DSA lecture : Probability & Statistics

Presented by: Dr. Sanam Narejo

Computer Systems department , MUET, Jamshoro

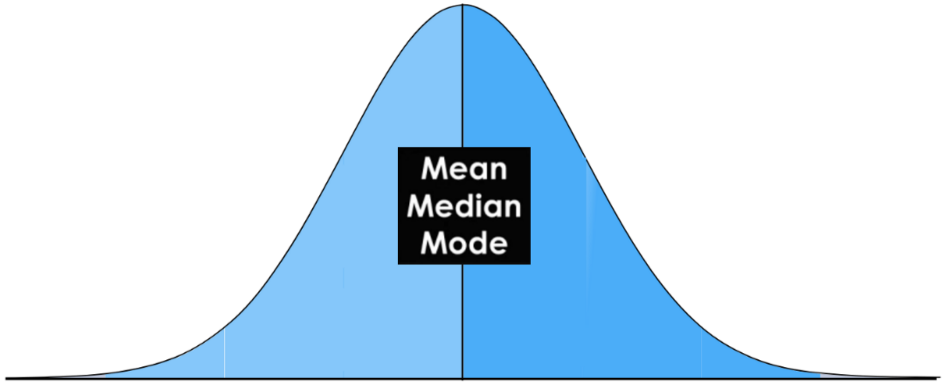
Fair Use Notice

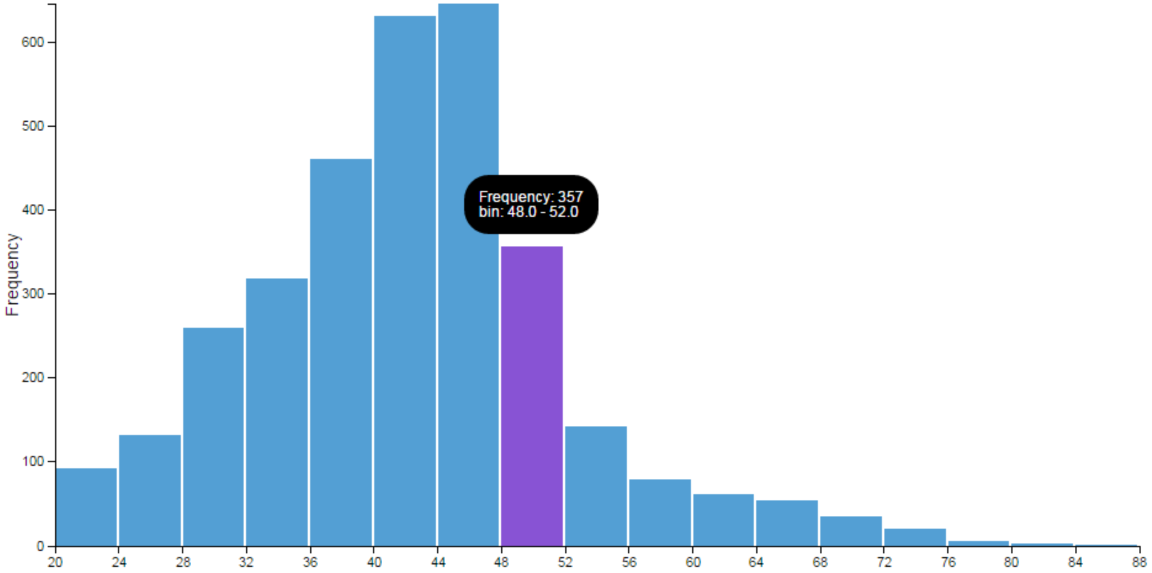
The material used in this presentation i.e., pictures/graphs/text, etc. is solely intended for educational/teaching purpose, offered free of cost to the students for use under special circumstances of Online Education due to COVID-19 Lockdown situation and may include copyrighted material - the use of which may not have been specifically authorised by Copyright Owners. It’s application constitutes Fair Use of any such copyrighted material as provided in globally accepted law of many countries. The contents of presentations are intended only for the attendees of the class being conducted by the presenter.

- Probability and Statistics form the basis of Data Science.

- The probability theory is very much helpful for making the prediction. Estimates and predictions form an important part of Data science. With the help of statistical methods, we make estimates for the further analysis.

- Thus, statistical methods are largely dependent on the theory of probability. And all of probability and statistics is dependent on Data.



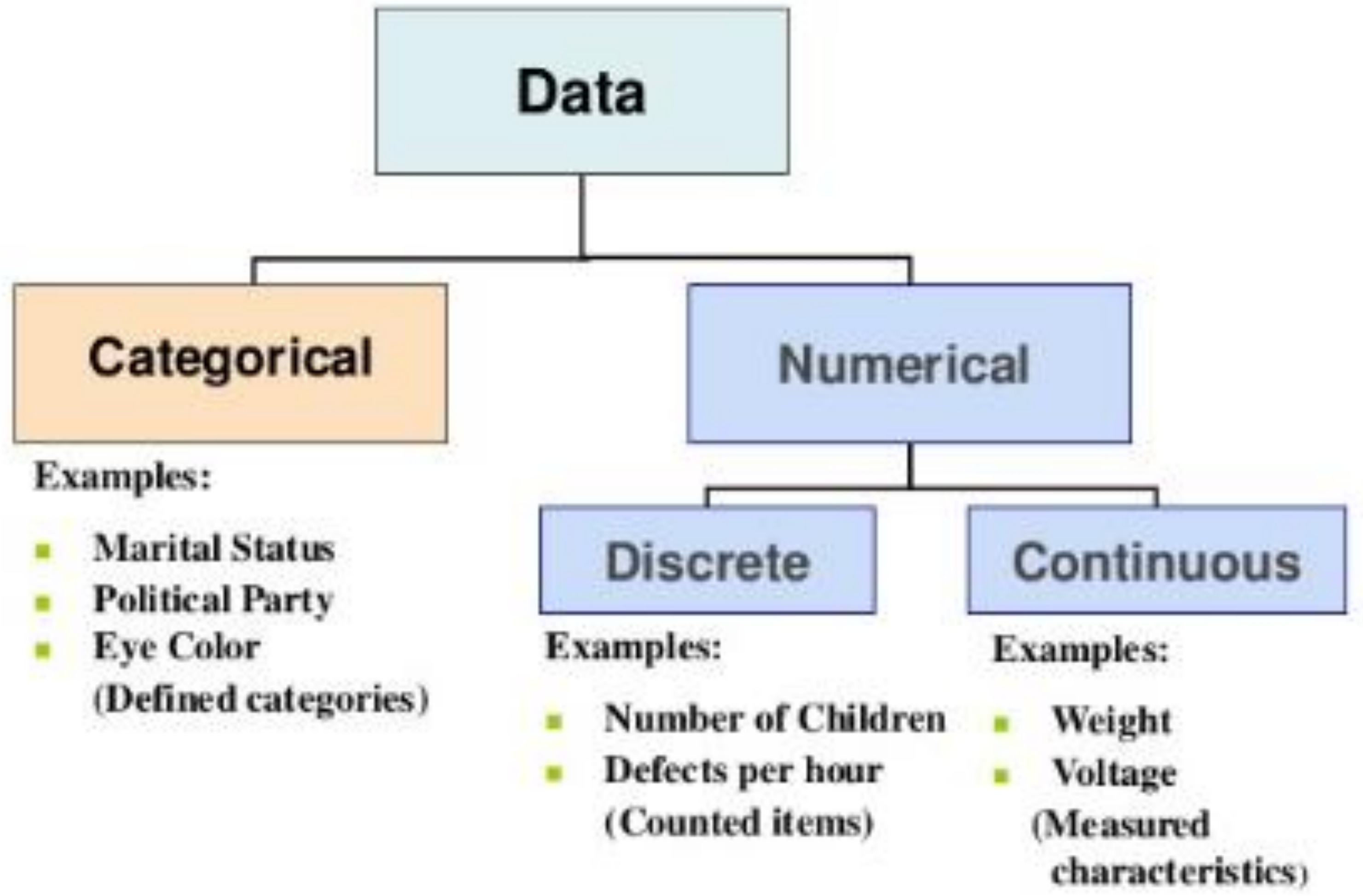


Data

- Data is the collected information(observations) we have about something or facts and statistics collected together for reference or analysis.

- Data — a collection of facts (numbers, words, measurements, observations, etc) that has been translated into a form that computers can process

- **Why does Data Matter?**
- Helps in understanding more about the data by identifying relationships that may exist between 2 variables.
- Helps in predicting the future or forecast based on the previous trend of data.
- Helps in determining patterns that may exist between data.
- Helps in detecting fraud by uncovering anomalies in the data.



Data

Categorical

Examples:

- Marital Status
- Political Party
- Eye Color

(Defined categories)

Numerical

Discrete

Examples:

- Number of Children
- Defects per hour

(Counted items)

Continuous

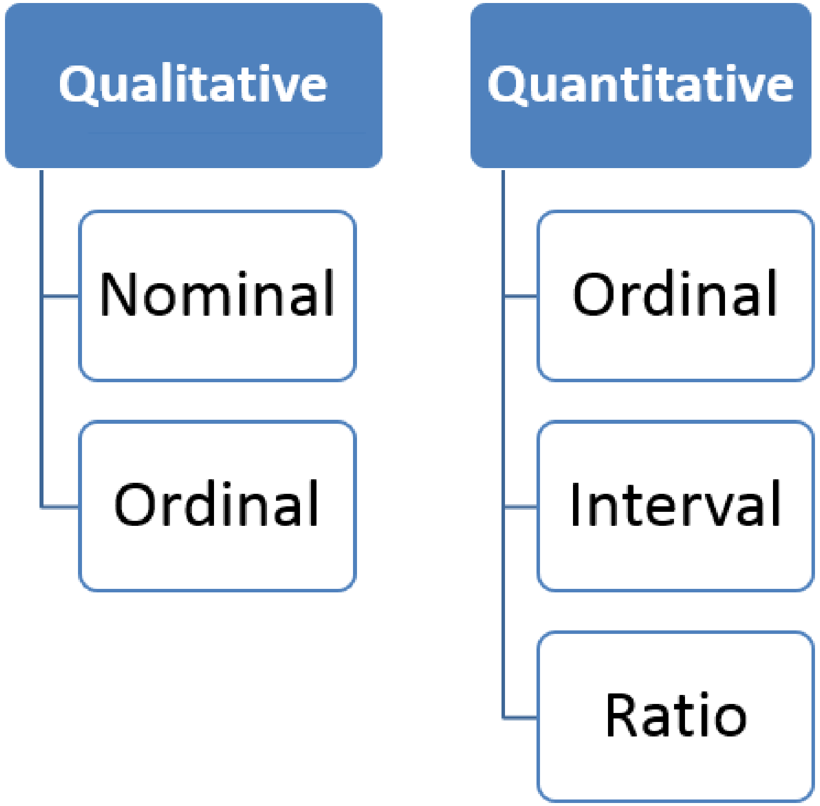
Examples:

- Weight
- Voltage

(Measured characteristics)

- Note: Categorical Data can be visualized by Bar Plot, Pie Chart, Numerical Data can be visualized by Histogram, Line Plot, Scatter Plot

Measurement level of Data



Qualitative

Nominal

Ordinal

Quantitative

Ordinal

Interval

Ratio

The qualitative and quantitative data is very much similar to the categorical and numerical data.

Measurement level of Data

- **Nominal**: Data at this level is categorized using names, labels or qualities. eg: Brand Name, ZipCode, Gender.

- **Ordinal**: Data at this level can be arranged in order or ranked and can be compared. eg: Grades, Star Reviews, Position in Race, Date

- **Interval**: Data at this level can be ordered as it is in a range of values and meaningful differences between the data points can be calculated. eg: Temperature in Celsius, Year of Birth

- **Ratio**: Data at this level is similar to interval level with added property of an inherent zero. Mathematical calculations can be performed on these data points. eg: Height, Age, Weight

QUALITATIVE DATA

A variable that cannot assume a numerical value but can be classified into two or more nonnumeric categories is called a qualitative or categorical variable. The data collected on such a variable are called **qualitative** data.

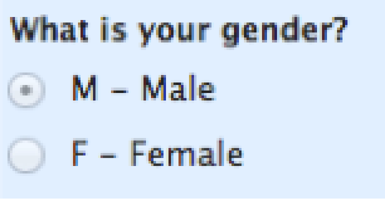
Types

There are **two types** of Qualitative variables:

1. **Nominal Variables** The values are not ordered. Example: Nationality, Gender etc.
2. **Ordinal Variables** - The values are ordered or ranked. Example: Satisfaction score (Not satisfied, Satisfied, Delighted), Spiciness of food (Less spicy mild & Hot

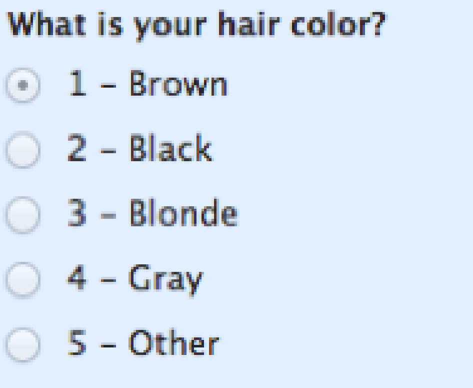
Nominal
Let’s start with the easiest one to understand. Nominal scales are used for ***labeling variables***, without any quantitative value.
“Nominal” scales could simply be called “labels.”

Here are some examples, below. Notice that all of these scales are mutually exclusive (no overlap) and none of them have any numerical significance. A good way to remember all of this is that “nominal” sounds a lot like “name” and nominal scales are kind of like “names” or labels.



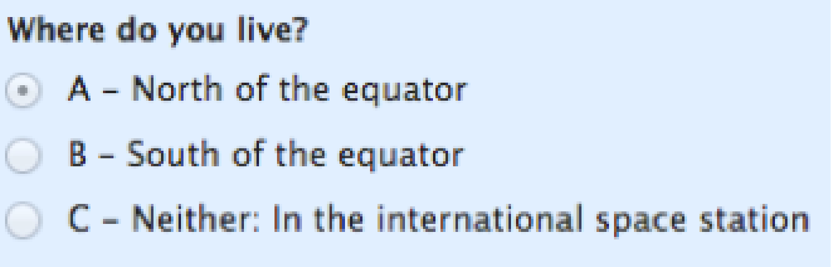
What is your gender?

● M – Male
○ F – Female



What is your hair color?

● 1 – Brown
○ 2 – Black
○ 3 – Blonde
○ 4 – Gray
○ 5 – Other



Where do you live?

● A – North of the equator
○ B – South of the equator
○ C – Neither: In the international space station

At this level, we cannot perform any quantitative mathematical operations, such as addition or division. These would not make any sense.
We can, however, do basic **counts** using pandas’ **value_counts** method.
•Because of our ability to count at the nominal level, graphs, like *bar charts*, *pie charts* are available to us.

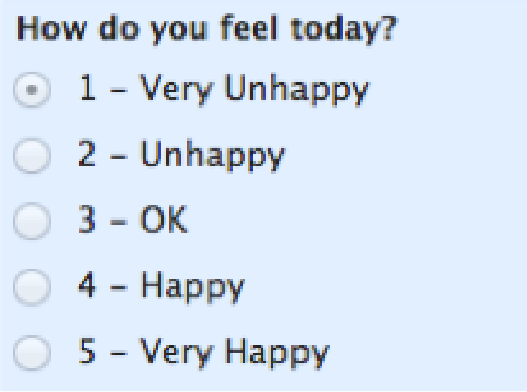
Ordinal
With ordinal scales, the ***order*** of the values is what's important and significant, but the *differences between each one is not really known*.

For example, is the difference between "OK" and "Unhappy" the same as the difference between "Very Happy" and "Happy?" We can't say.

•Ordinal scales are typically measures of non-numeric concepts like satisfaction, happiness, discomfort, etc.

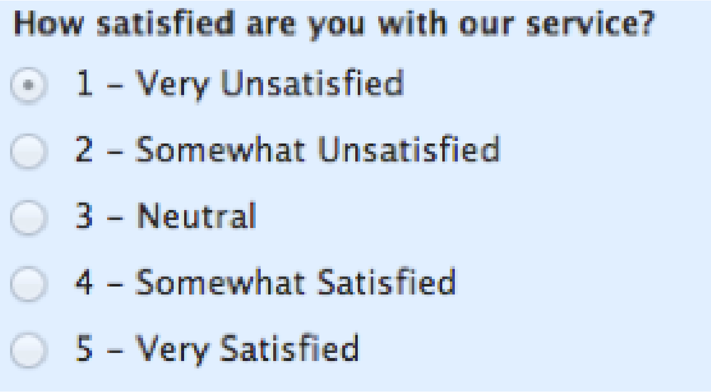
"Ordinal" is easy to remember because is sounds like "order" and that's the key to remember with "ordinal scales"–it is the *order* that matters, but that's all you really get from these.

Advanced note: The best way to determine *central tendency* on a set of ordinal data is to use the **mode** or **median**; a purist will tell you that the mean cannot be defined from an ordinal set.



How do you feel today?

● 1 – Very Unhappy
○ 2 – Unhappy
○ 3 – OK
○ 4 – Happy
○ 5 – Very Happy



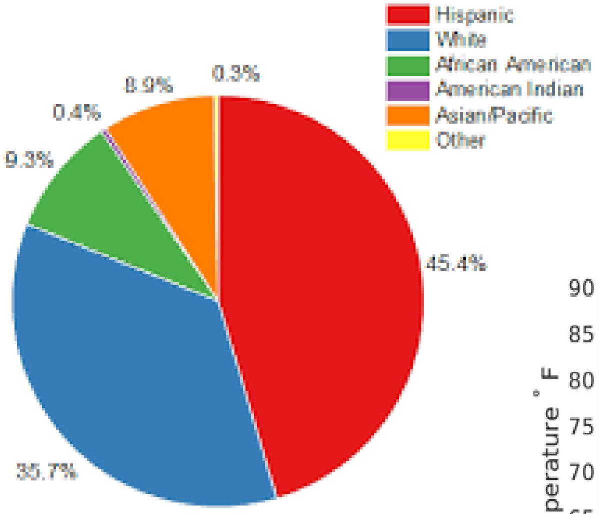
How satisfied are you with our service?

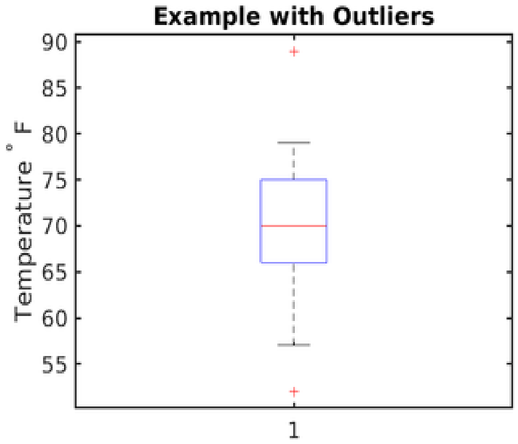
● 1 – Very Unsatisfied
○ 2 – Somewhat Unsatisfied
○ 3 – Neutral
○ 4 – Somewhat Satisfied
○ 5 – Very Satisfied

We can do basic counts as we do with nominal data. Also, for Ordinal data, we can have **comparisons** and **orderings**.

For this reason, we may utilize new graphs at this level. We may use bar and pie charts like we did at the nominal level, bu because we now have ordering and comparisons, *we can calculate **medians** and **percentiles***.

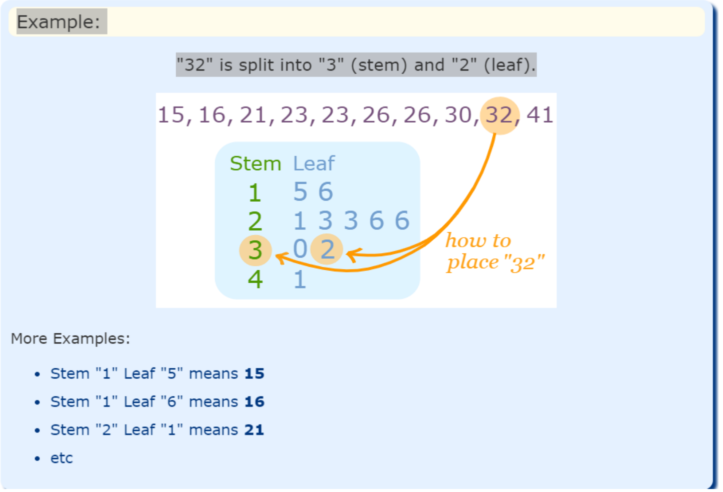
•With medians and percentiles, *stem-and-leaf plots*, as well as *box plots*, are possible.





In descriptive statistics, a box plot or boxplot is a method for graphically depicting groups of numerical data through their quartiles

A Stem and Leaf Plot is a special table where each data value is split into a "stem" (the first digit or digits) and a "leaf" (usually the last digit). Like in this example:



Example:

"32" is split into "3" (stem) and "2" (leaf).

15, 16, 21, 23, 23, 26, 26, 30, 32, 41

Stem	Leaf
1	5 6
2	1 3 3 6 6
3	0 2
4	1

More Examples:

- Stem "1" Leaf "5" means **15**
- Stem "1" Leaf "6" means **16**
- Stem "2" Leaf "1" means **21**
- etc

QUANTITATIVE DATA

A Variable that can be measured numerically is called a quantitative variable.
The data collected on a quantitative variable are called quantitative data.

TYPES

There are **two types** of Quantitative variables:

1. **Discrete Variables** - A variable whose values are countable is called a discrete variable. In other words, a discrete variable can assume only certain values with no intermediate values. Example: Number of heads in 10 tosses etc.

2. **Continuous Variables** - A variable that can assume any numerical value over a certain interval or intervals is called a continuous variable. Example: Height of person etc.

Interval

Interval scales are numeric scales in which we know both the order and the exact *differences between the values*. At Interval level, we will have meaningful differences between values.

•The classic example of an interval scale is Celsius temperature because the difference between each value is the same. For example, the difference between 60 and 50 degrees is a measurable 10 degrees, as is the difference between 80 and 70 degrees.

At the interval level, we have addition and subtraction to work with.

•With the ability to add values together, we may introduce two familiar concepts, the **arithmetic mean** (referred to simply as the mean) and **standard deviation**.

The most common graph to utilize starting at this level would be the **histogram**. This graph is a cousin of the bar graph and visualizes buckets of quantities and shows frequencies of these buckets.

One large advantage of having two or more columns of data at the interval level, is that it opens us up to using **scatter plots** where we can graph two columns of data on our axes and visualize data-points as literal points on the graph

Like the others, you can remember the key points of an “interval scale” pretty easily. “Interval” itself means “space in between,” which is the important thing to remember–interval scales not only tell us about order, but also about the value between each item.

Problem: Here’s the problem with interval scales: they don’t have a “true zero”. For example, there is no such thing as “no temperature,” at least not with Celsius. In the case of interval scales, zero doesn’t mean the absence of value, but is actually another number used on the scale, like 0 degrees Celsius. Negative numbers also have meaning.

Without a true zero, it is impossible to compute ratios. ***With interval data, we can add and subtract, but cannot multiply or divide***.

Confused? Ok, consider this: 10 degreesC + 10 degreesC = 20 degreesC. No problem there. 20 degreesC is not twice as hot as 10 degreesC. When converted to Fahrenheit, it’s clear: 10C=50F and 20C=68F, which is clearly not twice as hot. I hope that makes sense.
— Bottom line, interval scales are great, but we cannot calculate ratios, which brings us to our last measurement scale…

Ratio
Ratio scales are the *ultimate nirvana* when it comes to measurement scales because they tell us about the order, they tell us the exact value between units, AND they also have an **absolute zero**–which allows for a wide range of both descriptive and inferential statistics to be applied.

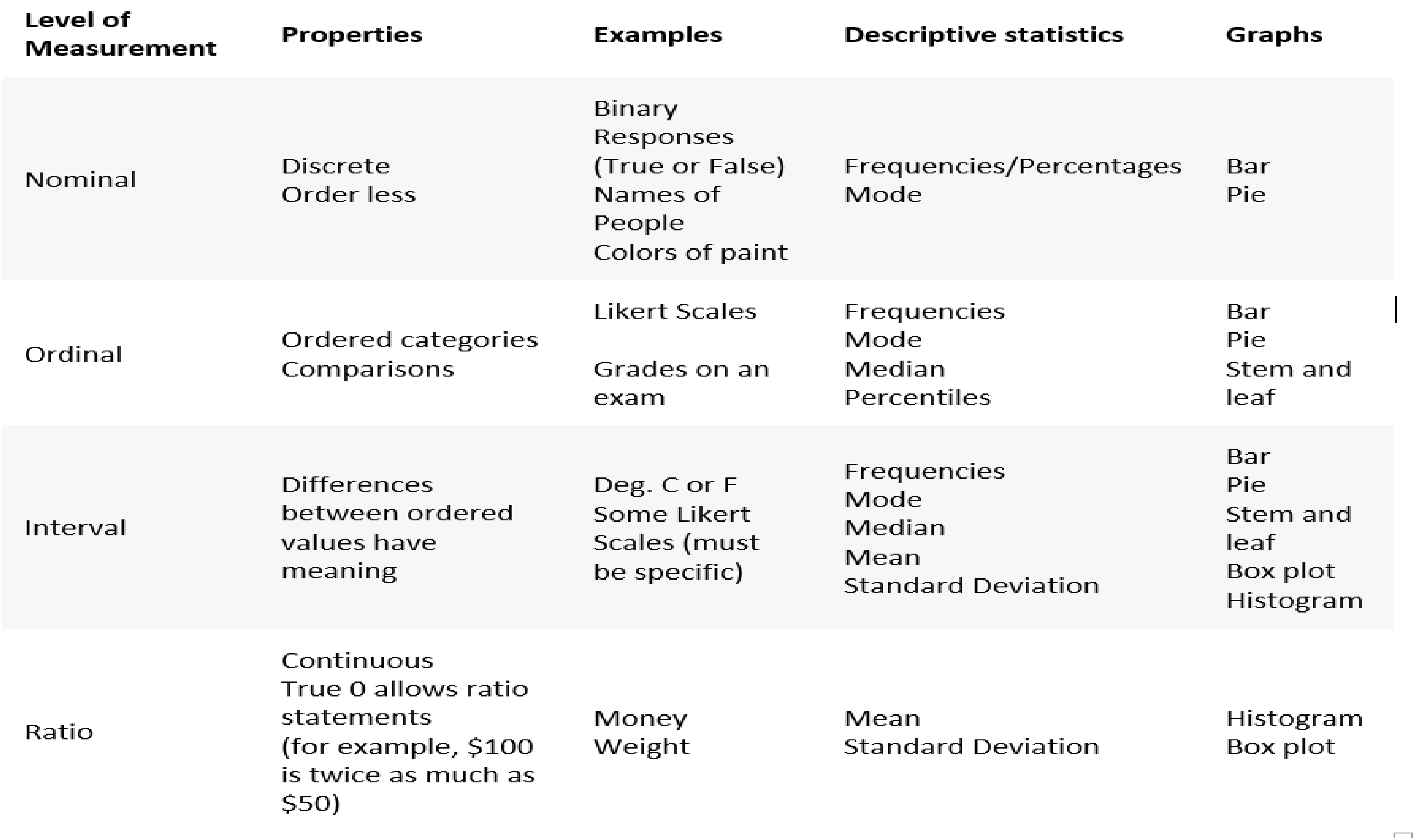
Good examples of ratio variables include height and weight.
Ratio scales provide a wealth of possibilities when it comes to statistical analysis. These variables can be meaningfully added, subtracted, multiplied, divided (ratios).

— Central tendency can be measured by mode, median, or mean; measures of dispersion, such as standard deviation and coefficient of variation can also be calculated from ratio scales.

Summary

In summary, **nominal** variables are used to “*name*,” or label a series of values. **Ordinal** scales provide good information about the *order* of choices, such as in a customer satisfaction survey. **Interval** scales give us the order of values + the ability to quantify *the difference between each one*. Finally, **Ratio** scales give us the ultimate–order, interval values, plus the *ability to calculate ratios* since a “true zero” can be defined.

Provides:	**Nominal**	**Ordinal**	**Interval**	**Ratio**
The “order” of values is known		✔	✔	✔
“Counts,” aka “Frequency of Distribution”	✔	✔	✔	✔
Mode	✔	✔	✔	✔
Median		✔	✔	✔
Mean			✔	✔
Can quantify the difference between each value			✔	✔
Can add or subtract values			✔	✔
Can multiple and divide values				✔
Has “true zero”				✔



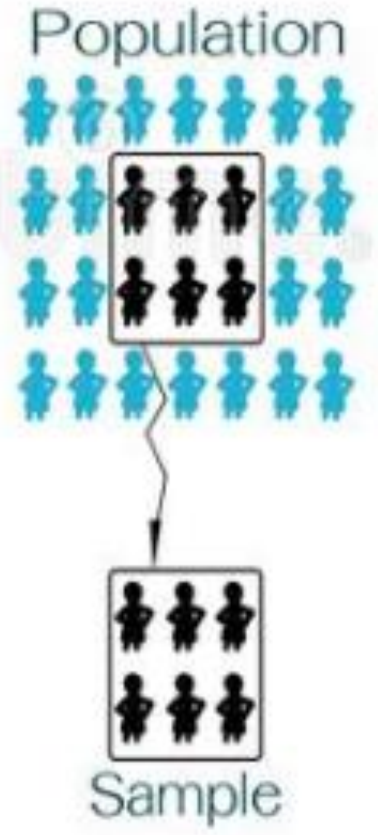
Level of Measurement	Properties	Examples	Descriptive statistics	Graphs
Nominal	Discrete Order less	Binary Responses (True or False) Names of People Colors of paint	Frequencies/Percentages Mode	Bar Pie
Ordinal	Ordered categories Comparisons	Likert Scales Grades on an exam	Frequencies Mode Median Percentiles	Bar Pie Stem and leaf
Interval	Differences between ordered values have meaning	Deg. C or F Some Likert Scales (must be specific)	Frequencies Mode Median Mean Standard Deviation	Bar Pie Stem and leaf Box plot Histogram
Ratio	Continuous True 0 allows ratio statements (for example, $100 is twice as much as $50)	Money Weight	Mean Standard Deviation	Histogram Box plot

Population or Sample Data

Before performing any analysis of data, we should determine if the data we're dealing with is population or sample.

Population: Collection of all items (N) and it includes each and every unit of our study. It is hard to define and the measure of characteristic such as mean, mode is called parameter.

Sample: Subset of the population (n) and it includes only a handful units of the population. It is selected at random and the measure of the characteristic is called as statistics.



For Example, say you want to know the mean income of the subscribers to a movie subscription service(parameter). We draw a random sample of 1000 subscribers and determine that their mean income(x̄) is $34,500 (statistic). We conclude that the population mean income (μ) is likely to be close to $34,500 as well.

Measures of Central Tendency

The measure of central tendency is a single value that attempts to describe a set of data by identifying the central position within that set of data. As such, measures of central tendency are sometimes called measures of central location. They are also classed as summary statistics.

Mean: The mean is equal to the sum of all the values in the data set divided by the number of values in the data set i.e the calculated average.

It susceptible to outliers when unusual values are added it gets skewed i.e deviates from the typical central value.

$$\bar{x} = \frac{(x_1 + x_2 + \cdots + x_n)}{n}$$

Measures of Central Tendency

Median: The median is the middle value for a dataset that has been arranged in order of magnitude.

Median is a better alternative to mean as it is less affected by outliers and skewness of the data.
The median value is much closer than the typical central value.
If the total number of values is odd then

$$\text{Median} = \left(\frac{n+1}{2}\right)^{th} \text{term}$$

If the total number of values is even then

$$\text{Median} = \left(\frac{\left(\frac{n}{2}\right)^{th} term + \left(\frac{n}{2}+1\right)^{th} term}{2}\right)^{\text{th}} \; term$$

Measures of Central Tendency

- **Mode:** The mode is the most commonly occurring value in the dataset. The mode can, therefore sometimes consider the mode as being the most popular option.

- For Example, In a dataset containing {13,35,54,54,55,56,57,67,85,89,96} values.

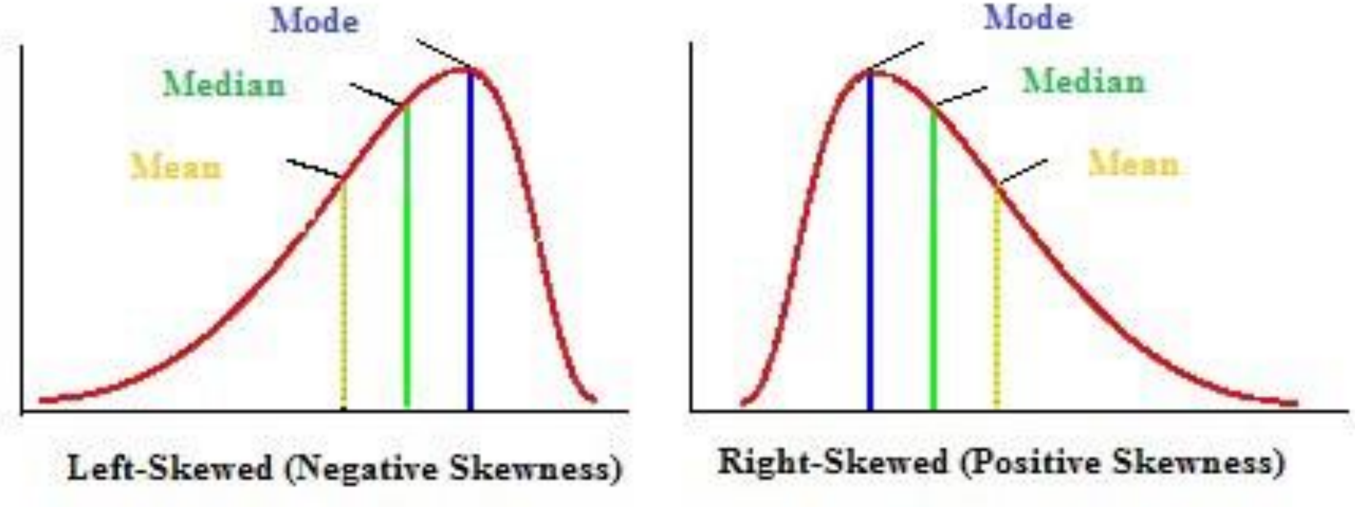
- Mean is 60.09. Median is 56. Mode is 54.

- **Mean** (or average) and **median** are statistical terms that have a somewhat similar role in terms of understanding the **central tendency** of a set of statistical scores. While an average has traditionally been a popular measure of a mid-point in a sample, it has the disadvantage of being affected by any single value being too high or too low compared to the rest of the sample. This is why a median is sometimes taken as a better measure of a mid point.

	Mean	Median
Definition	The mean is the arithmetic average of a set of numbers, or distribution. It is the most commonly used measure of central tendency of a set of numbers.	The median is described as the numeric value separating the higher half of a sample, a population, or a probability distribution, from the lower half.
Applicability	The mean is used for normal distributions.	The median is generally used for skewed distributions.
Relevance to the data set	The mean is not a robust tool since it is largely influenced by outliers.	The median is better suited for skewed distributions to derive at central tendency since it is much more robust and sensible.
How to calculate	A mean is computed by adding up all the values and dividing that score by the number of values.	The Median is the number found at the exact middle of the set of values. A median can be computed by listing all numbers in ascending order and then locating the number in the centre of that distribution.

Measures of Asymmetry

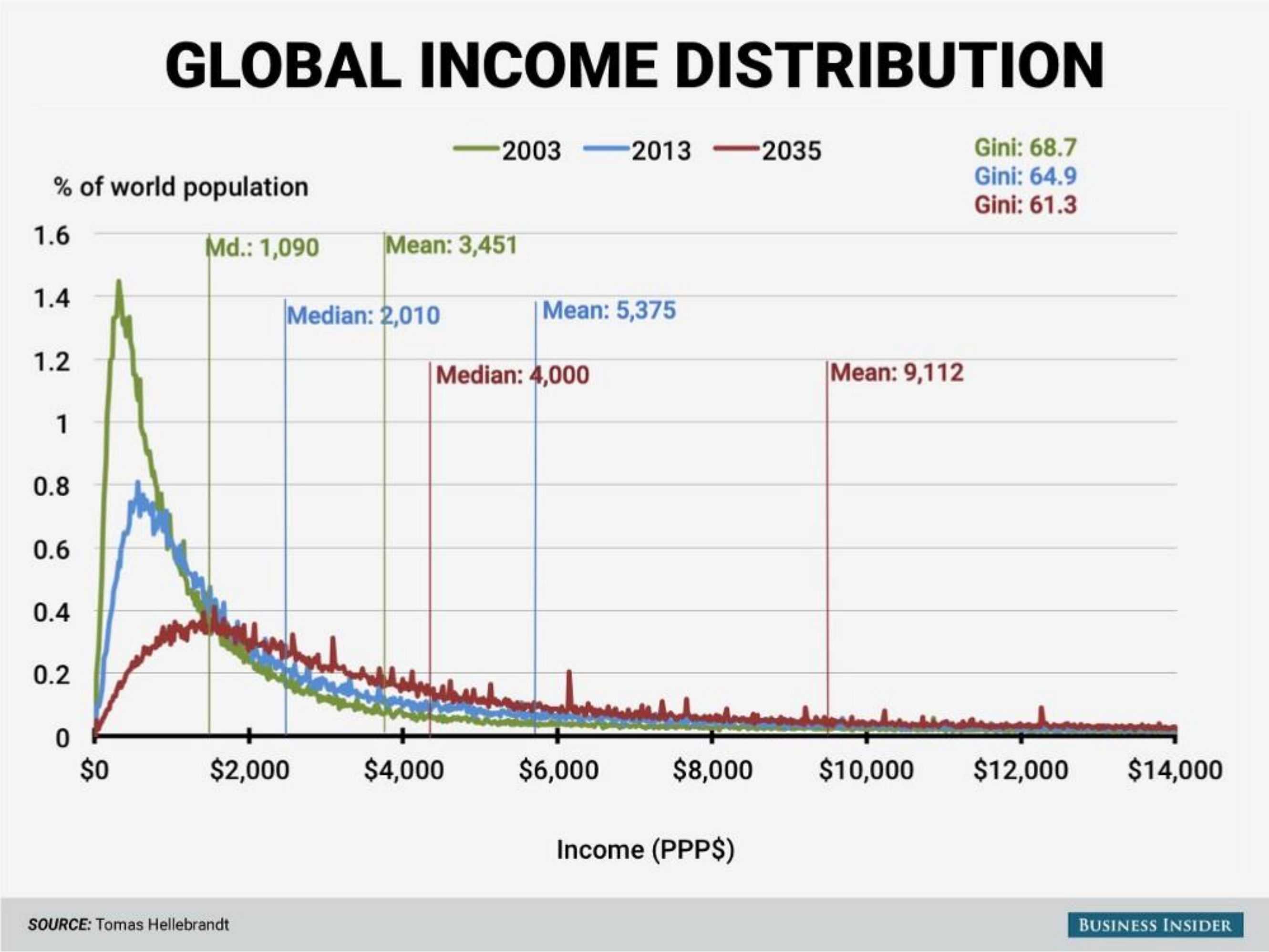
Skewness: Skewness is the asymmetry in a statistical distribution, in which the curve appears distorted or skewed towards to the left or to the right. Skewness indicates whether the data is concentrated on one side.


Mode
Median
Mean
Left-Skewed (Negative Skewness)
Mode
Median
Mean
Right-Skewed (Positive Skewness)

Positive Skewness: Positive Skewness is when the mean>median>mode. The outliers are skewed to the right i.e the tail is skewed to the right.
Negative Skewness: Negative Skewness is when the mean<median<mode. The outliers are skewed to the left i.e the tail is skewed to the left.
Skewness is important as it tells us about where the data is distributed.

For eg: Global Income Distribution in 2003 is highly right-skewed.We can see the mean $3,451 in 2003(green) is greater than the median $1,090.

It suggests that the global income is not evenly distributed. Most individuals incomes are less than $2,000 and less number of people with income above $14,000, so the skewness. But it seems in 2035 according to the forecast income inequality will decrease over time



SOURCE: Tomas Hellebrandt

Measures of Variability(Dispersion)

- The measure of central tendency gives a single value that represents the whole value; however, the central tendency cannot describe the observation fully. The measure of dispersion helps us to study the variability of the items i.e the spread of data.
- *Remember: Population Data has N data points and Sample Data has (n-1) data points. (n-1) is called Bessel's Correction and it is used to reduce bias.*
- **Range**: The difference between the largest and the smallest value of a data, is termed as the range of the distribution. Range does not consider all the values of a series, i.e. it takes only the extreme items and middle items are not considered significant. eg: For {13,33,45,67,70} the range is 57 i.e(70–13).

Measures of Variability(Dispersion)

- **Variance:** Variance measures how far is the sum of squared distances from each point to the mean i.e the dispersion around the mean.

Variance is the average of all squared deviations.

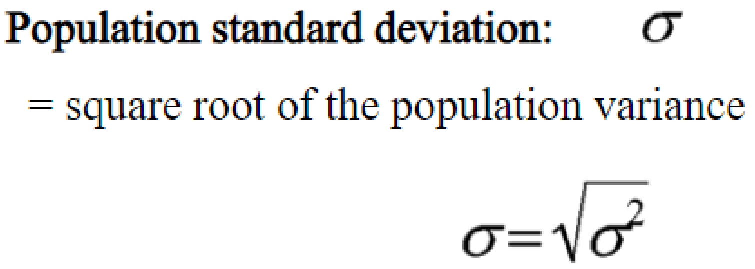
$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{n} \text{ for populations}$$

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} \text{ for samples}$$

Note: *The units of values and variance is not equal so we use another variability measure*

Measures of Variability(Dispersion)

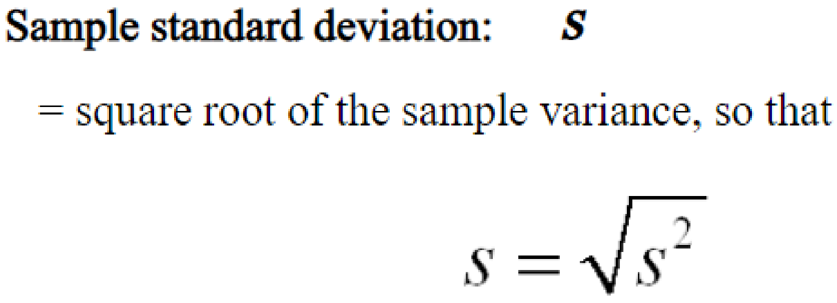
- **Standard Deviation:** As Variance suffers from unit difference so standard deviation is used. The square root of the variance is the standard deviation. It tells about the concentration of the data around the mean of the data set.



Population standard deviation: σ

= square root of the population variance

$$\sigma=\sqrt{\sigma^2}$$



Sample standard deviation: $\boldsymbol{s}$

= square root of the sample variance, so that

$$s=\sqrt{s^2}$$

For eg: {3,5,6,9,10} are the values in a dataset

$$\text{Mean} = \frac{3+5+6+9+10}{5} = 6.6$$

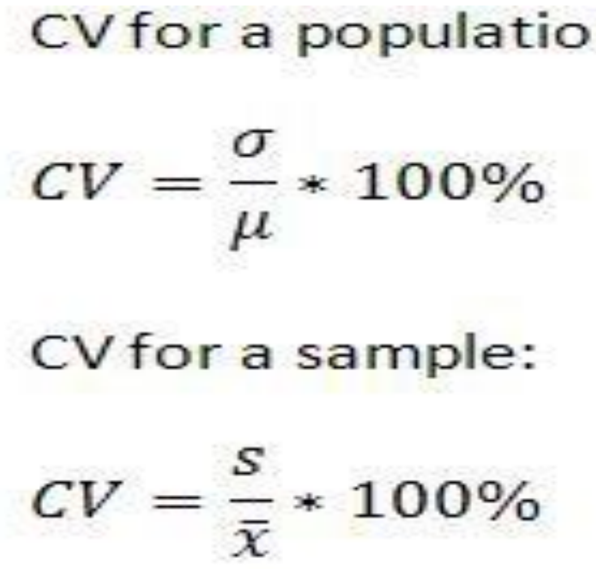
$$\text{Variance} = \frac{(3-6.6)^2 + (5-6.6)^2 + (6-6.6)^2 + (9-6.6)^2 + (10-6.6)^2}{5}$$

$$= \frac{12.96 + 2.56 + 0.36 + 5.76 + 11.56}{5} = \frac{33.2}{5} = 6.64$$

$$\text{Standard Deviation} = \sqrt{\text{Variance}} = \sqrt{6.64} = 2.576$$

Coefficient of Variation(CV): It is also called as the relative standard deviation. It is the ratio of standard deviation to the mean of the dataset.

Standard deviation is the variability of a single dataset. Whereas the coefficient of variance can be used for comparing 2 datasets



CV for a population:

$$CV = \frac{\sigma}{\mu} * 100\%$$

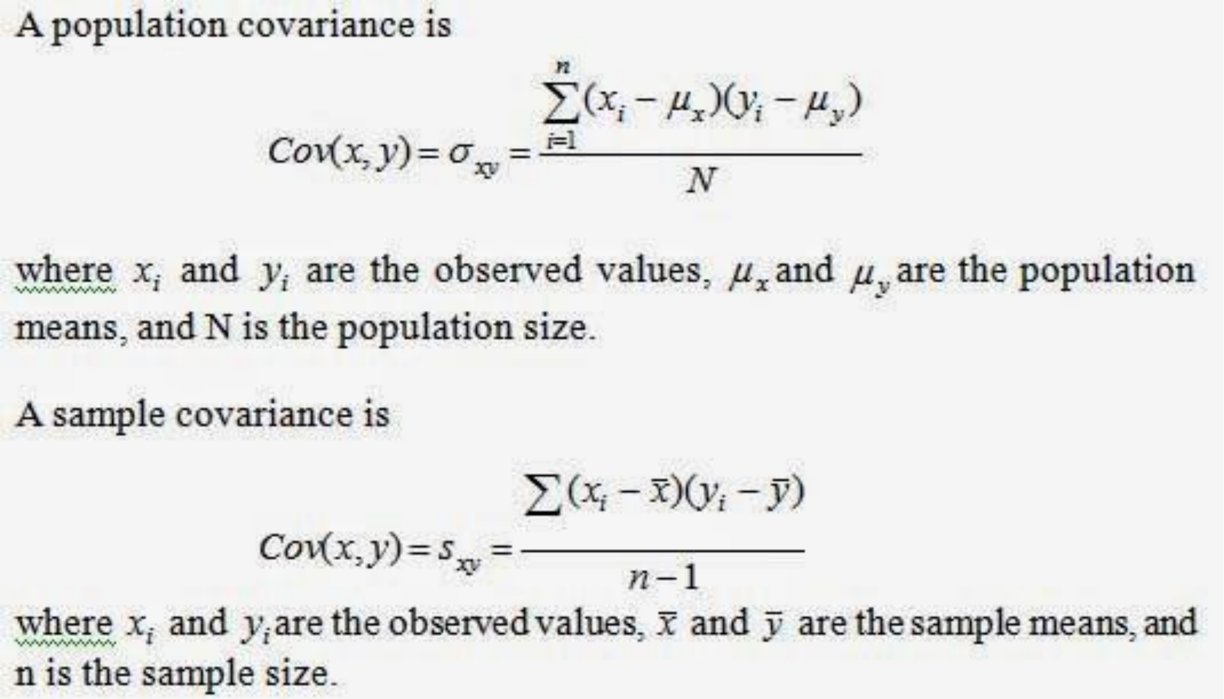
CV for a sample:

$$CV = \frac{s}{\bar{x}} * 100\%$$

Measures of Quartiles

- Quartiles are better at understanding as every data point considered. Measures of Relationship. Measures of relationship are used to find the comparison between 2 variables.
- **Covariance:** Covariance is a measure of the relationship between the variability of 2 variables i.e It measures the degree of change in the variables, when one variable changes, will there be the same/a similar change in the other variable.

Covariance does not give effective information about the relation between 2 variables as it is not normalized.



A population covariance is

$$Cov(x,y) = \sigma_{xy} = \frac{\sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y)}{N}$$

where x_i and y_i are the observed values, μ_x and μ_y are the population means, and N is the population size.

A sample covariance is

$$Cov(x,y) = s_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

where x_i and y_i are the observed values, $\bar{x}$ and $\bar{y}$ are the sample means, and n is the sample size.

- **Correlation:** Correlation gives a better understanding of covariance. It is normalized covariance. Correlation tells us how correlated the variables are to each other. It is also called as Pearson Correlation Coefficient.

$$Correlation = \rho = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y}$$

The value of correlation ranges from -1 to 1. -1 indicates negative correlation i.e with an increase in 1 variable independent there is a decrease in the other dependent variable.1 indicates positive correlation i.e with an increase in 1 variable independent there is an increase in the other dependent variable.0 indicates that the variables are independent of each other.

- In simple words, both the terms measure the **relationship** and the dependency **between** two variables. “**Covariance**” indicates the direction of the linear **relationship between** variables. “**Correlation**” on the other hand measures both the strength and direction of the linear **relationship between** two variables

Home work !!