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.

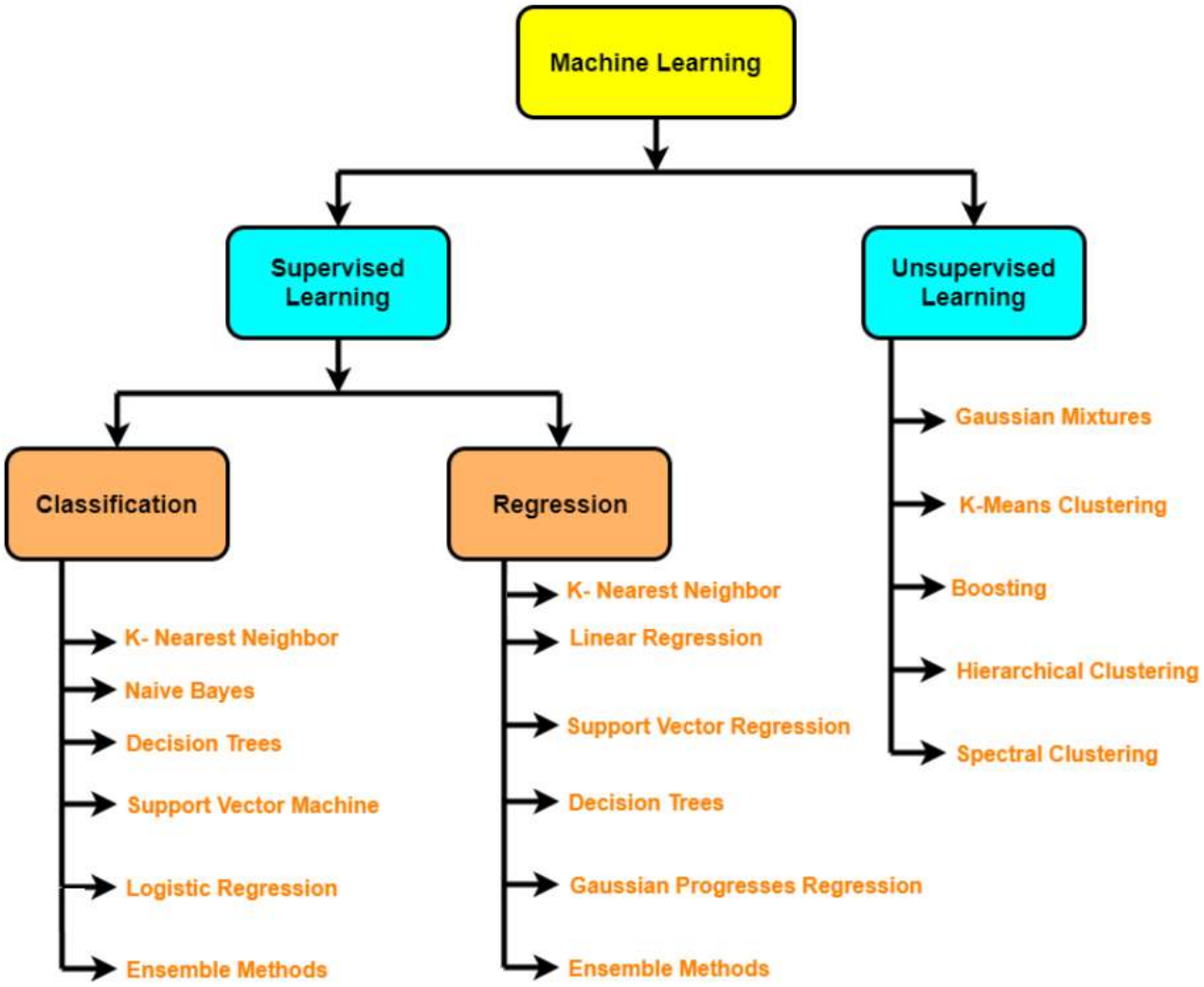
In other words, statistics is a method of pursuing truth. At a minimum, statistics can tell you the likelihood that your hunch is true in this time and place and with these sorts of people.

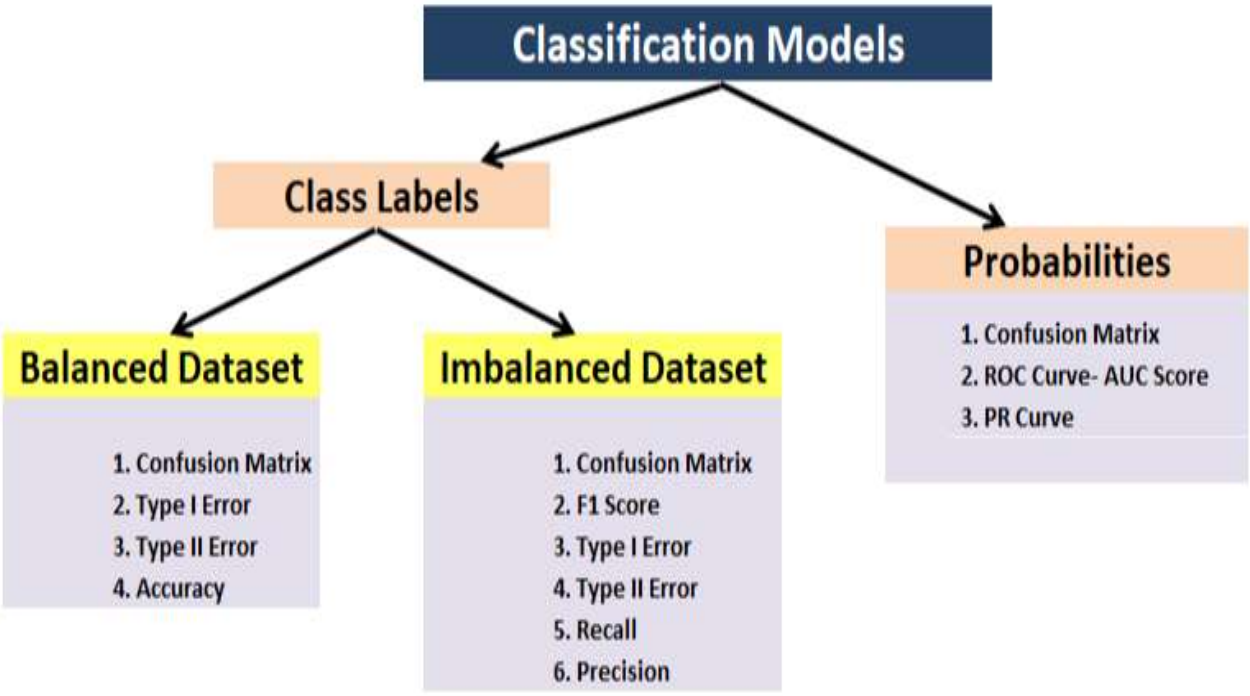
The word statistics comes from the Italian word statista, a person dealing with affairs of state (from stato, “state”). It was originally called “state arithmetic,” involving the tabulation of information about nations, especially for the purpose of taxation and planning the feasibility of wars.

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

Summary of statistics topics.:-

- **Understanding Data Central tendency**
- **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:-

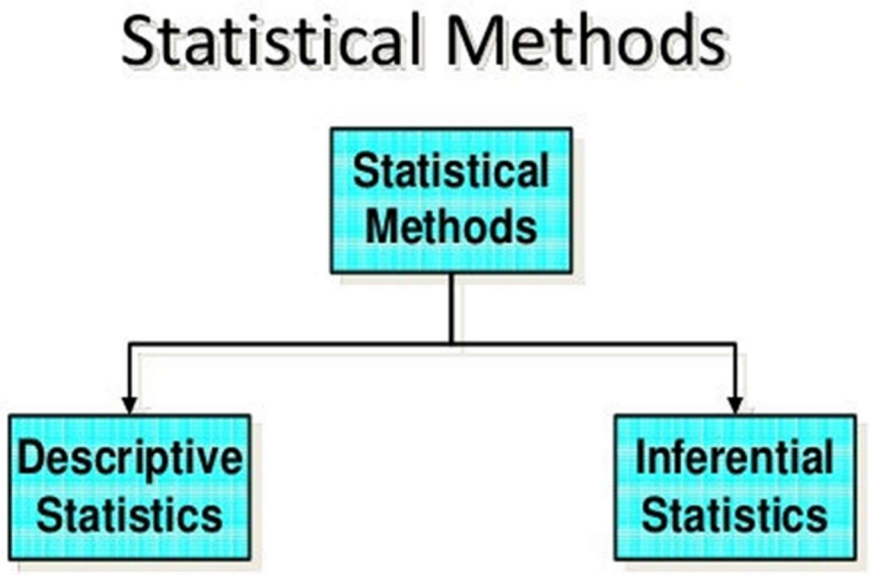
EXAMPLE 1:- “Vacation in Goa , you are Renting a basic Bike for one day local sight-seeing Rs.800, do you want to buy the Optional bike insurance for Rs. 300/per day?”

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?

The Two Branches of Statistical Methods – Descriptive statistics and Inferential Statistics



Descriptive statistics procedures for summarizing a group of data or otherwise making them more understandable.

inferential statistics procedures for drawing conclusions based on the scores collected in a research study but going beyond them

In this hyper-connected world, data are being generated and consumed at an unprecedented pace.

Data professionals need to be trained to use statistical methods to interpret numbers.

The core of machine learning is centered around statistics. You can't solve real-world problems with machine learning if you don't have a good grip of statistical fundamentals.

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?

All these are common and important questions that data teams have to answer on a daily basis.

The answers help us make decisions effectively. Statistical methods not only help us set up predictive modeling projects but also to interpret the results.

Basic Dataset Example:-

Cardio_vascular disease prediction

```
In [ ]: CARDIO_VASCULAR_disease_Prediction
```

Out[5]:

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	52	1	0	125	212	0	1	168	0	1.0	2	2	3	0
1	53	1	0	140	203	1	0	155	1	3.1	0	0	3	0
2	70	1	0	145	174	0	1	125	1	2.6	0	0	3	0
3	61	1	0	148	203	0	1	161	0	0.0	2	1	3	0
4	62	0	0	138	294	1	1	106	0	1.9	1	3	2	0

Cab Cancellation Prediction

```
In [8]: ('Cab_cancellation.csv')
```

Out[8]:

user_id	vehicle_model_id	package_id	travel_type_id	from_area_id	to_area_id	from_city_id	to_city_id	from_date	to_date
14364	12	NaN	2	1147.0	452.0	15.0	NaN	11/24/2013 18:00	41602.82020
48727	12	NaN	2	393.0	466.0	15.0	NaN	11/26/2013 1:00	41604.08037
48729	12	1.0	3	974.0	NaN	15.0	NaN	11/24/2013 18:30	41602.97917
30724	87	NaN	2	1263.0	542.0	15.0	NaN	11/24/2013 16:00	41602.69495
48730	12	NaN	2	689.0	393.0	15.0	NaN	11/25/2013 5:00	41603.25385

Distribution

Statistics is concerned with
the **frequency** and **pattern** of events in a population:

Summarizing Data

A **variable** can be any characteristic that differs from person to person, such as height, sex, smallpox vaccination status, or physical activity pattern. The **value** of a variable is the number or descriptor that applies to a particular person, such as 5’6” (168 cm), female, and never vaccinated.

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

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.

Examples of types of data	
Quantitative	
Continuous	**Discrete**
Blood pressure, height, weight, age	Number of children Number of attacks of asthma per week
Categorical	
Ordinal (Ordered categories)	**Nominal (Unordered categories)**
Grade of breast cancer Better, same, worse Disagree, neutral, agree	Sex (male/female) Alive or dead Blood group O, A, B, AB

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).

EDA ----- exploratory Data Analysis skill

Applied Statistics along with focuses on the numbers, math, and problems themselves. Applied statistics also thought of as “statistics-in-action” or using statistics with an eye toward real-world problems and what their solutions might be.

Problem solving process: -

Defining a Problem Statement

The most crucial part of predictive modeling is the actual definition of the problem that gives us the real objective to pursue.

This helps us decide the type of problem we're dealing with (that is, regression or classification). And it also helps us decide the structure and types of the inputs, outputs and metrics with regards to the objective.

But problem framing is not always straightforward, it may require significant exploration of the observations in the domain.

EDA ----- Initial Data Exploration

Data exploration involves gaining a deep understanding of both the distributions of variables and the relationships between variables in your data.

In part, domain expertise helps you gain this mastery over a specific type of variable.

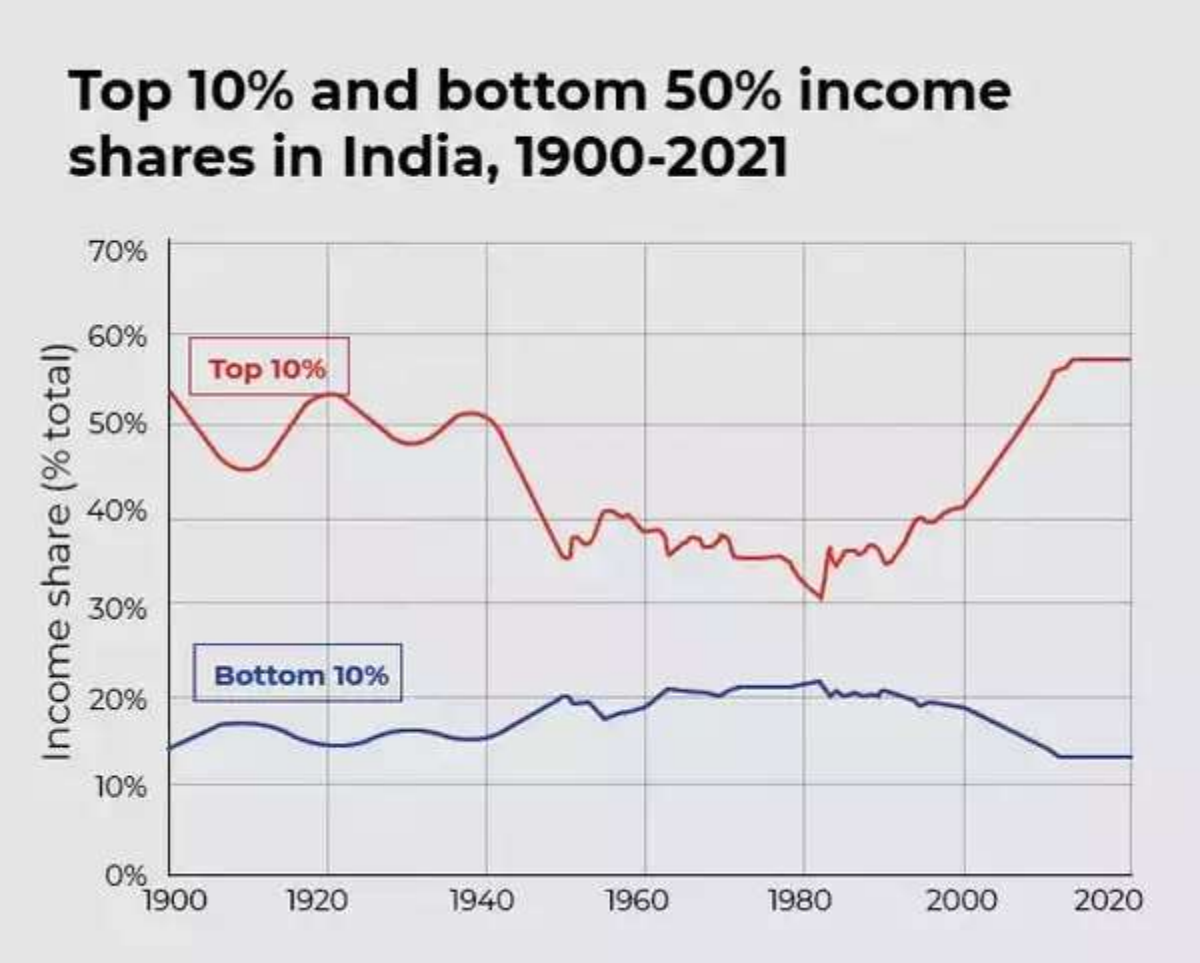
A visual presentation is good way to make a large group of scores easy to understand. A picture may be worth a thousand words, but it is also sometimes worth a thousand numbers.

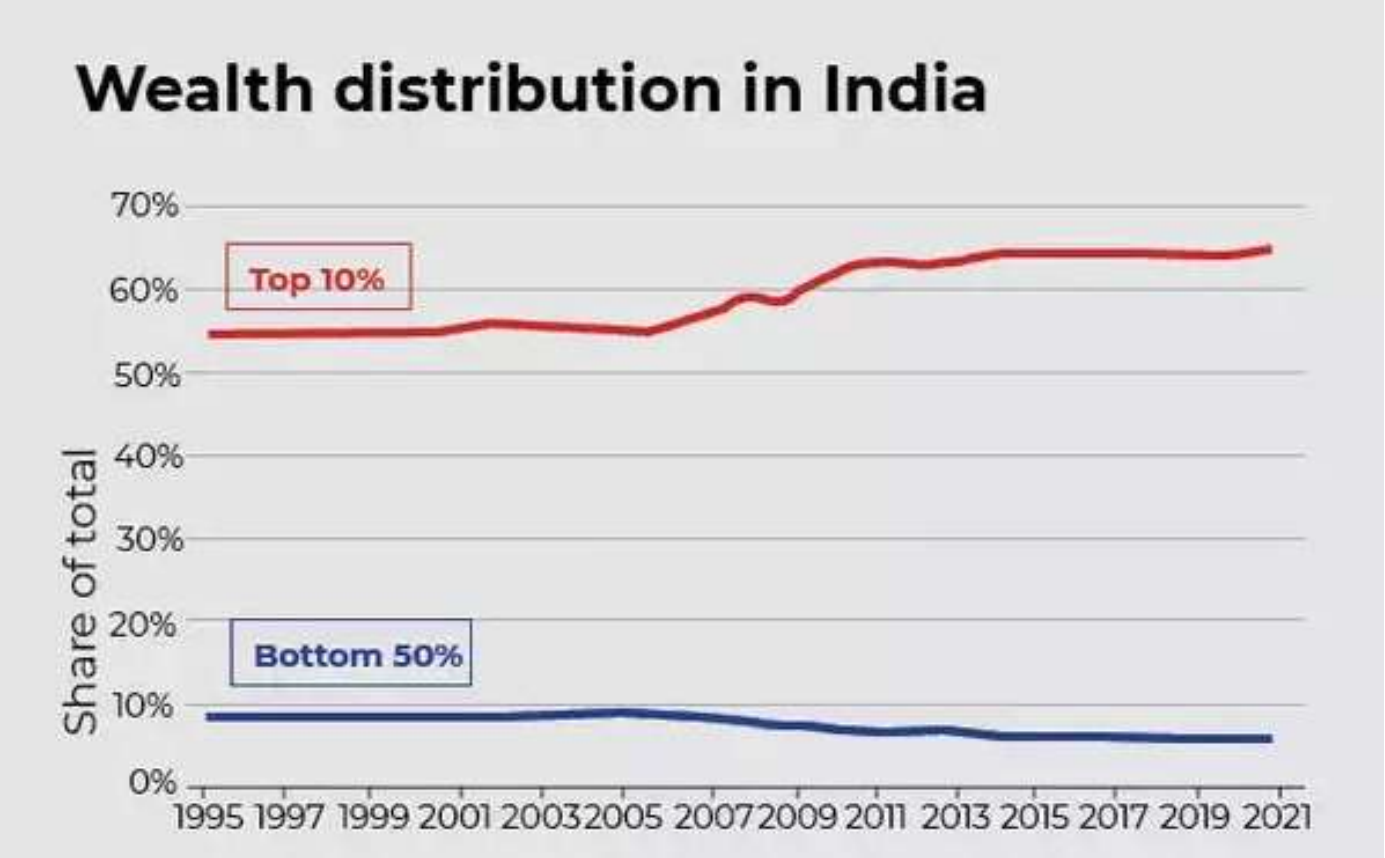
A straightforward approach is to make a graph of the frequency table. One kind of graph of the

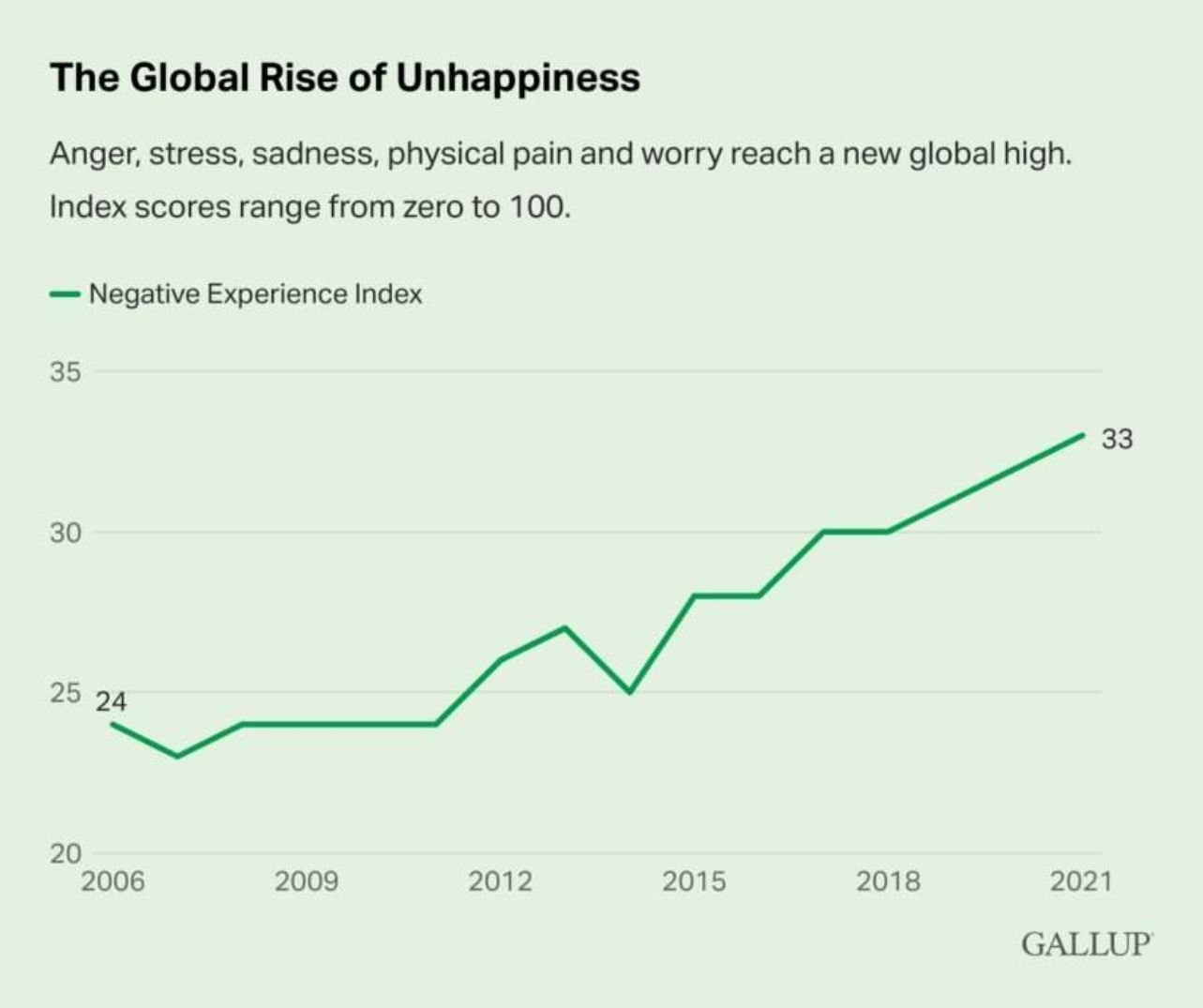
information in a frequency table is count plot, bar plot , histogram, scatter plot, pie plot etc.

Matplotlib

Seaborn




Wealth distribution in India
70%
60%
50%
40%
30%
20%
10%
0%
Share of total
Top 10%
Bottom 50%
1995 1997 1999 2001 2003 2005 2007 2009 2011 2013 2015 2017 2019 2021



Many things can make people unhappy, but the rise of global unhappiness has five main causes according to Gallup's research: **poverty, broken communities, hunger, loneliness and the scarcity of good work**.

Types of Statistics:

1. Descriptive statistics

   Summaries data by transforming raw observations into meaningful information that is easy to interpret and share.

   For example:-Mean, Median, Mode, Variance, Standard Deviation.

   (Without attempting to draw any inferences about population from it).

2. Inferential statistics

   To study experiments done on small samples of data and chalk out the inferences to the entire population (entire domain).uses mathematical tools to make forecasts and projections by analyzing the given data. (attempting to draw any inferences about the population from it).

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

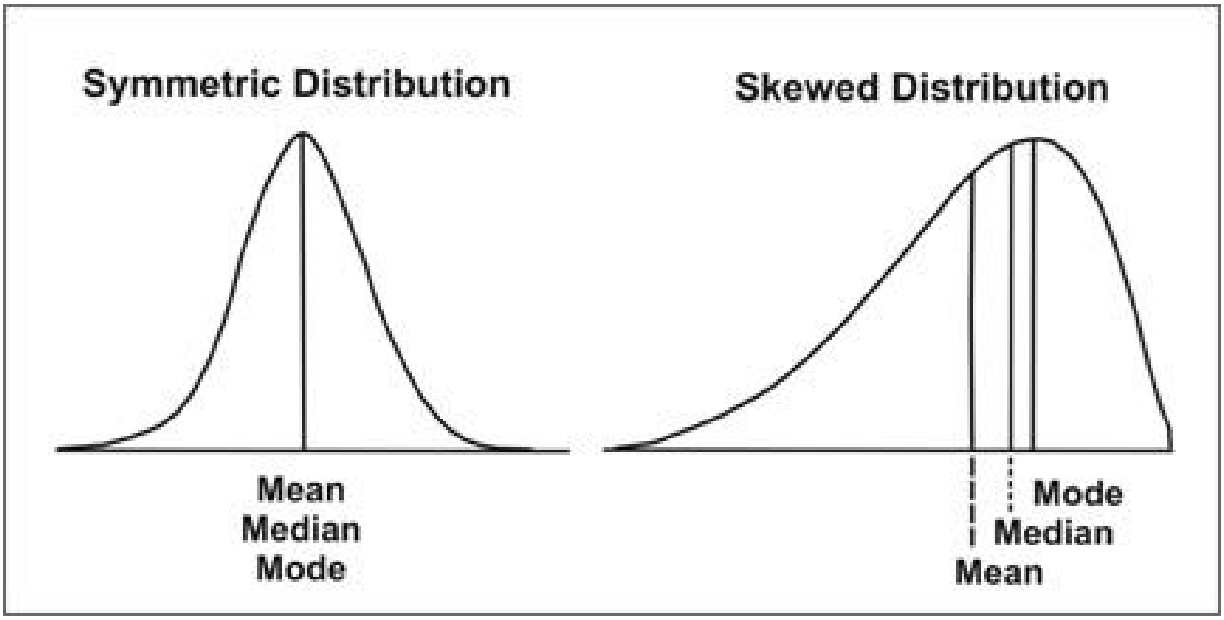
Player	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10
Home Runs	8	9	11	11	13	15	15	21	22	27

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

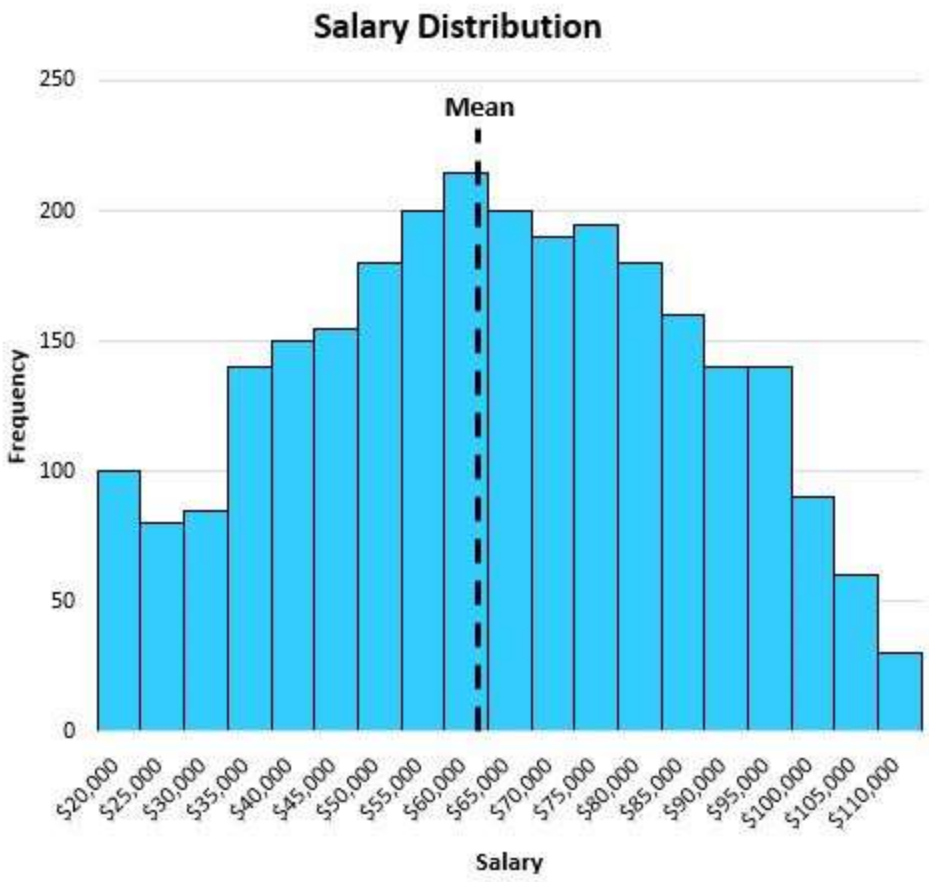
When to use the mean

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

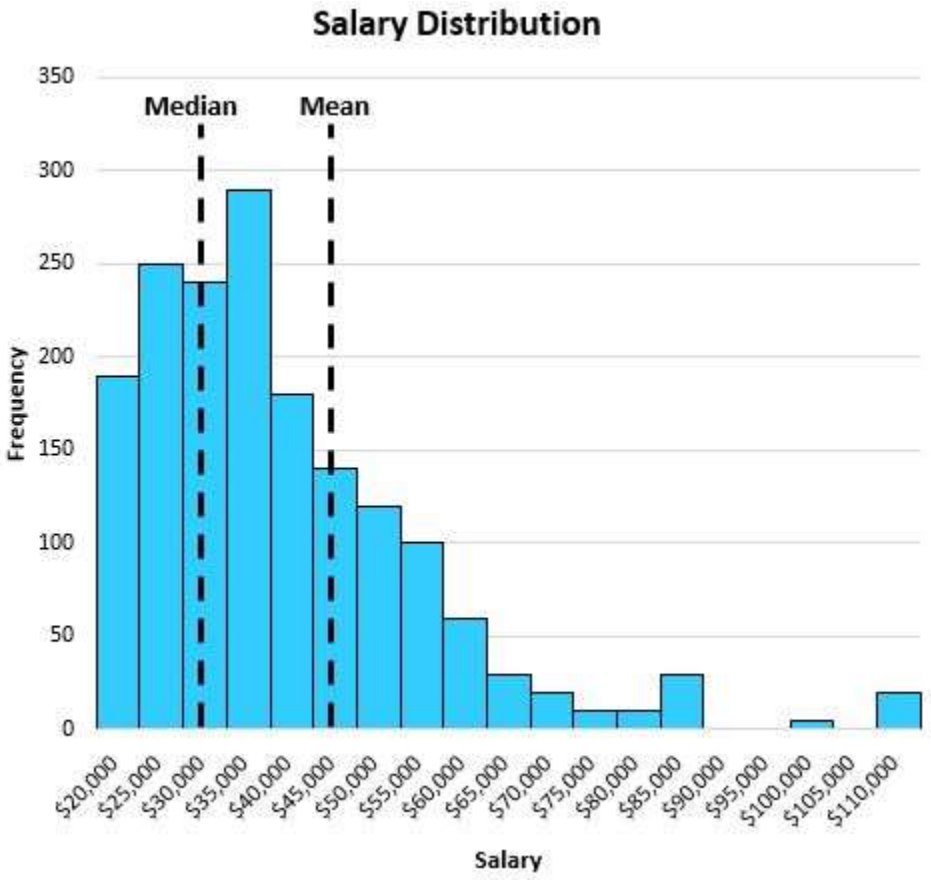
The mean turns out to be $63,000, which is roughly located in the center of the distribution:


Salary Distribution
Mean
250
200
150
100
50
0
Frequency
$20,000
$25,000
$30,000
$35,000
$40,000
$45,000
$50,000
$55,000
$60,000
$65,000
$70,000
$75,000
$80,000
$85,000
$90,000
$95,000
$100,000
$105,000
$110,000
Salary

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:

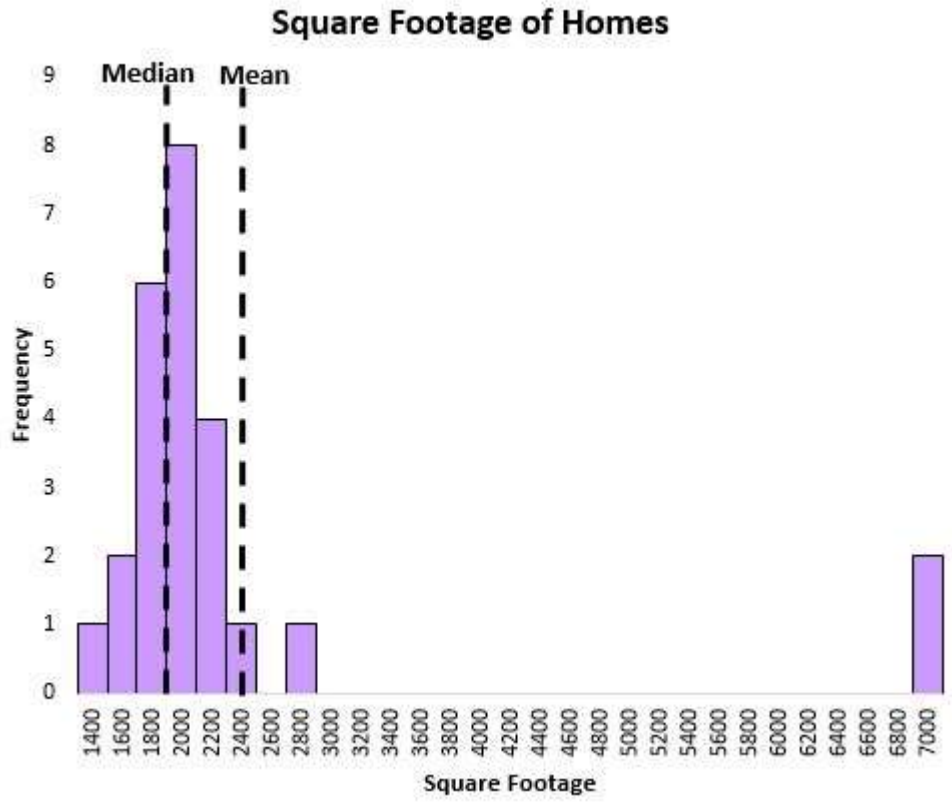


The median does a better job of capturing the “typical” salary of an individual than the mean.

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.

Outliers:

The median also does a better job of capturing the central location of a distribution when there are outliers present in the data. For example, consider the following chart that shows the square footage of houses on a certain street:

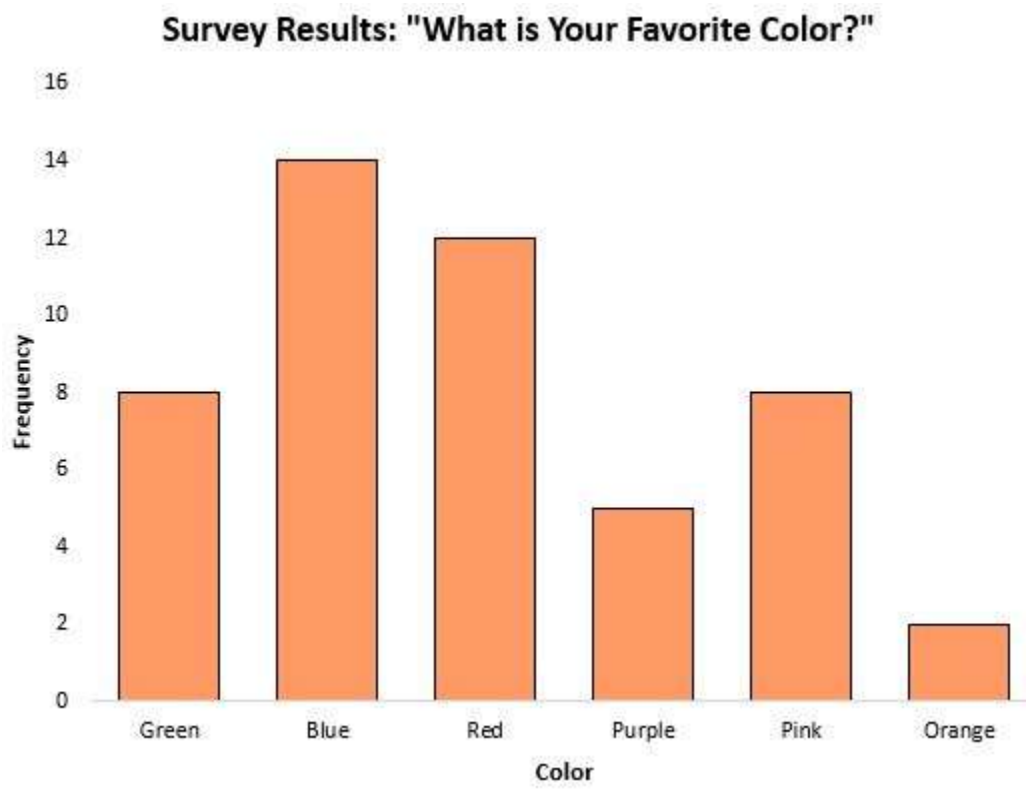


The mean is heavily influenced by a couple extremely large houses, while the median is not. Thus, the median does a better job of capturing the “typical” square footage of a house on this street compared to the mean.

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.

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

Variability :-

Researchers also want to know how spread out the scores are in a distribution. This shows the amount of variability in the distribution.

For example, suppose you were asked, “How old are the students in the statistics class?”, the mean age might be 29 years.

You could answer, “The average age of the students in my class is 29.” However, this would not tell the whole story. You could have a mean of 29 because every student in the class was exactly 29 years old.

You can think of the variability of a distribution as the amount of spread of the scores around the mean. In other words, how close or far from the mean are the scores in a distribution?

If the scores are mostly quite close to the mean, then the distribution has less variability than if the scores are further from the mean. Distributions with the same mean can have very different amounts of spread around the mean.

Measure of Spread / Dispersion

In statistics, the **standard deviation** (SD, also represented by the Greek letter sigma σ or the Latin letter s) is a measure that is used to quantify the amount of variation or dispersion of a set of data values. It is widely used measure of variability or diversity used in statistics and probability theory.

It shows how much variation or "dispersion" exists from the average (mean or expected value).

A low standard deviation indicates that the data points tend to be very close the mean, whereas high standard deviation indicate that the data points are spread out over a large range of values.

SD is the square root of its variance. A useful property of standard deviation is that , unlike variance , it is expressed in the same units as the data.

In addition to expressing the variability of a population, standard deviation is commonly used to measure confidence in statistical conclusions.

1. Standard deviation: - Standard deviation is the measurement of average distance between each quantity and mean. That is, how data is spread out from mean. A low standard deviation indicates that the data points tend to be close to the mean of the data set, while a high standard deviation indicates that the data points are spread out over a wider range of values

Sample SD

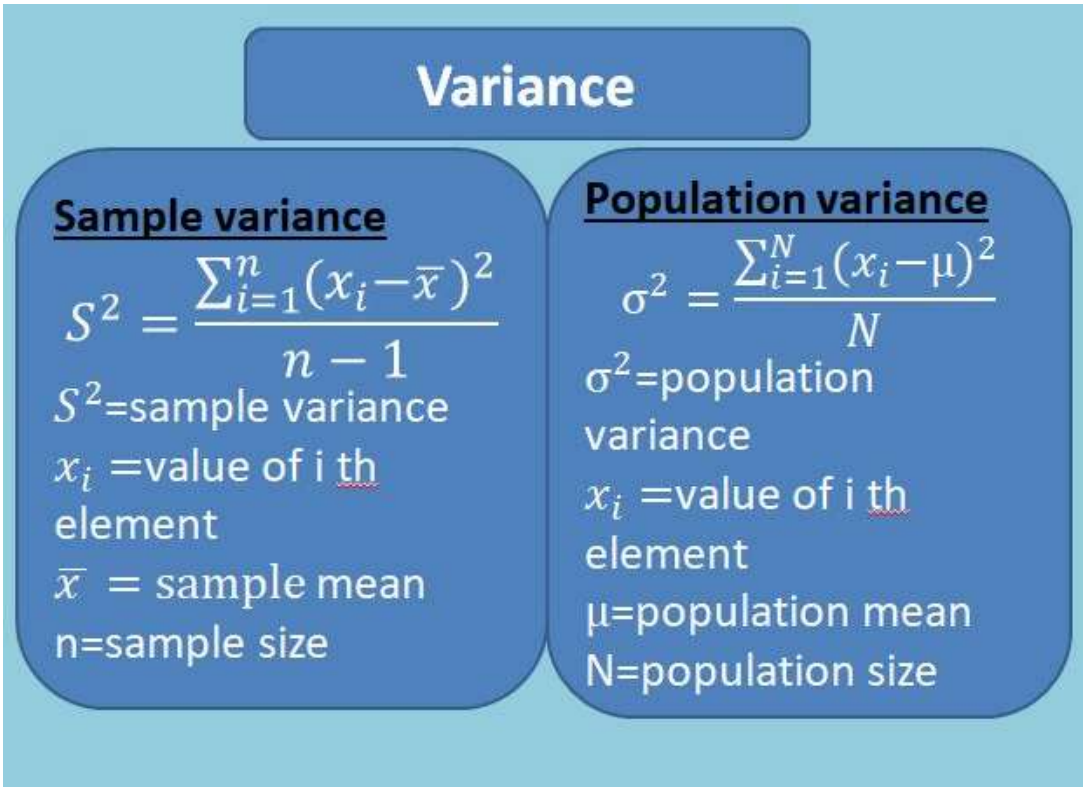
Population SD

Standard Deviation Formula	
Sample	**Population**
$s = \sqrt{\frac{\sum(X - \bar{x})^2}{n-1}}$	$\sigma = \sqrt{\frac{\sum(X - \mu)^2}{N}}$
X – The Value in the data distribution *$\bar{x}$ – The Sample Mean* *n - Total Number of Observations*	*X – The Value in the data distribution* *μ – The population Mean* *N – Total Number of Observations*

2. Variance: - Variance is a square of average distance between each quantity and mean. That is the square of standard deviation.

Sample Variance

Population Variance



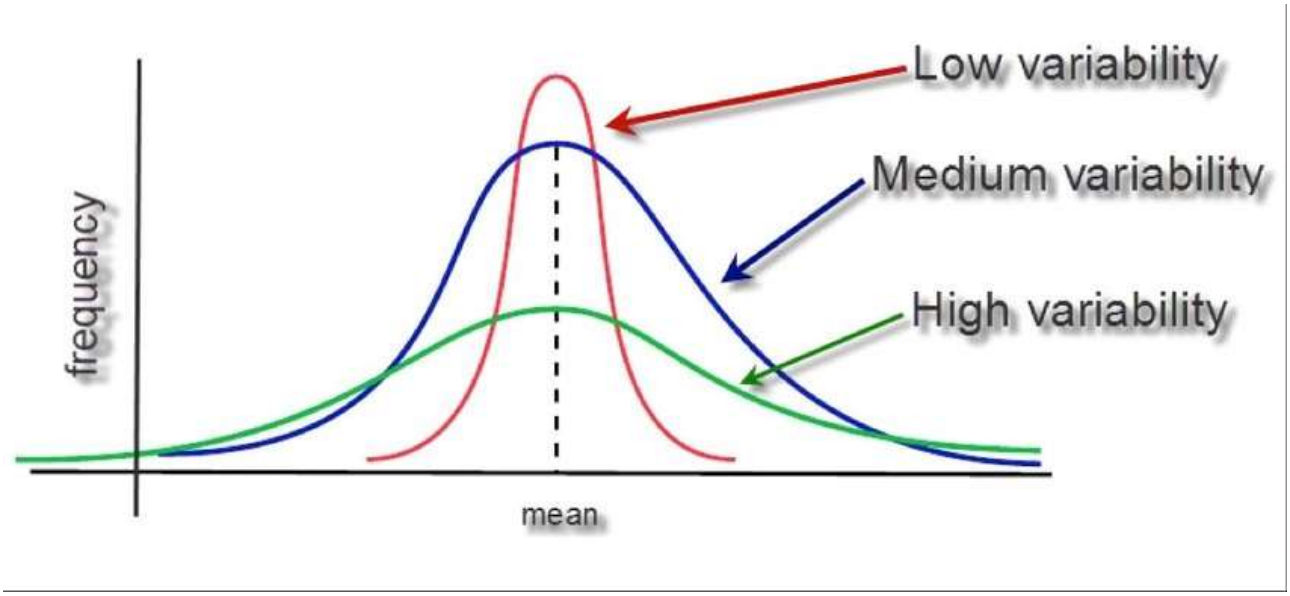
Example of Variance in Finance

Here's a hypothetical example to demonstrate how variance works. Let's say returns for stock in Company ABC are 10% in Year 1, 20% in Year 2, and −15% in Year 3. The average of these three returns is 5%. The differences between each return and the average are 5%, 15%, and −20% for each consecutive year.

Squaring these deviations yields 0.25%, 2.25%, and 4.00%, respectively. If we add these squared deviations, we get a total of 6.5%. When you divide the sum of 6.5% by one less the number of returns in the data set, as this is a sample (2 = 3-1), it gives us a variance of 3.25% (0.0325).

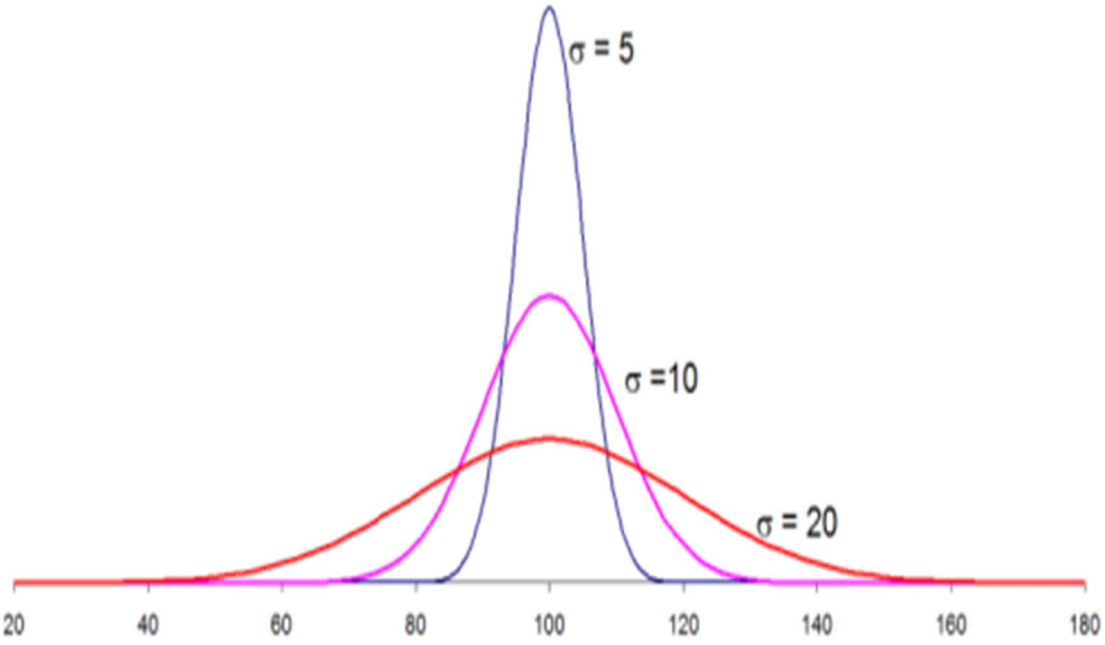
Taking the square root of the variance yields a standard deviation of 18% (√0.0325 = 0.180).

House price in Location A,B and C.

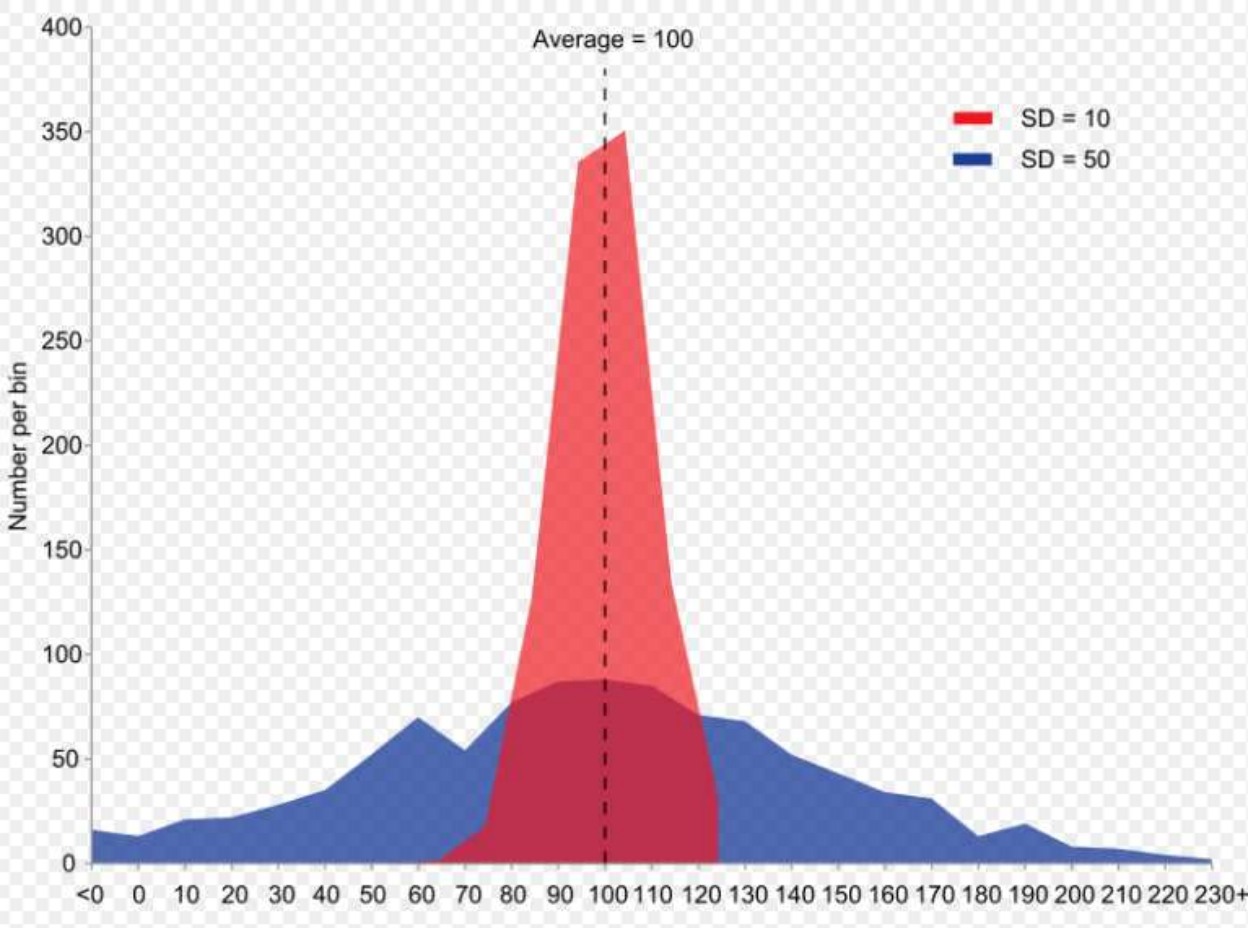


Example of three distributions with the same mean but different amounts of spread: housing prices for a neighborhood with diverse types of housing and for a neighborhood with a consistent type of housing.

Example: - Cricket Match player



Example: -Pizza Delivery time



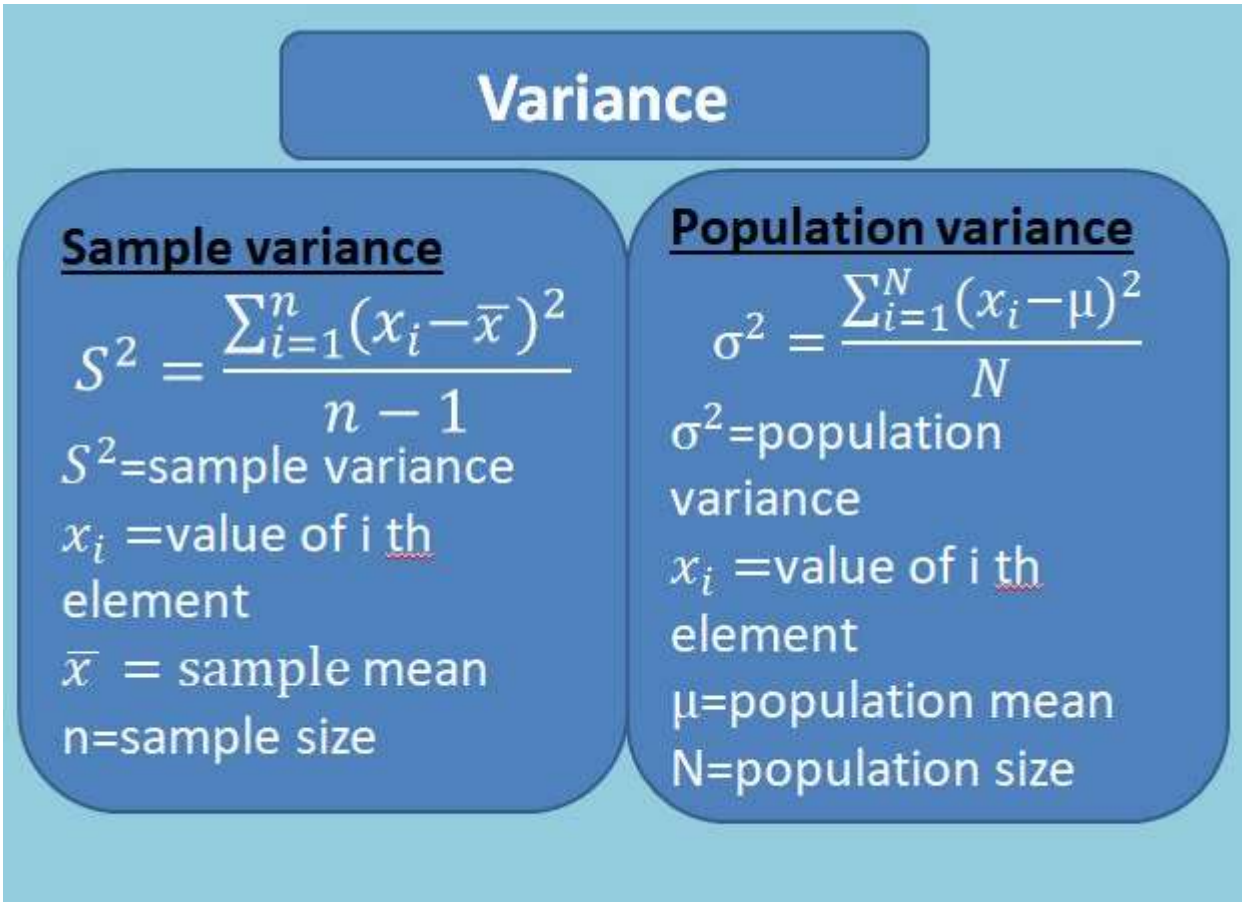
What Is Variance Used for?

Variance is essentially the degree of spread in a data set about the mean value It shows the amount of variation that exists among the data points. Visually, the larger the variance, the "fatter" a probability distribution will be. In finance, if something like an investment has a greater variance, it may be interpreted as more risky or volatile.

Why Is Standard Deviation Often Used More Than Variance?

Standard deviation is the square root of variance. It is sometimes more useful since taking the square root removes the units from the analysis. This allows for direct comparisons between different things that may have different units or different magnitudes.

So why do we subtract 1 when calculate the sample variance or standard deviation?



What is Bessel's Correction?

Bessels' correction refers to the "n-1" found in several formulas, including the sample variance and sample standard deviation formulas. This correction is made to correct for the fact that these sample statistics tend to underestimate the actual parameters found in the population.

As we can see, SAMPLE formulas have *n-1* in the denominator, where *n* is the sample size.

So why do we subtract 1 when using these formulas?

The simple answer: the calculations for both the sample standard deviation and the sample variance both contain a little bias (that's the statistics way of saying "error"). Bessel's correction (i.e. subtracting 1 from your sample size) corrects this bias. In other words, **you'll usually get a more accurate answer if you use *n-1* instead of *n*.**

Example calculations of standard deviation.

Calculate the standard deviation for the sample age data of the 15 students: -

22	23	25	27	28	35	32	28	30	40	24	26	27	29	31

Calculations: -

STEP – 1: Calculate the Mean

- The mean age of the 15 students is,
- Mean (Age) = (22 + 23 + 25 + 27 + 28 + 35 + 32 + 28 + 30 + 40 + 24 + 26 + 27 + 29 + 31) / 15 => (427 / 15)
- **Mean (Age) = 28.47**

STEP – 2: Calculate the Standard Deviation

- Let X be the Age of the 15 Students Sample,
- Then the Standard Deviation of sample X is,

$$s = \sqrt{\frac{\sum(X - \bar{x})^2}{n - 1}}$$

- Let us calculate the standard deviation.

X	$X - \bar{x}$	$(X - \bar{x})^2$
22	-6.47	41.86
23	-5.47	29.92
25	-3.47	12.04
27	-1.47	2.16
28	-0.47	0.22
35	6.53	42.64
32	3.53	12.46
28	-0.47	0.22
30	1.53	2.34
40	11.53	132.94
24	-4.47	19.98
26	-2.47	6.1
27	-1.47	2.16
29	0.53	0.28
31	2.53	6.4
	$\sum(X - \bar{x})^2$	**311.72**

The total number of observations, n = 15.

$$\sum(X - \bar{x})^2 = 311.72, \text{ n} = 15$$

$$s = \sqrt{\frac{311.72}{15-1}} = \sqrt{\frac{311.72}{14}} = \sqrt{22.47}$$

$$s = 4.72$$