STATISTICS - A Fun World! Part -I

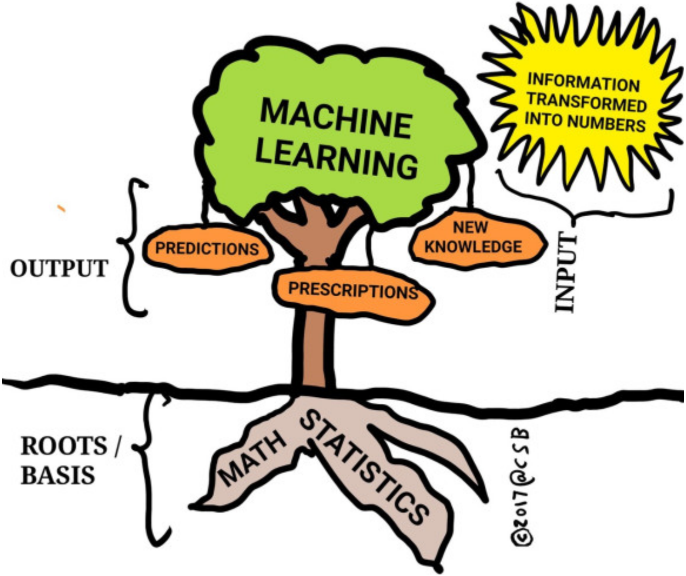
Premanand S
Machine Learning Enthusiast

May 16, 2021

Roadmap

- Is Statistics essential for AI | ML | DL
- Statistics
- Statistics - History
- Basic Terminologies - Statistics
- Types of Statistics
- Descriptive Statistics
- Inferential Statistics
- Sampling Techniques
- Exploratory Data Analysis
- Probability
- Probability Distribution
- Types of Probability
- Hypothesis Testing - Puzzling one!

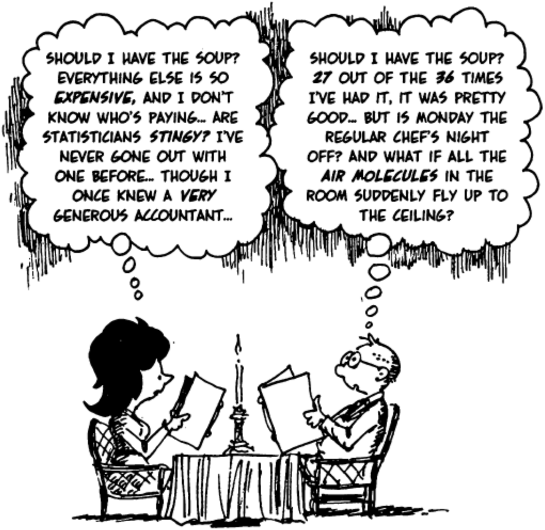
Is Statistics essential for AI | ML | DL



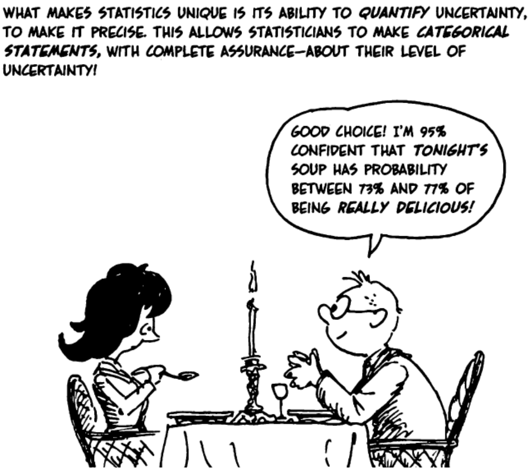
Need for Statistics

- Statistics : Problem $->$ Need to get DATA to solve
- Machine Learning or Deep Learning : DATA $->$ Need outcomes

Statistics - An Intro


SHOULD I HAVE THE SOUP? EVERYTHING ELSE IS SO EXPENSIVE, AND I DON'T KNOW WHO'S PAYING... ARE STATISTICIANS STINGY? I'VE NEVER GONE OUT WITH ONE BEFORE... THOUGH I ONCE KNEW A VERY GENEROUS ACCOUNTANT...
SHOULD I HAVE THE SOUP? 27 OUT OF THE 36 TIMES I'VE HAD IT, IT WAS PRETTY GOOD... BUT IS MONDAY THE REGULAR CHEF'S NIGHT OFF? AND WHAT IF ALL THE AIR MOLECULES IN THE ROOM SUDDENLY FLY UP TO THE CEILING?

Statistics - An Intro


WHAT MAKES STATISTICS UNIQUE IS ITS ABILITY TO QUANTIFY UNCERTAINTY, TO MAKE IT PRECISE. THIS ALLOWS STATISTICIANS TO MAKE CATEGORICAL STATEMENTS, WITH COMPLETE ASSURANCE—ABOUT THEIR LEVEL OF UNCERTAINTY!
GOOD CHOICE! I'M 95% CONFIDENT THAT TONIGHT'S SOUP HAS PROBABILITY BETWEEN 73% AND 77% OF BEING REALLY DELICIOUS!

Statistics - An Intro!

- '*status*' - latin - political state
- '*statista*' - Italian - government
- Science of learning from data
- Helps you to use the proper methods like,
  - employ the correct analyses through analysis and interpretation
  - effectively present the results
- Three disciplines - Statisticians rely on,
  - Data Analysis
  - Probability
  - Statisitcal Inference

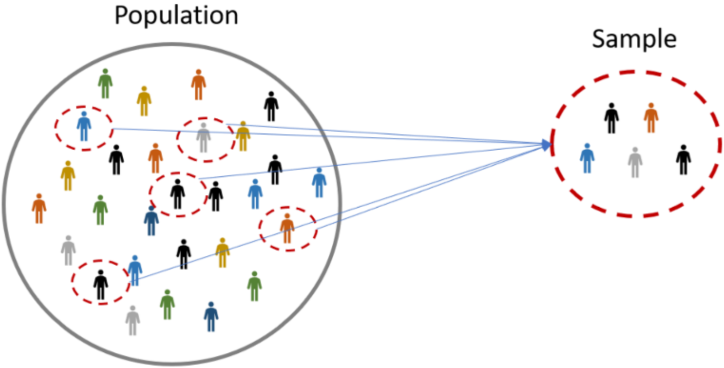
Statistics - History

- In 5th Century B.C - Athenians estimated the height of ladders necessary to scale the walls of Platea
- In 801 - 873 A.D - Al-Kindi - Manuscript on Deciphering Cryptographic Messages
- In 1532, Sir W. Petty - First weekly data on deaths in London
- In 1539, Start of data collection on baptisms, marriages, and deaths in France
- In 1662, J. Graunt, First published demographic study based on bills of mortality
- and many more. **If you still interested means hit me!**

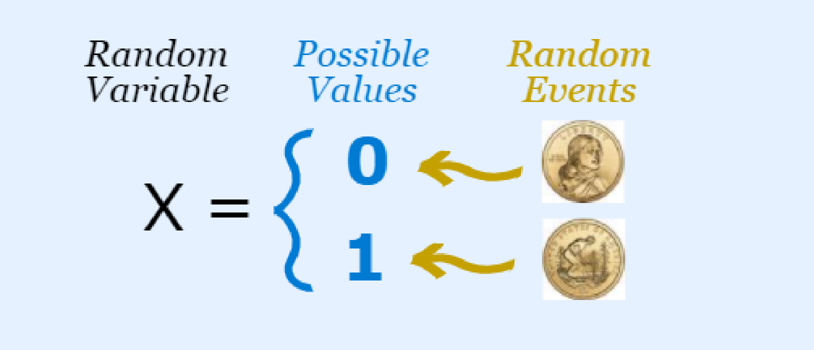
Basic Terminologies - Statistics

- Population
- Sample
- Random Variables

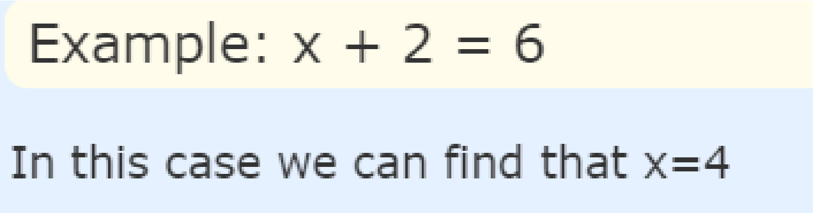
Population and Samples



Random Variables



Variables



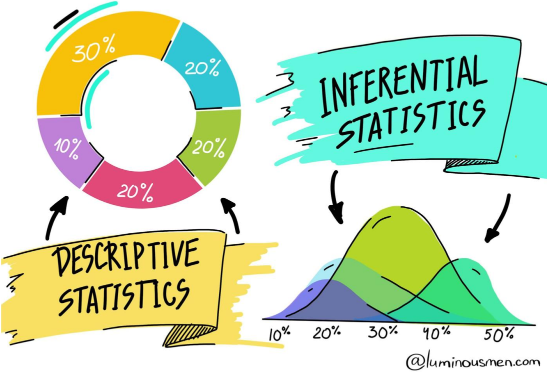
Example: $x + 2 = 6$

In this case we can find that $x=4$

Outliers

- Values that "**lie out**side" the other values
- Extreme values in the data
- Causes,
  - Experimental measure
  - Sampling problem
  - Natural variation
- Methods to identify,
  - Sorting the data
  - Graphical method (Boxplot, Scatter Plot)
  - Using Z score
  - Using IQR

Types of Statistics



- Descriptive Statistics - Summarizing (numbers), Organizing data in the form of visualization (graphs)
- Inferential Statistics - Drawing conclusion (through some tests) about parameter on basis of statistical inference

Descriptive Statistics

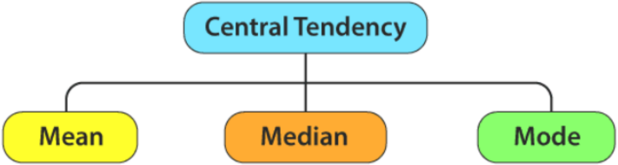
- Useful because they allow you to make sense of the data
- Helps exploring and making conclusion about the data in order to make rational decisions
- Includes calculating things such as average of the data, its spread and the shape it produces

Descriptive Statistics - Types

- Measure of Central Tendency
- Measure of Spread
- Measure of Shape or Measure of Asymmetry

Measure of Central Tendency

- Provide an exact representation of the entire collected data



Mean

- Arithmetic Mean : If values have the same units (normal)
- Geometric Mean : If values have differing units (nth root)
- Harmonic Mean: If the data values are ratios of two variables with different measures, called rates (reciprocal)
- Condition prefer: Symmetric distribution, Continuous data

Arithmetic mean

$$\frac{1}{n} \cdot \sum_{i=1}^{n} a_i$$

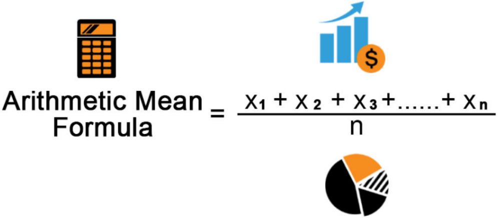
Geometric mean

$$\left(\prod_{i=1}^{n} a_i \right)^{\frac{1}{n}}$$

Harmonic mean

$$\left(\frac{1}{n} \cdot \sum_{i=1}^{n} a_i^{-1} \right)^{-1}$$

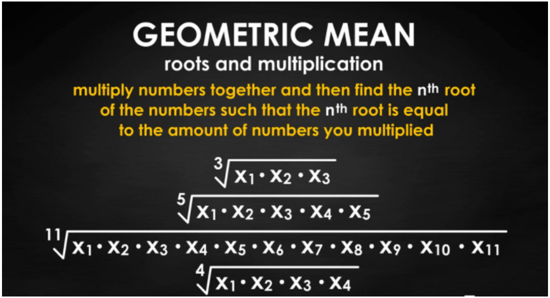
Arithematic Mean



$$\text{Arithmetic Mean Formula} = \frac{X_1 + X_2 + X_3 + \ldots\ldots + X_n}{n}$$

- Adding the numbers together and then dividing by the amount of numbers you were adding

Geometric Mean



GEOMETRIC MEAN

roots and multiplication

multiply numbers together and then find the n^th root of the numbers such that the n^th root is equal to the amount of numbers you multiplied

$$\sqrt[3]{x_1 \cdot x_2 \cdot x_3}$$

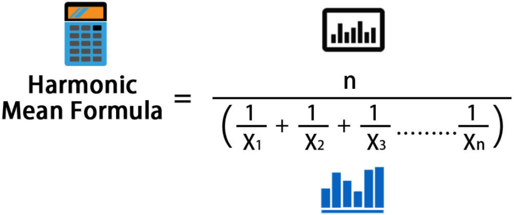
$$\sqrt[5]{x_1 \cdot x_2 \cdot x_3 \cdot x_4 \cdot x_5}$$

$$\sqrt[11]{x_1 \cdot x_2 \cdot x_3 \cdot x_4 \cdot x_5 \cdot x_6 \cdot x_7 \cdot x_8 \cdot x_9 \cdot x_{10} \cdot x_{11}}$$

$$\sqrt[4]{x_1 \cdot x_2 \cdot x_3 \cdot x_4}$$

- When we are trying to calculate where growth is determined by multiplication, not addition

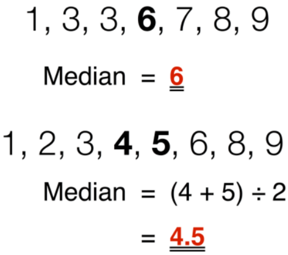
Harmonic Mean



$$\text{Harmonic Mean Formula} = \frac{n}{\left(\frac{1}{X_1} + \frac{1}{X_2} + \frac{1}{X_3} \dots\dots\dots \frac{1}{X_n}\right)}$$

- Used in specific situations or when dealing with averages of units, like average travel speed, area of finance to calculate price multiples like price-earnings ratio, price-sales ratio, etc.
- most appropriate when the set of numbers contains outliers that might skew the result

Median



1, 3, 3, **6**, 7, 8, 9

$$\text{Median} = \underline{\underline{6}}$$

1, 2, 3, **4**, **5**, 6, 8, 9

$$\begin{aligned}\text{Median} &= (4 + 5) \div 2 \\ &= \underline{\underline{4.5}}\end{aligned}$$

- Middle value of the dataset in which the dataset is arranged in the ascending order or in descending order
- Outliers and skewed data have a smaller effect on the median
- Condition prefer:Skewed distribution, Continuous data, Ordinal data

Mode

- Frequently occurring value in the dataset
- Sometimes the dataset may contain multiple modes and in some cases, it does not contain any mode at all
- Having two modes is called bimodal.
- Having more than two modes is called multimodal
- Find the mode for continuous data by locating the maximum value on a probability distribution plot
- Condition prefer: Categorical data, Ordinal data, Count data, Probability Distributions

Verdict: Measure of Central Tendency

- If you have a symmetrical distribution of continuous data, all the three measures of central tendency hold good. But most of the times, the analyst uses the mean because it involves all the values in the distribution or dataset
- If you have skewed distribution, the best measure of finding the central tendency is the median
- If you have the original data, then both the median and mode are the best choice of measuring the central tendency
- If you have categorical data, the mode is the best choice to find the central tendency

mail me: er.anandprem@gmail.com
ring me: +91 73586 79961
follow me: Linkedin
Website: tango-learning

Learning gives Creativity, Creativity leads to Thinking, Thinking provides Knowledge, and Knowledge makes you Great - Dr APJ Abdul Kalam