Dis 1: General Info; Data; Population vs. Sample

Relevant textbook chapters: 1 and 4
Ch 1 and 4 handout and solution offered by Dr. Pac can be accessed here: Handout Solution
This handout incorporates reviews with all exercises from Dr. Pac's original handout.

1 General Info

1. **Contact Me**

   You can reach me by **sending me an email** or **attending my office hours**.

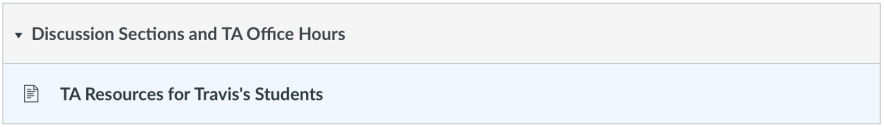
   - Email me at **travis.cao@wisc.edu** (please start the subject line with "Econ 310").
   - Office hours take place at the following times and locations:
     - Mondays, 9:15 - 10:15am, online via Zoom (link on Canvas TA page)
     - Wednesdays, 2:15 - 3:15pm, in person @ 7226 Social Sciences
     - Or by appointment

2. **Discussion Sections**

   - Attendance
     - Attendance is not required, but strongly encouraged.
     - Sections take place at the following times and locations:
       * Fridays, 9:55 - 10:45am @ 2104 Chamberlin
       * Fridays, 11:00 - 11:50am @ 2104 Chamberlin
       * Fridays, 12:05 - 12:55pm @ 2104 Chamberlin

       Due to the classroom capacity limit, please attend the section that you registered for.
     - Screen recording of the discussion section will be posted (I use my iPad to teach all sections, and my screen will be recorded and posted on Canvas after my last section of the week; all sections cover the same material, so only one section recording will be posted each week).
       * **I encourage you to attend discussion section in person every week**, since screen recording is, at best, an imperfect substitute for attending sections live.
       * Screen recording is intended as a resource to allow you to re-watch part of the section in case you didn't follow along at the time, or because you cannot make it to certain week's section due to any personal / health reason.
       * Bottom line: You all are adults. Make smart choices.
   - Discussion handout
     - If you notice the box underneath the title of this handout, you'll see that this handout differs from the version given by Dr. Pac.
       * I personally like to spend some time reviewing the concepts that we learned in the past week in sections, and then spend the remaining time going through exercises.
       * Dr. Pac's version of handout under the section "Lab Session Review Worksheets" on Canvas is great, as all the exercises come directly from him. However, I prefer to review the relevant material in a slightly different way to help students more easily digest the content, so I created this version of the handout that you are currently holding.

    - My handout contains the same exercises and reviews the same set of concepts as Dr. Pac's version, but includes additional material to help us efficiently and effectively work through the concepts together.
  - Bottom line: My version of handout is the only set of handout that you need for this course!
- Accessing discussion materials online
  - Handouts, solutions, section iPad screen recording, and any discussion material (including online office hour's Zoom link) are available on Canvas.
  - You can find them on "TA Resources for Travis's Students", which is under the module "Discussion Sections and TA Office Hours" (located at the bottom of your Canvas Econ 310 course home page).



2 Data

- What are we studying in this class about "statistics"?
  - "Statistics is the discipline that concerns the collection, organization, analysis, interpretation, and presentation of **data**." – Wikipedia
  - Data is at the core of the study of statistics, so let's first take a high level view of
    - How to categorize different types of data, and
    - What exactly can one do with a set of data
- **How to categorize different types of data?**
  - Commonly, data can be categorized based on the values recorded, or based on how much data points are collected.
  - If we categorize data **based on the values recorded**, we can divide data into 3 types:
    1. **Interval data**: The values recorded are actual numbers that make meaningful sense.
       e.g. Internet Speed = 20 Mbps, 41.3 Mbps, 537.234 Mbps, ...
    2. **Ordinal data**: The values recorded represent a ranked order.
       e.g. 1 = poor, 14 = okay, 36 = great, ...
    3. **Nominal data**: The values recorded are arbitrary (typically only used as identifier).
       e.g. 1 = Econ, 5 = History, 9 = Biology, ...

> Aside: Why does it matter to categorize data based on the values recorded?
> Consider the following three numbers that describe the central tendency of a set of data: mode, median, and mean (i.e., average).

It makes sense to calculate ... for ... type of data:

	Mode	Median	Mean
Interval data	Yes	Yes	Yes
Ordinal data	Yes	Yes	NO
Nominal data	Yes	NO	NO

* It makes sense to calculate mode for all three different data types: the mode number will reveal what's the most common observation for each data type; the most common internet speed (interval data), the most common survey response (ordinal data), and the most common major (nominal data) are all things that make sense for us to find.
* However, when finding the median, the median of a set of nominal data starts to not make much sense. All identifiers (IDs) in the nominal data are random, which means that the ID for, say, two majors can be switched easily. This switching of the ID could affect the median, but this switching is arbitrary to begin with, which means that finding the median of nominal data doesn't make sense.
* Lastly, when finding the mean, only interval data's mean still makes sense. This is because interval data records numbers that make sense on their own. The easiest way to see why the mean for ordinal data doesn't make sense is to consider the following: while we could have a set of ordinal data that represents 1=poor, 2=okay, 3=great, we could also change the numbers and use 1=poor, 14=okay, and 36=great to represent the same ranked orders between all options. This illustrates that the mean of ordinal data could easily be manipulated, so it's meaningless.

– If we categorize data **based on how much data points are collected**, we can divide data into 2 types:

1. **Population**: A set of data that records all items of interest.
2. **Sample**: A set of data that records only a subset of items of interest.

Exercise. For everyone sitting in the front row, their favorite numbers are:
Say that the favorite number of the front row students are 1, 3, 5, 8, 12, 41.
A set of data containing 1, 3, 5, 8, 12, 41 is the population data.
A set of data containing only 1, 3, 5 is a sample data.

- **What exactly can one do with a set of data?**

1. **Descriptive statistics**: A set of methods used to summarize or present your data.
   e.g. Making a bar graph from your data
   e.g. Calculating the mean (average) of your data
2. **Inferential statistics**: A set of methods used to draw conclusion or make inference about the population using a sample data.
   e.g. Say a sample from all Econ 310 students has been collected, and within the sample, 82% of them are sophomores.

Now, when asked to estimate the percentage of all students in Econ 310 that are sophomores, you might guess 82% based on the sample data.

3 Population vs. Sample

- We just mentioned that population and sample data differs based on how much data points are collected, and that two sets of methods – descriptive and inferential statistics – can be used to describe your data.
- Obviously, if one always has access to population data, then inferential statistics seem rather meaningless: you already have the population data, so there's no need to make inference about the population from a sample.
  - But, as you can intuitively see, this is likely not going to be the case: population data often is much harder to get, which is why inferential statistics matter.
  - Inferential statistics will be a big part of what we study for the rest of this semester. It's a harder set of methods compared with descriptive statistics (think about how can one tell that the inference made about the population makes sense), but it's more useful.
- For this first week, instead of looking at the more complex inferential statistics, let's look at a set of descriptive statistics that you can easily calculate: **measures of central tendency**.

Descriptive statistics for a population (Parameter)	**Descriptive statistics for a sample (Statistic)**
Population median	Sample median
Population mode	Sample mode
Population mean $\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$	Sample mean $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

- Question: Why do we sometimes refer to the mean of x as μ, and sometimes as $\bar{x}$?

  Answer: Because these two are different!

  μ (or denoted as μ_x) is the population mean (average of your population data), and $\bar{x}$ is the sample mean (average of your sample data).

  Typically, we use greek letter to refer to population parameter, and use Roman letter to refer to sample statistic. You'll see more examples of this distinction next week.

4 Exercises

1. A manufacturer claims that 1% of the artificial hearts it has ever produced are defective.

   When 1,000 hearts are randomly drawn, 1.5% are found to be defective.

(a) What is the population of interest?

Population data is defined as "a set of data that records all items of interest." Hence, the population of interest here is **all** artificial hearts ever produced by the manufacturer.

(b) What is the sample?

Sample data is defined as "a set of data that records only a subset of items of interest." Hence, the sample here is the 1,000 randomly drawn hearts (which is a subset of all artificial hearts ever produced).

(c) What is the parameter?

Parameter is a descriptive statistic for the population. In this case, the parameter records the true proportion of artificial hearts produced that are defective (which the manufacturer claims to be 1%).

(d) What is the statistic?

Statistic is a descriptive statistic for the sample. In this case, the statistic records the proportion of the 1,000 randomly drawn artificial hearts that are defective (which is 1.5% for this specific sample).

2. You are shown a coin. The owner of the coin claims it's "fair" (meaning that it will produce the same number of tails and heads when flipped a very large number of times).

(a) Describe an experiment to test the claim?

Since a "fair" coin means that the number of tails and heads should be the same when flipped for a very large number of times, the following experiment can be designed to test this claim:

Flip the coin for many times, and then count the number of heads and tails. If the number of heads roughly equal to the number of tails, then we can probably state that the coin is "fair."

For the rest of the question, suppose we do this 100 times.

(b) What is the population in your experiment?

Population records all items of interest. So in this case, the population in my experiment would record the outcome (head or tail from a flip) from an infinite number of flips.

(c) What is the sample?

Sample records a subset of items of interest. So in this case, the sample in my experiment would record the outcome (head or tail from a flip) from the 100 flips actually conducted (the outcome of 100 flips is a subset of the outcome from an infinite number of flips).

(d) What is the parameter?

Parameter is a descriptive statistic for the population. Here, we are interested in whether the coin is fair. To achieve this goal, you might instinctively want to record two parameters: number of heads from the infinite flips, and number of tails from the infinite flips, and then compare one against the other. But we can simplify this to record one set of measure by looking at the proportion. So in this case, let's consider the parameter to be the proportion of heads (or tails) if you were to flip the coin infinitely many times.

(e) What is the statistic?

Statistic is a descriptive statistic for the sample. Following the proportion example from parameter, we consider the statistic of interest here to be the proportion of heads (or tails) when you flip the coin 100 times.

(f) Recall your goal is to determine whether the coin is fair. What conclusion would you draw if 99 of the 100 flips came up heads? What conclusion would you draw if 50 of the 100 flips came up heads?

This questions is ultimately subjective. Later in the semester, we'll discuss how most statisticians would answer this question. For the time being, however, it suffices to say that if you observe 99 heads, it seems highly unlikely that the coin is fair. On the other hand, if you observe exactly 50 heads, then it seems quite plausible that the coin is indeed fair.

3. Consider grade data for the following sample of students (drawn randomly from the entire population of 350 students who took Econ 310 last semester):

Student	**Grade**
Tom	80
Sean	90
Ed	60
Ben	70
Nate	80

(a) What are N and n? Describe the difference between the two.

N is the number of observations in the entire population; in this case, $N = 350$.

n is the number of observations in the sample; in this case, $n = 5$.

(b) What are $\bar{x}$ and μ? Describe the difference between the two.

Let x be the grade of students in Econ 310.

$\bar{x}$ is the sample mean; in this case,

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{5} x_i = \frac{1}{n} \times (80 + 90 + 60 + 70 + 80) = 76$$

μ is the population mean; in this case, it is unknown, since we do not have data for the entire population (i.e., grades for all 350 Econ 310 students from last semester).

(c) Calculate the median, mode, and range.

To calculate these measures, it is helpful to first sort the data from the lowest to the highest, which looks like the following:

$$60, 70, 80, 80, 90$$

Median (specifically sample median in here) is the middle observation among the sorted data. Here, median $= 80$.

Mode (specifically sample mode in here) is the most common observation among the data. Here, mode $= 80$.

Range (specifically sample range in here) records the difference between the largest observation and the smallest observation. Here, range $=$ max $-$ min $= 90 - 60 = 30$.

4. Ten people in a room have an average height of 5 feet 6 inches. An 11th person, who is 6 feet 5 inches tall, enters the room. Find the average height of all 11 people?

   Note: 1 feet = 12 inches

   In this question, notice that two sets of units are used for measuring heights: feet, and inches. To simplify our problem, it is helpful to first convert everything to one set of unit. People tend to convert things to the smaller unit, so this tells us that,

   - Average height of 10 people in the room $= 5 \times 12 + 6 = 66$ inches
   - The height of the 11th person $= 6 \times 12 + 5 = 77$ inches

   With that out of the way, if you don't know where to start for solving the problem, it is always helpful to write down the information we already have explicitly. Here, we know that the average height among the first 10 people is 66 inches, which means that

$$\frac{1}{10}\sum_{i=1}^{10} x_i = 66$$
$$\sum_{i=1}^{10} x_i = 66 \times 10 = 660$$

   where x_i is defined as the height of the ith individual.

   Now, writing down the expression for what we are asked to solve (average height of all 11 people), we notice that some substitution can be done with the information we had previously:

$$\begin{aligned}
\text{average height of all 11 people} &= \frac{1}{11}\sum_{i=1}^{11} x_i \\
&= \frac{1}{11}\left[\left(\sum_{i=1}^{10} x_i\right) + x_{11}\right] \\
&= \frac{1}{11}\left[(660) + 77\right] \\
&\quad (x_{11} \text{ is the height of the 11th person, which we know is 77 inches}) \\
&= 67 \text{ inches} = 5'7''
\end{aligned}$$

   Thus, the average height of all 11 people is 67 inches (or 5′7″).