ANOVA (**An**alysis **o**f **Va**riance)

An ANOVA (“Analysis of Variance”) is a statistical technique that is used to determine whether there is a significant difference between the means of three or more independent groups.

The two most common types of ANOVAs are the one-way ANOVA and two-way ANOVA.

The **One-Way ANOVA** is used to determine how one factor impacts a response variable.

For example, we might want to know if three different studying techniques lead to different mean exam scores. To see if there is a statistically significant difference in mean exam scores, we can conduct a one-way ANOVA.

|  |  |  |
| --- | --- | --- |
| **Factors** |  | **Response Variable** |
| Studying Technique |  | Exam Score |

A **Two-Way ANOVA** is used to determine how two factors impact a response variable, and to determine whether there is an interaction between the two factors on the response variable.

For example, we might want to know how gender and how different levels of exercise impact average weight loss. We would conduct a two-way ANOVA to find out.

|  |  |  |
| --- | --- | --- |
| **Factors** |  | **Response Variable** |
| Level of Exercise |  | Weight Loss |
| Gender |  |  |

In ANOVA, Population sample must be

* Independent
* Normally distributed
* Equal variance
* Random sample

**How to perform One Way ANOVA Test?**

* **Null Hypothesis (Ho):** No significant differences between the means. µ1= µ2= µ3= …
* **Alternate Hypothesis (Hq):** There is significant differences between the means. µ1≠ µ2≠ µ3≠

|  |  |  |  |
| --- | --- | --- | --- |
| **Source of Variation** | **DF** | **MSS (Mean of Sum of Squares)** | **F** |
| Between the group | K-1 | MSSB = SSB / (K-1) | MSSB  MSSW |
| Within the group | N-K | MSSW = SSW / (N-K) |

Where,

K= Total no. of samples (columns)

N= Total no. of values in given samples

To calculate SSB (**S**um of **S**quare **B**etween groups)

To calculate SSW (**S**um of **S**quares **W**ithin groups)

**One way ANOVA Test.**

**Problem 1**

A pharmaceutical company conducts an experiment to test the effect of a new cholesterol medication. The company selects 15 subjects randomly from a larger population.

Each subject is randomly assigned to one of three treatment groups.

Within each treatment group, subjects receive a different dose of the new medication.

* In Group 1, subjects receive 0 mg/day;
* in Group 2, 50 mg/day; and
* in Group 3, 100 mg/day.

The treatment levels represent all the levels of interest to the experimenter, so this experiment used a fixed-effects model to select treatment levels for study.

After 30 days, doctors measure the cholesterol level of each subject. The results for all 15 subjects appear in the table below:

|  |  |  |
| --- | --- | --- |
| **DOSAGE** | | |
| **Group 1**  **0 mg** | **Group 2**  **50 mg** | **Group 3**  **100 mg** |
| 210 | 210 | 180 |
| 240 | 240 | 210 |
| 270 | 240 | 210 |
| 270 | 270 | 210 |
| 300 | 270 | 240 |

**Solution:**

**Null hypothesis:** The null hypothesis states that the independent variable (dosage level) has no effect on the dependent variable (cholesterol level) in any treatment group.

If the null hypothesis is true, the mean score (i.e., mean cholesterol level) in each treatment group should equal the population mean.

**Alternative hypothesis:** The alternative hypothesis states that the independent variable has an effect on the dependent variable in at least one treatment group.

If the null hypothesis is false, at least one pair of mean scores should be unequal.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **DOSAGE** | | |
| **Group 1**  **0 mg** | **Group 2**  **50 mg** | **Group 3**  **100 mg** |
| 210 | 210 | 180 |
| 240 | 240 | 210 |
| 270 | 240 | 210 |
| 270 | 270 | 210 |
| 300 | 270 | 240 |
| **Total** | **1290** | **1230** | **1050** |
| **Mean** | **258** | **246** | **210** |

= (210+210+180+240+240+210+270+270+210+300+270+240) / 15 = 238

SSB = 5(258 – 238)2 + 5(246 – 238)2 + 5(210 – 238)2 = 2000+320+3920 = 6240

SSW = [(210-258)2 + (240-258)2 + (270-258)2 + (270-258)2 + (300-258)2] +

[(210-246)2 + (240-246)2 + (240-246)2 + (270-246)2 + (270-246)2] +

[(180-210)2 + (210-210)2 +(210-210)2 +(210-210)2 +(240-210)2] = **9000**

|  |  |  |  |
| --- | --- | --- | --- |
| **Source of Variation** | **DF** | **MSS (Mean of Sum of Squares)** | **F** |
| Between the group | K-1 = 3 - 1 = 2 | MSSB = SSB / (K-1)  = 6240 / 2  = 3120 | MSSB  MSSW  = 3120 / 750  **= 4.16** |
| Within the group | N-K = 15 - 3 = 12 | MSSW = SSW / (N-K)  = 9000 / 12  = 750 |

Therefore,

Calculated value = **4.16**

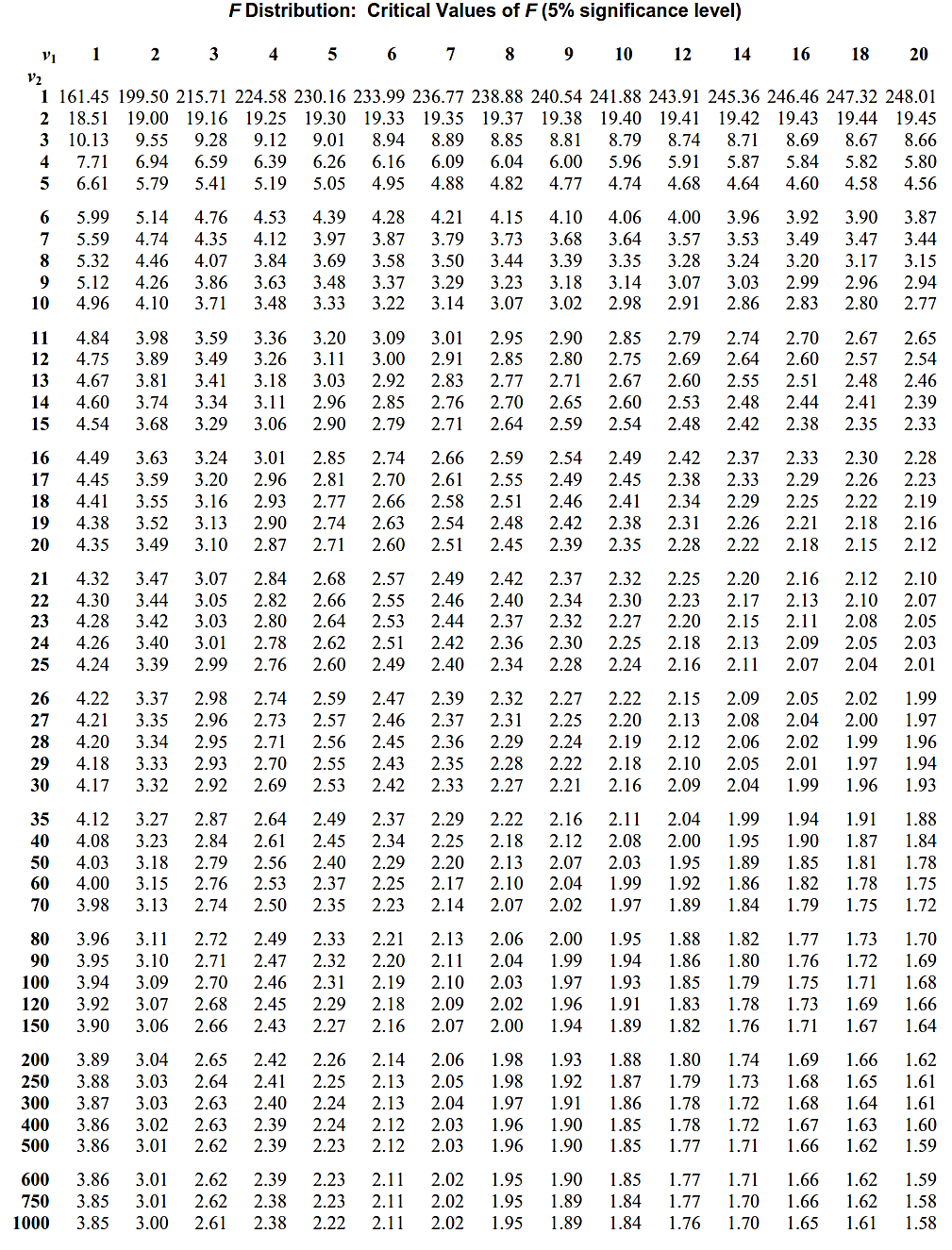
From table below for f(K-1, N-K) = f(2,12), see for 2 in v1 and 12 in v2 we get **3.89**

since

Calculated Value > Table Value

So,

**We reject the Null Hypothesis, and accepts the alternate hypothesis, which means cholesterol level is affected by dosage.**





**Two Way ANOVA Test**

A Two Way ANOVA is an extension of the One Way ANOVA. With a One Way, you have one independent variable affecting a dependent variable.

With a Two Way ANOVA, there are two independents.

Use a two-way ANOVA when you have one measurement variable (i.e. a quantitative variable) and two nominal variables. In other words, if your experiment has a quantitative outcome and you have two categorical explanatory variables, a two-way ANOVA is appropriate.

For example, you might want to find out if there is an interaction between income and gender for anxiety level at job interviews. The anxiety level is the outcome, or the variable that can be measured. Gender and Income are the two categorical variables. These categorical variables are also the independent variables, which are called factors in a Two Way ANOVA.

**Problem 2:**

The following table gives monthly sale (in thousands) of a certain firm in three states by its   
4 salesmen.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Salesmen** | | | |
| **States** | **X1** | **X2** | **X3** | **X4** |
| **Y1** | 6 | 5 | 3 | 8 |
| **Y2** | 8 | 9 | 6 | 5 |
| **Y3** | 10 | 7 | 8 | 7 |

Setup the ANOVA table and test whether there is any significant

1. difference between sales by the salesmen.
2. Difference between sales in the 3 states.

**Solution**:

**Null Hypothesis:** There is no significant difference in sales by salesmen as well as between states

**Alternate Hypothesis:** There is significant differences in sales between salesmen and between states.

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Salesmen** | | | | | **Total** | **nr = 4**  **(No of values in each row)** |  | | | |
| **States** | **X1** | **X2** | **X3** | **X4** | **(∑Y)** | | **X12** | **X22** | **X32** | **X42** |
| **Y1** | 6 | 5 | 3 | 8 | **22** | | 36 | 25 | 9 | 64 |
| **Y2** | 8 | 9 | 6 | 5 | **28** | | 64 | 81 | 36 | 25 |
| **Y3** | 10 | 7 | 8 | 7 | **32** | | 100 | 49 | 64 | 49 |
| **Total (∑X)** | **24** | **21** | **17** | **20** | **82** | | **(∑X12) 200** | **(∑X22)155** | **(∑X32)109** | **(∑X42)138** |
|  | **nc= 3**  **(No of values in each column)** | | | |  | |  |  |  |  |

N=12 (Total No of values) T = 6+5+3+8+8+9+6+5+10+7+8+7 =82 (sum of all values)

* **CF = 560.33**

TSS = ∑X12 + ∑X22+∑X32+∑X42 – CF = 200 +155 + 109 + 138 – 560.33 = 41.67

* **Total Sum of Squares (TSS) = 41.67**

**SSC (Sum of Squares of Columns)**

* **SSC = 8.37**

**SSR (Sum of Squares of Rows)**

* **SSR = 12.37**

**Residual Error (SSE)**

SSE = TSS – (SSC+SSR)

SSE = 41.67 – (8.37+12.37) = 20.93

* **SSE = 20.93**

**ANOVA Table Format**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Source of variation** | **Sum of Squares** | **DOF** | **Mean Sum of Squares** | **F - Ratio** |
| **Between Columns** | **SSC** | **Nc-1** |  | **For columns**  **or** |
| **Between Rows** | **SSR** | **Nr-1** |  |
| **Residual Error** | **SSE** | **(Nc-1)( Nr-1)** |  | **For Rows**  **or** |
| **Total** | **TSS** | **N-1** |  |

**\*Important Note: Always Numerator should be greater than denominator while calculating Fc and Fr**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Source of variation** | **Sum of Squares** | **DOF** | **Mean Sum of Squares** | **F - Ratio** |
| **Between Columns** | **SSC = 8.37** | **Nc-1 = 4-1 = 3** |  | **For columns\***  **1.254** |
| **Between Rows** | **SSR = 12.37** | **Nr-1 = 3-1 = 2** |  |
| **Residual (Error)** | **SSE = 20.93** | **(Nc-1)( Nr-1)**  **3\*2 = 6** |  | **For Rows\***  **= 1.77** |
| **Total** | **TSS = 41.67** | **N-1 = 11** |  |

**DOF for ‘Between columns’: (6,3) (DOF of numerator, DOF of denominator)**

*\*\* Whichever the value we took in numerator, the corresponding dof should be taken. In this case, to calculate Fc MSE is taken in the numerator. So, the corresponding dof is 6.*

**DOF for ‘Between Rows’: (2,6)**

Table value for Fc(6,3) at 5% level = 8.94 🡪 Calculated Value < Table Value

Table value for Fr(2,6) at 5% level = 5.14 🡪 Calculated Value < Table Value

**So, we accept the null hypothesis, which states that there is no significant differences between sales and salesmen as well as between states.**

**Comparison between T-test and ANOVA**

|  |  |  |
| --- | --- | --- |
| **Comparison variable** | **T-TEST** | **ANOVA** |
| **Definition** | t-test is statistical hypothesis test used to compare the means of two population groups. | ANOVA is an observable technique used to compare the means of more than two population groups. |
| **Feature** | t-test compares two sample sizes (n) both below 30. | ANOVA equates three or more such groups. |
| **Error** | t-test is less likely to commit an error. | ANOVA has more error risks. |
| **Example** | Sample from class A and B students have given a mathematics course may have different mean and standard deviation. | When one crop is being cultivated from various seed varieties. |
| **Test** | t-test can be performed in a double-sided or single-sided test. | ANOVA is one-sided test due to no negative variance. |
| **Population** | t-test is used when the population is less than 30. | ANOVA is used for huge population counts. |
| **Types** | Common types include one-sample t-test, two-sample t-test, and paired t-test. | The two types are one-way and two-way ANOVA. |
| **Value indicates** | If the t-score or t-value is small, the groups or samples are similar, whereas if the t-score is large, the groups or samples are different. | The higher the F value, there exist significant variation between sample or group means, and a low F-value indicates low variability. |