

# AI235AT Statistics for Data Science

*Dr. Arulalan Rajan,*

*Founder & Director, vidyākośā, Bangalore.*

*Faculty, Proficiency Programme, CCE, IISc, Bangalore.*

*Adjunct Professor, Dept of AIML, RVCE, Bangalore.*

January 1, 2024

# Statistics

- ▶ What?
- ▶ Why?
- ▶ How?

# Course Contents<sup>1</sup>

- ▶ **Descriptive Statistics** - Describing data sets - Frequency tables and graphs, relative frequency tables and graphs, Grouped data, histograms, Summarising data sets - Sample Mean, sample median, sample mode, sample variance and sample standard deviation, percentiles and box-plots
- ▶ **Sampling and Sampling Distributions** - Types of sampling, Sample Mean, Sample Variance, Sampling distributions from a normal population, sampling from a finite population, Normal Distribution, approximating binomial, Poisson distributions using normal distribution
- ▶ **Correlation, Covariance and Independent Random Variables** - Joint behavior of random variables, Correlation, Covariance, variance-covariance matrix, Independent random variables, Sums of independent random variables, Law of Large Numbers, CLT

<sup>1</sup>The Instructor acknowledges authors of various articles available on the web and other resources from which some of the materials presented here are taken

## Course Contents . . .

- ▶ **Large Sample Estimation** - Statistical Inference, Types of Estimators, Point estimation - Point estimation of a population parameter, Interval Estimation - Constructing a confidence interval, Large-Sample Confidence Interval for a Population Mean Interpreting the confidence interval, Large sample confidence interval for a population proportion, Estimating the difference between two population means, Estimating the difference between two binomial distributions, One-sided confidence bounds, Choosing the Sample size.
- ▶ **Hypothesis Testing** - Testing of hypothesis about population parameters, Statistical Test of hypothesis, A large-sample test about the population mean - Essentials of the test, calculating the p-value, two types of errors, power of a statistical test, A large-sample test of hypothesis for the difference between two population means - Hypothesis testing and confidence intervals, Hypothesis testing for the binomial, Some comments on testing of

## About the course

[<+>]

- ▶ Instructors - Dr. Prasad Sudhakar (GE Healthcare) & Dr. Arulalan Rajan
- ▶ Teaching Assistant for the course: Abhijith Kamath (PhD@ IISc)
- ▶ Mentors from your senior batch
- ▶ Problem Solving & Assignments
- ▶ **Focus - Concepts and making you think and reason out logically**
- ▶ **Exams will focus on testing your concepts and NEVER DESCRIPTIVE and COMPUTATION INTENSIVE**
- ▶ *Natural Intelligence what makes Artificial Intelligence work* - So, don your thinking hat!!!
- ▶ Notes are what you write down in class. Carry pens, papers or notebooks, tablets with stylus

## Textbooks and References


- ▶ Sheldon M. Ross, Introduction to Probability and Statistics for Engineers and Scientists, 5th Edition, Academic Press, 2014
- ▶ David Freedman, Robert Pisani and Roger Purves, Statistics, 4th Edition, Norton Company, 2007
- ▶ Richard A. Johnson, Miller Freund's - Probability And Statistics For Engineers, 9th Edition, Pearson, 2018
- ▶ William Mendenhall, R J Beaver, B M Beaver, Introduction to Probability and Statistics - Cengage Learning, 2019

So there we go!!!

Welcome to the course!!!

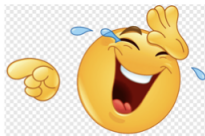


## Some famous quotes on Statistics



The statistics on sanity are that one out of every four Americans is suffering from some form of mental illness. Think of your three best friends. If they're okay, then it's you

There are lies,  
Damned lies and  
Statistics





## Few Questions . . .

- ▶ How/Why did you choose BTech/BE?
- ▶ How/Why did you choose RVCE?
- ▶ How/Why did you choose AIML?
- ▶ IPL and . . .
- ▶ Ola, Uber and Google Maps - Bangalore Traffic
- ▶ Purchases on Amazon, Flipkart etc!

# Role of Statistics

- ▶ Analyse and model data
- ▶ Extremely important in making informed decisions
- ▶ Allow us to answer critical questions like
  - ▶ why things happen?
  - ▶ when things happen?
  - ▶ can we predict reoccurrence?

## How does Statistics do all those mentioned before?

- Presents facts in numerical figures

Most wins in ODI this decade (2010-2019) (min 100 matches played)								
Team	Mat	Won	Lost	Tied	NR	Highest Total	Lowest total	Win %
India	249	157	79	6	7	418	88	63.05
Australia	216	125	79	1	11	417	74	57.87
England	218	123	82	4	9	481	99	56.42
South Africa	188	114	68	1	5	439	118	60.64
Sri Lanka	256	113	127	2	14	377	43	44.14
Pakistan	217	104	106	2	5	399	74	47.93
New Zealand	192	98	82	2	10	398	79	51.04
Bangladesh	162	70	87	0	5	333	58	43.21
West Indies	196	69	114	5	8	389	61	35.20
Afghanistan	123	57	62	1	3	338	58	46.34
Ireland	112	50	55	2	5	331	82	44.64
Zimbabwe	159	44	111	2	2	334	54	27.67

*Courtesy: Hindustan Times, Dec. 23, 2019*

- Questions?

## Some Classic Examples

- ▶ Weather Prediction - Statistical Models compare prior weather with current weather to predict future
- ▶ Bank Loans based on individual application
- ▶ Vaccines
- ▶ Predicting disease based on available data
- ▶ Data you give in Amazon, Flipkart, Myntra or any other websites
- ▶ Facebook, Google, Insta

## Facebook 10 year challenge



Data from the challenge could be used by companies like Facebook or Amazon to train facial recognition algorithms.

# Google . . .

## Your activity



We collect information about your activity in our services, which we use to do things like recommend a YouTube video that you might like. The activity information that we collect may include:

- Terms that you search for
- Videos that you watch
- [Views and interactions with content and ads](#)
- [Voice and audio information](#)
- Purchase activity

## Google Privacy & Terms

[Overview](#)[Privacy Policy](#)[Terms of Service](#)[Technologies](#)[FAQ](#)[Introduction](#)[Information that Google collects](#)[Why Google collects data](#)[Your privacy controls](#)[Sharing your information](#)[Keeping your information secure](#)[Exporting & deleting your information](#)[Retaining your information](#)[Compliance & cooperation with regulators](#)

## We want you to understand the types of information we collect as you use our services

We collect information to provide better services to all our users – from figuring out basic stuff such as which language you speak, to more complex things like which [ads you'll find most useful](#), [the people who matter most to you online](#) or which YouTube videos you might like. The information Google collects, and how that information is used, depends on how you use our services and how you manage your privacy controls.

When you're not signed in to a Google Account, we store the information that we collect with [unique identifiers](#) tied to the browser, application or [device](#) that you're using. This allows us to do things like maintain your preferences across browsing sessions, such as your preferred language or whether to show you more relevant search results or ads based on your activity.

When you're signed in, we also collect information that we store with your Google Account, which we treat as [personal information](#).

# Google . . .

- People with whom you communicate or share content
- Activity on third-party sites and apps that use our services
- Chrome browsing history that you've [synced with your Google Account](#)

If you use our [services to make and receive calls or send and receive messages](#), we may collect call and message log information like your phone number, calling-party number, receiving-party number, forwarding numbers, sender and recipient email address, time and date of calls and messages, duration of calls, routing information and types and volumes of calls and messages.

You can visit your Google Account to find and manage activity information that's saved in your account.



## Google . . .

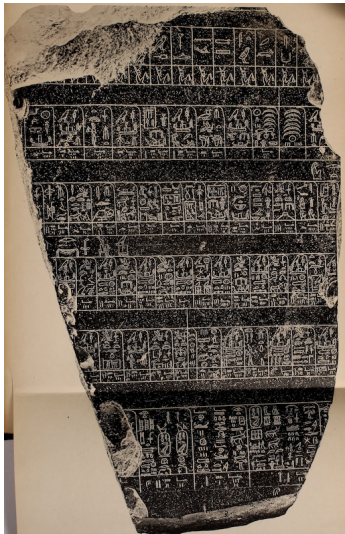
- People with whom you communicate or share content
- Activity on third-party sites and apps that use our services
- Chrome browsing history that you've synced with your Google Account

If you use our services to make and receive calls or send and receive messages, we may collect call and message log information like your phone number, calling-party number, receiving-party number, forwarding numbers, sender and recipient email address, time and date of calls and messages, duration of calls, routing information and types and volumes of calls and messages.

You can visit your Google Account to find and manage activity information that's saved in your account.

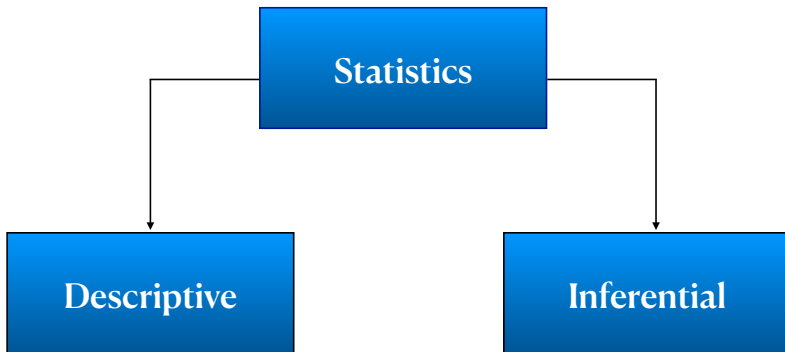
There are over 1.8 billion Gmail users worldwide as of 2023.

# Is Statistics a recent phenomena?



- ▶ 25th Century BC - 4500 Years ago!!!
- ▶ List of Egyptian Pharaohs of the 1st 5 dynasties
- ▶ Records of taxation, religious ceremonies, levels of flood in river Nile
- ▶ Levels of flood in Nile through Nilometers - Used till 1960's
  - ▶ Low water low - Inference?
  - ▶ High level - Inference?

# Statistics - Heart of Data Analytics/Data Science



# Descriptive Statistics

- ▶ Characteristics of data
- ▶ Describe the group characteristics using summary statistics and graphs
- ▶ No Inference, only description
- ▶ Requirement: Sizeable amount of data points - Sample!
- ▶ Obtain insights and visualize data
- ▶ Describe both population and sample
- ▶ Frequency Distribution, Central Tendency, Variability/Dispersion

# Inferential Statistics

- ▶ Make inferences about population based on samples
- ▶ Given a subset of population, how do we draw conclusions about the full set?
- ▶ Inferences based on principles of evidence employ sample statistics
- ▶ Requirement: How accurate the sample data represents the population!
- ▶ Make Predictions and hence involve probability
- ▶ Random Sampling - Essential
- ▶ Regression Analysis, Hypothesis Testing, Confidence Intervals

# Estimates of Location

- ▶ Basic step in exploring data - Obtaining a typical value for each feature
- ▶ Estimate of where most of the data are located - Central tendency
- ▶ Physics Example!!!

- ▶ Mean / Average:  $\bar{x} = \frac{\sum_{i=1}^N x_i}{N}$

## Estimates of Location - Trimmed Mean

- ▶ Trimming gets rid of tails of the distribution

### Trimmed Mean:

$$\bar{x} = \frac{\sum_{i=p+1}^{N-p} x_i}{N - 2p}$$

- ▶ Example: Consider the data: 5, 4, 7, 8, 0, 20, 2, 6, 10, 9
- ▶ Ordered Data: 0, 2, 4, 5, 6, 7, 8, 9, 10, 20
- ▶ The 10% Trimmed Mean =

$$\bar{x}_{0.1} = \frac{2 + 4 + 5 + 6 + 7 + 8 + 9 + 10}{8} = 6.375$$

- ▶ What did we do here?
- ▶ Famous Trimmed Mean - Median - Extreme Trimmed Mean - All observations are removed except one or two

## Estimates of Location - Trimmed Mean



- ▶ Eliminates the influence of extreme values
- ▶ Robust estimators of central tendency -
- ▶ Least affected by outliers

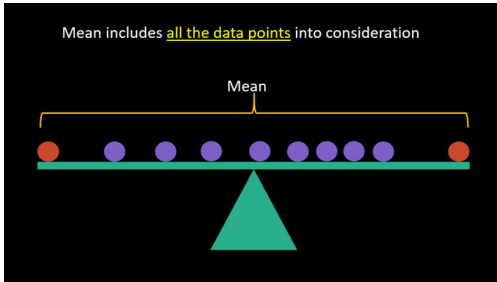


- ▶ Possibility of different amounts of trimming and report the one that gives significant results!

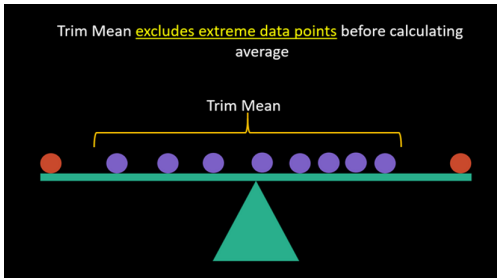


# Mean and Trimmed Mean<sup>2</sup>

Mean includes all the data points into consideration



Trim Mean excludes extreme data points before calculating average



## Estimates of Location - Weighted Mean

- ▶ Average computed by giving different weights to some of the individual values
- ▶ Equal weights imply average or mean

### Weighted Mean:

$$\bar{x}_w = \frac{\sum_{i=1}^N w_i x_i}{\sum_{i=1}^N w_i}$$

- ▶ What did we do here?
- ▶ How do we allot weights?
- ▶ Lower weights to highly varying observations
- ▶ Accounts for data from different groups/sources/methods

## Some Information - Rules of the game!

**Table:** Evaluation Scheme for Statistics and Data Science Course

Components	Marks
Quiz 1 (10 Marks) + Quiz 2 (10 Marks)	20
Test 1 (50 Marks) + Test 2 (50 Marks)	100
Experiential Learning	40
Semester End Exam	100

**Table:** Weights

CIE	SEE
50%	50%

$$\text{CIE} = \text{Quiz 1} + \text{Quiz 2} + 0.4(\text{Test 1} + \text{Test 2}) + \text{EL} = 20 + 40 + 40 = 100$$

## Median and Robust Estimates

- ▶ Sort the data first!
- ▶ Pick the middle number in the sorted data - **Median**
- ▶ If the number of data values is even, median is the average of the two values that divide the sorted data into two halves
- ▶ Median - Robust to outliers
- ▶ Outlier - Skews the result - Value that is very distant / far away from all the other values in a dataset
- ▶ Outlier could result from
  - ▶ data errors like mixing of data
  - ▶ bad measurements!
- ▶ Outliers - Could give better insights! - **Small Data - The Tiny Clues that Uncover Huge Trends, *Martin Lindstrom*, 2016, St. Martin's Press.**

# Estimates of Variability

- ▶ More than location, we often would want to know if the data is clustered about a value or spread out
- ▶ Variability reduction, deciding in the presence of variability, identifying sources of variability etc

## Deviation/ Errors/ Residuals

Difference between the observed values and the estimate of the location

## Estimates of Variability

- Some more definitions

### Variance/Mean Squared Error

Sum of squared deviations from the mean divided by  $N-1$ , where  $N$  is the total number of data points

$$\text{Variance} = s^2 = \frac{\sum_{i=1}^N (x_i - \bar{x})^2}{N - 1}$$

### Standard Deviation/ $l_2$ - norm/ Euclidean norm

Square root of variance =  $s = \sqrt{\text{Variance}}$

- Neither variance nor standard deviation are robust to outliers!
- Sensitive to outliers due to squared deviations!

## Estimates of Variability

### Mean Absolute Deviation / l1 norm/ Manhattan Norm

Mean of absolute value of deviations from the mean

$$\text{Mean Absolute Deviation} = \frac{\sum_{i=1}^N |x_i - \bar{x}|}{N}$$

### Median Absolute Deviation from the Median / MAD

Median of the absolute value of the deviations from the mean

$$\text{MAD} = \text{Median}(|x_1 - m|, |x_2 - m|, \dots, |x_N - m|)$$

# Estimates of Variability

## Range

Difference between the largest and smallest value in a data set

## Order Statistics / Ranks

Metrics based on the data values sorted from the smallest to the largest

## Percentile/Quantile

Value such that  $P\%$  of the values take on this value or less and  $100-P\%$  takes on this value or more

## Interquartile Range / IQR

Difference between 75th percentile and 25th percentile



## Estimates based on Percentiles

- ▶ Dispersion can be estimated from the spread of the sorted data
- ▶ Statistics based on sorted data - **Ordered Statistics**
- ▶ **Range** - Difference between the largest and smallest number
- ▶ Range - Sensitive to outliers and hence not suited as measure of dispersion of the data