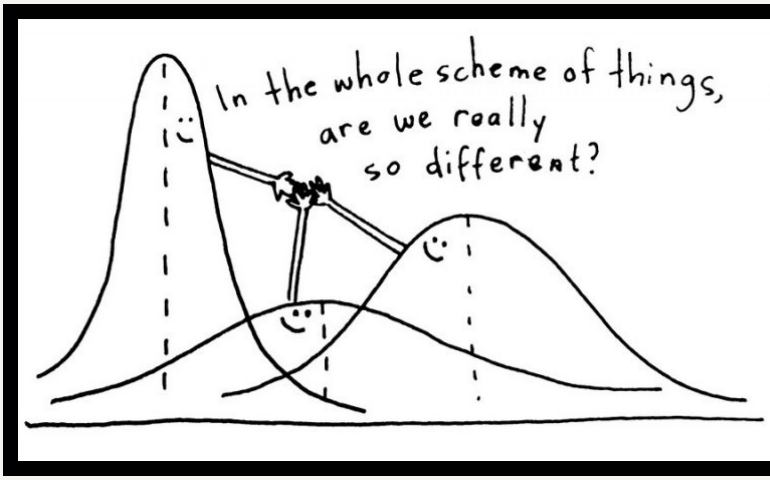
ANOVA

ANalysis Of VAriance

Def:

ANOVA or analysis of Variance is a group of statistical models to test if there exists a significant difference between means. It tests whether the means of various groups are equal or not. In ANOVA, the variance observed in a particular variable is partitioned into different components based on the sources of variation.



Types of ANOVA:

- One-Way ANOVA
- Two-Way ANOVA
- N-Way ANOVA

One-Way ANOVA

A one-way ANOVA has just one independent variable. For example, difference in IQ can be assessed by Country, and County can have 2, 20, or more different categories to compare.

Two-Way ANOVA

A two-way ANOVA refers to an ANOVA using two independent variables. Expanding the example above, a 2-way ANOVA can examine differences in IQ scores (the dependent variable) by Country (independent variable 1) and Gender (independent variable 2). Two-way ANOVA can be used to examine the interaction between the two independent variables. Interactions indicate that differences are not uniform across all categories of the independent variables. For example, females may have higher IQ scores overall compared to males, but this difference could be greater (or less) in European countries compared to North American countries. Two-way ANOVAs are also called factorial ANOVAs.

N-Way ANOVA

A researcher can also use more than two independent variables, and this is an n-way ANOVA (with n being the number of independent variables you have). For example, potential differences in IQ scores can be examined by Country, Gender, Age group, Ethnicity, etc, simultaneously.

- When we have two samples, t-test and ANOVA give the same results, but using a t-test would not be reliable in cases where there are more than two samples to be compared.
- In such cases, ANOVA is most effective to compare the means.

Assumptions:

(i) Subjects are chosen via a simple random sample.

(ii) Within each group/population, the response variable is normally distributed.

(iii) While the population means may be different from one group to the next, the population standard deviation is the same for all groups

1 Hypothesis

The hypotheses of interest in an ANOVA are as follows:

- H0: μ1 = μ2 = μ3 ... = μk
  All the means are equal.
- H1: At least one of the means is not equal.

2 Calculation:

In one way ANOVA the total variability is split into two sources:

1. Variability between group means

$$SSG := \sum_{i=1}^{k} n_i(\bar{x}_i - \bar{x})^2$$

2. Variability within groups means

$$SSE := \sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)^2 = \sum_{i=1}^{k}(n_i - 1)s_i^2.$$

One-Way ANOVA table:

Source	SS	df	MS	F
Model/Group	SSG	$k-1$	$\text{MSG} = \frac{SSG}{k-1}$	$\frac{MSG}{MSE}$
Residual/Error	SSE	$n-k$	$\text{MSE} = \frac{SSE}{n-k}$	
Total	SST	$n-1$		

- **SSG** = sum of squares between Groups (or) Treatments (or) Samples.
- **SSE** or **RSS** = Residual (or) Error Sum of Squares.
- **SST** = Total Sum of Squares.
- **Degrees of freedom** can be defined as the minimum number of independent coordinates that can specify the position of the system completely.
- **MS** = **Mean Square** = SS/df
  This is like a standard deviation. Its numerator was a sum of squared deviations , and it was divided by the appropriate number of degrees of freedom.
- The **F statistic** = MSG/MSE
  F = variation between group means / variation within the groups
  If we're hoping to show that the means are different, it's good when the within-group variance is low.

Two-Way ANOVA table:

Source	df	SS	MS	F	p-value
A	$I-1$	SSA	MSA	MSA/MSE	
B	$J-1$	SSB	MSB	MSB/MSE	
A × B	$(I-1)(J-1)$	SSAB	MSAB	MSAB/MSE	
Error	$n-IJ$	SSE	MSE		
Total	$n-1$	SST			

- This time we are dividing the variation into four components:
- 1. the variation explained by factor A
  2. the variation explained by factor B
  3. the variation explained by the interaction of A and B
  4. the variation explained by randomness
- **Similarly in N-Way ANOVA '(2^N)-N-1' no. of interactions will be present along with 'N' no. of main factors.**
- **P value**: The P value, or calculated probability, is the probability of finding the observed, or more extreme, results when the null hypothesis of a study question is true.
- P value can be calculated from the f statistic using the F table or any statistical software.
- Statistically significant if P < 0.05 and statistically highly significant as P < 0.001 (less than one in a thousand chance of being wrong) i.e., accept Alternate hypothesis and conclude that there is a significant difference among given means of Groups (or) Treatments (or) Samples.

3 Conclusion:

- Reject Ho if F > Critical value
  Critical value = F(n-k,n-1) -> derived from F table values.
- Else accept Ho.
- Or we can use the p- value for the decision
  Reject Ho if p-value < **α**
- Else accept Ho
- The significance level, also denoted as alpha or "**α**", is the probability of rejecting the null hypothesis when it is true. For example, a significance level of 0.05 indicates a 5% risk of concluding that a difference exists when there is no actual difference.

Examples:

- Ex 1): Suppose a Phone vendor has 4 top selling models, By using ANOVA we can check whether a single model has a higher sale.
- Ex 2): To check whether a certain fertilizer among two or many fertilizers gives a better yield.
- Ex 3): To find out which mode of teaching(outdoor or indoor or electronic etc) has better impact on students marks.

Case Study

- A research study was conducted to examine the impact of eating a high protein breakfast on adolescents' performance during a physical education physical fitness test. Half of the subjects received a high protein breakfast and half were given a low protein breakfast. All of the adolescents, both male and female, were given a fitness test with high scores representing better performance. Test scores are recorded below.

Group	High Protein	Low Protein
Males	10 7 9 6 8 Mean=8.0	5 4 7 4 5 Mean=5.0
Females	5 4 6 3 2 Mean=4.0	3 4 5 1 2 Mean=3.0

Factor 1: Protein level (High, Low)
Factor 2: Gender (Males, Females)

Hence we use two way ANOVA.

Case Study

1) Hypothesis:

Null Hypothesis H0:

a) There is no significant difference in fitness level between High Protein and Low Protein diet i.e. High Protein diet does not result in better fitness level than Low Protein

b) There is no significant difference in fitness level between Males and Females i.e. Males do not have better fitness than Females.

Alternate Hypothesis H1: There is a difference in fitness level between High Protein and Low Protein diet (and/or) Males and Females.

2) Calculation:

ANOVA table

Source	SS	df	MS	F	P
Protein Level	20	1	20	8.89	0.008811
Gender	45	1	45	20	0.000385
Protein Level x Gender	5	1	5	2.22	0.157441
Within	36	16	2.25		
Total	106	19			

Case Study

Source	F	F Criticial	P	α
Protein Level	8.89	> 3.0481	0.008811	< .01 Therefore reject H0
Gender	20	> 3.0481	0.000385	< .01 Therefore reject H0
Protein Level x Gender	2.22	< 3.04811	0.157441	> .01 Therefore accept H0

F critical value for df1=1 and df2 = 16 can be derived from the f distribution table values.
Which is 3.04811

3) Conclusion:

> There appears to be significant main effects for both protein level (F=8.89 (1,16), p<.01) and gender (F=20.00 (1,16), p<.01). There was not a significant interaction effect (F=2.22 (1,16), not significant).

> Based on this data, it appears that a high protein diet results in a better fitness test score. Additionally, young men seem to have a significantly higher fitness test score than women.

SUBJECT ANOVA

AUTHOR **SAMUEL**