Mithilesh singh

Introduction:

Statistics

It is a mathematical science including methods of collecting, organizing and analyzing data in such a way that meaningful conclusions can be drawn from them.

→ Science of learning from data.

“science of making decisions under uncertainty."

Think of statistics as a tool that has evolved from a basic thinking process employed by every human.

In other words, statistics is a method of pursuing truth. At a minimum, statistics can tell you the likelihood that your hunch is true in this time and place and with these sorts of people.

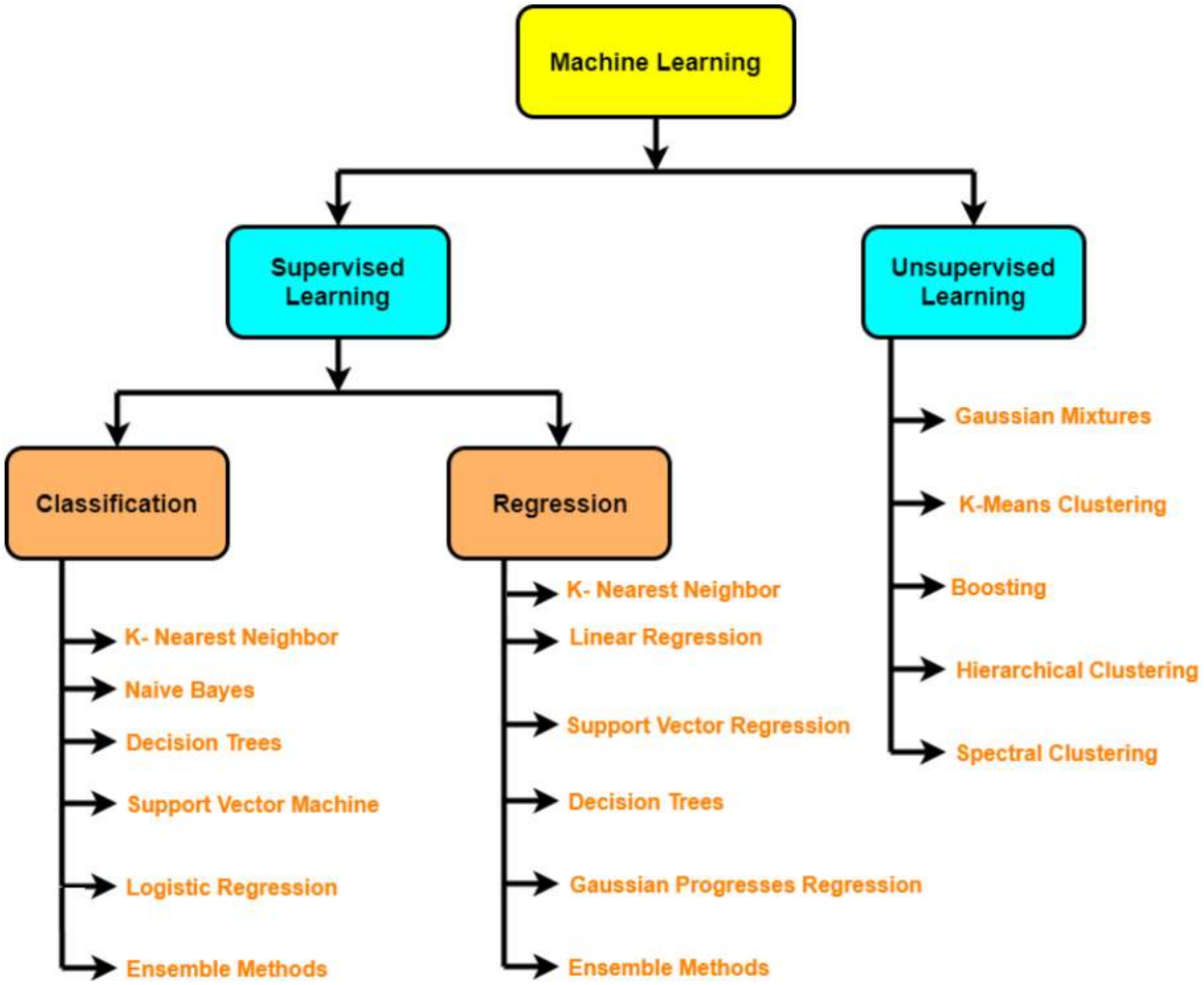
The word statistics comes from the Italian word statista, a person dealing with affairs of state (from stato, “state”). It was originally called “state

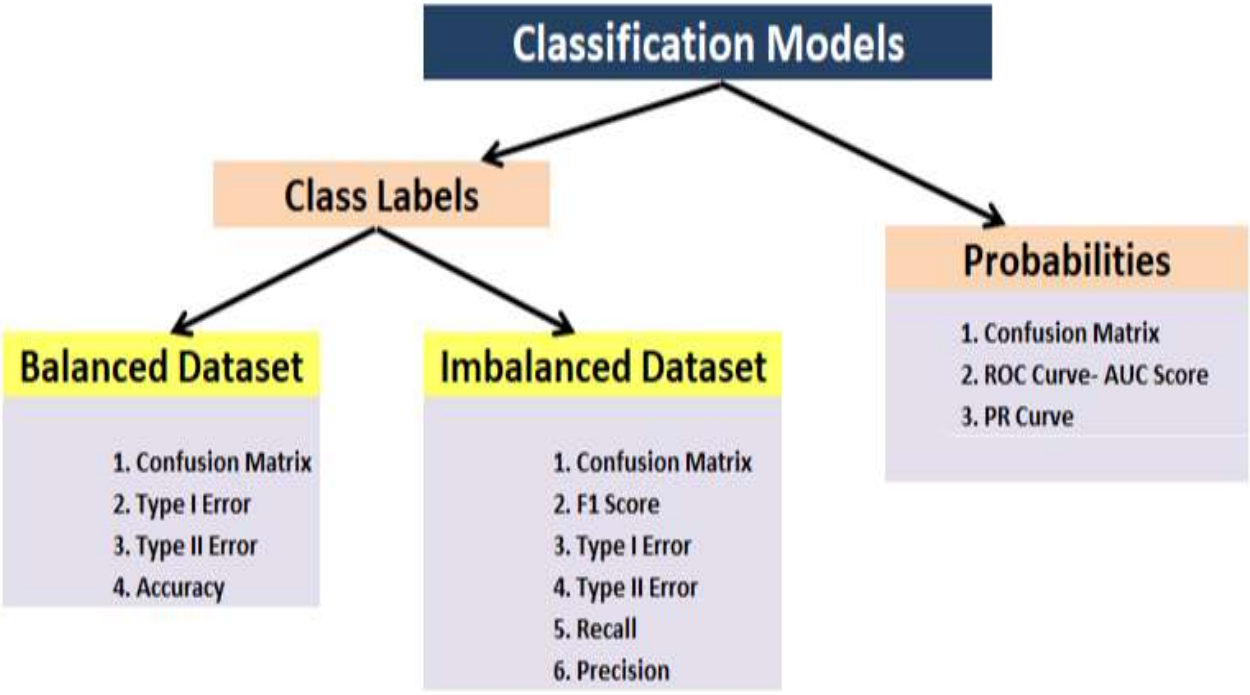
arithmetic," involving the tabulation of information about nations, especially for the purpose of taxation and planning the feasibility of wars.

Making a good foundation in Statistics- needed for understanding problem statement and solution interpretation

Summary of statistics topics.:-

- **Descriptive Statistics-Data Central tendency, dispersion.**
- **Inferential stats: - Hypothesis Testing**
- **Correlation and Co-Variance**
- **Probability theorem**
- **Preprocessing etc.**




Classification Models
Class Labels
Probabilities
Balanced Dataset
1. Confusion Matrix
2. Type I Error
3. Type II Error
4. Accuracy
Imbalanced Dataset
1. Confusion Matrix
2. F1 Score
3. Type I Error
4. Type II Error
5. Recall
6. Precision
1. Confusion Matrix
2. ROC Curve- AUC Score
3. PR Curve

Statistical Thinking :-

Day to day life example:-

EXAMPLE 1:- “Vacation in Goa , you are Renting a basic Bike for one day local sight-seeing Rs.800, do you want to buy the Optional bike insurance for Rs. 300/per day?”

Example 2:- My friend wise grandfather smoked 5 packs a day and drank a quart of scotch a day, but was always healthy and died peacefully in his sleep when he was 90, Do you conclude that all the health warnings about cigarettes are wrong?

Example 3:- A celebrity advertised Amul ice-cream in Month May last year, due to that ice-cream sales increase 25% in following three months. Thus the advertisement was effective.

Example 4:- The more Liquor shop in the city , the more crime there is , so the liquor shop is the responsible for crime?

Some Basic Concepts about data Variables---
Kind of variables
Numerical (Quantitative variables)
Categorical variables

```
In [13]: df.head()#Car Performance dataset
```

Out[13]:

	origin	cylinders	model_year	mpg_level	car_company	mpg	displacement	horsepower	weight	acceleration
0	usa	8	70	medium	chevrolet	18.0	307.0	130.0	3504	12.0
1	usa	8	70	low	buick	15.0	350.0	165.0	3693	11.5
2	usa	8	70	medium	plymouth	18.0	318.0	150.0	3436	11.0
3	usa	8	70	low	amc	16.0	304.0	150.0	3433	12.0
4	usa	8	70	medium	ford	17.0	302.0	140.0	3449	10.5

Measure types(variable)

The basic distinction is between

QUANTITATIVE DATA :-(for which one asks “how much?”)

Data that measure in Numbers. Like Height, weight, Score, Salary.

CATEGORICAL DATA (for which one asks “what type?”).

Type of city, color , Department, Education field.

(1) Quantitative variables

(a)Discrete number- counted as whole no.

exp:- No. of kids 2, 3, Training session 1,2,3 its always integer not a float.(2.5)

(b)Continuous: - Number can be infinite precision.

Weight :- 105, 89, 73.5 kg,

Example-you can sleep 7 hr 30 min & 20 sec, distance can be 70.97 meters

(2) Categorical variables

The only measure of central tendency can be use is *the mode*.

Types: -

Categorical- Ordinal

Some examples of variables that can be measured on an ordinal scale include:

- **Satisfaction:** unsatisfied, neutral, satisfied, very satisfied
- **Socioeconomic status:** Low income, medium income, high income
- **Workplace status:** Entry Analyst, Analyst I, Analyst II, Lead Analyst
- **Degree of pain:** Small amount of pain, medium amount of pain, high amount of pain

Categorical-nominal

No order can be defined. No hierarchy.

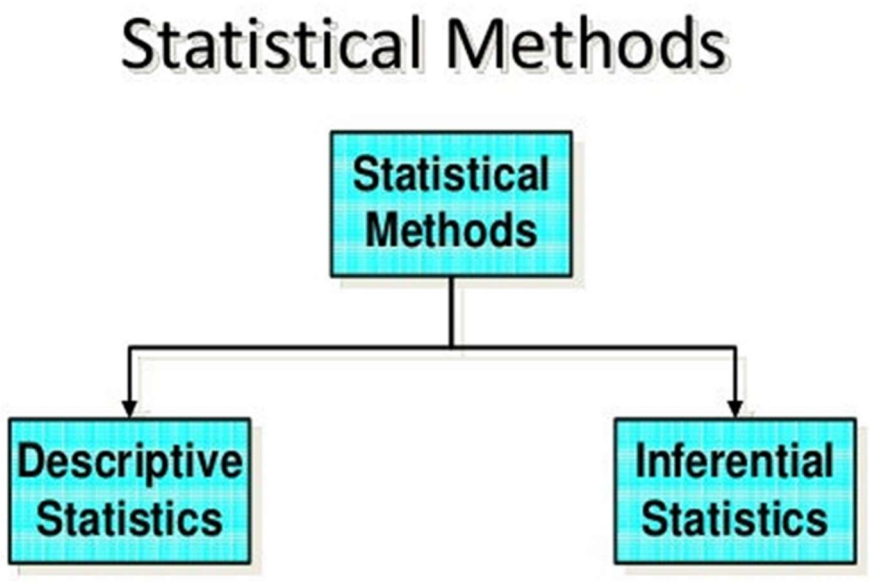
Some examples of variables that can be measured on a nominal scale include:

- **Gender:** Male, female
- **Eye color:** Blue, green, brown
- **Blood type:** O-, O+, A-, A+, B-, B+, AB-, AB+
- **City you live:** Mumbai, Bangalore, Delhi, Chennai, Kolkata

When we transform Categorical to numerical—if the category type is
Ordinal- we can apply map approach.
Nominal- Label encoding or one hot encoding approach.

Binary or dichotomous variable: -

Binary or dichotomous variable include gender (male of female), smoker (yes or no), disease status (present or absent),property type (Residential or Commercial).



Descriptive statistics procedures for summarizing a group of data or otherwise making them more understandable. For example: -Mean, Median, Mode, Variance, Standard Deviation.

Inferential statistics procedures for drawing conclusions based on the Descriptive Statistics scores collected in a research study but going beyond them make inference on Larger population.

The field of **statistics** is concerned with collecting, analyzing, interpreting, and presenting data.

Every Industry organization is striving to become data-driven. Statistics are widely use in almost all fields engineering, economics, biology, social sciences, business, agriculture, Prediction of any events, Elections, communications, Medicine.

Statistics helps answer questions like…

- What features are the most important?
- How should we design the experiment to develop our product strategy?
- What performance metrics should we measure?
- What is the most common and expected outcome?
- How do we differentiate between noise and valid data?

For example: - field of healthcare,

1: Statistics allows healthcare professionals to monitor the vital parameters prepare Central tendency, dispersion.

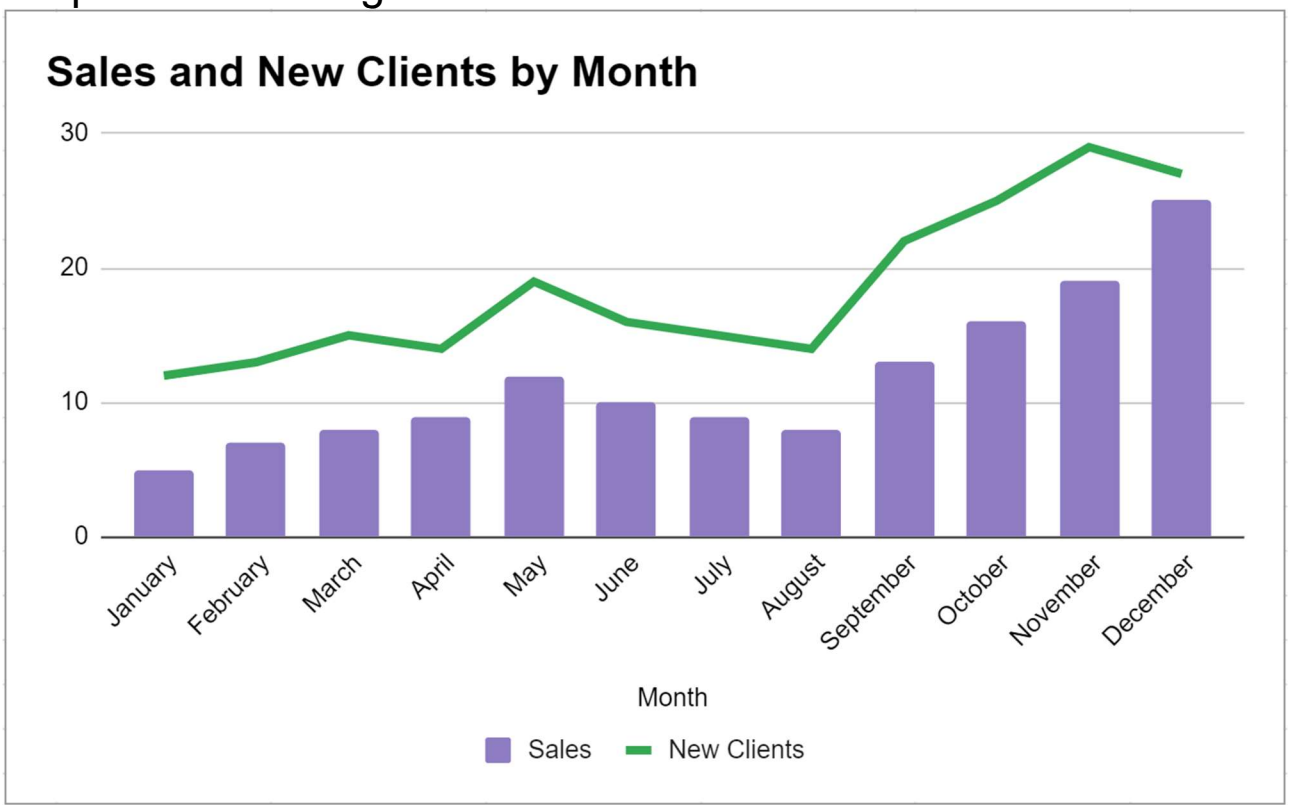
2: Quantify the relationship between variables using regression models.

3: Compare the effectiveness of different medical procedures using hypothesis tests (Inferential statistics)

Business:- statistics is important for the following reasons:

1: understand consumer behavior better using descriptive statistics.

2: spot trends using data visualization.



3: understand the relationship between different variables using regression models.

4: segment consumers into groups using cluster analysis.

Descriptive Statistics: -

Measures of Central Tendency:

A **measure of central tendency** - a single value that represents the center point of a dataset.

three common measures of central tendency:

- **The mean**
- **The median**
- **The mode**

Each of these measures finds the central location of a dataset using different methods depending on the type of data.

Mean: The average value in a dataset.

Median: The middle value in a dataset.

Mode:-The most frequently occurring value.

What is the difference between mean, median, and mode?

The mean is the average that appears in a set of data.

The median is the midway point above (below) where 50% of the values in the data sits.

The mode refers to the most frequently observed value in the data (the one that occurs the most).

Mode = 3 Median – 2 Mean

Mean or Average

Mean = (sum of all values) / (total no. of values)

Student	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10
Age in year	8	15	22	21	12	9	11	27	14	13

The mean is an indicator that can be used to gauge performance over time. Specific to investing, the mean is used to understand the performance of a company's stock price over a period of days, months, or years.

An analyst who wants to measure the trajectory of a company's stock value in, say, the last 10 days would sum up the closing price of the stock in each of the 10 days. The sum total would then be divided by the number of days to get the arithmetic mean.

Median

The **median** is the middle value in a dataset.

arranging all the values in ascending order and finding the middle value. If there are an odd number of values, the median is the middle value. If there are an even number of values, the median is the average of the two middle values.

Student	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10
Age in year	8	15	22	21	12	9	11	27	14	13

The Mode

The **mode** is the value that occurs most often in a dataset.

Player	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10
Home Runs	8	9	11	12	13	14	15	21	22	27

Player	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10
Home Runs	8	9	11	13	13	15	19	21	22	27

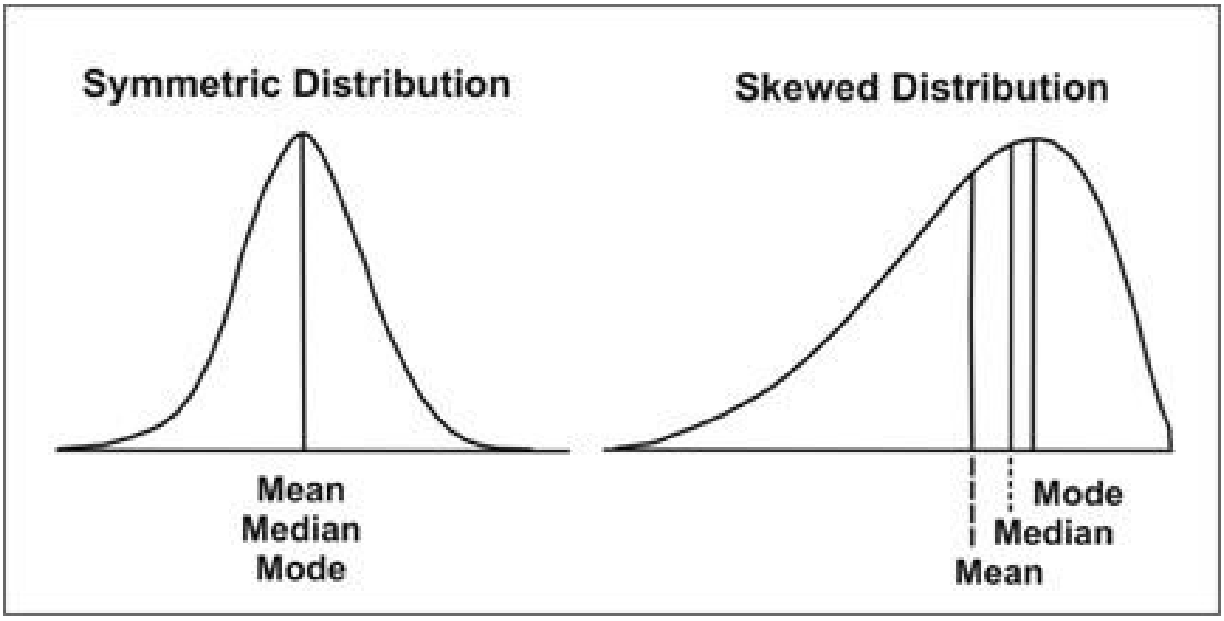
Player	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10
Home Runs	8	9	11	11	13	15	15	21	22	27

Player	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10
Home Runs	8	8	11	11	15	15	17	19	19	27

The mode can be a particularly helpful measure of central tendency when working with categorical data.

A dataset can have no mode (if no value repeats), one mode, or multiple modes.

When to Use the Mean, Median, and Mode

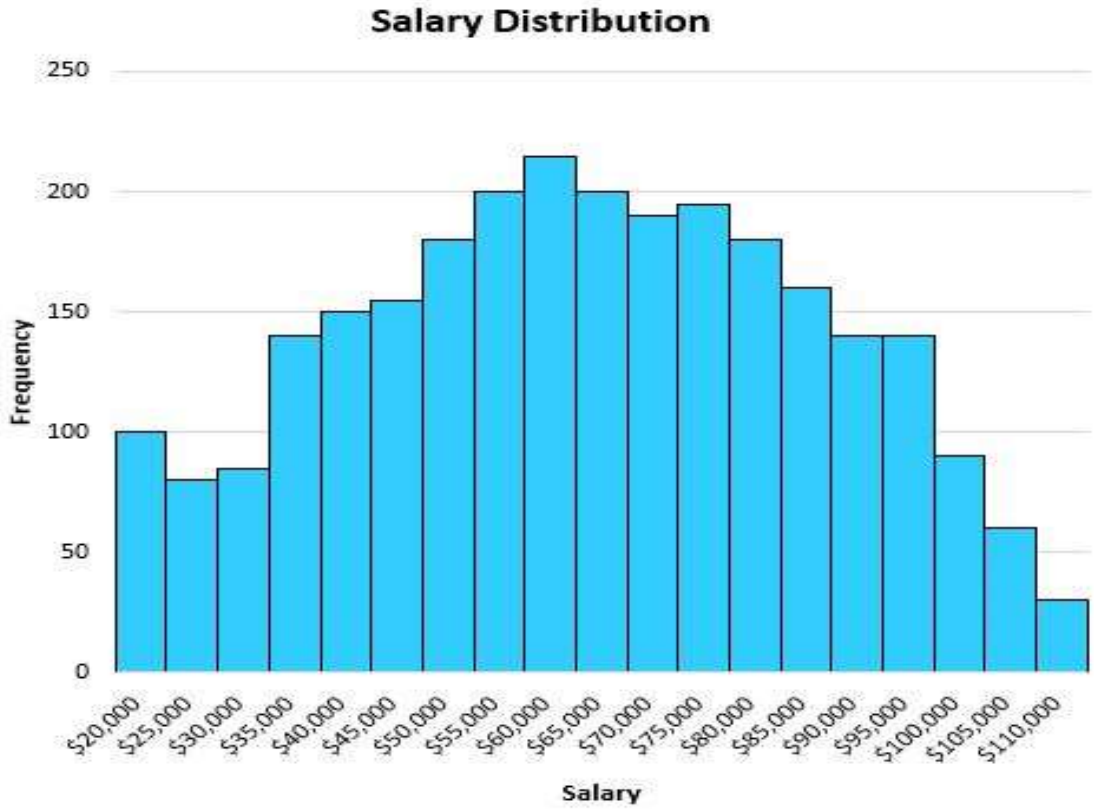


Description: In a symmetric distribution the mean, median, and mode are all located in the center of the x-axis. In a skewed distribution, the mean, median, and mode are not located together. If skewed to the right, the mean occurs to the left of the median; mode occurs to the right of the median

When to use the mean

It is best to use the mean when the distribution of the data is fairly symmetrical and there are no outliers.

For example, suppose we have the following distribution that shows the salaries of individuals in a certain town:



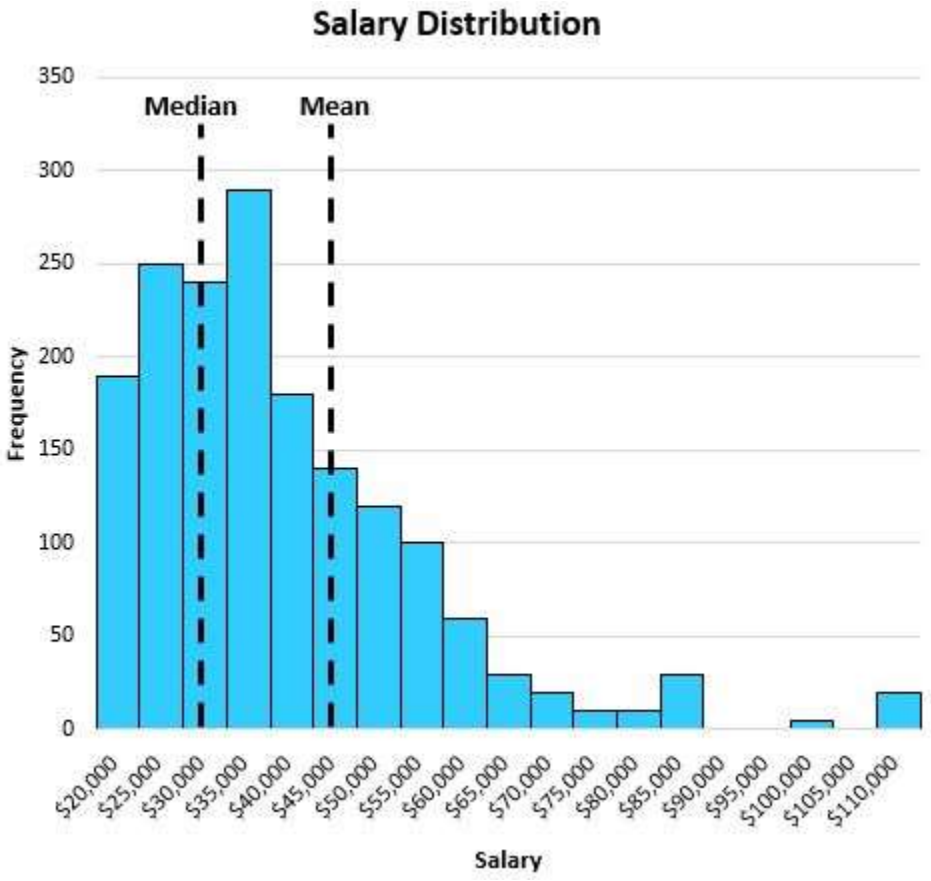
Since this distribution is fairly symmetrical (i.e. if you split it down the middle, each half would look roughly equal) and there are no outliers (i.e. no extremely high salaries

The mean turns out to be $63,000, which is roughly located in the center of the distribution:

When to use the median

It is best to use the median when the distribution of the data is either skewed or there are outliers present.

Skewed data:



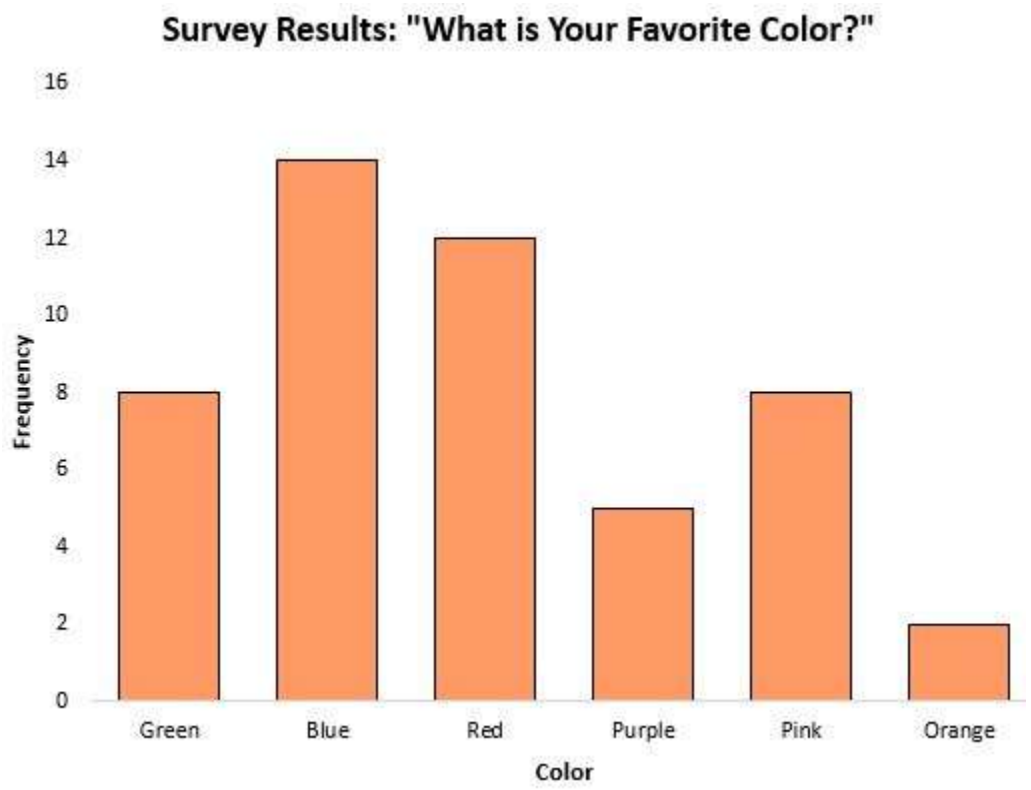
The median does a better job of capturing the “typical” salary of an individual than the mean.

In this particular example, the mean tells us that the typical individual earns about $47,000 per year in this town while the median tells us that the typical individual only earns about $32,000 per year, which is much more representative of the typical individual.

When to use the mode

It is best to use the mode when you are working with categorical data and you want to know which category occurs most frequently.

You conduct a survey about people's preferences among three choices for a website design :-



working with categorical data then it's not even possible to calculate the median or mean, mode is the measure of central tendency.

Note: It's important to note that if a dataset is *perfectly* normally distributed, then the mean, median, and mode are all the same value.

Quiz: -

Q1.Suppose all household incomes in California increase by 5%. How does that change the mean household income?

Q2. Suppose all household incomes in California increase by 5%. How does that change the median household income?

Q3. Suppose all household incomes in California increase by $5,000. How does that change the mean household income?

Q4. Suppose all household incomes in California decrease by $5,000. How does that change the median household income?

Q5. The median sales price for houses in a certain county during the last year was $342,000. What can we say about the percentage of sales represented by the houses that sold for more than $342,000?

(a) the houses that sold for more than $342,000 represent more than 50% of all sales.
(b) the houses that sold for more than $342,000 represent exactly 50% of all sales
(c) the houses that sold for more than $342,000 represent less than 50% of all sales