University of Science, VNU-HCM
Faculty of Information Technology

Introduction to Data Science Course

Basic Statistics

Le Ngoc Thanh
lnthanh@fit.hcmus.edu.vn
Department of Computer Science

Ho Chi Minh City

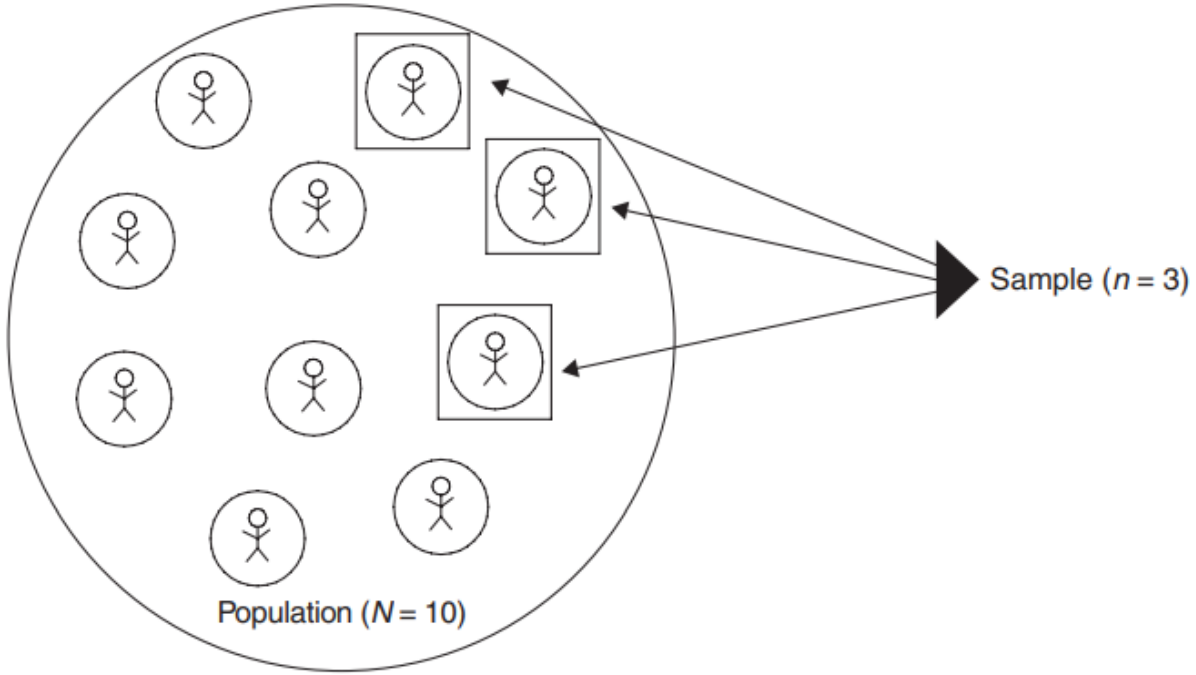
Contents

Population vs. Sample

- What is the average income of citizens in Ho Chi Minh City?
  - Get a list of all citizens in HCM City and find out the income, so on…
  - Pros/Cons?
  - Is there any other ways better?

Population vs. Sample

- A population is an individual or group that represents all the members of a certain group or category of interest.
- A sample is a subset drawn from the larger population.
  - Randomly select a subset is known as a random sample.



Quiz. Which is population and sample?

Mr.Thanh wanted to know the average height of the students in his data science class this term. He collects height of each student and then calculate on them.

- Which type of this collection? (population/sample)

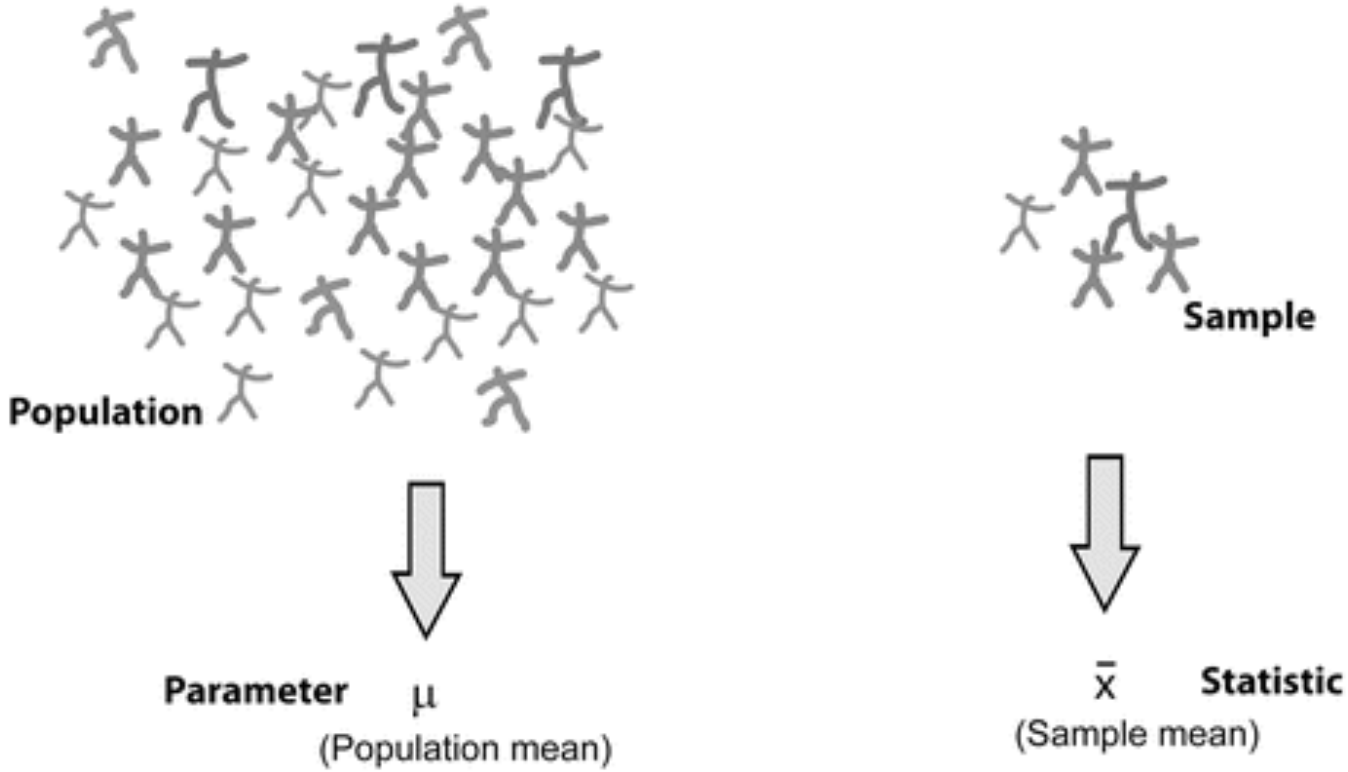
Quiz. Which is population and sample?

Lucio wants to know whether the food he serves in his restaurant is within a safe range of temperatures. He randomly selects 70 entrees and measures their temperatures just before he serves them to his customers.

A. The population is all of the hot entrees Lucio serves; the sample is the entrees that are a safe temperature.
B. The population is the 70 selected entrees; the sample is the entrees that are a safe temperature.
C. The population is all of the entrees Lucio serves; the sample is the 70 selected entrees.

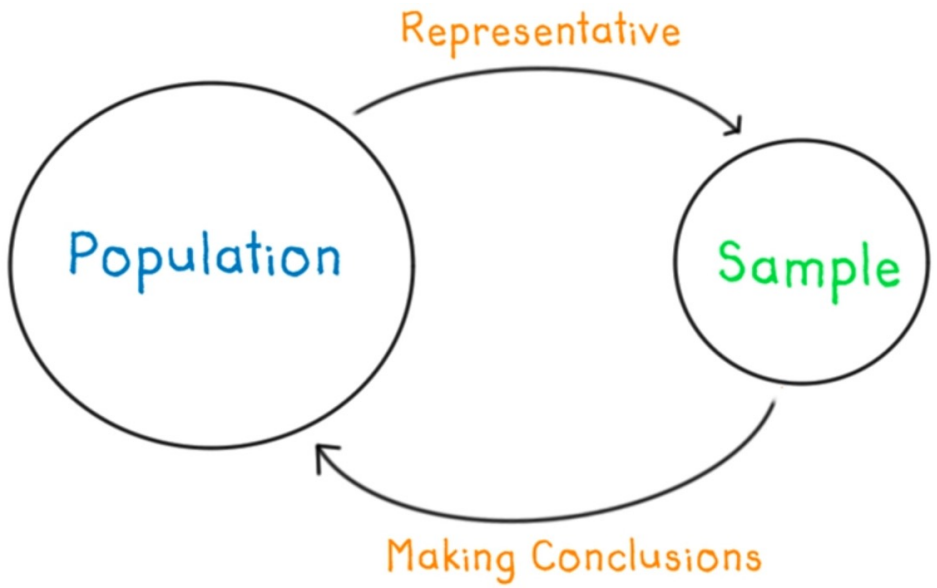
Population vs. Sample

- A population…
  - Parameters are numerical descriptions of population characteristics.
- A sample ...
  - Statistics are numerical descriptions from sample data.



Descriptive vs Inferential Statistics

- Which kind of statistics for following scenarios?
  1. The average weekly income for all students in our class is $405
  2. The average weekly income for a sample of 450 college students is $325.
  3. A recent survey of a sample of 450 college students reported that the average weekly income for students is $325.



Descriptive vs Inferential Statistics

- Scenario:

  In a recent study, volunteers who had less than 6 hours for sleep were four times more likely to answer incorrectly on a science test than were participants who had at least 8 hours of sleep.

- Decide which part is a the descriptive statistic and what conclusion might be drawn using inferential statistics.

What if …?

- What if the sample is not truly representative of the population?
  - We cannot be confident that conclusions based on our sample data will apply to the larger population.

➔ Sampling Issues

Sampling

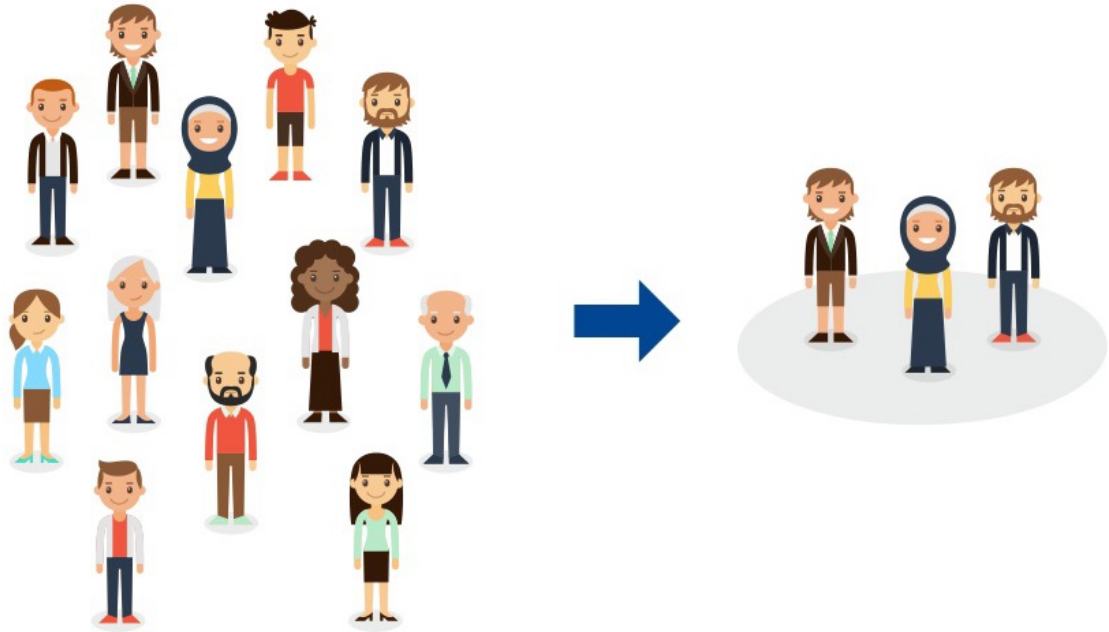
- Ways to select samples:
  - Random sampling: the most useful, but also the most difficult
    - *Random* means that every member of a defined population has an equal chance of being selected into a sample.
    - Unbiased or biased sampling?

Sampling

- Ways to select samples:
  - Representative sampling: selects cases so that they will match the larger population on specific characteristics.
  - Unbiased or biased sampling?

Sampling

- If you want to conduct a study examining the average annual income of adults in San Francisco, how to random sample or representative sample?
  - This population includes a number of subgroups (e.g., different ethnic and racial groups, men and women, retired adults, disabled adults, parents, single adults, etc.).
  - These different subgroups may be expected to have different incomes.



Example

- Identify color of leaves in September with some following ways. Which is biased or unbiased sampling?
  - 100 fallen leaves collected from the ground
  - 100 leaves on tree branches
  - 50 fallen leaves and 50 leaves on branches
  - 50 fallen oak leaves and 50 oak leaves on branches

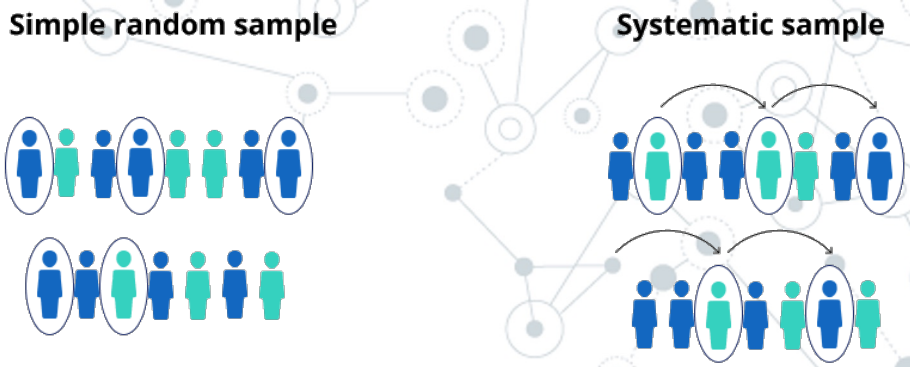
Practical examples about sampling

- To test for pollution levels in the ocean of Southern California, researchers took a sample of water and test it. If the pollution levels are too high in the sample, the beach is declared unsafe and is closed.
  - How about this process?

Practical examples about sampling influences results

- (Doshi, 2015) For decades, doctors and medical researchers considered heart disease to be a problem only for men.
- Reason:
  - Doctors were less likely to order testing for heart disease for their female patients than their male patients.
  - The symptoms of heart disease and cardiac failure among women, which are often different from those of men, were not understood.

Other views about sampling

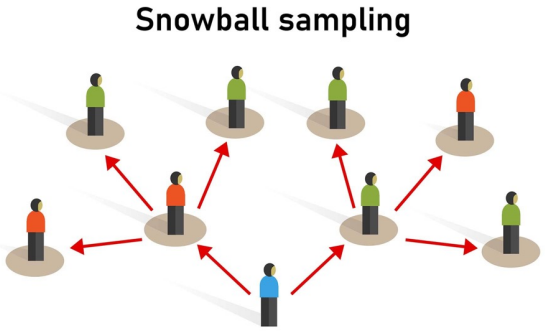


Probability Sampling Techniques:

- Random sampling
- Systematic sampling: select members of the population at a regular interval.
- Stratified sampling: select specific proportions of individuals from various subpopulations (strata).

Non-probability Sampling Techniques:

- Convenience sampling: select members which is close to hand.
- Snowball sampling: participants recruit other participants for a test.
- Quota sampling: decide quotas so that the samples can be useful in collecting data

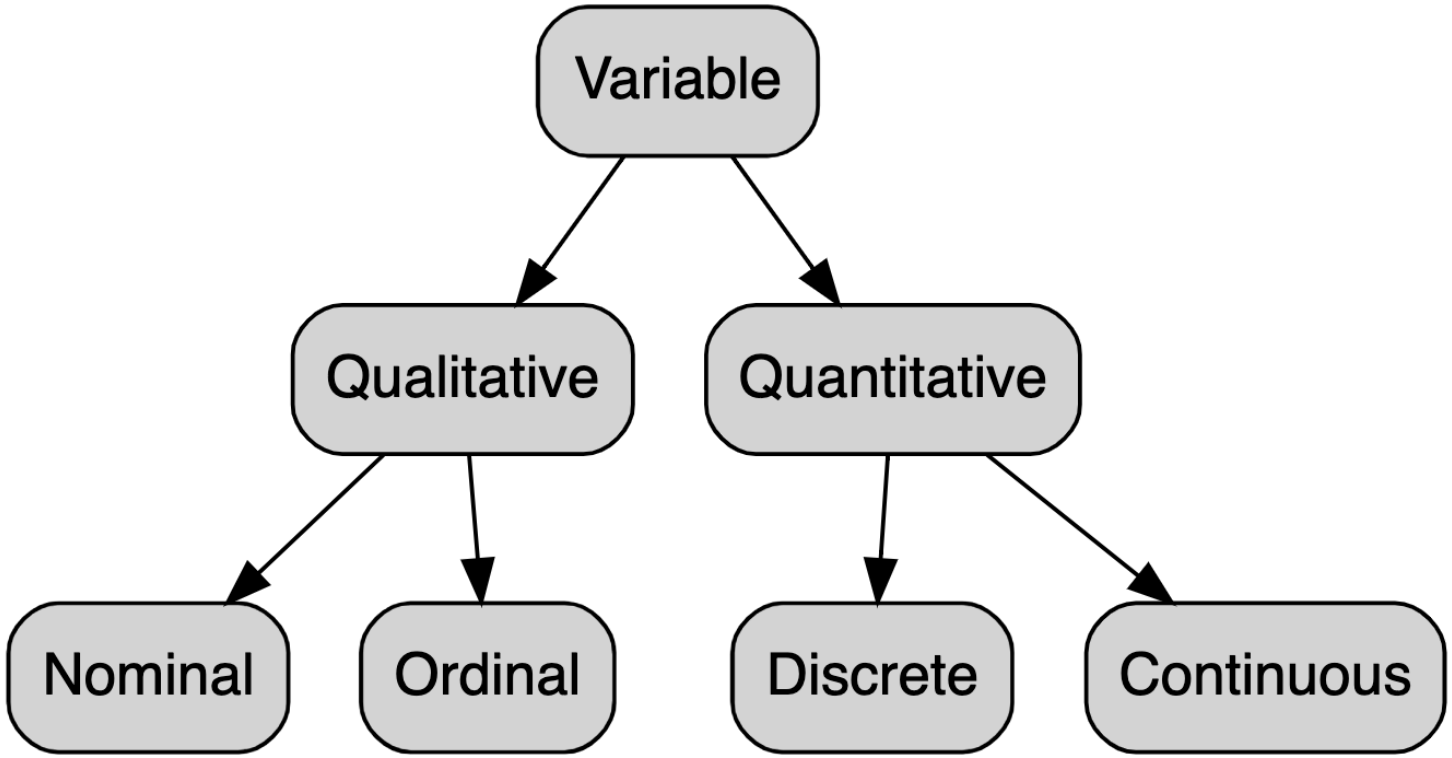


Variable and constant

- A variable is pretty much anything that can be codified and have more than a single value.
  - Example: income, gender, age, height, etc.
- A constant has only a single score.
  - Example: male, female, 25, etc.

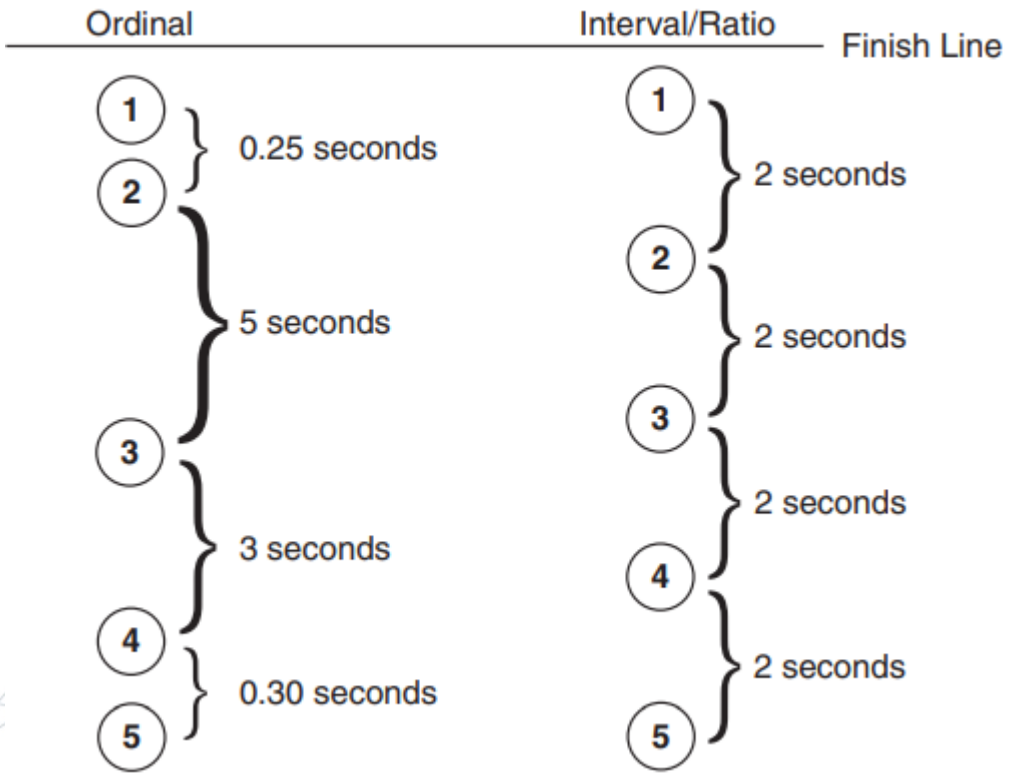
Types of variables

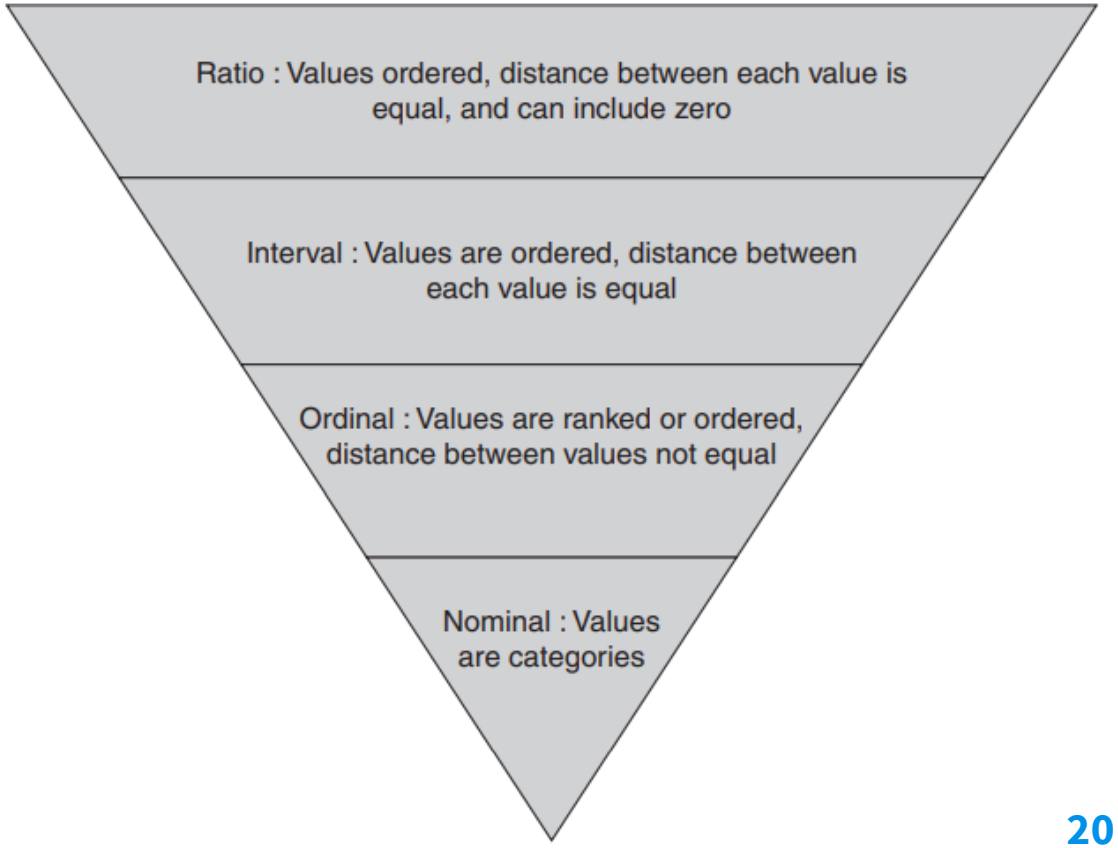
- Types of variables include qualitative (nominal, ordinal categorical) and quantitative (discrete, continuous).



Scales of measurement

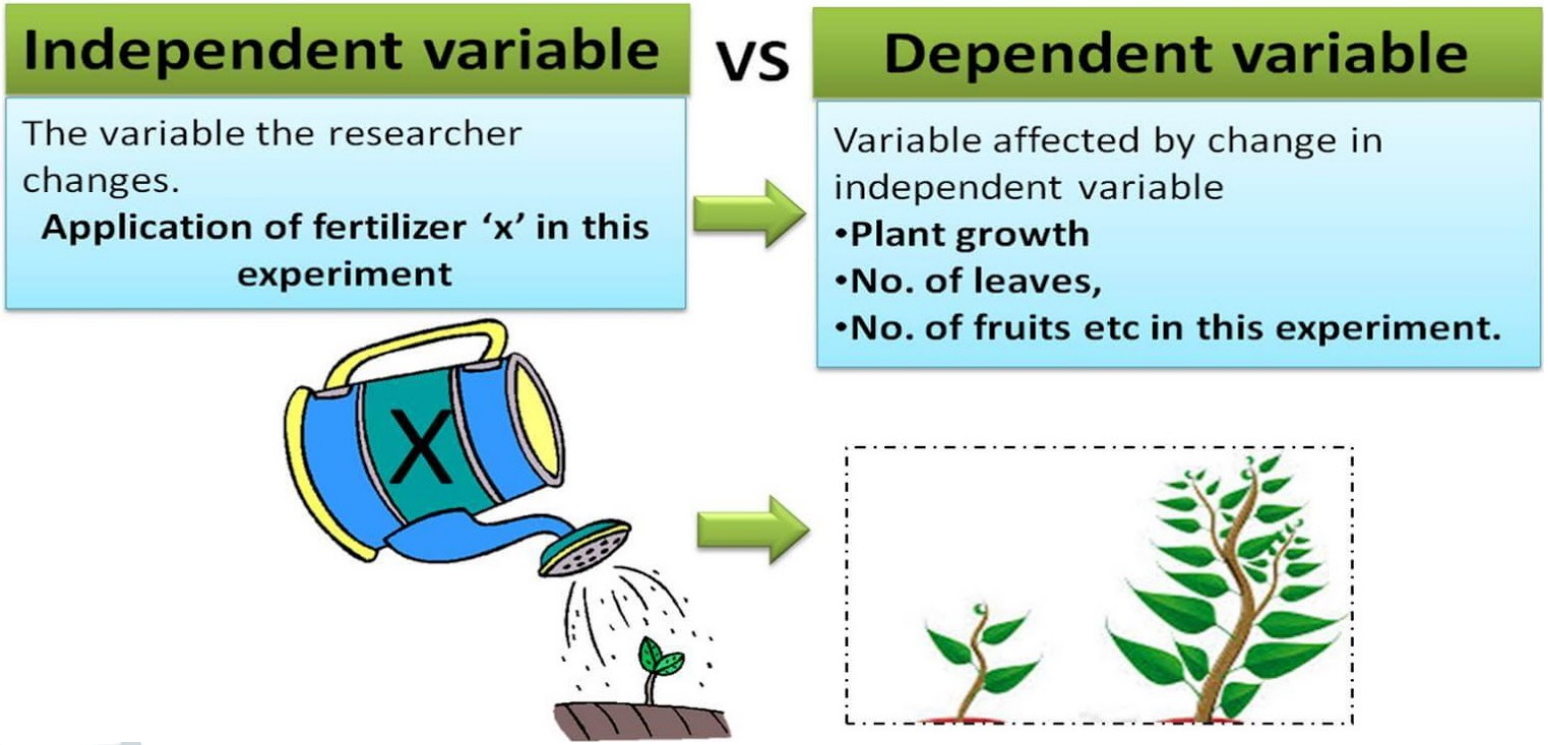
- Qualitative variables has 4 scales: nominal, ordinal, interval, and ratio.





Independent vs dependent variables

- The independent variable is the variable the experimenter manipulates or changes, and is assumed to have a direct effect on the dependent variable.



Research Design

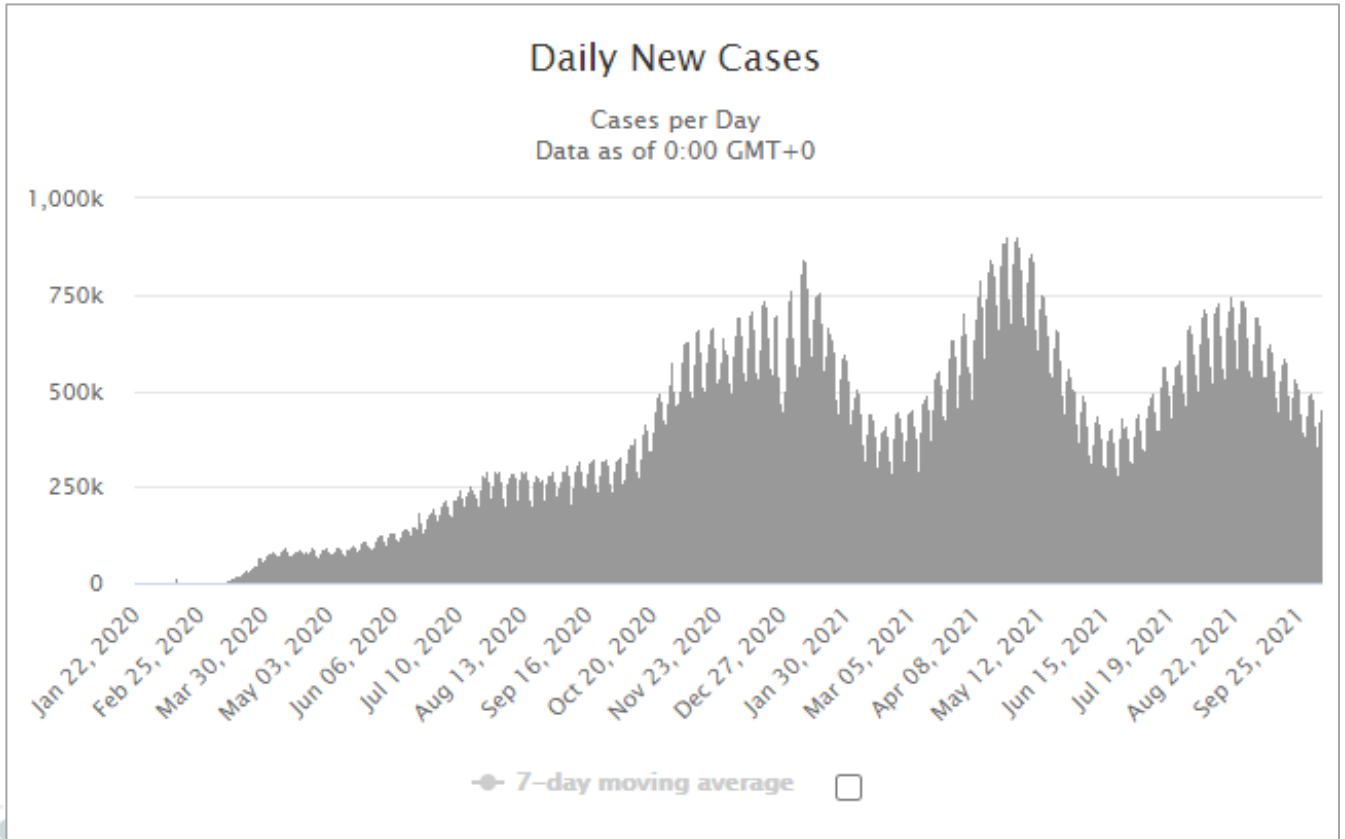
- Some research designs:
  - Experimental design: divides the cases in the sample into different groups and then compares the groups on one or more variables of interest.
  - Quasi-experimental research design: a experiment occurs outside of the lab, in a naturally occurring setting.
  - Correlational research designs: participants are not usually randomly assigned to groups, the researcher typically does not actually manipulate anything, the researcher simply collects data on several variables and then conducts some statistical analyses to determine how strongly different variables are related to each other.

After collect data

- Whenever you collect data, you end up with a group of values/scores on one or more variables.
- What do you do next?
  - Preprocessing
  - …

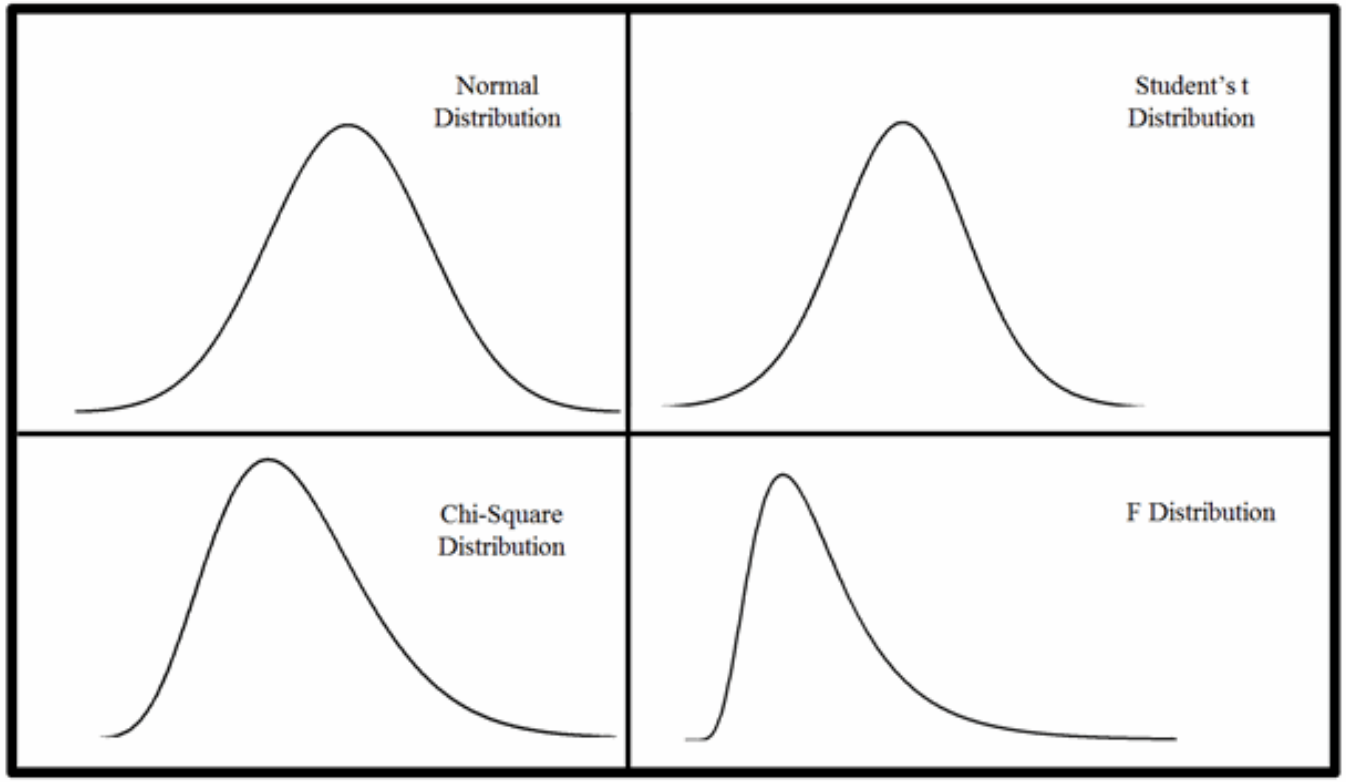
Distribution

- A distribution is simply a collection of data, or scores, on a variable which are arranged in order from smallest to largest and then they can be presented graphically.



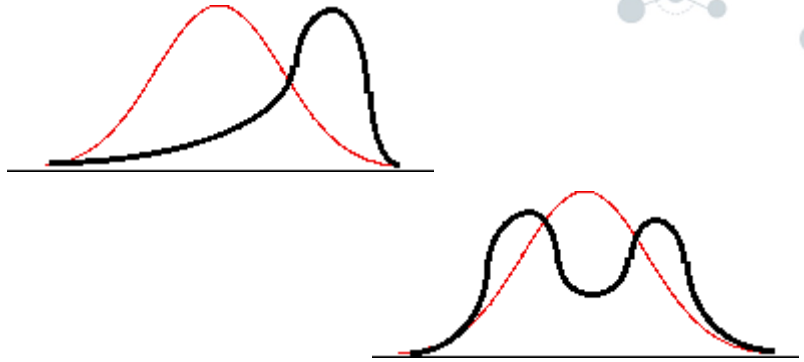
Distribution Types

- There are many distribution types:
  - Nomal distribution
  - t distribution
  - F distribution
  - chi-square distribution
  - …



Distribution characteristics

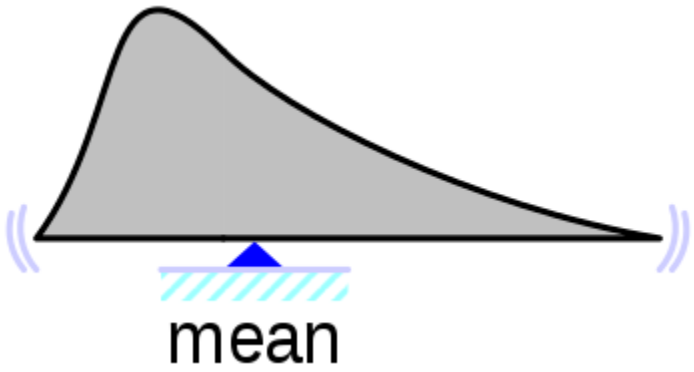
- Some distribution characteristics includes:
  - Shape
  - How spread out the value are (dispersion)
  - max, min
  - …



- However, interested characteristics are central tendency and variability.
  - Central tendency: mean, median, and mode.
  - Variability: range, variance, and standard deviation.

Mean

- **Mean** is the arithmetic average of a distribution of values.
  - Calculating the mean involves adding, or summing, all of the values in a distribution and dividing by the number of values.



$$\mu = \frac{\Sigma X}{N}$$

or

$$\bar{X} = \frac{\Sigma X}{n}$$

where $\bar{X}$ is the sample mean,
μ is the population mean,
Σ means "the sum of,"
X is an individual score in the distribution,
n is the number of scores in the sample,
N is the number of scores in the population.

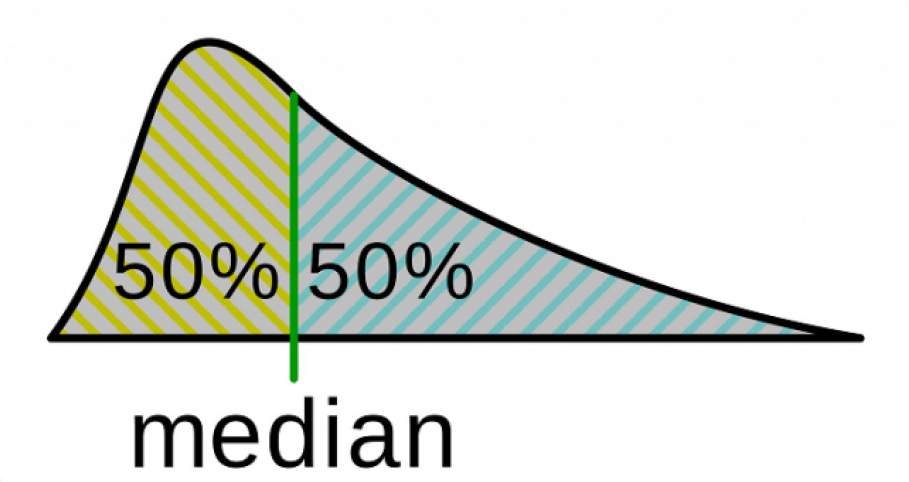
X: 86 90 95 100 100 100 110 110 115 120

$\sum X = 86 + 90 + 95 + 100 + 100 + 100 + 110 + 110 + 115 + 120 = 1026$

1026/10 = 102.6

Median

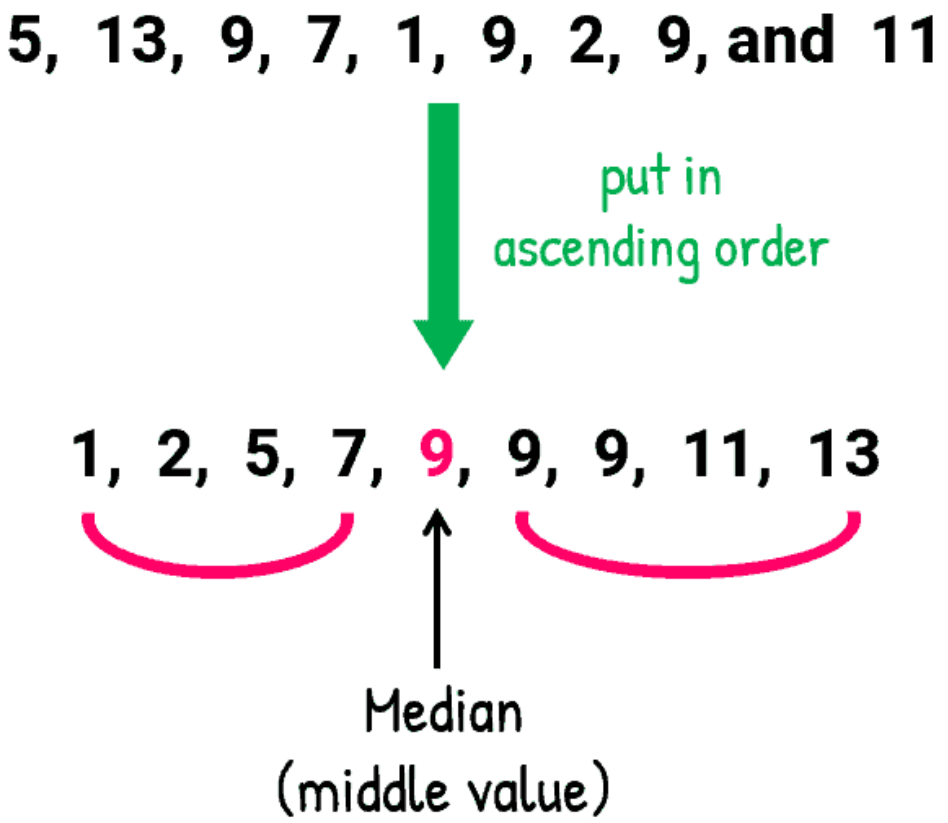
- The median is the score in the distribution that marks the 50th percentile.
  - 50 percent of the scores in the distribution fall above the median and 50 percent fall below it.



X: 86 90 95 100 100 100 110 110 115 120
Because two scores in the middle:
Median = (100 + 100)/2 = 100

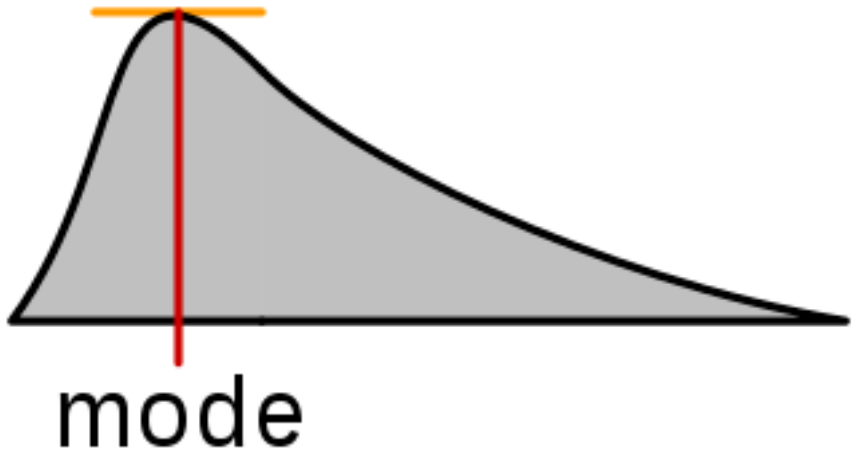
Median

- To find the median (P50 and Mdn) of a distribution:
  - Arrange all of the scores in the distribution in order, from smallest to largest.
  - Find the middle score in the distribution if there is an odd number.
  - If there is an even number of scores in the distribution, the median is the average of the two scores in the middle of the distribution



Mode

- The mode is simply the category in the distribution that has the highest number of scores, or the highest frequency.



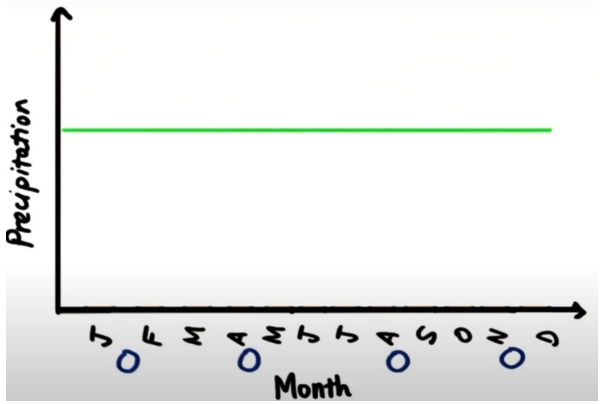
2, 4, 5, 5, 4, 5

→ 2, 4, 4, 5, 5, 5

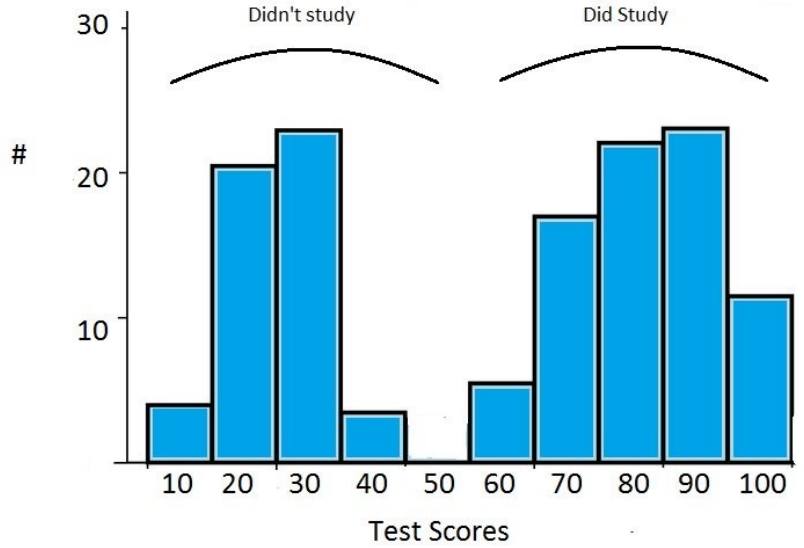
1 2 3

MODE = 5

Mode

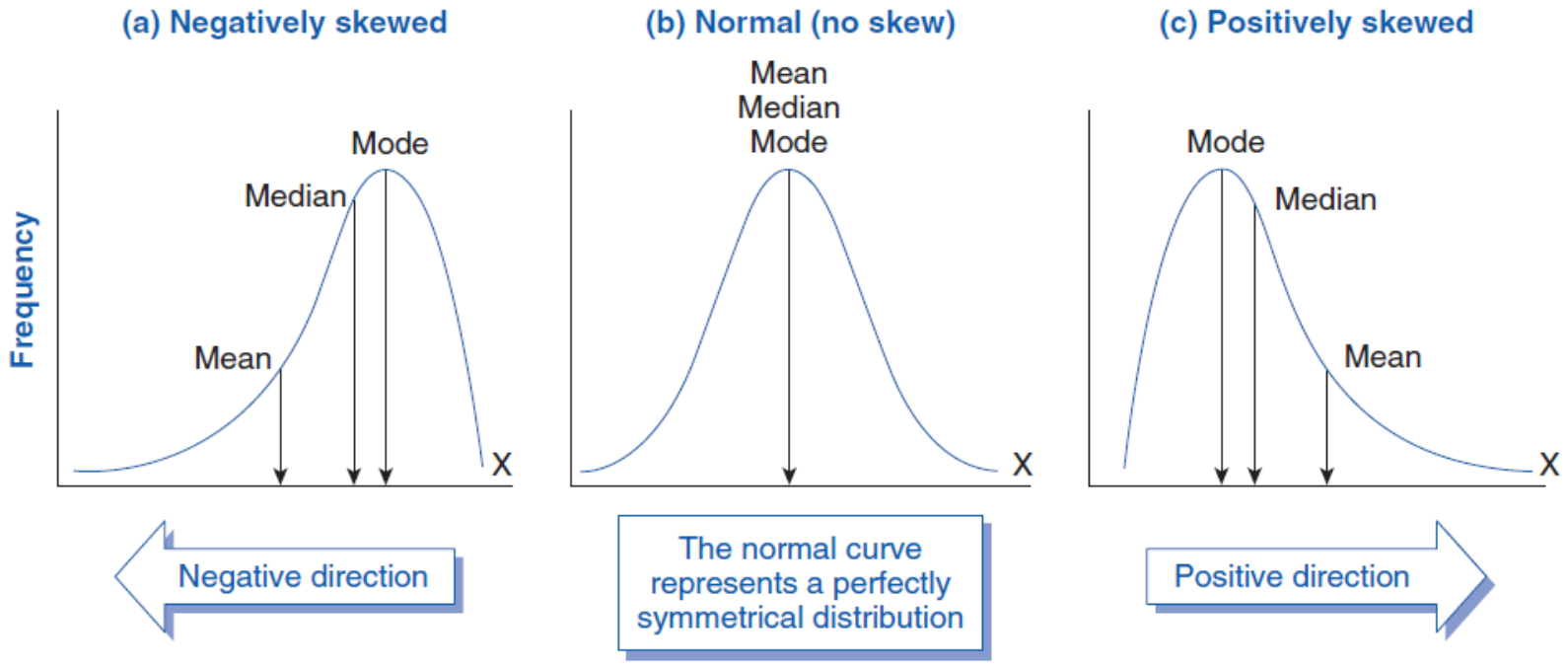


- No Mode
- Single Mode
- Multimodal: if a distribution has more than one category with the most common score, the distribution has multiple modes.
  - BiModal



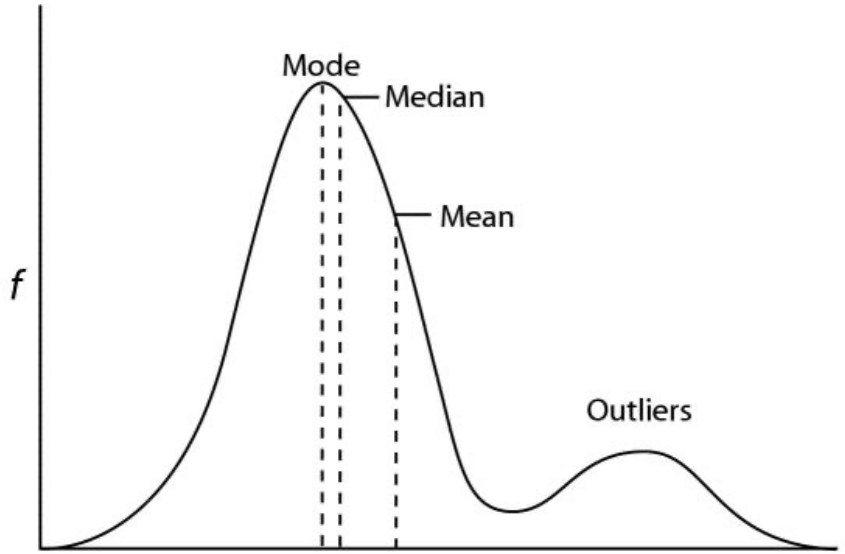
Skewed Distribution

- If one tail is longer than another, the distribution is skewed.
  - These distributions are sometimes called asymmetric or asymmetrical distributions.

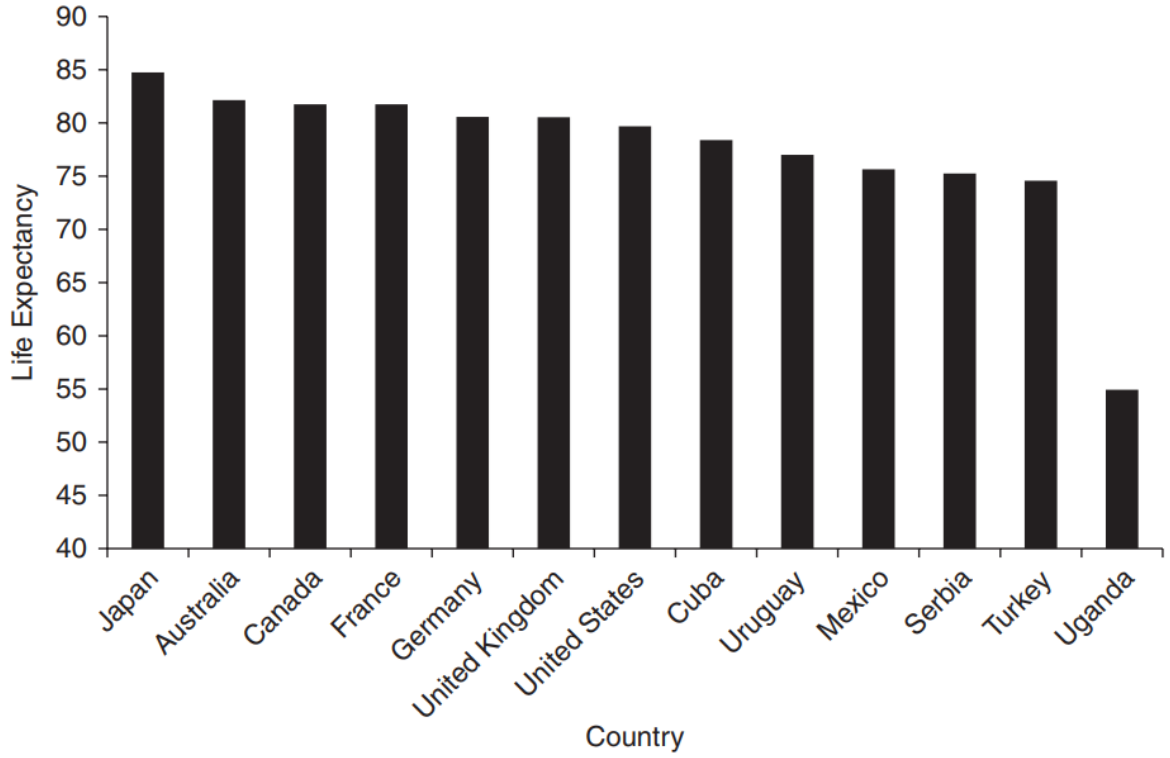


Outliers

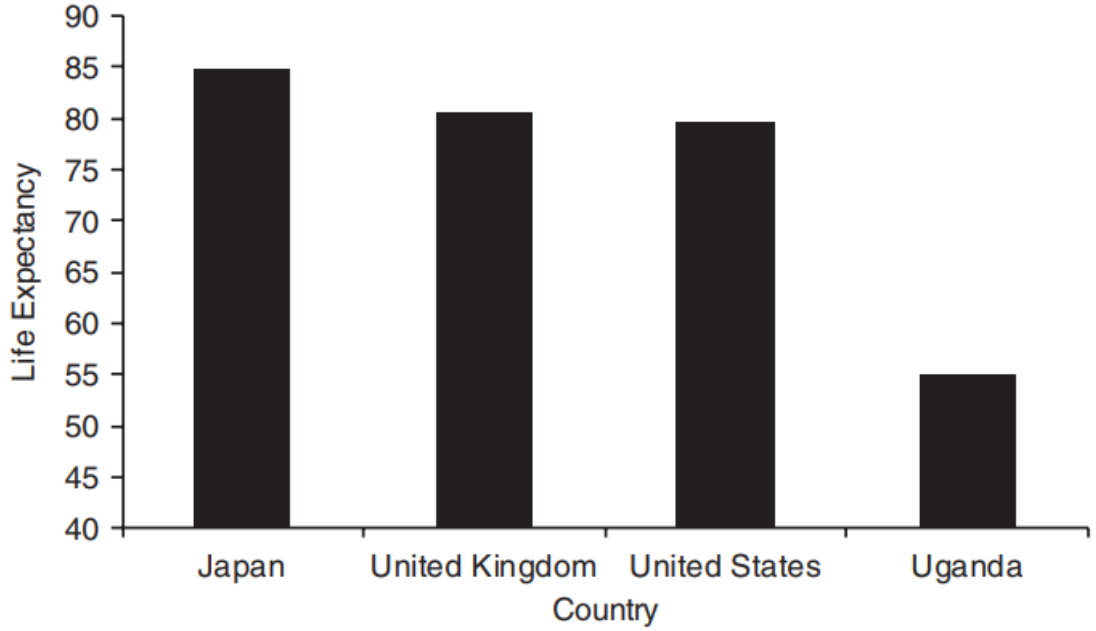
- Distributions have a few scores that are very far away from the mean, they are called outliers.
  - They can have a dramatic effect on the mean of the distribution.
  - When most of the outliers are at one side of a distribution, they pull the mean toward that end of the distribution.



Outliers



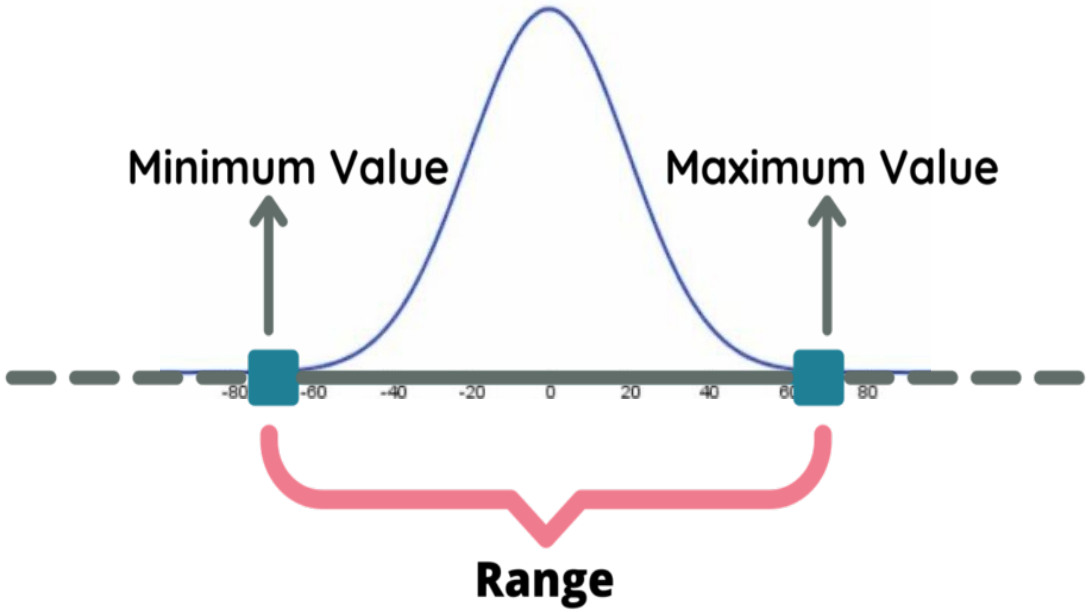
79.34 vs 77.46



81.66 vs 74.98

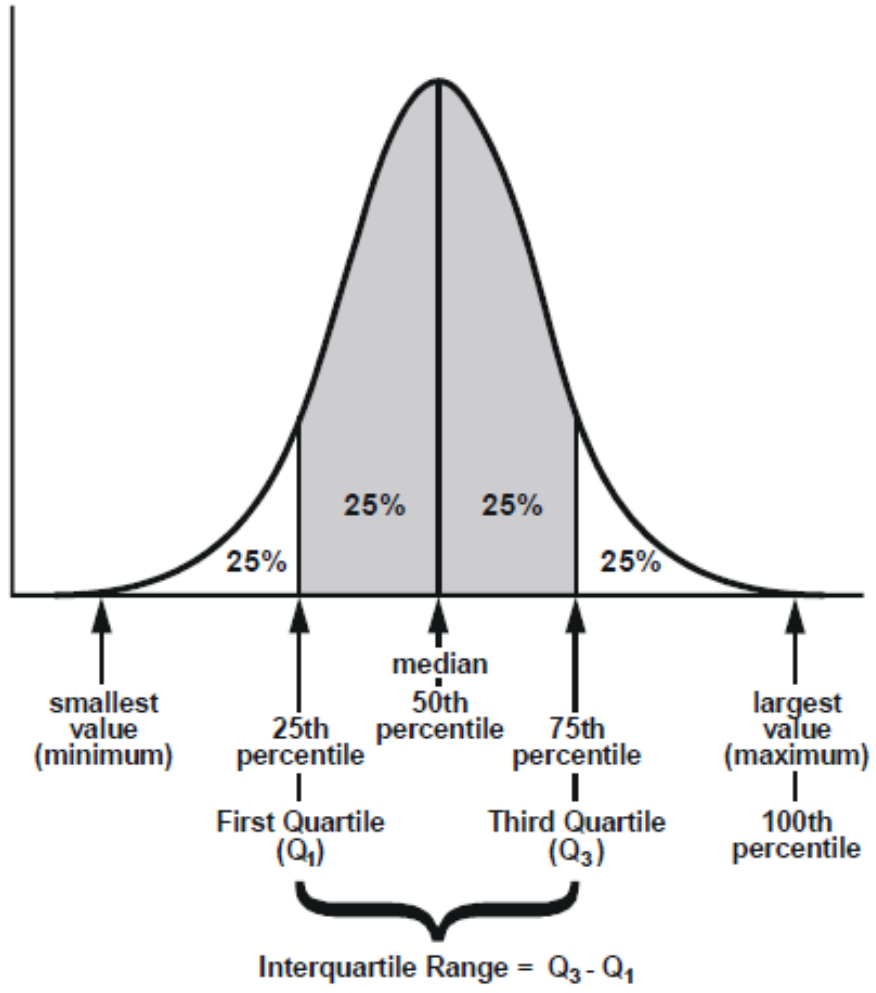
Range

- The range is the difference between the largest score (the maximum value) and the smallest score (the minimum value) of a distribution.



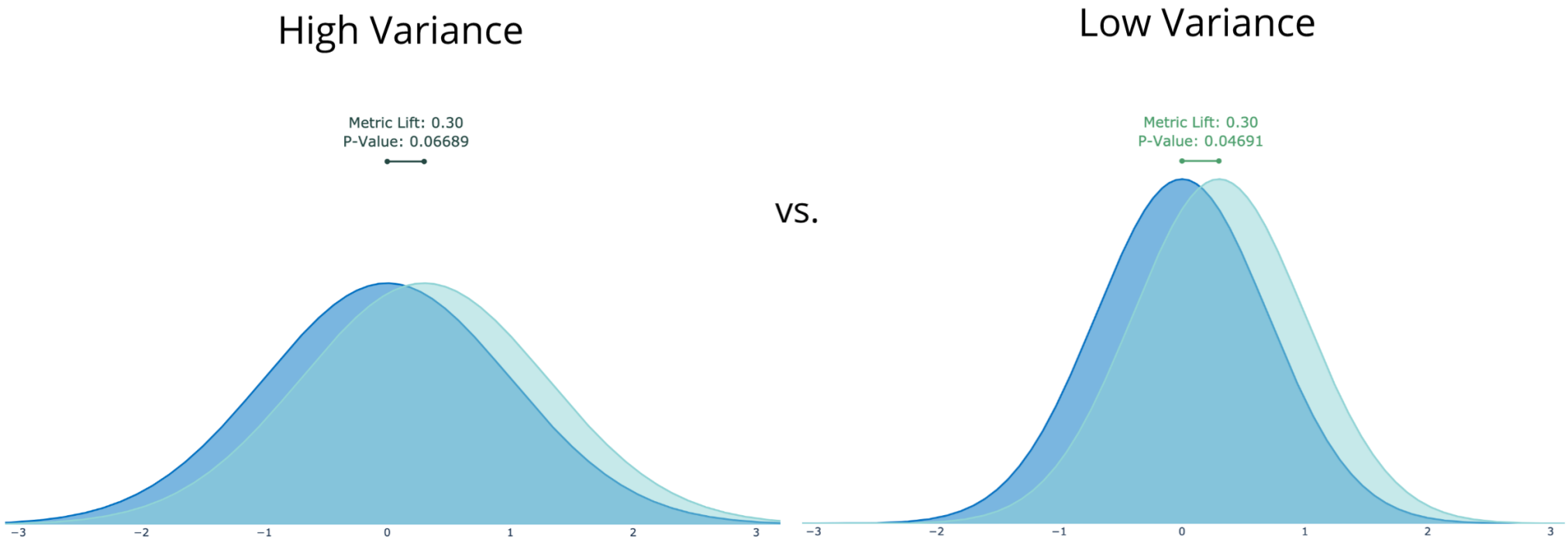
Interquartile Range

- The interquartile range (IQR) is the difference between the score that marks the 75th percentile (the third quartile) and the score that marks the 25th percentile (the first quartile).



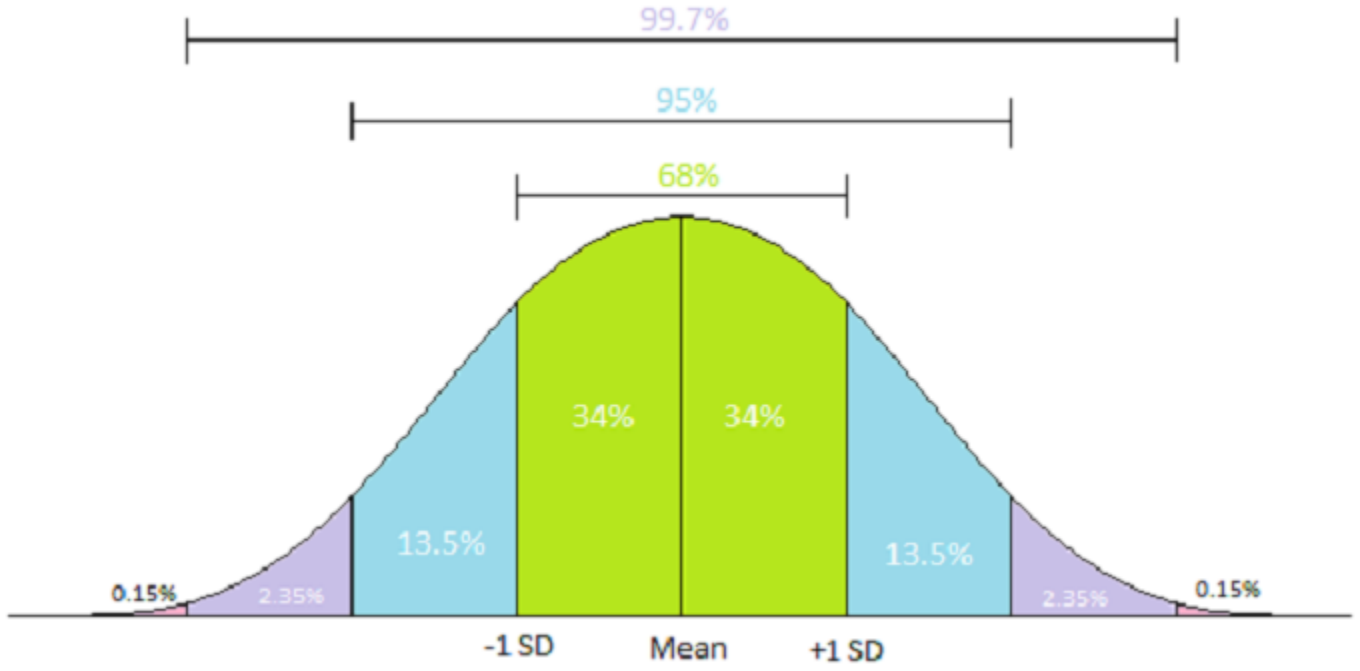
Variance

- The variance provides a statistical average of the amount of dispersion in a distribution of scores.



Standard Deviation

- A standard deviation is the typical, or average, deviation between individual scores in a distribution and the mean for the distribution.



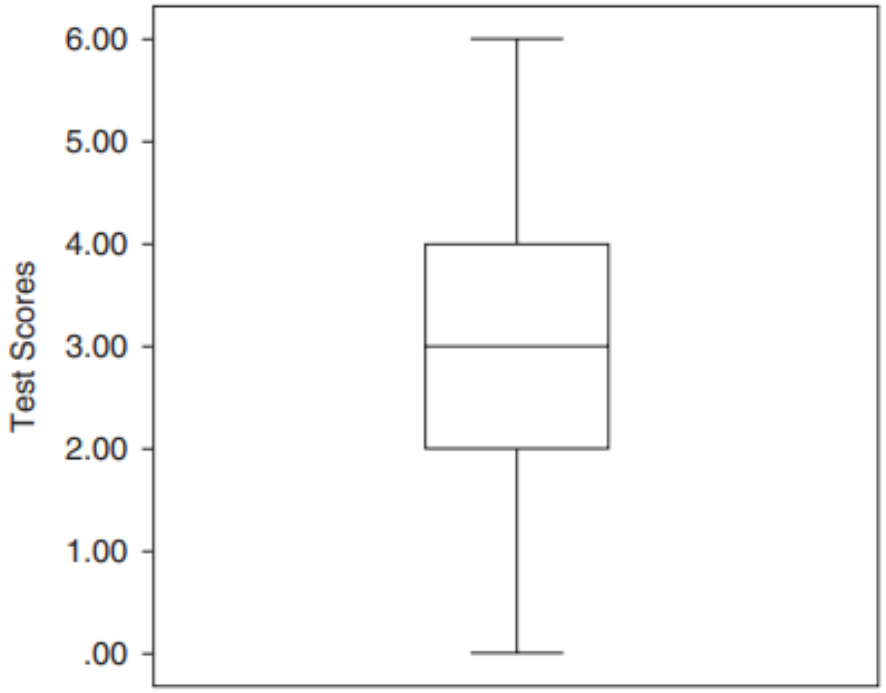
Variance and Standard Deviation in Population and Sample

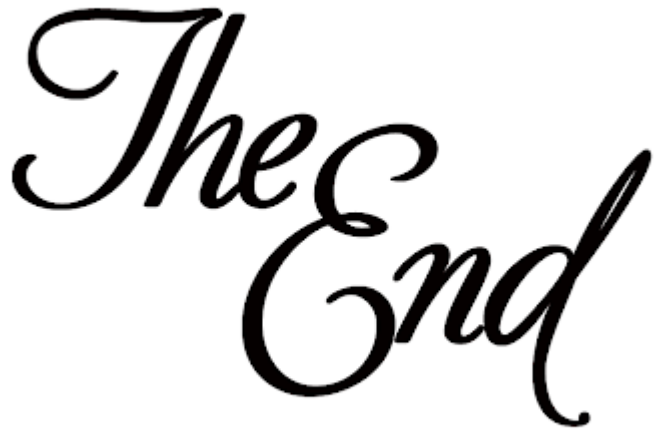
	Population	Estimate Based on a Sample
Variance	$\sigma^2 = \frac{\Sigma(X-\mu)^2}{N}$ where σ^2 is population variance, Σ is to sum, X is each score in the distribution, μ is the population mean, N is the number of cases in the population.	$s^2 = \frac{\Sigma(X-\bar{X})^2}{n-1}$ where s^2 is sample variance, Σ is to sum, X is each score in the distribution, $\bar{X}$ is the sample mean, n is the number of cases in the sample.
Standard Deviation	$\sigma = \sqrt{\frac{\Sigma(X-\mu)^2}{N}}$ where σ is population standard deviation, Σ is to sum, X is each score in the distribution, μ is the population mean, N is the number of cases in the population.	$s = \sqrt{\frac{\Sigma(X-\bar{X})^2}{n-1}}$ where s is sample standard deviation, Σ is to sum, X is each score in the distribution, $\bar{X}$ is the sample mean, n is the number of cases in the sample.

Effect of sample size on Standard Deviation

	N or $n = 500$	N or $n = 10$
Population	$\sigma = \sqrt{\frac{100}{500}} = .44721$	$\sigma = \sqrt{\frac{100}{10}} = 3.16$
Sample	$s = \sqrt{\frac{100}{499}} = .44766$	$s = \sqrt{\frac{100}{9}} = 3.33$

Boxplot




The End