Chapter 1: Collecting and Analyzing Data

Vocabulary

Data: Information in all forms.

Population: The collection of all people or objects to be studied.

Census: Collecting data from everyone in a population.

Sample: Collecting data from a small subgroup of the population.

Bias: When data does not represent the population.

Introduction: The goal of collecting and analyzing data is to understand the world around us. How data is collected is very important. The goal of collecting data is to get "unbiased" data that represents the population. Analyzing biased data may result in incorrect conclusions and lead to a misguided view of the world around us. It is also important to have a goal in mind when you collect data. Are we trying to find a population percentage from categorical data or a population average from quantitative data? Are we trying to show that two variables are related or are we trying to show cause and effect? Data needs to be collected differently depending on what goal you have in mind.

Section 1A – Two Types of Data – Categorical and Quantitative

One of the most important factors when analyzing data is to determine what type of data you have and how many variables you are analyzing. Let us start with the type of data.

There are two general types of data, categorical and quantitative.

Categorical Data

Categorical data (or qualitative data) are generally words or labels that describe the people or objects in the data set. For example: The country a person was born in, the brand of car they drive, or who they will vote for in the next presidential election.

Usually, categorical data is made up of words. (Do you smoke? yes or no.) Occasionally a number can be used as a category. For example, the numbers "1" and "0" may be used instead of yes or no. A zip code can be used instead of describing the area a person lives in. It is important to remember that a number in place of a written description is still considered categorical.

Country of Birth	Brand of Car	U.S. 2024 Presidential Vote	Smoke? Yes = 1, No = 0	Los Angeles County Zip Code
USA	Ford	Kamala Harris	1	93591
Canada	Toyota	Donald Trump	0	91506
Mexico	Chevrolet	Kamala Harris	0	91390
USA	Honda	Chase Oliver	1	91505
France	Tesla	Donald Trump	0	91741
Spain	Dodge	Cornel West	1	91331
Australia	Cadillac	Jill Stein	0	91355

Nominal Verses Ordinal Categorical Data

Categorical data can also be described as either nominal or ordinal.

Nominal Categorical Data: Most categorical data are nominal. Nominal means the words or descriptions in the categorical data do not have a specific ranking and the order we put them in does not really matter. The data values above for Country of Birth, Brand of Car, and Presidential vote are considered nominal categorical data. They do not have a natural ranking and you can list them in any order. Yes or no questions are also nominal even if they are described as ones and zeros. Would it matter if yes answers came first on a graph or if no answers came first? It does not matter. Does it matter if a graph of brands of cars shows Honda first or Ford last? It does not matter. You still get the same information.

Students often ask about putting categories in alphabetical order. Putting nominal categories in alphabetical order does not stop them from being nominal. They still do not have a natural ranking and we still get the same information even if they are not in alphabetical order.

Ordinal Categorical Data: Sometimes, categorical data is ordinal. This means that the categories have a natural ranking or order. A graph or summary of the data should have one category before another. For example, suppose we asked people how much they like apples. We can rank the answers with "none, low, medium or high". This would be considered ordinal. The graph would look weird if we put "medium" first, "none" second, "high" third, and "low" last. Also, we could represent none as 0, low as 1, medium as 2, and high as 3. Remember replacing a written description with a number does not stop it from being categorical. Notice in ordinal there is no specific difference between categories. We can not quantify how much of a difference there is between liking apples a little bit and liking apples a medium amount. Even if the data is represented by the numbers 0,1,2 and 3, those numbers represent categories.

How much do you like apples?

I do not like apples at all. = 0
I like apples a little bit. = 1
I like apples a medium amount. = 2
I like apples a lot. = 3

Quantitative Data

Quantitative data are numbers that measure or count something. They usually have units and taking an average makes sense. Examples include the heights of people in inches, the weights of cats in kilograms, or how much coffee or tea someone drinks per day in cups. Notice that in each of these cases, the data are numbers that measure or count something, and an average seems appropriate in the context. We can find the average height, the average weight, or the average number of cups of coffee or tea people drink per day.

Units: Students are often confused about the units. What are the units in quantitative data? Units are also called the unit of measurement. The scale by which we measure and compare. In the height data, the units are inches. In the weight data, the units are kilograms. In the coffee/tea data, the units are the number of cups per day.

Height of People (Inches)	Weight of Domestic Cats (kilograms)	Number of cups of coffee or tea drink per day
65.0	5.17	2
61.8	3.76	1
68.0	3.67	3
67.0	4.04	0
57.0	4.17	1
70.8	4.22	2
66.2	4.81	4
71.7	4.13	2
68.7	3.63	8
67.6	4.40	1
