BASIC STATISTICS FOR DATA ANALYSIS

Why Statistics?

Statistical methods are mainly useful to ensure that your data are interpreted correctly. And that apparent relationships are really “significant” or meaningful and it is not simply happen by chance. Actually, the statistical analysis helps to find meaning to the meaningless numbers.

So, a “statistic” is nothing but some numerical value to that can describe certain property of your data set. There are few well know statistics are the average (or “mean”) value, and the “standard deviation” etc. Standard deviation is the variability within a data set around the mean value. The “variance” is the square of the standard deviation. The linear trend is another example of a data “statistic”.

Steps in the Data Analysis Process

Before staring Data Analysis pipeline you should know there are mainly five steps involved into it.

Step 1: Decide on the objectives or Pose a Question

The first step of the data analysis pipeline is to decide on objectives. These objectives may usually require significant data collection and analysis.

Step 2: What to Measure and How to Measures

Measurement generally refers to the assigning of numbers to indicate different values of variables. Suppose, through your research you are trying to find if there was a relationship between height and weight of human, it would make sense to measure the height and weight of dogs using a scale.

Step 3: Data Collection

Once you know what types of data you need for your statistical study then you can determine whether your data can be gathered from existing sources/databases or not. If data is not sufficient the you have to collect new data. Even if you have existing data, it is very important to know how the data was collected? This will helps you to understand you ca determine the limitations of the generalizability of results and conduct a proper analysis.

The more data you have, the more better correlations, building better models and finding more actionable insights is easy for you. Especially data from more diverse sources helps to do this job easier way.

Step 4: Data Cleaning

This is another crucial step in data analysis pipeline is to improve data quality for your existing data. Too often Data scientists correct spelling mistakes, handle missing values and remove useless information. This is the most critical step because junk data may generate inappropriate results and mislead the business.

Step 5: Summarizing and Visualizing Data

Exploratory data analysis helps to understand the data better. Because a picture is really worth a thousand words as many people understand pictures better than a lecture. Likewise, Measures of Variance indicate the distribution of the data around the center. Correlation refers to the degree to which two variable move in sync with one another.

Step 6: Data Modeling

Now build models that correlate the data with your business outcomes and make recommendations. This is where the unique expertise of data scientists becomes important to business success. Correlating the data and building models that predict business outcomes

Step 7: Optimize and Repeat

The data analysis is a repeatable process and sometime leads to continuous improvements, both to the business and to the data value chain itself.

Now you know steps involved in Data Analysis pipeline. Before advancing to more sophisticated techniques, I suggest starting your data analysis journey with the following statistics fundamentals –

Here is a road map for getting started with Data Analysis. Before starting any statistical data analysis, we need to explore data more and more. To explore data below topics are very useful.

Basic Statistics

- Cases, Variables, Types of Variables
- Matrix and Frequency Table
- Graphs and Shapes of Distributions
- Mode, Median and Mean
- Range, Interquartile Range and Box Plot
- Variance and Standard deviation
- Z-scores
- Contingency Table, Scatterplot, Pearson's r
- Basics of Regression
- Elementary Probability
- Random Variables and Probability Distributions
- Normal Distribution, Binomial Distribution & Poisson Distribution

Inferential Statistics

- Observational Studies and Experiments
- Sample and Population
- Population Distribution, Sample Distribution and Sampling Distribution
- Central Limit Theorem
- Point Estimates
- Confidence Intervals
- Introduction to Hypothesis Testing

EXPLORE YOUR DATA: CASES, VARIABLES, TYPES OF VARIABLES

A data set contains informations about a sample. A Dataset consists of **cases.** Cases are nothing but the objects in the collection. Each case has one or more attributes or qualities, called **variables** which are characteristics of cases.

Example:

Suppose you are collecting information about breast cancer patients. Now for each and every cancer patient you want to know the below information

1. Sample code number: id number
2. Clump Thickness: 1 – 10
3. Uniformity of Cell Size: 1 – 10
4. Uniformity of Cell Shape: 1 – 10
5. Marginal Adhesion: 1 – 10
6. Single Epithelial Cell Size: 1 – 10
7. Bare Nuclei: 1 – 10
8. Bland Chromatin: 1 – 10
9. Normal Nucleoli: 1 – 10
10. Mitoses: 1 – 10
11. Class: (2 for benign, 4 for malignant)

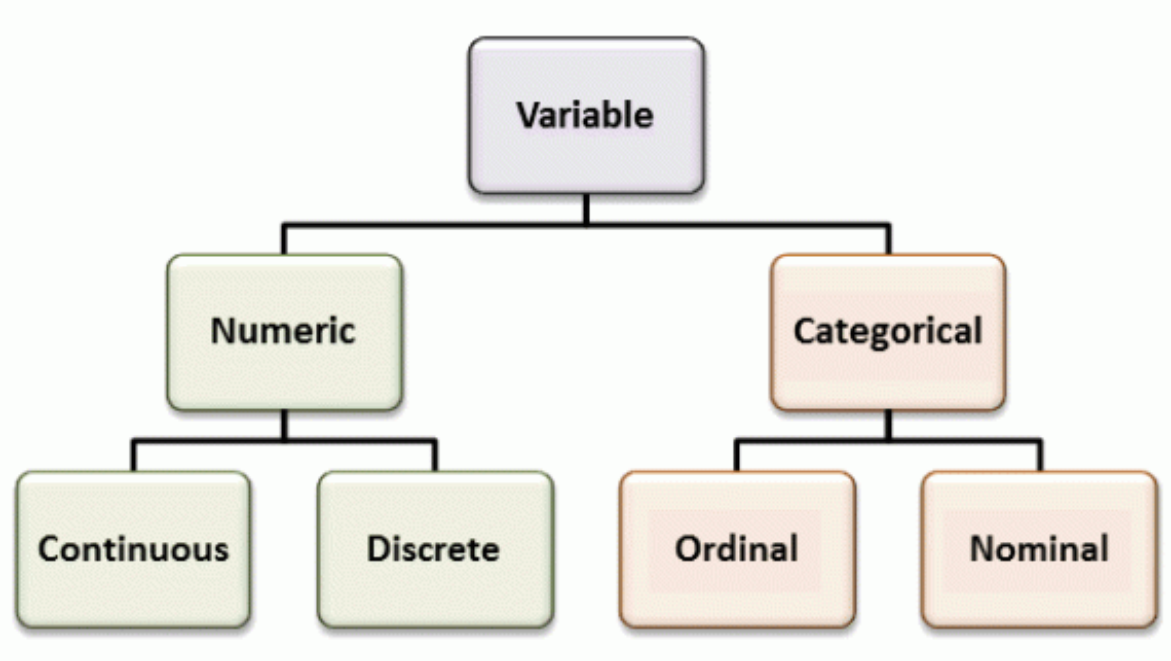
These features were taken from UCI Breast Cancer Dataset. You can find it here

https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(original)

For this example, breast cancer patients themselves are **cases** and all these characteristics of the patients are **variables.**

In a study, cases can be many different things. They can be individual patients and group of patients. But they can also be, for instance, companies, schools or countries etc.

we can have many, different kinds of variables, representing different characteristics. Because of this reason there are various level of measurements or different types of variables.



Categorical Variables:

Both nominal and ordinal variables can be called categorical variables.

1. Nominal Variable:

A nominal variable is made up of various categories which has no order.

Example:

Gender of a patient may be Male or Female or State where they live in. Here each category differs from each other but there is no ranking order.

2. Ordinal Variable:

The second level of measurement is the ordinal level. There is not only a difference between the categories of a variable; there is also an order. An example might be Highest paid, Average Paid and Lowest Paid employee.

Quantitative/ Numerical Variables:

1. Continuous Variable:

A variable is continuous if the possible values of the variable form an interval. An example is, again, the height of a patient. Someone can be 172 centimeters tall and 174 centimeters tall. But also, for instance, 170.2461. We don't have a set of separate numbers, but an infinite region of values.

2. Discrete Variable:

A variable is discrete if its possible categories form a set of separate numbers.

For the above breast cancer data **Uniformity of Cell Size: 1 – 10** is an example of discrete variable.

DATA MATRIX AND FREQUENCY TABLE

If you're conducting a study, you should think about your data in terms of cases and variables.

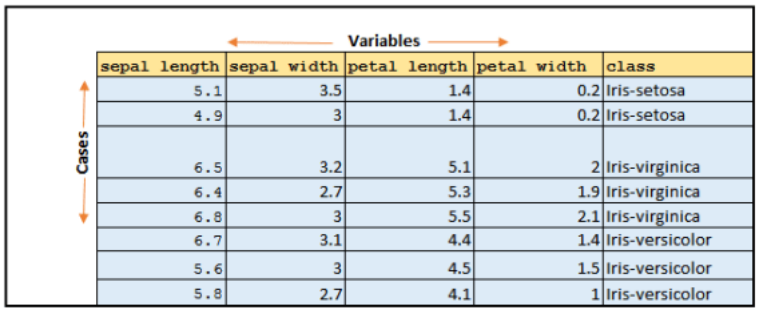
Cases are the persons, animals or things in your study, and variables are the characteristics of interest. Here, I will discuss how you can order and present your cases and variables. Lets take an example, imagine you are interested in the "Primera División", the top football competition in Spain. Here, the cases you're interested in are individual football players within the league, and the variables you focus on are age, body weight, goals scored, team membership and hair color. The best way to order all this information is by means of a data matrix.

So, Data Matrix is the tabular format representation of cases and variables of your statistical study. Each row of a data matrix represents a case and each column represent a variable.

A complete Data Matrix may contain thousands or lakhs or even more cases.

Sample from IRIS Dataset has shown below. You can get it from UCI Repository.

https://archive.ics.uci.edu/ml/datasets/iris



sepal length	sepal width	petal length	petal width	class
5.1	3.5	1.4	0.2	Iris-setosa
4.9	3	1.4	0.2	Iris-setosa
6.5	3.2	5.1	2	Iris-virginica
6.4	2.7	5.3	1.9	Iris-virginica
6.8	3	5.5	2.1	Iris-virginica
6.7	3.1	4.4	1.4	Iris-versicolor
5.6	3	4.5	1.5	Iris-versicolor
5.8	2.7	4.1	1	Iris-versicolor

To get more insight, summarization of the information is very useful. A good way to do that is to make a frequency table. A frequency table shows how the values of a variable are distributed over the cases. Consider this following example to consider that. We can get the frequency of items and then percentage or even calculating cumulative percentage.

Class	Frequency	Percentage	Cumulative Percentage
Iris-setosa	2	25%	25%
Iris-virginica	3	38%	63%
Iris-versicolor	3	38%	100%
Total	8	100%	

Here we have total 8 cases and among 8 cases 2 cases (25 % cases) belongs to Iris-Setosa.

3 cases which means 38% cases belongs to Iris-Virginia and similarly another 38% are Iris Versicolor.

Above example is for a categorical variable called **class.** But think if your variable is **quantitative** then computing percentage for every specific value does not make sense**. In that case first bring your data into some ordinal categories, by using intervals.** Then do the rest of the things.

EXPLORE YOUR DATA: GRAPHS AND SHAPES OF DISTRIBUTIONS

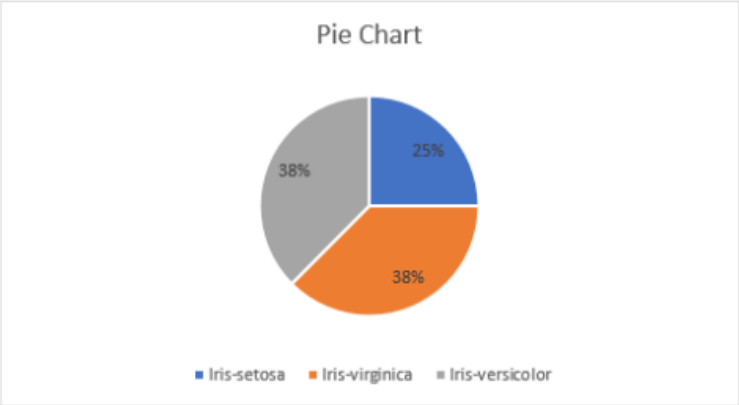
For Categorical Variables:

If variable of interest is categorical then generally **Pie char**t or **Bar Graph** is the best representation.

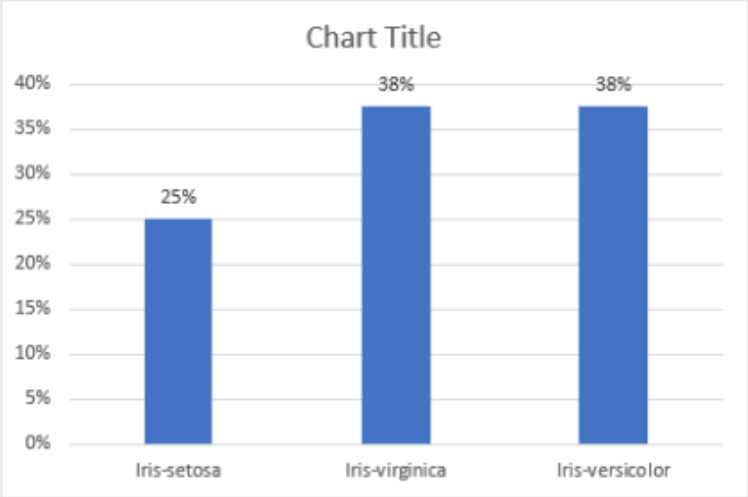
If too many categories are there then pie chart might be messy but Bar Graph will give a clear view**. So, Bar Graph is having advantage over Pie Chart when no of variables are too high.**

Class	Frequency	Percentage	Cumulative Percentage
Iris-setosa	2	25%	25%
Iris-virginica	3	38%	63%
Iris-versicolor	3	38%	100%
Total	8	100%	

For the above table if we draw pie chart for percentage column it will be like this.



For the same table if we draw bar graph for percentage column it will be like this.



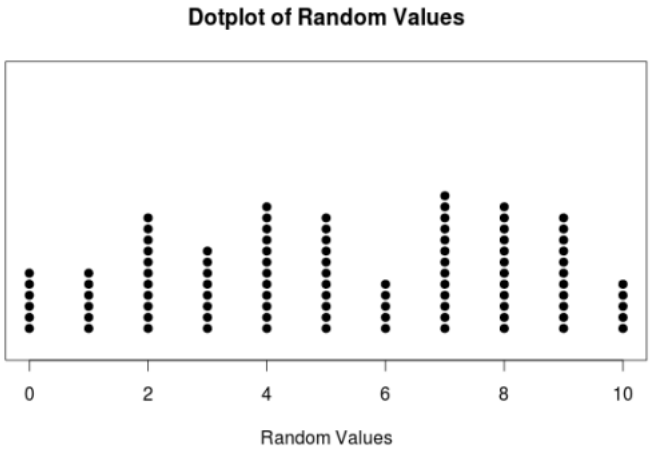
Basically, for categorical variable you can do these many representations listed below.

- Bar plot
- Pie Chart
- Frequency Table
- Contingency Table
- Segmented Bar Plot.
- Relative Frequency
- Mosaic Plot
- Side by Side Box Plot

For Quantitative Variables:

Dot Plot:

If you are working with quantitative variables or numerical variable then Dot Plot is one kind of representation that can be used. A dot plot looks like this. Plot each and every point into the graphs after drawing a horizontal line and label the possible values on it, in regular intervals.

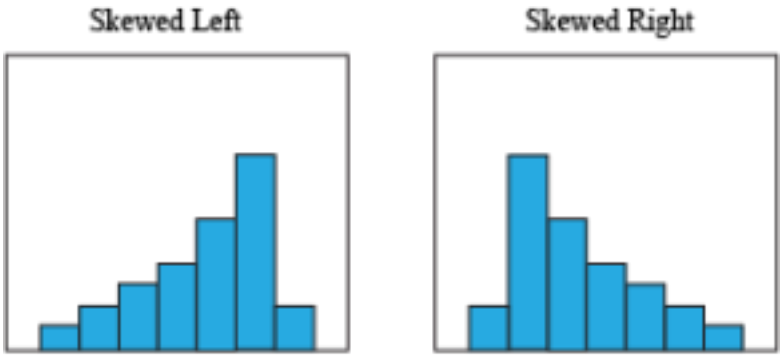


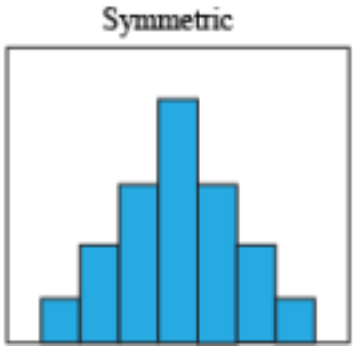
But If you have a very large sample then dot plot may looks messy.

So, another kind of representation called Histogram might be useful for that case.

Histogram:

A histogram is similar to a bar graph in the sense that it uses bars to portray the frequencies or relative frequencies of the possible values of a variable. However, there is one important difference. That difference is that the bars in a histogram touch each other. This touching represents that the values of an interval/ratio variable represent an underlying continuous scale. Below is an example of Histogram.



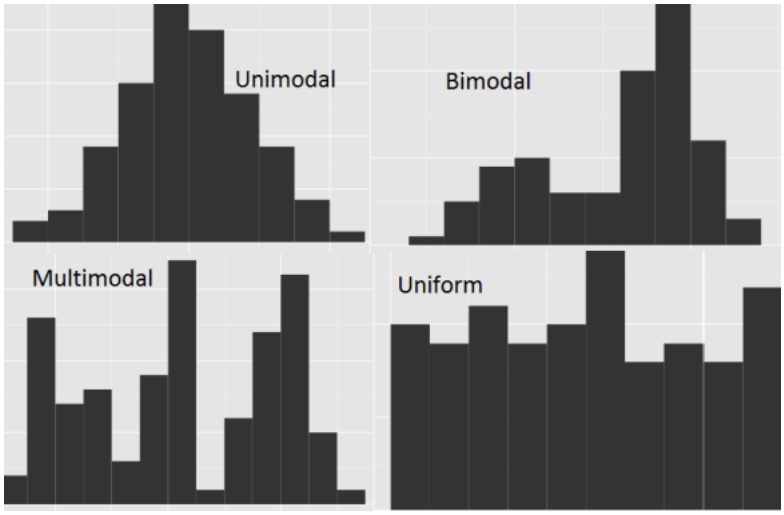


Observe the above histogram and see the distribution. There are three kind of shapes.

1. Middle one has the shape of a bell curve, has one peak, and is approximately symmetric.
2. Left one is left skewed and unimodal
3. Right one is right skewed and unimodal

Four kind of modalities are there

- Unimodal: It has only one peak
- Bimodal: It has two peak
- Multimodal: It has many peak
- Uniform: All are distributed uniformly

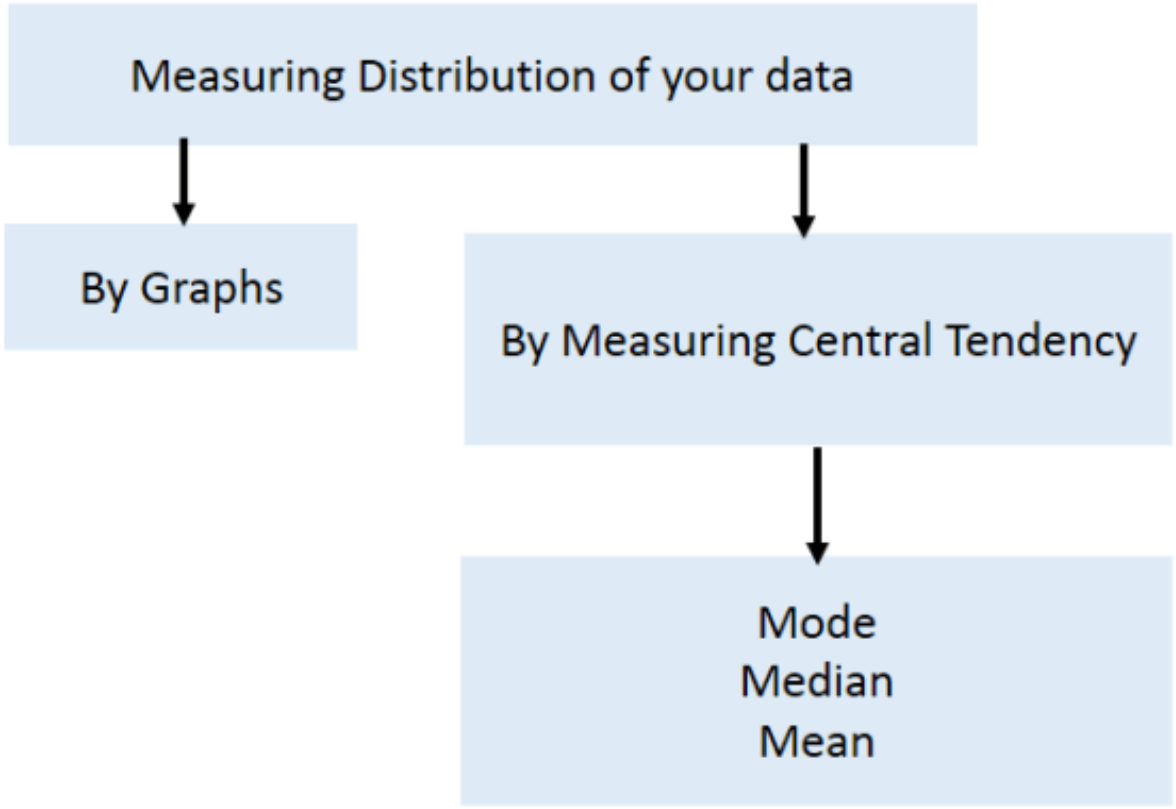


Whenever working with any data don't forget to observe shape of a distribution. As it has essential importance because it could affect the statistical methods you are going to employ later.

MODE, MEDIAN AND MEAN

In "Graphs and shapes of Distributions" section it is explained how to **summarizing a distribution** of your data in terms of graphs. Now it's time to measure the **center of your distribution**. Once we talk about measuring central tendency of a variable then 3 M's come into picture.

1. Mode
2. Median
3. Mean

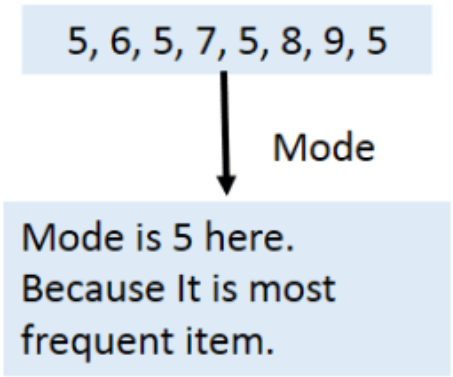


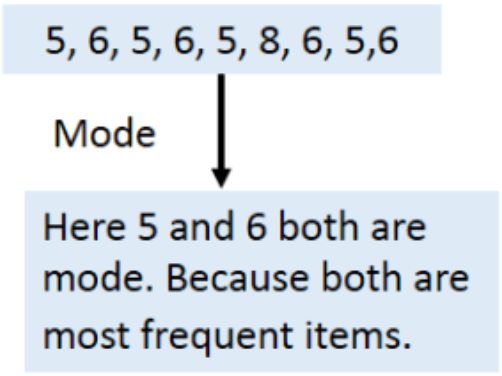
Mode:

If your variable of interest is measured in nominal or ordinal (Categorical) level then Mode is the most often used technique to measure the central tendency of your data.

Finding the mode is easy. Basically, it is the value that occurs most frequently. In other words, mode is the most common outcome. Mode is the name of the category that occurs more often.

There is a chance of having more than one mode in your variable.





Here you have two modes.

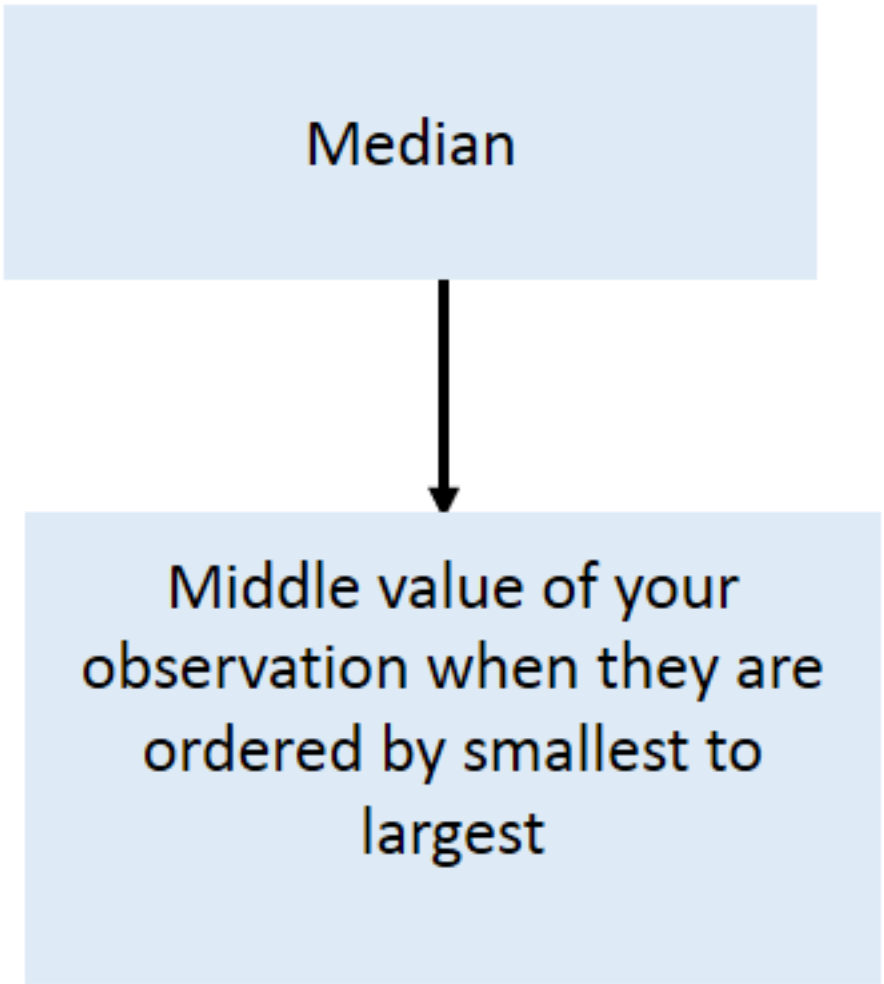
Example:

If you still don't understand and want to know you to calculate mode step by step then follow the below link

http://www.purplemath.com/modules/meanmode.htm

Median:

The second measure of central tendency is the median. The median is nothing more than the middle value of your observations when they are order from the smallest to the largest.



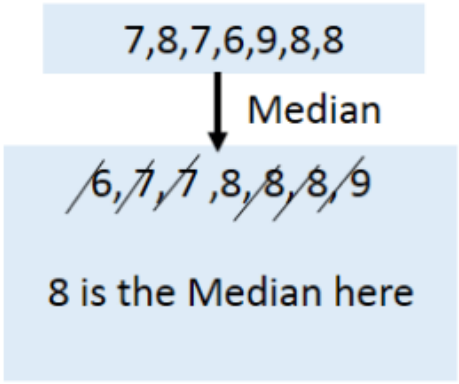
It involves two steps:

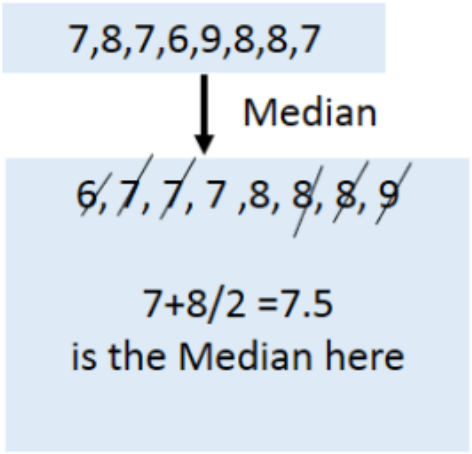
1. Oder your cases from smallest to largest
2. Find the middle Value

- If you have odd number of cases then finding middle value is easy. Let's think you have 5 cases. So, after ordering always 3rd position is the middle value.

- If you have even number no cases (let's think 6 cases). In this case there is no single middle value. Then how do we calculate median? Well, we just take the average of the two middle values.

Example:

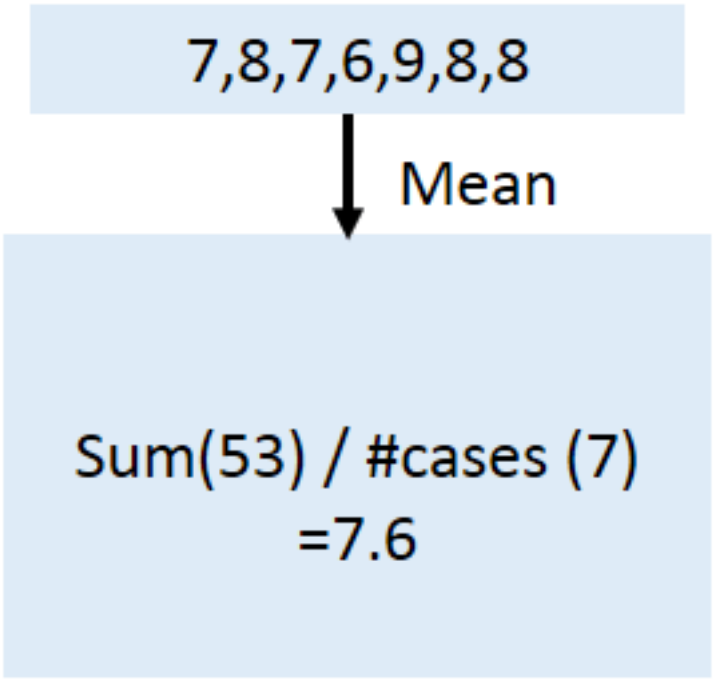




Mean:

The third measure of central tendency is the most often used one, and also the one you most probably already know quite well: the mean. The mean is the sum of all the values divided by the number of observations. It is nothing but the average value.

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

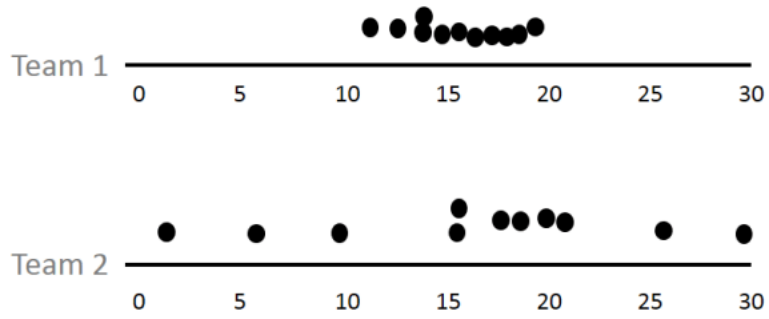


Now the question is When to use what measurement of central tendency?

- If data is Categorical (Nominal or Ordinal) it is impossible to calculate mean or median. So, go for mode.
- If your data is quantitative then go for mean or median. Basically, if your data is having some influential outliers or data is highly skewed then median is the best measurement for finding central tendency. Otherwise go for Mean.

RANGE, INTERQUARTILE RANGE AND BOX PLOT

Let's think, in certain cases, you are comparing two groups. You have already calculated the central tendency of your data i.e. Mean, Median and Mode for both the groups. Sometimes it may happen that mean, median, and mode are same for both groups. Let's take the below example:



If you consider both the team their Mode= 14.1, Median=15 and Mean=15

This indicates that, if you adequately describe a distribution some time it may need **more information than the measures of central tendency.**

In this situation **measures of variability** comes into picture. They are

- Range
- Interquartile range.
- Box Plot to get good indication of how the values in a distribution are spread out.

Range:

The most simple measure of variability is the range. It is the difference between the highest and the lowest value.

For the above Example range will be:

Range(team1) = 19.3 – 10.8 = 8.5

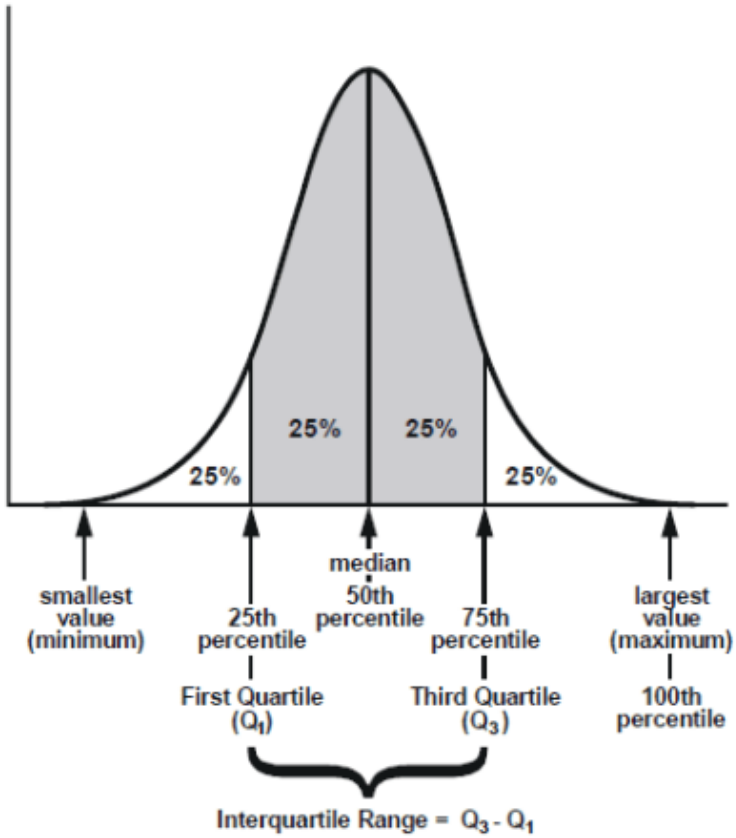
Range(team2) = 27.7-0 = 27.7

As ranges takes only the count of extreme values sometimes it may not give you a good impact on variability. In this case, you can go for another measure of variability called interquartile range (IQR).

Interquartile Range (IQR):

Interquartile range gives another measure of variability. It is a better measure of dispersion than **range** because **it leaves out the extreme values**. It equally divides the distribution into four equal parts called quartiles. First 25% is 1st quartile (Q1), last one is 3rd quartile (Q3) and middle one is 2nd quartile (Q2).

2nd quartile (Q2) divides the distribution into two equal parts of 50%. So, basically it is same as **Median.**



The interquartile range is the distance between the third and the first quartile, or, in other words, IQR equals Q3 minus Q1

IQR = Q3- Q1

How to calculate IQR

Step 1: Order from low to high

Step 2: Find the median or in other words Q2

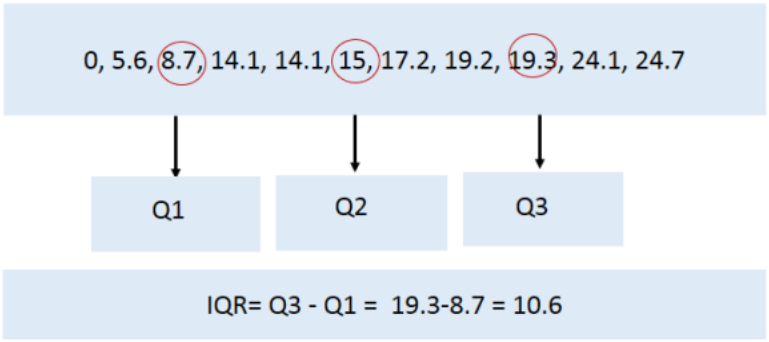
Step 3: Then find Q1 by looking the median of the left side of Q2

Steps 4: Similarly find Q3 by looking the median of the right of Q2

Steps 5: Now subtract Q1 from Q3 to get IQR.

Example:

Consider the below example to get clear idea.



Consider another example to get better understanding.

Consider the following numbers: 1, 3, 4, 5, 5, 6, 7, 11. Q1 is the middle value in the first half of the data set. Since there are an even number of data points in the first half of the data set, the middle value is the average of the two middle values; that is, Q1 = (3 + 4)/2 or Q1 = 3.5. Q3 is the middle value in the second half of the data set. Again, since the second half of the data set has an even number of observations, the middle value is the average of the two middle values; that is, Q3 = (6 + 7)/2 or Q3 = 6.5. The interquartile range is Q3 minus Q1, so IQR = 6.5 – 3.5 = 3.

Advantage of IQR:

- The main advantage of the IQR is that it is not affected by outliers because it doesn't take into account observations below Q1 or above Q3.
- It might still be useful to look for possible outliers in your study.
- As a rule of thumb, observations can be qualified as outliers when they lie more than 1.5 IQR below the first quartile or 1.5 IQR above the third quartile.

Outliers = Q1 – 1.5* IQR OR

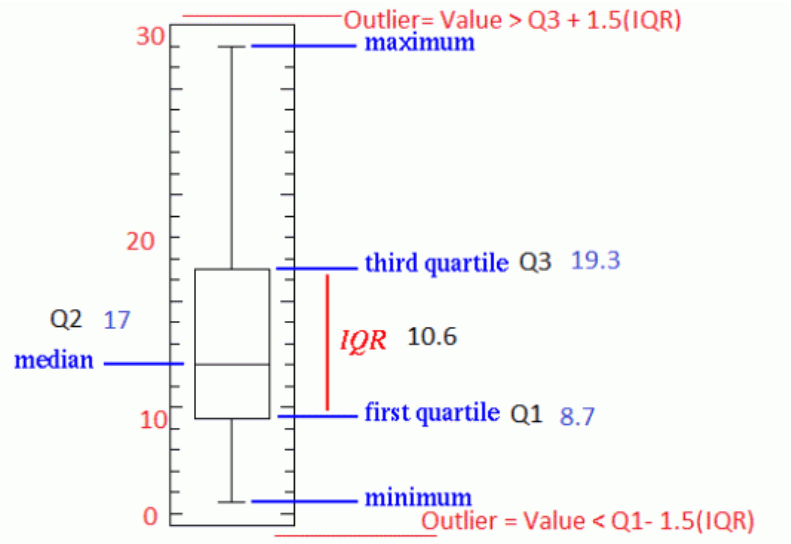
=Q3 + 1.5*IQR

Box Plot:

There is one graph that is mainly used when you are describing center and variability of your data.

It is also useful for detecting outliers in the data.

Carefully, observe the above first IQR example when it is plotted in a boxplot.



VARIANCE AND STANDARD DEVIATION

In "Range, Interquartile Range and Box Plot" section, it is explained that **Range, Interquartile Range (IQR) and Box plot** are very useful to measure the **variability of the data**.

There are two other kind of variability that a statistician use very often for their study.

1. **Variance**
2. **Standard Deviation**

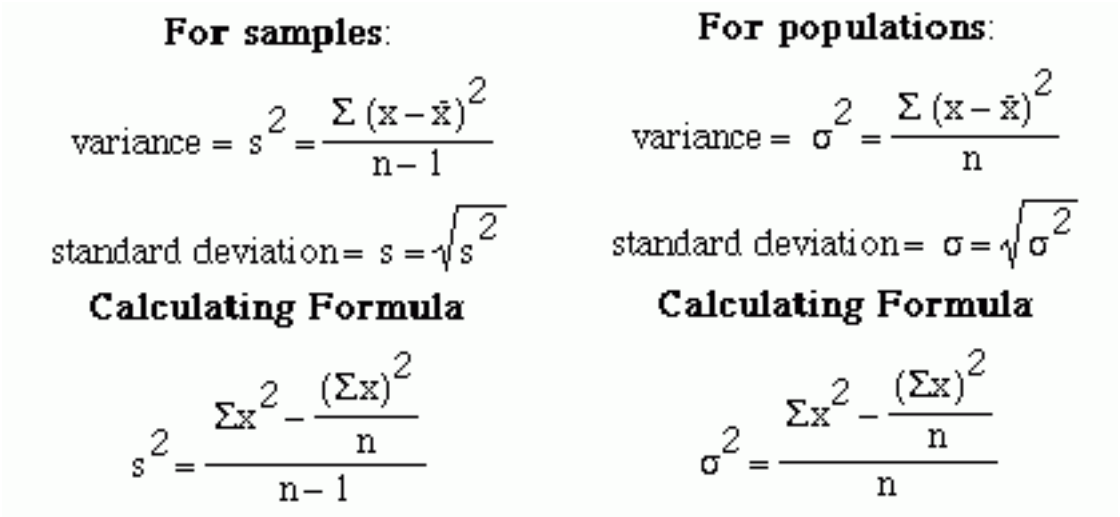
Why variance and Standard Deviation are good measures of variability?

Because variance and standard deviation consider **all the values of a variable** to calculate the variability of your data.

There are two types of variance and standard deviation in terms of Sample and Population. First their formula has been given. Then, what is the difference between sample and population has been discussed below.

Variance:

Here is the formula for sample and population variance and standard deviation. There is slight difference observe them carefully.



Where

- X is individual one value
- N is size of population
- $\bar{x}$ is the mean of population

How to calculate variance step by step:

1. Calculate the mean $\bar{x}$.
2. Subtract the mean from each observation. X- $\bar{x}$
3. Square each of the resulting observations. (X- $\bar{x}$) ^2
4. Add these squared results together.
5. Divide this total by the number of observations n (in case of population) to get variance $\mathbf{S^2}$. If you are calculating sample variance then divide by n-1.
6. Use the positive square root to get standard deviation **S**.

X	$X - \bar{X}$	$(X - \bar{X})^2$
0	-15	225
24.1	9.1	82.81
5.6	-9.4	88.36
14.1	-0.9	0.81
17.2	2.2	4.84
8.7	-6.3	39.69
19.2	4.2	17.64
14.1	-0.9	0.81
27.7	12.7	161.29
15	0	0
19.3	4.3	18.49
		639.74

Here,

N =11

N-1=10

Mean ($\bar{x}$) =15

Sample variance (s^2) = 639.74/10 = 63.97

Population (σ^2) = 639.74/11 = 58.16

S = 8.00

σ = 7.6

Intuition:

1. If variance is high, that means you have larger variability in your dataset. In the other way, we can say more values are spread out around your mean value.
2. Standard deviation represents the average distance of an observation from the mean
3. The larger the standard deviation, larger the variability of the data.

Standard Deviation:

The Standard Deviation is a measure of how spread out numbers are. Its symbol is σ (the greek letter sigma) for population standard deviation and S for sample standard deviation. It is the square root of the Variance.

Population vs. Sample Variance and Standard Deviation

The primary task of inferential statistics (or estimating or forecasting) is making an opinion about something by using only an incomplete sample of data.

In statistics, it is very important to distinguish between population and sample. A population is defined as all members (e.g. occurrences, prices, annual returns) of a specified group. Population is the whole group.

A sample is a part of a population that is used to describe the characteristics (e.g. mean or standard deviation) of the whole population. The size of a sample can be less than 1%, or 10%, or 60% of the population, but it is never the whole population. As both sample and population are not same thing therefore slight difference is there in their formula.

A question may raise that at the time of calculating Variance why we do square the difference?

To get rid of negatives so that negative and positive don't cancel each other when added together.

+5 -5 = 0