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.
    2. **Ordinal data**: The values recorded represent a ranked order.
       e.g.
    3. **Nominal / categorical data**: The values recorded are arbitrary (typically only used as identifier).
       e.g.
  - 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.

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

- **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.

     e.g.

  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 _____

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.

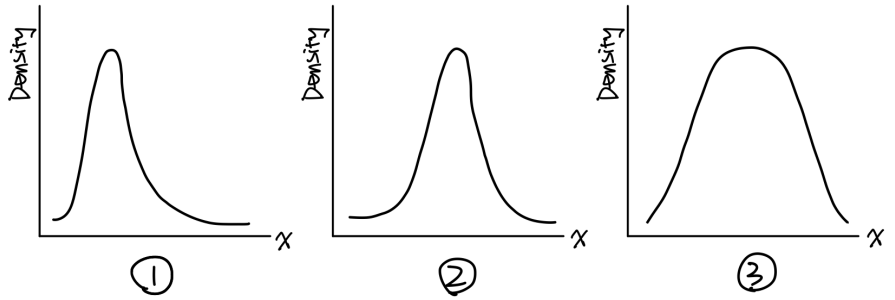
Descriptive statistics for a population (Parameter)	**Descriptive statistics for a sample (Statistic)**
Population median	Sample median
Population mode	Sample mode
Population mean $\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$	Sample mean $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

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

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

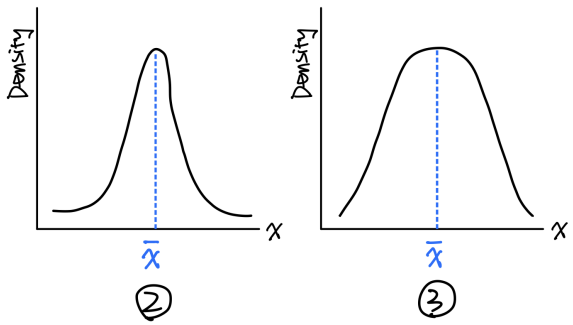
Name	**Population Notation**	**Sample Notation**	**Formula**
Variance	σ^2 or σ_x^2	s^2 or s_x^2	Parameter: $\sigma_x^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_x)^2$ $= \left(\frac{1}{N}\sum_{i=1}^{N} x_i^2\right) - \mu_x^2$ Statistic: $s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ $= \frac{1}{n-1}\left[\sum_{i=1}^{n}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

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?

   (b) What is the sample?

(c) What is the parameter?

(d) What is the statistic?

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.

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

Student	Grade
Kendall	80
Shiv	90
Roman	60
Logan	70
Marcia	80

(c) Calculate the median and mode.

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