Process of Data Analysis & Common different statistical Test

Prof. Dr. Kazi Yesmin
Department of Microbiology
Green Life Medical College

Objectives

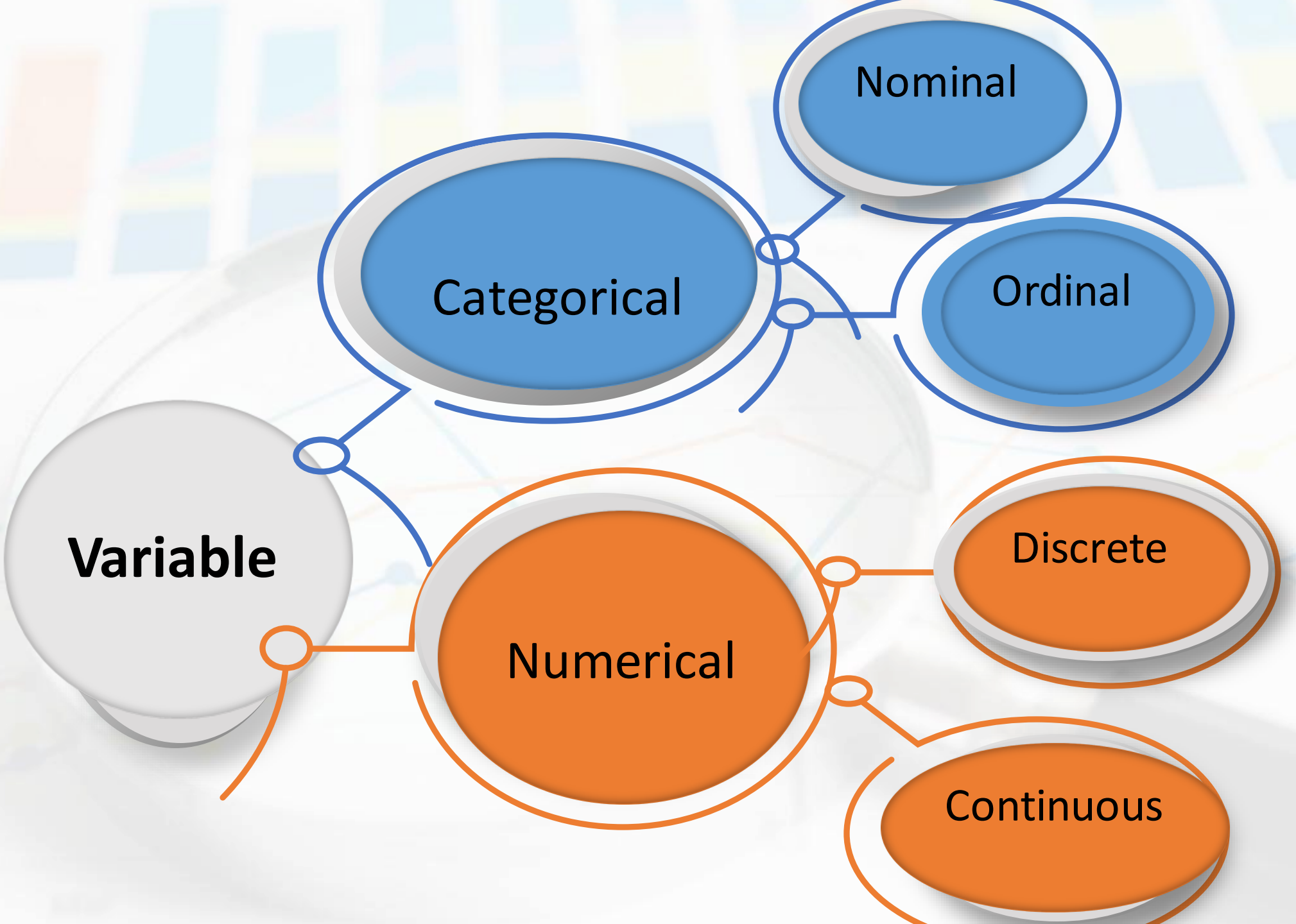
- Define variables & Data and different types of variables
- Independent variables & dependent variables
- Normal distribution curve
- Process of Data analysis
- Categorical data and numerical data analysis
- Descriptive analysis
- Inferential analysis
- Hypothesis –types & steps
- Types of Statistical test
- Hypothesis testing

Variable & Data

A variable is a characteristic whose value varies from person to person, object to object or from phenomenon to phenomenon.

Data are the facts and figures collected, analyzed, and summarized for presentation and interpretation.

All the data collected in a particular study are referred to as the data set for the study.


Variable
Categorical
Nominal
Ordinal
Numerical
Discrete
Continuous

Categorical Data

Nominal

Sex of study participant
Religion
Resident

Ordinal

Education level
Income level
Satisfaction ranking

Numerical Data

Discrete

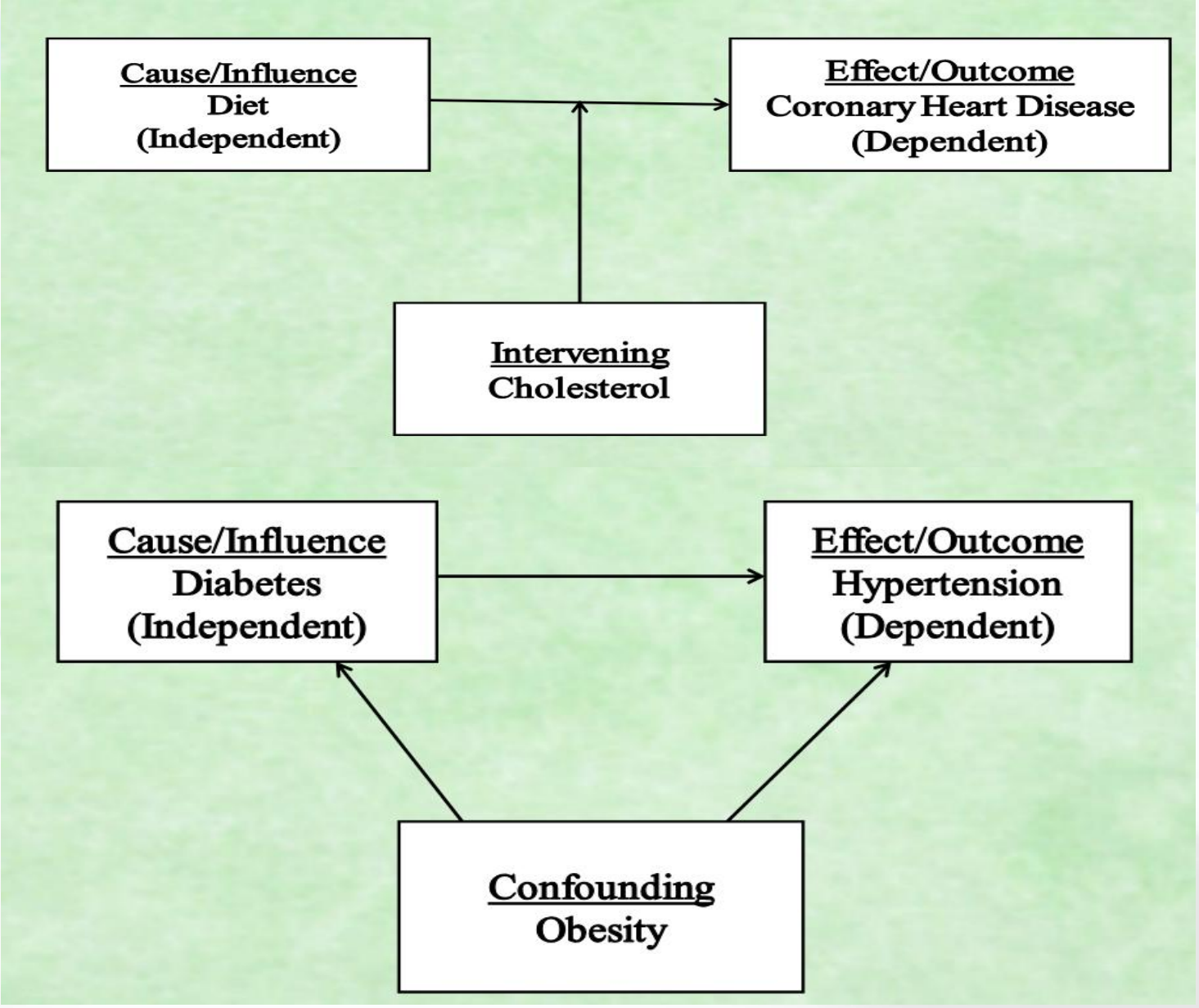
No. of children
No. of hospital beds
No. of family member

Continuous

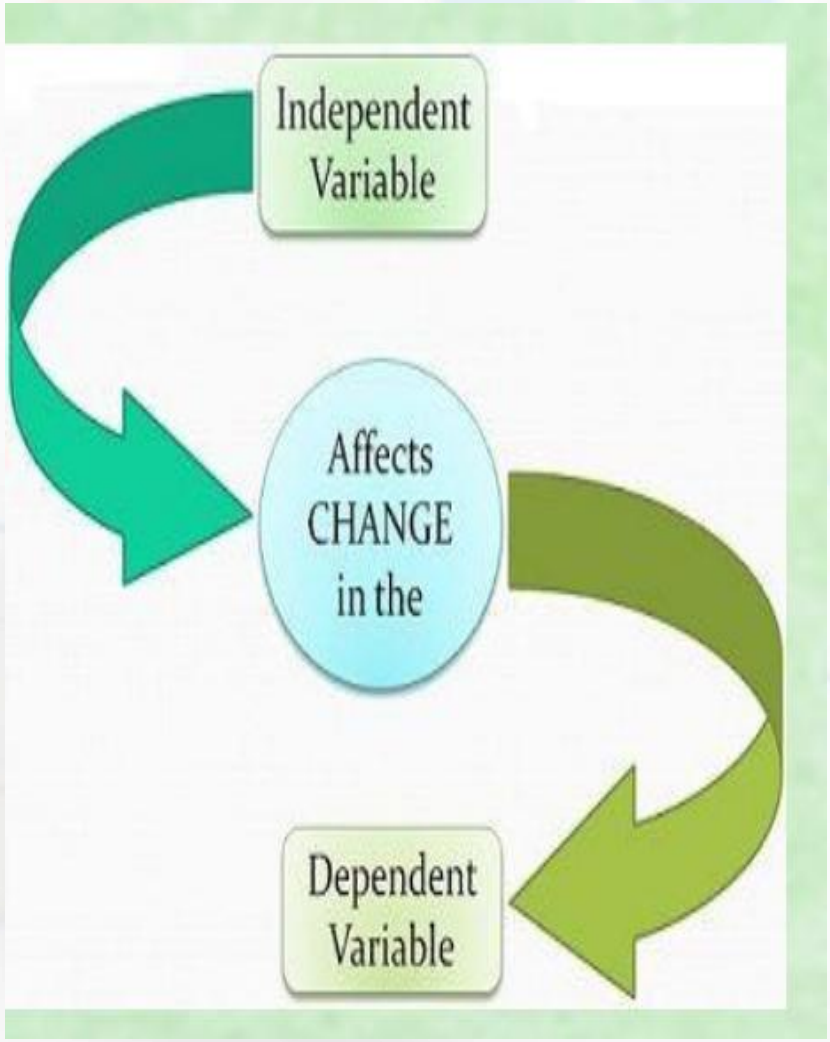
Age, height, weight, blood sugar etc.

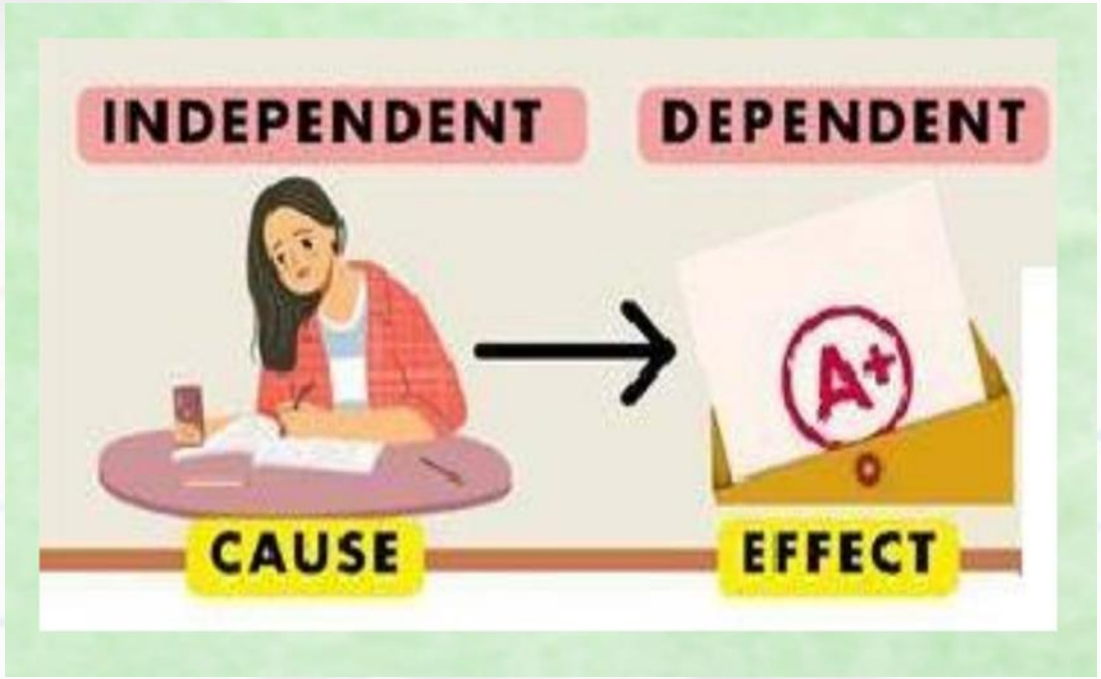
Other variables

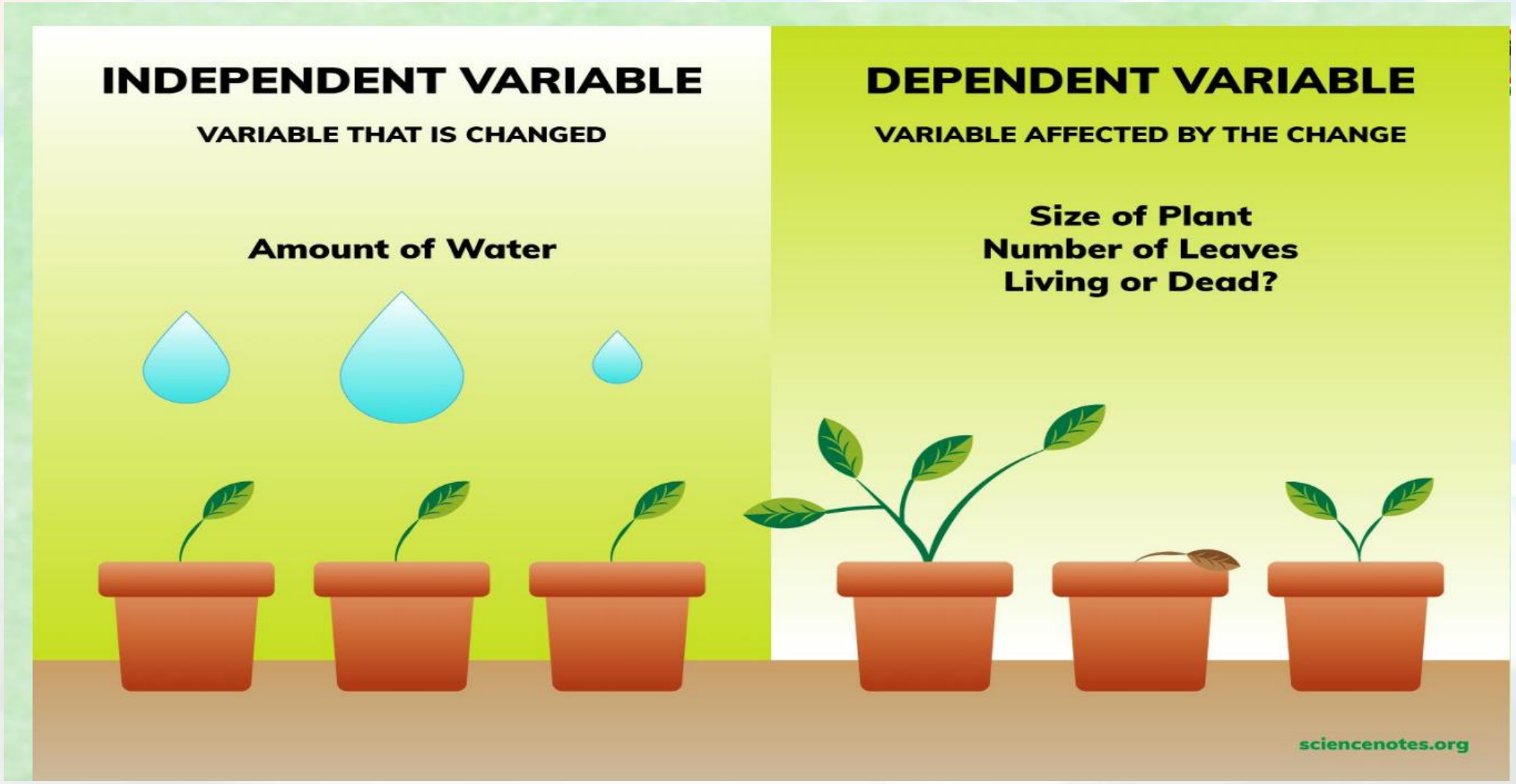
- Independent variables /Predictor variables
- Dependent variables /Outcome variables
- Confounding variables
- Background variables
- Intervening variables



Independent and dependent variable



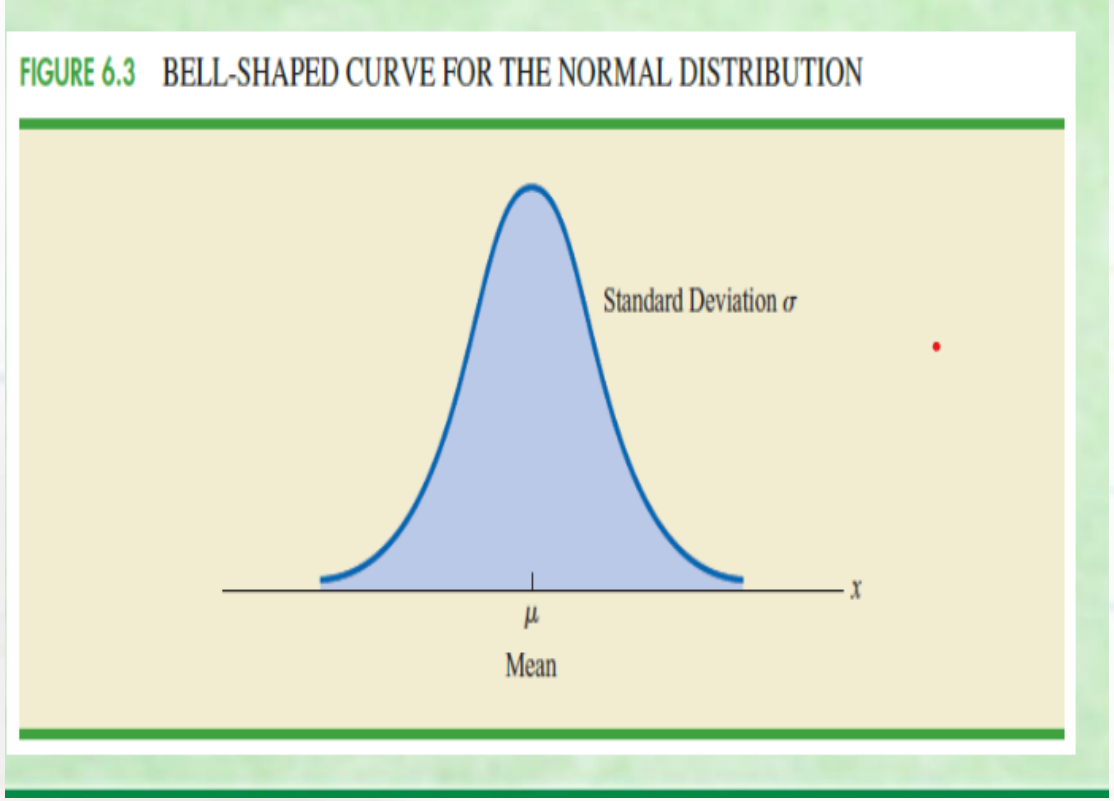


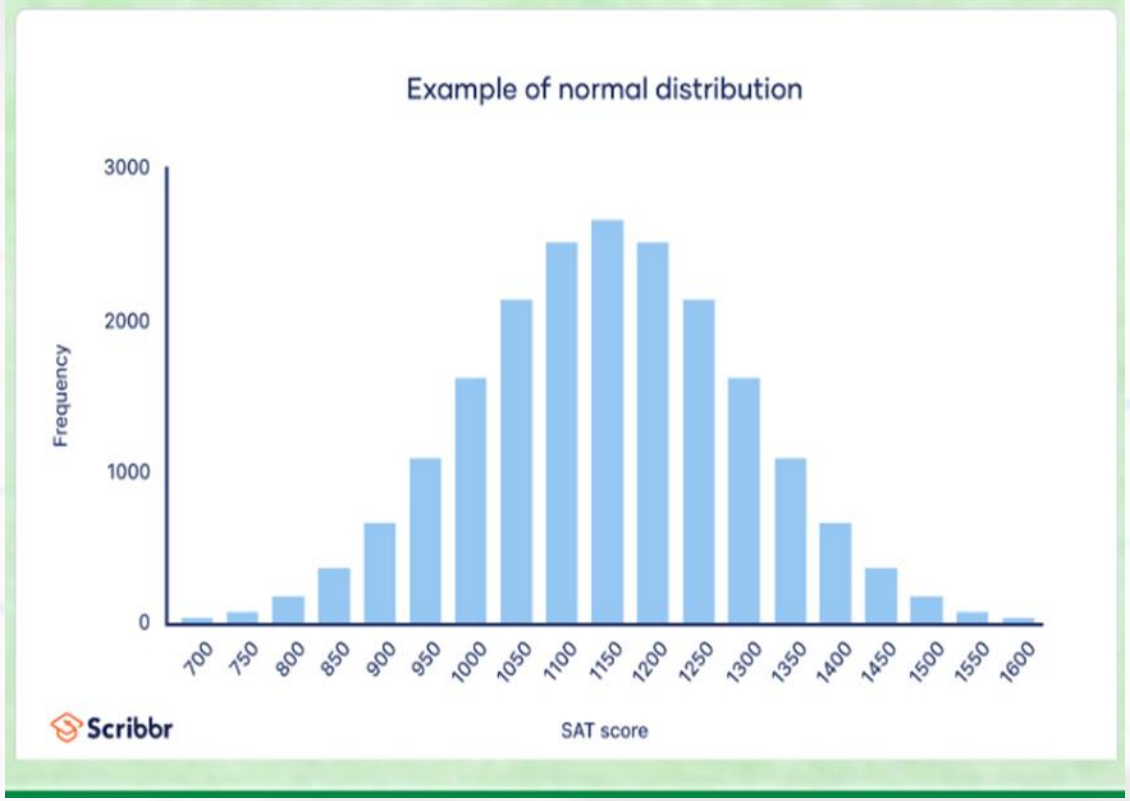

INDEPENDENT VARIABLE
VARIABLE THAT IS CHANGED
Amount of Water
DEPENDENT VARIABLE
VARIABLE AFFECTED BY THE CHANGE
Size of Plant
Number of Leaves
Living or Dead?
sciencenotes.org

Normal distribution curve

- Data is symmetrically distributed with no skew.
- When plotted on a graph the data follows a bell shape with most values clustering around a central region.
- It is a probability distribution.
- The mean, median and mode are all equal.

Normal distribution curve





Process of data analysis

- Descriptive analysis
- Inferential analysis
- Predictive analysis
- Exploratory analysis
- Causal analysis
- Time-series analysis
- Survival analysis

Common tools for analysis

1. R
2. Stata
3. SPSS etc.

Descriptive analysis

1. Frequency distribution

   group data for table, figure, histogram etc.

2. Measurement of Central tendency

   i) Mean

   ii) Median

   iii) Mode

3. Measurement of Dispersion

   i) Standard deviation

   ii) Range- minimum & maximum

   iii) Percentile & Quartile

Presentation of categorical data

1. Frequency distribution – by tables, bar & pie charts
2. Percentages
3. Mode
4. Contingency tables, cross tabulation

Table 1: Distribution of study participant by gender (n=2100)

Gerder	Frequency (n)	Percentage (%)
Male	1176	56.0
Female	924	44.0
Total	2100	100.0

Presentation of categorical data

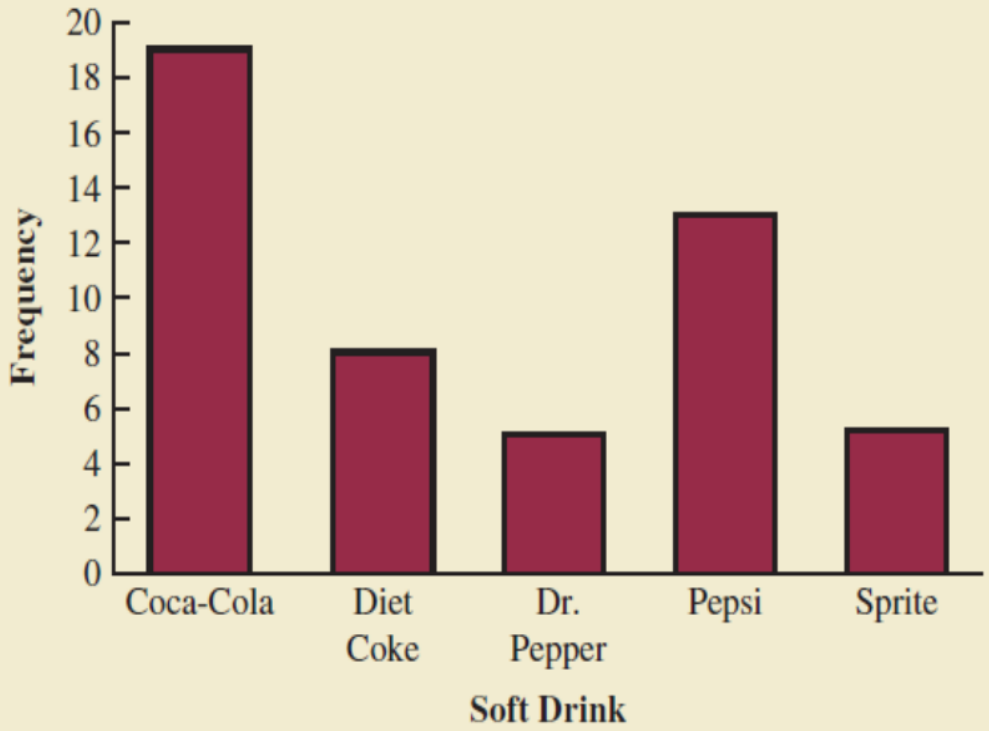
Table 1. Distribution of the Respondents by their Religions (n=423)

Religions	Frequency (n)	Percentage (%)
Islam	373	88.2
Hindu	47	11.1
Christian	3	0.7
Total	**423**	**100**

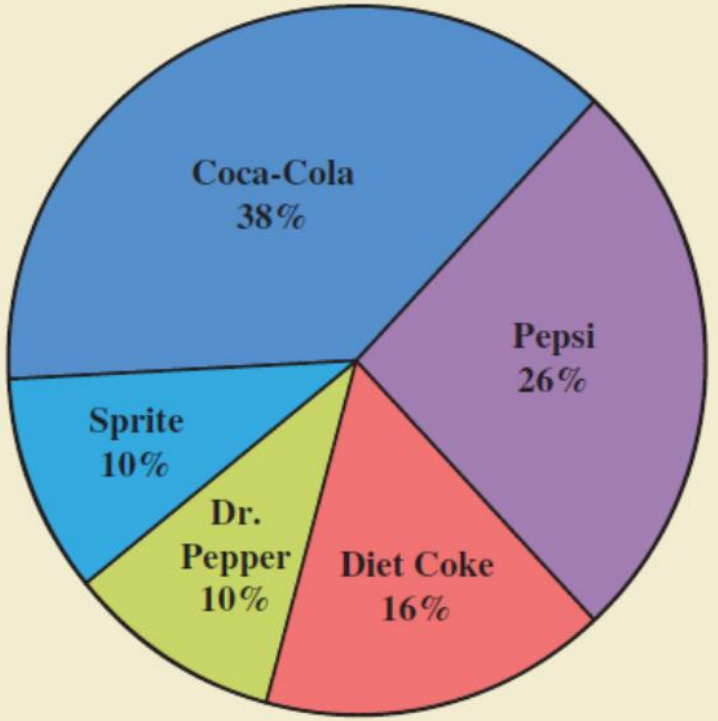
Table 1 shows the religious affiliation of the respondents. The vast majority (88.2%) identified as Muslim, while 11.1% were Hindu and only 0.7% were Christian. These findings indicate that **Islam was the predominant religion** among the participants, which aligns with the national religious composition of Bangladesh.

Presentation of categorical data

BAR CHART OF SOFT DRINK PURCHASES



PIE CHART OF SOFT DRINK PURCHASES



Presentation of numerical data

If data follows normal distribution

- Frequency distribution- frequency table, bar and pie charts
- Measurement of central tendency- Mean, median & mode
- Range, variance, co-variance & Standard deviation
- Line chart
- Histogram and

If data does not follow normal distribution

- Median
- Mode
- Box & Whisker plot
- Ogive
- Stem-and-leaf display

Measurement of central tendency

- The mean (or arithmetic mean) of a set of data is the measure of center found by adding all of the data values and dividing the total by the number of data values.

For a sample of size n:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + \square + x_n}{n}$$

Observed values

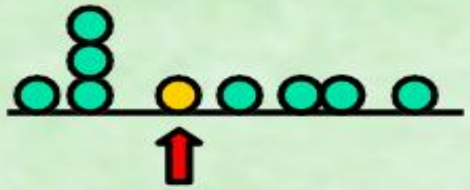
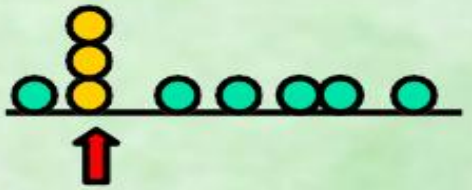
Sample size

- The median of a data set is the measure of center that is the middle value when the original data values are arranged in order of increasing (or decreasing) magnitude.
- The mode of a data set is the value(s) that occurs with the greatest frequency.

Measures of Central Tendency

Overview

Central Tendency

- **Mean**
  - $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$
  - Arithmetic average
- **Median**
  - 
  - Midpoint of ranked values
- **Mode**
  - 
  - Most frequently observed value

Standard Deviation

- The standard deviation of a set of sample values, denoted by s, is a measure of how much data values deviate away from the mean.

- Sample standard deviation:

$$S = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$$

Presentation of numerical data

Table 2: Distribution of participant by age category (n=2100)

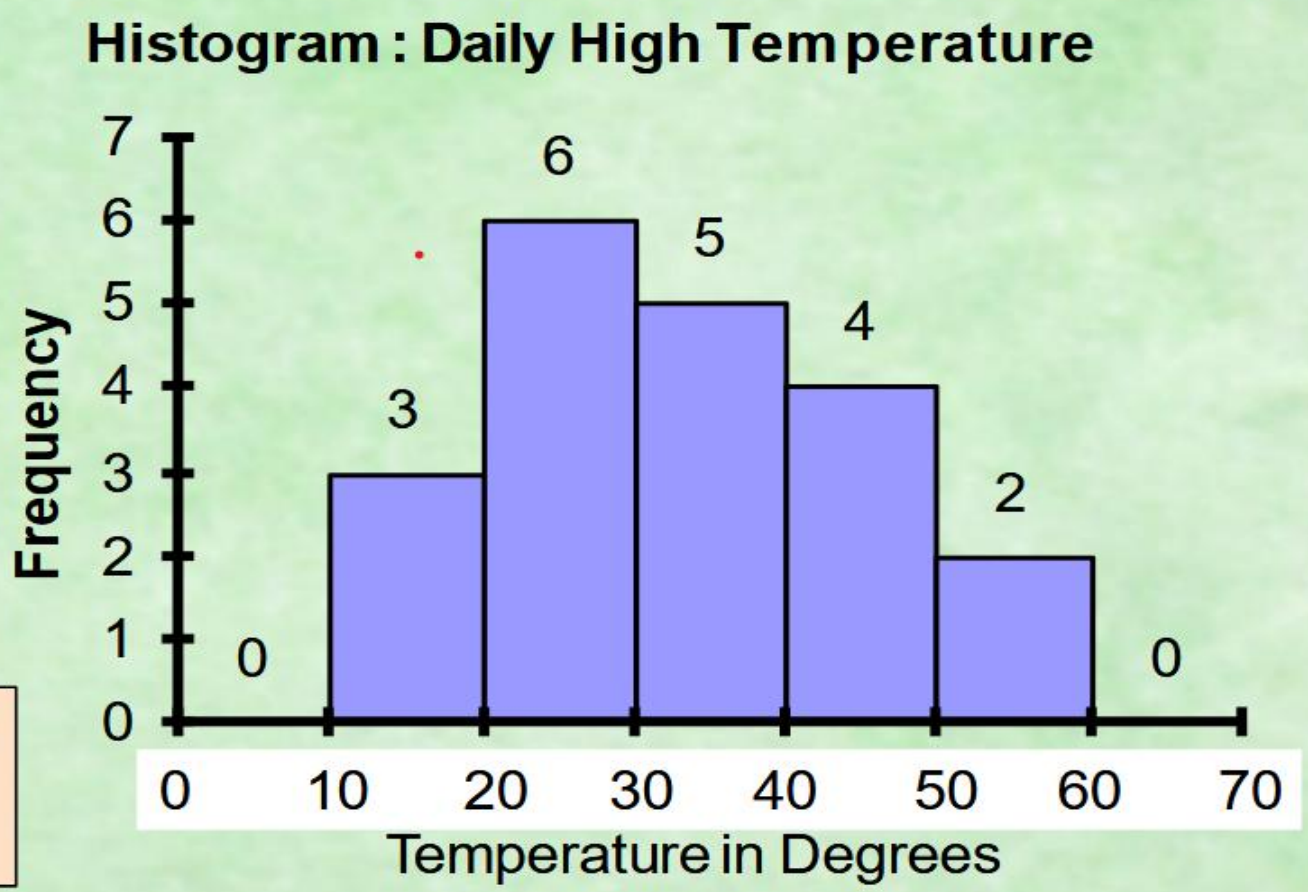
Age category	Frequency (n)	Percentage (%)
≤22 years	1099	52.3
> 22 years	1001	47.7
Total	2100	100.0
Mean± SD= 22.58± 2.224, Max= 35 Min =18		

Line chart


Magazine Subscriptions by Year
Thousands of subscribers
350
300
250
200
150
100
50
0
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006

Presentation of numerical data

Interval	Frequency
10 but less than 20	3
20 but less than 30	6
30 but less than 40	5
40 but less than 50	4
50 but less than 60	2



(No gaps between bars)

Inferential analysis

Hypothesis testing

- Deciding about the value of a parameter based on preconceived hypothesis.
- **Parametric tests**
- **Non-parametric test**

Estimation

- Estimating the value of the parameter
- **Point estimation**
- **Interval estimation**

Hypothesis

- A hypothesis is a tentative statement about the relationship between two or more variables/ What the researchers predict the relationship between two or more variables?

- **Two types of hypothesis-**

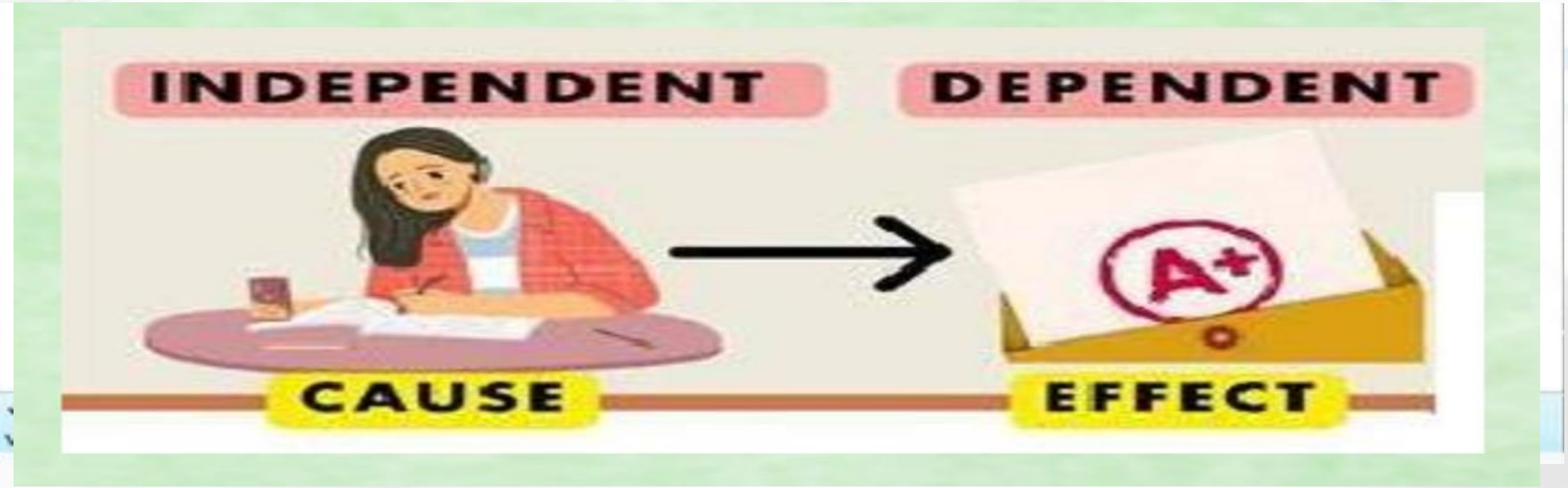
1. Null hypothesis
2. Alternate / Research hypothesis

Null hypothesis

- There is no association/ relationship between two or more variables
- It is denoted by Ho

Alternate hypothesis

- There is association/ relationship between two or more variables
- It is denoted by Ha/ H_1



Steps of Hypothesis test

STEP 01
Formulating Ho hypothesis

STEP 02
Selecting a significance level

STEP 03
Choosing an appropriate statistical test

STEP 04
Computing the test statistic

STEP 05
Determining the critical region & p-value

STEP 06
Making decision & draw conclusion

Interpretation of Hypothesis testing

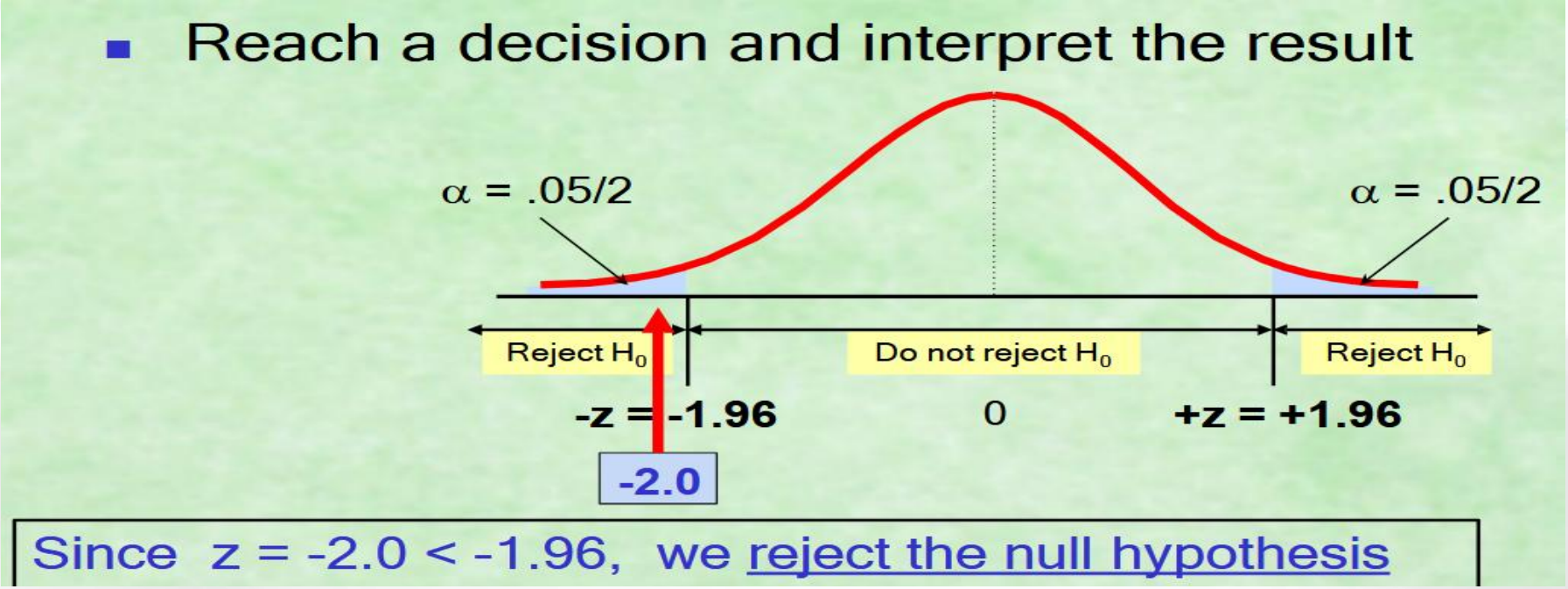
- By p value
- By calculated test value such as t test, Z test, Chi-square test

P-value

- Small p (≤0.05) reject null hypothesis. This is strong evidence.
- Large p (>0.05) don’t reject the null hypothesis.

Interpretation of hypothesis test

- If p value ≤ 0.05 ⟹ null hypothesis is rejected & accept alternate hypothesis at 95% confidence interval in the significance level. So, there is an association /relationship between variables.
- If p value > 0.05 ⟹ not reject the null hypothesis at 95% confidence interval. So, there is no association between variables



Statistical test

Parametric test

- One sample t test
- Independent t test
- Pair t test
- Z test
- One-way/two-way ANOVA
- ANCOVA
- MANCOVA
- Pearson correlation

Non-parametric test

- Chi-square test
- Sign test
- Mann-Whitney U test
- Wilcoxon test
- McNemar
- Kruskal-Wallis test
- Kolmogorov-Smirnov test
- Spearman correlation

Parametric vs. Non-Parametric Tests

Parametric Tests

Assumptions

- Normality
- Homogeneity of variance
- Independence

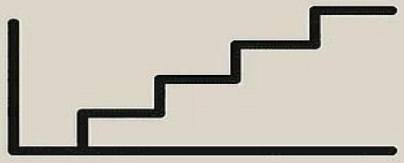
Examples

- T-tests
- ANOVA
- Pearson correlation

Non-Parametric Tests

When to Use

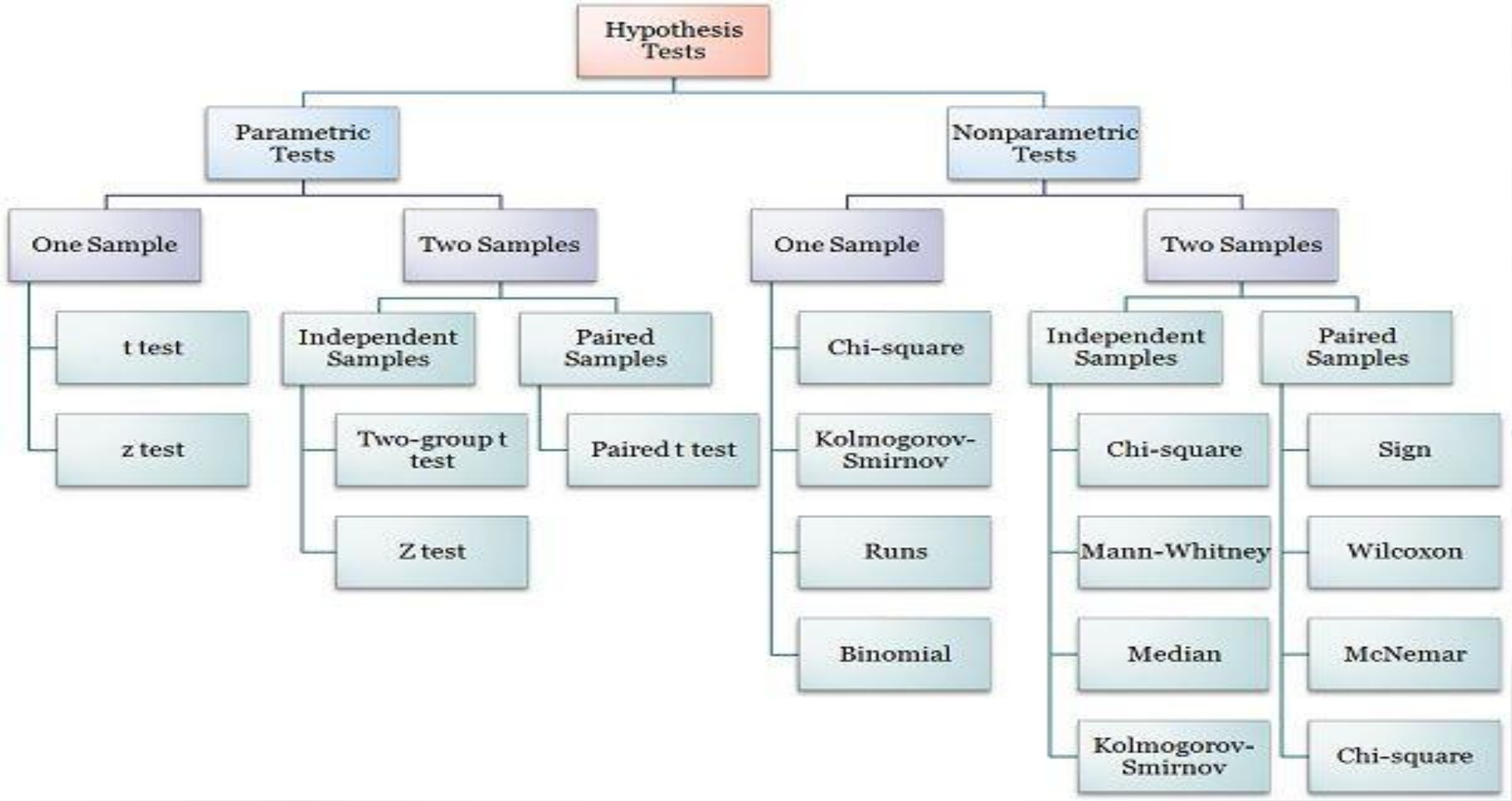
- Data is not normally distributed
- Small sample size
- Ordinal/nominal data
- Presence of outliers



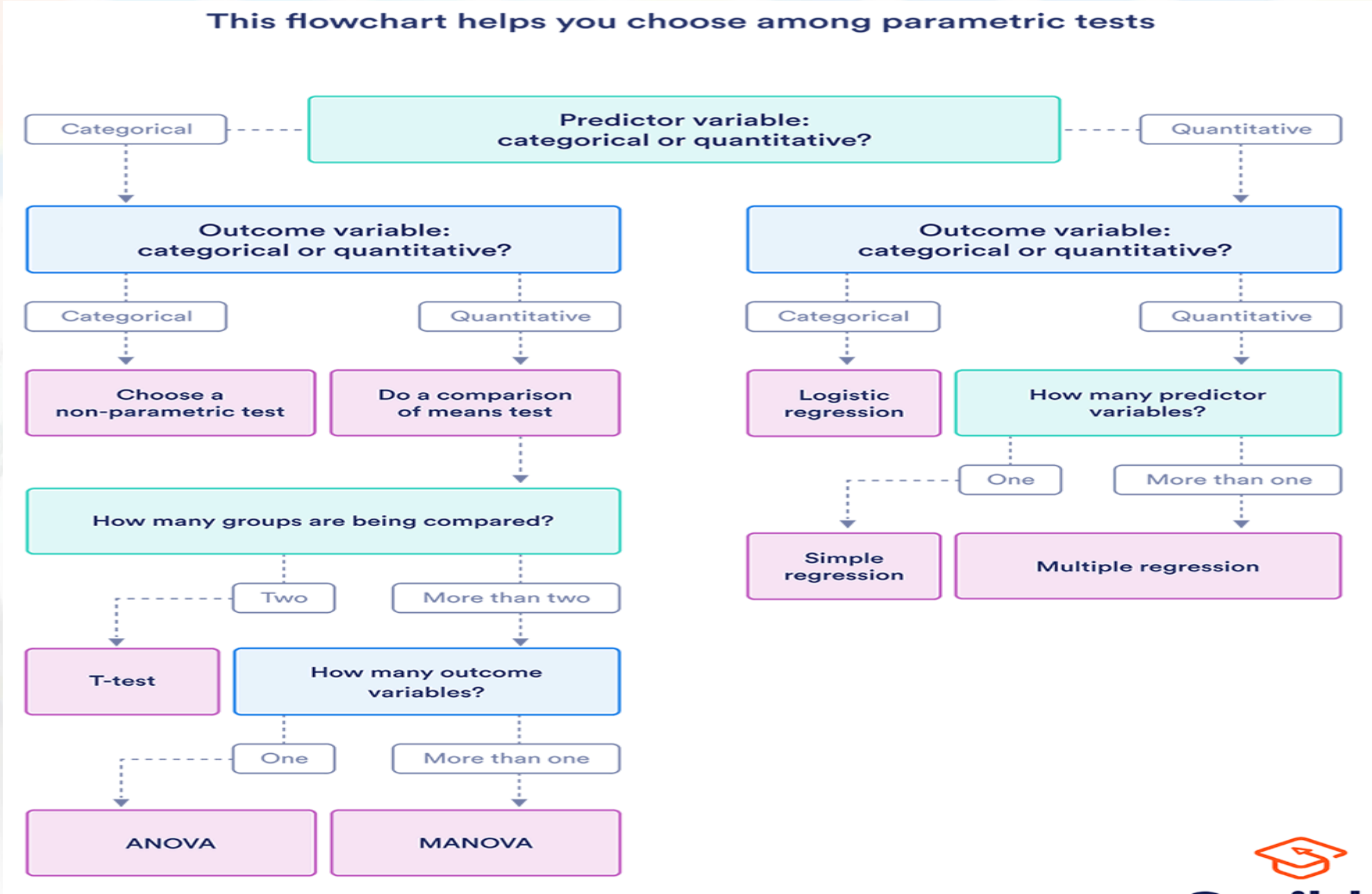
Examples

- Mann-Whitney U test
- Wilcoxon signed-rank test
- Kruskal-Wallis test
- Spearman’s rank correlation

Hypothesis Test Hierarch



Choosing a statistical test



Data set	Parametric test	Non-parametric test
1 variable, 2 categories Between subjects	Independent t test	Mann-Whitney U test
1 variable, 2 categories within subjects	Pair t- test	Wilcoxon test
1 variable, > categories Between subjects	One-way ANOVA	Kruskal Wallis test
1 variable, >2 categories within subjects	Repeated measures ANOVA	Friedman test
2 variables (continuous)	Pearson,s correlation	Spearson,s correlation

t-test

- t-test is used when the population standard deviation is unknown or when the sample size is small (typically n < 30).
- It is used to test hypotheses about the mean of a population based on sample data.
- t-test assumes that the population follows a normal distribution
- Example: Comparing the mean test scores of two groups of patient (e.g., treatment group vs. control group)

t Table

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
one-tail	0.50	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
two-tails	1.00	0.50	0.40	0.30	0.20	0.10	0.05	0.02	0.01	0.002	0.001
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.000	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.000	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.000	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.000	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.000	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.000	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.000	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.000	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	0.000	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232	3.460
80	0.000	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195	3.416
100	0.000	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174	3.390
1000	0.000	0.675	0.842	1.037	1.282	1.646	1.962	2.330	2.581	3.098	3.300
z	0.000	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.291
	0%	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
	Confidence Level										

Z-test

STANDARD NORMAL DISTRIBUTION

Z Score Table

Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-1.4	.08076	.07927	.07780	.07636	.07493	.07353	.07215	.07078	.06944	.06811
-1.3	.09680	.09510	.09342	.09176	.09012	.08851	.08691	.08534	.08379	.08226
-1.2	.11507	.11314	.11123	.10935	.10749	.10565	.10383	.10204	.10027	.09853
-1.1	.13567	.13350	.13136	.12924	.12714	.12507	.12302	.12100	.11900	.11702
-1.0	.15866	.15625	.15386	.15151	.14917	.14686	.14457	.14231	.14007	.13786
-0.9	.18406	.18141	.17879	.17619	.17361	.17106	.16853	.16602	.16354	.16109
-0.8	.21186	.20897	.20611	.20327	.20045	.19766	.19489	.19215	.18943	.18673
-0.7	.24196	.23885	.23576	.23270	.22965	.22663	.22363	.22065	.21770	.21476
-0.6	.27425	.27093	.26763	.26435	.26109	.25785	.25463	.25143	.24825	.24510
-0.5	.30854	.30503	.30153	.29806	.29460	.29116	.28774	.28434	.28096	.27760
-0.4	.34458	.34090	.33724	.33360	.32997	.32636	.32276	.31918	.31561	.31207
-0.3	.38209	.37828	.37448	.37070	.36693	.36317	.35942	.35569	.35197	.34827
-0.2	.42074	.41683	.41294	.40905	.40517	.40129	.39743	.39358	.38974	.38591
-0.1	.46017	.45620	.45224	.44828	.44433	.44038	.43644	.43251	.42858	.42465
0.0	.50000	.50399	.50798	.51197	.51595	.51994	.52392	.52790	.53188	.53586
0.1	.53983	.54380	.54776	.55172	.55567	.55962	.56356	.56749	.57142	.57535
0.2	.57926	.58317	.58706	.59095	.59483	.59871	.60257	.60642	.61026	.61409
0.3	.61791	.62172	.62552	.62930	.63307	.63683	.64058	.64431	.64803	.65173
0.4	.65542	.65910	.66276	.66640	.67003	.67364	.67724	.68082	.68439	.68793
0.5	.69146	.69497	.69847	.70194	.70540	.70884	.71226	.71566	.71904	.72240
0.6	.72575	.72907	.73237	.73565	.73891	.74215	.74537	.74857	.75175	.75490
0.7	.75804	.76115	.76424	.76730	.77035	.77337	.77637	.77935	.78230	.78524
0.8	.78814	.79103	.79389	.79673	.79955	.80234	.80511	.80785	.81057	.81327
0.9	.81594	.81859	.82121	.82381	.82639	.82894	.83147	.83398	.83646	.83891
1.0	.84134	.84375	.84614	.84849	.85083	.85314	.85543	.85769	.85993	.86214
1.1	.86433	.86650	.86864	.87076	.87286	.87493	.87698	.87900	.88100	.88298
1.2	.88493	.88686	.88877	.89065	.89251	.89435	.89617	.89796	.89973	.90147
1.3	.90320	.90490	.90658	.90824	.90988	.91149	.91309	.91466	.91621	.91774
1.4	.91924	.92073	.92220	.92364	.92507	.92647	.92785	.92922	.93056	.93189

Negative Z Score Table Positive Z Score Table

Chi-square test

Critical values of the Chi-square distribution with d degrees of freedom

	Probability of exceeding the critical value						
d	0.05	0.01	0.001	d	0.05	0.01	0.001
1	3.841	6.635	10.828	11	19.675	24.725	31.264
2	5.991	9.210	13.816	12	21.026	26.217	32.910
3	7.815	11.345	16.266	13	22.362	27.688	34.528
4	9.488	13.277	18.467	14	23.685	29.141	36.123
5	11.070	15.086	20.515	15	24.996	30.578	37.697
6	12.592	16.812	22.458	16	26.296	32.000	39.252
7	14.067	18.475	24.322	17	27.587	33.409	40.790
8	15.507	20.090	26.125	18	28.869	34.805	42.312
9	16.919	21.666	27.877	19	30.144	36.191	43.820
10	18.307	23.209	29.588	20	31.410	37.566	45.315

INTRODUCTION TO POPULATION GENETICS, Table D.1

Hypothesis testing (t-test or chi-square test)

- **Ho**= There is no differences in knowledge regarding Biomedical waste management between undergraduate nursing students and intern nurses
- **H_1**=There is differences in knowledge regarding Biomedical waste management between undergraduate nursing students and intern nurses

- Table 2. Mean Score and Level of knowledge regarding biomedical waste management among nursing students and interns nurse (N=315)

Knowledge of Biomedical Waste Management	Interns (n = 86)		Nursing Students (n = 229)		Total (N = 315)		Test of sig(*P*)
	No.	%	No.	%	No.	%	
Knowledge Level							
Inadequate knowledge(<10)	27	31.4	154	67.2	181	57.5	χ^2=32.880
Adequate knowledge (11-19)	59	68.6	75	32.8	134	42.5	(<0.001*)
	Mean ± SD		**Mean ± SD**		**Mean ± SD**		
Overall Knowledge	**64.20 ± 18.77**		**44.01 ± 22.92**		**49.52 ± 23.62**		t= 7.983(<0.001*)
General Information	53.65 ± 26.47		33.19 ± 33.19		38.78 ± 30.87		t= 5.851(<0.001*)
Waste Management	69.27 ± 21.97		45.85 ± 28.32		52.24 ± 28.67		t= 7.757(<0.001*)
Color Coding	62.21 ± 42.66		32.10 ± 40.91		40.32 ± 43.46		t= 5.752(<0.001*)
Risks associated	78.29 ± 30.15		72.93 ± 29.03		74.39 ± 29.39		t= 1.447(<0.149)

χ^2: Chi square test t: Student t-test

p: p value for comparing between the studied groups *: Statistically significant at p ≤ 0.05

Interpretation of Hypothesis testing

- Significant statistical differences between the two studied groups concerning their overall knowledge of BMW management ($t= 7.983$, $P<0.001$).

- Null hypothesis is rejected so alternative hypothesis is accepted. Intern nurses had higher mean (64.20 ± 18.77) compared to undergraduate nursing students (44.01 ± 22.92).

- A higher proportion of the interns had adequate knowledge (68.6%) than that of nursing students (32.8%) ($\chi2=32.880$, $p<0.001$).

ANOVA

- ANOVA stands for Analysis of Variance
- Testing groups to see if there's a difference between them.
- One-way ANOVA (or One-factor ANOVA) is used to determine whether there are any statistically significant differences between the means of two or more independent (unrelated) groups.
- Dependent variable is continuous (i.e., interval or ratio level)
- Independent variable is categorical (i.e., two or more groups)
- Example- compares the three courses (beginner, intermediate, advanced)

ANOVA

Time

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	91.467	2	45.733	4.467	.021
Within Groups	276.400	27	10.237		
Total	367.867	29			

Correlation

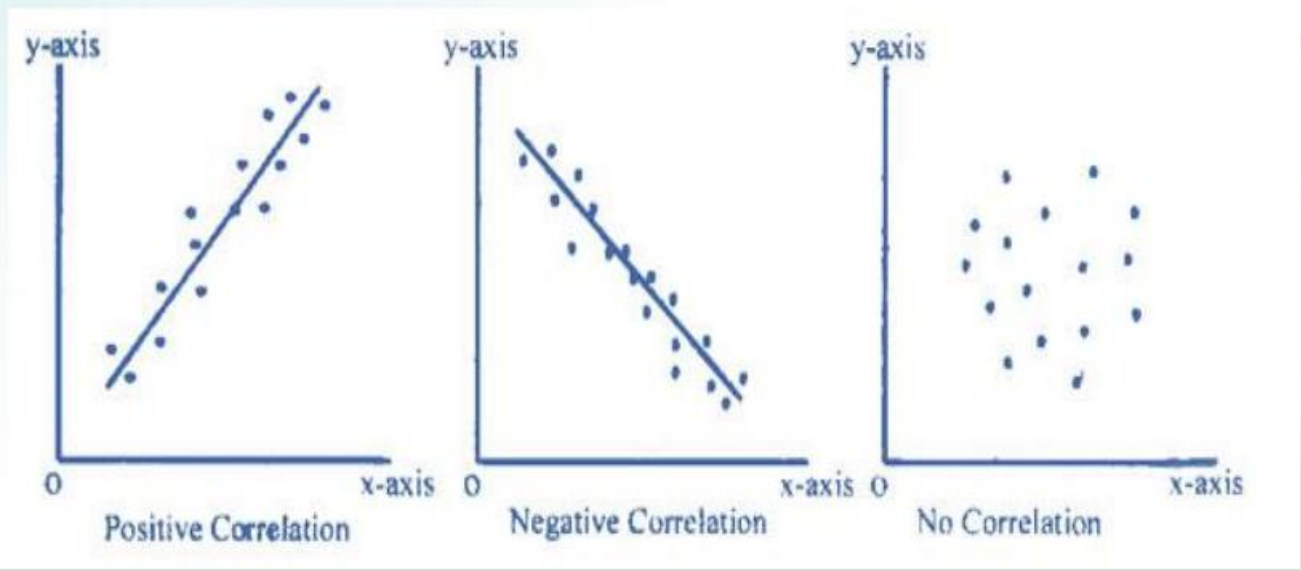
Both independent and dependent variables are continuous/numerical

Used to measure the strength of association between two variables

- Positive correlation - As one variable increases so does the other
- Negative correlation - As one variable increases, the other decreases
- No correlation - No apparent relationship between the variables

Example:

- Height and weight
- Exam score and study hour

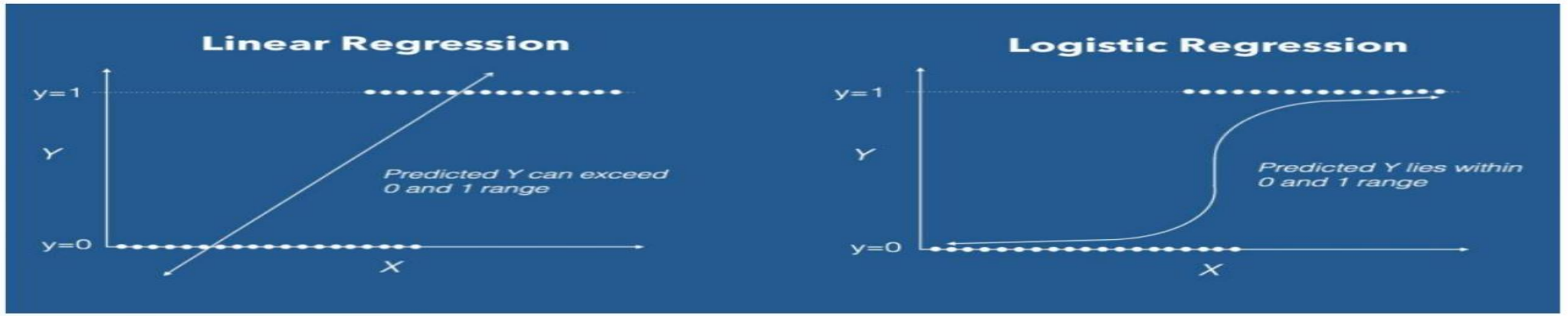


Regression test

- Use to infer or describe the relationship between a dependent variable and one or more independent variables
- Predict the value of dependent variable based on the value of independent variable(s)

Example:

- Amount of eat and gain weigh
- Income and Expenditure
- Smoking status and Having lung cancer



Take Home Messages

- Before data analysis must know about data and variables
- Data is categorical or numerical
- Identify independent variables and dependent variables
- Identify the group or sample
- Select the process of data analysis
- For inferential analysis- Formulating hypothesis- null & alternate hypothesis
- Selecting a significance level
- Choosing an appropriate statistical test
- Done the test and take decision


THANK YOU!