution of the data.

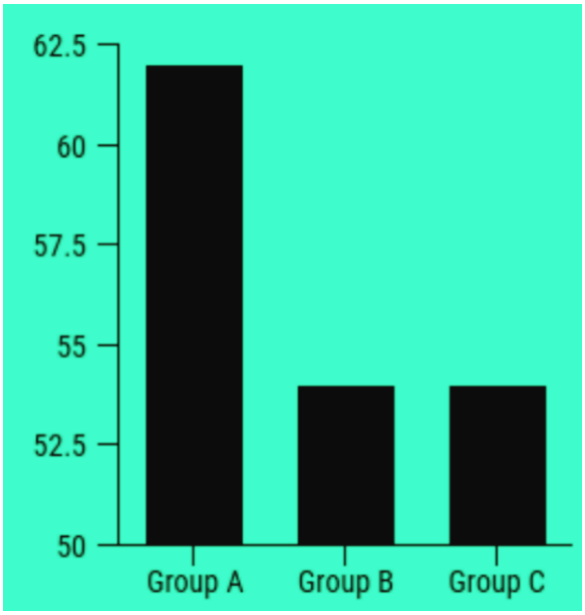
Bar Chart or Bar Graph – a graph that displays parallel bars for each category with the length of each bar indicating the frequency of that category.

The following bar chart intends to show the percentage of Female students that are in Groups A, B and C. We are able to conclude that the highest percentage of females are in Group A. We can see that the percentage of females is the same in groups B and C. We also can see clearly that Group A is only 8 percentage points higher than the other groups.

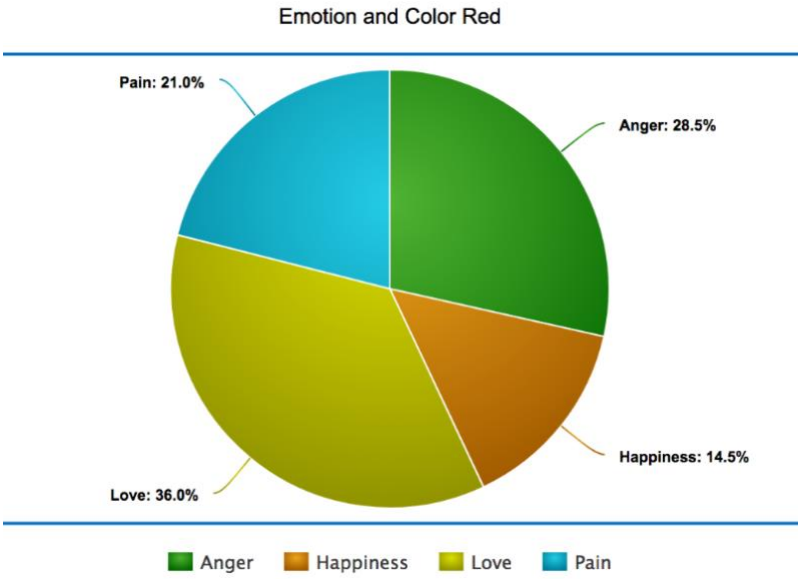


MISLEADING BAR CHART: This bar chart is built poorly since the vertical axis does not go down to zero. As a result, the percentage differences between group A and B/C seems much larger.

To learn more about misleading graphs, visit the website where I found this example.



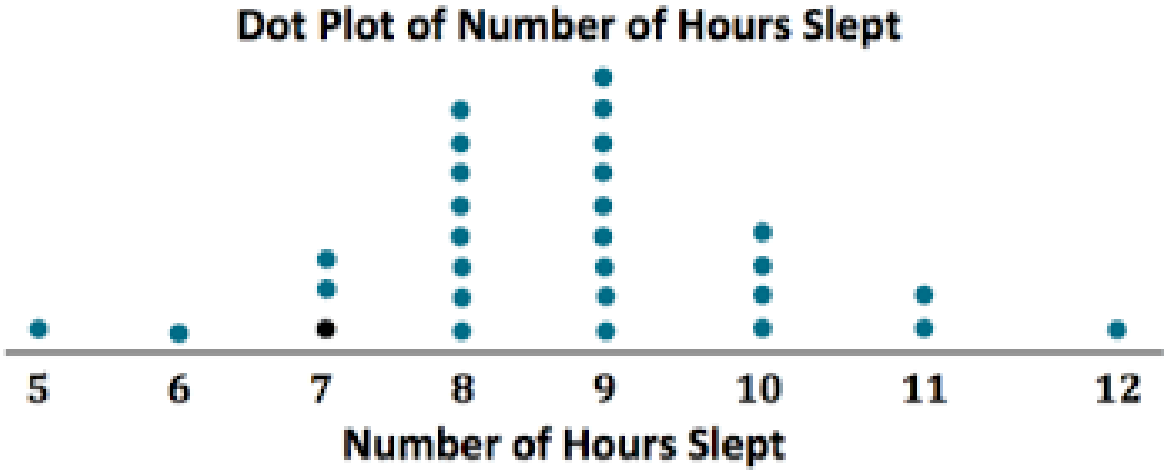
Let's see the pie chart associated with the frequency distribution from the previous slide. I used a free online pie-chart creator linked here and produced the following pie chart.



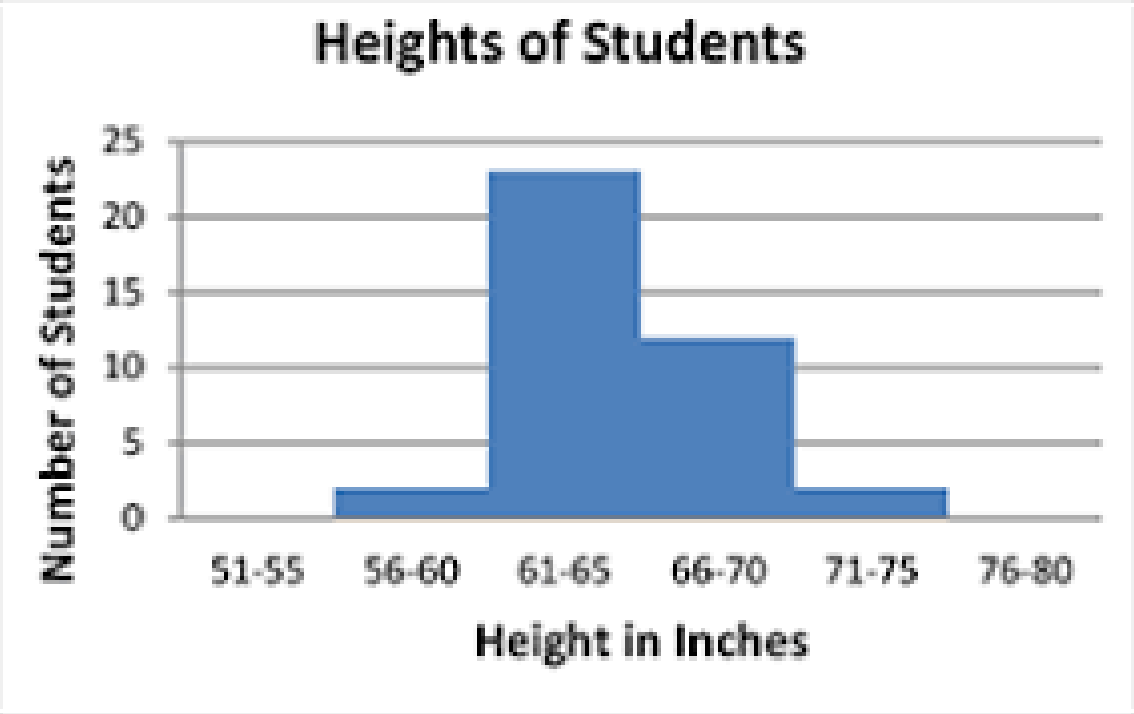
This graphical display is appropriate for the data collected since it represents all the responses from a group of people. We can clearly see that the color red is associated with "love" for the greatest proportion of people who responded.

Dot plots and Histograms

Discrete, numerical data can be visualized nicely in a dot plot. Notice the horizontal axis represents the discrete variable 'number of hours slept' for a sample of students. You can find what might be considered 'typical' for that sample of data by finding the highest frequency of dots. Based on the graph, it appears that nine hours is a typical response for this sample since the highest frequency of responses occurred with 9 hours. In future lessons, we will learn to describe the graph more completely.



Histograms can be built with numerical data---discrete or continuous--and bins are calculated to best group the data. In this histogram of the heights for a sample of students, you can see that the frequency of heights is grouped together into the following bins: 51 inches to 55 inches, 56 inches to 60 inches, etc. No one in the sample had a height in the first bin and that is why the graph does not exist there. The typical height for this sample of students is between 61 and 65 inches, since that bin has the highest frequency of responses. In future lessons, we will learn to describe the graph more completely.



Examples of Identifying the Population, Sample, Type of Data, Type of Display and more…

Example 4: Penn State is reviewing the undergraduate enrollment numbers by campus for Fall 2017 and constructed this table.

Tally

Campus	Count	Percent
University Park	40835	48.5%
Commonwealth Campuses	29388	34.9%
PA College of Technology	5465	6.5%
World Campus	8513	10.1%
Total	84201	100.0%

Penn State Fall 2017 Undergraduate Enrollments

1. Identify the objects in the population and in the sample.
2. Identify the variable being studied.
3. Identify the type of graphical display and describe the display in one sentence.

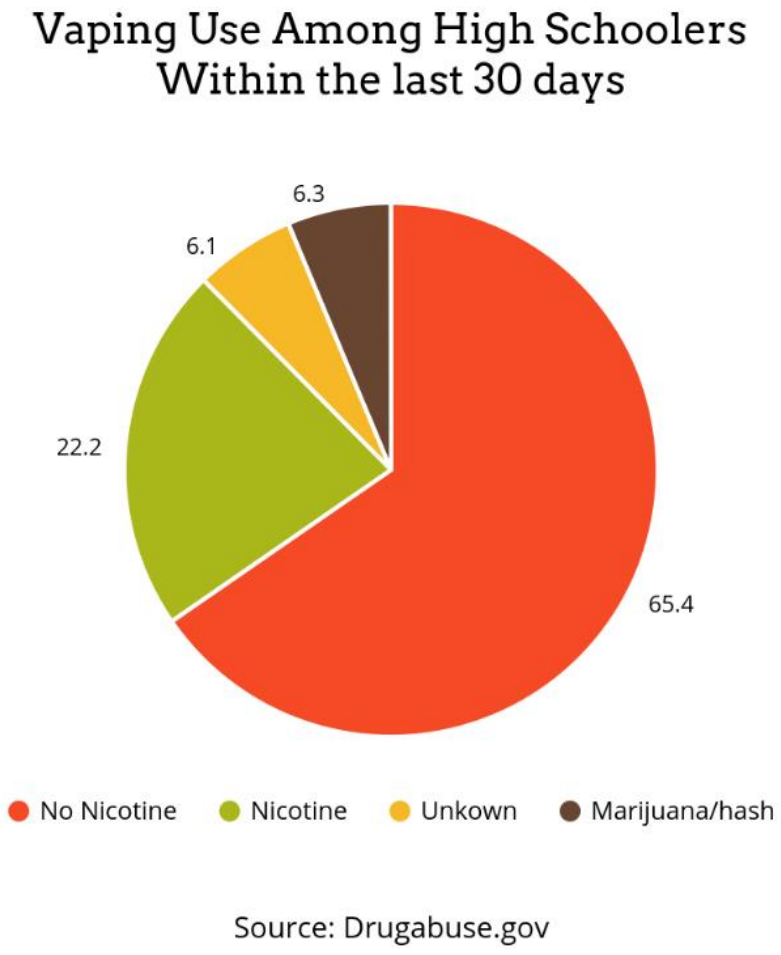
Answer:

1. The population is all undergraduate students at Penn State in Fall 2017. Since we have all the data, no sampling has taken place.

2. The variable is which campus the student is enrolled; it is a categorical variable (By the way, there are 4 levels to the variable...4 different Penn State campuses)

3. The data are being represented in a frequency distribution (otherwise called a frequency table). After reviewing the undergrad enrollments on the different Penn State campuses in 2017, the one sentence description could be: "Of all the undergrads at Penn State, almost 50% of undergraduates are enrolled at University Park". Another reasonable sentence could be: "The smallest percentage of Penn State undergrads are located at PA College of Technology".

Example 5: A random sample of US high schoolers were asked numerous survey questions including the type of e-liquid they are using while vaping over the last 30 days. The following display was produced.



Source: Drugabuse.gov

1. Identify the objects in the population and in the sample.
2. Identify the variable being studied.
3. Identify the type of graphical display and describe the display in one sentence.

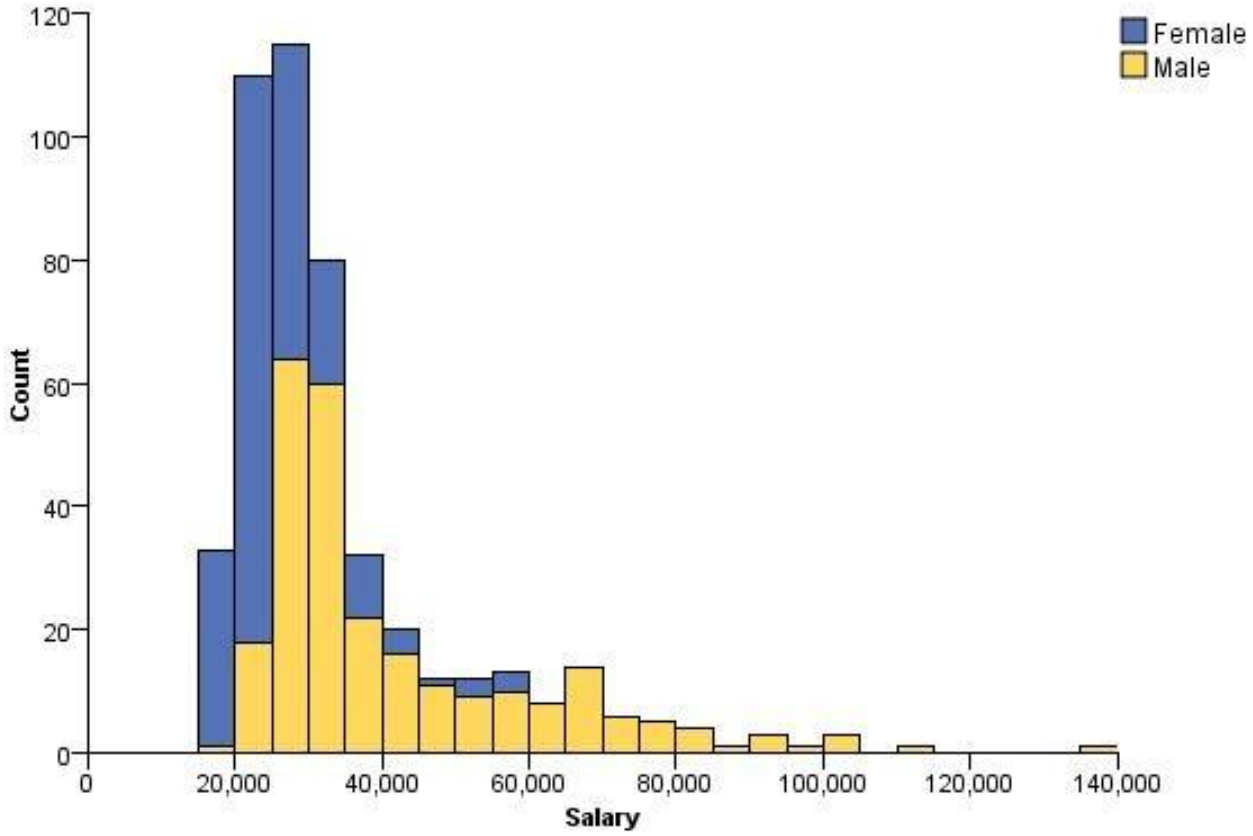
ANSWER:

1. The objects in the population would be all US high schoolers. The sample consists of this US high schoolers that were randomly selected to take the survey.

2. The variable is the type of e-liquid used when vaping. The variable is categorical. (There are 4 levels to the variable...i.e. 4 different types of e-liquid choices as responses)

3. The data are summarized in a pie chart. One sentence might be: Based on the sample taken, a majority of teenagers are vaping an e-liquid that has no nicotine.

Example 6: A recent survey was conducted on a random sample of New York State residents. Many questions were asked and the graphical display below was produced.



1. Identify the objects in the population and in the sample.
2. Identify the variable being studied.
3. Identify the type of graphical display and describe the display in one sentence.

ANSWER:

1. The objects in the population are all New York State residents. The sample consists of randomly selected NYS residents.

2. The main variable being studied is salary (in dollars). Salary is a numerical variable. If only salary were studied, we would have one histogram. Two histograms were built by separating the salary by gender. Gender is a categorical variable. Ultimately two of the variables from the survey are being studied: salary by gender.

3. This is a histogram (it can be called a stacked histogram since the two histograms are stacked on top of one another). One possible sentence: According to the sample of NYS residents gathered, the typical salaries for both men and women are around $30,000.