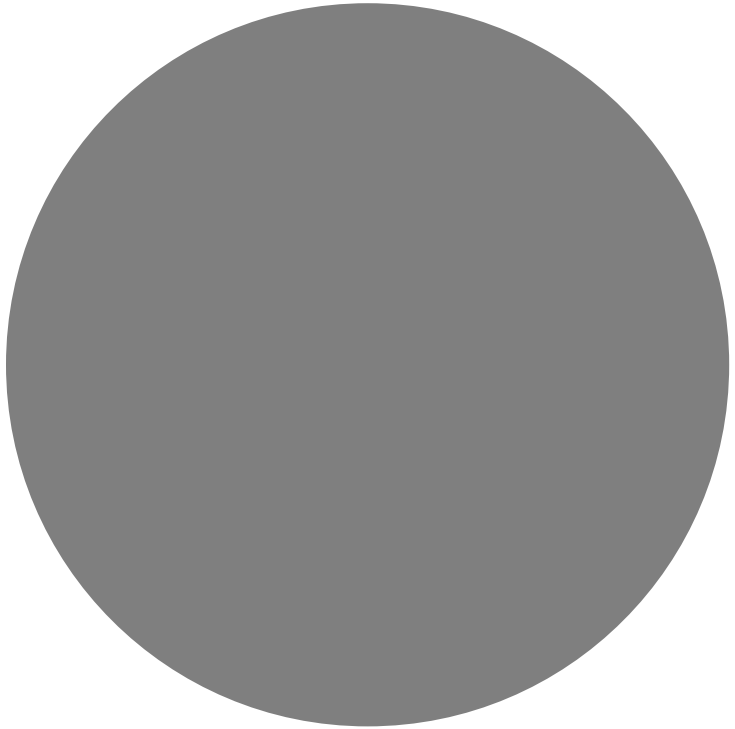
Mohamed Noordeen Alaudeen



Senior Data Scientist - Logitech

Application of ANOVA

- The purpose of ANOVA (Analysis of Variance) is to test for significant differences between means of different groups.
- **A pharmaceutical company tested 3 formulations of a migraine relief drug. 27 volunteers were randomly grouped in 3 groups. Each group was given a different drug formulation. The participants took the drug when they had the next migraine attack and recorded the pain on a scale of 1 to 10, 1 being no pain and 10 being extreme pain 30 minutes after taking the medicine.**

- **We want to understand if the differences are due to within group differences or between group differences.**

Application of ANOVA

Group 1			Group 2			Group 3		
3	4	3	3	5	7	5	5	5
2	5	5	6	7	6	6	5	7
4	3	3	4	4	8	7	6	6

$\bar{X}_1 = 3.56$ $\bar{X}_2 = 5.56$ $\bar{X}_3 = 5.78$

$$\bar{\bar{X}} = \frac{134}{27} = 4.96$$

$Total\ Sum\ of\ Squares, SST$

$= (2 - 4.96)^2 + 5 * (3 - 4.96)^2 + 4 * (4 - 4.96)^2 + 7 * (5 - 4.96)^2 + 5 * (6 - 4.96)^2$

$+ 4 * (7 - 4.96)^2 + (8 - 4.96)^2 = \mathbf{62.96}$

Application of ANOVA

When there are *m* groups and *n* members in each group, the degrees of freedom are $\boldsymbol{mn - 1}$, since we can calculate one member knowing the overall mean.

How much of this variation is coming from within the groups and how much from between the groups?

Application of ANOVA

Group 1			Group 2			Group 3		
3	4	3	3	5	7	5	5	5
2	5	5	6	7	6	6	5	7
4	3	3	4	4	8	7	6	6

$\bar{X}_1 = 3.56 \quad \bar{X}_2 = 5.56 \quad \bar{X}_3 = 5.78 \quad \bar{\bar{X}} = \frac{134}{27} = 4.96$

$Total\ Sum\ of\ Squares\ Within, SSW$
$= (2-3.56)^2 + 4*(3-3.56)^2 + 2*(4-3.56)^2 + 2*(5-3.56)^2 + (3-5.56)^2 + 2*(4-5.56)^2$
$+ (5-5.56)^2 + 2*(6-5.56)^2 + 2*(7-5.56)^2 + (8-5.56)^2 + 4*(5-5.78)^2 + 3*(6-5.78)^2$
$+ 2*(7-5.78)^2 = \mathbf{36.00}$

When there are *m* groups and *n* members in each group, the degrees of freedom are $\boldsymbol{m(n-1)}$, since we can calculate one member knowing the group mean.

$Total\ Sum\ of\ Squares\ Between, SSB$
$= 9*(3.56-4.96)^2 + 9*(5.56-4.96)^2 + 9*(5.78-4.96)^2 = \mathbf{26.96}$

When there are *m* groups, the degrees of freedom are $\boldsymbol{m-1}$.

SST = SSW + SSB

Also, for degrees of freedom, $\boldsymbol{mn-1 = m(n-1) + (m-1)}$

Application of ANOVA

- What is the null hypothesis?
- The population means of the 3 groups from which the samples were taken have the same mean, i.e., the drug formulations do not have an impact on relieving migraine headache. $\mu 1 = \mu 2 = \mu 3$. Let us also have a significance level, $\alpha = 0.10$.
- What is the alternate hypothesis?
  The drug formulations have an impact on migraine pain relief.

Application of ANOVA

$$F - statistic = \frac{\frac{SSB}{df_{SSB}}}{\frac{SSW}{df_{SSW}}} = \frac{\frac{26.96}{2}}{\frac{36}{24}} = 8.9876$$

If numerator is much bigger than the denominator, it means variation **between** means has bigger impact than variation **within**, thus rejecting the null hypothesis.

F Table for $\alpha = 0.10$

\	df_1=1	2	3	4	5	6	7	8	9	10	12	
df_2=1	39.86346	49.50000	53.59324	55.83296	57.24008	58.20442	58.90595	59.43898	59.85759	60.19498	60.70521	61
2	8.52632	9.00000	9.16179	9.24342	9.29263	9.32553	9.34908	9.36677	9.38054	9.39157	9.40813	9
3	5.53832	5.46238	5.39077	5.34264	5.30916	5.28473	5.26619	5.25167	5.24000	5.23041	5.21562	5
4	4.54477	4.32456	4.19086	4.10725	4.05058	4.00975	3.97897	3.95494	3.93567	3.91988	3.89553	3
5	4.06042	3.77972	3.61948	3.52020	3.45298	3.40451	3.36790	3.33928	3.31628	3.29740	3.26824	3
6	3.77595	3.46330	3.28876	3.18076	3.10751	3.05455	3.01446	2.98304	2.95774	2.93693	2.90472	2
7	3.58943	3.25744	3.07407	2.96053	2.88334	2.82739	2.78493	2.75158	2.72468	2.70251	2.66811	2
8	3.45792	3.11312	2.92380	2.80643	2.72645	2.66833	2.62413	2.58935	2.56124	2.53804	2.50196	2
9	3.36030	3.00645	2.81286	2.69268	2.61061	2.55086	2.50531	2.46941	2.44034	2.41632	2.37888	2
10	3.28502	2.92447	2.72767	2.60534	2.52164	2.46058	2.41397	2.37715	2.34731	2.32260	2.28405	2
11	3.22520	2.85951	2.66023	2.53619	2.45118	2.38907	2.34157	2.30400	2.27350	2.24823	2.20873	2
12	3.17655	2.80680	2.60552	2.48010	2.39402	2.33102	2.28278	2.24457	2.21352	2.18776	2.14744	2
13	3.13621	2.76317	2.56027	2.43371	2.34672	2.28298	2.23410	2.19535	2.16382	2.13763	2.09659	2
14	3.10221	2.72647	2.52222	2.39469	2.30694	2.24256	2.19313	2.15390	2.12195	2.09540	2.05371	2
15	3.07319	2.69517	2.48979	2.36143	2.27302	2.20808	2.15818	2.11853	2.08621	2.05932	2.01707	1
16	3.04811	2.66817	2.46181	2.33274	2.24376	2.17833	2.12800	2.08798	2.05533	2.02815	1.98539	1
17	3.02623	2.64464	2.43743	2.30775	2.21825	2.15239	2.10169	2.06134	2.02839	2.00094	1.95772	1
18	3.00698	2.62395	2.41601	2.28577	2.19583	2.12958	2.07854	2.03789	2.00467	1.97698	1.93334	1
19	2.98990	2.60561	2.39702	2.26630	2.17596	2.10936	2.05802	2.01710	1.98364	1.95573	1.91170	1
20	2.97465	2.58925	2.38009	2.24893	2.15823	2.09132	2.03970	1.99853	1.96485	1.93674	1.89236	1
21	2.96096	2.57457	2.36489	2.23334	2.14231	2.07512	2.02325	1.98186	1.94797	1.91967	1.87497	1
22	2.94858	2.56131	2.35117	2.21927	2.12794	2.06050	2.00840	1.96680	1.93273	1.90425	1.85925	1
23	2.93736	2.54929	2.33873	2.20651	2.11491	2.04723	1.99492	1.95312	1.91888	1.89025	1.84497	1
24	2.92712	2.53833	2.32739	2.19488	2.10303	2.03513	1.98263	1.94066	1.90625	1.87748	1.83194	1
25	2.91774	2.52831	2.31702	2.18424	2.09216	2.02406	1.97138	1.92925	1.89469	1.86578	1.82000	1

The *df* are 2 for numerator and 24 for denominator.

F_c, the critical F-statistic, therefore, is 2.53833. 8.9876 is way higher than this and hence we reject the null hypothesis. That means the drug formulations do have an impact on migraine pain relief.

Application of ANOVA

- **STOCK MARKET EXAMPLE**
- A stock analyst randomly selected 8 stocks from each of 3 industries, viz., Financial, Energy and Utilities. She compiled the 5- year rate of return for each stock.
- The analyst wants to know if, at 0.05 Significance Level, there is a difference in the rate of return for any of the industries.

Application of ANOVA

STOCK MARKET EXAMPLE

	5-Year Rates of Return		
	Financial	Energy	Utilities
	10.76	12.72	11.88
	15.05	13.91	5.86
	17.01	6.43	13.46
	5.07	11.19	9.9
	19.5	18.79	3.95
	8.16	20.73	3.44
	10.38	9.6	7.11
	6.75	17.4	15.7
xbar	**11.585**	**13.846**	**8.913**
s	**5.124**	**4.867**	**4.530**

- What is the null hypothesis?
- $\mu 1 = \mu 2 = \mu 3$. $\alpha = 0.05$
  All 3 industries have the same average rate of return.
- What is the alternate hypothesis?
- At least one of the industries has a different rate of return than the others.

Application of ANOVA

Anova: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
Financial	8	92.68	11.585	26.2528
Energy	8	110.77	13.8463	23.6879
Utilities	8	71.3	8.9125	20.5247

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	97.593	2	48.7965	2.07747	0.1502	3.4668
Within Groups	493.26	21	23.4885			
Total	590.85	23				

Thought Process on When to Use a Particular Test

What do you want to do?

- – Description – Summary statistics, Various plots, Correlations
- – Prediction – Linear regression, Logistic regression
- – Intervention (differences between groups) – *t*-test, Chi-square, ANOVA

Is the dependent variable Categorical or Numerical?

- – Nominal – Chi-square, Logistic regression
- – Ordinal – Chi-Square
- – Dichotomous – Logistic regression
- – Numerical – *t*-test, ANOVA, Correlation, Multiple regression