Week 2: Descriptive Statistics and Causality

Jennifer Lin

2021-09-29

Descriptive Statistics

When we have data, we want to know what our data look like and what story it tells. The most simple way to quickly glance at our data is using **descriptive statistics** such as the mean, median and mode concepts we came to love since elementary school. These statistics are simple models of our data. But before calculating any means, medians or modes, it is important to consider the way we constructed the measurements so we know how to best build our model, however simple.

Suppose we want to model ice cream tastes of students at Northwestern. We can do this in many ways. Here are three very simple ways. To illustrate the examples that follow, I constructed a simple data set that we will use to follow along. This data set includes a hypothetical sample of 100 Northwestern Wildcats.

Nominal Scales

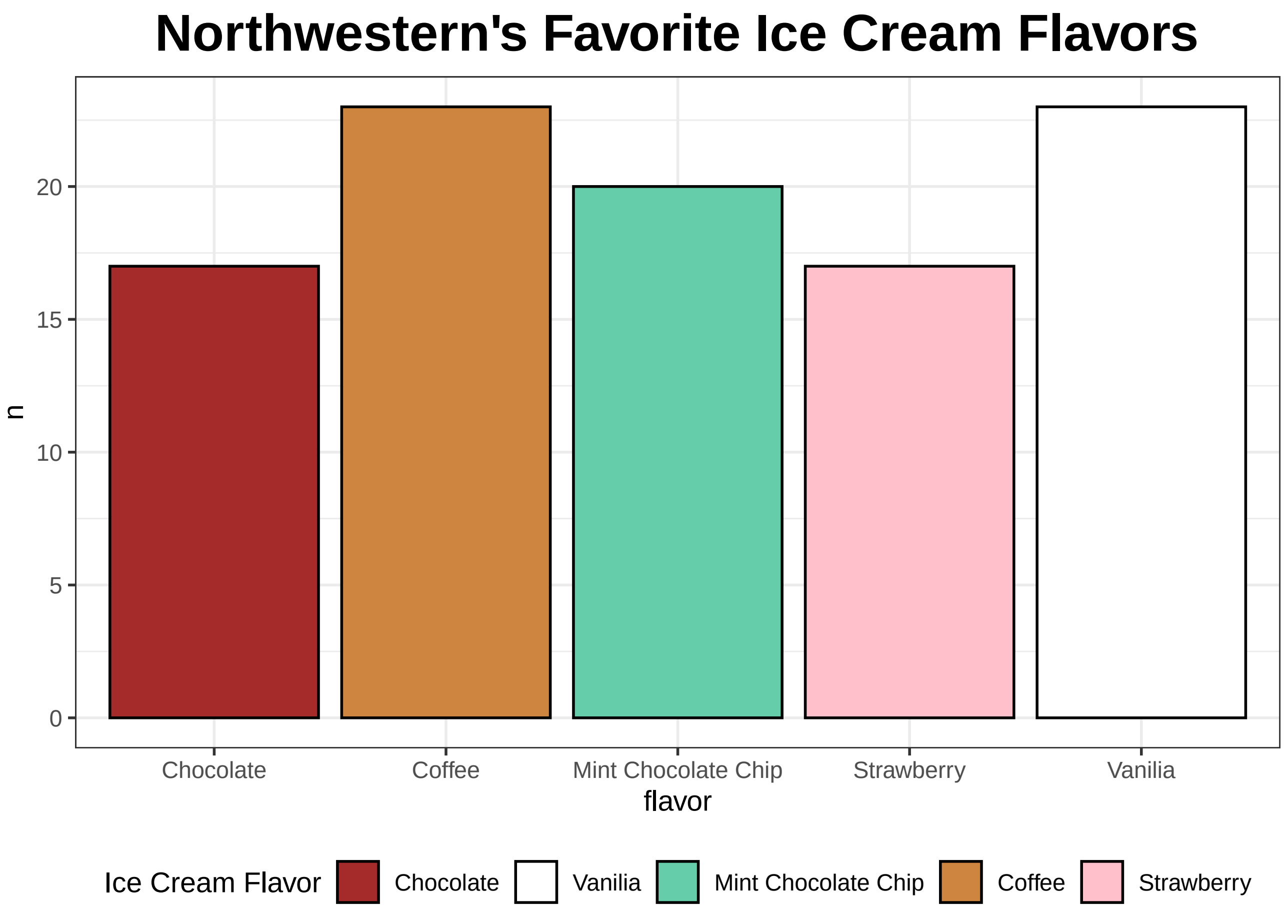
Nominal scales are categorical. In survey questions, they represent items that sort into categories, often descriptive.

What is your favorite ice cream flavor?

- Chocolate
- Vanilla
- Mint Chocolate Chip
- Coffee
- Strawberry

These data are often described using a bar graph or a pie chart. Given the small number of categories, you are likely better off just summarizing them in a table. Nonetheless, here is a bar graph of our results

flavor	n
Chocolate	17
Coffee	23
Mint Chocolate Chip	20
Strawberry	17
Vanilia	23


Northwestern's Favorite Ice Cream Flavors
n
0
5
10
15
20
Chocolate
Coffee
Mint Chocolate Chip
Strawberry
Vanilia
flavor
Ice Cream Flavor
Chocolate
Vanilia
Mint Chocolate Chip
Coffee
Strawberry

Ordinal Scales

In the case of **ordinal scales**, we are interested in questions that generate an ordered scale as the response options. These can include items such as "From a scale of 0 - 10, how much do you like ice cream?" or "Do you agree or disagree with the notion that ice cream is a good desert?" For our ice cream example, here is a question to consider. The options range from "Strongly disagree" to "Strongly agree".

> *Do you agree or disagree with the fact that Vanilla Ice Cream is your favorite ice cream flavor*

- Strongly Disagree
- Disagree
- Neither Agree nor Disagree
- Agree
- Strongly Agree

We can summarize these results in a similar way that we did the nominal results. Alternatively, we can cover them to numbers and generate means and medians.

agreement	agree	n
Strongly Disagree	1	17
Disagree	2	17
Neither Agree nor Disagree	3	27
Agree	4	15
Strongly Agree	5	24

Interval Scales

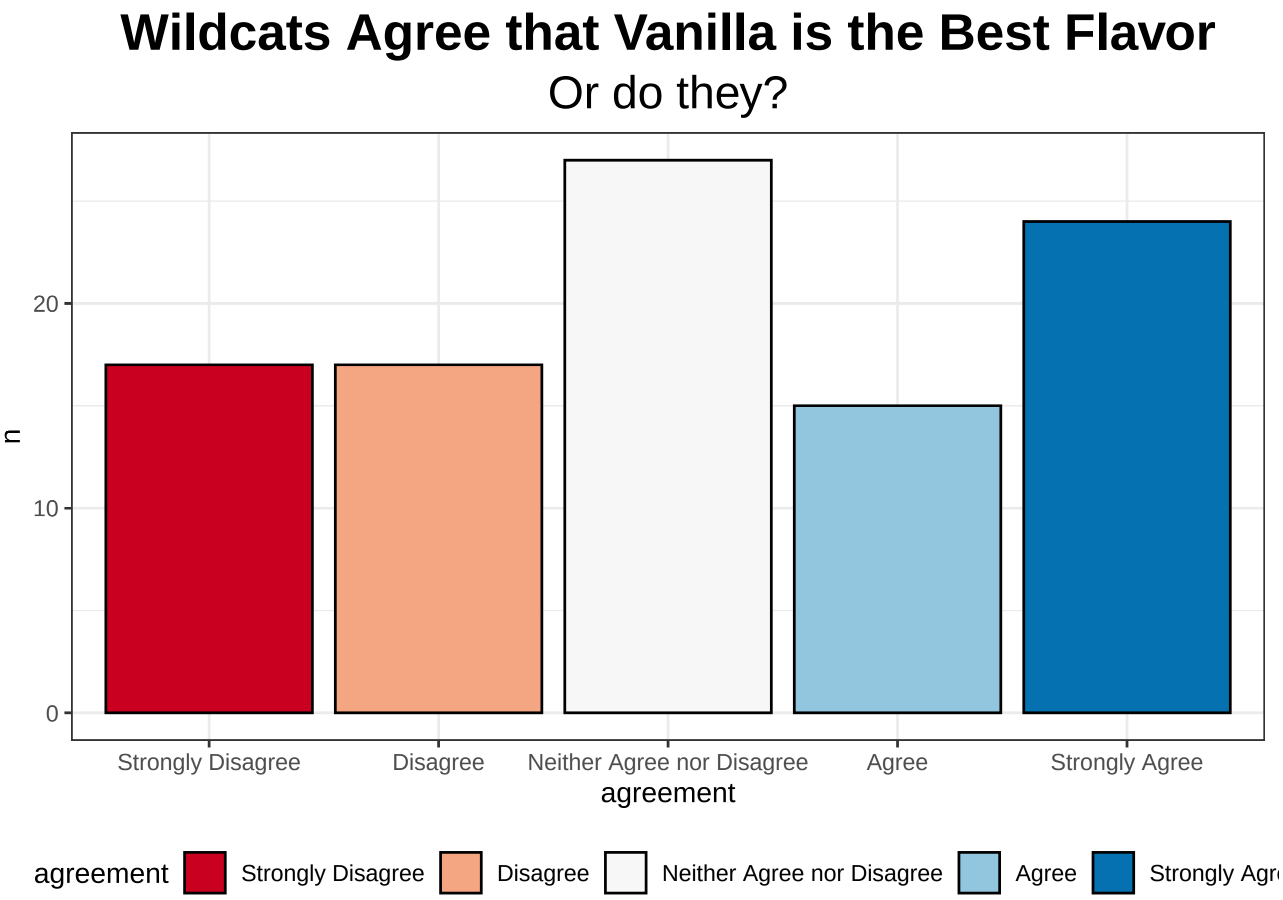
Finally, **Interval Scales** are scales that use a numeric interval in the response options. Often, these have a meaningful zero and the numbers are evenly spaced apart. When you ask your friend what the temperature is outside, they likely will give you a number on the Celsius or Fahrenheit scales that both have meaningful zeros[1]. When you buy something, the cashier would measure the amount you owe in dollars and cents, where there is also a meaningful zero. In our ice cream study, we can consider the following question.

> *How many scoops of ice cream do you eat each day?*

Instead of the graphs, we can better understand our data with means and medians.

mean	med
4.69	4

[1] I guess the Celsius more so than Fahrenheit since 0°C is the freezing point of water.


Wildcats Agree that Vanilla is the Best Flavor
Or do they?
n
20
10
0
Strongly Disagree
Disagree
Neither Agree nor Disagree
Agree
Strongly Agree
agreement
agreement
Strongly Disagree
Disagree
Neither Agree nor Disagree
Agree
Strongly Agr

Introduction to Causality

The Basics

On a hot summer day, we can reasonably expect ourselves to become dehydrated, or our ice cream to melt almost instantly. Here, we are able to pinpoint direct causes of effects that we want to observe.

However, social science is not always as straightforward. If a tree fell in the woods, and no one was there to witness it, (a) do we know it fell and (b) can we reasonably make suggestions as to how it fell?

Causal inference is about pinpointing, with *pretty good* certainty, how the tree fell. It is drawing a direct relationship between two variables X and Y. Notice the emphasis on *pretty good*. This is because we cannot gain a full sense of cause in social science research, due to the error or noise that is bound to arise in human behavior.

Let's turn to a case study on voting. Why do people go vote? Can it be based on any of the following?

- Interest in politics
- Knowledge of the candidates and issues
- Desire to make a change in the community
- Tagging along with friends and family
- Why not?

But at the same time, can any of these be an effect of voting?

The Four Questions

Using our voting example, let's focus on the "Desire to make change in the community" aspect. Can this desire lead people to vote?

- **Is there a relationship between X and Y?** – For voting, sure! Desires to make change leads to voting.
- **Could Y cause X?** – Can voting lead to greater desires to make change? Likely, since you now began a journey of political engagement
- **Is there a causal pathway from X to Y?** – Likely, if we design a model to isolate the correct variables
- **Could there be confounding variables?** – OH... ABSOLUTELY!

Experiments

Experiments are one of the best ways to identify causal mechanisms. You create two comparable groups using random assignment, give one group the treatment and another no treatment. Now, the only difference between the groups is the treatment, since both groups are, in theory, comparable (or the same).

$$
\begin{aligned}
Y_{i,T} &= \text{Treatment} \\
Y_{i,C} &= \text{Control} \\
\text{Causal Effect} &= Y_{i,T} - Y_{i,C}
\end{aligned}
$$

Suppose we want to know whether **desire to effect change** in society really causes **voter turnout**. We con conduct an experiment using the following conditions.

1. $Y_{i,T}$: Prime Desire to Effect Change in society by telling people about things in their neighborhood that need to change
2. $Y_{i,C}$: Control message – Recycling is good for the planet

If a desire to effect change causes turnout, then the treatment conditions will see a higher turnout rate than the control condition.