ANOVA

ANalysis Of VAriance

ANOVA

The purpose of ANOVA (Analysis of Variance) is to test for significant differences between means of different groups.

Consider 3 empty decks and 3 cards are placed into the each deck with their values. It is understand if the differences are due to within group differences or between group differences.

Group 1	Group 2	Group 3
6	10	6
4	6	8
2	8	4
$\bar{X}_1 = 4$	$\bar{X}_2 = 8$	$\bar{X}_3 = 6$

$$\bar{X} = \frac{6+4+2+10+6+8+6+8+4}{9} = 6$$

Total sum of square, SST

$= (6-4)^2 + (4-4)^2 + (2-4)^2 + (10-8)^2 + (6-8)^2 + (8-8)^2 + (6-6)^2 + (8-6)^2 + (8-6)^2$

$= 24$

When there are ***m*** *deckss* and ***n*** cards in each group, the degrees of freedom are 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?

Group 1	Group 2	Group 3
6	10	6
4	6	8
2	8	4
$\bar{X}_1 = 4$	$\bar{X}_2 = 8$	$\bar{X}_3 = 6$

$$\bar{X} = \frac{6+4+2+10+6+8+6+8+4}{9} = 6$$

Total sum of square Within, SSW

$$= (6-4)^2 + (4-4)^2 + (2-4)^2 + (10-8)^2 + (6-8)^2 + (8-8)^2 + (6-6)^2 + (8-6)^2 + (8-6)^2$$

$$= 24$$

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

Total sum of square between, SSB = $3(4-6)^2 + 3(8-6)^2 + 3(6-6)^2 = 24$

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

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

Group 1	Group 2	Group 3
6	10	6
4	6	8
2	8	4
$\bar{X}_1 = 4$	$\bar{X}_2 = 8$	$\bar{X}_3 = 6$

$$\bar{X} = \frac{6+4+2+10+6+8+6+8+4}{9} = 6$$

Given that mean of deck 2 is highest and that of deck 1 lowest, can we conclude that the cards placed in deck 2 had a larger impact or is it just variation within the deck?

Let us have a null hypothesis that the population means of the 3 decks from which the samples were taken have the same mean, i.e., the cards do not have an impact on the performance in the game. $\mu 1 = \mu 2 = \mu 3$. Let us also have a significance level, $\alpha = 0.10$.

- What is the alternate hypothesis?
- The cards have an impact on performance.

Group 1	Group 2	Group 3
6	10	6
4	6	8
2	8	4
$\bar{X}_1 = 4$	$\bar{X}_2 = 8$	$\bar{X}_3 = 6$

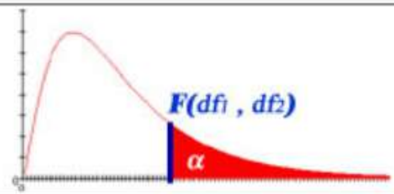
$$\bar{X} = \frac{6+4+2+10+6+8+6+8+4}{9} = 6$$

The test statistic used is F-statistic.

$$F - statistic = \frac{\frac{SSB}{df_{ssb}}}{\frac{SSW}{df_{ssw}}} = \frac{\frac{24}{2}}{\frac{24}{6}} = 3$$

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 α = 0.10



\	df$_1$=1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60
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	61.22034	61.74029	62.00205	62.26497	62.52905	62.79428
2	8.52632	9.00000	9.16179	9.24342	9.29263	9.32553	9.34908	9.36677	9.38054	9.39157	9.40813	9.42471	9.44131	9.44962	9.45793	9.46624	9.47456
3	5.53832	5.46238	5.39077	5.34264	5.30916	5.28473	5.26619	5.25167	5.24000	5.23041	5.21562	5.20031	5.18448	5.17636	5.16811	5.15972	5.15119
4	4.54477	4.32456	4.19086	4.10725	4.05058	4.00975	3.97897	3.95494	3.93567	3.91988	3.89553	3.87036	3.84434	3.83099	3.81742	3.80361	3.78957
5	4.06042	3.77972	3.61948	3.52020	3.45298	3.40451	3.36790	3.33928	3.31628	3.29740	3.26824	3.23801	3.20665	3.19052	3.17408	3.15732	3.14023
6	3.77595	3.46330	3.28876	3.18076	3.10751	3.05455	3.01446	2.98304	2.95774	2.93693	2.90472	2.87122	2.83634	2.81834	2.79996	2.78117	2.76195
7	3.58943	3.25744	3.07407	2.96053	2.88334	2.82739	2.78493	2.75158	2.72468	2.70251	2.66811	2.63223	2.59473	2.57533	2.55546	2.53510	2.51422

The df are 2 for numerator and 6 for denominator.

F_c, the critical F-statistic, therefore, is 3.46330. 12 is way higher than this and hence we reject the null hypothesis. That means the pills do have an impact on the performance.

ANOTHER EXAMPLE

A IOT device developer is testing 3 different designs of Lift Alarms. He picks 3 different locations in a building to install each device. The maximum load (KG) is measured for 9 consecutive days in each location.

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

Model 1			Model 2			Model 3		
253	254	253	253	255	257	255	255	255
252	255	255	256	257	256	256	255	257
254	253	253	254	254	258	257	256	256
$\bar{X}_1 = 253.56\,MW$			$\overline{X_2} = 255.56\,MW$			$\overline{X_3} = 255.78\,MW$		

$$\bar{X} = \frac{134}{27} = 254.96\,KG$$

Total Sum of Squares, SST = 62.93

Total Sum of Squares Within, SSW = 36.00

Total Sum of Squares Between, SSB = 26.93

What is the null hypothesis?

All 3 locations from which the samples were taken have the same population-mean, i.e., the devices does not have any impact with the power consumption. That is

$$\mu_1 = \mu_2 = \mu_3$$

Let us also specify a significance level, $\alpha = 0.10$.

What is the alternate hypothesis?
The lifts does impact the power output.

Compute the statistics

$$F\ -statistic = \frac{\frac{SSB}{df_{SSB}}}{\frac{SSW}{df_{SSW}}} = \frac{\frac{26.93}{2}}{\frac{36}{24}} = 8.97$$

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 α = 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.
Fc, the critical F-statistic, therefore, is 2.53833.

Our F=8.97 is way higher than this and hence we reject the null hypothesis. That means the IOT device does have an impact on the power production.

Anova: Single Factor

SUMMARY

Groups	Count	Sum	Average	Variance
Group1	9	2282	253.5556	1.027778
Group2	9	2300	255.5556	2.777778
Group3	9	2302	255.7778	0.694444

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	26.93	2	13.4815	8.987654	0.0012	2.5383
Within Groups	36	24	1.5			
Total	62.93	26				

Reference

- Head First Statistics