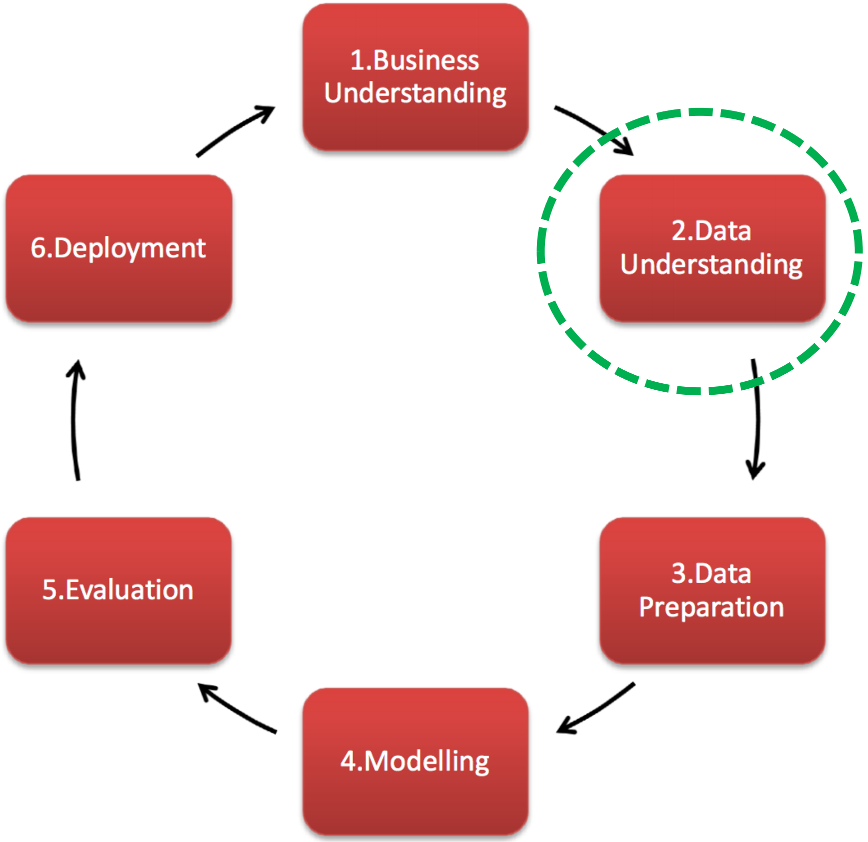


http://www.cheat-sheets.org/saved-copy/427513-Statistics-Reference-Cheatsheet.pdf

Time For Data Understanding

Business Problem

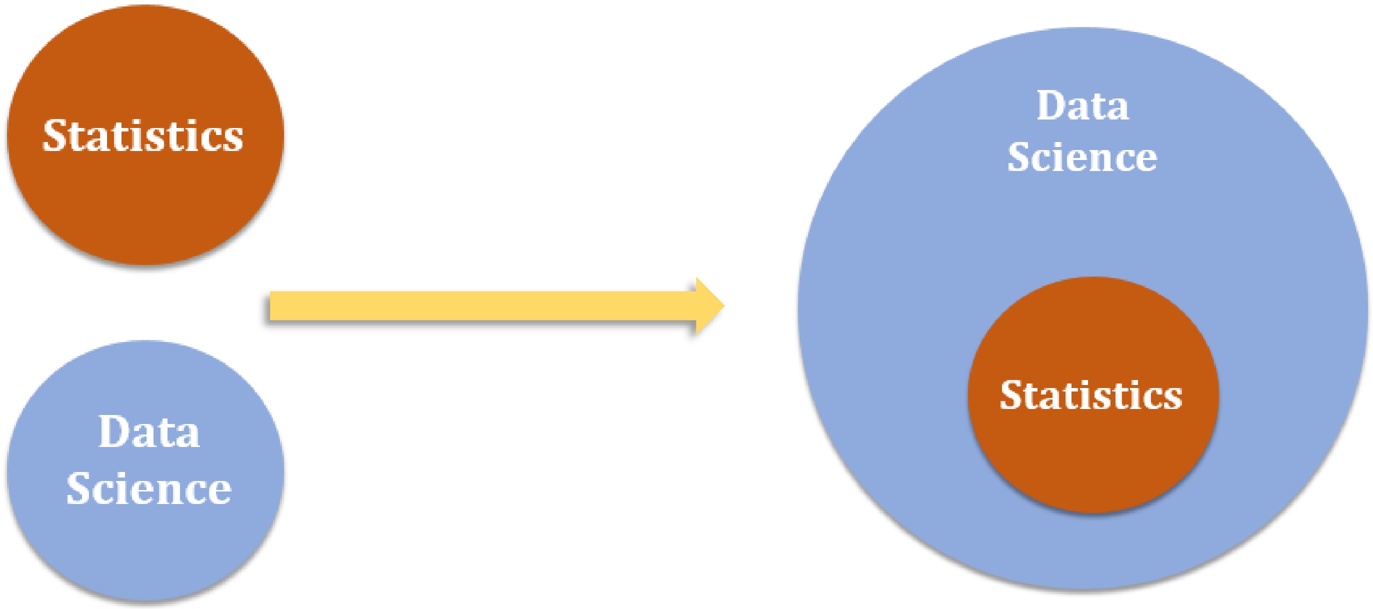
Data

Probability Of Getting Approval

Yes Or No

- Variable Identification
- Univariate Analysis
- Bivariate Analysis
- Missing Values
- Outliers

Data Science Vs Statistics



- Both fields use data, but they have different motivations and different goals
- They have different backgrounds, and they operate within different contexts

Hence Most Data Scientists Are Not Trained As Statisticians

Why Statistics

Statistics uses methods analyze the numbers and transform them into useful information for <u>making right decisions</u>

- Discover something new in the organization
- To Understand the situation
- Understand results with Biased View
- Which numbers are helpful in which situation

To Be A Good Statistical Person, Always Ask Yourself

- Where did the given numbers come from?
- How are they calculated?
- Are those the right numbers needed to make this decision?

http://www.mathworksheetscenter.com/mathtips/statsareimportant.html

Topics
Fundamentals of Statistics

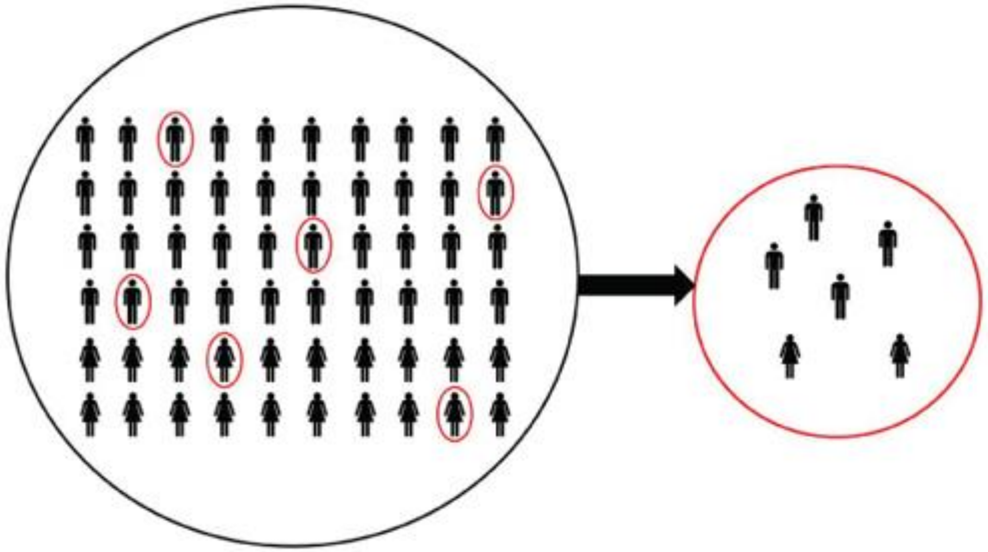
1. Population and Sample
2. Variable
3. Types of Data
4. Branches of Statistics

Fundamentals of Statistics
1. Population and Sample

What Data I have?

N Population

n Sample



Parameters

Statistics

Food Sampling

Your Questions Should Be:

Will it not be a wrong assumption if we take only samples?

You are not expected to get all the data

Fundamentals of Statistics

2. Variable

Variable

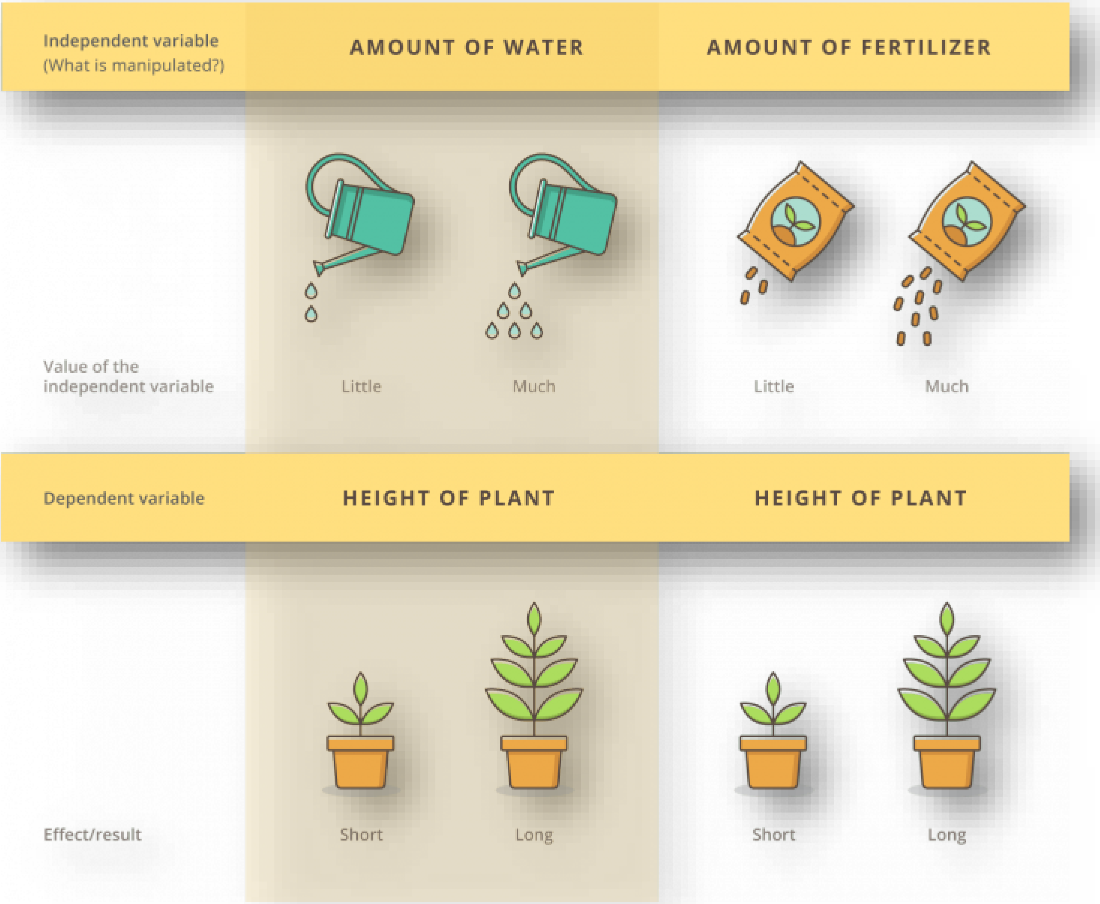
Records

Name	Maths	Science	English	Physics
David	85	87	88	92
Richard	91	81	78	71
John	81	86	88	84
Tony	84	86	87	82
Scott	71	79	82	88

Data is a combination of many variables

Fundamentals of Statistics
Types Of Variables

Dependent (Y) and Independent Variables (X's)



https://explorable.com/sites/default/files/independent-and-dependent-variable2.png

Fundamentals of Statistics
Types Of Variables

Dependent (Y) and Independent Variables (X's)

Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status
LP001003	Male	Yes	1	Graduate	No	4583	1508	128	360	1	Rural	N
LP001005	Male	Yes	0	Graduate	Yes	3000	0	66	360	1	Urban	Y
LP001006	Male	Yes	0	Not Graduate	No	2583	2358	120	360	1	Urban	Y
LP001008	Male	No	0	Graduate	No	6000	0	141	360	1	Urban	Y
LP001011	Male	Yes	2	Graduate	Yes	5417	4196	267	360	1	Urban	Y
LP001013	Male	Yes	0	Not Graduate	No	2333	1516	95	360	1	Urban	Y
LP001014	Male	Yes	3+	Graduate	No	3036	2504	158	360	0	Semiurban	N
LP001018	Male	Yes	2	Graduate	No	4006	1526	168	360	1	Urban	Y
LP001020	Male	Yes	1	Graduate	No	12841	10968	349	360	1	Semiurban	N
LP001024	Male	Yes	2	Graduate	No	3200	700	70	360	1	Urban	Y
LP001028	Male	Yes	2	Graduate	No	3073	8106	200	360	1	Urban	Y
LP001029	Male	No	0	Graduate	No	1853	2840	114	360	1	Rural	N
LP001030	Male	Yes	2	Graduate	No	1299	1086	17	120	1	Urban	Y
LP001032	Male	No	0	Graduate	No	4950	0	125	360	1	Urban	Y
LP001036	Female	No	0	Graduate	No	3510	0	76	360	0	Urban	N
LP001038	Male	Yes	0	Not Graduate	No	4887	0	133	360	1	Rural	N
LP001043	Male	Yes	0	Not Graduate	No	7660	0	104	360	0	Urban	N
LP001046	Male	Yes	1	Graduate	No	5955	5625	315	360	1	Urban	Y
LP001047	Male	Yes	0	Not Graduate	No	2600	1911	116	360	0	Semiurban	N

IV's **DV**

Fundamentals of Statistics
Types Of Variables

Variables

Categorical

YES NO

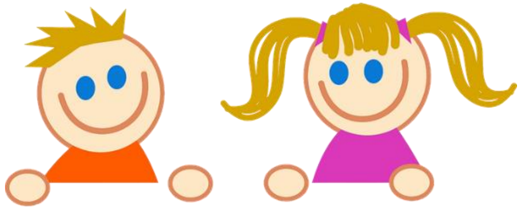
Ordinal

Nominal

Qualitative

Numerical

Discrete



Continuous

Quantitative

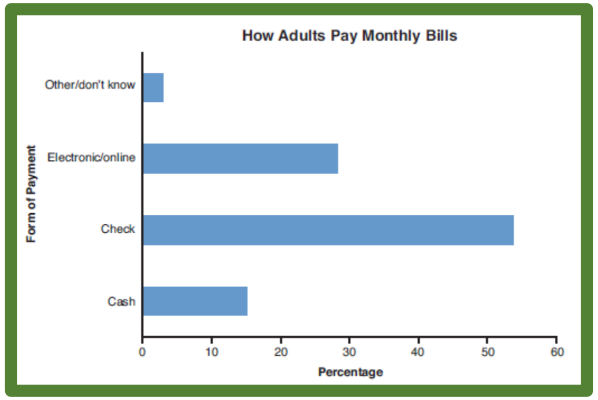
Visualization
Presenting Categorical Variables

What can be represented in it:

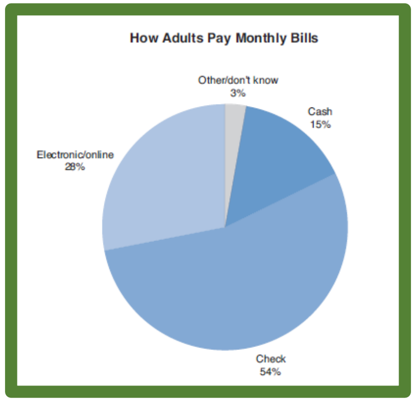
- Count
- Amount
- Percentage

Univariate

The Bar Chart



The Pie Chart



Bi-variate

Two-Way Cross-Classification Table

		Wafer Condition		
		Good	Bad	Total
Particles Found	Yes	14	36	50
	No	320	80	400
	Total	334	116	450

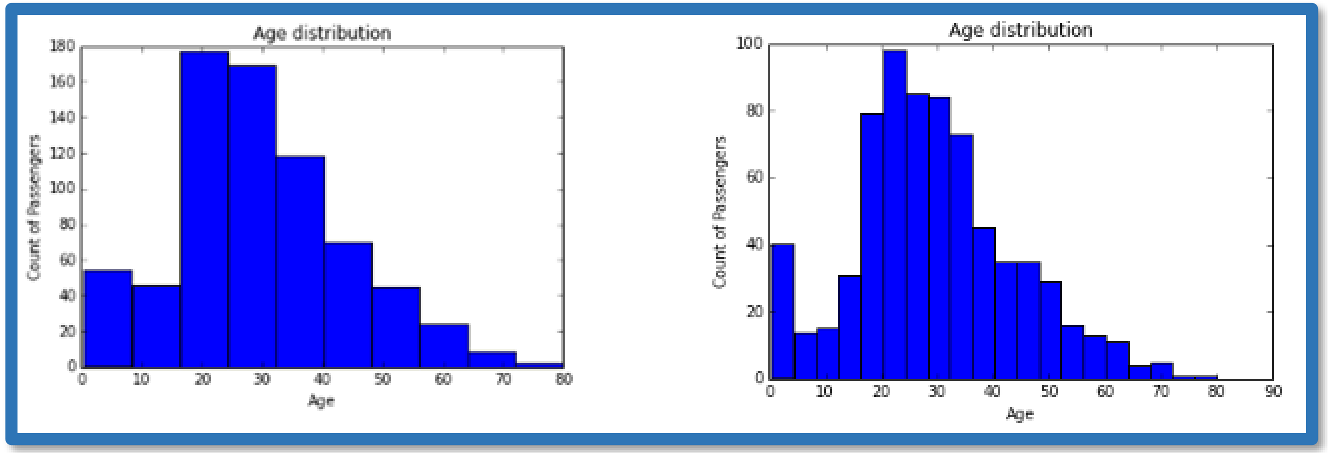
Visualization
Presenting Numerical Variables

What we look at?

- Centre
- Spread
- Outliers

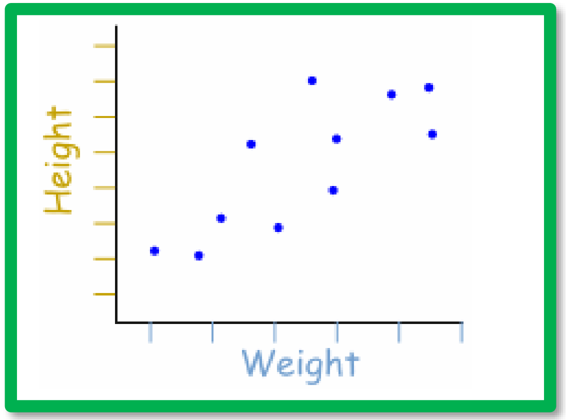
Univariate

Histogram



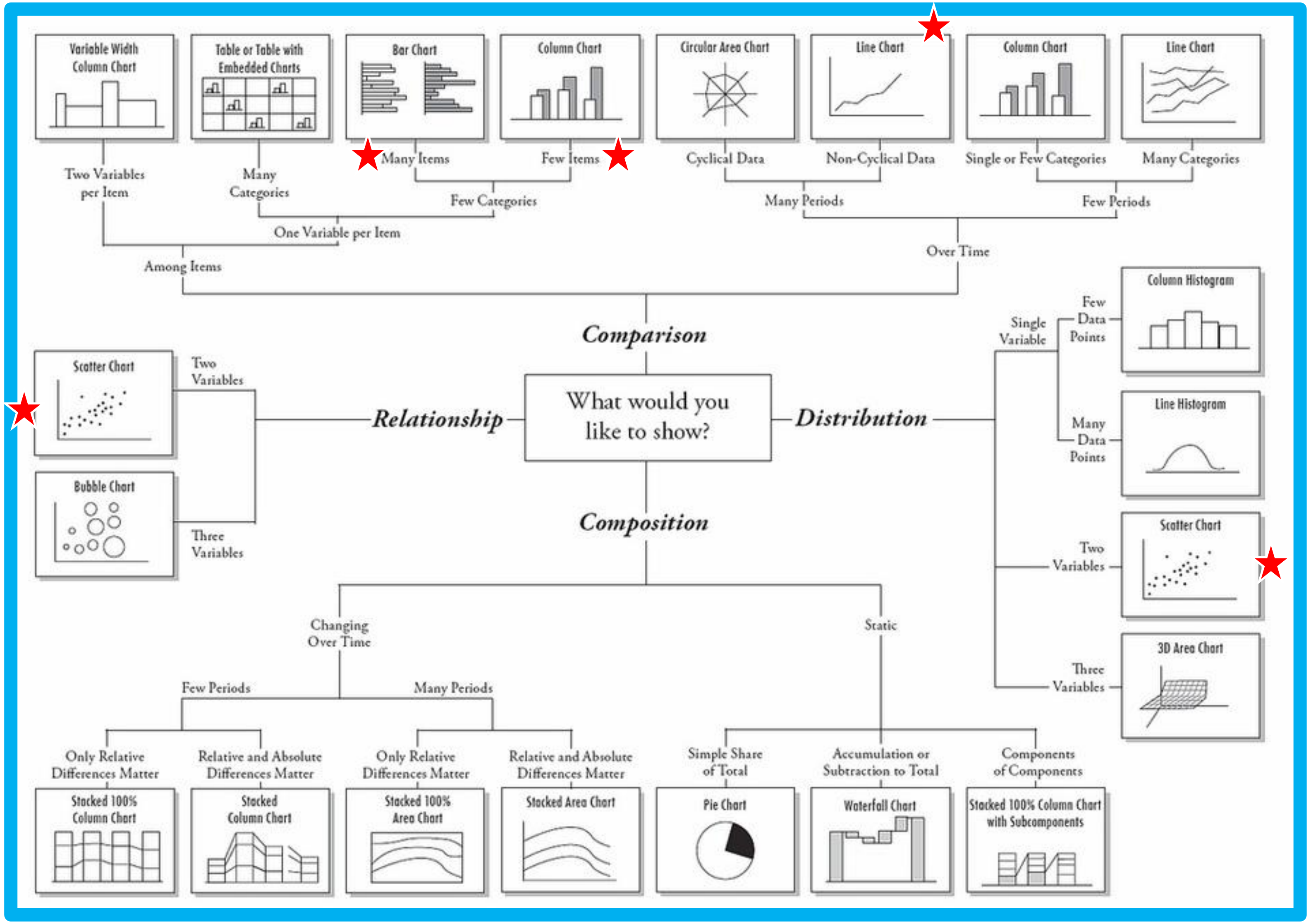
Distribution of Count of Passengers vs Age (Into Bins)

Bi-variate



Scatter Plot

Visualization
Presenting Data in Charts and Tables



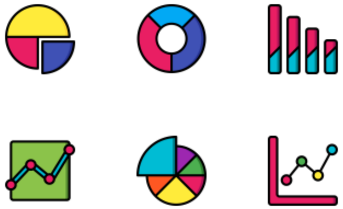
Fundamentals of Statistics
4. Branches of Statistics

Descriptive Statistics

Large amount of data

Collecting

Summarizing and Presenting



http://www.statisticshowto.com/probability-and-statistics/descriptive-statistics/

Inferential Statistics

Sample data

Analyzes **sample data**

Conclusions about a population

http://www.statisticshowto.com/inferential-statistics/

Topics

Fundamentals of Statistics

1. Population and Sample
2. Variable
3. Types of Data
4. Branches of Statistics
    a) Descriptive Statistics
    b) Inferential Statistics

Descriptive Statistics

1. Measures of Central Tendency
2. Measures of Spread

Descriptive Statistics
1. Measures of Central Tendency

glucose
148
85
183
89
137
116
78
115
197
125
110
168
139

Mean = 156

Three commonly used measures of
Central Tendency

1. Mean
2. Median
3. Mode

Descriptive Statistics
1. Measures of Central Tendency

Mean
Simple Average

Concept

$$\mu = \frac{\sum x}{N}$$

$$\bar{X} = \frac{\sum x}{n}$$

Disadvantage

Mean is Effected by Outliers

http://www.statisticshowto.com/probability-and-statistics/statistics-definitions/mean-median-mode/

Descriptive Statistics
1. Measures of Central Tendency
Median

The median is the middle value for a set of data
when a set of the data values have been ordered from lowest to higher value.

65	55	89	56	35	14	56	55	87	45	92

14	35	45	55	55	**56**	56	65	87	89	92

14	35	45	55	**55**	**56**	56	65	87	89

Concept

$$\frac{n+1}{2}$$

Advantage

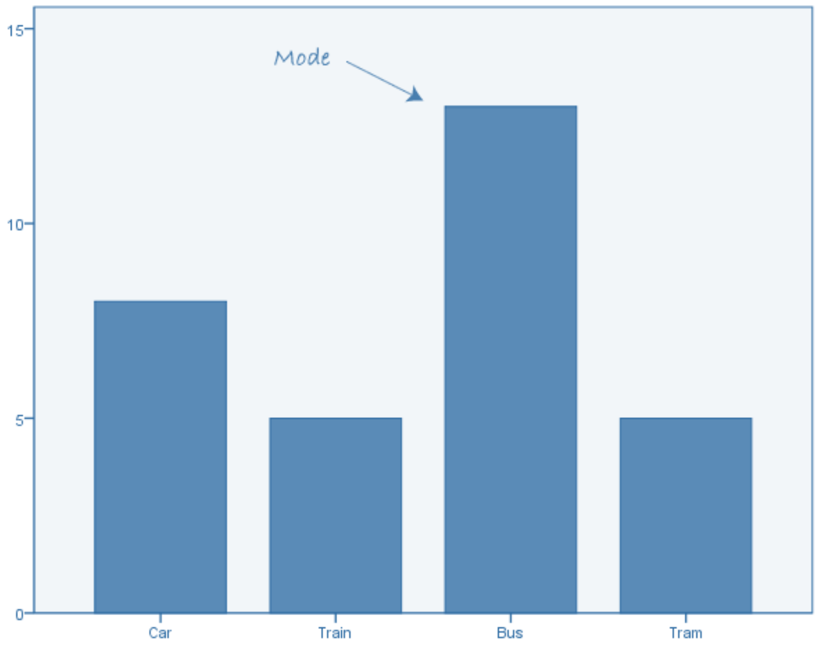
Median is Less Effected by Outliers

http://www.statisticshowto.com/probability-and-statistics/statistics-definitions/mean-median-mode/

Descriptive Statistics
1. Measures of Central Tendency

Mode

A value (or values) in a set of data that appears most frequently, the **Mode is generally used for categorical data**



- Not Effected by Outliers
- Not good most common mark is far away from the rest of the data

http://www.statisticshowto.com/probability-and-statistics/statistics-definitions/mean-median-mode/

Your Questions Should Be:

Which measure is the best?

There is no best method

At least we must try two methods

Don’t ever depend on single method

Descriptive Statistics
2.Measures of Spread

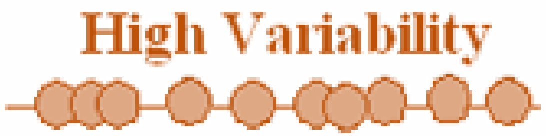
How Many Teeth

29, 30, 31 → 32

Your IQ

82, 84, 86, 89, 90, 91, 93, 95, 99, 101, 103

More spread out



http://www.softschools.com/math/probability_and_statistics/measures_of_spread_range_standard_deviation_and_variance/

Descriptive Statistics
2.Measures of Spread

Measures of spread help us to summarize how spread out these scores are!

	A	B	C	D
1	ID	Name	Course	Marks
2	1	Jack	Software Engineering	60
3	2	Billy	Requirement Engineering	90
4	3	Mcfaden	Multivariate Calculus	34
5	4	Steven Shwimmer	Software Architecture	96
6	5	Ruby jason	Relational DBMS	70
7	6	Mark Dyne	PHP development	34
8	7	Philip namdaf	Microsoft Dot Net Platform	78
9	8	Erik Bawn	HTMl & Scripting	87
10	9	Ricky ben	Data communication	78
11	10	Miecky	Computer Networks	89

Measure of Spread
Can Be Explained Through

1. Range
2. Quartile and Interquartile Range
3. Variance
4. Standard Deviation

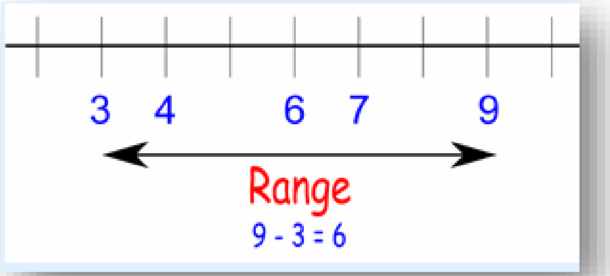
Measures of Spread is used to describe the variability in a sample or population and is connected with measure of central tendency

Descriptive Statistics

2.Measures of Spread

Range

Difference between the lowest and highest values



Meal cost Range in City = 62
Suburban Meal Cost Range = 42

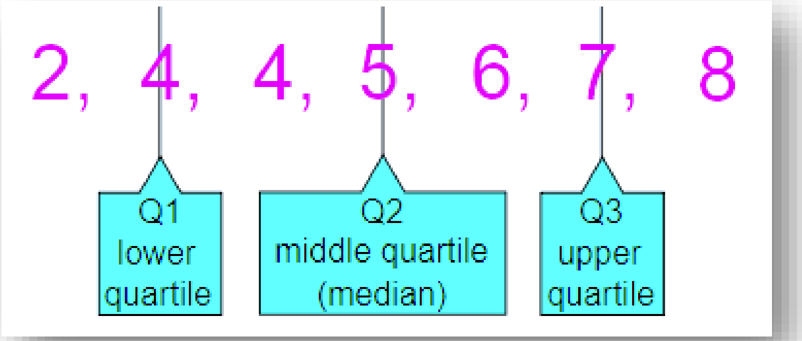
Meal costs in the city show much more variation than suburban meal costs.

Disadvantage

Misleading, when there are extremely high or low values

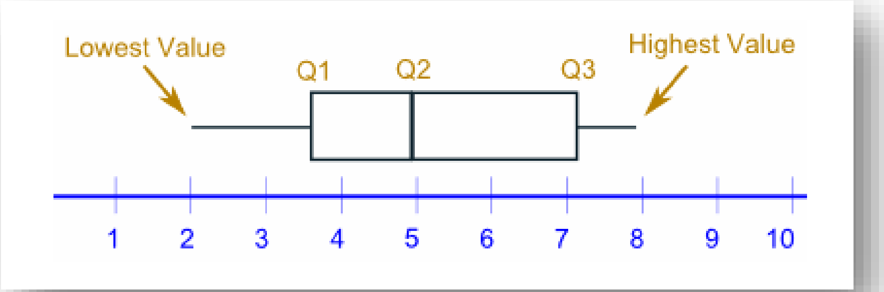
Descriptive Statistics
2.Measures of Spread
Quartiles and Interquartile Range

Quartiles tell us about the spread of a data set by breaking the data set into quarters



$Q_1 = \frac{n+1}{4} th$ $\quad$ $\text{Median} = \frac{n+1}{2}$ $\quad$ $Q_3 = \frac{3(n+1)}{4} th$

Quartiles useful:
- Useful to measure of spread
- Data that isn't symmetrically distributed
- Data set that has outliers

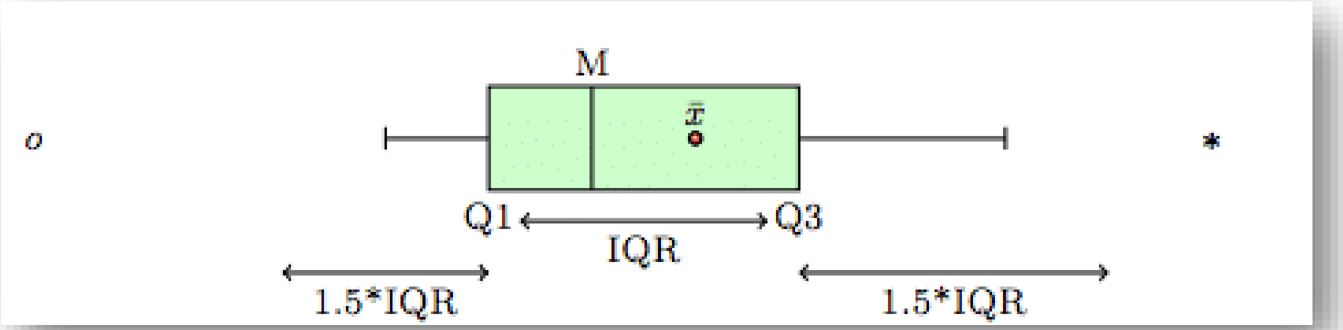


https://www.mathsisfun.com/data/quartiles.html

Descriptive Statistics
2.Measures of Spread

Interquartile range or IQR

Distance between the first quartile and the third quartile.



IQR useful for: For calculating outliers

- Lower boundary is Q1 – 1.5 * IQR
- Upper boundary is Q3 + 1.5 * IQR

https://en.wikipedia.org/wiki/Interquartile_range

Descriptive Statistics
2.Measures of Spread

Variation

Two measures that tell you how a set of data values fluctuate around the mean of the variable.

1. Variance **2. Standard Deviation**

Descriptive Statistics
Measures of Spread
1. Variance

$$\text{Sample variance} = S^2 = \frac{\Sigma(X_i - \overline{X})^2}{n-1}$$

Steps to Calculate Variance

1. Find Mean
2. Calculate the difference between each of the 10 individual times and the mean (which is 39.6 minutes)
3. Square those differences
4. Sum the squared differences

Day	Time	Difference: Time Minus Mean (39.6)	Square of Difference
1	39	−0.6	0.36
2	29	−10.6	112.36
3	43	3.4	11.56
4	52	12.4	153.76
5	39	−0.6	0.36
6	44	4.4	19.36
7	40	0.4	0.16
8	31	−8.6	73.96
9	44	4.4	19.36
10	35	−4.6	21.16
		Sum of Squares:	412.40

$$\frac{412.40}{9} \longrightarrow$$ **Sample Variance**

The square root of 45.82 is sample standard deviation

$$S = \sqrt{\frac{\Sigma(X_i - \overline{X})^2}{n-1}}$$

Descriptive Statistics
Measures of Spread
Standard Deviation

Promotion to a Cricketer

						Average
33	31	32	36	31	31	32.3
22	34	58	52	10	21	32.8

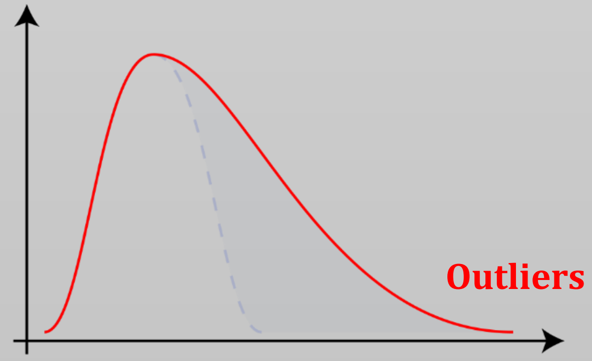
						Average	STDEV	
33	31	32	36	31	31	32.3	2.0	✓
22	34	58	52	10	21	32.8	18.9	✗ → Extremely volatile

Conclude that between 30.3 (32.3 – 2.0) runs and 34.3 (32.3 + 2.0)

Descriptive Statistics
Shape of Distributions

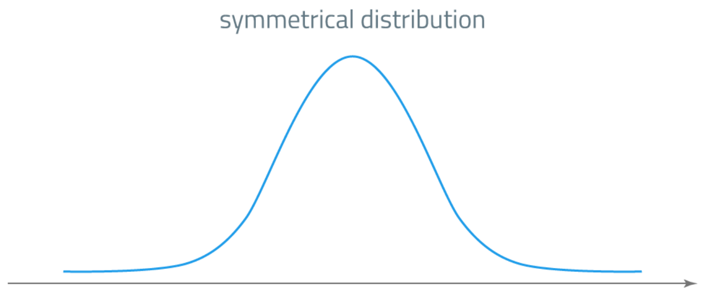
For Numerical Data

More Points on Left and Tail Leading to Right Side



Positive Skew

Right Skewness
Outliers are to the right
Mean > Median



Highest point is defined by mode

More Points on Right and Tail Leading to Left Side

Negative Skew

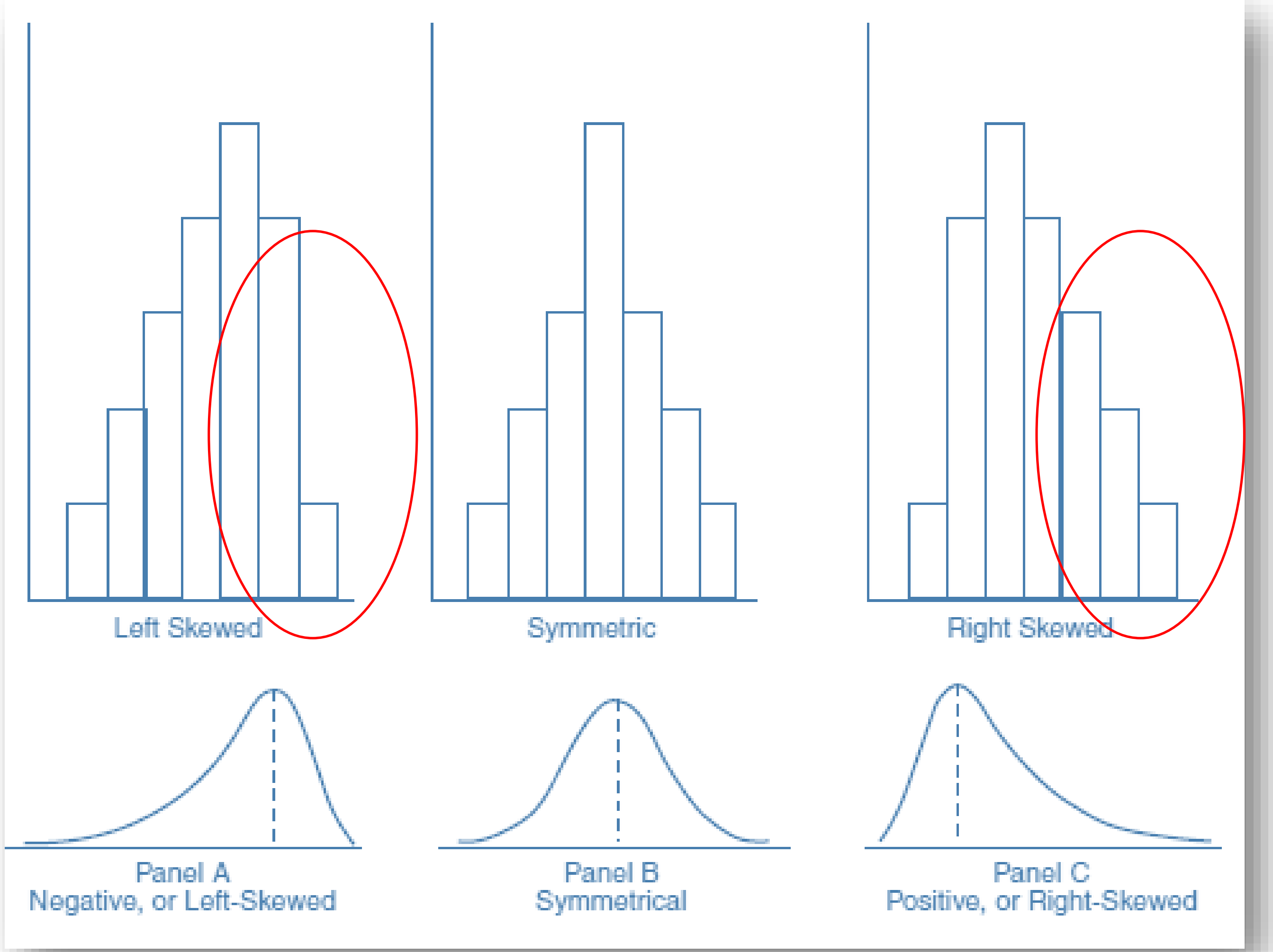
Left Skewness
Outliers are to the left
Mean < Median

Scores on an exam in which most students score between 10 to 60

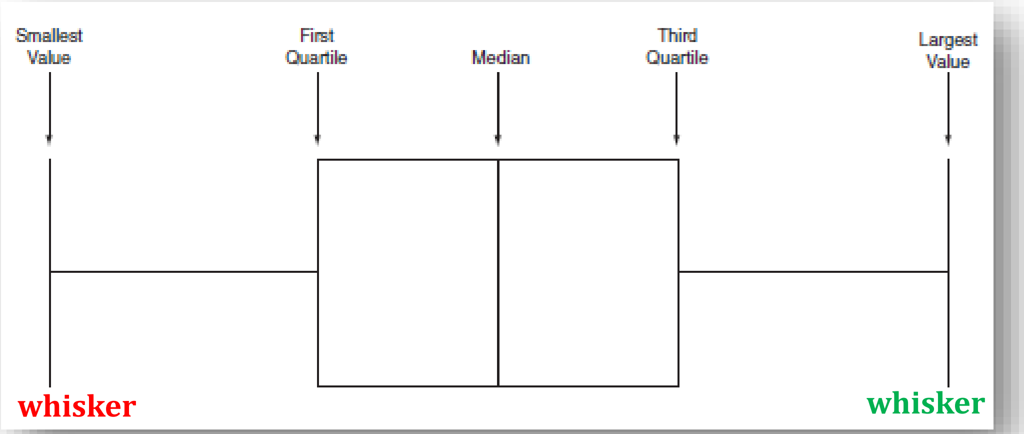
Scores on an exam in which most students score between 80 and 100

Descriptive Statistics
Shape of Distributions

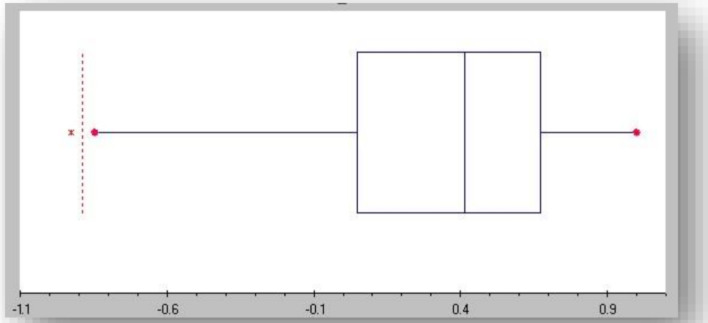
Identifying Shape



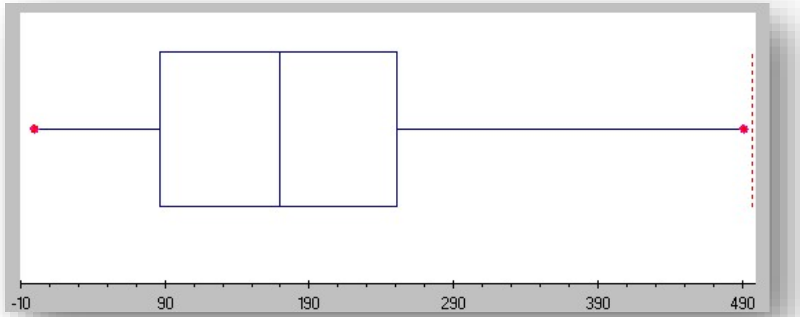
Descriptive Statistics
Shape of Distributions
Box Plot



Symmetrical



Left-skewed



Right-skewed

Descriptive Statistics

1. Measures of Central Tendency
2. Measures of Spread
3. Measure of Symmetry

Univariate Measures

What if we are working with more than one variable

Descriptive Statistics

1. Measures of Central Tendency
2. Measures of Spread
3. Measure of Symmetry
4. Covariance
5. Correlation

Bivariate Measures

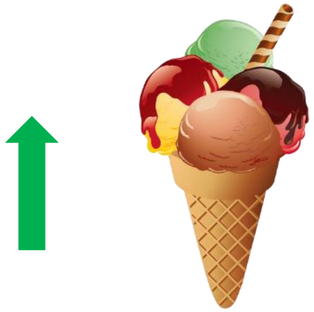
To understand the measure of relations between variables

Check The Below Images

Correlation Coefficient is 1
Absolutely dependent on each other

Correlation Coefficient is 0
Absolutely Independent on each other

Correlation Coefficient is -1
Negative relation with each other

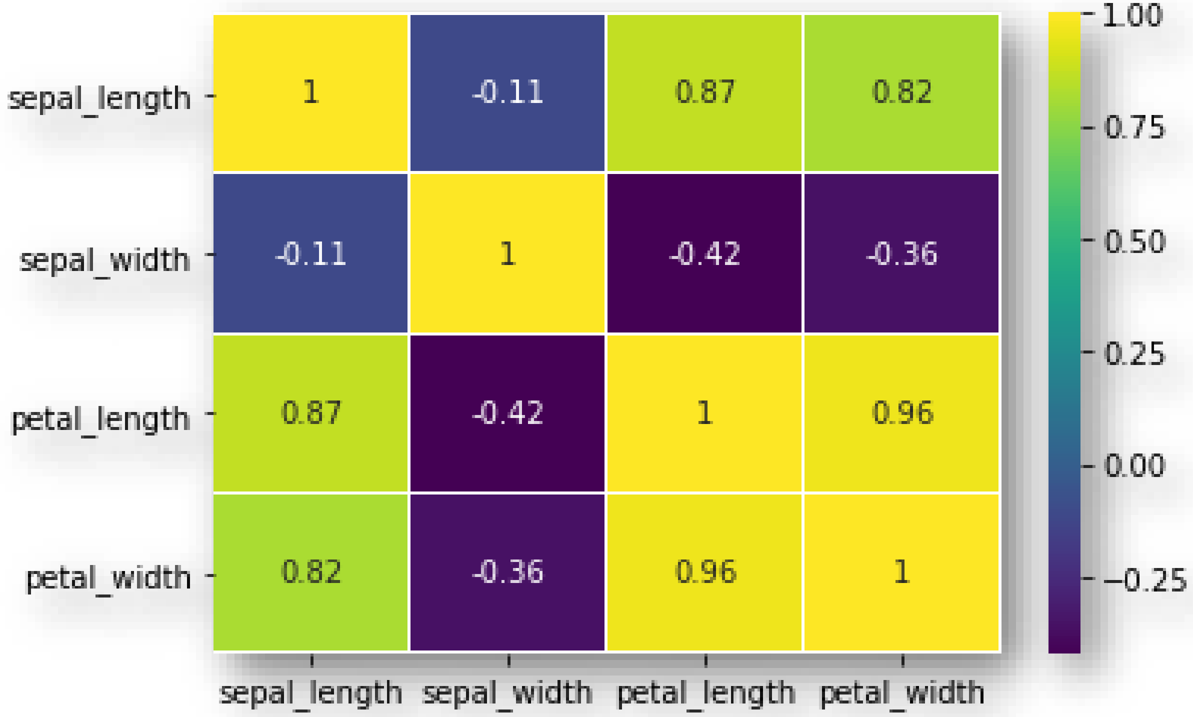


But with how much once variable is related with other? Number

Functionality	Covariance	Correlation
Used for	Extent to which two random variables change	How strongly two random variables are related
Values	-∞ and +∞	-1 and +1
Scaling down to understandable values	A measure of correlation	Scaled form of covariance
Change in scale effect	Covariance is affected by the change in scale	Correlation is not influenced by the change in scale.
Hence it is	Unit bound measure	A unit-free measure

Correlation is preferred over covariance

Descriptive Statistics
Correlation Coefficient



For linear models , multicollinearity will decreases the model performance

Probability

How likely is it that some event will occur?

N Trials = Experiment

Experiment

Q. What is the probability of getting two when I throw a die You

A. We have six possibilities, two is one of them, so it is 1/6th

Event

Probability Calculating

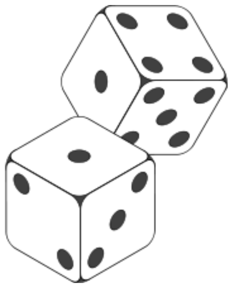
$$p(\text{Event}) = \frac{\text{Number of Elementary Outcomes in Event}}{\text{Number of Elementary Outcomes in the Sample Space}}$$

Example:

Calculate the probability of an even number, 2, 4, or 6 on the toss of a die

Probability Calculating

How many elementary outcomes for a pair dice

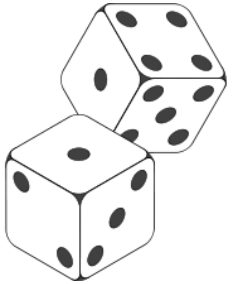


- 6 elementary outcomes on die number one
- 6 elementary outcomes on die number two
- 6*6 = 36 elementary outcomes in the sample space

Now tell me:

What's the probability of tossing a pair of dice and having a 5 come up?

Compound Events



Tossing a pair of dice and having a 5 come up?

Compound Events

Here we have 2 events

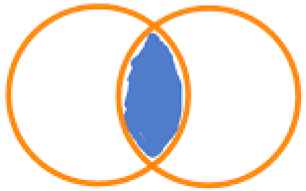
Event A and Event B (combinations of events)

Combine events in either of two ways

Union

$A \cup B$

Intersection



$A \cap B$

Compound Events

Probability of a 1 **or** a 5 on the toss of a single fair die.

$$p(1 \cup 5) = \frac{\text{Number of Elementary Outcomes in Tossing a 1 OR 5}}{\text{Number of Elementary Outcomes in the Sample Space}} = \frac{2}{6} = .33$$

Probability of tossing between 1 and 3 or between 2 and 4?

$$p(A \cup B) = \frac{(\text{\# of outcomes in } A) + (\text{\# of outcomes in } B) - (\text{\# of outcomes in } A \cap B)}{\text{Number of Elementary Outcomes in the Sample Space}} = \frac{3+3-2}{6} = .67$$

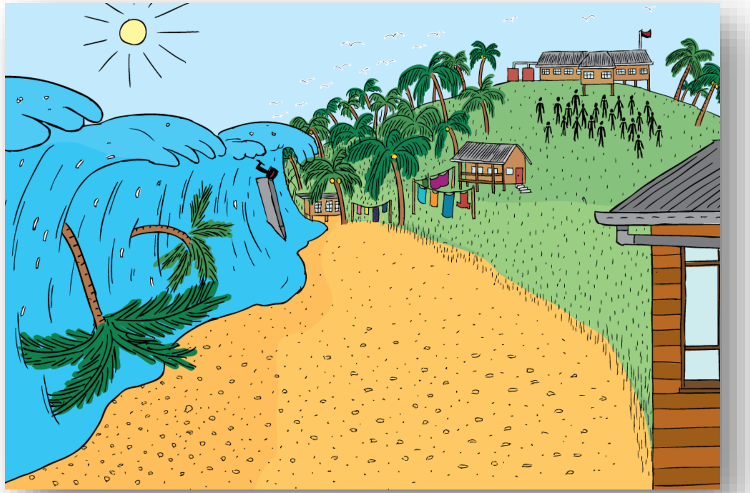
Conditional Probability

Let's think about two independent experiments.

Event (B) influences the probability of occurrence another event (A)

↓

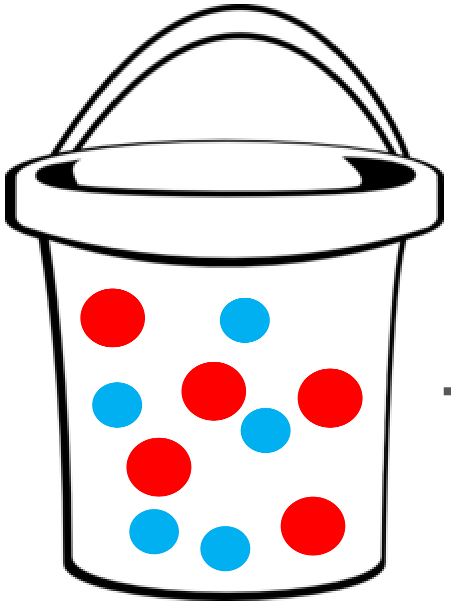
Conditional Probability



Tsunami

$$p(A \mid B) = \frac{p(A \cap B)}{p(B)}$$

Conditional Probability



2 balls are drawn without replacement

If 1st ball be blue

Probability that the second ball is also blue?

Bayesian Probability

It's based on a theorem of refining the probability of A given B

$$p(A \mid B) = \frac{p(B \mid A)\, p(A)}{p(B \mid A)\, p(A) + p(B \mid \bar{A})\, p(\bar{A})}$$

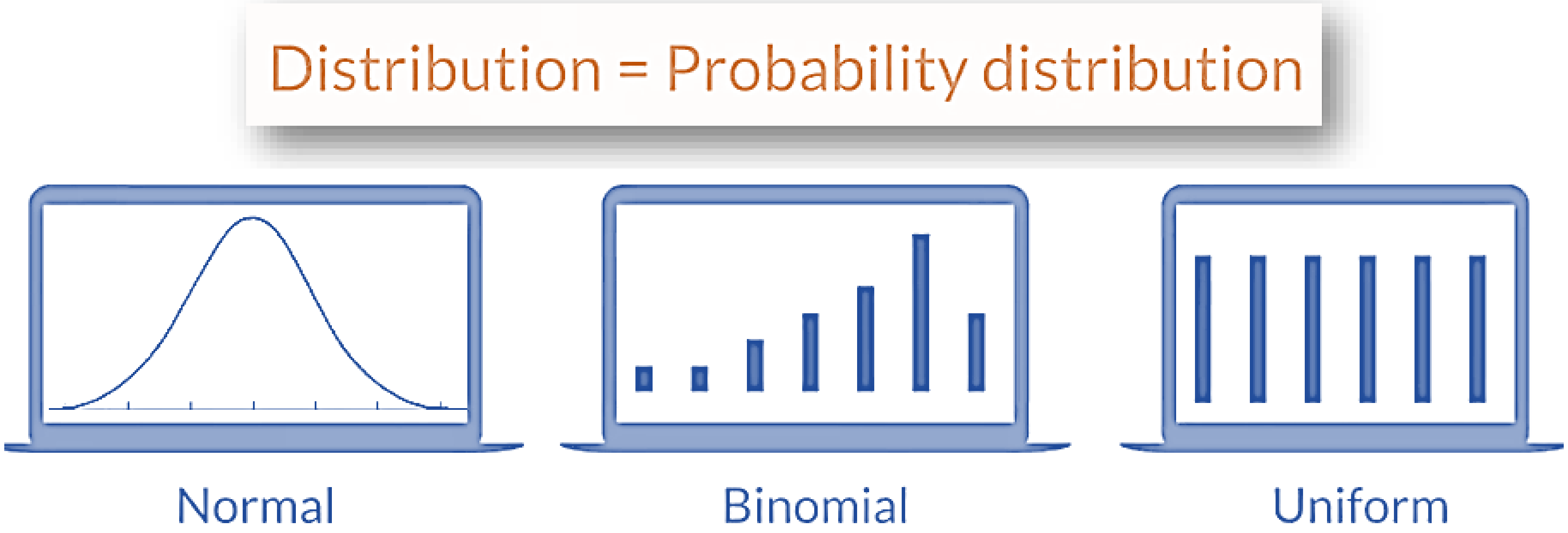
$$P(A|B) = \frac{P(A)\ P(B|A)}{P(B)}$$

Another Version

Distribution

(Probability Distribution)

- Distribution is another name for a set of numbers.
- A distribution is a function that shows the possible values for a variable and how often they occur



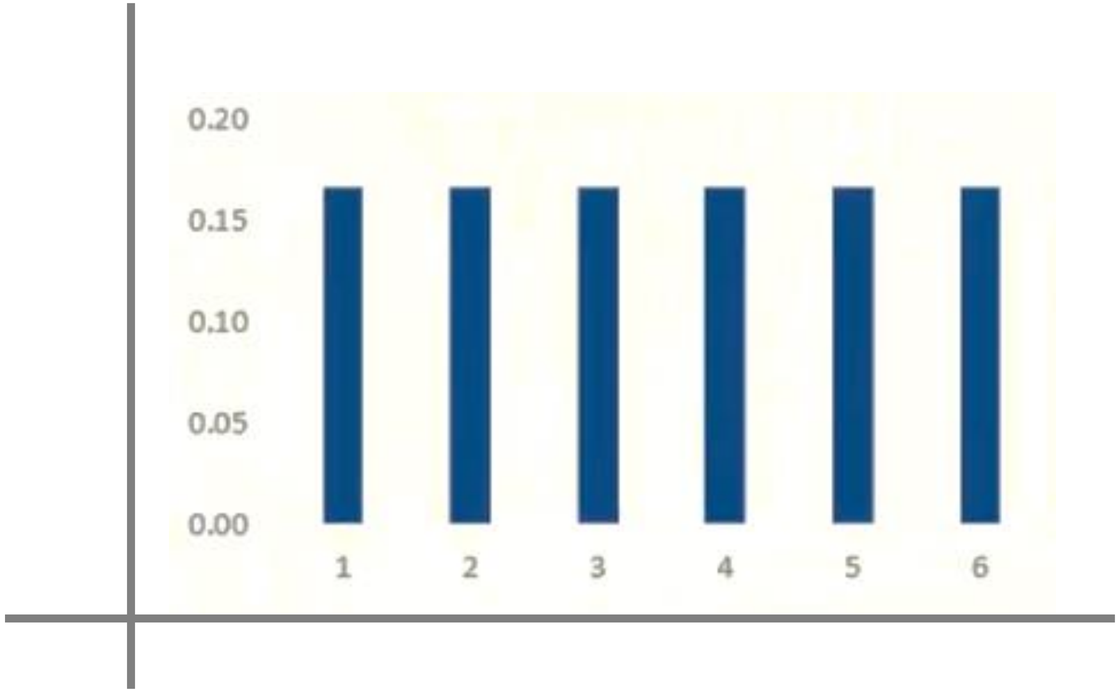
Different Distributions: http://www.stat.tamu.edu/~twehrly/611/distab.pdf

Rolling A die

Probability distribution is associated with a graph describing the likelihood of occurrence of every event

Outcome	Probability	
1	1/6	0.17
2	1/6	0.17
3	1/6	0.17
4	1/6	0.17
5	1/6	0.17
6	1/6	0.17
7	0	

Likelihood of occurrence of every event



Uniform Distribution

Rolling Two Dice

Discrete Distribution

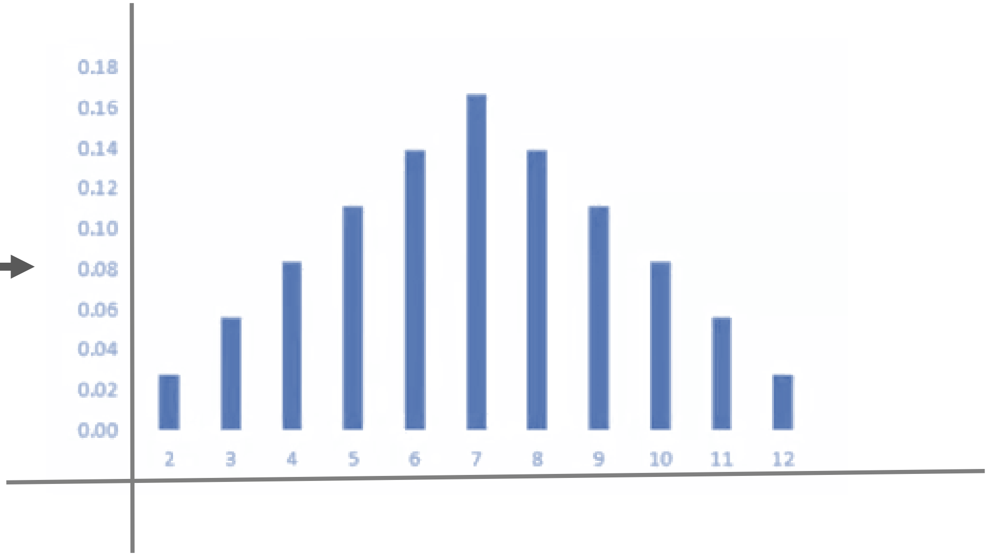
Possibilities?

(1,1) (2,1) (3,1) (4,1) (5,1) (6,1)
(1,2) (2,2) (3,2) (4,2) (5,2) (6,2)
(1,3) (2,3) (3,3) (4,3) (5,3) (6,3)
(1,4) (2,4) (3,4) (4,4) (5,4) (6,4)
(1,5) (2,5) (3,5) (4,5) (5,5) (6,5)
(1,6) (2,6) (3,6) (4,6) (5,6) (6,6)

Possible Combinations

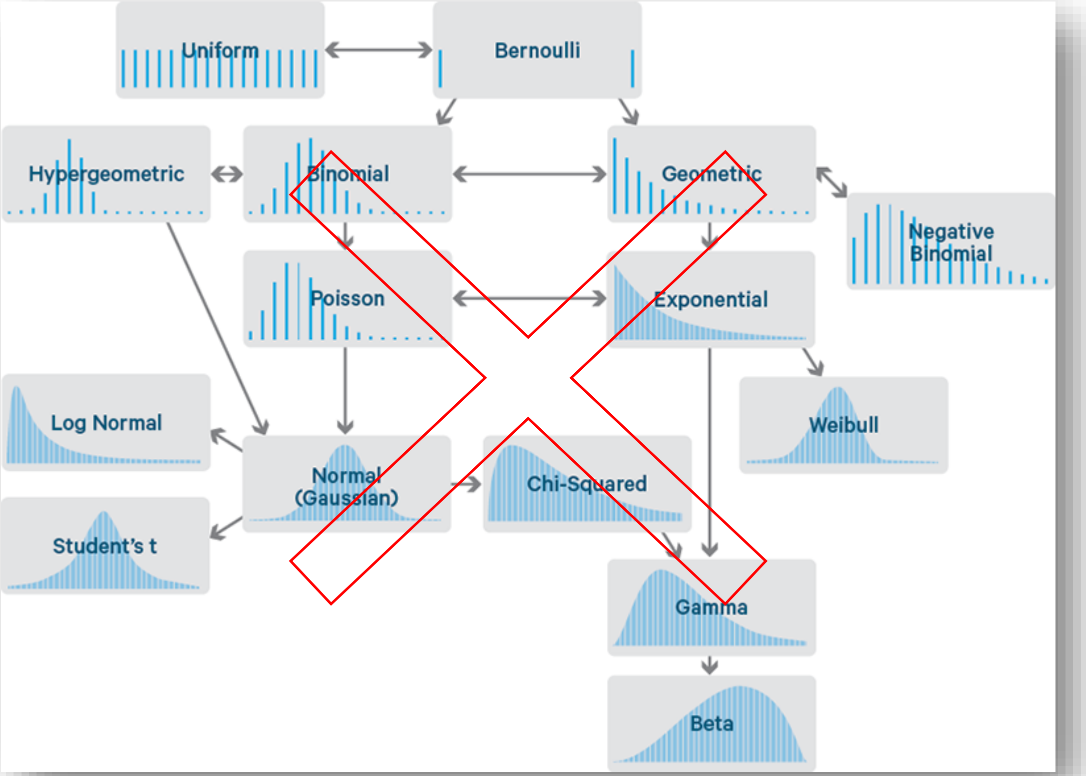
Outcome	Probability
2	0.03
3	0.06
4	0.08
5	0.11
6	0.14
7	0.17
8	0.14
9	0.11
10	0.08
11	0.06
12	0.03
All else	0

Sum of two dice



Binomial Distribution

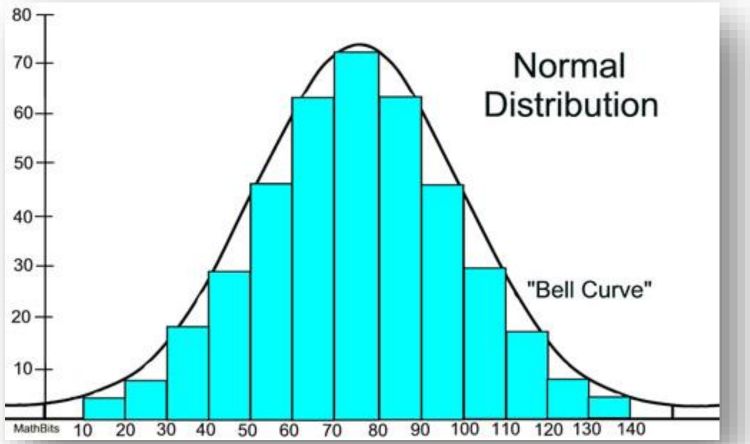
Distribution Of Our Focus



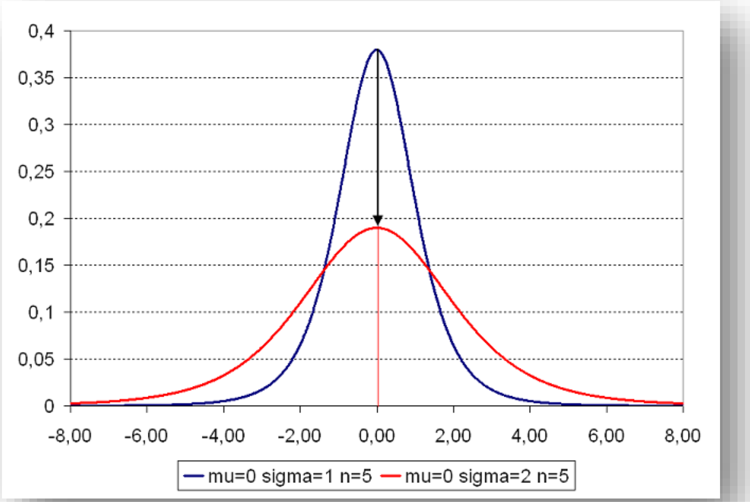
- Normal Distributions
- Students Distributions

Distribution Of Our Focus

(Continuous Distribution)



Normal Distribution



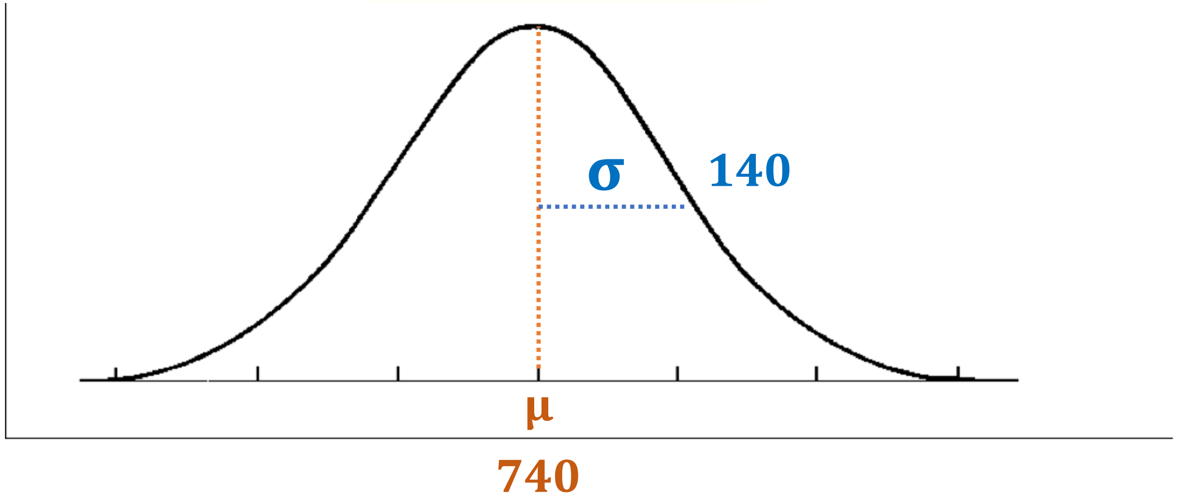
Students T Distribution

- Decisions based on ND have good track record
- Distributions of sample means with large enough samples size could be approximate to normal.

Distribution

(Normal Distribution)

$N \sim (\mu, \sigma^2)$



Gaussian Distribution
Bell Curve

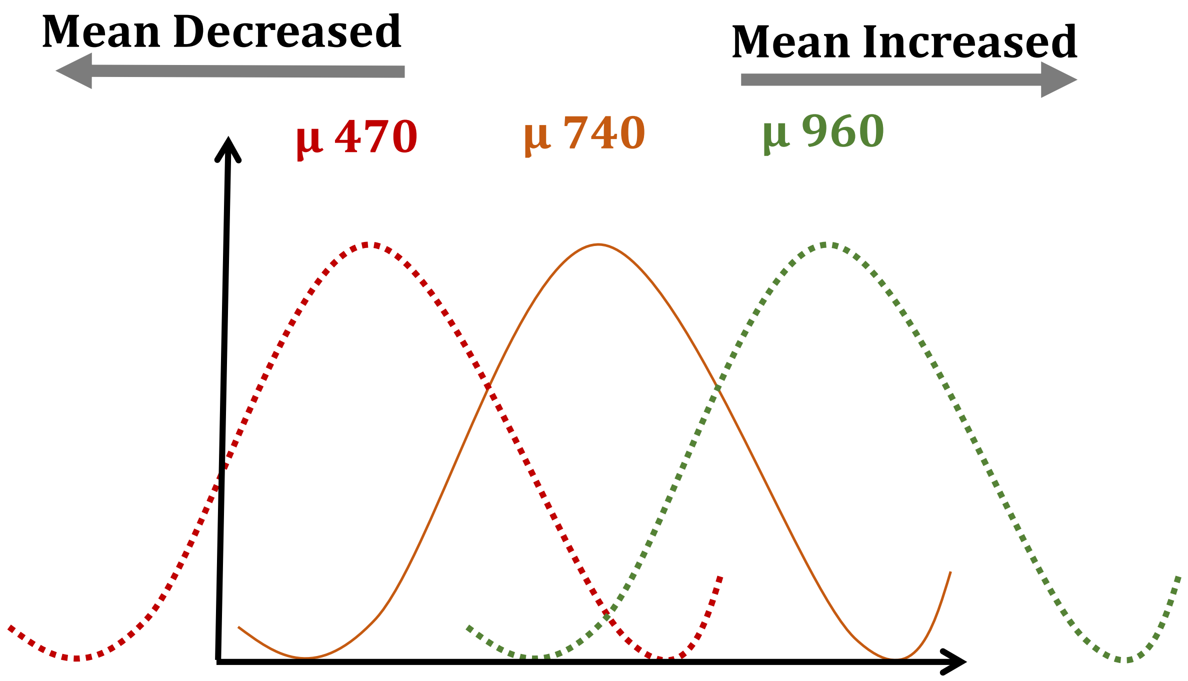
Symmetrical
No Skew
Mean Median Mode are equal

Imagine the above data plot has mean 740 and standard deviation 140

What if Mean is smaller or larger

Normal Distribution

(Controlling the Standard Deviation)

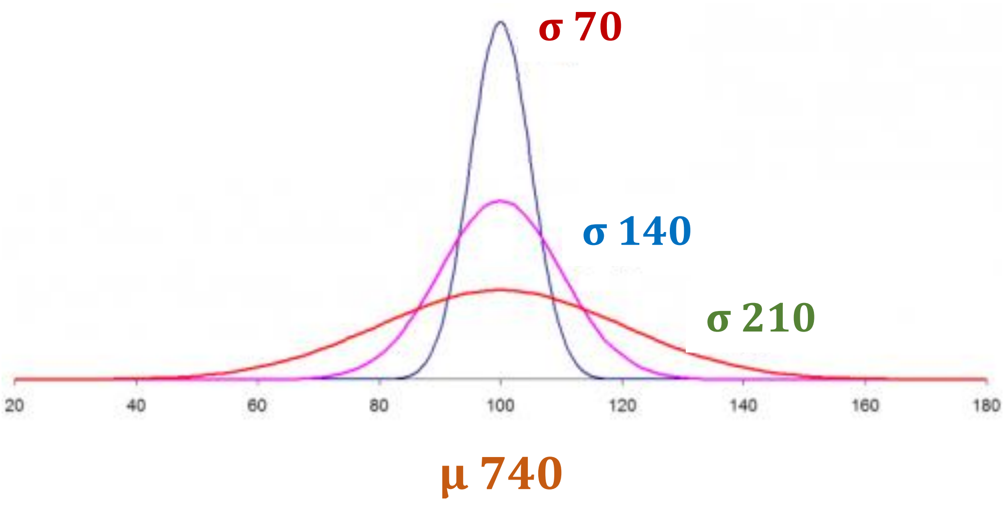


Mean Decreased: Distribution moves towards left
Mean Increased: Distribution moves towards right

What if Standard Deviation is smaller or larger

Normal Distribution

(Controlling the Mean)

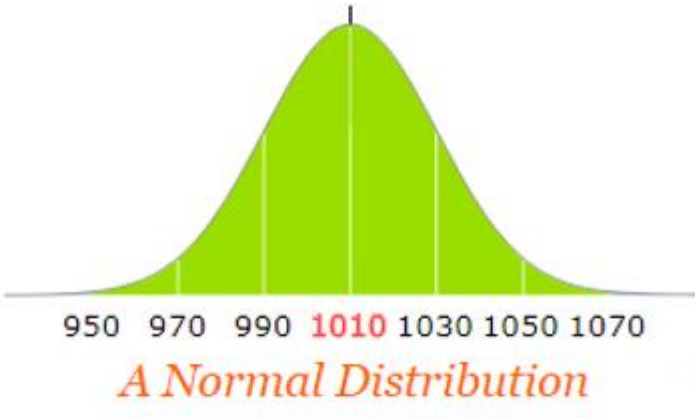


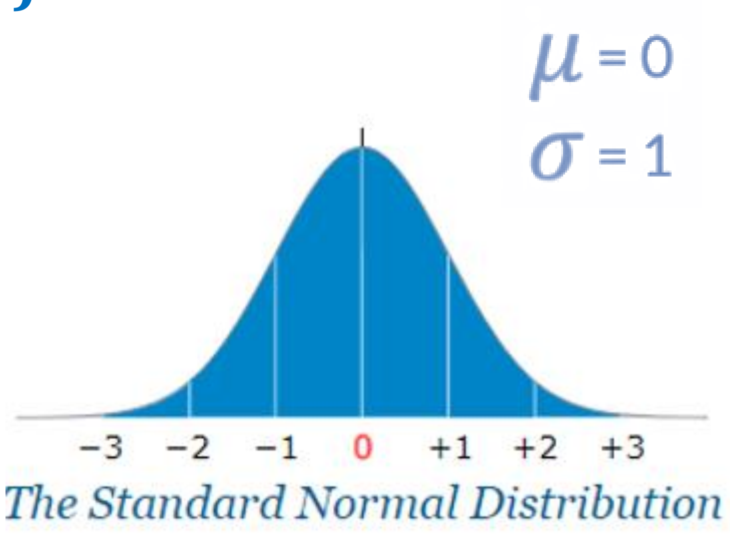
Observation:
No Moment but changing the shape

Std. Deviation Decreased: More data will be in the middle, sharp tail
Std. Deviation Increased: Less data will be in the middle, flat tail

Distribution

(Standard Normal Distribution)





Normal Distribution Variable

Standard Normal Distribution Variable
Z score Variable

$$Z\ score = \frac{Original\ Variable - Mean}{Standard\ Deviation}$$

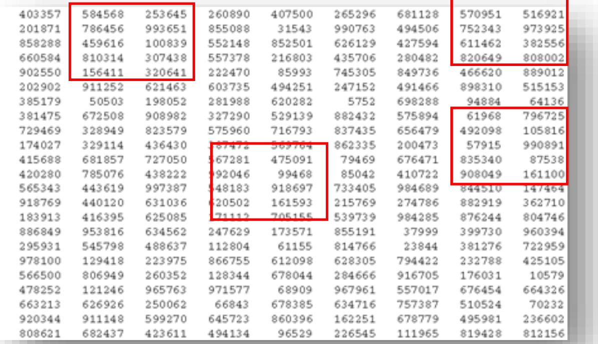
https://www.mathsisfun.com/data/standard-normal-distribution.html

Why Standardize?

- Compare different normally distributed datasets
- Detect normality
- Detect outliers
- Create confidence intervals
- Test hypotheses
- Perform regression analysis

Central Limit Theorem

Data



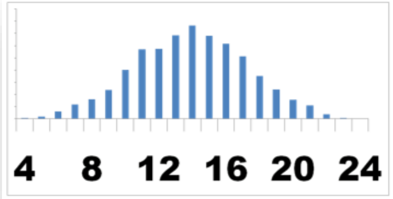
403357	584568	253645	260890	407500	265296	681128	570951	516921
201871	786456	993651	855088	31543	990763	494506	752343	973925
858288	459616	100839	552148	852501	626129	427594	611462	382556
660584	810314	307438	557378	216803	435706	280482	820649	808002
902550	156411	320641	222470	85993	745305	849736	466620	889012
202902	911252	621463	603735	494251	247152	491466	898310	515153
385179	50503	198052	281988	620282	5752	698288	94884	64136
381475	672508	908982	327290	529139	882432	575894	61968	796725
729469	328949	823579	575960	716793	837435	656479	492098	105816
174027	329114	436430	[illegible]	[illegible]	862335	200473	57915	990891
415688	681857	727050	567281	475091	79469	676471	835340	87538
420280	785076	438222	992046	99468	85042	410722	908049	161100
565343	443619	997387	548183	918697	733405	984689	844510	147464
918769	440120	631036	620502	161593	215769	274786	882919	362710
183913	416395	625085	171112	705155	539739	984285	876244	804746
886849	953816	634562	247629	173571	855191	37999	399730	960394
295931	545798	488637	112804	61155	814766	23844	381276	722959
978100	129418	223975	866755	612098	628305	794422	232788	425105
566500	806949	260352	128344	678044	284666	916705	176031	10579
478252	121246	965763	971577	68909	967961	557017	676454	664326
663213	626926	250062	66843	678385	634716	757387	510524	70232
920344	911148	599270	645723	860396	162251	678779	495981	236602
808621	682437	423611	494134	96529	226545	111965	819428	812156

Take Samples

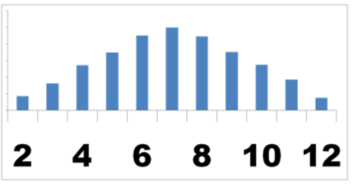
Calculate Sample Mean

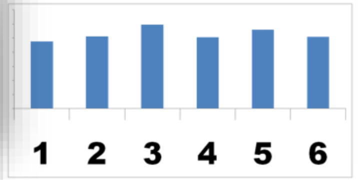
Mean of Samples

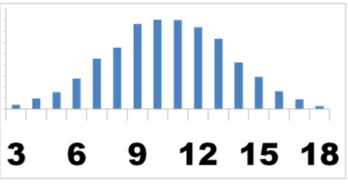
μ1 470
μ2 513
μ3 600
μ4 490
μ5 555
μ6 601



Normal Distribution





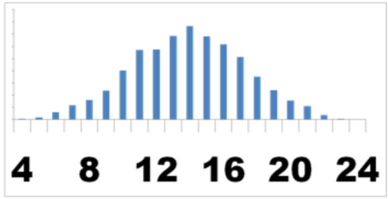


No matter the what distribution

More **k** More **n**

Mean of All Samples

Approximate to



Normal Distribution

Standard Error

Mean of Samples

μ1 470
μ2 513
μ3 600
μ4 490
μ5 555
μ6 601

Samples Distribution

Standard Deviation between the sample means

More the Standard Deviation More the Standard Error

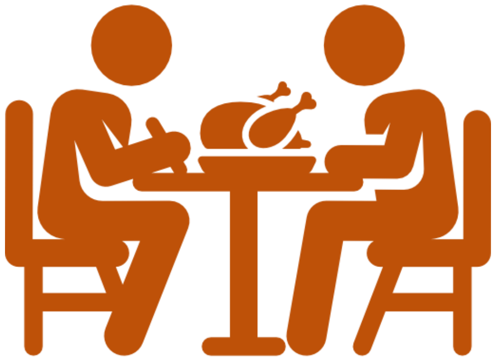
Sample Size ↑ Standard Error ↓

Confidence Interval

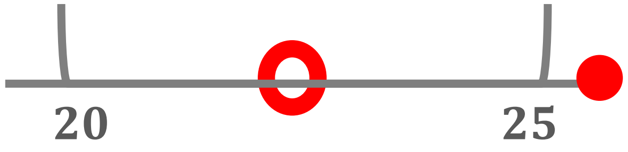
A Confidence Interval is a range of values we are fairly sure our true value lies in

Hyderabad Restaurant

You visit 5% of the restaurants



22.50 Rupees



→ 95 % confident that the population parameter lies between 20 and 25

Level of Confidence

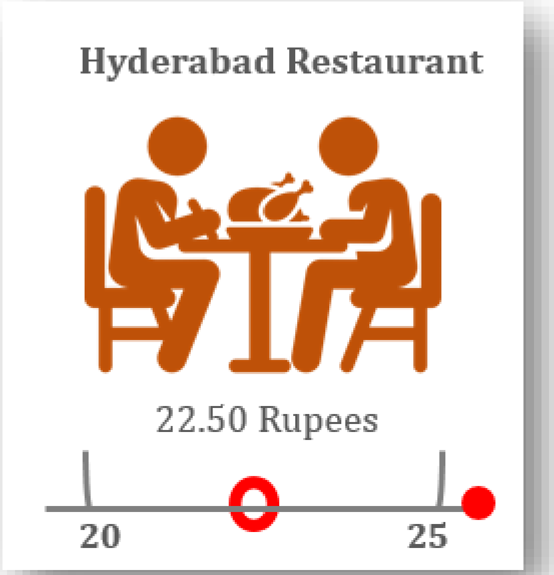
Denotation

$$1-\alpha$$

$$0 \leq \alpha \leq 1$$

α	Confidence level
5%	95%
1%	99%

Hypothesis Testing



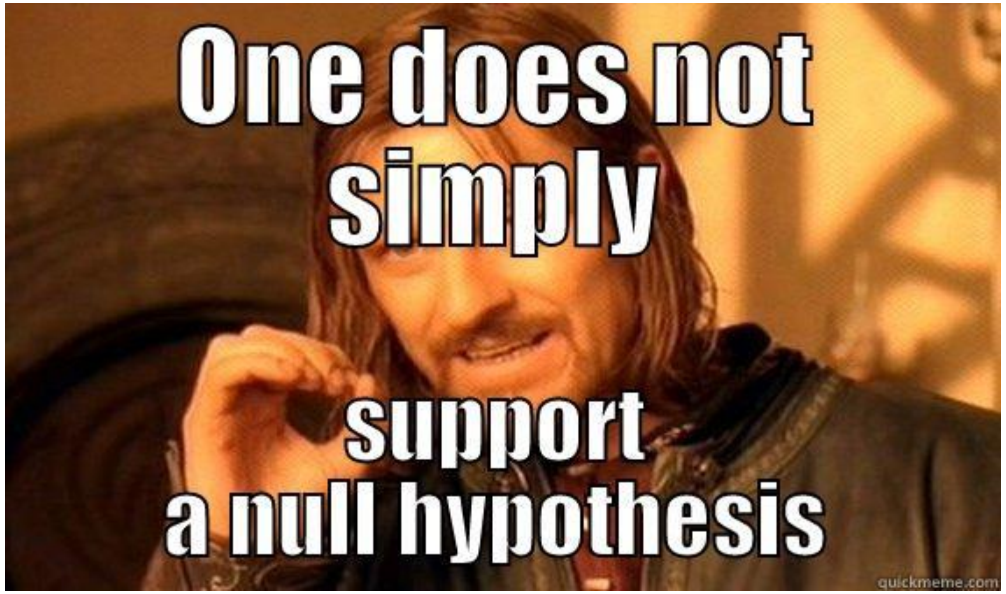
Hypothesis Testing

Hypothesis Testing

A hypothesis is an idea that can be tested

↓

Evaluates a claim made about the value of a population parameter by using a sample statistic.



Hypothesis Testing

Apples in Hyderabad are expensive → Not Testable

Price of Apples in Hyderabad > 100/- → Hypothesis

→ Compare Employment rate or Administration in Obama period and George bush period

Hypothesis Testing

Hypotheses	Notation
Null hypothesis	H_0
Alternative hypothesis	H_1 or H_A

$H_0: \mu_0 = \$\ 113{,}000$

$H_1: \mu_0 \neq \$\ 113{,}000$

Data Scientist

Mean data scientists salary the US is 113000$

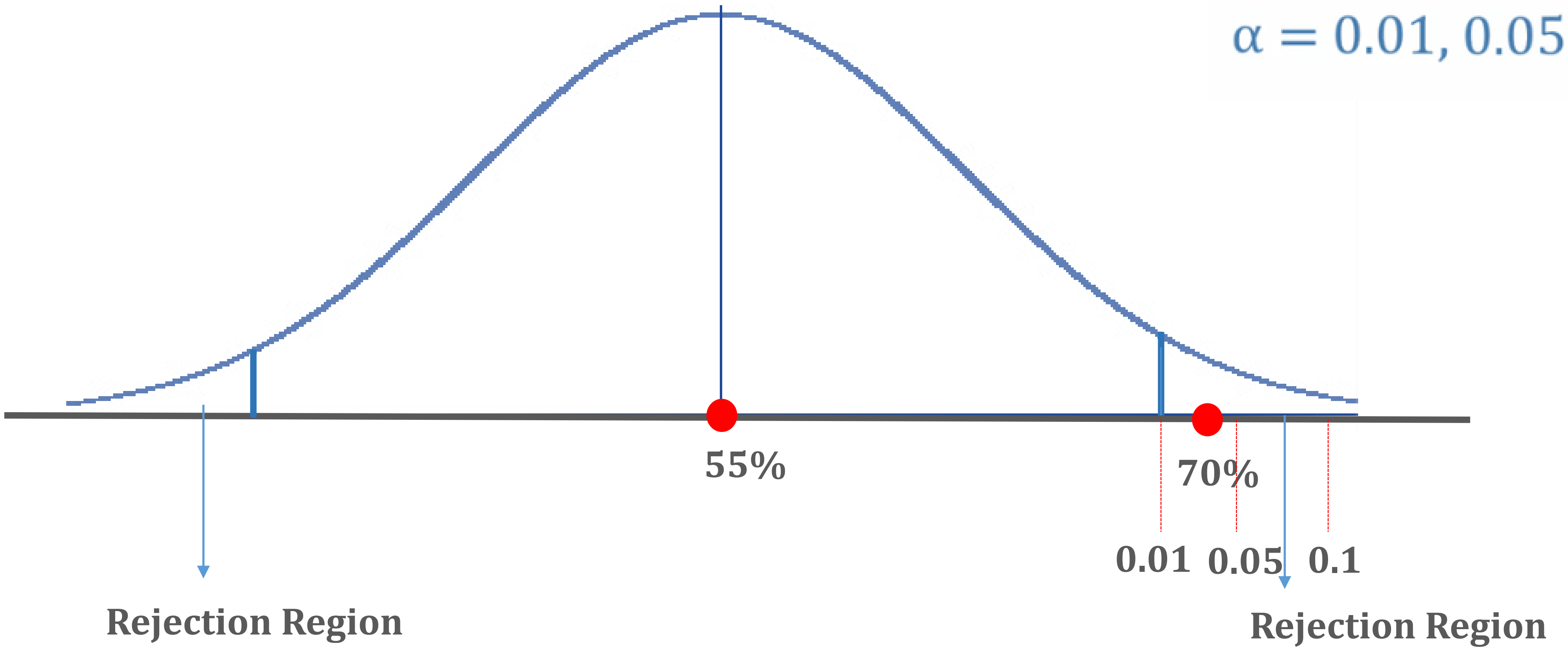
Generally the researcher will be trying to reject the null hypothesis

Hypothesis Testing

$H_0\colon \mu_{grade} = 70\%$

$H_1\colon \mu_{grade} \neq 70\%$

$\alpha = 0.01, 0.05, 0.1$



Hypothesis Testing

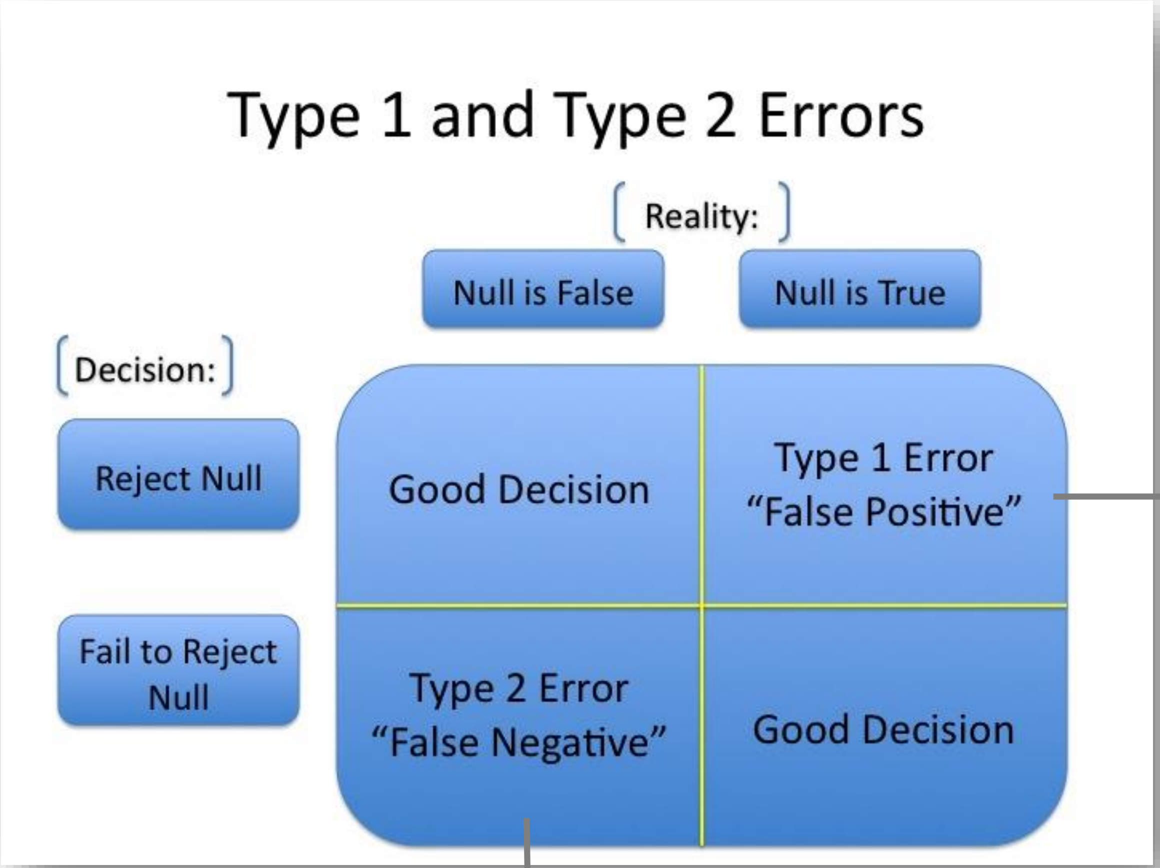
Low significance levels

High significance levels

Hypothesis Testing Errors

Type I Error	Type II Error
Reject a true null hypothesis	Accept a false null hypothesis
False Positives	False Negative
Alpha	Beta
You are Responsible for this error	Depends mainly on sample size and population variance

Hypothesis Testing



Has Disease
But Predicts No Disease

Has No Disease
But Predicts Disease

To Understand The Correlation

Type of Test	Use of the Test	Data Type
Pearson correlation	To understand association between two continuous variables	Continues Variables
Spearman correlation	To understand association between two ordinal variables	Ordinal Variable (Categorical)
Chi-square	To understand association between two categorical variables	Categorical

To understand difference between the means of variables

Type of Test	Use of the Test	Data Type
Paired T-test	Tests for the difference between two related variables	Continues Variables
Independent T-Test	Tests for the difference between two independent variables	Continues Variables
ANOVA	Tests the difference between group means (More than two) Differences in the means of 3+ independent groups for one variable	Continues Variables

https://www.sheffield.ac.uk/mash/what_test

BASIS FOR COMPARISON	T-TEST	Z-TEST
Meaning	T-test refers to a type of parametric test that is applied to identify, how the means of two sets of data differ from one another when variance is not given.	Z-test implies a hypothesis test which ascertains if the means of two datasets are different from each other when variance is given.
Based on	Student-t distribution	Normal distribution
Population variance	Unknown	Known
Sample Size	Small	Large

Ex:
Measuring the average diameter of shafts from a certain machine when you have a small sample

Ex:
Comparing the average engineering salaries of men versus women

Regression: Assess if change in one variable predicts change in another variable

Type of Test	Use of the Test	Data Type
Simple Regression	Tests how change in the predictor variable predicts the level of change in the outcome variable	Continues Variables
Multiple regression	Tests how change in the combination of two or more predictor variables predict the level of change in the outcome variable	Continues Variables


Namaste