INTRODUCTION TO DATA ANALYSIS

2nd Sem, MCA

CONTENT

- Introduction to Data Analysis
  - Hypothesis Testing
    - ✓ Bivariate Analysis: Correlation Test
      - Correlation coefficient
      - Chi square test
      - ANOVA
      - Summary tables, contingency tables, visualization
    - ✓ Multivariate Analysis
  - Grouping
    - Association rule mining
    - Market Basket Analysis
    - Recommendation system
    - Apriori algorithm
    - FP Growth Algorithm

DATA ANALYSIS – HYPOTHESIS TESTING

- Main purpose of statistics is to test a hypothesis.
- **Hypothesis**: *Educated guess* about something (should be testable).
  - Proposed explanation made on basis of ***limited*** evidence as a **starting point** for further investigation.
- Hypothesis testing in statistics is a way to test results to see if results are meaningful.
- *Null hypothesis* are generally accepted as being true (initially).
- *Alternative hypothesis* is effectively the opposite *(not always)* of a null hypothesis.
  - *H0: There is no relationship between X and Y variable.*
  - *H1: There is a relationship between X and Y variable.*
- Steps in hypothesis Testing:
  - *State null hypothesis,*
  - *Choose what kind of test to perform,*
  - *Either support or reject null hypothesis.*

DATA ANALYSIS – HYPOTHESIS TESTING

- H_0: null hypothesis; No variation between two variables(population); two variables have same distribution.
- H_a: two populations (variables) are not equal.
- ***p-value***: if ***p-value*** is less than a specified significance level **α** (*alpha value;* usually 0.05); difference is significant and null hypothesis H_0 is rejected.
  - *P-value (probability value) tells how likely a particular set of observations occurs if null hypothesis were true.*
  - *Smaller the p-value, more likely to reject null hypothesis.*
  - *P-value will never reach zero, because there's always a possibility.*
- H_0 is rejected: two variables are not from same distribution.

Example: significance level 0.05; degrees for freedom = 2; test result = 0.7533

- *95 times out of 100, survey that agrees with a sample will have a distribution value of 5.99 or less.*
- *0.7533 is less than 5.99 → accept null hypothesis with 0.05 significance level*

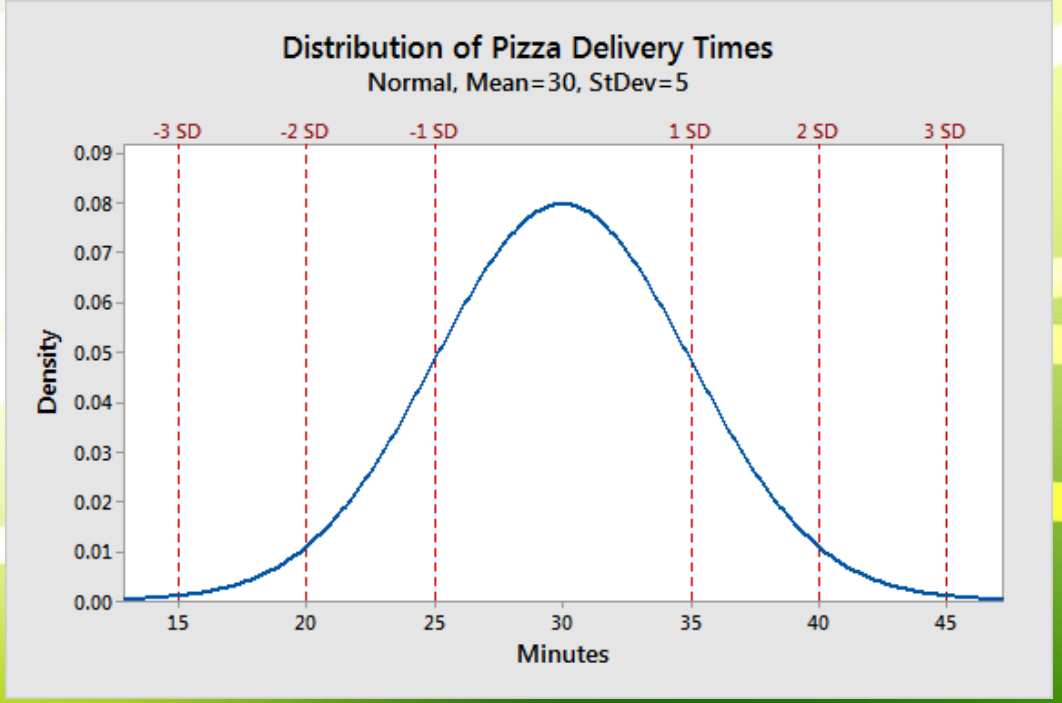
	Probability level (alpha)					
df	0.5	0.10	0.05	0.02	0.01	0.001
1	0.455	2.706	3.841	5.412	6.635	10.827
2	1.386	4.605	5.991	7.824	9.210	13.815
3	2.366	6.251	7.815	9.837	11.345	16.268
4	3.357	7.779	9.488	11.668	13.277	18.465
5	4.351	9.236	11.070	13.388	15.086	20.517

DATA ANALYSIS – HYPOTHESIS TESTING

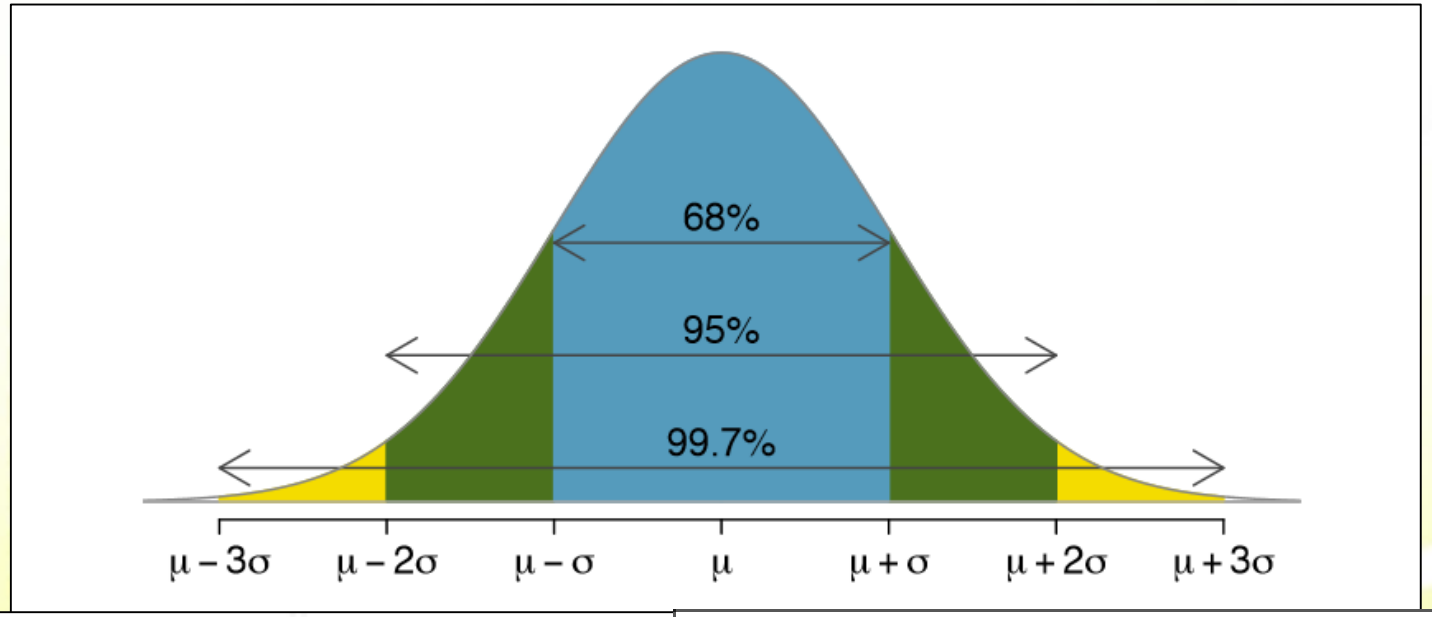
- Normal/Gaussian/bell-shaped distribution: continuous probability distribution i.e. symmetrical around its mean.
- Most observations cluster around central peak
- Probabilities for values further away from mean taper off (equally) in both directions.
- Extreme values in both tails of distribution are similarly unlikely.
- *Box–Cox* transformation

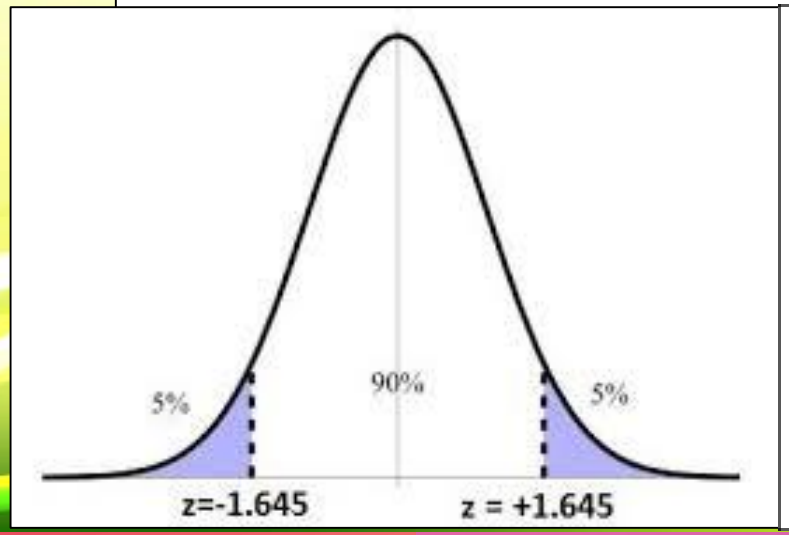
$$x_i' = \frac{x_i^\lambda - 1}{\lambda}$$

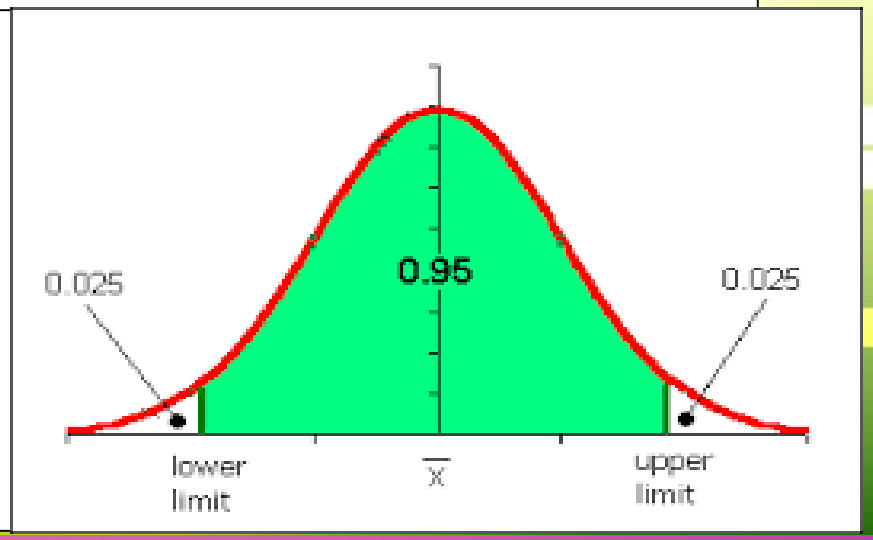
Mean +/- standard deviations	Percentage of data contained
1	68%
2	95%
3	99.7%



DATA ANALYSIS – HYPOTHESIS TESTING







DATA ANALYSIS – HYPOTHESIS TESTING

- In statistics, **confidence interval** refers to probability that a population parameter will fall between a set of values for a certain proportion (percentage) of times.
- Confidence intervals measure the degree of uncertainty or certainty.
- Most common are 95% or 99% *confidence/significance* level.

Confidence level = 100 × (1 − α)

	Probability level (alpha)					
df	0.5	0.10	0.05	0.02	0.01	0.001
1	0.455	2.706	3.841	5.412	6.635	10.827
2	1.386	4.605	5.991	7.824	9.210	13.815
3	2.366	6.251	7.815	9.837	11.345	16.268
4	3.357	7.779	9.488	11.668	13.277	18.465
5	4.351	9.236	11.070	13.388	15.086	20.517

- For 90% confidence level alpha is 0.1; for 95% confidence level alpha is 0.05; for 99% confidence level α is 0.01.
- Confidence level means that; if experiment is repeated over and over again, 95% times results will match.

- **Example**, a survey conducted on group of pet owners to see how many cans of dog food they purchase a year. Testing the statistic at 99% confidence level gives a confidence interval of (200,300) → they buy between 200 and 300 cans a year (with a very high probability 99%)

DATA ANALYSIS – HYPOTHESIS TESTING

- *Confidence Interval (CI)* is a *range of values* we are fairly sure our *true value* lies in.

- CI can be constructed with
  - *t-distribution*

    $\mu \pm t * \sigma / (\sqrt{n})$

    $$\overline{X} \pm t \frac{s}{\sqrt{n}}$$

  - *Normal or z-distribution*

    $\mu \pm z * \sigma / (\sqrt{n})$

- *standard error of the sampling distribution* = $\sigma / (\sqrt{n})$

- Since size 'n' is in denominator and standard deviation 's' is in numerator
  → small samples with large variations increase standard error,
  this reduces confidence that sample statistic is a close approximation of the population parameter.

DATA ANALYSIS – HYPOTHESIS TESTING

T-Distribution Table

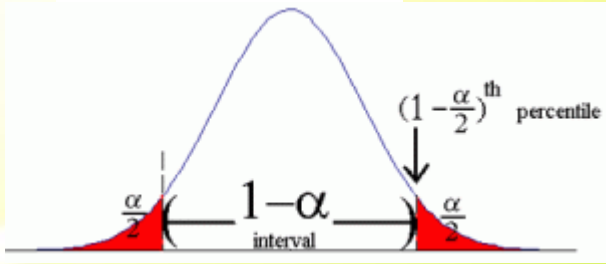
df	a = 0.1	0.05	0.025	0.01	0.005
∞	t_a = 1.282	1.645	1.960	2.326	2.576
1	3.078	6.314	12.706	31.821	63.656
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169

Confidence Interval	Z-score
80%	1.282
85%	1.440
90%	1.645
95%	1.960
99%	2.576
99.5%	2.807
99.9%	3.291

DATA ANALYSIS – HYPOTHESIS TESTING

***Example**: Construct a 98% Confidence Interval based on the following data: 45, 55, 67, 45, 68, 79, 98, 87, 84, 82.*

$$\bar{X} \pm t\frac{s}{\sqrt{n}}$$



- ***Step 1**: Find mean, μ and standard deviation, σ for the data.*

  σ: 18.172; μ: 71

- ***Step 2:** Subtract 1 from sample size to find degrees of freedom (df).*

  df = 10 – 1 = 9

- ***Step 3:** Find alpha level; Subtract confidence level from 1, then divide by two. (1 – .98) / 2 = .01*
- ***Step 4**: Look up df and a in t-distribution table. For df = 9 and a = .01, table gives 2.821*
- ***Step 5:** Apply CI formula for t-distribution μ ± t * σ / (√n)*

  Lower end of CI range, 71 – 16.22075 = 54.77925

  Upper end of CI range 71 + 16.22075 = 87.22075

  98% CI is (54.78, 87.22)

DATA ANALYSIS – HYPOTHESIS TESTING

Example: *Construct a 95 % confidence interval an experiment that found the sample mean temperature for a certain city in August was 101.82, with a population standard deviation of 1.2. There were 6 samples in this experiment*

- ***Step 1****: Subtract confidence level (Given as 95 percent in question) from 1 and then divide the result by two.*

  alpha level (represents area in one tail) = (1 – .95) / 2 = .025

- ***Step 2****: Find z-score from z-table : z score = 1.96.*
- ***Step 3****: Plug the numbers into the second part of the formula and solve:* $z * \sigma / (\sqrt{n})$

  *= 1.96 * 1.2/√(6) = 1.96 * 0.49 = 0.96*

- ***Step 4****: Find the CI:*

  Lower end of CI range, subtract step 3 from mean = 101.82 – 0.96 = 100.86

  Upper end of CI range, add step 3 to mean = 101.82 + 0.96 = 102.78.

  CI is (100.86,102.78)

DATA ANALYSIS - CORRELATION

- **Bivariate Analysis:** Analysis of any concurrent relation between two variables or attributes.
  - Consists of a group of statistical techniques that examine relationship between two variables.
  - Bivariate analysis forms foundation of multivariate analysis.
- **Correlation:** Relation between two variables.
- **Bivariate correlation Test**: Statistical technique to determine existence of relationships/association between two different variables (X, Y)
  - *whether/how much X will change when there is a change in Y.*

Types of tests:

- *Correlation*: check the association between variables.
- *Comparison of means*: check the differences between means of variables.
- *Regression*: check if one variable predicts changes in another variable.
- *Non-Parametric*: tests that are used when data does not meet the assumptions of parametric tests.

DATA ANALYSIS - CORRELATION

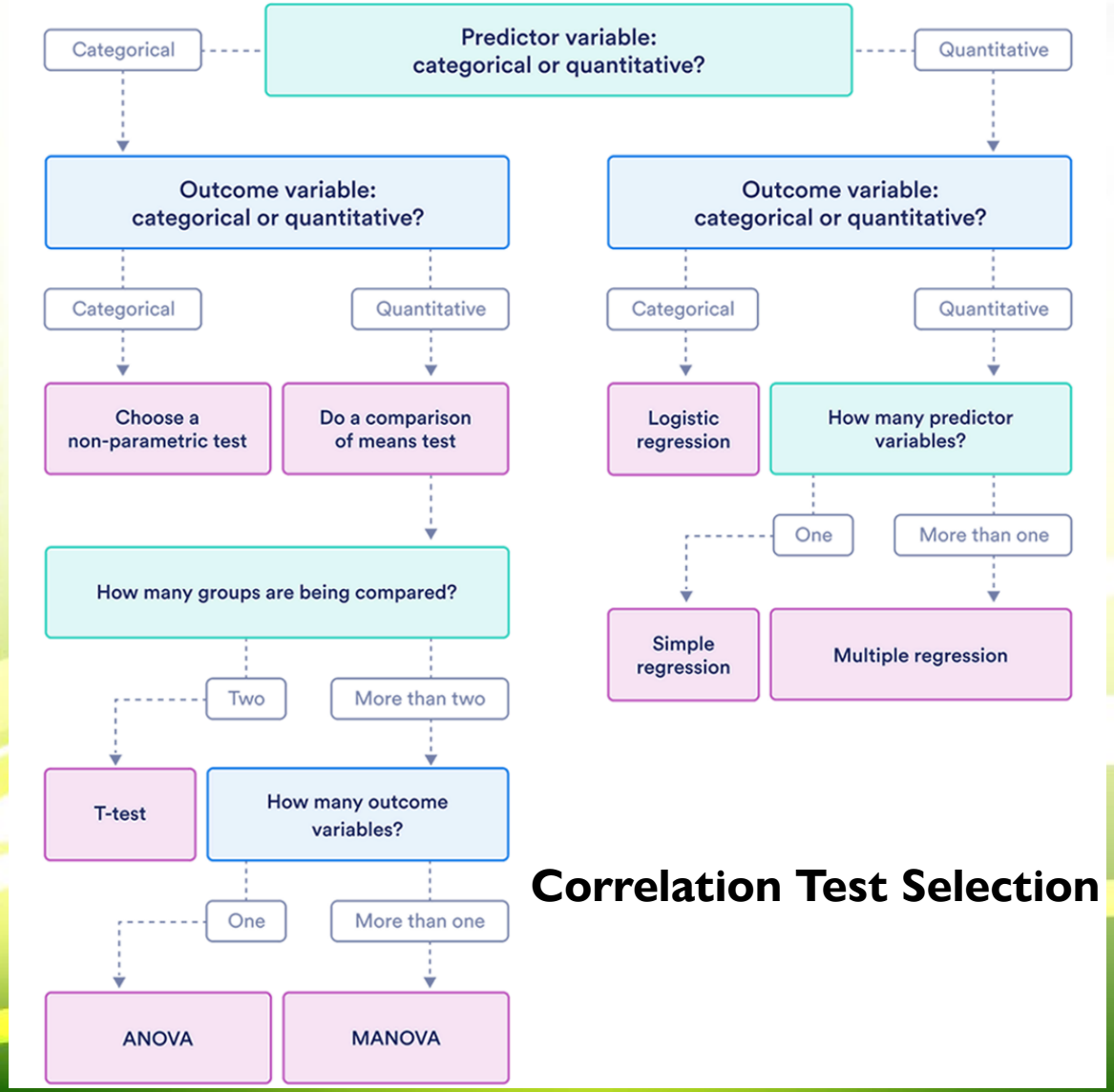
Parametric Tests

- Prior knowledge of population distribution (normal) is available.
- Fixed set of parameters used to determine probabilistic model.
- Parameters used in normal distribution: Mean, Standard Deviation
- T-test; Z-test; F-test; ANOVA (post-hoc test)

Non-parametric Tests

- No fixed set of parameters available, and also there is no distribution (normal) knowledge available for use.
- No assumption made about parameters for given population.
- Referred to as **distribution-free tests**.
- More popular; Easy to apply and understand; less complex.
- Chi-square test; Mann-Whitney U-test; Kruskal-Wallis H-test

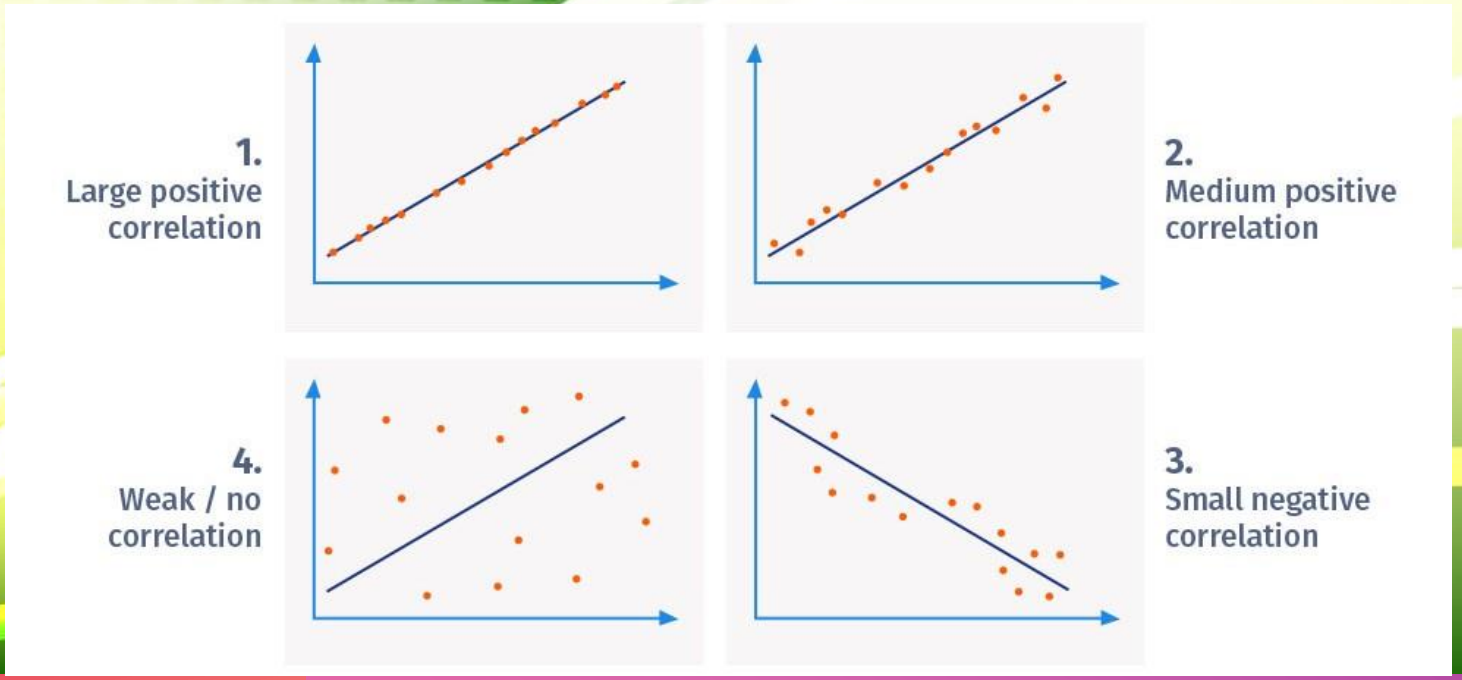
DATA ANALYSIS - CORRELATION



Correlation Test Selection

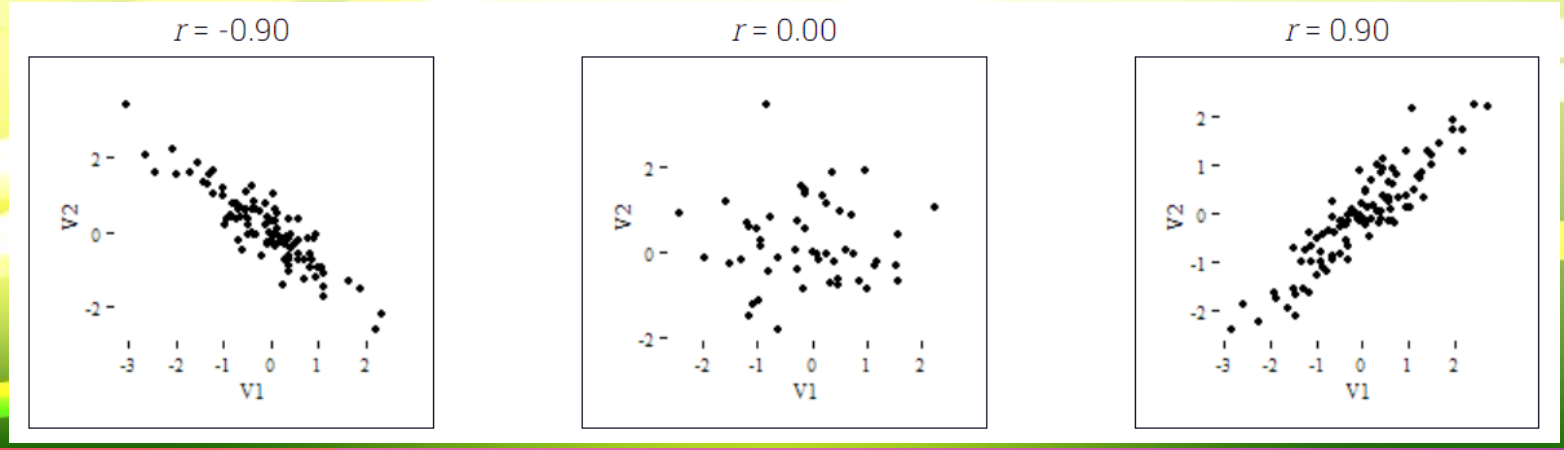
DATA ANALYSIS - CORRELATION

- **Positive correlation**: both variables move in same direction → increase in one variable leads to increase in other variable and vice versa.
  - spending more time on a treadmill burns more calories.
- **Negative correlation**: two variables move in opposite directions → increase in one variable leads to decrease in other variable and vice versa.
  - increasing speed of a vehicle decreases time to reach destination.
- **Weak/Zero correlation**: one variable does not affect other.
  - no correlation between number of years of school a person has attended and letters in his/her name.



DATA ANALYSIS - CORRELATION

- **Correlation coefficient (r)** measures strength of association/co-occurrence (between -1 to +1).
- **Pearson ('r' or product-moment) correlation coefficient**: Between two continuous-level variables.
  - Positive correlation shows direct relationship between two variables (the larger A, the larger B).
  - Negative correlation shows inverse relationship (the larger A, the smaller B).
  - Zero correlation coefficient indicates no relationship between the variables at all.
  - .1 < | r | < .3 … small / weak correlation
  - .3 < | r | < .5 … medium / moderate correlation
  - .5 < | r | …… large / strong correlation



DATA ANALYSIS - CORRELATION

Advantages of correlation analysis

- **Observe relationships**: correlation helps to identify absence/presence of relationship between two variables.
- **Good starting point for research/analysis**.
- **Uses for further studies**: Guides to identify direction and strength of relationship between two variables and later narrow the findings down in later studies.
- **Simple metrics**: findings are simple to classify (range from -1.00 to 1.00). Only three potential broad outcomes of the analysis.

DATA ANALYSIS - CORRELATION

Bessel's correction

- Use of ‘n − 1’ instead of ‘n’ in the formula for sample variance and sample standard deviation.
- Corrects the bias in estimation of population variance and population standard deviation.
- Except for rare cases (sample mean = population mean), data will be closer to sample mean than it will be to the true population mean.
    - So the value on denominator will probably be a bit smaller than what it would be if used the true population mean. To make up for this, divide by ‘n-1’ (a smaller value) rather than ‘n’.

$$\sigma = \sqrt{\frac{\sum(x_i - \mu)^2}{n}}$$

$$s = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n - 1}}$$

DATA ANALYSIS – CORRELATION

Pearson r correlation:

$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}$$

r = Pearson r correlation coefficient between x and y
n = number of observations
x_i = value of x (for ith observation)
y_i = value of y (for ith observation)

S_x, S_y = S.D. for x and y

DATA ANALYSIS - CORRELATION

Pearson r correlation:

r_{xy} = Pearson r correlation coefficient between x and y
n = number of observations
x_i = value of x (for ith observation)
y_i = value of y (for ith observation)

$$r = \frac{L_{xy}}{\sqrt{L_{xx} L_{yy}}}$$

$$r_{xy} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\sqrt{n\sum y_i^2 - (\sum y_i)^2}}$$

$$L_{xx} = \sum_i x_i^2 - \frac{(\sum_i x_i)^2}{n}$$

$$L_{xy} = \sum_i x_i y_i - \frac{(\sum_i x_i)(\sum_i y_i)}{n}$$

$$L_{yy} = \sum_i y_i^2 - \frac{(\sum_i y_i)^2}{n}$$

Population Correlation Coefficient

$$P_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \frac{\sum (xi - \bar{x})(yi - \bar{y})}{\sqrt{\left(\sum (xi - \bar{x})^2\right)\left(\sum (yi - \bar{y})^2\right)}}$$

Where,
$\sigma_x, \sigma_y \rightarrow$ Population Standard Deviation
$\sigma_{xy} \rightarrow$ Population Covariance
$\bar{X}, \bar{Y} \rightarrow$ Population Mean

$$r_{xy} = \frac{\mathrm{cov}(x, y)}{\sqrt{\mathrm{var}(x)} \cdot \sqrt{\mathrm{var}(y)}}$$

$$\mathrm{Cov}\,(X, Y) = \frac{\Sigma(X_i - \bar{X})(Y_j - \bar{Y})}{n}$$

$$\sigma^2 = \frac{\Sigma(xi - \bar{x})^2}{N}$$

DATA ANALYSIS – CORRELATION

CATEGORICAL DATA ENCODING

- ***Categorical data****:* variables contain label values rather than numeric values.
- Number of possible values is often limited to a fixed set.
- Each value represents a different category.
- Categorical variables also called nominal *(ordinal, if ordered).*
  - variable "pet" with values: "dog" and "cat".
  - Variable "color" with values: "red", "green" and "blue".
  - variable "rank" with values: "first", "second" and "third" *(ordinal)*
- Machine learning algorithms (data analytics) cannot operate on label data directly *(all input/output variables must be numeric).*

DATA ANALYSIS – CORRELATION

CATEGORICAL DATA ENCODING

Integer/Label Encoding

- Each unique category value is assigned an integer value.
  - "red" is 1, "green" is 2, and "blue" is 3.
- Easily reversible.
- Such integer values have a natural ordered relationship between each other → machine learning algorithms tend to understand and harness this relationship.
  - For some variables/analysis (ordinal), this may be enough/good.
- For nominal data label encoding is not enough/good.
  - using such encoding and allowing the model to assume natural ordering between categories may result in poor performance or unexpected results.

DATA ANALYSIS – CORRELATION

CATEGORICAL DATA ENCODING

One-Hot Encoding

- New binary variable is added for each unique categorical data value.
- Original variable is discarded.
  - In "color" variable example, there are 3 categories.
  - 3 binary variables are added.
  - "1" value is placed in the binary variable for respective color and "0" values for all other color variables.

Color
Red
Green
Blue

Red	Green	Blue
1	0	0
0	1	0
0	0	1

DATA ANALYSIS - CORRELATION

Summary Table

- Visualization that summarizes statistical information about data in table form.

Column	Sum	Avg	Min	Max	Median	StdDev
Sales	3956	18	8	35	18	7
Cost	3194	15	6	29	13	6

DATA ANALYSIS - CORRELATION

Contingency table:

- **crosstabs or two-way tables**
- Tabular representation of categorical data.
- Used in statistics to summarize relationship between several categorical variables.
- Special type of frequency distribution table, where two variables are shown simultaneously.
- Usually shows frequencies for particular combinations of values of two discrete random variable s X and Y.
- Each cell in the table represents a mutually exclusive combination of X-Y values.

Gender	Result
Male	Pass
Female	Pass
Male	Fail
Male	Fail
Male	Pass
Female	Pass
Female	Fail

	Pass	Fail	
Male	2	2	4
Female	2	1	3
	4	3	

DATA ANALYSIS – CHI SQUARE

- Pearson's chi-square test.
- Primary use of chi-square test is to **examine whether** two variables are independent (not related) or not.
  - If two variables are correlated, their values tend to move together, either in same or opposite direction.
  - One variable is "not correlated with" or "independent of" other if increase in one variable is not associated with increase in another.
- Chi-Square statistic is based on the *difference between what is actually observed data and what would be expected if there was truly no relationship between the variables.*
- Null and alternative Hypothesis:
  - H0: There is no relationship between X and Y variable.
  - H1: There is a relationship between X and Y variable.

DATA ANALYSIS – CHI SQUARE

- Calculation of Chi-Square statistic: $X^2 = \sum(O_i - E_i)^2/E_i$

$$X^2 = \sum \frac{(\text{Observed value} - \text{Expected value})^2}{\text{Expected value}}$$

  - *Oi = observed frequency (observed counts in the cells)*
  - *Ei = expected frequency (if NO relationship existed between the variables)* **E_i = row total*column total/sample size**
- Chi-square statistic can't be negative (WHTHER related or not; doesn't indicate directionality)
- *degrees of freedom* = (r-1)*(c-1). ***(Number of response categories)***
  - r, c: number of rows, columns in considered dataset *(contingency table)*
- Compare statistical value for degree of freedom (d) & critical/alpha value (p) from Chi-square distribution table with calculated Chi-square statistical value to decide whether variables are related or not.
  - Accept/reject hypothesis
  - *Chi-square calculated value > Chi-square critical value* → reject the null hypothesis.

DATA ANALYSIS – CHI SQUARE

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

DATA ANALYSIS - CHI SQUARE

- Is gender independent of education level? A random sample of 395 people were surveyed and each person was asked to report the highest education level they obtained. The data that resulted from the survey is summarized in the following table:

	High School	Bachelors	Masters	Ph.d.	Total
Female	60	54	46	41	201
Male	40	44	53	57	194
Total	100	98	99	98	395

- **Question**: Are gender and education level dependent at 95% level of significance?
- In other words, given the data collected above, is there a relationship between the gender of an individual and the level of education that they have obtained?

DATA ANALYSIS - CHI SQUARE

Actual Data

	High School	Bachelors	Masters	Ph.d.	Total
Female	60	54	46	41	201
Male	40	44	53	57	194
Total	100	98	99	98	395

Expected Data

	High School	Bachelors	Masters	Ph.d.	Total
Female	50.886	49.868	50.377	49.868	201
Male	49.114	48.132	48.623	48.132	194
Total	100	98	99	98	395

$$\chi^2 = \frac{(60 - 50.886)^2}{50.886} + \cdots + \frac{(57 - 48.132)^2}{48.132} = 8.006$$

- H0: There is no relationship between X and Y variable.
- H1: There is a relationship between X and Y variable.

- Critical value of $\chi 2$ with 3 degree of freedom is 7.815.
- 8.006 > 7.815 → reject the null hypothesis.
- Education level depends on gender at a 95% level of significance.