INTRODUCTION TO DATA ANALYSIS

2nd Sem, MCA

CONTENT

- Introduction to Data Analysis
  - Hypothesis Testing
    - ✓ Bivariate Analysis: Correlation Test
      - Correlation coefficient
      - Chi square test
      - T - 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

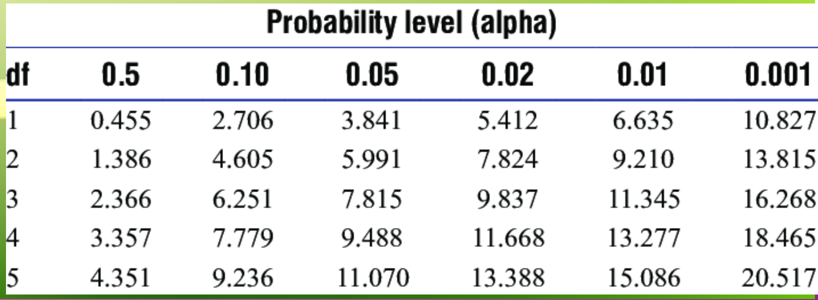
- Main purpose of statistics is to test a hypothesis.
- **Hypothesis**: *Educated guess* about something (should be testable).
  - Proposed explanation made on basis of ***limited*** evidence as a **starting point** for further investigation.
- Hypothesis testing in statistics is a way to test results 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)

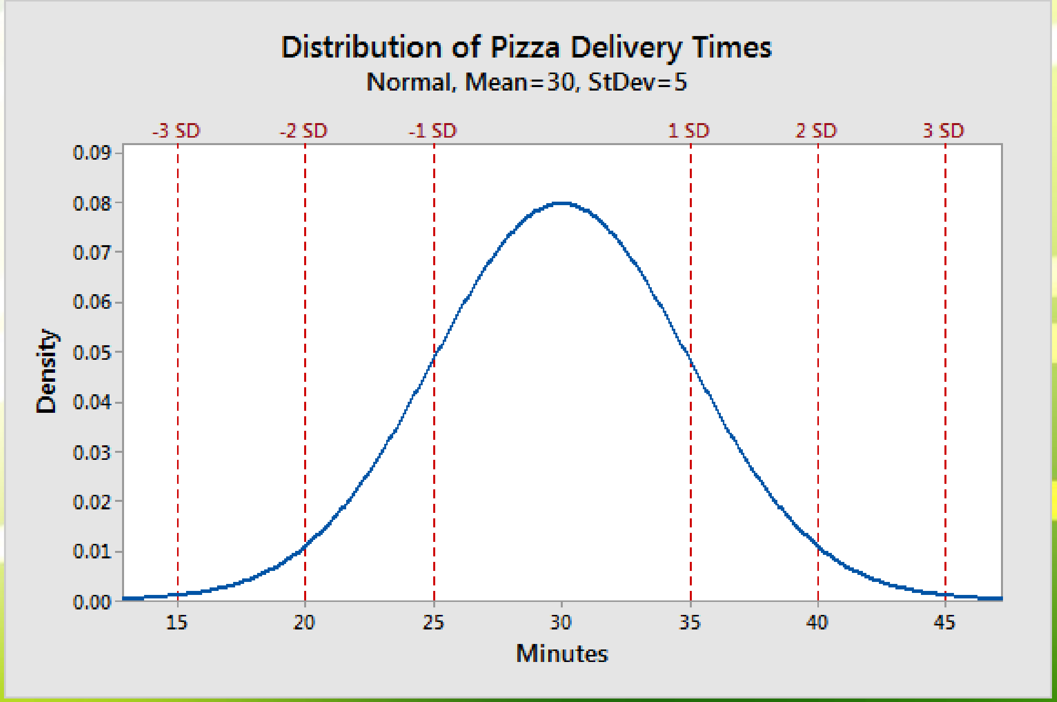
df	0.5	0.10	0.05	0.02	0.01	0.001
1	0.455	2.706	3.841	5.412	6.635	10.827
2	1.386	4.605	5.991	7.824	9.210	13.815
3	2.366	6.251	7.815	9.837	11.345	16.268
4	3.357	7.779	9.488	11.668	13.277	18.465
5	4.351	9.236	11.070	13.388	15.086	20.517

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

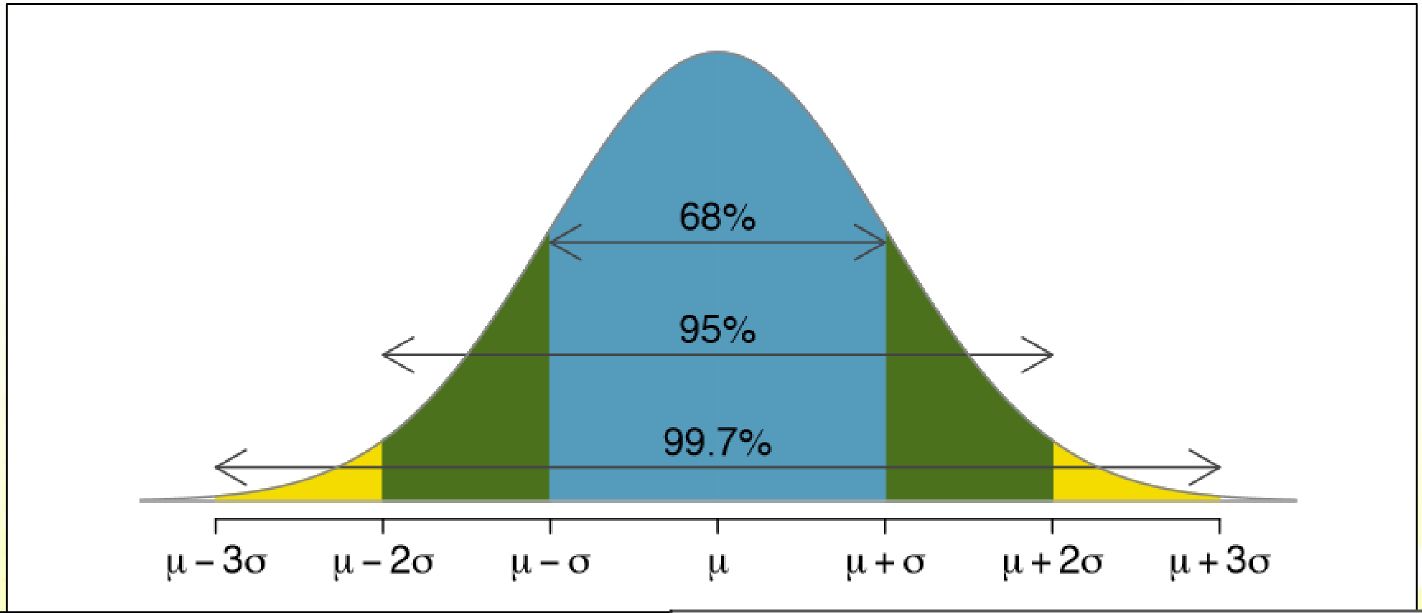
$$x'_i = \frac{x_i^{\lambda} - 1}{\lambda}$$

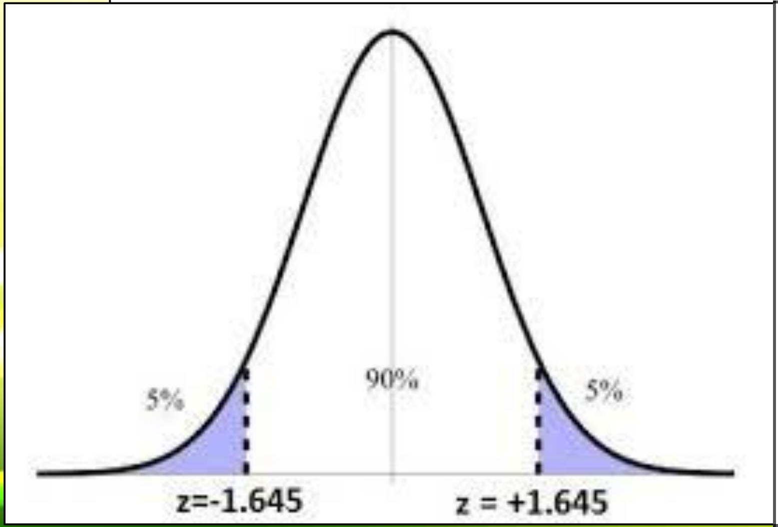
Mean +/- standard deviations	Percentage of data contained
1	68%
2	95%
3	99.7%

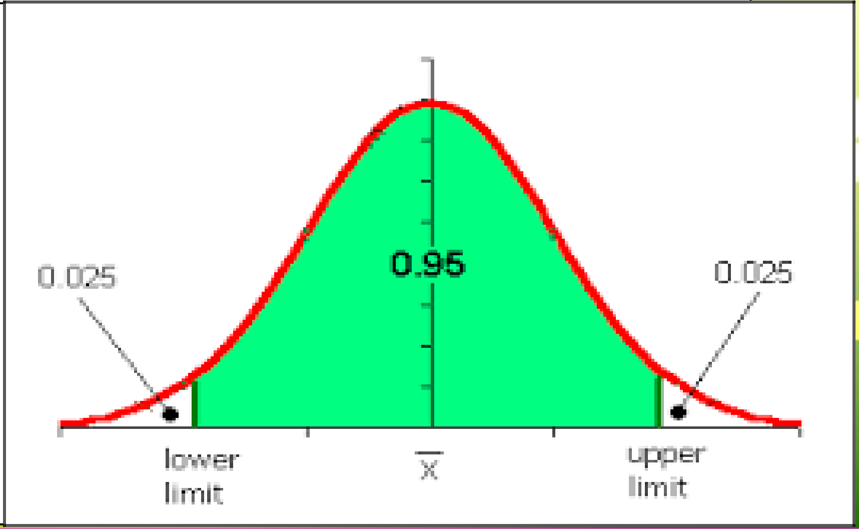

Distribution of Pizza Delivery Times

Normal, Mean=30, StDev=5

DATA ANALYSIS – HYPOTHESIS TESTING



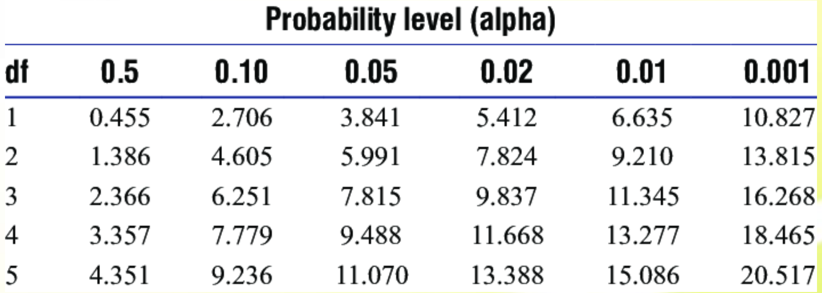




DATA ANALYSIS – HYPOTHESIS TESTING

- 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.

$$\textbf{\textit{Confidence level}} = 100 \times (1 - \alpha)$$



	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*

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

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

  - *Normal or z-distribution*

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

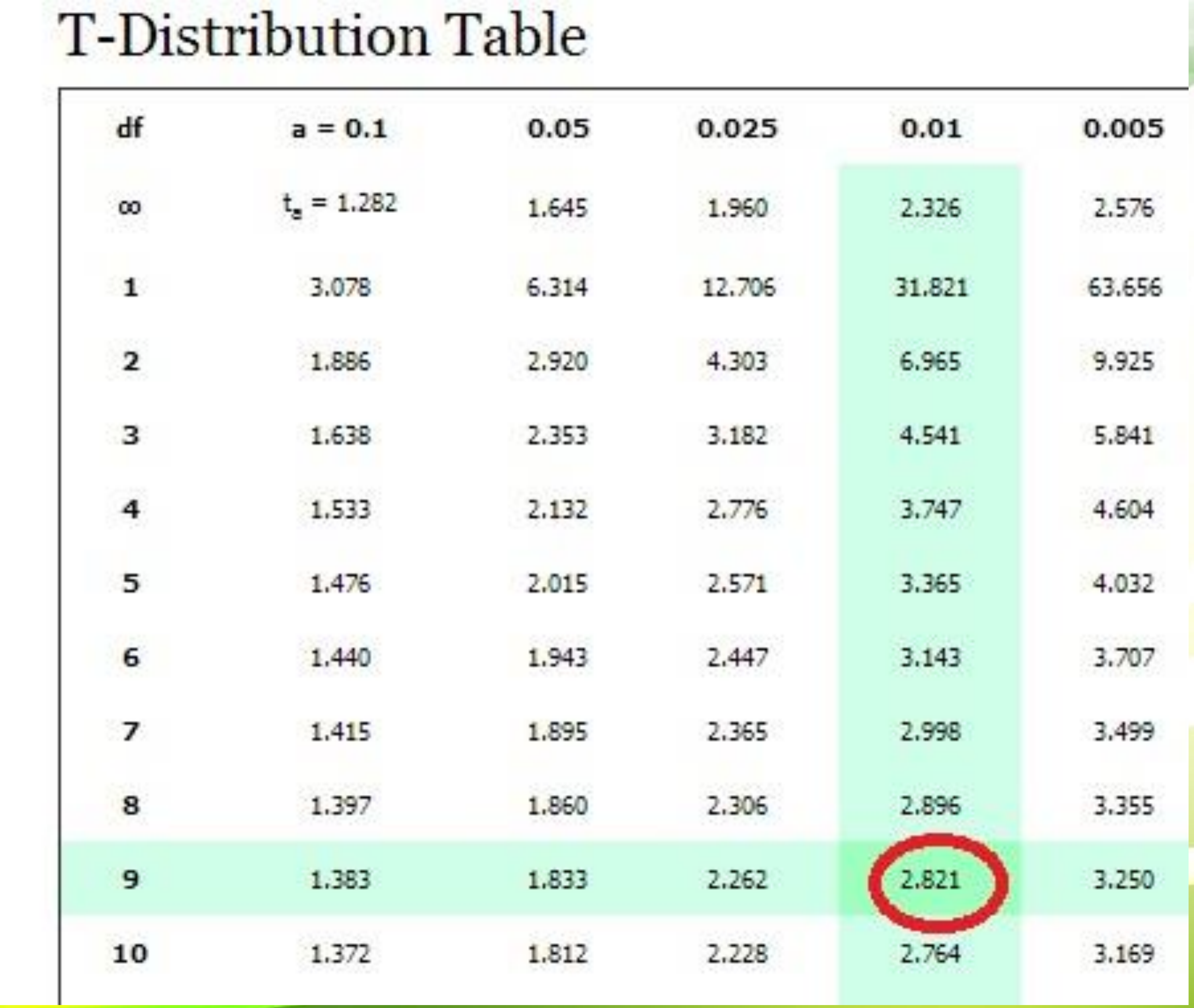
- *standard error of the sampling distribution* = $\boldsymbol{\sigma / (\sqrt{n})}$

- 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

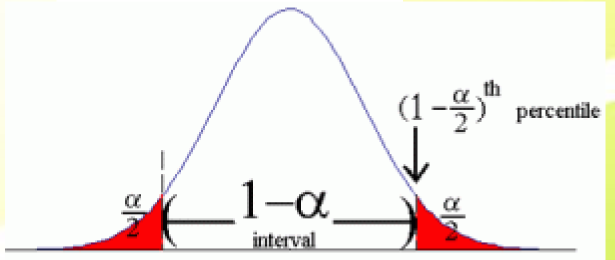
df	a = 0.1	0.05	0.025	0.01	0.005
∞	t_a = 1.282	1.645	1.960	2.326	2.576
1	3.078	6.314	12.706	31.821	63.656
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	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 * σ / (√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

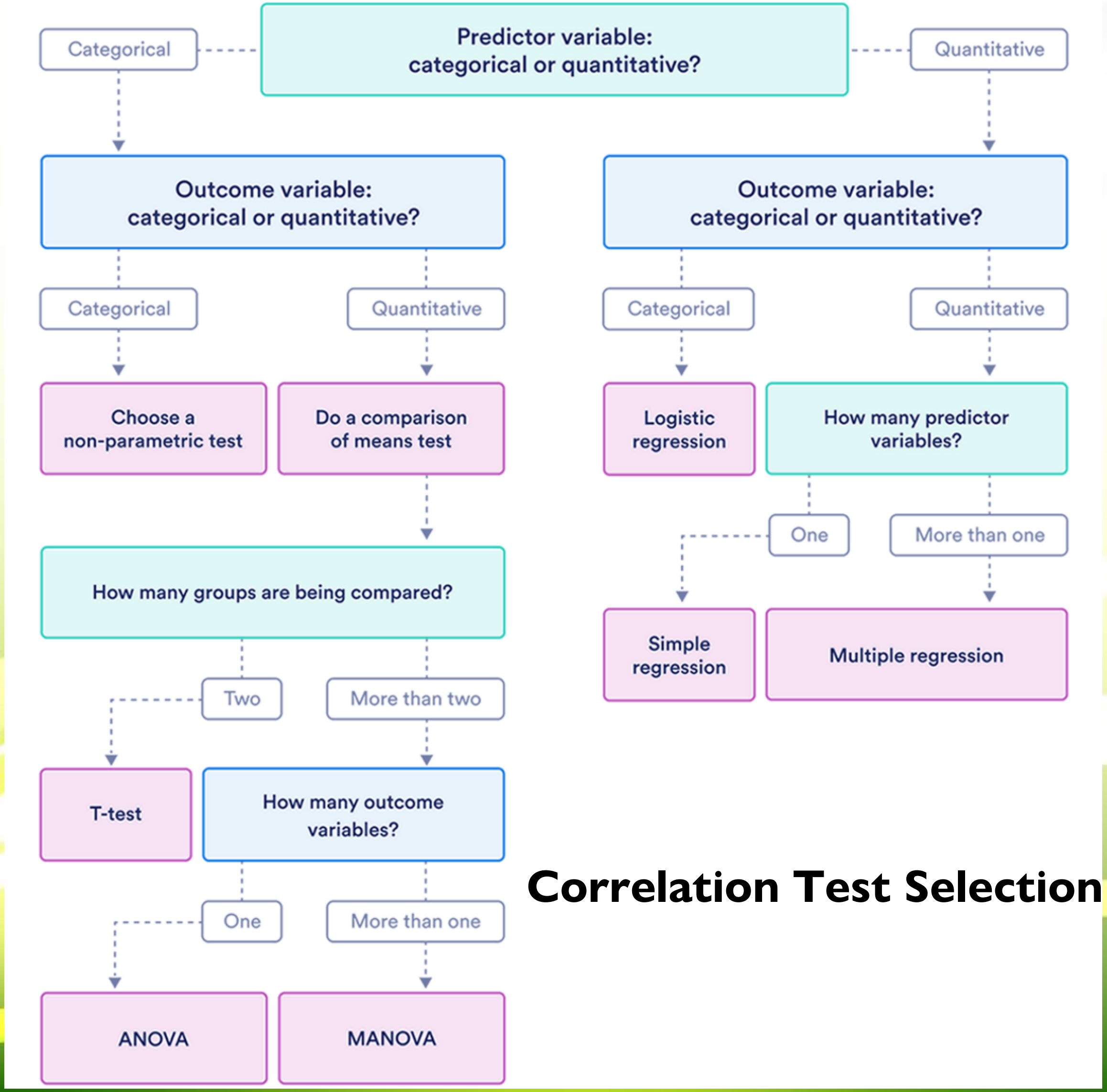
Parametric Tests

- 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

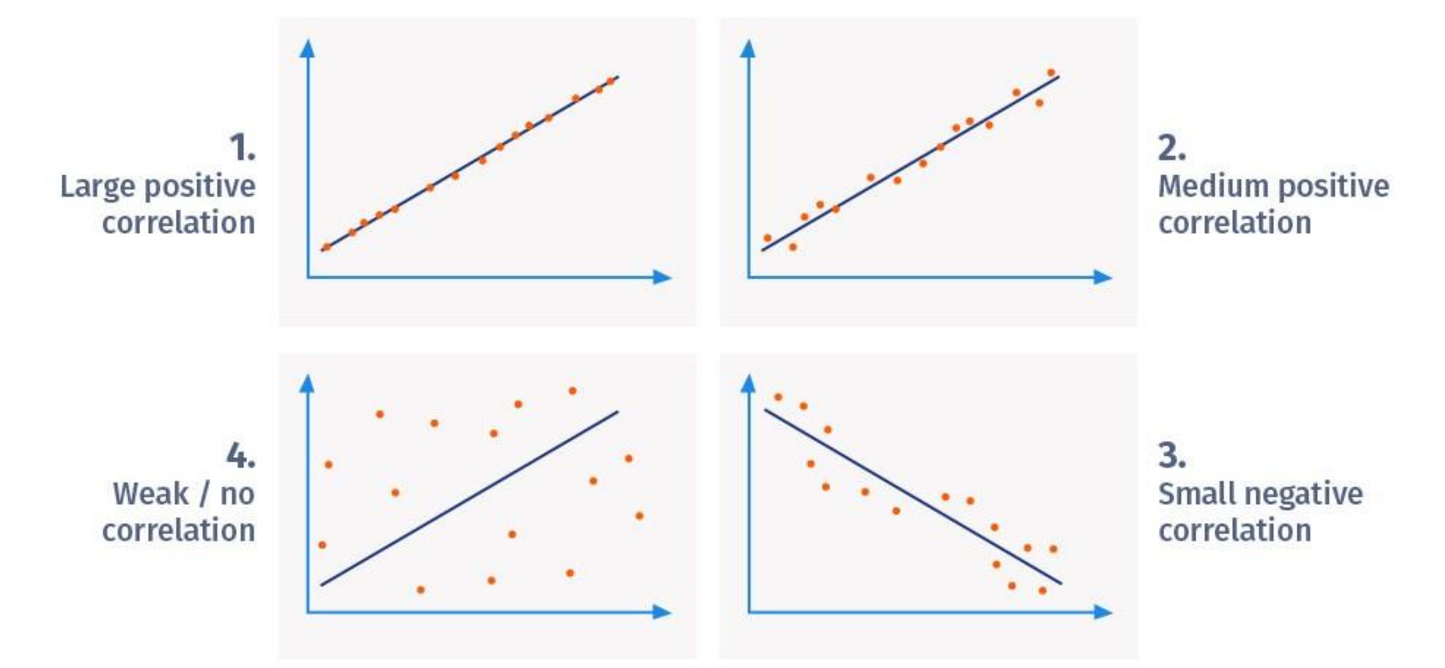
DATA ANALYSIS - CORRELATION


Predictor variable: categorical or quantitative?
Categorical
Quantitative
Outcome variable: categorical or quantitative?
Categorical
Quantitative
Choose a non-parametric test
Do a comparison of means test
How many groups are being compared?
Two
More than two
T-test
How many outcome variables?
One
More than one
ANOVA
MANOVA
Outcome variable: categorical or quantitative?
Categorical
Quantitative
Logistic regression
How many predictor variables?
One
More than one
Simple regression
Multiple regression

Correlation Test Selection

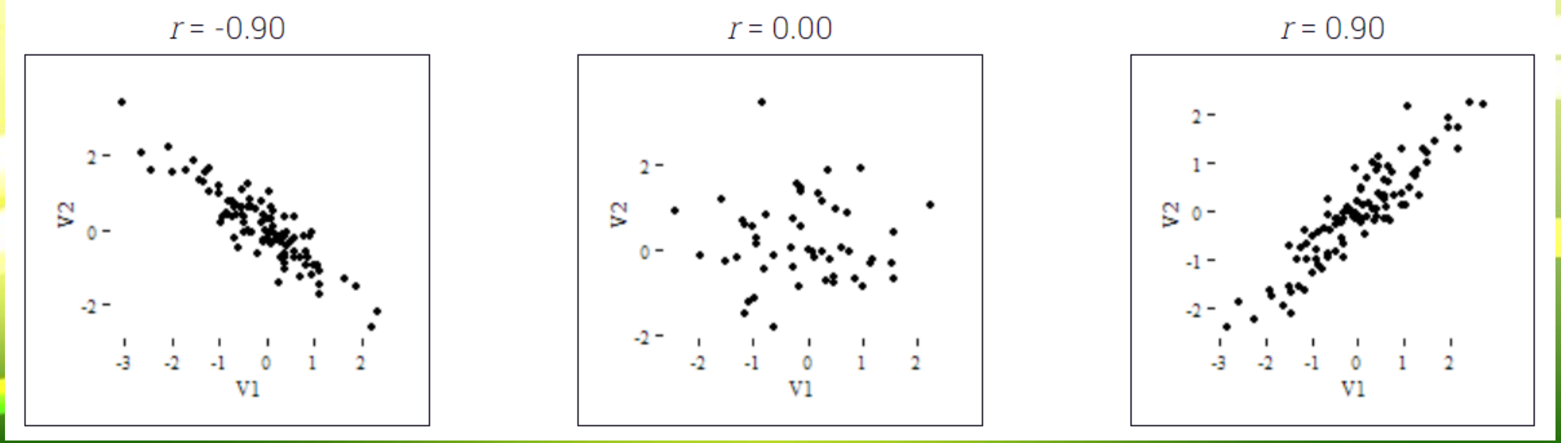
DATA ANALYSIS - CORRELATION

- **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, σ_x, σ_y → Population Standard Deviation

σ_{xy} → Population Covariance

$\bar{X}, \bar{Y}$ → Population Mean

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

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

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

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.

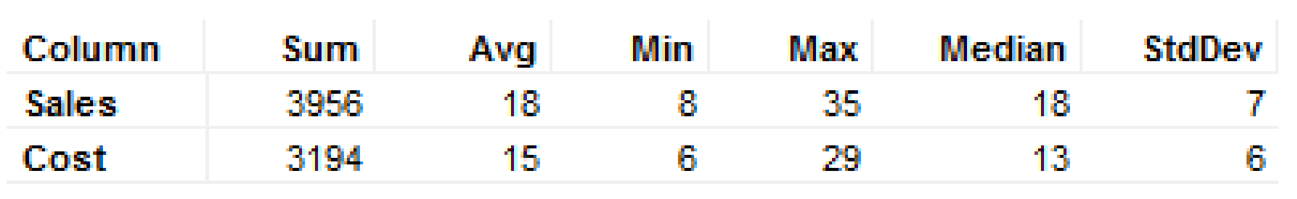
Color
Red
Green
Blue

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.

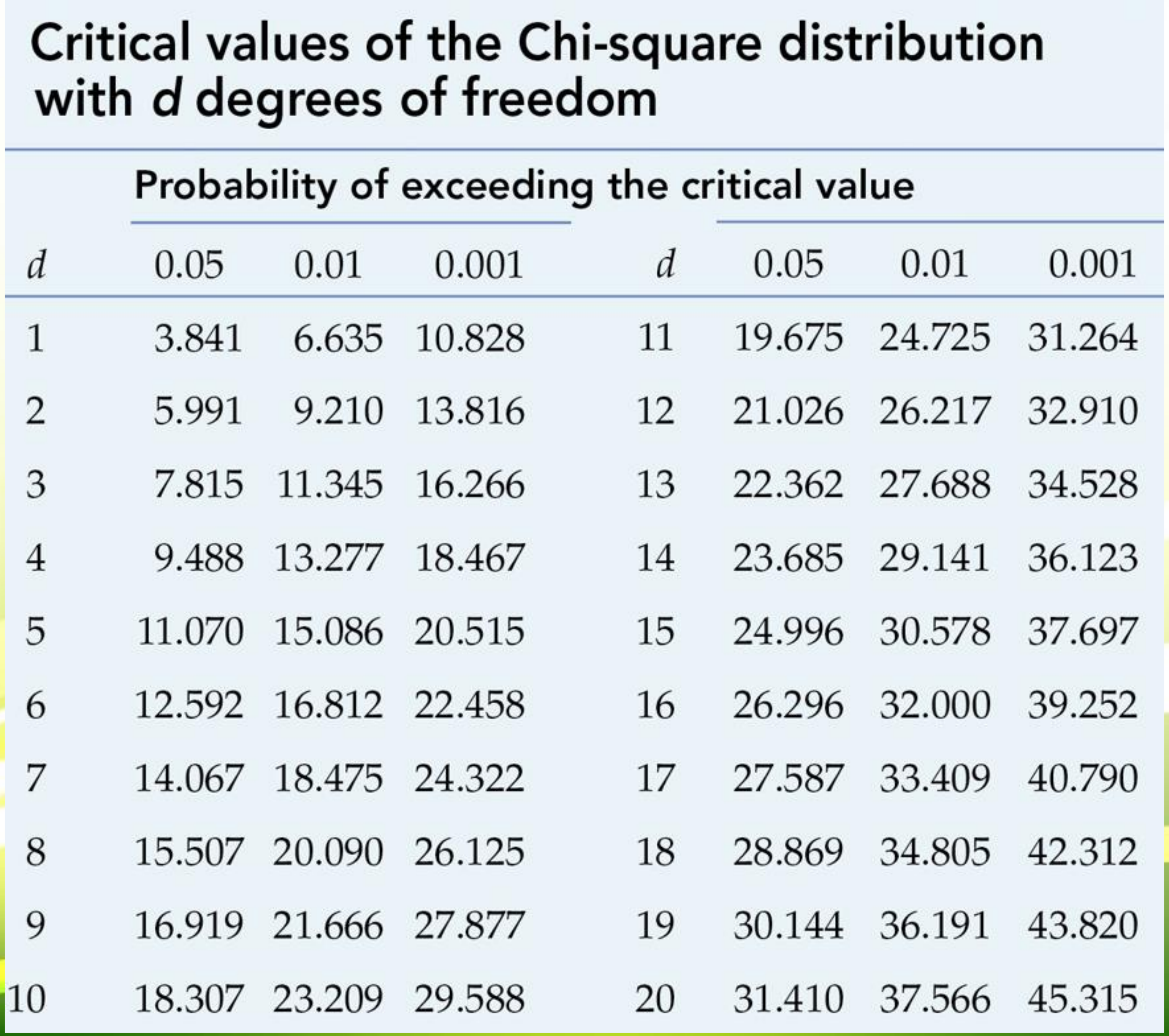
DATA ANALYSIS – CHI SQUARE

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

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

  - *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

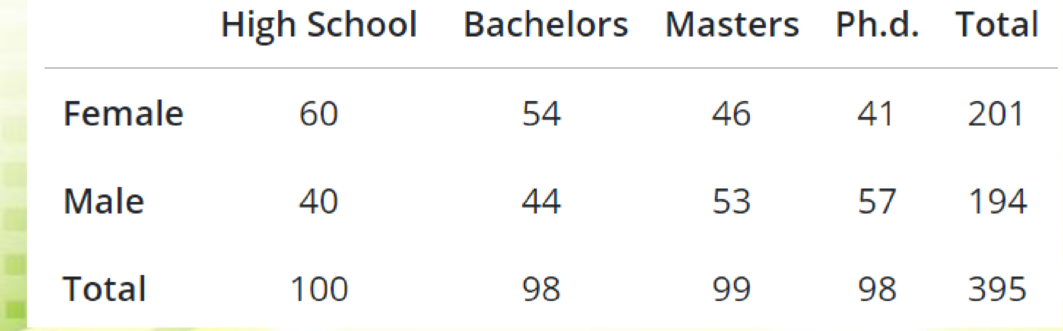


Critical values of the Chi-square distribution with *d* degrees of freedom

	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:

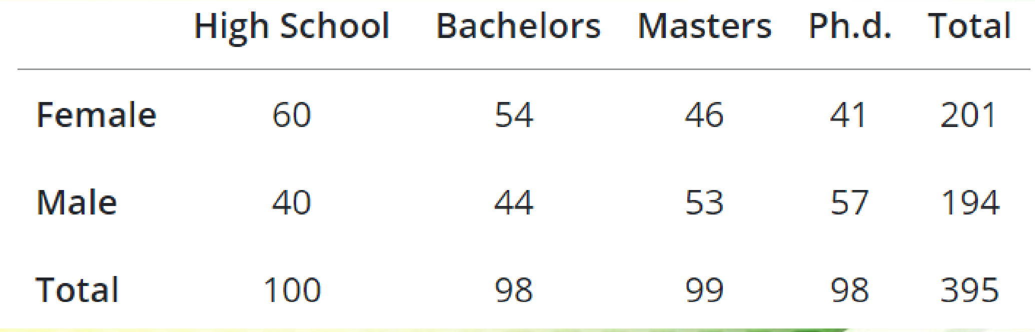


	High School	Bachelors	Masters	Ph.d.	Total
Female	60	54	46	41	201
Male	40	44	53	57	194
Total	100	98	99	98	395

- **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?

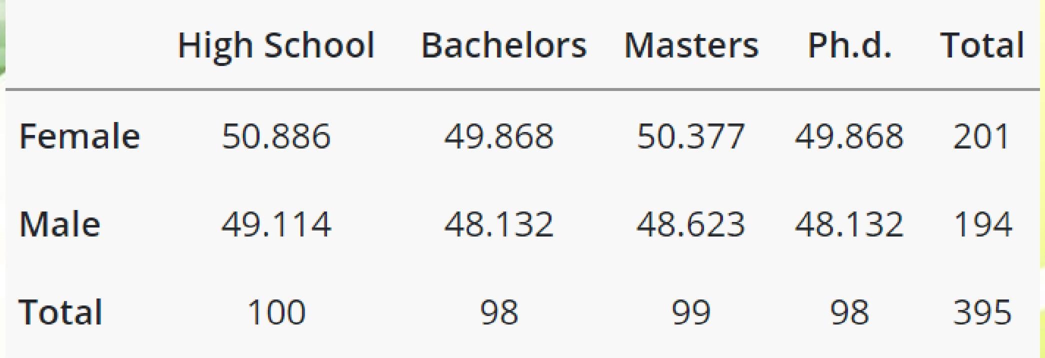
DATA ANALYSIS - CHI SQUARE

Actual Data

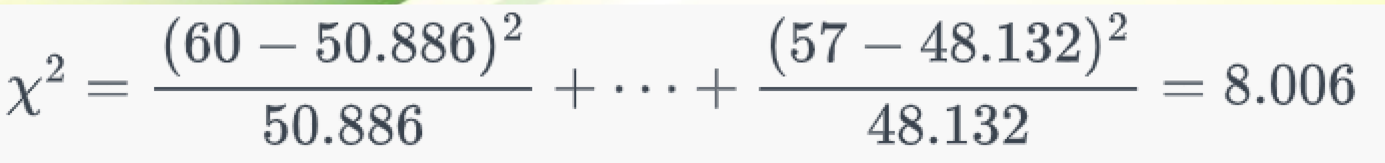


	High School	Bachelors	Masters	Ph.d.	Total
Female	60	54	46	41	201
Male	40	44	53	57	194
Total	100	98	99	98	395

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 χ2 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.

DATA ANALYSIS – T-TEST

One-Sample T – test

- Compares the mean of sample data to a known value.
  - **Example**, one might want to know how sample mean compares to population mean.
- One sample t-test used when population standard deviation not known or sample size is small.
- H0: $\mu = \bar{x}$ (there is no difference in sample and population mean)
- H1: $\mu > \bar{x}$ (there is a difference in sample and population mean)

  - $\bar{x}$: sample mean
  - µ : population mean
  - S : sample standard deviation
  - N : Number of observations

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

DATA ANALYSIS – T-TEST

Example: your company wants to improve sales. Past sales data indicate that the average sale was \$100 per transaction. After training your sales force, recent sales data (taken from a sample of 25 salesmen) indicates an average sale of \$130, with a standard deviation of \$15. Did the training work? Test your hypothesis at a 95% confidence lelve.

H0: $\mu = \bar{x}$
H_1: $\mu > \bar{x}$

sample mean($\bar{x}$) \$130.
population mean(μ) \$100 (from past data).
sample standard deviation(s) = \$15.
Number of observations(n) = 25.

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

calculated t-value = (130 – 100) / ((15 / √(25)) = t = (30 / 3) = 10

degrees of freedom: 25 – 1 = 24.
Alpha = 0.05
Critical t-value = 1.711

calculated t-value > Critical t-value → reject null hypothesis (it’s highly likely that sample mean of sale is greater → sales training was probably a success)

DATA ANALYSIS – T-TEST

Example: A company wants to test the claim that their batteries last more than 40 hours. Using a simple random sample of 15 batteries yielded a mean of 44.9 hours, with a standard deviation of 8.9 hours. Test this claim using a significance level of 0.05..

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

H0: μ = 40
H_1: μ > 40

$$\hat{x} = 44.9\,, \quad \mu = 40 \quad s = 8.9\,, \quad n = 15\,, \; df = n - 1 \rightarrow df = 15 - 1 = 14$$

$$test\ \ statistic: \quad t = \frac{44.9 - 40}{\left(\frac{8.9}{\sqrt{15}}\right)} = 2.13$$

calculated t-value **>** Critical t-value → reject null hypothesis

DATA ANALYSIS – T-TEST

Two-Sample T – test

- Compares the means of two sample data (means).
  - Test the difference (d_0) between two sample means.
  - To determine whether the means are equal.
  - **Example**; Compare the mean scores of two section (sample) of class (population).

- H0: $\mu_1 = \mu_2$ (there is no difference in sample means)
- H1: $\mu_1 \neq \mu_2$ (there is a difference in sample means)

DATA ANALYSIS – T-TEST

Two-Sample T – test

Assuming unequal variances in two sample;

$$T = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_1}}}$$

- $\bar{x}_1$ $\bar{x}_2$: sample means
- S_1 S_2 : sample variances
- n_1 n_2: number of observations in the two sample

Assuming equal variances in two sample;

$$T = \frac{\bar{x}_1 - \bar{x}_2}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

- $\bar{x}_1$ $\bar{x}_2$: sample means
- S_p : pooled sample standard deviation
- n_1 n_2: number of observations in the two sample
- S_1 S_2 : sample variances

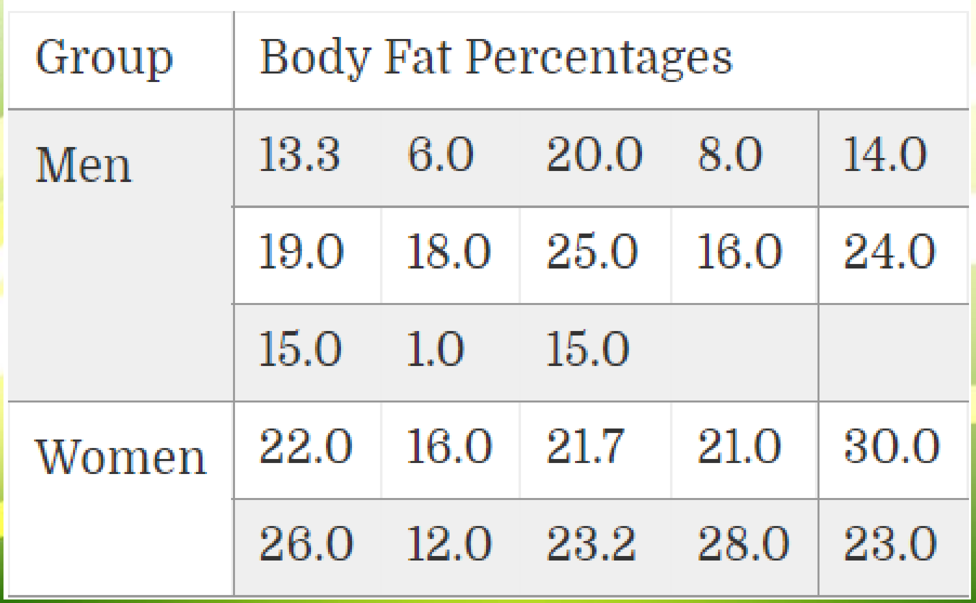
- **degrees of freedom;** $\mathrm{df} = n_1 + n_2 - 2$

DATA ANALYSIS – T-TEST

Example: Average body fat percentages measures a person's fitness and it vary by age. Some study tells, normal range for men is 15-20% body fat, and the normal range for women is 20-25%. Sample data collected from a group of men and women.

There are some overlap in data and also some differences. Just by looking at the data, it's hard to draw any solid conclusions about whether the underlying populations of men and women have the same mean body fat.

Check it statistically.

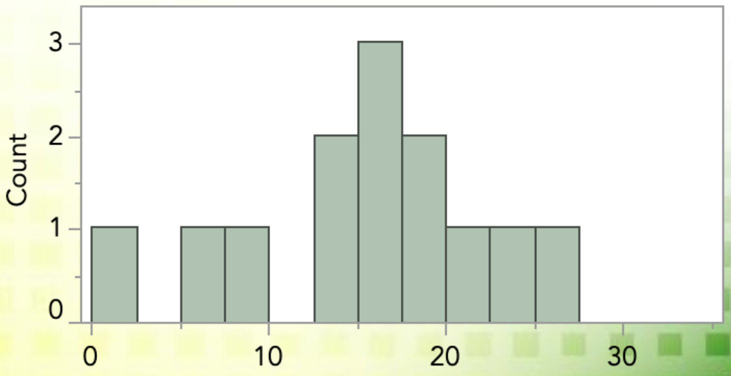


Group	Body Fat Percentages				
Men	13.3	6.0	20.0	8.0	14.0
	19.0	18.0	25.0	16.0	24.0
	15.0	1.0	15.0		
Women	22.0	16.0	21.7	21.0	30.0
	26.0	12.0	23.2	28.0	23.0

DATA ANALYSIS – T-TEST

Group=Men

Body Fat Percentage

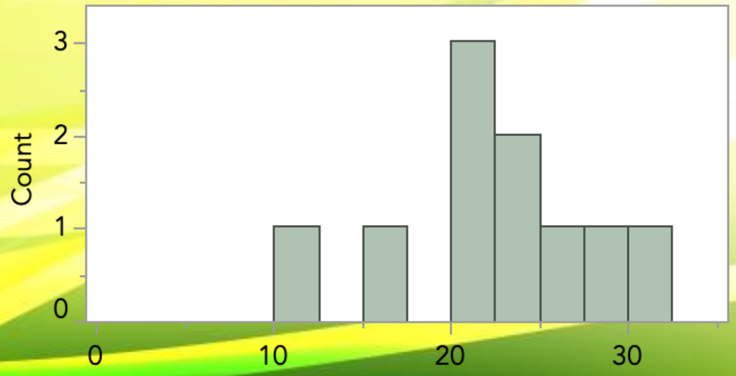


Summary Statistics

Mean	14.95
Std Dev	6.84
Std Err Mean	1.90
Upper 95% Mean	19.08
Lower 95% Mean	10.81
N	13.00

Group=Women

Body Fat Percentage

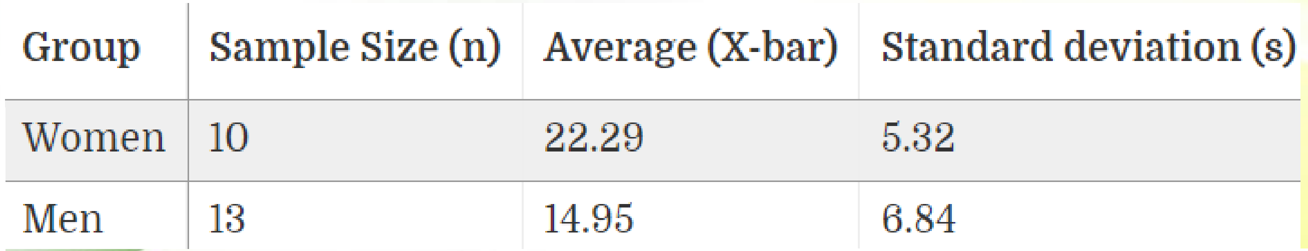


Summary Statistics

Mean	22.29
Std Dev	5.32
Std Err Mean	1.68
Upper 95% Mean	26.10
Lower 95% Mean	18.48
N	10.00

- Two histograms are on same scale.
- There are no very unusual points (*outliers)*.
- data look roughly bell-shaped (normal distribution seems reasonable).
- Examining summary statistics, standard deviations looks similar → supports the idea of equal variances.
  - ○ THIS can also be checked using **test for variances**.

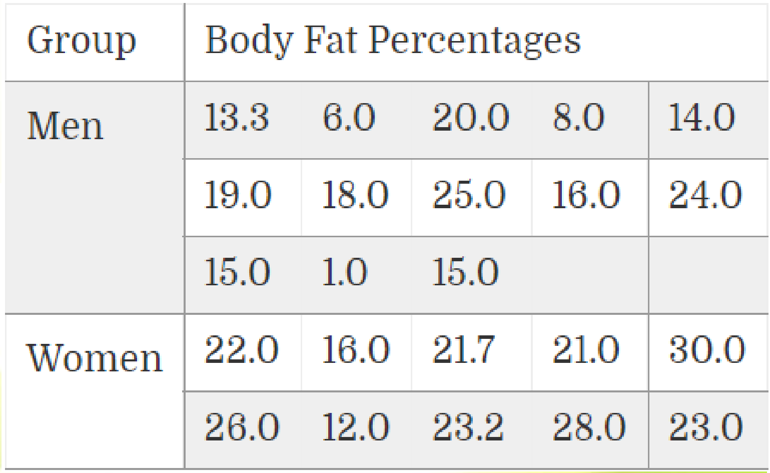
DATA ANALYSIS – T-TEST



Group	Sample Size (n)	Average (X-bar)	Standard deviation (s)
Women	10	22.29	5.32
Men	13	14.95	6.84

Assuming equal variances in two sample

$$T = \frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \qquad s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$



Group	Body Fat Percentages				
Men	13.3	6.0	20.0	8.0	14.0
	19.0	18.0	25.0	16.0	24.0
	15.0	1.0	15.0		
Women	22.0	16.0	21.7	21.0	30.0
	26.0	12.0	23.2	28.0	23.0

$$t = \frac{\text{difference of group averages}}{\text{standard error of difference}} = \frac{7.34}{(6.24\times\sqrt{(1/10+1/13)})} = \frac{7.34}{2.62} = 2.80$$

H0: $\mu_1 = \mu_2$ H1: $\mu_1 \neq \mu_2$

degrees of freedom: df=n1+n2−2=10+13−2=21 Alpha = 0.05

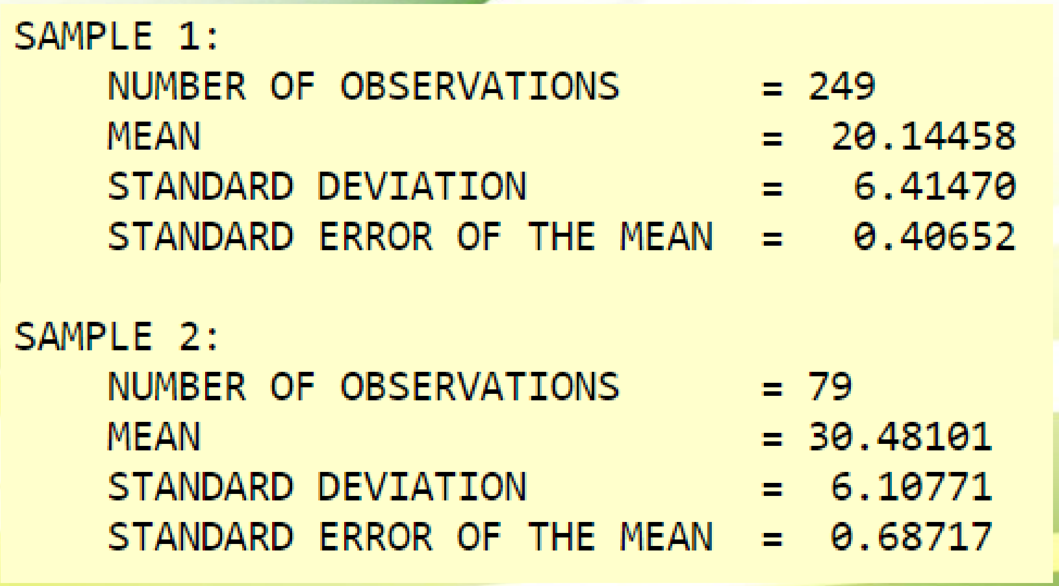
Critical t-value = 2.080

calculated t-value > Critical t-value → reject null hypothesis → mean body fat for men and women are NOT equal.

DATA ANALYSIS – T-TEST

Example: One data set contains miles per gallon for U.S. cars (sample 1) and for Japanese cars (sample 2); the summary statistics for each sample are shown below.

Apply t-test to conclude whether fuel consumptions in both countries are identical (at alpha = 0.05).



```
SAMPLE 1:
    NUMBER OF OBSERVATIONS      = 249
    MEAN                        =  20.14458
    STANDARD DEVIATION          =   6.41470
    STANDARD ERROR OF THE MEAN  =   0.40652

SAMPLE 2:
    NUMBER OF OBSERVATIONS      = 79
    MEAN                        = 30.48101
    STANDARD DEVIATION          =  6.10771
    STANDARD ERROR OF THE MEAN  =  0.68717
```

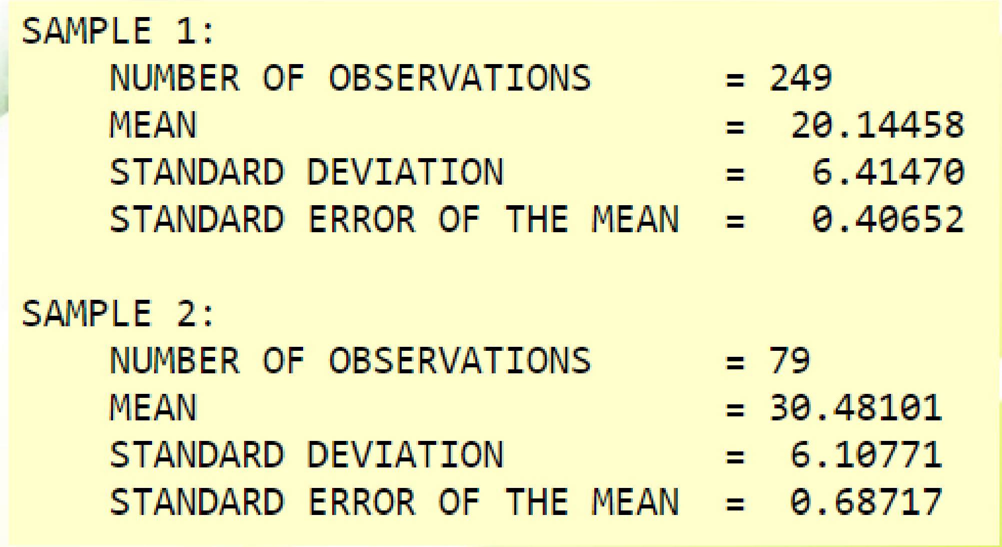
DATA ANALYSIS – T-TEST

Hypothesis to test that the means are equal for two samples.
We assume that the variances for the two samples are equal.

H0: μ1 = μ2
Ha: μ1 ≠ μ2

Test statistic: T = -12.62059
Pooled standard deviation: sp = 6.34260
Degrees of freedom: ν = 326
Significance level: α = 0.05
Critical value = 1.9673
Critical region: Reject H0 if |T| > 1.9673

absolute value of test statistic (12.62059) > critical value (1.9673) → reject null hypothesis
conclude that two sample means are different at 0.05 significance level.



```
SAMPLE 1:
    NUMBER OF OBSERVATIONS      = 249
    MEAN                        =  20.14458
    STANDARD DEVIATION          =   6.41470
    STANDARD ERROR OF THE MEAN  =   0.40652

SAMPLE 2:
    NUMBER OF OBSERVATIONS      = 79
    MEAN                        = 30.48101
    STANDARD DEVIATION          =  6.10771
    STANDARD ERROR OF THE MEAN  =  0.68717
```

$$T = \frac{\bar{x}_1 - \bar{x}_2}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

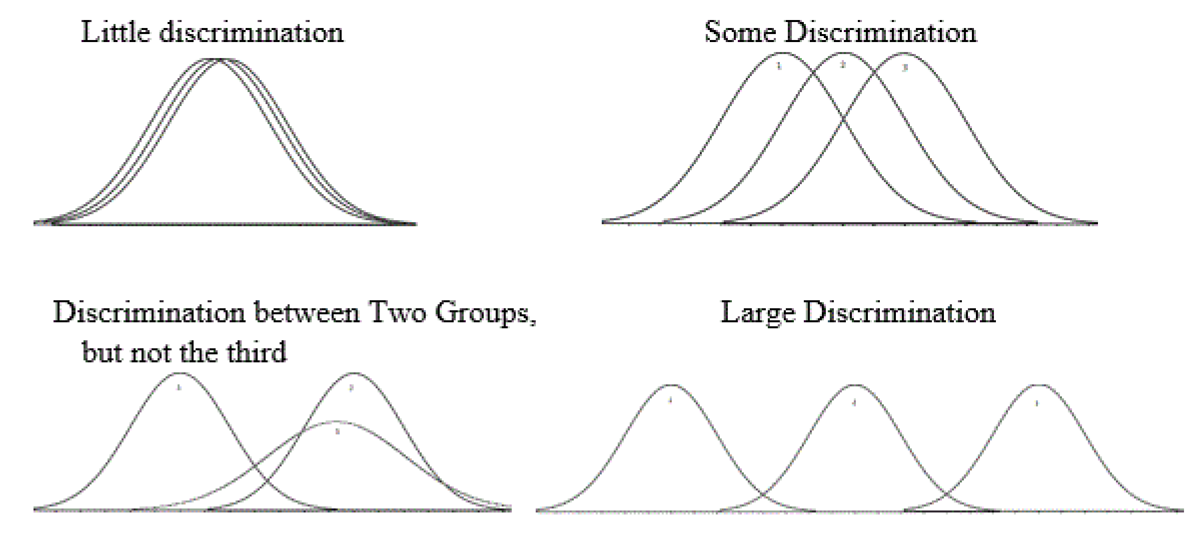
DATA ANALYSIS - ANOVA

- **AN**alysis **O**f **VA**riance
- Statistical method to examine how a dependent variable changes as value of independent variable changes.
- ANOVA checks impact of one or more factors by comparing the means of different samples.
- Three types of ANOVA:
    - **One-way ANOVA** (more common): One independent variable.
    - **Two-way (or factorial) ANOVA**: Use of two independent variables.
    - **N-way ANOVA**: more than two (n) independent variables

DATA ANALYSIS - ANOVA

- **Mean**: Simple or arithmetic average of a range of values.
- **Sample mean** (μ_1, μ_2, μ_3) : Mean for single sample (individually)
- **Grand mean** (μ) : Mean of sample means.
- ***Null hypothesis***: All sample means are equal (they don't have any significant difference). $H_o : \mu_1 = \mu_2 = \cdots = \mu_L$
  - They can be considered as a part of a larger set of population.
- *Alternate hypothesis:* At least one of the sample means is different from rest of the sample means. $H_1 : \mu_l \neq \mu_m$
- **Between-group variability:** Variations between the distributions.

  - As samples overlap, their individual means won't differ by great margin → difference between individual means and grand mean won't be significant enough.
  - As samples differ from each other by big margin, their individual means would also differ → difference between individual means and grand mean would be significant.



Little discrimination

Some Discrimination

Discrimination between Two Groups, but not the third

Large Discrimination

DATA ANALYSIS - ANOVA

Sum-of-squares for between-group variability:

- Deviation is given greater weight if it’s from larger sample
- Multiply each squared deviation by each sample size and add them up

$$\mathrm{SS}_{\mathrm{between}} = \mathrm{n}_1(\overline{x_1} - \overline{x_G})^2 + \mathrm{n}_2(\overline{x_2} - \overline{x_G})^2 + \mathrm{n}_3(\overline{x_3} - \overline{x_G})^2 + \dots.\ \mathrm{n}_k(\overline{x_k} - \overline{x_k})^2$$

$$\mathrm{SSB} = \sum \mathrm{n}_j(\overline{\mathrm{X}}_{j} - \overline{\mathrm{X}})^2.$$

Mean-square between-group variability:

- Find each squared deviation → weigh them by their sample size → sum them up → divide by degrees of freedom.

$$\mathrm{MS}_{\mathrm{between}} = \frac{\mathrm{n}1(\overline{x_1} - \overline{x_G})2 + \mathrm{n}2(\overline{x_2} - \overline{x_G})2 + \mathrm{n}3(\overline{x_3} - \overline{x_G})2 + \dots.\mathrm{nk}(\overline{x_k} - \overline{x_k})2}{k-1}$$

$$\mathrm{MSB} = \mathrm{SSB} / (\mathrm{k} - 1)$$

Degrees of freedom (df)

$\mathrm{df}_{\mathrm{within}}$ = n - k

$\mathrm{df}_{\mathrm{between}}$ = k - 1

- *N: Sum of the sample sizes*
- *K: number of samples*

DATA ANALYSIS - ANOVA

Within Group Variability: Variations within a sample (error).

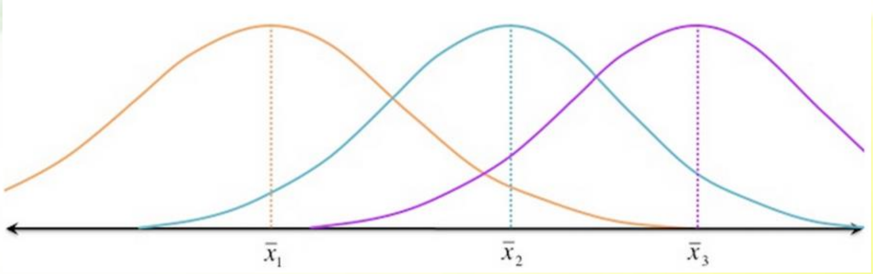
- Variations caused by differences within individual groups.
- *As the spread (variability) of each sample is increased, their distributions overlap and they become part of a big population.*
- *For distribution of samples with less variability (though means of samples are same), they seem to belong to different populations.*
- **Sum of squares for within-group variability** → how much each value in each sample differs from its respective sample mean.

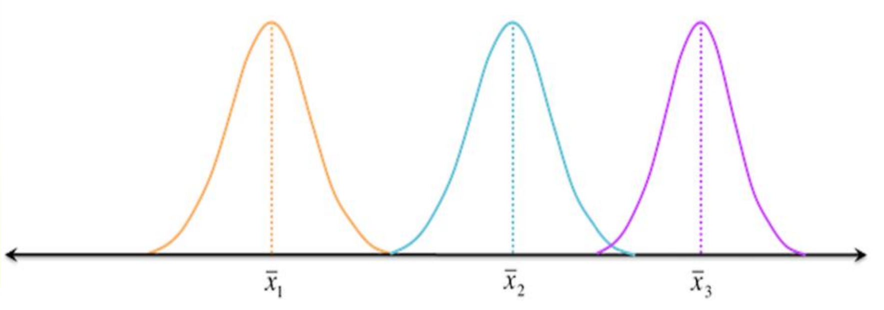
$$SS_{within} = \Sigma(x_{i1} - \bar{x}_1)^2 + \Sigma(x_{i2} - \bar{x}_2)^2 + \ldots + \Sigma(x_{ik} - \bar{x}_k)^2$$
$$= \Sigma(x_{ij} - \bar{x}_j)^2$$

- **Mean square within-group variability:**

$$MS_{within} = \sum(x_{ij} - \overline{x_j})^2 / (N - k)$$

$$MSE = SSE / (N - k)$$





$$SSE = \sum\sum(X - \overline{X}_j)^2$$

$$\frac{\sum_{i=1}^{k}(n_i - 1)s_i^2}{N - k}$$

Total sum of squares, SST = SSB + SSE

DATA ANALYSIS - ANOVA

F-Statistic

- *F-Ratio* measures whether means of different samples are significantly different or not.
- Lower the F-Ratio, more similar are the sample means *(Accept null hypothesis)*.

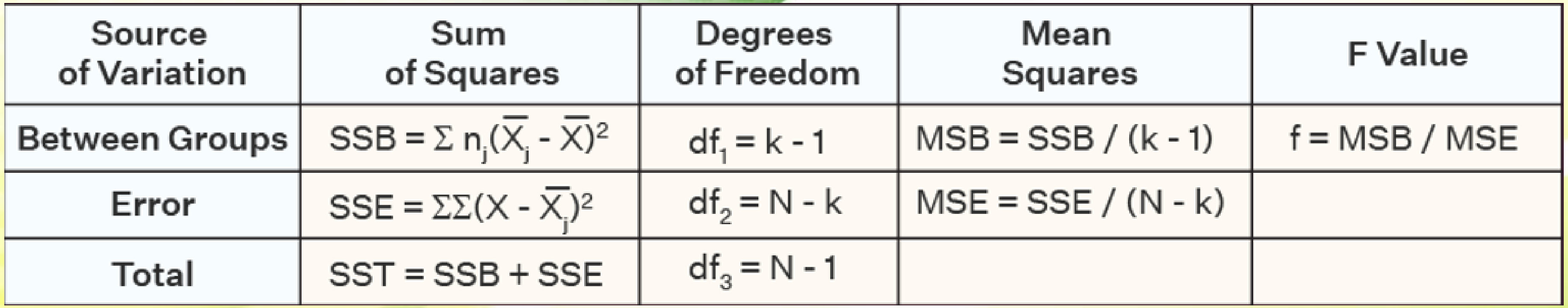
f = Between group variability / Within group variability

ANOVA test statistic, f = MSB / MSE

- Calculated F-statistic is compared with F-critical value for making a conclusion.
- If calculated F-statistic > F-critical value (for a specific α/significance level) → reject null hypothesis.
- F-distribution does not have any negative values *(because; between and within-group variability are always positive due to squaring each deviation)*.

DATA ANALYSIS - ANOVA

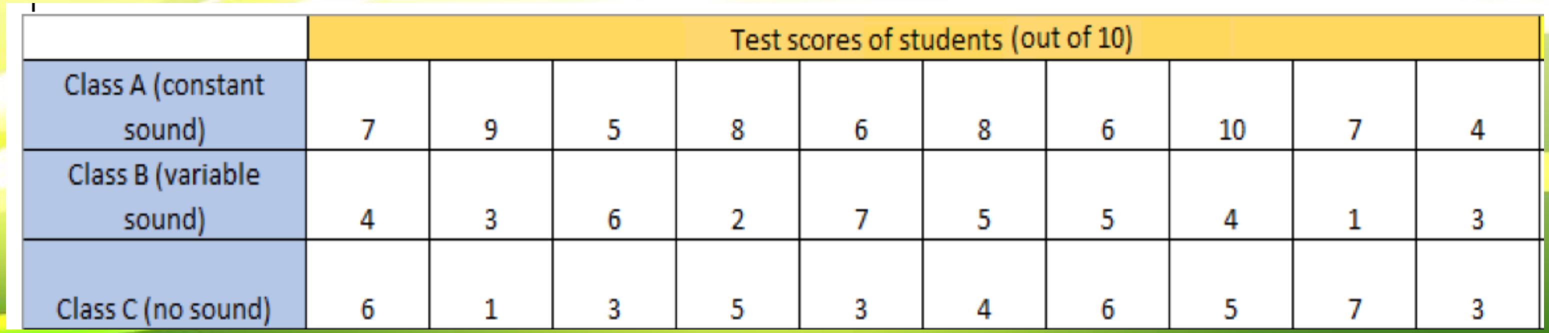
Anova Table



Source of Variation	Sum of Squares	Degrees of Freedom	Mean Squares	F Value
Between Groups	$SSB = \Sigma\, n_j(\bar{X}_j - \bar{X})^2$	$df_1 = k - 1$	MSB = SSB / (k - 1)	f = MSB / MSE
Error	$SSE = \Sigma\Sigma(X - \bar{X}_j)^2$	$df_2 = N - k$	MSE = SSE / (N - k)	
Total	SST = SSB + SSE	$df_3 = N - 1$		

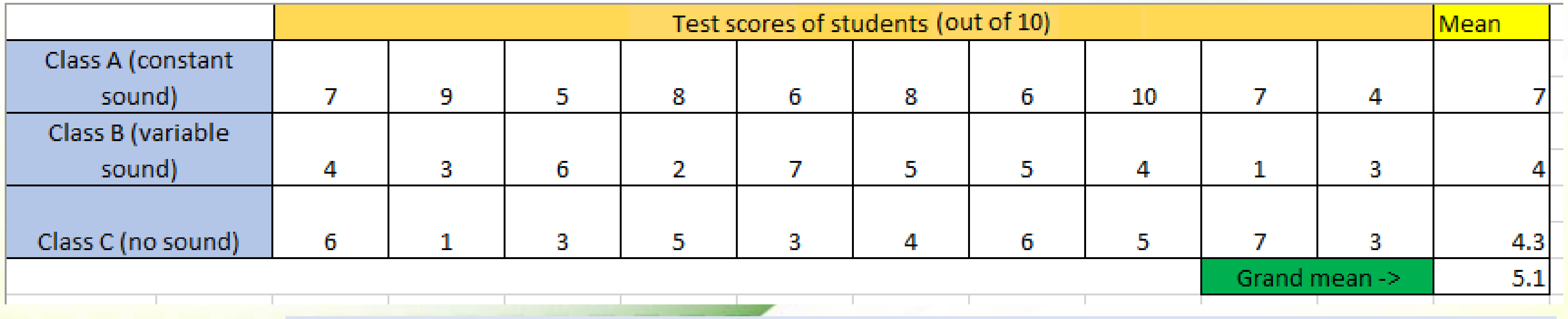
DATA ANALYSIS - ANOVA

- **Claim**: using music in class enhances concentration and consequently helps students absorb more information.
- **Reaction**: What if it affected results of students in negative way? What kind of music would be a good choice for this?
- **Approach**: *Implement it on smaller group of randomly selected students from three different classes & Analyze.*
  - Take three different groups of ten randomly selected students (*same age*) from three different classrooms.
  - Each classroom was provided with a different environment for students to study.
  - Classroom A had constant music being played in the background, classroom B had variable music being played and classroom C was a regular class with NO music playing.
  - After one month, conduct a test for all the three groups and collect their test scores.



	Test scores of students (out of 10)									
Class A (constant sound)	7	9	5	8	6	8	6	10	7	4
Class B (variable sound)	4	3	6	2	7	5	5	4	1	3
Class C (no sound)	6	1	3	5	3	4	6	5	7	3

DATA ANALYSIS - ANOVA



	Test scores of students (out of 10)										Mean
Class A (constant sound)	7	9	5	8	6	8	6	10	7	4	7
Class B (variable sound)	4	3	6	2	7	5	5	4	1	3	4
Class C (no sound)	6	1	3	5	3	4	6	5	7	3	4.3
									Grand mean ->		5.1

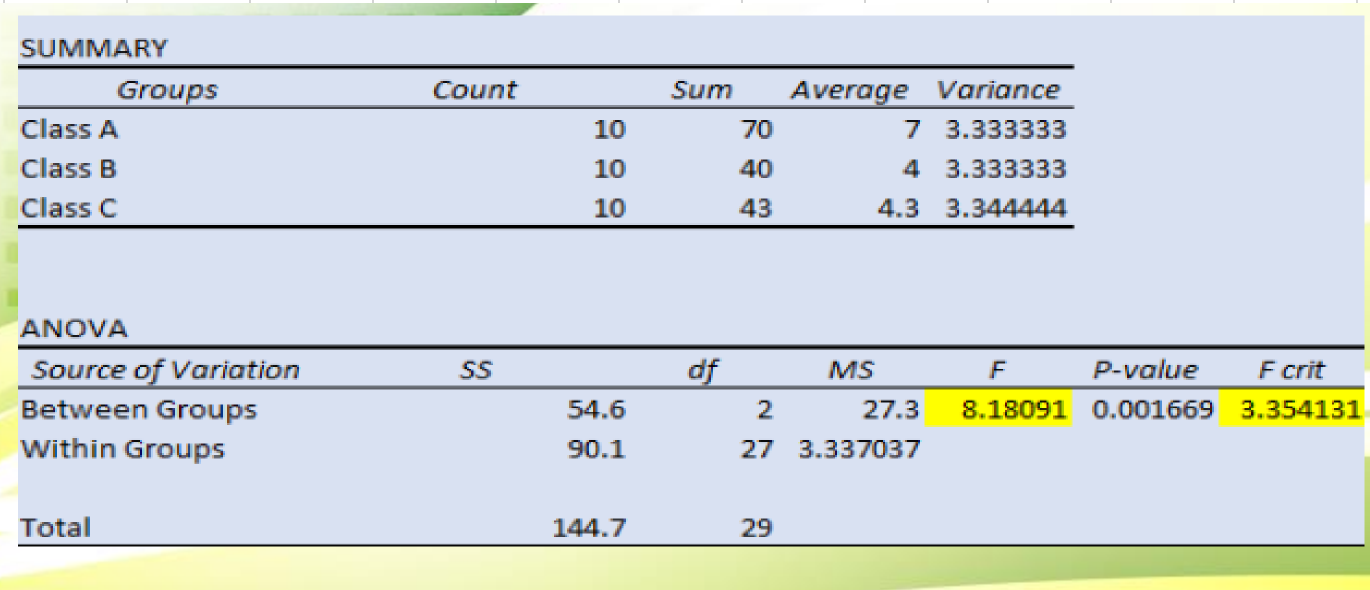
Music treatment was helpful in improving test results of students.

H_0: The means are equal.

H_a: The means are not equal.

P or α = 0.05 (5%)

$\mu_1 = 7, \mu_2 = 4,\ \mu_3 = 4.3\ \&\ \mu = 5.1.$



SUMMARY

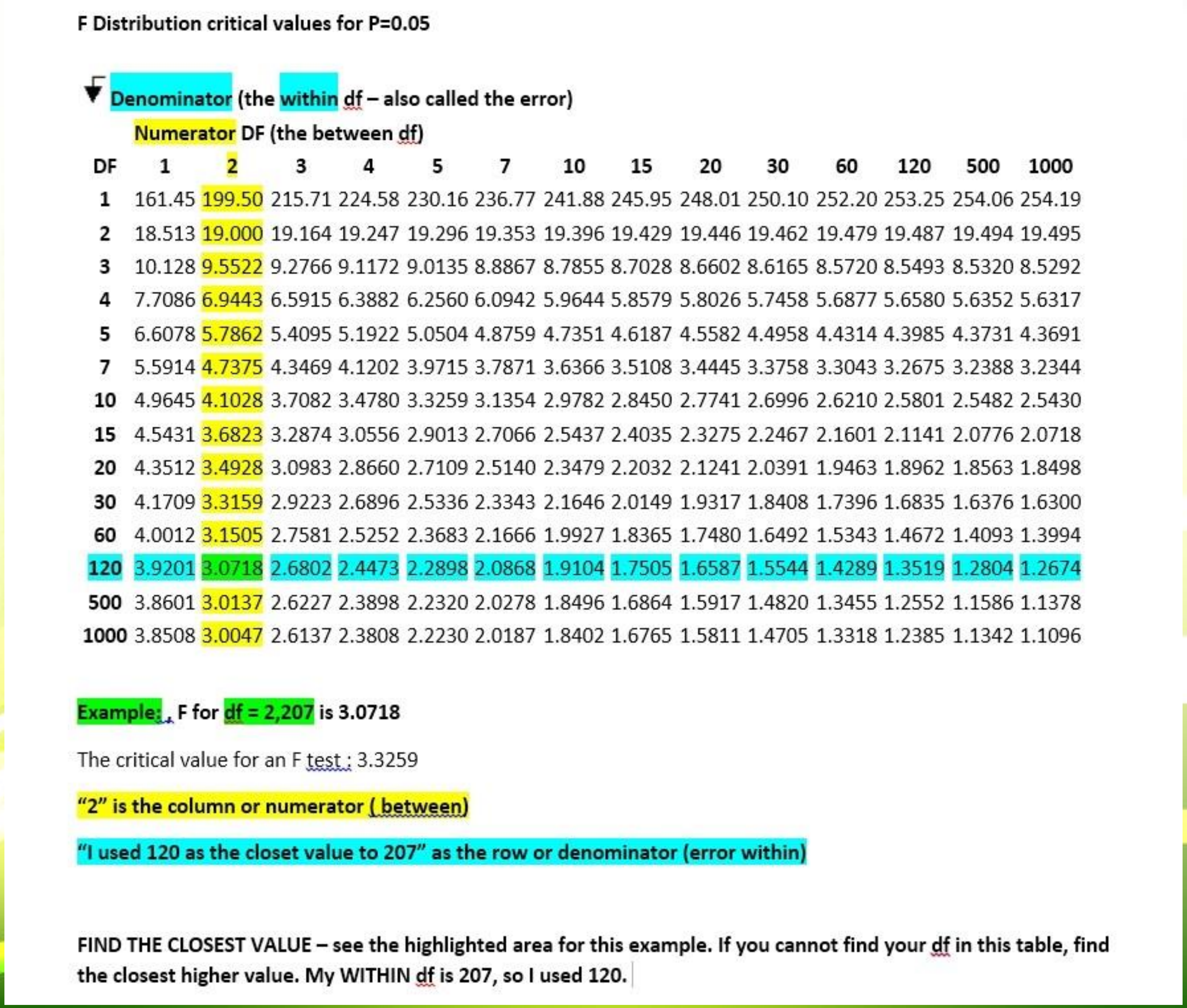
Groups	*Count*	*Sum*	*Average*	*Variance*
Class A	10	70	7	3.333333
Class B	10	40	4	3.333333
Class C	10	43	4.3	3.344444

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	54.6	2	27.3	8.18091	0.001669	3.354131
Within Groups	90.1	27	3.337037			
Total	144.7	29				

- F-value > F-critical value for alpha level selected (0.05) → **reject null hypothesis.**
- At least one of the three samples have significantly different means → belong to an entirely different population.

DATA ANALYSIS - ANOVA



F Distribution critical values for P=0.05

Denominator (the within df – also called the error)

Numerator DF (the between df)

DF	1	2	3	4	5	7	10	15	20	30	60	120	500	1000
1	161.45	199.50	215.71	224.58	230.16	236.77	241.88	245.95	248.01	250.10	252.20	253.25	254.06	254.19
2	18.513	19.000	19.164	19.247	19.296	19.353	19.396	19.429	19.446	19.462	19.479	19.487	19.494	19.495
3	10.128	9.5522	9.2766	9.1172	9.0135	8.8867	8.7855	8.7028	8.6602	8.6165	8.5720	8.5493	8.5320	8.5292
4	7.7086	6.9443	6.5915	6.3882	6.2560	6.0942	5.9644	5.8579	5.8026	5.7458	5.6877	5.6580	5.6352	5.6317
5	6.6078	5.7862	5.4095	5.1922	5.0504	4.8759	4.7351	4.6187	4.5582	4.4958	4.4314	4.3985	4.3731	4.3691
7	5.5914	4.7375	4.3469	4.1202	3.9715	3.7871	3.6366	3.5108	3.4445	3.3758	3.3043	3.2675	3.2388	3.2344
10	4.9645	4.1028	3.7082	3.4780	3.3259	3.1354	2.9782	2.8450	2.7741	2.6996	2.6210	2.5801	2.5482	2.5430
15	4.5431	3.6823	3.2874	3.0556	2.9013	2.7066	2.5437	2.4035	2.3275	2.2467	2.1601	2.1141	2.0776	2.0718
20	4.3512	3.4928	3.0983	2.8660	2.7109	2.5140	2.3479	2.2032	2.1241	2.0391	1.9463	1.8962	1.8563	1.8498
30	4.1709	3.3159	2.9223	2.6896	2.5336	2.3343	2.1646	2.0149	1.9317	1.8408	1.7396	1.6835	1.6376	1.6300
60	4.0012	3.1505	2.7581	2.5252	2.3683	2.1666	1.9927	1.8365	1.7480	1.6492	1.5343	1.4672	1.4093	1.3994
120	3.9201	3.0718	2.6802	2.4473	2.2898	2.0868	1.9104	1.7505	1.6587	1.5544	1.4289	1.3519	1.2804	1.2674
500	3.8601	3.0137	2.6227	2.3898	2.2320	2.0278	1.8496	1.6864	1.5917	1.4820	1.3455	1.2552	1.1586	1.1378
1000	3.8508	3.0047	2.6137	2.3808	2.2230	2.0187	1.8402	1.6765	1.5811	1.4705	1.3318	1.2385	1.1342	1.1096

Example: F for df = 2,207 is 3.0718

The critical value for an F test : 3.3259

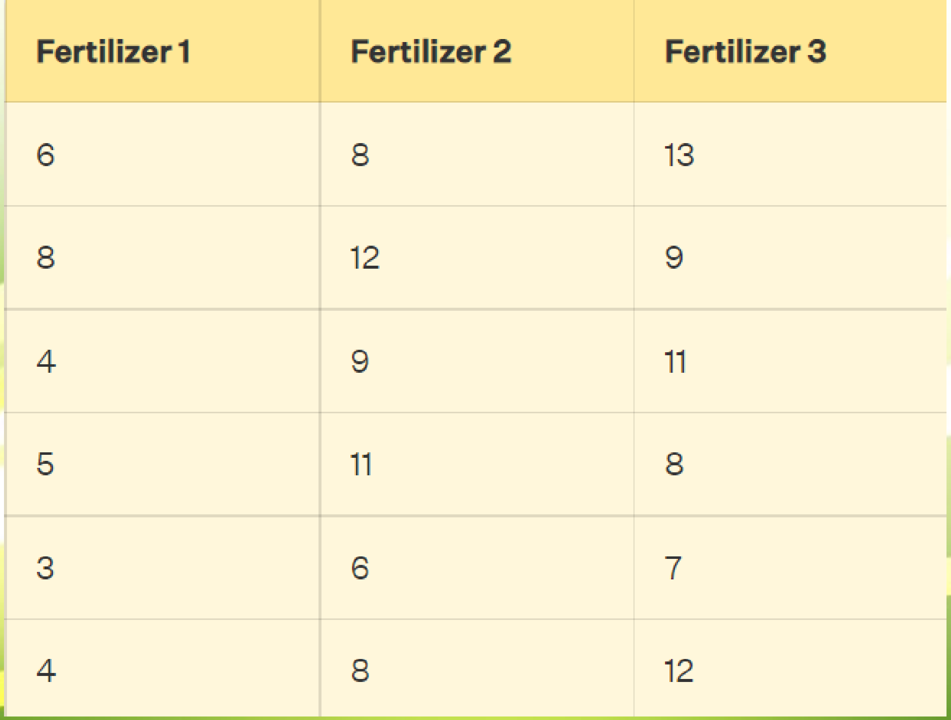
"2" is the column or numerator (between)

"I used 120 as the closet value to 207" as the row or denominator (error within)

FIND THE CLOSEST VALUE – see the highlighted area for this example. If you cannot find your df in this table, find the closest higher value. My WITHIN df is 207, so I used 120.

DATA ANALYSIS - ANOVA

Example 1: Three types of fertilizers are used on three groups of plants for 5 weeks. We want to check if there is a difference in the mean growth of each group. Using the data given below apply a one way ANOVA test at 0.05 significant level.

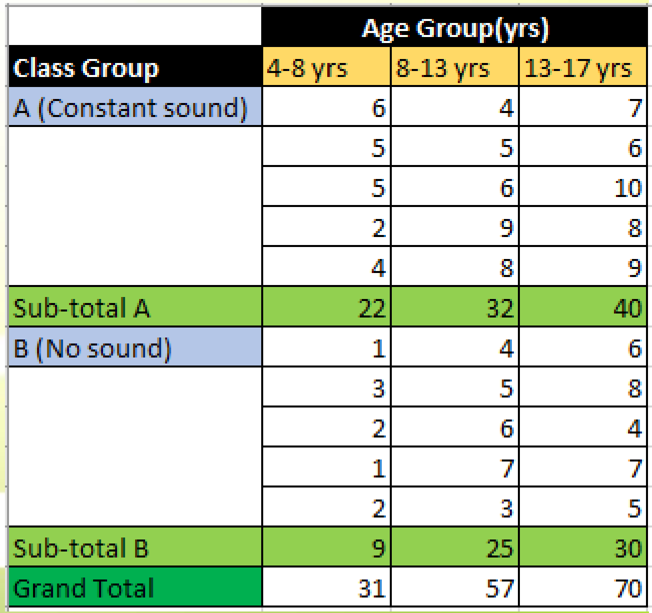


Fertilizer 1	Fertilizer 2	Fertilizer 3
6	8	13
8	12	9
4	9	11
5	11	8
3	6	7
4	8	12

DATA ANALYSIS - ANOVA

Two-Way ANOVA:

- Using one-way ANOVA → music treatment was helpful in improving test results of students.
  - It was conducted on students of **same age**.
- What if the process was to affect different age groups of students in different ways? Maybe the process had varying effects depending upon the teacher who taught the class?
- How to be sure as to which factor(s) is affecting results of students more? Maybe the age group is a more dominant factor responsible for student's performance than musical process?
- *In one-way ANOVA test, group subjected to 'variable music' and 'no music at all' performed almost equally.*
  - *It means that variable music did not have any significant effect on the students.*
  - *So, while performing two-way ANOVA, do not consider "variable music". Rather a new factor, age, will be introduced to find out how music affects when applied to students of different age groups.*



	Age Group(yrs)		
Class Group	4-8 yrs	8-13 yrs	13-17 yrs
A (Constant sound)	6	4	7
	5	5	6
	5	6	10
	2	9	8
	4	8	9
Sub-total A	22	32	40
B (No sound)	1	4	6
	3	5	8
	2	6	4
	1	7	7
	2	3	5
Sub-total B	9	25	30
Grand Total	31	57	70

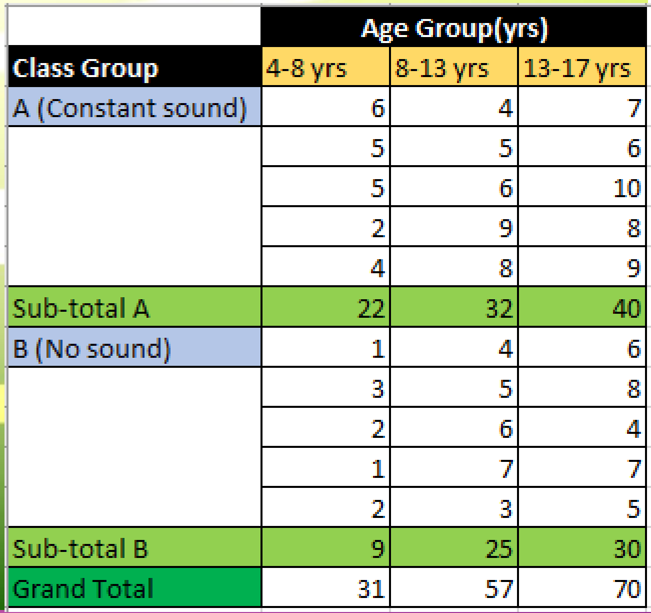
DATA ANALYSIS - ANOVA

- There are two factors – class group and age group with two and three levels respectively.
- Total six different groups of students based on different permutations of class groups and age groups and each different group has a sample size of 5 students.
- Two-way ANOVA tells about main effect and interaction effect.
  - Main effect is similar to one-way ANOVA, where effect of music and age would be measured separately.
  - Interaction effect is where both music and age are considered at the same time.
- So, two-way ANOVA can have up to three hypotheses.
  - Two null hypotheses will be tested if only one observation placed in each cell.

    H1: *All the music treatment groups have equal mean score.*

    H2: *All the age groups have equal mean score.*
  - For multiple observations in cells, third hypothesis can be:

    H3: *The factors are independent or the interaction effect does not exist.*



	Age Group(yrs)		
Class Group	4-8 yrs	8-13 yrs	13-17 yrs
A (Constant sound)	6	4	7
	5	5	6
	5	6	10
	2	9	8
	4	8	9
Sub-total A	22	32	40
B (No sound)	1	4	6
	3	5	8
	2	6	4
	1	7	7
	2	3	5
Sub-total B	9	25	30
Grand Total	31	57	70

DATA ANALYSIS - ANOVA

6 combination of hypotheses:

- H01 : *All the music treatment groups have equal mean score.*
- H11 : *All the music treatment groups do not have equal mean score.*

- H02 : *All the age groups have equal mean score.*
- H12 : *All the age groups do not have equal mean score.*

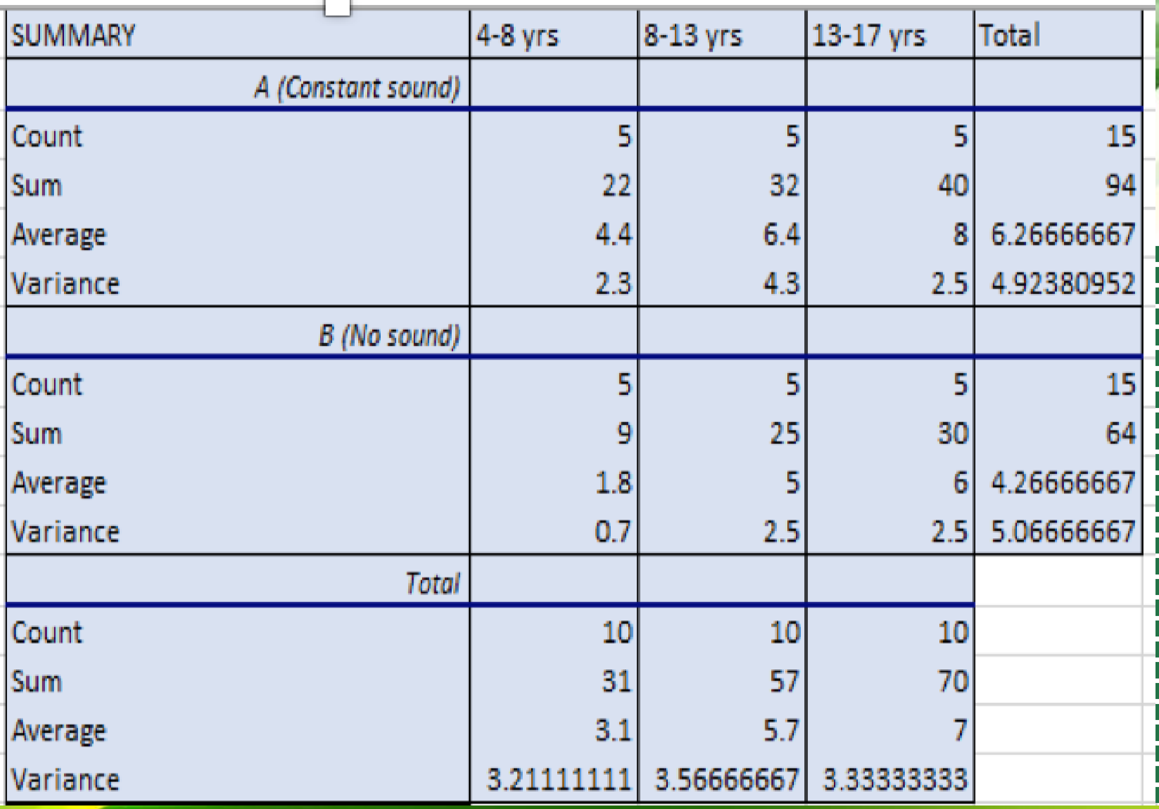
- H03 : *Interaction effect does not exist… The factors are independent.*
- H13 : *Interaction effect exists… Factors are not independent.*

- F-statistic is computed for each pair of hypothesis (H0 or H1 pair)
- (For a particular effect), if F value > respective F-critical value, then reject null hypothesis for that particular effect.

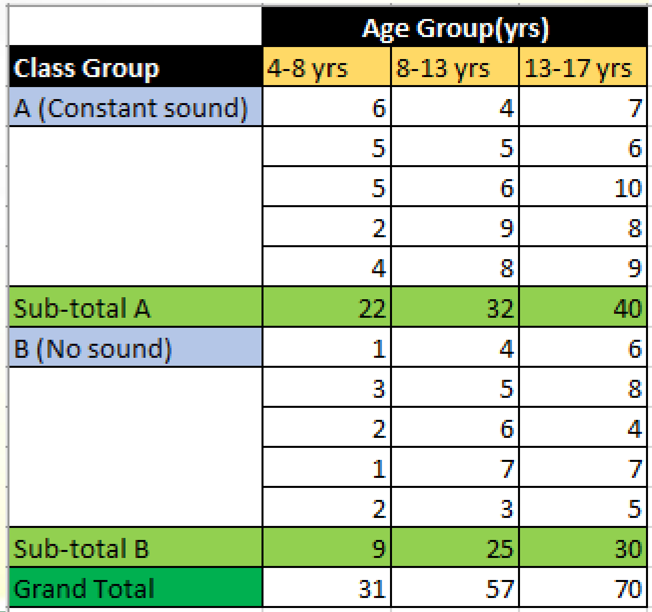
DATA ANALYSIS - ANOVA

Contingency & Summary tables:

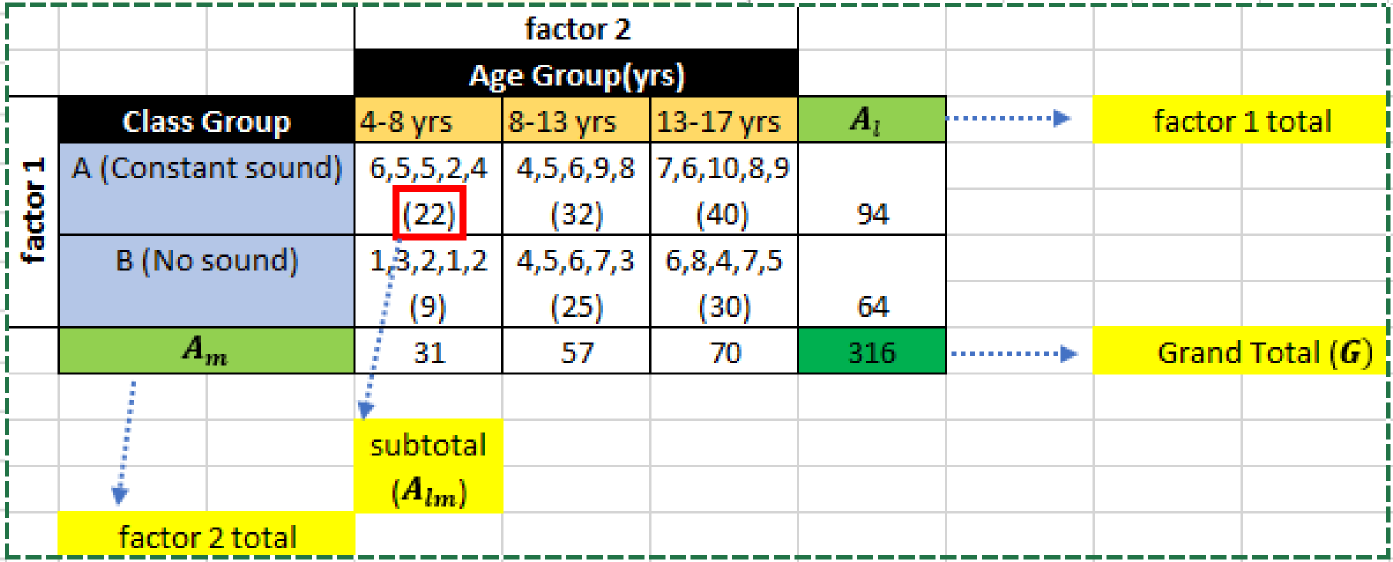
- *Represents statistical measure of the samples based only on factor 1 & 2, separately.*



SUMMARY	4-8 yrs	8-13 yrs	13-17 yrs	Total
A (Constant sound)				
Count	5	5	5	15
Sum	22	32	40	94
Average	4.4	6.4	8	6.26666667
Variance	2.3	4.3	2.5	4.92380952
B (No sound)				
Count	5	5	5	15
Sum	9	25	30	64
Average	1.8	5	6	4.26666667
Variance	0.7	2.5	2.5	5.06666667
Total				
Count	10	10	10	
Sum	31	57	70	
Average	3.1	5.7	7	
Variance	3.21111111	3.56666667	3.33333333	



	Age Group(yrs)		
Class Group	4-8 yrs	8-13 yrs	13-17 yrs
A (Constant sound)	6	4	7
	5	5	6
	5	6	10
	2	9	8
	4	8	9
Sub-total A	22	32	40
B (No sound)	1	4	6
	3	5	8
	2	6	4
	1	7	7
	2	3	5
Sub-total B	9	25	30
Grand Total	31	57	70



		factor 2				
		Age Group(yrs)				
	Class Group	4-8 yrs	8-13 yrs	13-17 yrs	A_l	factor 1 total
factor 1	A (Constant sound)	6,5,5,2,4 (22)	4,5,6,9,8 (32)	7,6,10,8,9 (40)	94	
	B (No sound)	1,3,2,1,2 (9)	4,5,6,7,3 (25)	6,8,4,7,5 (30)	64	
	A_m	31	57	70	316	Grand Total (G)

subtotal (A_{lm})

factor 2 total

DATA ANALYSIS - ANOVA

$$SS_{within} = \Sigma(x_{i1} - \bar{x}_1)^2 + \Sigma(x_{i2} - \bar{x}_2)^2 + \ldots + \Sigma(x_{ik} - \bar{x}_k)^2$$
$$= \Sigma(x_{ij} - \bar{x}_j)^2$$

- **SS**$_{\textbf{interaction}}$ and **df**$_{\textbf{interaction}}$ defines the combined effect of the two factors.

$$SS_{within} = \sum_{i=1}^{2}\sum_{j=1}^{3}\sum_{k=1}^{5}\left(Y_{ijk} - \bar{Y_{ij\cdot}}\right)^2 = 59.2$$

$$df_{within} = (r-1) * a * b = 4 * 2 * 3 = 24$$

$$MS_{within} = SS_{within} / df_{within} = 59.2/24 = 2.46$$

$$SS_{sound} = r.b.\sum_{i=1}^{2}(\bar{Y_i} - \bar{Y})^2 = 30$$

$$df_{sound} = 2\text{-}1 = 1$$

$$SS_{age} = r.a.\sum_{i=1}^{3}(\bar{Y_i} - \bar{Y})^2 = 78.86$$

$$df_{age} = 3\text{-}1 = 2$$

$$SS_{interaction} = r \times \sum_{i=1}^{2}\sum_{j=1}^{3}\left(\bar{Y_{ij\cdot}} - \bar{Y_{i\cdot\cdot}} - \bar{Y_{\cdot j\cdot}} + \bar{Y_{\cdot\cdot\cdot}}\right)^2 = 1.8$$

$$df_{interaction} = (a-1) \times (b-1) = 2$$

$$MS_{interaction} = SS_{interaction} / df_{interaction} = 0.9$$

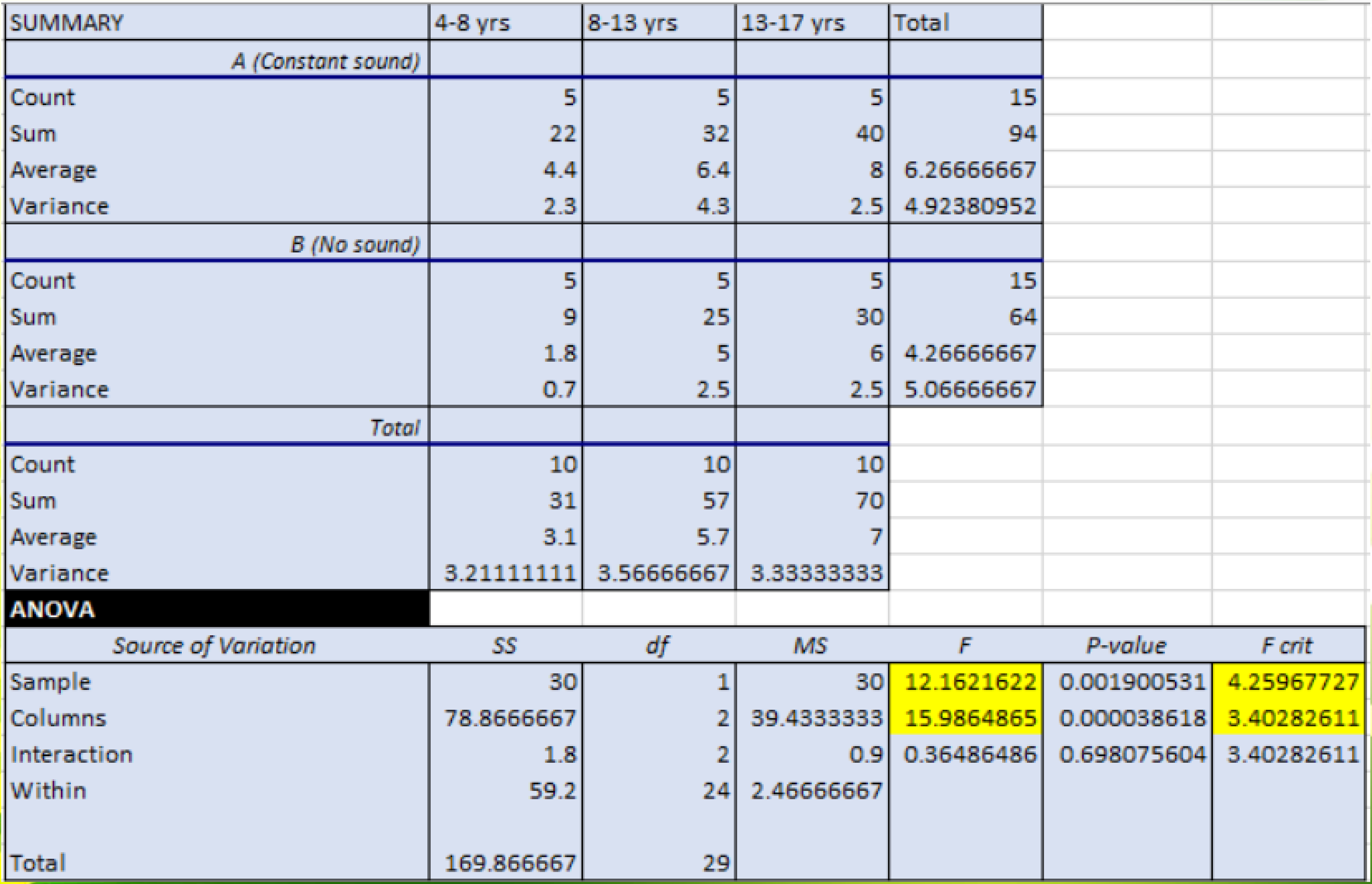
- Since more than one source of variation (main effects and interaction effects) → there will be more than one F-statistic.
- F-statistic for main and interaction effect:

  F1 = 12.16; **F2** = 15.98; **F12** = 0.36

- F-critical values from F-table (p = 0.05)

  Fcrit1 = 4.25; **Fcrit2** = 3.40; **Fcrit12** = 3.40

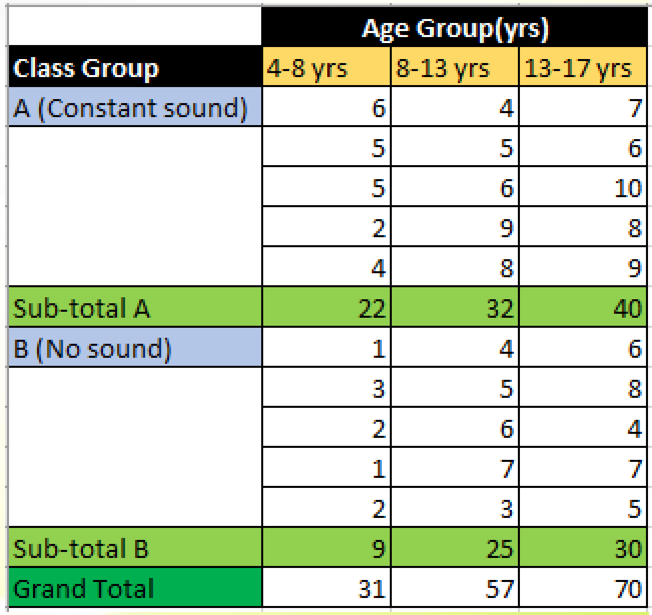
DATA ANALYSIS - ANOVA



SUMMARY	4-8 yrs	8-13 yrs	13-17 yrs	Total
A (Constant sound)				
Count	5	5	5	15
Sum	22	32	40	94
Average	4.4	6.4	8	6.26666667
Variance	2.3	4.3	2.5	4.92380952
B (No sound)				
Count	5	5	5	15
Sum	9	25	30	64
Average	1.8	5	6	4.26666667
Variance	0.7	2.5	2.5	5.06666667
Total				
Count	10	10	10	
Sum	31	57	70	
Average	3.1	5.7	7	
Variance	3.21111111	3.56666667	3.33333333	

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Sample	30	1	30	12.1621622	0.001900531	4.25967727
Columns	78.8666667	2	39.4333333	15.9864865	0.000038618	3.40282611
Interaction	1.8	2	0.9	0.36486486	0.698075604	3.40282611
Within	59.2	24	2.46666667			
Total	169.866667	29				



	Age Group(yrs)		
Class Group	4-8 yrs	8-13 yrs	13-17 yrs
A (Constant sound)	6	4	7
	5	5	6
	5	6	10
	2	9	8
	4	8	9
Sub-total A	22	32	40
B (No sound)	1	4	6
	3	5	8
	2	6	4
	1	7	7
	2	3	5
Sub-total B	9	25	30
Grand Total	31	57	70

- F-value for sample and column, i.e. factor 1 (music) and factor 2 (age) respectively.
- **F1 & F2 > their F-critical values → reject null hypothesis for factors.**
- F-value for interaction effect < F-critical value → Accept 3rd null hypothesis (music and age did NOT have any combined effect on population.)

DATA ANALYSIS - MULTIVARIATE

- **Multivariate:** involving multiple dependent variables resulting in one outcome.
- **Multivariate analysis** (**MVA**): Statistical procedure for analysis of data involving more than one type of measurement or observation.
- MVA techniques :
  - Multiple linear/nonlinear regression
  - Multiple linear correlation
  - Multivariate Analysis of Variance (MANOVA)
  - Interdependent analysis
  - Discriminant analysis
  - Classification and cluster analysis
  - Principal component analysis (PCA)
  - Factor analysis
  - Canonical correlation analysis

DATA ANALYSIS - MULTIVARIATE

Dependence methods

- Used when one or some variables are dependent on others (cause and effect)
- Values of two or more independent variables used to describe/predict value of another dependent variable.
  - dependent variable “weight” might be predicted by independent variables such as “height” and “age.”
- Focus on Effect of certain variables on others.

Interdependence methods

- Variables cannot be classified as either dependent or independent.
- Used to understand the structural makeup and underlying patterns within a dataset.
- Seek to give meaning to a set of variables or group them together in meaningful ways.
- Focus on Structure of the dataset.

DATA ANALYSIS - MULTIVARIATE

<u>**Objective of multivariate analysis**</u>

- **Data reduction or structural simplification**: helps data to get simplified as possible without sacrificing valuable information; making interpretation easier.
- **Sorting and grouping**: For multiple variables, Groups of "similar" objects or variables are created, based upon measured characteristics.
- **Investigation of dependence among variables**: Are all the variables mutually independent or are one or more variables dependent on others?
- **Prediction Relationships between variables**: must be determined for the purpose of predicting values of one or more variables based on observations on other variables.
- **Hypothesis construction and testing**. Specific statistical hypotheses, formulated in terms of parameters of multivariate populations, are tested → to validate assumptions or to reinforce prior convictions.