Lec 1*: Data; Population vs. Sample; Descriptive Statistics

1 Data

- What are we studying when we say we are studying "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, 2 = okay, 3 = great, ...
    3. **Nominal / categorical data**: The values recorded are arbitrary (typically only used as identifier).
       e.g. 1 = Sociology, 2 = Econ, 3 = History, ...
  - 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.

    e.g. 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 first-year Ph.D. statistics class's students has been collected, and within the sample, 82% of them are from the midwest.
     Now, when asked to estimate the percentage of all students in this class that are from the midwest, you might guess 82% based on the sample data.

*Some exercise questions are taken from or slightly modified based on Dr. Gregory Pac's Econ 310 discussion handout.

2 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 guess, 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 some descriptive statistics that you can use.

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!

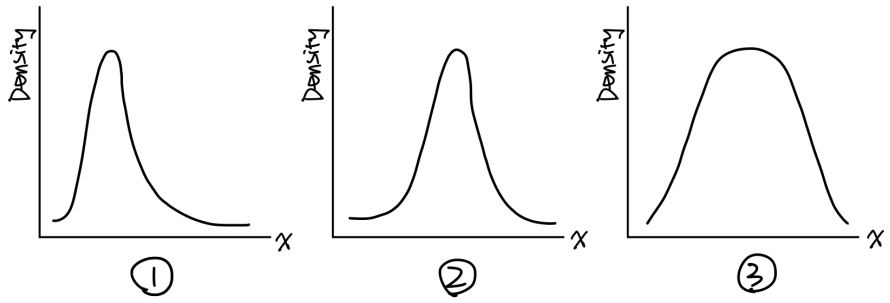
  μ (or denoted as μ_x) is the population mean (average of your population data), and $\bar{x}$ is the sample mean (average of your sample data).

  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.

3 Descriptive Statistics

3.1 For a single variable (x)

- Recall that descriptive statistics are a set of methods that summarize or present your data.
- For example, say that you have three different sets of data distributed in the following way:



How can we tell the three data apart?

- **Method 1**: Use measures of central tendency

Name	**Population Notation**	**Sample Notation**	**Formula**
Mean	μ or μ_x	$\bar{x}$	Parameter: $\mu_x = \frac{1}{N}\sum_{i=1}^{N} x_i$ Statistic: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
Median	-	-	Parameter and statistic: Sort all data, and take the middle one's value (or the average of the two middles)
Mode	-	-	Parameter and statistic: The most common observation(s)

- **Method 2**: Use measures of variation

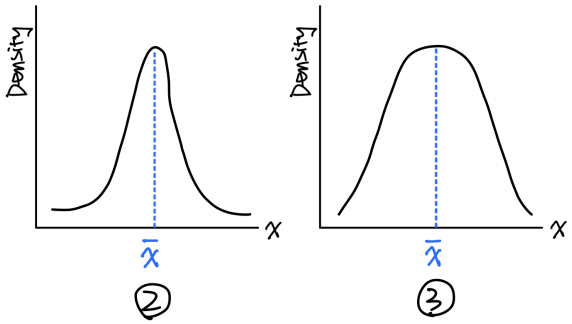
Name	**Population Notation**	**Sample Notation**	**Formula**
Variance	σ^2 or σ_x^2	s^2 or s_x^2	Parameter: $\sigma_x^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_x)^2$ $= \left(\frac{1}{N}\sum_{i=1}^{N} x_i^2\right) - \mu_x^2$ Statistic: $s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ $= \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}\right]$
Standard deviation (std)	σ or σ_x	s or s_x	Parameter: $\sigma_x = \sqrt{\sigma_x^2}$ Statistic: $s_x = \sqrt{s_x^2}$

Side note: How does variance measure the variation of x?

Recall that for sample data, variance formula is

$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Looking at the second and third data from earlier,



Clearly, data #3 has greater variation compared with data #2, so how is this reflected in the variance formula?

Notice that #3 and #2 have similar sample means ($\bar{x}$). To simplify things, let's just say that their sample means are exactly the same.

Based on the sample variance formula, we then need to look at how far each data point x_i is from the sample mean $\bar{x}$, and then sum up all these differences.

For #2, each x_i is relatively close to $\bar{x}$. For #3, each x_i is relatively far away from $\bar{x}$. Thus, the sum of these differences would be much bigger for #3 than it is for #2.

Finally, the sample variance formula divides the sum of differences by the degree of freedom $(n-1)$. With #2 and #3's data in a similar range, the number of x_is are very close. Say that these ns are the same. Thus, the division of $(n-1)$ does not change things. So the sample variance in #3 is bigger than in #2 because of the bigger sum of differences between x_i and $\bar{x}$, which reflects the fact that #3 has greater variation in x than #2.

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. Consider grade data for the following sample of students (drawn randomly from the entire population of 350 students who took Intro to Statistics class last semester):

Student	**Grade**
Kendall	80
Shiv	90
Roman	60
Logan	70
Marcia	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 Intro to Statistics.

   $\bar{x}$ 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 Intro to Statistics students from last semester).

   (c) Calculate the median and mode.

   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 $= 80$.

3. 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:

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

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