ANALYSIS OF VARIANCE (ANOVA)

One-Way ANOVA

Introduction

- *ANOVA is a tool to test equality of more than two means simultaneously.*

OR

- *It is a tool for analyzing how the mean value of a quantitative response variable is related to one or more categorical explanatory factors.*

- F-test is used to determine the significance difference among three or means.
- It was developed by Sir R. A. Fisher, an English Statistician.

Examples

- To determine significant differences for mean time of solving a computer problem by four groups of students, using C, C #, C++ and Python.

- To determine significant differences for Software Effort among different phases of SDLC.

- To determine significant differences for software metrics such as: Defect metric, process metric, KSLOC, FPs, etc.

- To determine interaction effect of testing technique, software type, expertise level etc.

F-test (Definition – I)

- If χ_u^2 and χ_υ^2 are two independent chi-square random variables with u and υ degrees of freedom, respectively, then the ratio:

$$F_{u,\upsilon} = \frac{\chi_u^2/u}{\chi_\upsilon^2/\upsilon}$$

follows the F distribution with u numerator degrees of freedom and υ denominator degrees of freedom.

F-test (Definition – II)

- Consider two independent normal populations with common variance σ^2 . If random samples of sizes n_1 & n_2 are drawn from these populations then:

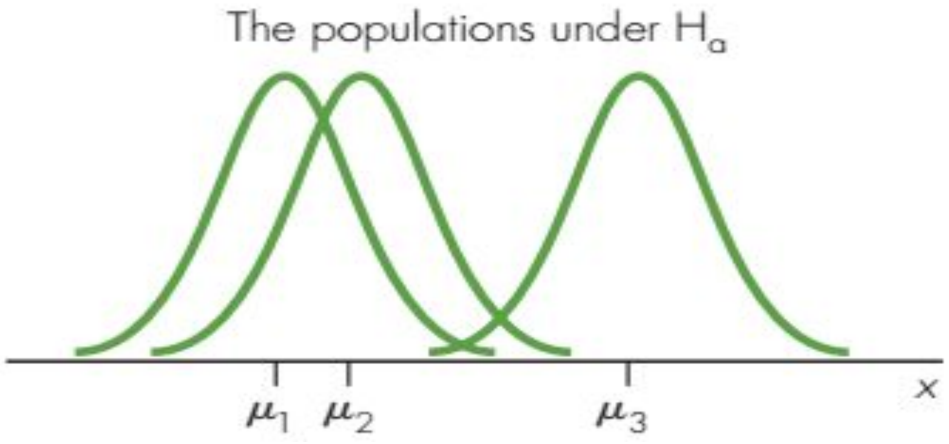
$$\frac{S_1^2}{S_2^2} \sim F_{n_1-1,\, n_2-1}$$

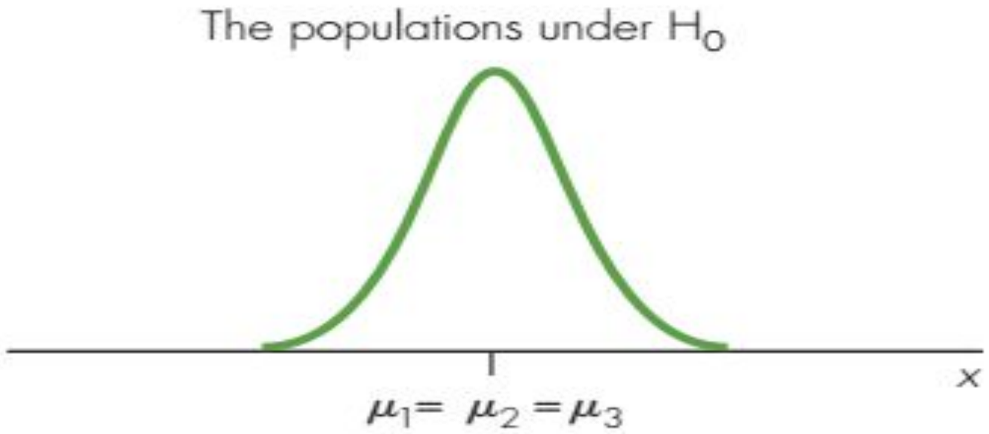
where S_1^2 and S_2^2 are two sample variances.

*(**Note:** the larger of the variances is placed in the numerator of the F formula)*

Assumptions & Conditions for F-test

- The samples are independent random samples.
- The distribution of the response variable is a normal curve within each population.
- The different populations may have different means.
- All populations have the same standard deviation, σ.





Sensitivity of F-statistic

- The *F*-statistic is sensitive to differences among a set of sample means.
  - The greater the variation among the sample means, the larger is the value of the test statistic.
  - The smaller the variation among the observed means, the smaller the value of the test statistic.

F-distribution

- If **x** is an F random variable with u numerator and υ denominator degrees of freedom, then the PDF of **x** is:

$$h(x) = \frac{\Gamma\left(\frac{u + v}{2}\right)\left(\frac{u}{v}\right)^{u/2} x^{(u/2)-1}}{\Gamma\left(\frac{u}{x}\right)\Gamma\left(\frac{v}{2}\right)\left[\left(\frac{u}{v}\right)x + 1\right]^{(u+v)/2}} \qquad 0 < x < \infty$$

Why not t-test ?

Why t-test should not be done while comparing several means taking two at a time?

- when one is comparing two means at a time, the rest of the means under study are ignored.
- the more means there are to compare, the more *t* tests are needed.
  - For the comparison of 5 means two at a time, 10 tests are required.
  - for the comparison of 10 means two at a time, 45 tests are required.
- the more *t* tests that are conducted, the greater is the likelihood of getting significant differences by chance alone.

Why the procedure is called ANOVA?

- The name **Analysis of Variance** is derived from a partitioning of total variability into its component parts.

Measure of Total Variability

$$\boxed{SS_T} = \boxed{SS_{\text{Treatments}}} + \boxed{SS_E}$$

Total corrected Sum of Squares

Sum of Squares b/w treatments

Sum of Sq. due to error (within treatments)

One-Way ANOVA

- Random samples of size n are selected from each of k populations. The k different populations are classified on the basis of a **single criterion** such as different treatments or groups.

- It is assumed that the k populations are independent and normally distributed with means $\mu_1, \mu_2, \ldots \mu_k$ and common variance σ^2.

Model for One-way ANOVA

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij},$$

- μ is the Grand Mean of all μ_i, $\mu = \frac{1}{k}\sum_{i=1}^{k} \mu_i,$

  α_i

- The ϵ_{ij}-term represents random error (within group variation).
- α_i is the effect of i^{th} treatment with constraint $\sum_{i=1}^{k} \alpha_i = 0.$

Summary procedure of ANOVA

H_0: $\mu_1 = \mu_2 = \cdots = \mu_k$

H_1: At least one mean is different from the others.

α: 0.05, 0.01, or 0.10

ANOVA Table

Source of Variation	Sum of Squares	d.f.	Mean Square	F_o
Between Treatment (groups)	$k\sum_{i=1}^{k}(\bar{y}_{i.} - \bar{y}_{..})^2$	k – 1	$MS_{Treatments}$	$Fo = MS_{Treatments} / MS_E$
Error (within treatment)	$SS_E = SS_T - SS_{treatment}$	N – k	MS_E	
Total	$\sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{..})^2$	N – 1		

Example # 1

- A researcher wishes to try three different techniques to lower the blood pressure of individuals diagnosed with high blood pressure. The subjects are randomly assigned to three groups; the first group takes medication, the second group exercises, and the third group follows a special diet. After four weeks, the reduction in each person's blood pressure is recorded. At $\alpha = 0.05$, test the claim that there is no difference among the means. The data follow.

Medication	Exercise	Diet
10	6	5
12	8	9
9	3	12
15	0	8
13	2	4

Solution (Example 01)

- Find Grand Mean (GM) as: $\overline{X}_{\text{GM}} = \frac{\Sigma X}{N} = \frac{10 + 12 + 9 + \cdot \cdot \cdot + 4}{15} = \frac{116}{15} = 7.73$

- Find b/w group variance as:

$$s_B^2 = \frac{\Sigma\, n_i\,(\overline{X}_i - \overline{X}_{\text{GM}})^2}{k-1}$$
$$= \frac{5(11.8 - 7.73)^2 + 5(3.8 - 7.73)^2 + 5(7.6 - 7.73)^2}{3-1}$$
$$= \frac{160.13}{2} = 80.07$$

- Find within group variance as:

$$s_W^2 = \frac{\Sigma\,(n_i - 1)s_i^2}{\Sigma\,(n_i - 1)}$$
$$= \frac{(5-1)(5.7) + (5-1)(10.2) + (5-1)(10.3)}{(5-1) + (5-1) + (5-1)}$$
$$= \frac{104.80}{12} = 8.73$$

- F-test as: $F = \frac{s_B^2}{s_W^2} = \frac{80.07}{8.73} = 9.17$

Solution (Example 01, Contd.)

S. V.	SS	d.f.	MS	F-ratio
Between	160.13	2	80.07	9.17
Within (error)	104.80	12	8.73	
Total	264.93	14		

- Decision: the decision is to reject H_o as : $9.17 > 3.89.$
  - There is enough evidence to reject the claim and conclude that at least one mean is different from the others.

Example # 02 [Plasma Etching Experiment]

Recall that the engineer is interested in determining if the RF power setting affects the etch rate, and she has run a completely randomized experiment with four levels of RF power and five replicates.

RF Power (W)	Observed Etch Rate (A/min)				
	1	2	3	4	5
160	575	542	530	539	570
180	565	593	590	579	610
200	600	651	610	637	629
220	725	700	715	685	710

Solution (Example 02)

$$SS_T = \sum_{i=1}^{4}\sum_{j=1}^{5} y_{ij}^2 - \frac{y_{..}^2}{N}$$
$$= (575)^2 + (542)^2 + \cdots + (710)^2 - \frac{(12{,}355)^2}{20}$$
$$= 72{,}209.75$$

$$SS_{\text{Treatments}} = \frac{1}{n}\sum_{i=1}^{4} y_{i.}^2 - \frac{y_{..}^2}{N}$$
$$= \frac{1}{5}[(2756)^2 + \cdots + (3535)^2] - \frac{(12{,}355)^2}{20}$$
$$= 66{,}870.55$$

$$SS_E = SS_T - SS_{\text{Treatments}}$$
$$= 72{,}209.75 - 66{,}870.55 = 5339.20$$

S. V	SS	d. f.	MS	F_o	Decision
RF Power	66870.55	3	22290.18	66.80	Rejecct H_o
Error	5339.20	16	333.70		
Total	72209.75	19			

Example # 03

Suppose in an industrial experiment that an engineer is interested in how the mean absorption of moisture in concrete varies among 5 different concrete aggregates. The samples are exposed to moisture for 48 hours. It is decided that 6 samples are to be tested for each aggregate, requiring a total of 30 samples to be tested. The data are recorded in Table 13.1.

Test the hypothesis $\mu_1 = \mu_2 = \cdots = \mu_5$ at the 0.05 level of significance for the data of Table 13.1 on absorption of moisture by various types of cement aggregates.

Table 13.1: Absorption of Moisture in Concrete Aggregates

Aggregate:	**1**	**2**	**3**	**4**	**5**	
	551	595	639	417	563	
	457	580	615	449	631	
	450	508	511	517	522	
	731	583	573	438	613	
	499	633	648	415	656	
	632	517	677	555	679	
Total	3320	3416	3663	2791	3664	16,854
Mean	553.33	569.33	610.50	465.17	610.67	561.80

Solution (Example 04)

H_0: $\mu_1 = \mu_2 = \cdots = \mu_5$,

H_1: At least two of the means are not equal.

$\alpha = 0.05$.

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	85356.4667	21339.1167	4.30	0.0088
Error	25	124020.3333	4960.8133		
Corrected Total	29	209376.8000			

Practice Questions

Q1) **13.1** Six different machines are being considered for use in manufacturing rubber seals. The machines are being compared with respect to tensile strength of the product. A random sample of four seals from each machine is used to determine whether the mean tensile strength varies from machine to machine. The following are the tensile-strength measurements in kilograms per square centimeter $\times 10^{-1}$:

Machine

1	**2**	**3**	**4**	**5**	**6**
17.5	16.4	20.3	14.6	17.5	18.3
16.9	19.2	15.7	16.7	19.2	16.2
15.8	17.7	17.8	20.8	16.5	17.5
18.6	15.4	18.9	18.9	20.5	20.1

Perform the analysis of variance at the 0.05 level of significance and indicate whether or not the mean tensile strengths differ significantly for the six machines.

Q2)

13.2 The data in the following table represent the number of hours of relief provided by five different brands of headache tablets administered to 25 subjects experiencing fevers of 38°C or more. Perform the analysis of variance and test the hypothesis at the 0.05 level of significance that the mean number of hours of relief provided by the tablets is the same for all five brands. Discuss the results.

Tablet

A	*B*	*C*	*D*	*E*
5.2	9.1	3.2	2.4	7.1
4.7	7.1	5.8	3.4	6.6
8.1	8.2	2.2	4.1	9.3
6.2	6.0	3.1	1.0	4.2
3.0	9.1	7.2	4.0	7.6