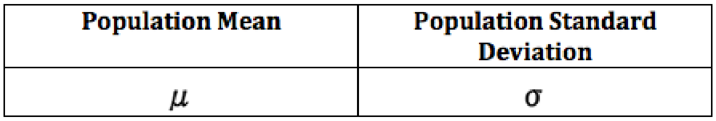
Intro to Distributions of Data Notes[1]

Stanford GSE Math Camp

Do Not Distribute Outside GSE

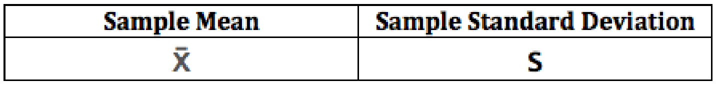
1 Population vs. Sample

- **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.



Population Mean	Population Standard Deviation
μ	σ

- **Sample:** It is often impractical to study the WHOLE population, so we take 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

2 Data Types

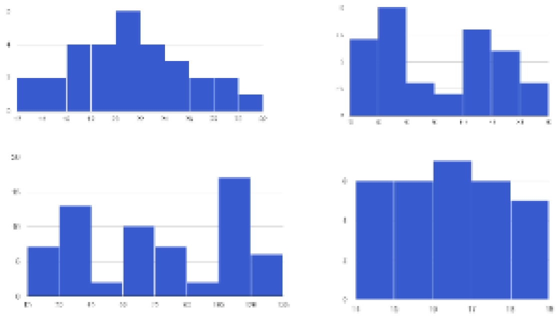
- There are different types of data that you will encounter. Different types of data necessitate different analyses and graphic representation. The two types we’ll focus on are categorical and quantitative.
  - **Categorical Data**: type of data that can be divided into groups. A **categorical variable** 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.
  - **Quantitative Data**: type of data that is numerical and reflect measurement. A **quantitative variable** has units and answers questions about the quantity of what is being measured. Age, grade, height, and SAT scores are quantitative.

3 Distributions of Quantitative Data

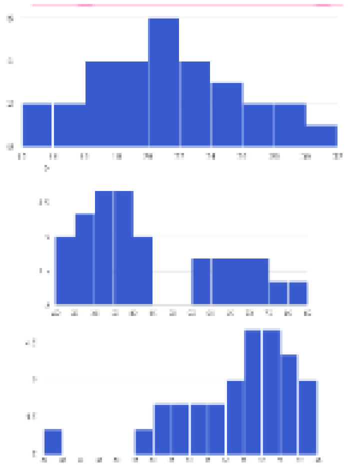
- Sample statistics describe the *distribution of the sample.* This is simply the distribution of the values in your sample, i.e. what you would see if you made a histogram or when you calculate the summary statistics for that sampled data. These distributions have certain characteristics that are important to note when analyzing and describing them.

[1] Contributor(s): Kelly Boles and Erin Fahle. If you find errors, please let us know so that we may correct them. Thanks!

- **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).



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.

median

The value that is right 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.

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

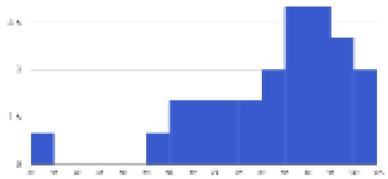
standard deviation

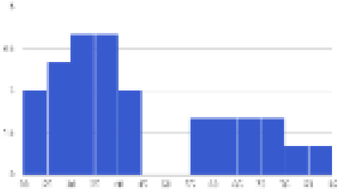
$$SD_x = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

range

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

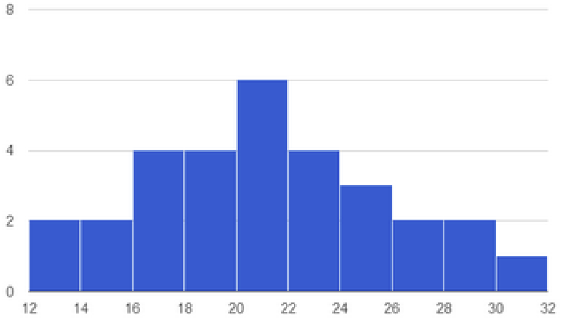
- **Unusual Features**: Distributions sometimes have interesting features that can help us further examine and understand the data. Outliers and gaps are two such features.



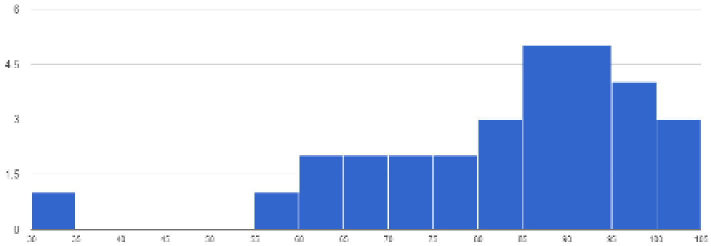


4 Appropriate and/or Resistant Measures

- It is not always appropriate to use just *any* measure of center or spread at any time. Rather, the shape and features of the distribution dictate which are appropriate.
  - Unimodal, Approximately Symmetric Distribution: mean = median. We usually use mean and standard deviation to describe these distributions.



  - Skewed Distributions, Distributions with Outliers: If the distribution has a high outlier or is skewed right, then the mean is *greater than* the median. If the distribution has a low outlier or is skewed left, then the mean is *less than* the median. In these instances, we typically use the median and range as measures of center and spread.



- The median and range are considered **resistant measures** of center and spread, respectively, since their values are less affected by skew or outliers.

5 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.