



Introduction to  
**STATISTICS**

# Objective of today's class



**“Build basic understanding of Statistics”**

The foundation of Data Analytics and Machine Learning

## Agenda

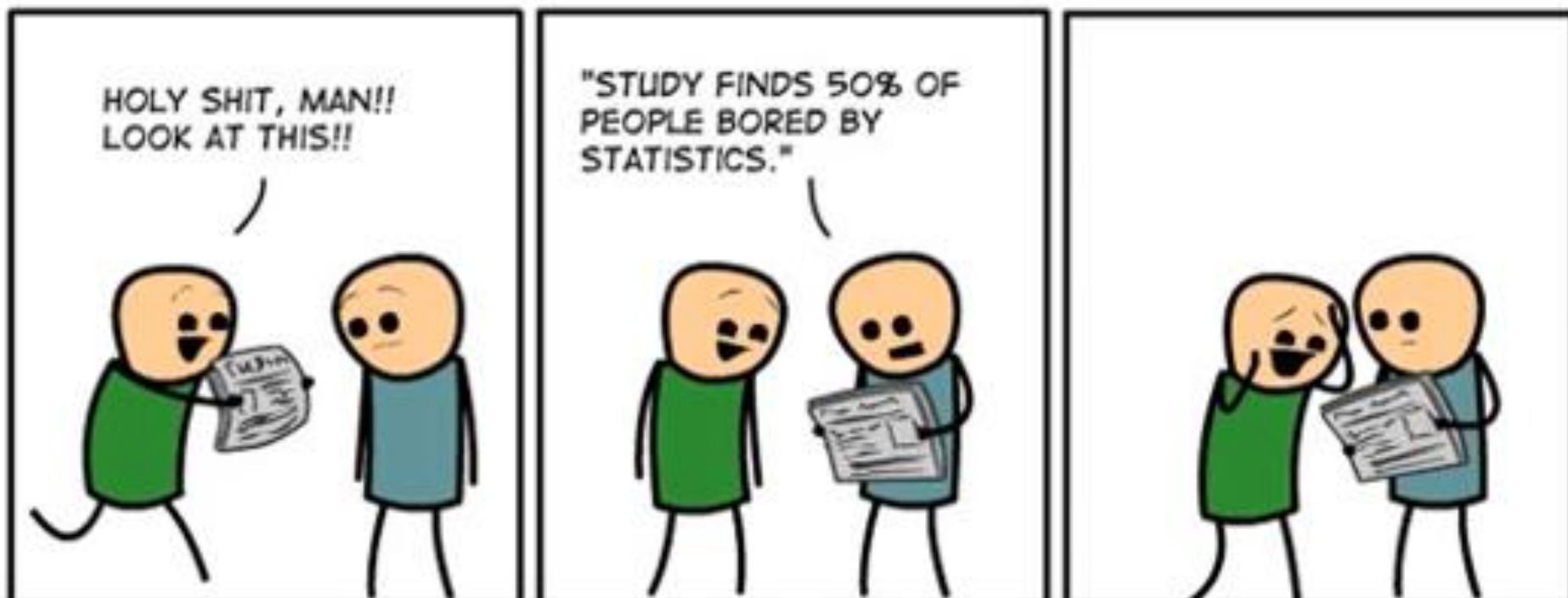
1. Introduction to Statistics
2. Importance of Stats in Data Science & ML
3. Data in Statistics - Types & Sources
4. Types of Statistics
5. Intro to Descriptive Stats
6. Plots

## Introduction to Statistics

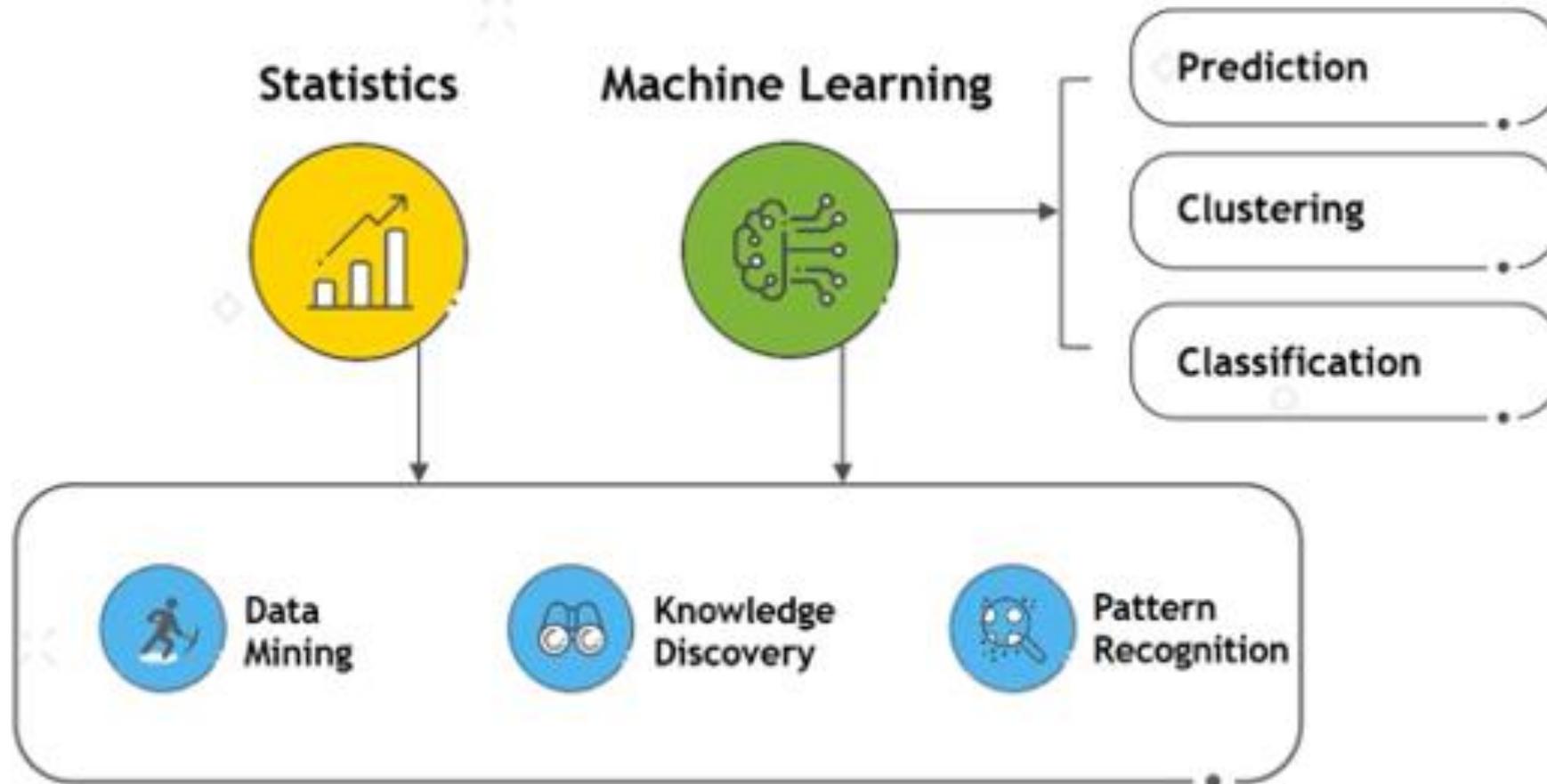


# What is Statistics?

Statistics is the science of conducting studies to collect, organize, summarize, analyze and draw conclusions from the data.



# Importance of Stats in Data Science



# Where is Statistics applied?



## Medicine

- Drug cure rate over 100 patients
- WHO research on epidemic spread
- Measure of improvement in TB cases across the world



## Stock Market

- Understand performance of a stock over time
- Average price of automotive stocks
- Study Infosys stocks over a period of 52 weeks



@amity1415

# Where is Statistics applied?



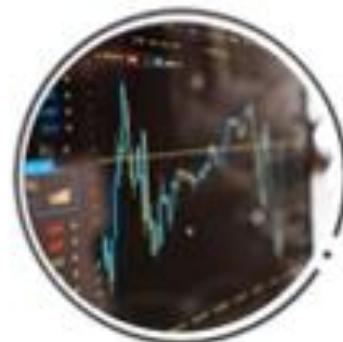
Medicine



Business



Weather  
Forecast



Stock  
Market



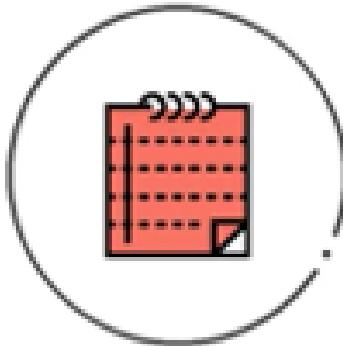
Health &  
Social  
Sciences

## What is Data?

Data is a collection of information such as facts, numbers, words, measurements, observations or even just descriptions of things



# Types of Data



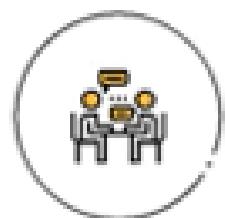
**Primary**



**Secondary**

# Primary Data

Primary Data is data that has been collected first hand by the researcher for addressing the population at hand



Interviews



Observation



Questionnaires



Case Studies



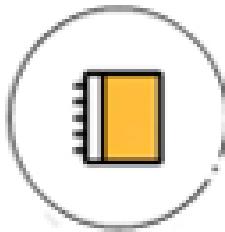
Focused Group  
Discussions

# Secondary Data

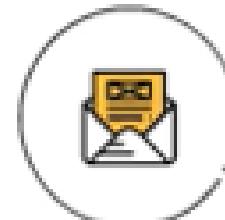
Secondary data is the data that has been already collected by and readily available from other sources



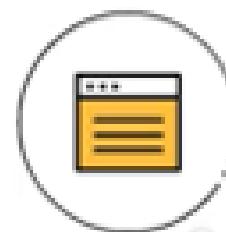
Previous  
Research



Diaries



Letters



Web Info



Census  
Data

# Types of Data

## Quantitative

### a. Discrete



Two horses



### b. Continuous



Height

## Qualitative

### a. Nominal



Male



Female

### b. Ordinal



Customer Service

## Interval



Time scale

## Ratio

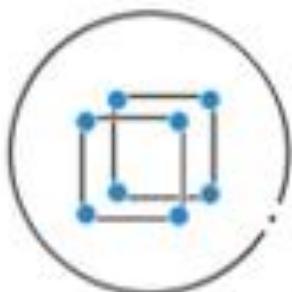


Weight

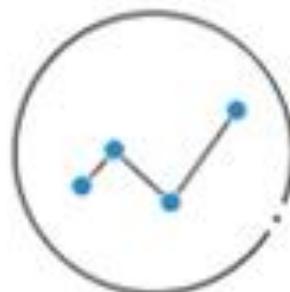
@amity1415

# Data Issues, Population and Sample

# Data Quality Issues



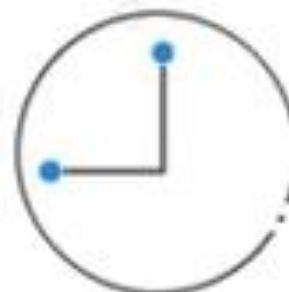
**Duplicity**  
Redundancy leading to resource wastage



**Inconsistency**  
Withdrawal of INR 10/- not reflecting in Net Banking



**Correctness**  
Age/Income as a negative number



**Timeliness**  
Feedback forms given to students from instructor



**Missing values**  
Stock prices risen, but displaying low on front-end

# Population and Sample

We want to know about this



$\mu$  Population Mean

$\sigma$  Population Standard Deviation

$\pi$  Population Standard Proportion

Parameter

Random Selection

We use this



$\bar{x}$  Sample Mean

$s$  Sample Standard Deviation

$p$  Sample Standard Proportion

Statistic

Draw inferences on

# Population and Sample: Case



## Population

The wait time for all 566 cars passing through the drive-thru

CASE



## Sample

The wait time for a subset of 100 cars

## Parameter

Average waiting time for population of 566 cars

## Statistic

Average waiting time for a sample



# Agenda

@amity1415

- 1 Introduction to Statistics
- 2 Importance of Stats in Data Science & ML
- 3 Data in Statistics - Types & Sources
- 4 Types of Statistics
- 5 Intro to Descriptive Stats
- 6 Plots

## Types of Statistics



# Types of Statistics

@amity1415



**Descriptive**



**Inferential**

# Descriptive Statistics

@amity1415

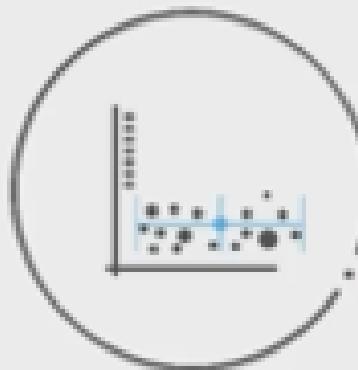


- 1 Measures of Central Tendency
- 2 Measures of Spread
- 3 Measures of Shape

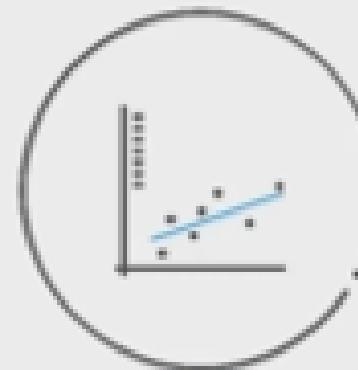
@amity1415

# Measurement of Central Tendency

# Measures of Central Tendency



**Mean**



**Median**



**Mode**

# Mean

Average value in a dataset



99 kg



68 kg



45 kg

$$\text{Mean} = (99 + 68 + 45)/3$$

70.67 kg

# Mean

Apple selects sample of 4 recent placements from 100 new hires



\$ 144,000



\$ 155,000



\$ 98,000



\$ 316,000

$$\text{Sample mean salary} = \bar{x} = \frac{\sum x}{n} = \frac{713,000}{4} = \$ 178,250$$

# Mean affected by extreme values

Mean provides a misleading balance point because of an outlier



\$ 144,000



\$ 155,000



\$ 316,000



\$ 1,000,000

$$\text{Sample mean salary} = \bar{x} = \frac{\sum x}{n} = \frac{1,615,000}{4} = \$ 403,750$$

# Median

The median is a positional average and refers to the middle value in a distribution



28 kg



30 kg



32 kg



38 kg



40 kg

# Median for even size of sample

The median is the average of the middle 2 values



28 kg



30 kg



32 kg



34 kg



38 kg



65 kg

$$\text{Median} = (32 + 34)/2 = 33 \text{ Kg}$$

# Mode

The mode is the most frequently occurring value in the dataset



37 kg



32 kg



32 kg



32 kg



30 kg

# Mode

Order values from least to greatest & locate value that occurs the most

3, 4, 5, 5, 5, 6, 6, 7,  
8, 8, 9

**Unimodal**

The dataset has 1 mode

3, 4, 5, 5, 5, 6, 6, 6,  
8, 8, 9

**Bimodal**

The dataset has 2 modes

1, 2, 3, 4, 5, 6, 7, 8,  
9, 10

**No mode**

The dataset has no mode

**Trimodal**

The dataset has 3 modes

**Multimodal**

The dataset has more than 1 mode

# Pizza table example

Deciding the seating arrangement of a restaurant by figuring out the most frequently occurring group size



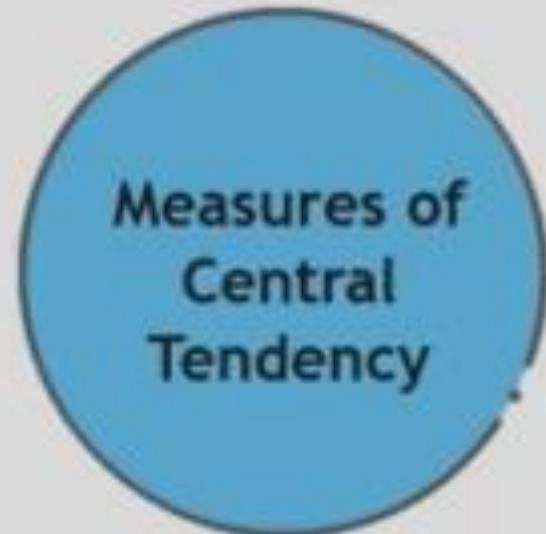
A sample of 20 groups is selected at random:

People=[2, 4, 1, 2, 3, 2, 4, 2, 3, 6, 8, 4, 2, 1, 7, 4, 2, 4, 4, 3]

There are 2 modes, each occurring six times - 2, 4

Basis the above, the manager will decide on 2 seater and 4 seater tables being kept in the patio

# Practical uses



- a Understand data spread around the centre in Data Analysis
- b Replace missing values in Data Analysis
- c Showcase performance of players across their career
- d Study salary spread across the company

# Descriptive Statistics



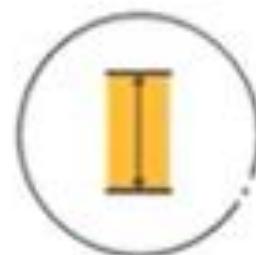
- 1 Measures of Central Tendency
- 2 Measures of Spread
- 3 Measures of Shape

# Measures of Spread

Measures of **variability** provide information about the degree to which individual scores deviate from the average value in a distribution



**Range**



**Variance**



**Interquartile Range**



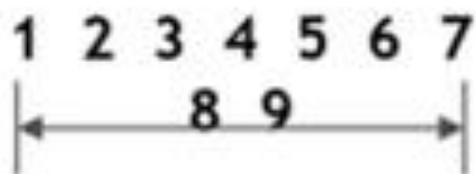
**Standard Deviation**

# Range

Range is the difference between the highest and lowest score in a distribution



Range



# Range: Case

Two plant managers are asked to record their plant production output for five days



PLANT A	PLANT B
15	23
25	26
35	25
20	24
30	27



Plant B

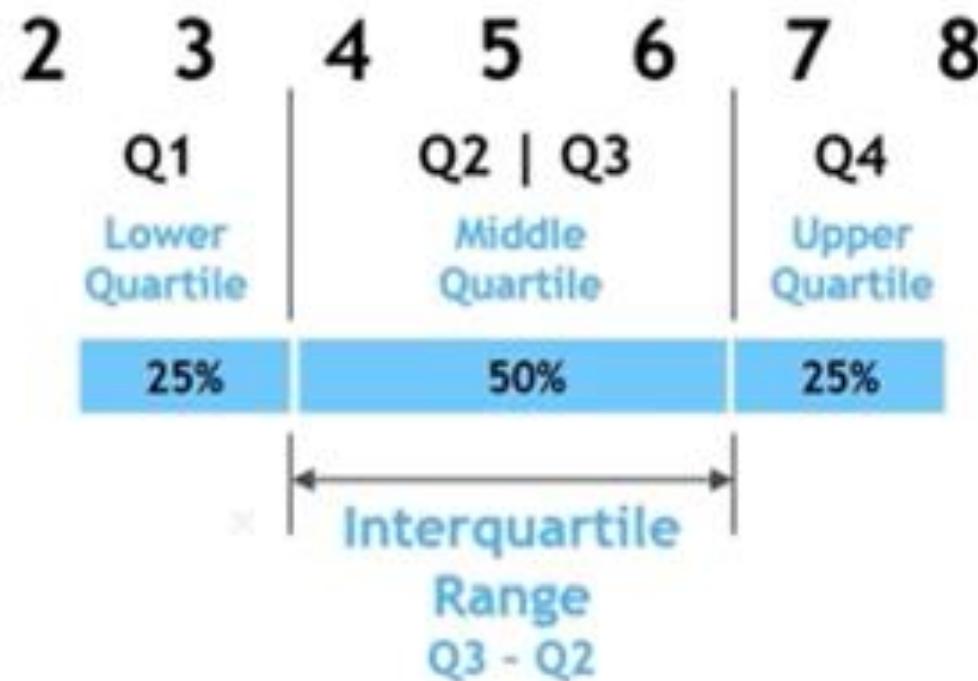
Range,  $R = \text{Maximum} - \text{Minimum}$

$$R = 27 - 23$$

$$R = 4$$

# Interquartile Range

The IQR describes the middle 50% of values when ordered from lowest to highest



## Interquartile Range: Example

1 4 6 12 14 17 20 25 40 41 46

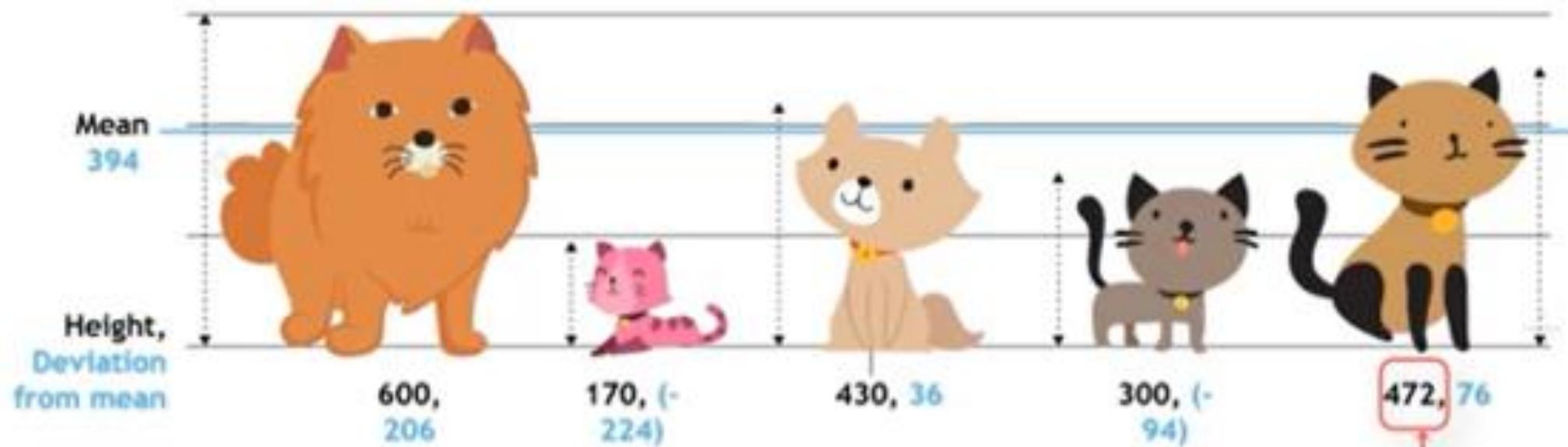
Median = 17

1 4 6 12 14 17 20 25 40 41 46  
Q1 Q2 Q3 Q4

IQR = Q3 - Q1 = 40 - 6 = 34

# Variance

Variance is the average of the squared distances from the mean

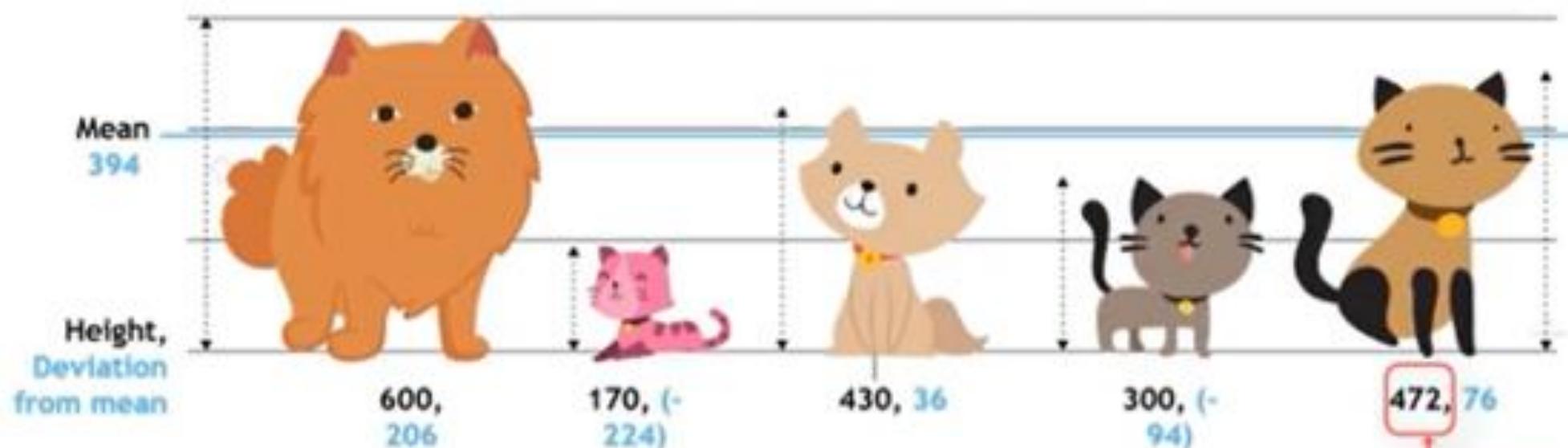


$$\text{Variance} = \frac{\sum 206^2 + 76^2 + (-224)^2 + 36^2 + (-94)^2}{5} = 21704 \text{ mm}^2$$

Note: The value should be 470

# Standard Deviation

Standard Deviation is the square root of variance



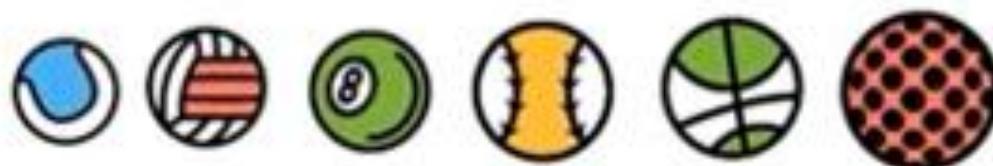
$$SD = \sqrt{21704} = 147.3 \text{ mm}$$

Note: The value should be 470

# Standard Deviation

Standard deviation tells how close the values in a dataset are to the mean

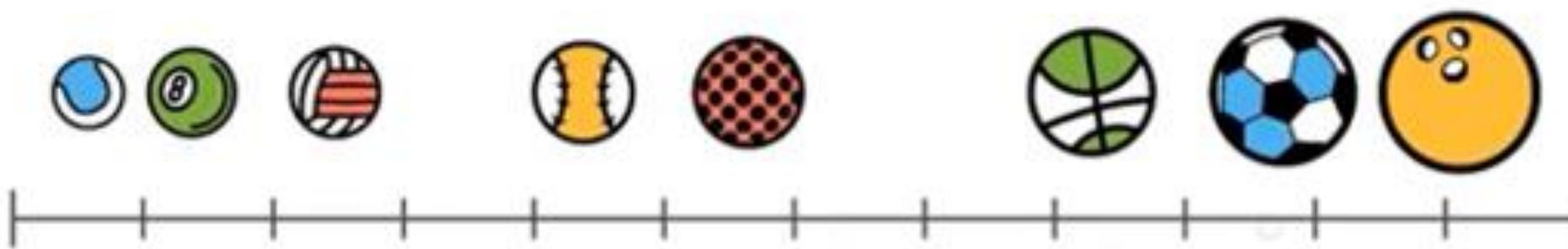
(Data with low standard deviation)



Radius of differently sized balls in set A

# Standard Deviation

(Data with high standard deviation)



Radius of differently sized balls in set B

# Practical uses: Measures of Spread

A measure of spread gives us an idea of how well the mean, for example, represents the data



Smaller the spread, better the mean represents the data



Larger the spread, worse is the mean at representing data

# Practical uses: Measures of Spread



Figure outliers in data



Standardize data and interpret wrt mean



Monitor control systems for not normal behaviour

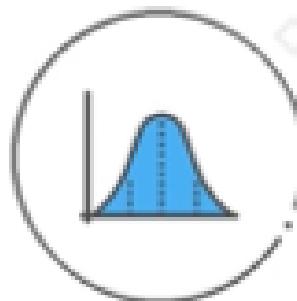
# Descriptive Statistics



- 1 Measures of Central Tendency
- 2 Measures of Spread
- 3 **Measures of Shape**

# Measures of Shape

Measures of shape describe the distribution (or pattern) of the data within a dataset



Symmetric



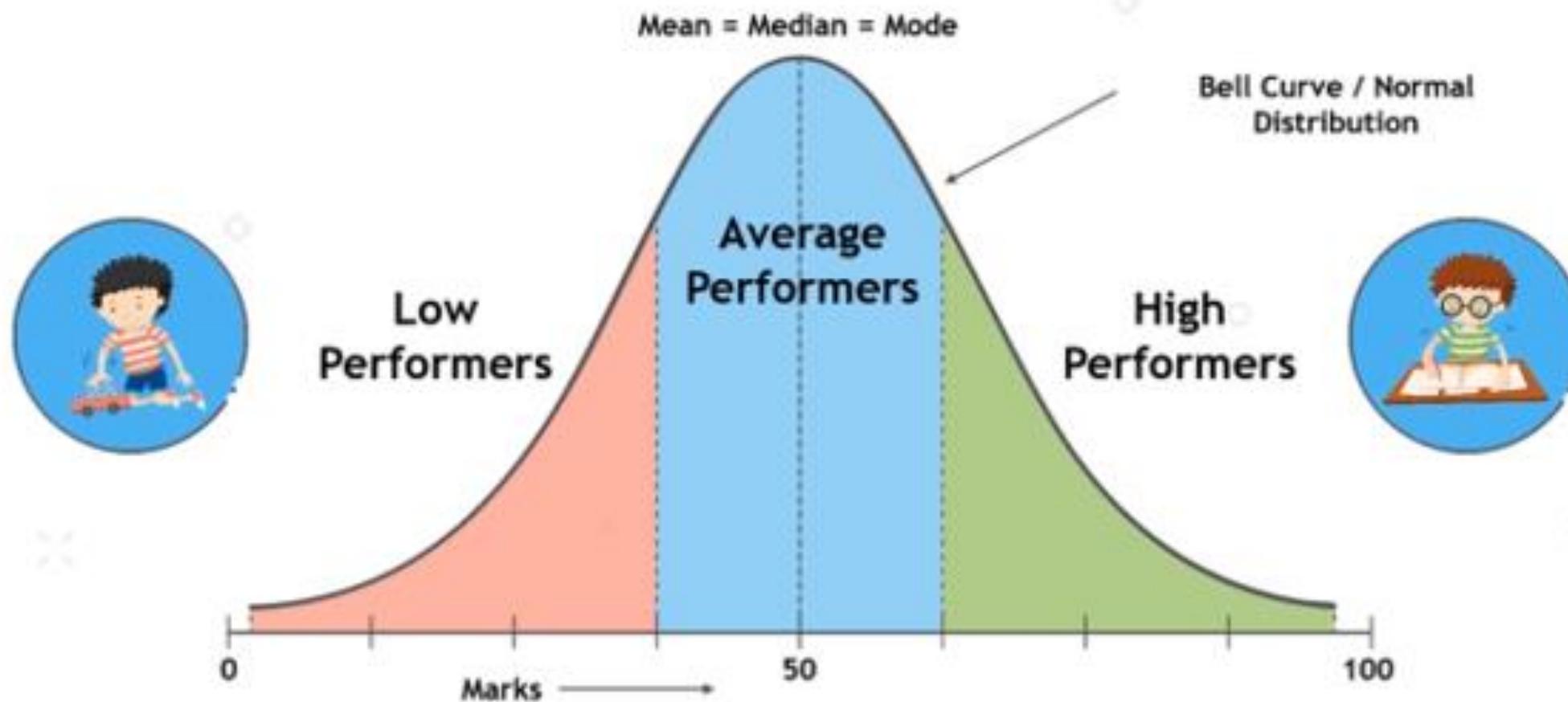
Skewed



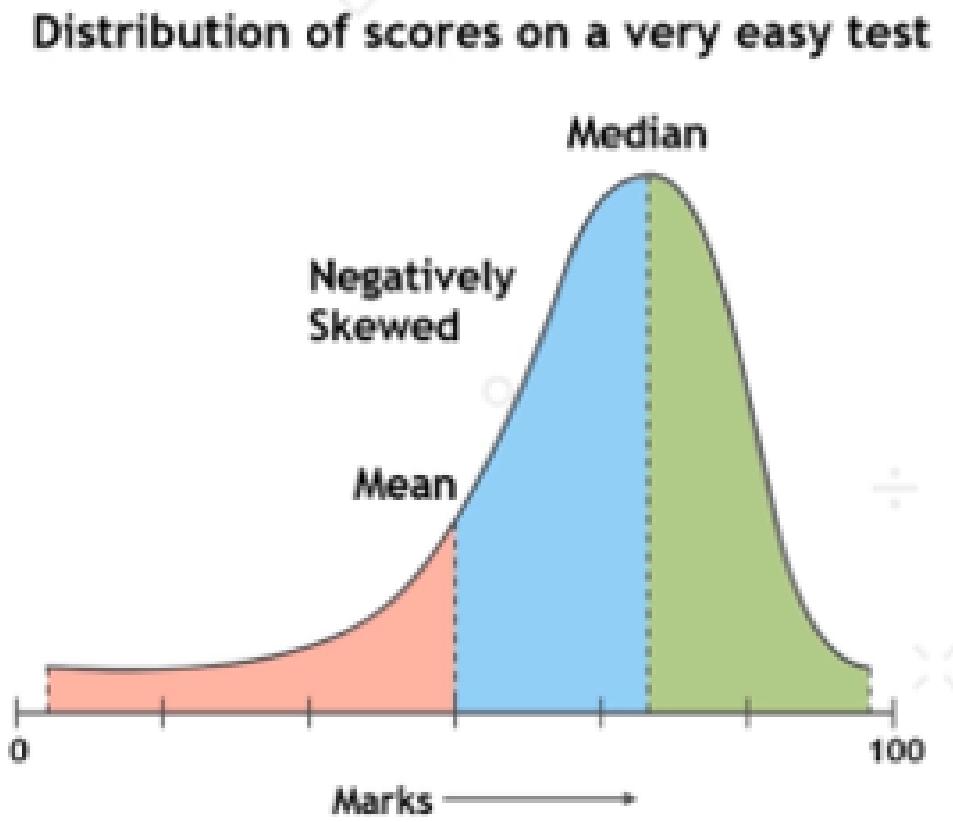
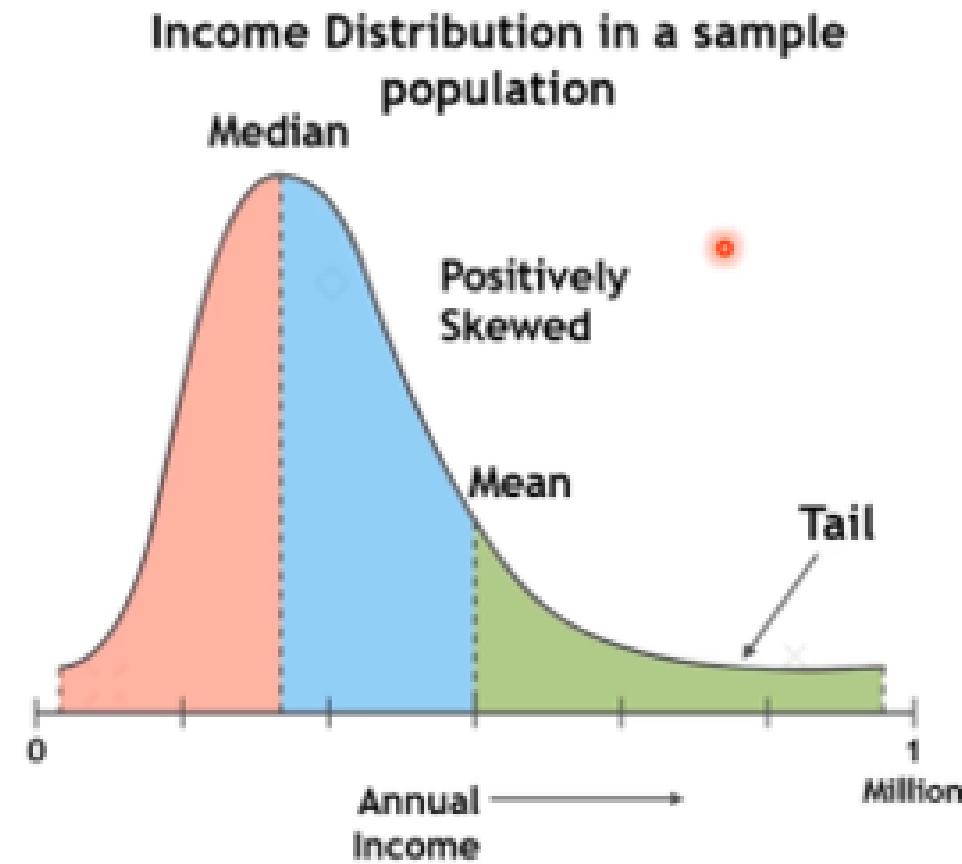
Kurtosis

# Symmetric

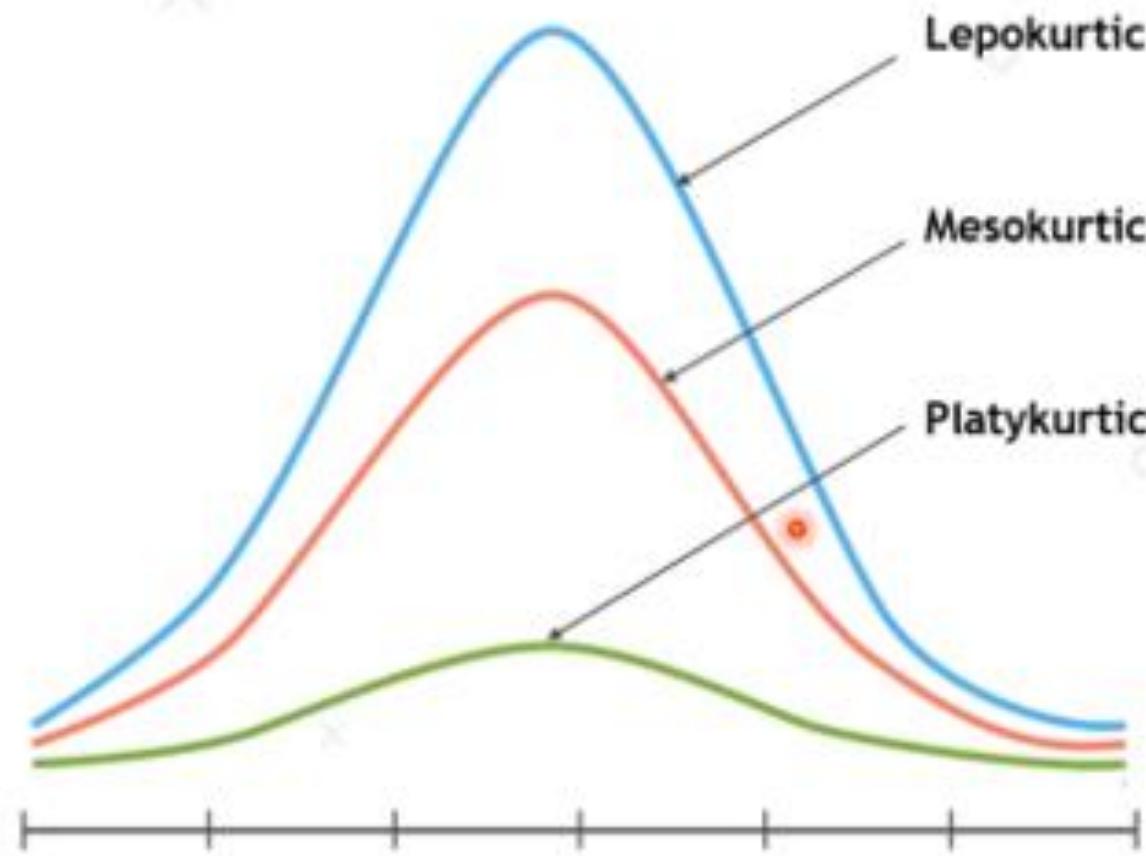
Distribution of marks received by 100 students in a math test



# Skewness

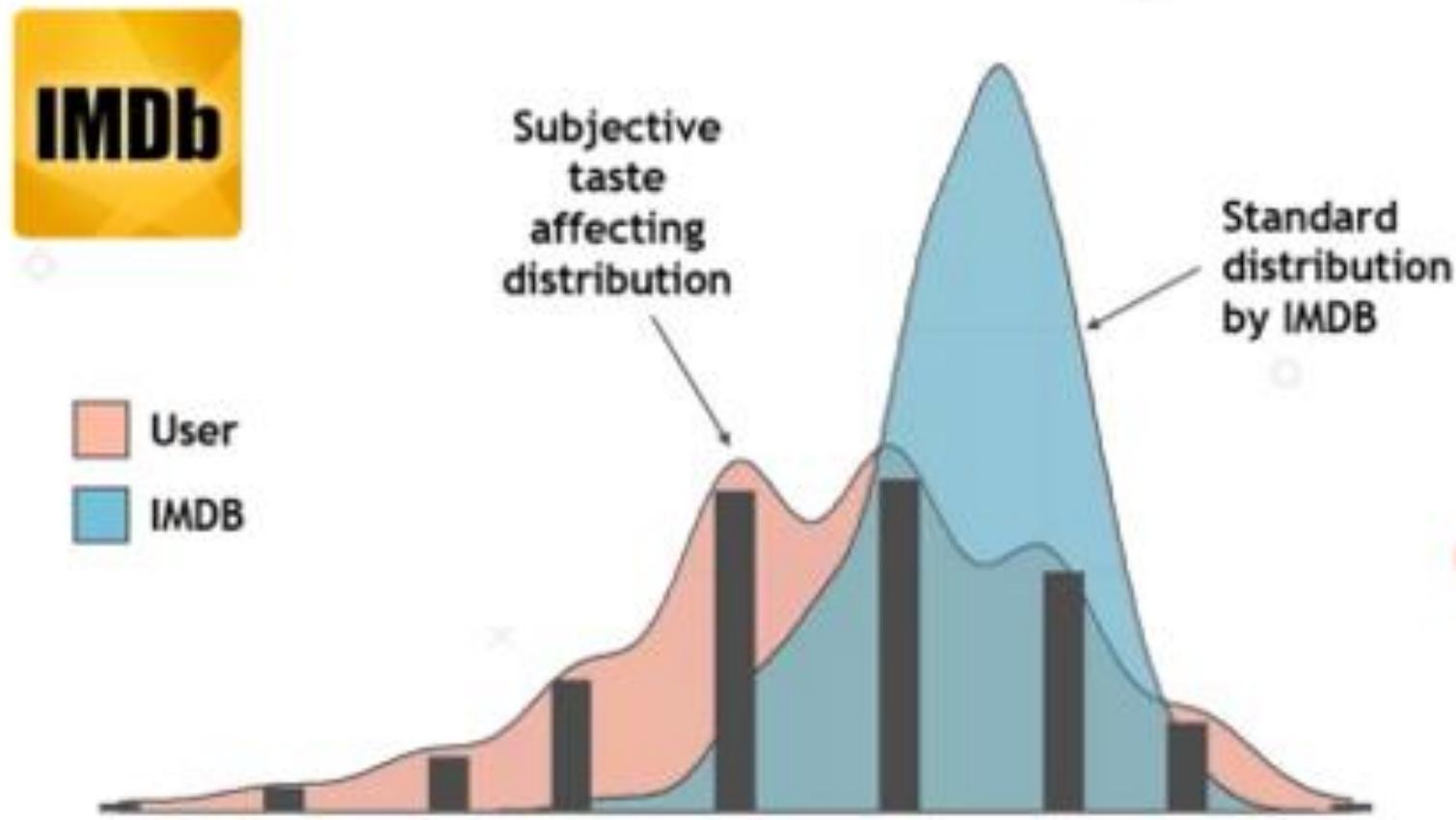


# Kurtosis



# Kurtosis: Example

Comparison of 400+ movie ratings by IMDB and a single user



# Measures of Shape: Practical uses

A measure of the extent to which a frequency distribution is concentrated about its mean

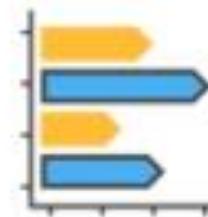
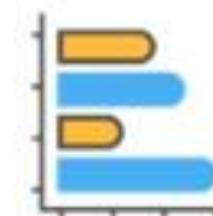


Measures  
outliers (tails)



Study effect  
of  
outliers on the  
overall data

# Plots



# Frequency distribution

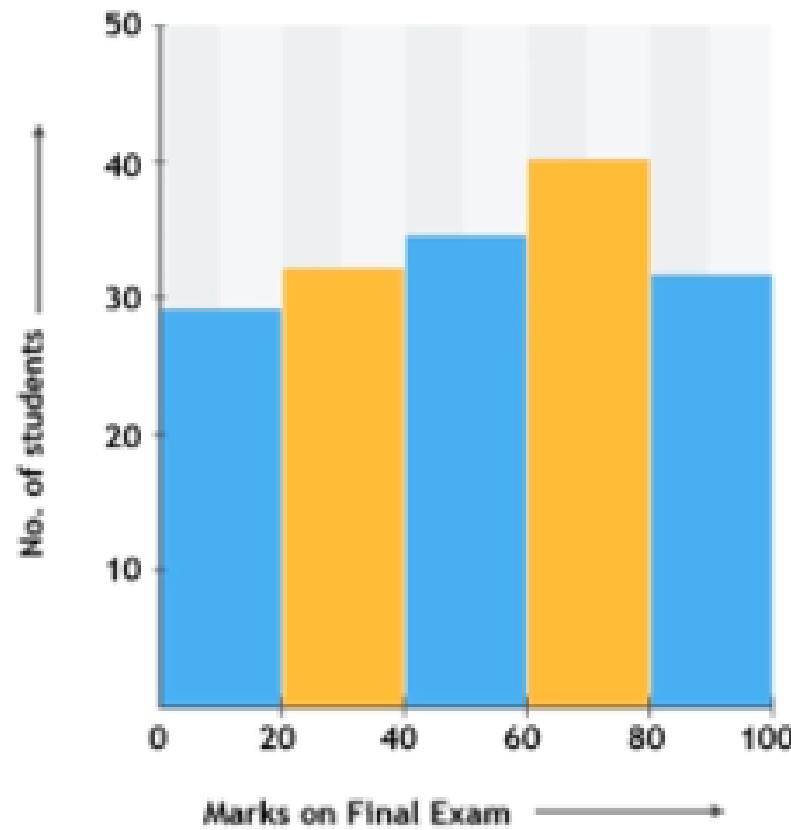
Compiling make of vehicles driven by families in a community

Car Brand	Frequency
Ford	2
Mercedes	2
Toyota	2
Honda	3
Maruti	1
Total	10



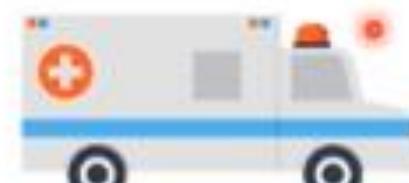
# Histogram

Histograms help us in gaining understanding of data spread

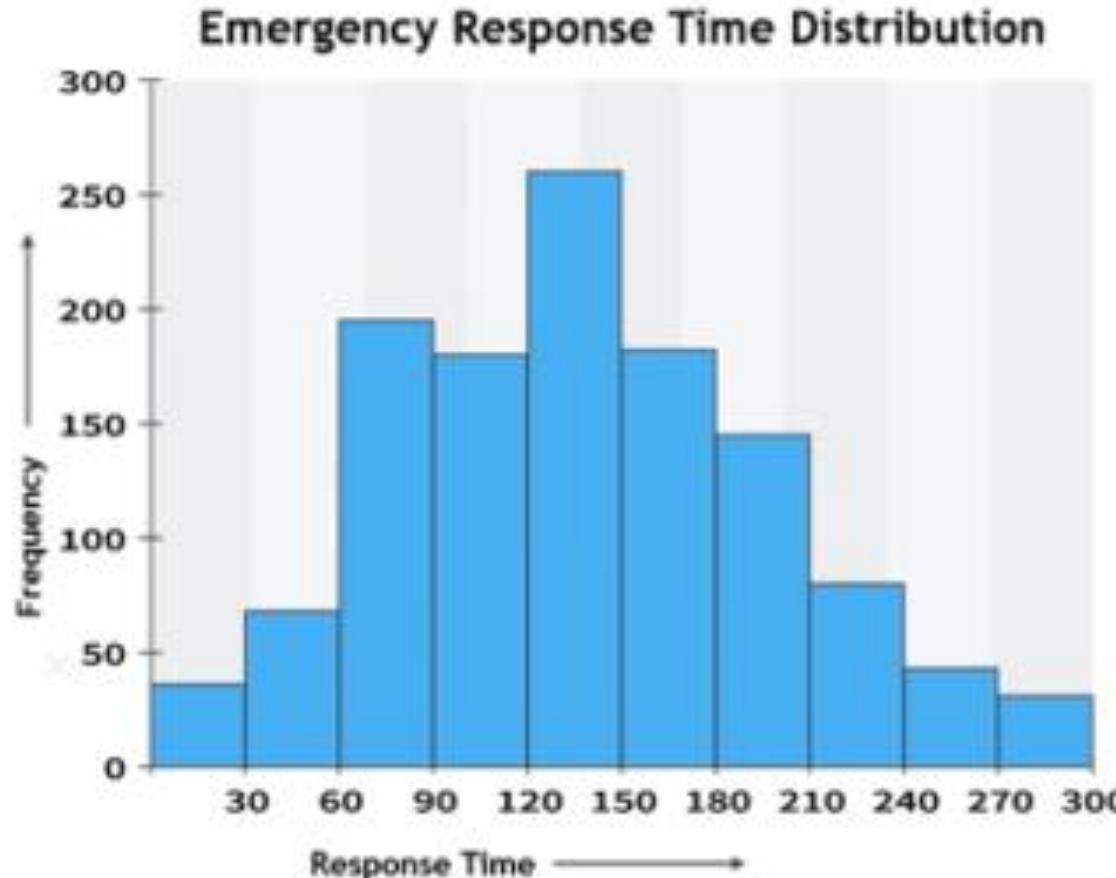


# Histogram: Example

The emergency response times of an ambulance are recorded

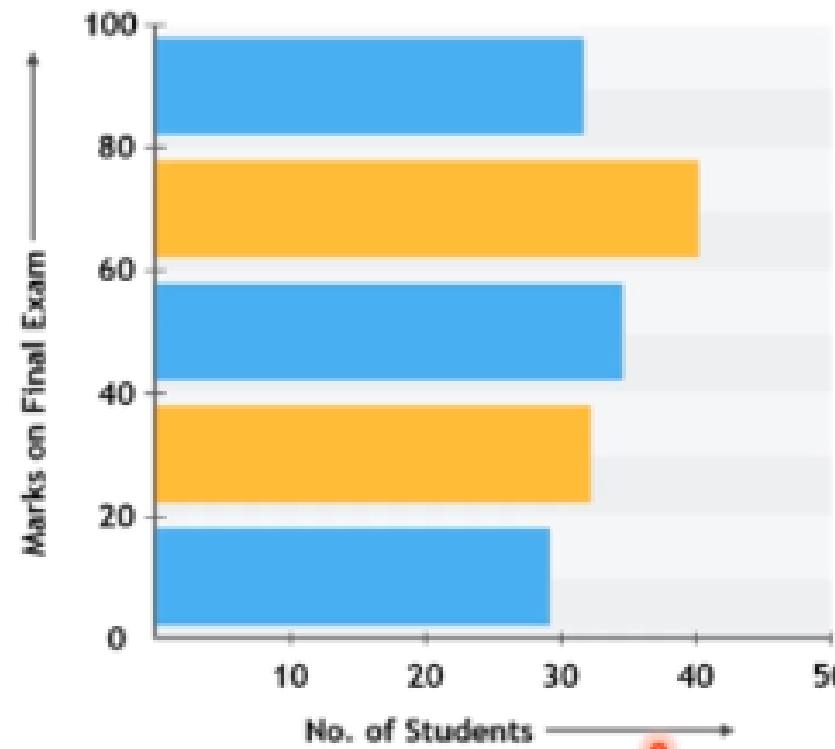


Response Time (sec)	Frequency
0-30	36
30-60	68
60-90	195
90-120	180
120-150	260
150-180	182
180-210	145
210-240	80
240-270	43
270-300	31



# Bar Charts

Bar charts can be vertical or horizontal

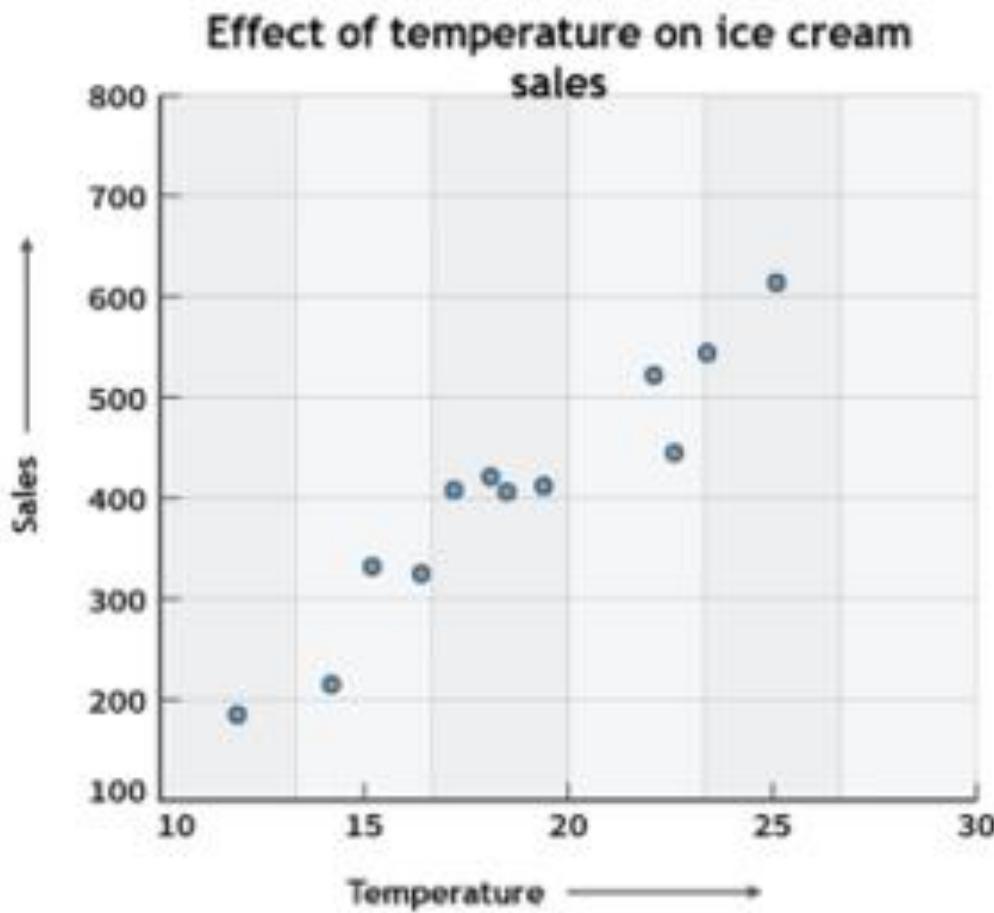


# Scatter plots

Scatter plots show how much one variable is affected by another

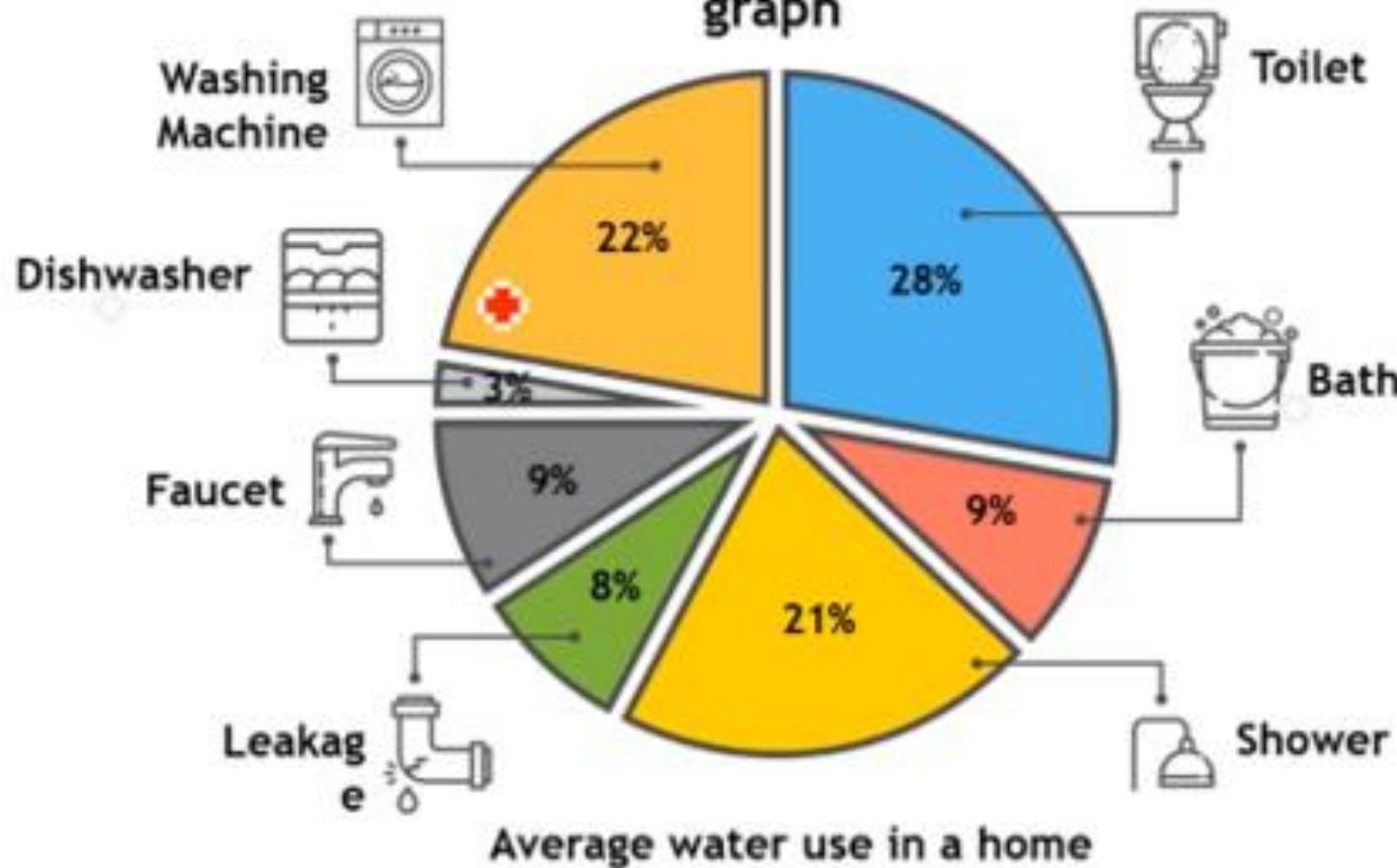


Ice Cream Sales vs Temperature	
Temperature °C	Ice Cream Sales
14.2°	\$215
16.4°	\$325
11.9°	\$185
15.2°	\$332
18.5°	\$406
22.1°	\$522
19.4°	\$412
25.1°	\$614
23.4°	\$544
18.1°	\$421
22.6°	\$445
17.2°	\$408



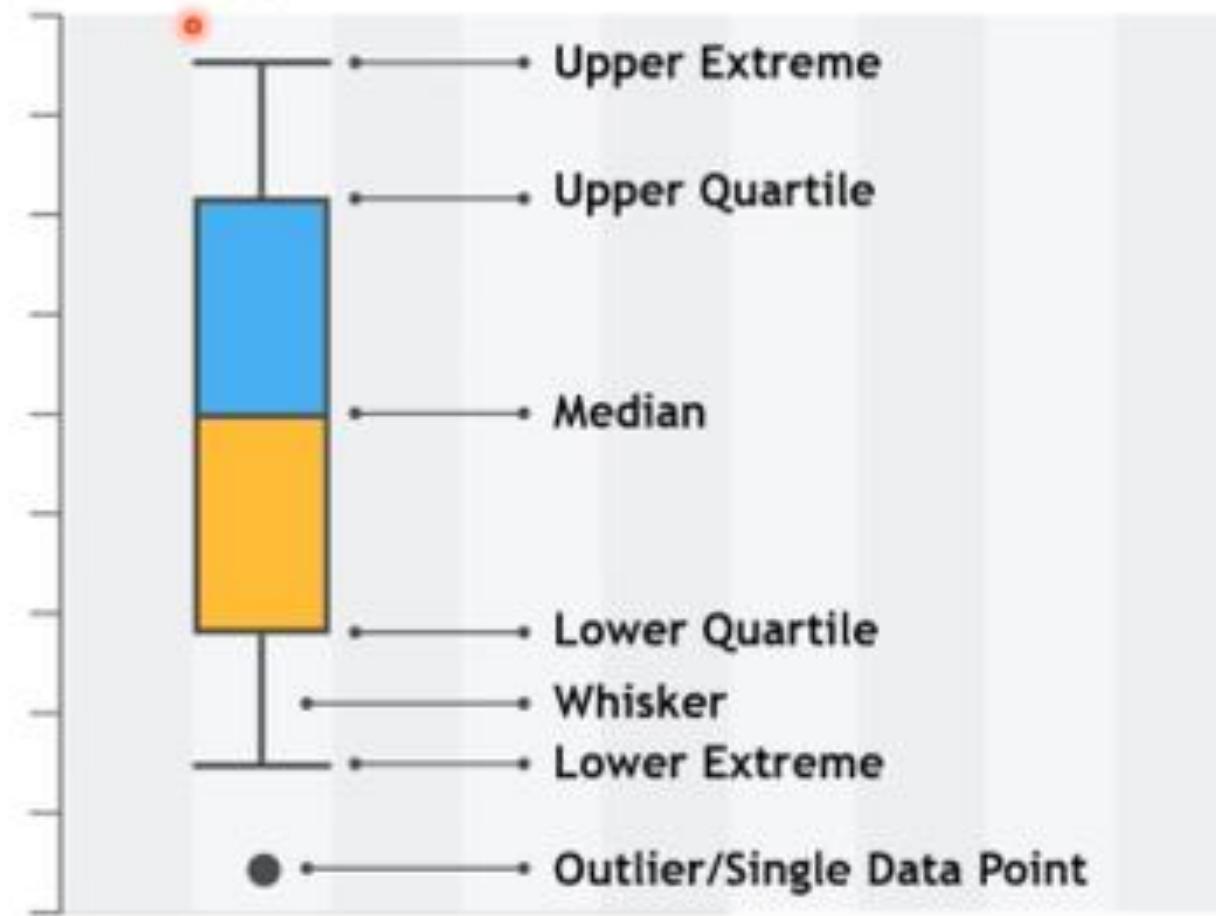
# Pie Charts

- A Pie Chart is a type of graph that displays data in a circular graph



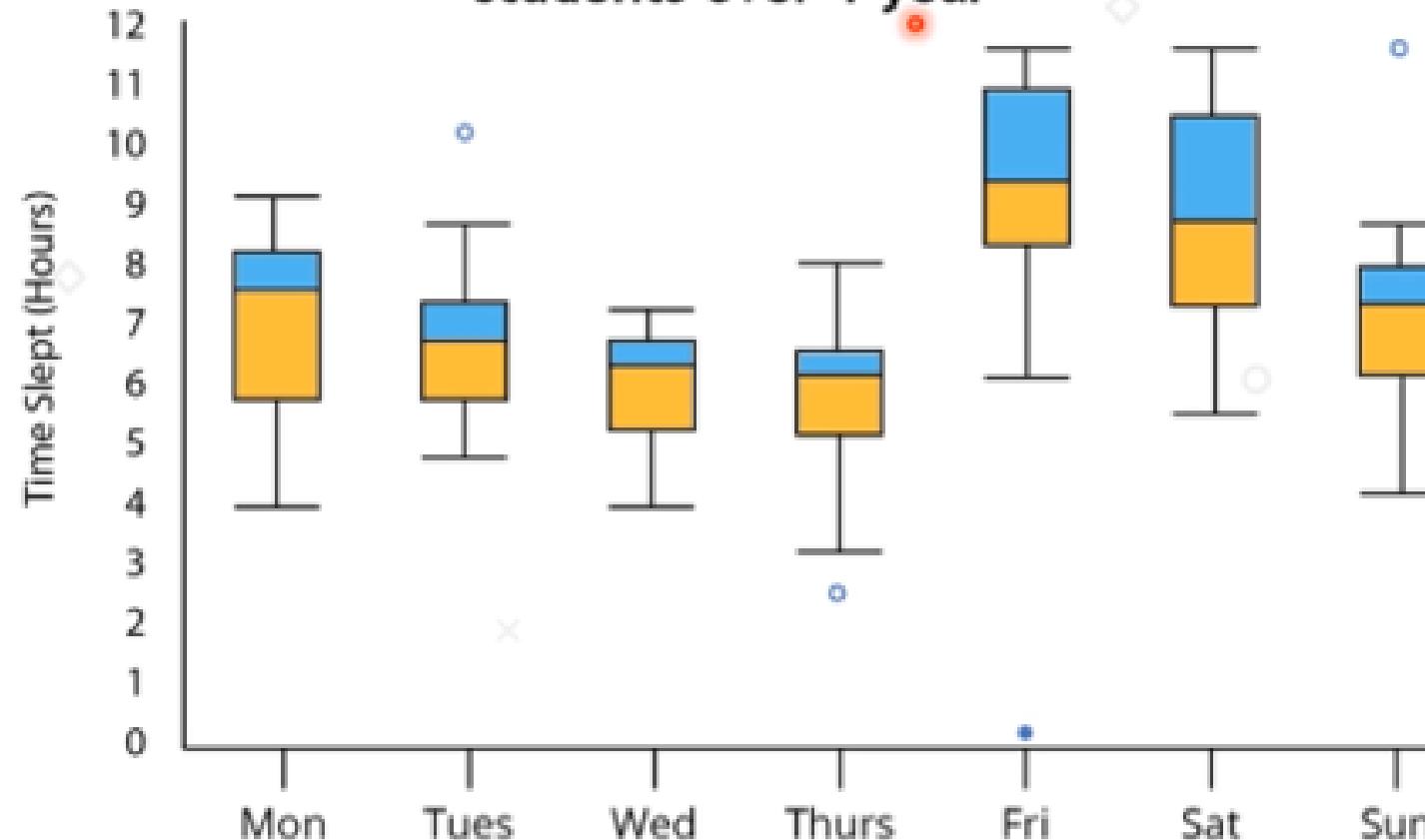
# Box Plots

Box plots graphically depict groups of numerical data through their quartiles



# Box Plots

Representing the hours slept each day of the week by 20 students over 1 year



# Recap Slide



**Introduction  
to Statistics**



**Data in Stats -  
Types &  
Sources**



**Intro to  
Descriptive  
Statistics**



**Stats Importance  
In Data Analysis  
& ML**



**Types of  
Statistics**



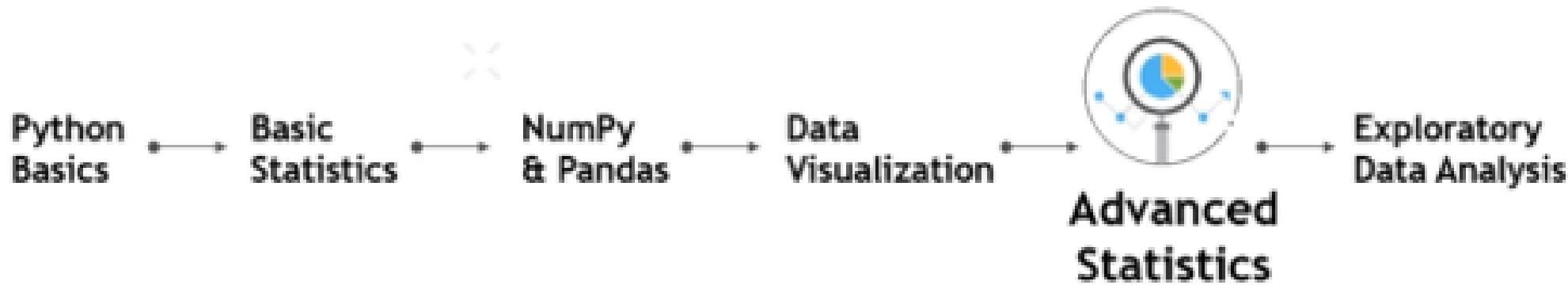
**Plots**

@amity1415

## Inferential Statistics



# Takeaways from the session



- 1 Basic understanding of inferential statistics
- 2 Probability - An important component influencing decisions
- 3 Validate a hypothesis for decision making using hypothesis testing
- 4 Calculate Degree of certainty in decisions using confidence intervals

# Contribution of Statistics to DS



**Understand underlying data**  
*Spread of data around mean*



**Draw inferences from data**  
*Sample to population*  
*E.g. A/B testing*



**Make predictions**  
*Predicting class of objects*  
*E.g. Spam/Not Spam*

# Contribution of Statistics to DS

Infer population insights from sample statistics

Population mean of  
10,000 insurance  
agents



How much insurance  
is sold by each  
agent?



Random Sample

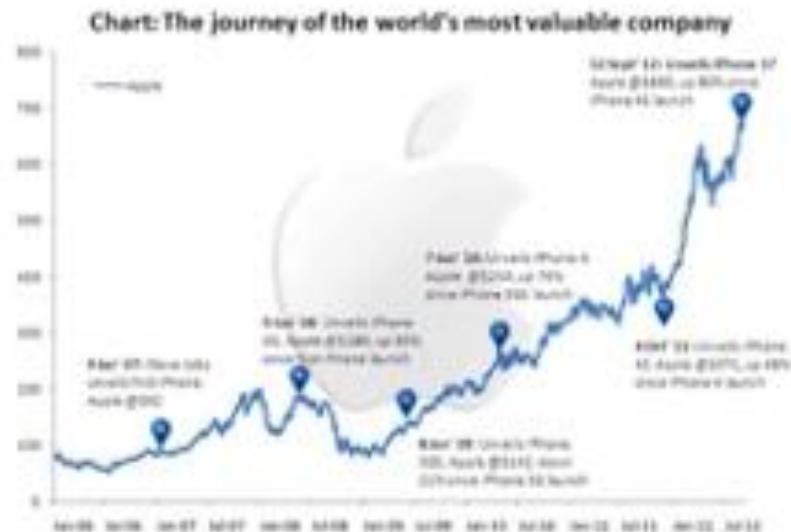
## Random variables - Outcomes of random phenomenon

A random variable  $X$ , is a variable whose possible values are numerical outcomes of a random phenomenon



### Discrete Random Variable

*Number of houses sold by a real estate agent in a month*



### Random Variable

*Stock prices of Apple in a month*

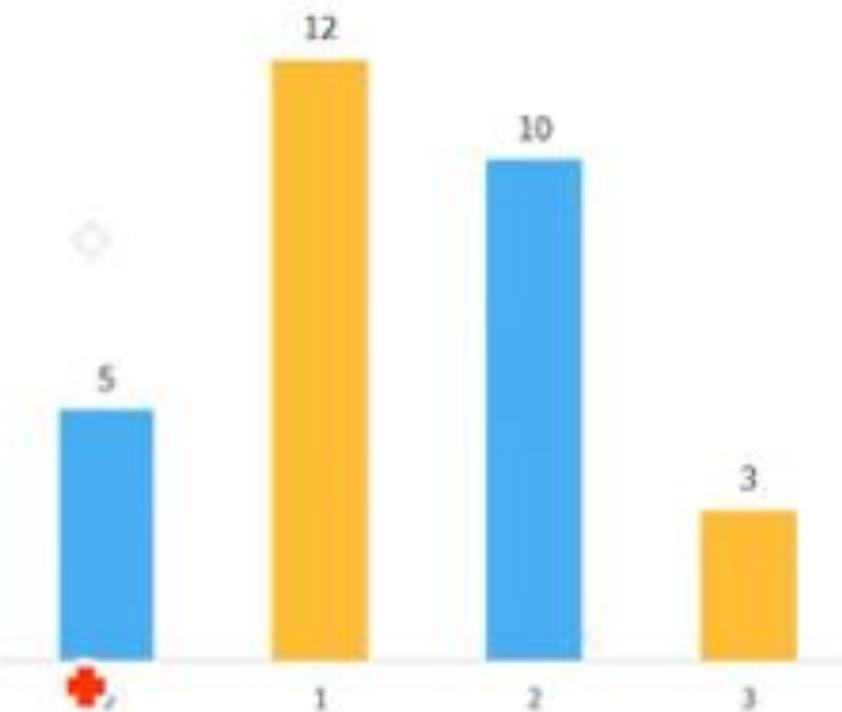
# Distribution of Real Estate Sales



Number of houses sold	Frequency (days)	Probability
0	5	5/30
1	12	12/30
2	10	10/30
3	3	3/30
Total	30	1

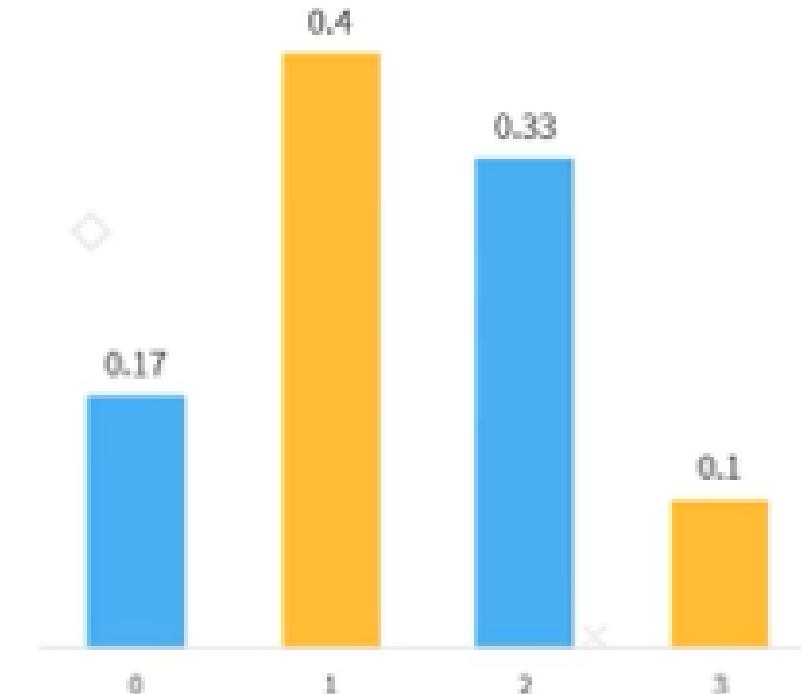
# Frequency Distribution of Real Estate Sales

Number of houses sold by a Real Estate Agent



Number of houses sold	Frequency
0	5
1	12
2	10
3	3

# Probability Distribution of Real Estate Sales



Number of houses sold	Probability
0	0.17
1	0.40
2	0.33
3	0.10

# Expected value of a probability distribution



Expected value,  
 $(R)=1.36$   
The average  
number of sales  
is 1.36 in a day

Number of houses sold	Probability
0	0.17
1	0.40
2	0.33
3	0.10

# Normal Probability distribution

@amity1415

Normal probability distribution models many natural processes, manufacturing processes and human endeavors



IQ scores



Stocks



Flight bookings

# How does Normal Distribution look?

@amity1415

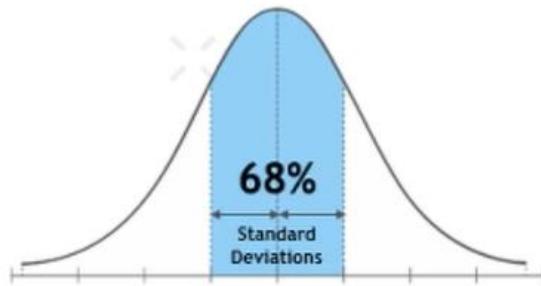
- Single Peak
- Symmetric
- Centered around

Probability = 0.5

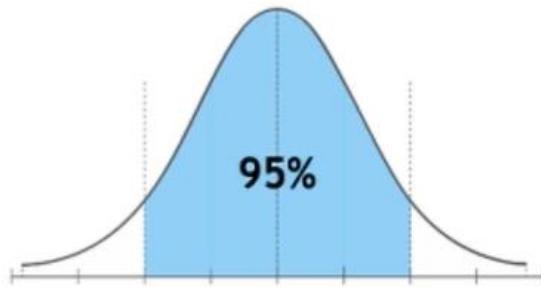
Probability = 0.5

Mean  
Median  
Mode

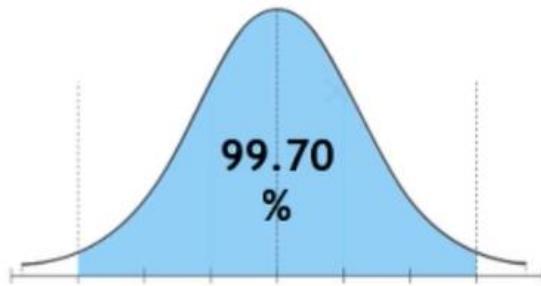
# Significance of values lying across 68-95-99.7 rule



68% of values are  
within 1 standard  
deviations from the  
mean



95% of values are  
within 2 standard  
deviations from the  
mean

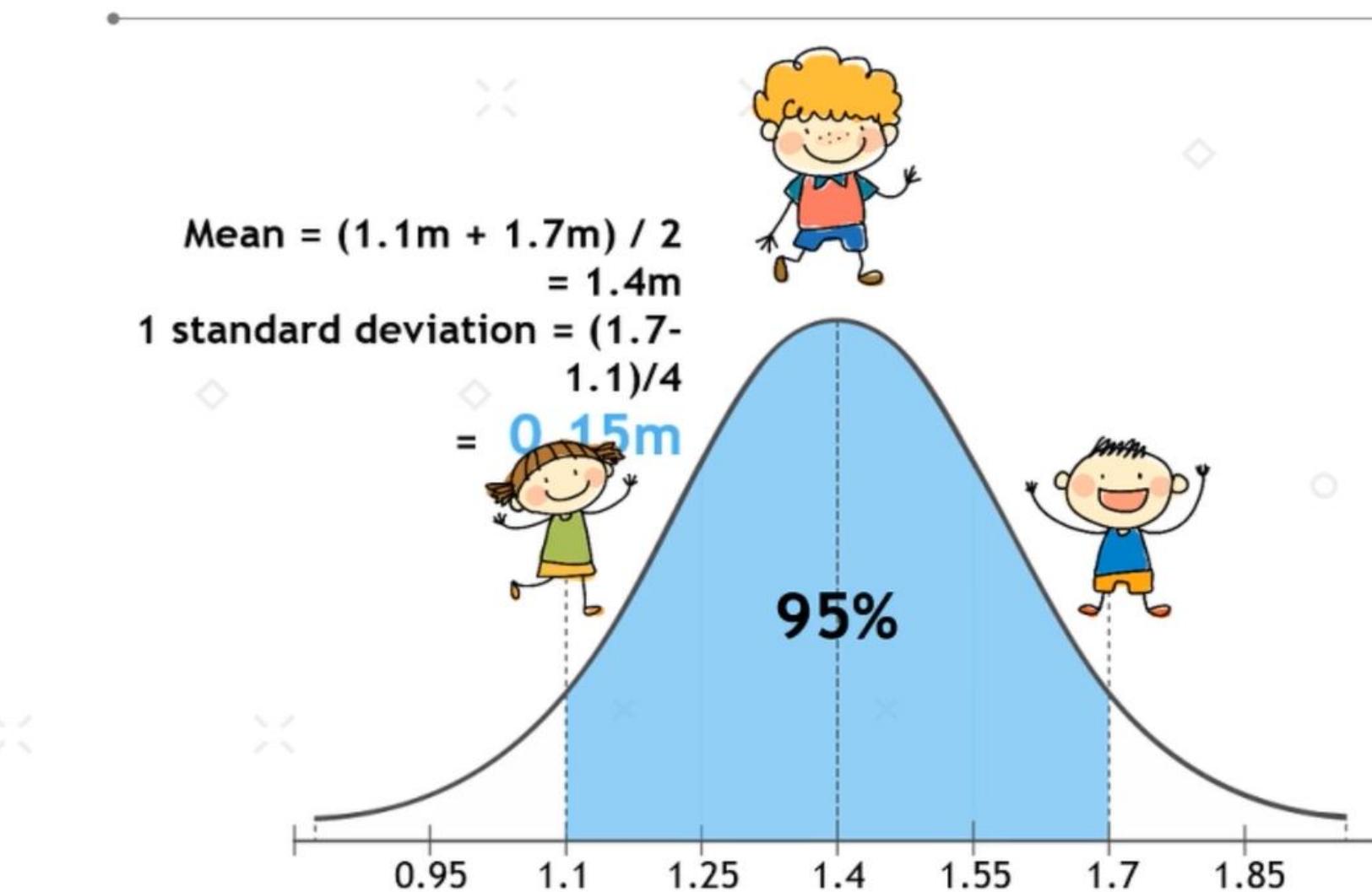


99.70% of values are  
within 3 standard  
deviations from the mean

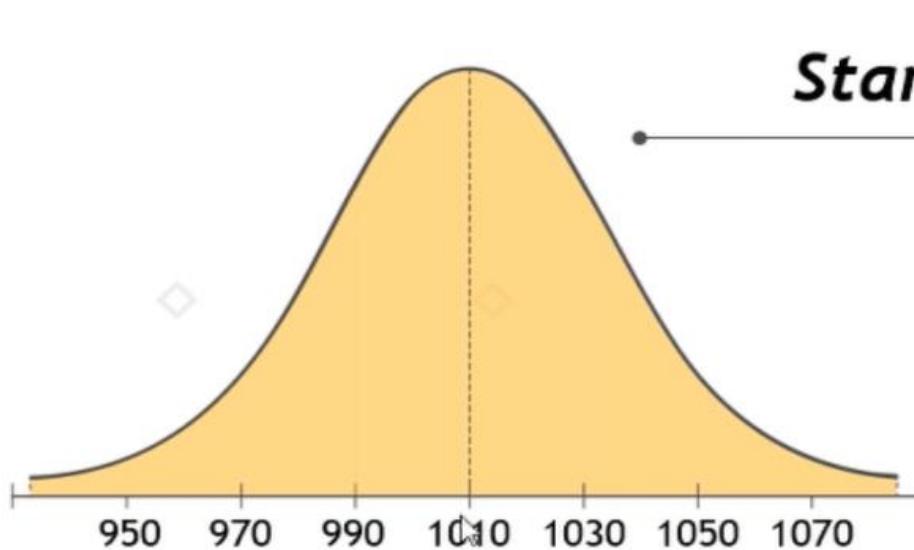
# Normal distribution of heights in a school

$$\text{Mean} = (1.1\text{m} + 1.7\text{m}) / 2 \\ = 1.4\text{m}$$

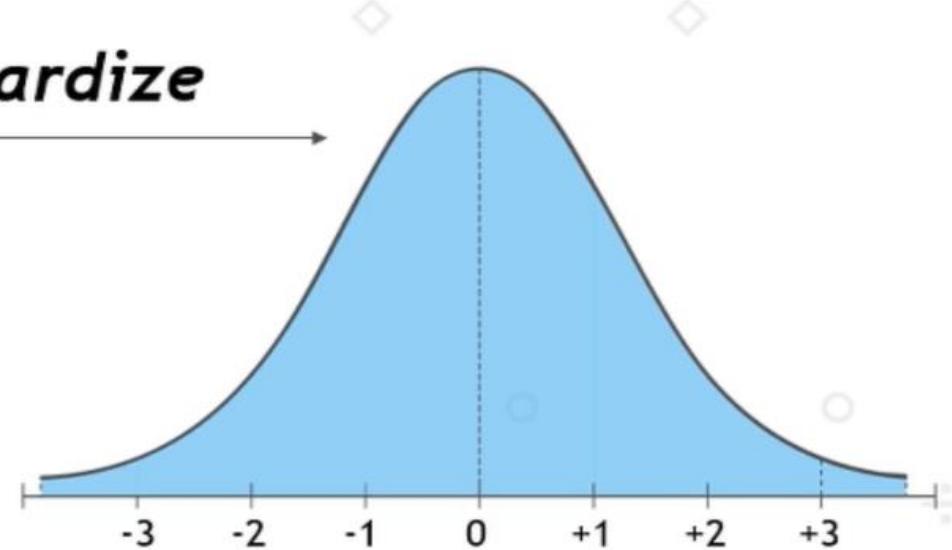
$$\text{1 standard deviation} = (1.7 - 1.1) / 4 \\ = 0.15\text{m}$$



# Standardizing data values



**Normal Distribution**



**Standard Normal Distribution**

# Why standardize data values? ▾



$$Z = \frac{X - \mu}{\sigma}$$

# Luggage loading time in the Airport

$$\begin{aligned}\mu &= 15 \\ \sigma &= 3.5\end{aligned}$$

$$Z = \frac{X - \mu}{\sigma}$$

Convert normal distribution to standard normal distribution through a Z-score

Probability that a flight will take 22 minutes or more

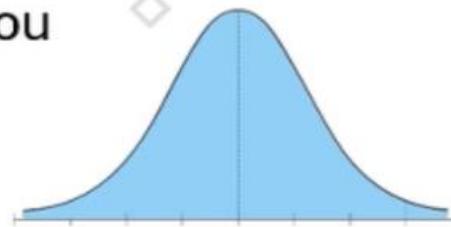
$$P(X \geq 22) = P(z \geq 2) = 0.5 - 0.4772 = 0.0228$$



# Normal Distribution Table

The normal curve table gives the percentage of data starting from the middle. For  $z = 1.28$ , you get **0.3997**.

This means 39.97% of the data in the normal curve is found between  $z = 0$  and  $z = 1.28$

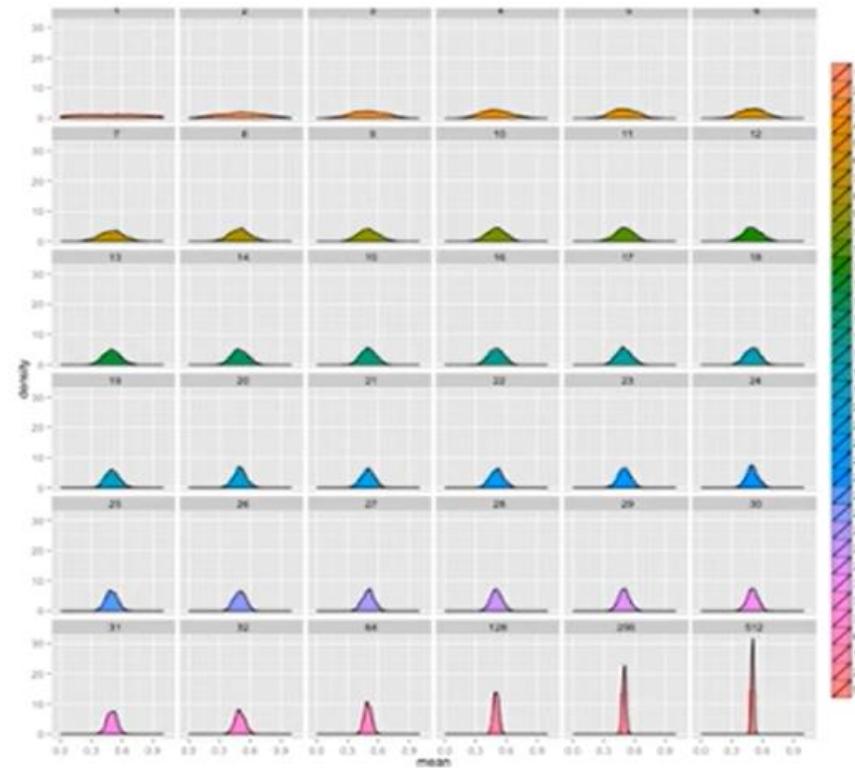


	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0	0	0.004	0.008	0.012	0.016	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.091	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.148	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.17	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.195	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.219	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.258	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.291	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.334	0.3365	0.3389
1	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.377	0.379	0.381	0.383
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.398	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177

# Central Limit theorem

The Central Limit Theorem, or CLT, states that:

- Any data set that is randomly sampled from a population repeatedly for a couple of times,
- Distribution of the sample mean will be approximately normal IF the sample size is large, regardless of its original distribution



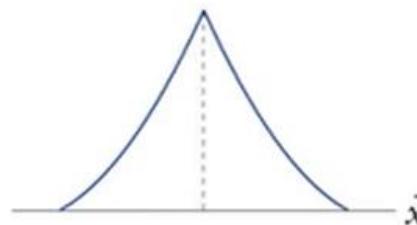
# Central Limit theorem

@amity1415

Population distribution



Sampling distribution of  $\bar{X}$  with  $n = 5$



Sampling distribution of  $\bar{X}$  with  $n = 30$

