GSE Math Camp: Stats Week!

Lecture 1 - Introduction to Data Distributions

1. Population vs. Sample

Population Parameters

- **Population:** The entire set of individuals about whom the researcher is interested in learning
- **Population Parameters:** These parameters are the “true” parameters that we are interested in estimating with our sample.

- Ex: Census

Population Mean	Population Standard Deviation
μ	σ

Sample Statistics

- **Sample:** A subset of the population for study, called a sample. There are multiple ways to sample, but we will focus on random samples. Randomly selected samples, on average, reflect the characteristics of the whole population so we do not have to worry about bias.
- **Sample Statistics:** Any summary calculation based on the data is a sample statistic, BUT we are particularly interested in those that estimate the population parameters.

Sample Mean	Sample Standard Deviation
$\bar{X}$	S

Population

- **Entire** group of interest
  - ALL U.S. school children
  - ALL Palo Alto teachers

Parameter

- What we want to know about the population
- Usually difficult to know, so we estimate it

Sample

- People/objects we actually talk to or measure
  - 300 U.S. school children
  - 20 Palo Alto teachers

Statistics

- Value we calculate from our sample
- Used as an estimate for our parameter

Group Task #1

In groups of 2-3, identify the population, parameter, sample, and statistic for the two scenarios provided.

** See Lecture 1 Tasks file on GitHub.

2. Data Types

Two Types of Data: Categorical & Quantitative

- A **categorical variable** names categories and answers questions about how cases fall into those categories. Race, sex, age group, and educational level are examples of categorical variables. Note that age (rather than age group) and grade (rather than educational level) would not be categorical.
- A **quantitative variable** has units and answers questions about the quantity of what is being measured. Age, grade, height, and SAT scores are quantitative.

Group Task #2

In groups of 2-3, determine whether the given variables are categorical or quantitative. For any quantitative variable, also name the unit in which it was measure, if given.

** See Lecture 1 Tasks file on GitHub.

3. Distributions of Quantitative Data

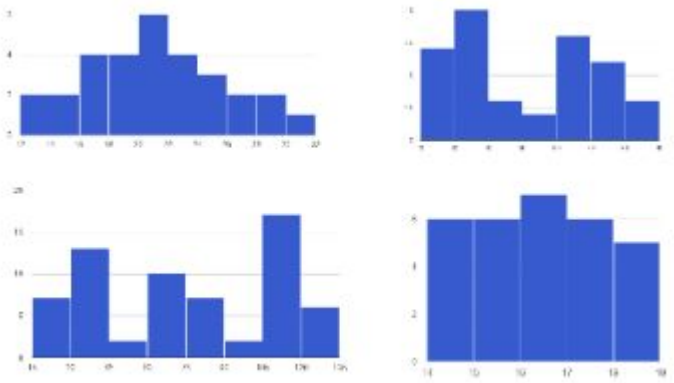
Data Distributions

- Data distributions are what you would see if you made a histogram or are represented when you calculate the summary statistics for that sampled data.
- These distributions have certain characteristics that are important to note when analyzing and describing them.
  - Shape
  - Center
  - Spread
  - Unusual Features

Shape

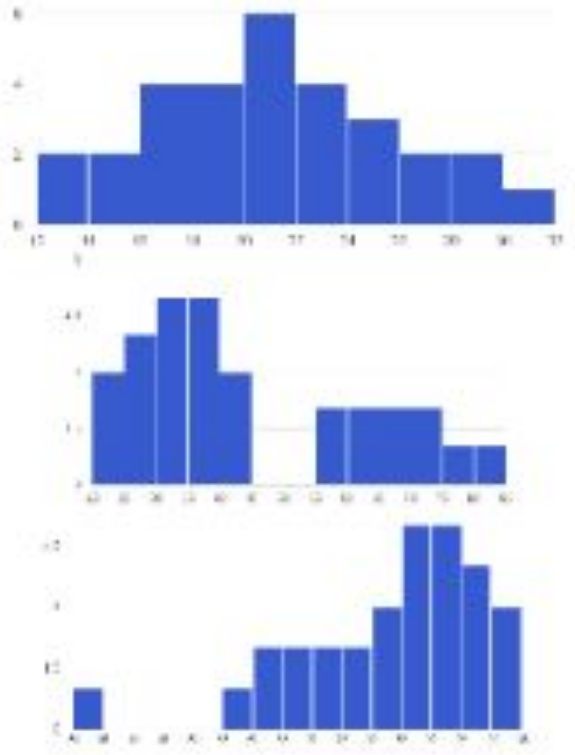
A distribution's shape is described both by its modality and its symmetry.

- Modality - can be thought of as the number of "humps" a distribution may have. These "humps" are called modes. We can see distributions that are **unimodal** (one hump), **bimodal** (two humps), **multimodal** (3+ humps) or **uniform** (no apparent humps).



Shape

- Symmetry - is considered by imagining a vertical line through the middle of the distribution. If the distribution is then folded, would the halves line up? If so, the distribution is **approximately symmetric**. If not, the distribution is **skewed**. We determine the type of skew (right or left) by the longest tail of the distribution.



Center

When thinking of a typical value for a distribution, we often consider its center. There are two measures of centers that are often used.

Mean

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

**Note: n is the number of observations in our sample AKA sample size.

Median

Value in the middle of the ordered data set. If you have an even number of observations, the median is the average of the two middle observations.

Individual Task #1

Calculate the mean and the median for the following datasets:

#1) {7, 9, 8, 5, 8, 6, 7, 7, 6}

#2) {6, 3, 2, 4, 3, 3, 5, 7, 2}

Spread

When numerically describing a distribution, we always report a measure of spread along with a measure of center. There are two measures of spread that we'll consider for now.

Standard Deviation

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

Range

The range is the difference between the maximum number in the dataset and the minimum number in the dataset.

Individual Task #2

Calculate the sd and the range for the following datasets:

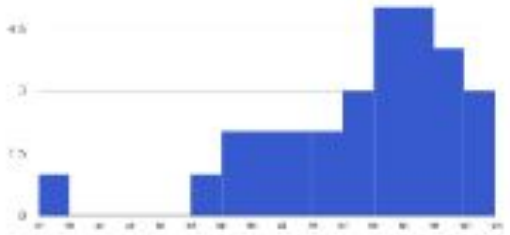
#1) {7, 9, 8, 5, 8, 6, 7, 7, 6}

#2) {6, 3, 2, 4, 3, 3, 5, 7, 2}

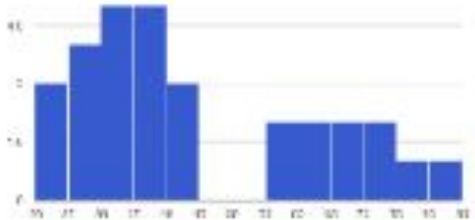
Unusual Features

Distributions sometimes have interesting features that can help us further examine and understand the data.

Outliers



Gaps



Group Task #3

In groups of 2-3, using your assigned dataset, quickly sketch a histogram or dotplot of the data, find their shapes, means, medians, standard deviations, ranges, and identify any unusual features.

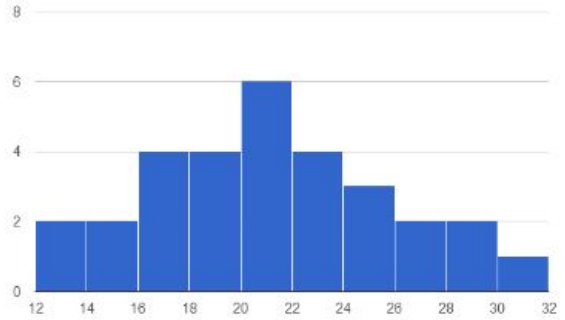
** See Lecture 1 Tasks file on GitHub.

4. Appropriate and/or Resistant Measures

Appropriate and/or Resistant Measures

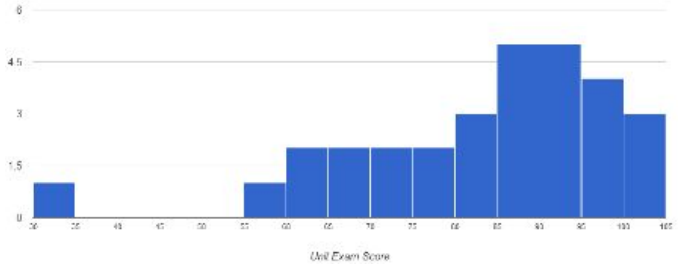
- Not always appropriate to use just *any* measure of center or spread. Rather, the shape and features of the distribution dictate which are appropriate.

Unimodal, Appx Symmetric



Mean = Median
** Use mean & standard deviation

Skewed or w/ Outliers



Right Skew (or Outlier): **mean > median**
Left Skew (or Outlier): **mean < median**
** Use median

Why?

Group Task #4

In groups of 2-3, work through the task, calculating the mean, median, range, and standard deviation of the two datasets.

#1) {1,2,3,4,5,66}

Oh no! There was a typo in your dataset.

#2) {1,2,3,4,5,6}

Resistant Measures of Center and Spread

- Median is a **resistant** measure of center. Mean is not.
- There are resistant measures of spread. Standard Deviation is not one of them.

Group Task #5

In groups of 2-3, return to your previous datasets from Mrs. X's classroom. Using the appropriate measures of center and spread, describe the distribution in a short paragraph.

5. Comparing Distributions

Comparing Distributions

- Often we need to compare sets of data to one another. Although there are lots of stylistic choices with which to do this, it is typical to compare distributions on each or most of their characteristics.
  - When comparing **two** distributions, use comparative, "-er," language.
  - When comparing **three or more** distributions, use superlative, "-est," language.
- Also, comparing distributions of data requires that you compare the same statistics. That is, a comparison of Distribution A's range to Distribution B's standard deviation would be of little use.