# Chapter 8:

## The analysis of variance (ANOVA)

November 30, 2021

The analysis of variance (ANOVA) is a hypothesis-testing technique used to test the claim that three or more populations (or treatment) means are equal by examining the variances of samples that are taken. This is an extension of the two independent samples t-test.

ANOVA is based on comparing the variance (or variation) between the data samples to variation within each particular sample. If the between variation is much larger than the within variation, the means of different samples will not be equal. If the between and within variations are approximately have the same size, then there will be no significant difference between sample means.

In the one and two-way ANOVA models, some of the names and what they refer to are listed as follows:

1. Response: the dependent variable (interval or ratio scale)
2. Factor(s): the independent variable(s) (factors): nominal or ordinal scale with more than 2 categories. In the One-way ANOVA: one factor is involved while in the two-way ANOVA: two factors are involved.
3. Levels: the possible values of a factor.
   Treatments: another name for levels in one-way ANOVA, but there will be a distinction between levels and treatments when we discuss two-way ANOVA.
   The term treatments derive from medicine, where the different treatments were the drugs or procedures being tested on patients, and agriculture, where the treatments were the different fertilizers or pesticides being tested on crops.
4. Unit: person, animal, piece of material, etc. that is subjected to treatment(s) and provides a response.

One-way ANOVA = one factor = one independent variable with two or more levels/conditions

In the one-way ANOVA model (Single Factor Analysis), the dependent (or response) variable is quantitative, but the independent (or factor) variable is qualitative. Conditions or Assumptions:

1. All populations involved follow a normal distribution.
2. All populations have the same variance (or standard deviation).
3. The samples are randomly selected and independent of one another.

Since ANOVA assumes the populations involved follow a normal distribution, ANOVA falls into a category of hypothesis tests known as parametric tests. If the populations involved did not follow a normal distribution, an ANOVA test could not be used to examine the equality of the sample means. Instead, one would have to use a non-parametric test (or distribution-free test), which is a more general form of hypothesis testing that does not rely on distributional assumptions.

## HYPOTHESIS TEST:

The null hypothesis for a one-way ANOVA always assumes the population means for the k samples drawn (one from each population) are equal. Hence, we may write the null hypothesis as:

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

This is equivalent to saying that the k treatments have no differential effect upon the value of the response. Since the null hypothesis assumes all the means are equal, we could reject the null hypothesis if only mean is not equal. Thus, the alternative hypothesis is:

$$H_1 : \text{At least one of the means different.}$$

The ANOVA doesn't test that one mean is less than another, only whether they're all equal or at least one is different.

## COMPUTING ONE WAY ANOVA:

We will assume the model

$$y_{ij} = \mu_i + e_{ij},$$

where

- $i = 1, \ldots, k$,
- $j = 1, \ldots, n_i$, and $n = \sum_{i=1}^{k} n_i$. $n_i$ is the number of units in the $i-$th treatment and $n$ is the total of observations.
- $y_{ij}$ denotes the value of the response variable in the $j-$th unit and $i-$th treatment.
- $\mu_i$ denotes the parameter of the unknown population mean for the $i-$th treatment.
- $e_{ij}$ is independent random errors with $N(0, \sigma^2)$.

| | Treatment | | | |
|---|---|---|---|---|
| | 1 | 2 | ... | k |
| | $y_{11}$ | $y_{21}$ | $\cdots$ | $y_{k1}$ |
| | $y_{12}$ | $\vdots$ | $\cdots$ | $y_{k2}$ |
| | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| | $y_{1n_1}$ | $y_{2n_2}$ | $\cdots$ | $y_{kn_k}$ |
| sample Mean | $\bar{y}_{1.}$ | $\bar{y}_{2.}$ | $\cdots$ | $\bar{y}_{k.}$ |
| sample Variance | $S_1^2$ | $S_2^2$ | $\cdots$ | $S_k^2$ |

To carry out ANOVA, the mean and standard deviation for each of the groups involved in the study must be calculated. The analysis of variance will then tell us three things:

1. Whether any of the specific groups differ from each other. There are more than one possible pairs of groups when you have more than two. This is provided by using a comparison technique.
2. Whether the differences are relatively big or small. Measures of explained variance will tell us this. The within-group variance is related to sampling error, subject differences, etc. It's like the variance we previously discussed. The between-group variance is due to the differences between the groups. If the means of the groups differ significantly, then there will be an associated between groups variance that is high.
3. Whether there are any significant differences of means among all of the groups provided by the F-ratio. The F-ratio is similar to the t-ratio, however, the F-ratio compares the variance between the groups with the variance within the groups:

$$F = \frac{(\text{variance between groups})}{(\text{variance within groups})}$$

If the F-ratio is small, then groups are probably not significantly different from each other. However, if it is big, then some (two or more) might be significantly different from each other. The F-table is consulted to determine if there is a significant difference among the means.
In order to carry out the one-way ANOVA, we will do the following steps:

1. State the null and alternative hypothesis.
2. Compute SS (sums of squares).
3. Compute degrees of freedom (df).
4. Compute MS (mean squares).
5. Compute F.
6. Take a decision.

Example:

An instructor divided the classroom into three rows: first, second, and third. The first in the front row, the second in the middle row and the third in the back row. The instructor noticed that the further the students were from him, the more likely they were to miss class or use an instant messenger during class. A random sample of the students in each row was taken. The score for those students on the second exam was recorded:

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| First (F) | 82 | 83 | 97 | 93 | 55 | 67 | 53 | | |
| Second (S) | 83 | 78 | 68 | 61 | 77 | 54 | 69 | 51 | 63 |
| Third (T) | 38 | 59 | 55 | 66 | 45 | 52 | 52 | 61 | |

At level of significance of $\alpha = 0.05$, the instructor wanted to see are the students further away did worse on the exams?

Solution:

1. State the null and alternative hypothesis.

The null hypothesis of this example is:

$$H_0 : \mu_F = \mu_S = \mu_T$$

The alternative hypothesis is:

$$H_A : \text{At least one of the means differ.}$$

The summary statistics for the grades of each row are shown in the table below

| | First (F) | Second (S) | Third (T) |
|---|---|---|---|
| Sample size | 7 | 9 | 8 |
| Mean | 75.71 | 67.11 | 53.50 |
| St. Dev. | 17.63 | 10.95 | 8.96 |
| Variance | 310.90 | 119.86 | 80.29 |

## VARIATION:

Variation is the sum of the squares of the deviations between a value and the mean of the value. Sum of squares, which is another name for variation, is abbreviated by SS and often followed by a variable in parentheses such as SS(B) or SS(W) so we know which sum of squares we're talking about.

- ARE ALL OF THE VALUES IDENTICAL? No, so there is some variation in the data. This is called the total variation and denoted by SS(T).
- ARE ALL OF THE SAMPLE MEANS IDENTICAL? No, so there is some variation between the groups. This is called variation between groups. Sometimes called the variation due to the factor and denoted by SS(B).
- ARE EACH OF THE VALUES WITHIN EACH GROUP IDENTICAL? No, there is some variation within the groups. This is called variation within groups. Sometimes called the error variation and denoted by SS(W).

As a result, there are two sources of variation:

- The variation between the groups, SS(B), or the variation due to the factor.
- The variation within the groups, SS(W), or the variation that can't be explained by the factor so it's called the error variation and denoted sometimes by SS(E).

THE GRAND MEAN:

The grand mean is the average of all the values when the factor is ignored. It is a weighted average of the individual sample means.

$$\bar{y}_{..} = \frac{\sum_{i=1}^{k} n_i \bar{y}_{i.}}{\sum_{i=1}^{k} n_i}$$

$$\bar{y}_{..} = \frac{n_1\bar{y}_{1.} + n_2\bar{y}_{2.} + \ldots + n_k\bar{y}_{k.}}{n_1 + n_2 + \ldots + n_k}$$

Grand Mean for our example is:

$$\bar{y}_{..} = \frac{7(75.71) + 9(67.11) + 8(53.50)}{7 + 9 + 8} = 65.08$$

2. Compute SS (sums of squares)

a. BETWEEN GROUP VARIATION, SS(B): The between group variation is the variation between each sample mean and the grand mean. Each individual variation is weighted by the sample size.

$$SS(B) = \sum_{i=1}^{k} n_i(\bar{y}_{i.} - \bar{y}_{..})^2$$
$$= n_1(\bar{y}_{1.} - \bar{y}_{..})^2 + n_2(\bar{y}_{2.} - \bar{y}_{..})^2 + \ldots + n_k(\bar{y}_{k.} - \bar{y}_{..})^2$$

The Between Group Variation for our example is:
$SS(B) = 7(75.71 - 65.08)^2 + 9(67.11 - 65.08)^2 + 8(53.50 - 65.08)^2 = 1900.84.$

b. WITHIN GROUP VARIATION, SS(W): The Within Group Variation is the weighted total of the individual variations. The weighting is done with the degrees of freedom. The df for each sample is one less than the sample size for that sample. Within Group Variation is:

$$\begin{aligned} SS(W) &= \sum_{j=1}^{k} df_j S_j^2 \\ &= df_1 S_1^2 + df_2 S_2^2 + \ldots + df_k S_k^2 \end{aligned}$$

The within group variation for our example is:
$SS(W) = 6(310.90) + 8(119.86) + 7(80.29) = 3386.31.$

c. After filling in the sum of squares, we have:

| Source | SS | df | MS | F | $F_\alpha$ | P-value |
|---|---|---|---|---|---|---|
| Between | 1900.84 | | | | | |
| Within | 3386.31 | | | | | |
| Total | 5287.15 | | | | | |

3. Compute degrees of freedom (df)

A degree of freedom occurs for each value that can vary before the rest of the values are predetermined. For example, if you had six numbers that had an average of 40, you would know that the total had to be 240. Five of the six numbers could be anything, but once the first five is known, the last one is fixed so the sum is 240. The df in this case would be 6-1=5.

The Between Group df is one less than the number of groups. We have three groups, so df(B) = 2.

The Within Group DF is the sum of the individual df's of each group

The sample sizes are 7, 9, and 8

$$df(W) = 6 + 8 + 7 = 21$$

The total df is one less than the sample size

$$df(\text{Total}) = 24–1 = 23$$

Filling in the degrees of freedom gives this:

| Source | SS | DF | MS | F | $F_\alpha$ | P-value |
|---|---|---|---|---|---|---|
| Between | 1900.84 | 2 | | | | |
| Within | 3386.31 | 21 | | | | |
| Total | 5287.15 | 23 | | | | |

## 4. VARIANCES

The variances are also called the Mean of the Squares and abbreviated by MS, often with an accompanying variable MS(B) or MS(W). They are an average squared deviation from the mean and are found by dividing the variation by the degrees of freedom

$$MS = \frac{SS}{df}$$

-MS(B)= 1900.84 / 2 = 950.42
-MS(W)= 3386.31 / 21 = 161.253

Notice that the MS(Total) is NOT the sum of MS(Between) and MS(Within). This works for the sum of squares SS(Total), but not the mean square MS(Total). The MS(Total) isn't usually shown.

- Completing the MS gives

| Source | SS | df | MS | F | $F_\alpha$ | P-value |
|---|---|---|---|---|---|---|
| Between | 1900.84 | 2 | 950.42 | | | |
| Within | 3386.31 | 21 | 161.253 | | | |
| Total | 5287.15 | 23 | | | | |

Special Variances

The MS(Within) is also known as the pooled estimate of the variance since it is a weighted average of the individual variances. The MS(Total) is the variance of the response variable, not technically part of ANOVA table.

## 5. Compute F test statistic

An F test statistic is the ratio of two sample variances. The MS(B) and MS(W) are two sample variances and that's what we divide to find F.

$$F = MS(B) / MS(W)$$

For our data, F = 951.055 / 161.2 = 5.898 Adding F to the table

| Source | SS | DF | MS | F | $F_\alpha$ | P-value |
|---|---|---|---|---|---|---|
| Between | 1900.84 | 2 | 950.42 | 5.894 | $F_{0.05}^{(2),(21)} = 3.47$ | |
| Within | 3386.31 | 21 | 161.253 | | | |
| Total | 5287.15 | 23 | | | | |

The F test is a right tail test. The F test statistic has an F distribution with df(B) numerator df and df(W) denominator df. The tabulated value is $F_{0.05,2,21} = 3.47$. Since the calculated F value, F=5.894, is greater than the tabulated F value, $F_{0.05,2,21} = 3.47$, we reject $H_0$.

The p-value is the area to the right of the test statistic: $P(F_{df(B),df(W)} > F)$ Using any statistical package, one have: $P(F_{2,21} > 5.9) = 0.009$.
Completing the table with the p-value

| Source | SS | DF | MS | F | $F_\alpha$ | P-value |
|---|---|---|---|---|---|---|
| Between | 1900.84 | 2 | 950.42 | 5.894 | $F_{0.05}^{(2),(21)} = 3.47$ | 0.009 |
| Within | 3386.31 | 21 | 161.253 | | | |
| Total | 5287.15 | 23 | | | | |

The p-value is 0.009, which is less than the significance level of 0.05, so we reject the null hypothesis. The null hypothesis is that the means of the three rows in class were the same, but we reject that, so at least one row has a different mean.

There is enough evidence to support the claim that there is a difference in the mean scores of the front, middle, and back rows in class. The ANOVA doesn't tell which treatment is different, you would need to look at confidence intervals or run post hoc tests to determine that. (out of the course)

In General, one-way ANOVA table is as bellow:

| Source of variation (Source) | Sum of squares (SS) | degree of freedom (df) | Mean Square (MS) | F |
|---|---|---|---|---|
| BETWEEN (B) | SSB | k-1 | MSB | $F_B = \frac{MSB}{MSW(E)}$ |
| Within(Error) | SSW(E) | n-k | MSW(E) | |
| Total | SSTOT | n-1 | | |

One way to improve the precision of the treatment comparisons is to reduce variability among the units. We can group units into blocks so that each block contains relatively homogeneous units. Within each block, we randomly assign treatment to units (we do a separate random assignment for each block). The number of units per block is a multiple of number of the number of factor combinations. The most common isto use each treatment onece in each block.
Example: Toxicology Assays
Treatment: 6 levels of toxin
units:24 mice
Blocks: make 4 blocks with 6 mice per block, based on weight.
Clinical trails:
Comparison of 3 treatments for controlling blood pressure
Treatment: 3 drugs units: 30 human subjects
Blocks: make 10 blocks with 3 subjects per block. Match on weight, sex, smoking status, diet, initial blood pressure data.

Benefits of Blocking

- Reduction in variability of estimators for treatment means.
    - Improved power for F test.
    - smaller values for MSE.

The data of Completely Randomized Designs (CRD) are arranged in two way table

| Treatments | Blocks 1 | 2 | ... | j | ... | b | Treatment means |
|---|---|---|---|---|---|---|---|
| 1 | $y_{11}$ | $y_{12}$ | ... | $y_{1j}$ | ... | $y_{1b}$ | $\bar{y}_{1.}$ |
| 2 | $y_{21}$ | $y_{22}$ | ... | $y_{2j}$ | ... | $y_{2b}$ | $\bar{y}_{2.}$ |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| i | $y_{i1}$ | $y_{i2}$ | ... | $y_{ij}$ | ... | $y_{ib}$ | $\bar{y}_{i.}$ |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| k | $y_{k1}$ | $y_{k2}$ | ... | $y_{kj}$ | ... | $y_{kb}$ | $\bar{y}_{k.}$ |
| Block means | $\bar{y}_{.1}$ | $\bar{y}_{.2}$ | ... | $\bar{y}_{.j}$ | ... | $\bar{y}_{.b}$ | $\bar{y}_{..}$ |

- $y_{ij}$= the response the response obtained by using treatment $i$ in block $j$,
- $\bar{y}_{i.}$= mean of the response for the $i-$th treatment level,
- $\bar{y}_{.j}$= mean of the response for the $j-$th block,
- $\bar{y}_{..}$= mean of all kb response observations.

The analysis of variance model in CRD is based on the following:

$$y_{ij} = \mu_{ij} + e_{ij}, i = 1, \ldots, k, j = 1, \ldots, b.$$

where $y_{ij}$ represents the observation of the $i-$th treatment in the $j-$th block, $\mu_{ij}$ is the mean response of the $i-$th treatment in the $j-$th block and $e_{ij}$ are independent random errors with $N(0, \sigma^2)$. The response mean for the level $i$ of the factor is $\mu_{i.} = E(\bar{y}_{i.})$ and the response mean for the block $j$ is $\mu_{.j} = E(\bar{y}_{.j})$ and the overall mean is $\mu_{..} = E(\bar{y}_{..})$. The hypothesis to be tested is as follows:

- for treatment effects

  $H_0$ :$\mu_{1.} = \mu_{2.} = \ldots = \mu_{k.}$

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

- for block effects

  $H_0$ :$\mu_{.1} = \mu_{.2} = \ldots = \mu_{.b}$

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

Sum of squares quantities:
We have

$$\text{SST} = \text{SSA} + \text{SSB} + \text{SSE},$$

where

$$\begin{aligned}
\text{SST(TOT)} &= \sum_{i=1}^{k} \sum_{j=1}^{b} (y_{ij} - \bar{y}_{..})^2 \\
\text{SSA} &= \text{b} \sum_{i=1}^{k} (\bar{y}_{i.} - \bar{y}_{..})^2 \\
\text{SSB} &= \text{k} \sum_{j=1}^{b} (\bar{y}_{.j} - \bar{y}_{..})^2 \\
\text{SSE} &= \sum_{i=1}^{k} \sum_{j}^{b} (y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..})^2
\end{aligned}$$

| Source | degree of freedom | Sum of Squares |
|---|---|---|
| Treatment | k-1 | SSA |
| Block | b-1 | SSB |
| Residual(error) | (k-1)(b-1) | SSE |
| Total | kb-1 | SST(TOT) |

The null hypothesis of no treatment effect difference

$$H_0 : \mu_{1.} = \mu_{2.} = \ldots = \mu_{k.}$$

or

$$H_0^{\text{treat}} : \mu_1 = \mu_2 = \ldots = \mu_k$$

can be tested can be tested by using the F statistic

$$F = \frac{SSA/(k-1)}{SSE/(k-1)(b-1)} = \frac{MSA}{MSE},$$

where SSA and SSE are the treatment and error sums of squares.
The F test rejects $H_0$ at level of significance $\alpha$ if the F statistic value in exceeds $F_{\alpha,k-1,(k-1)(b-1)}$ (reject $H_0$ if $F > F_{\alpha,k-1,(k-1)(b-1)}$)

The null hypothesis of no block effect difference

$$H_0 : \mu_{.1} = \mu_{.2} = \ldots = \mu_{.b}$$

or

$$H_0^{\text{block}} : \mu_1 = \mu_2 = \ldots = \mu_b$$

can be tested can be tested by using the F statistic

$$F = \frac{SSB/(b-1)}{SSE/(k-1)(b-1)} = \frac{MSB}{MSE},$$

where SSB and SSE are the block and error sums of squares. The F test rejects $H_0$ at level of significance $\alpha$ if the F statistic value in exceeds $F_{\alpha,b-1,(k-1)(b-1)}$ (reject $H_0$ if $F > F_{\alpha,b-1,(k-1)(b-1)}$)

# Anova Table for Completely Randomized Block Design:

| Source of variation (Source) | Sum of squares (SS) | degree of freedom (df) | Mean Square (MS) | F |
|---|---|---|---|---|
| Treatments (A) | SSA | k-1 | MSA | $F_A = \frac{MSA}{MSE}$ |
| Blocks (B) | SSB | b-1 | MSB | $F_B = \frac{MSB}{MSE}$ |
| Error | SSE | (k-1)(b-1) | MSE | |
| Total | SST | kb-1 | | |

Example:

Four methods of manufacturing penicillin were compare in a randomized block design. The block are blends of the raw material, corn steep liquor, known to be quite variable. The yield of each method for five blends is given below.

| | blend (block) | | | | |
|---|---|---|---|---|---|
| Method | 1 | 2 | 3 | 4 | 5 |
| A | 89 | 84 | 81 | 87 | 79 |
| B | 88 | 77 | 87 | 92 | 81 |
| C | 97 | 92 | 87 | 89 | 80 |
| D | 94 | 79 | 85 | 84 | 88 |

At level $\alpha = 0.05$

1. are there significant differences between the blends (blocks)?
2. are there significant differences between the Methods (treatments)?

| Method | blend (block) 1 | 2 | 3 | 4 | 5 | Method mean |
|---|---|---|---|---|---|---|
| A | 89 | 84 | 81 | 87 | 79 | $\bar{y}_{1.} = 84$ |
| B | 88 | 77 | 87 | 92 | 81 | $\bar{y}_{2.} = 85$ |
| C | 97 | 92 | 87 | 89 | 80 | $\bar{y}_{3.} = 89$ |
| D | 94 | 79 | 85 | 84 | 88 | $\bar{y}_{4.} = 86$ |
| blend mean | $\bar{y}_{.1} = 92$ | $\bar{y}_{.2} = 83$ | $\bar{y}_{.3} = 85$ | $\bar{y}_{.4} = 88$ | $\bar{y}_{.5} = 82$ | $\bar{y}_{..} = 86$ |

We have $k = 4, b = 5, \bar{y}_{..} = \frac{1}{4\times 5}\sum_{i=1}^{4}\sum_{j=1}^{5} y_{ij} = \frac{89+84+81+87+79+\ldots+94+79+85+84+88}{20} = \frac{1720}{20}$.

$SSA = b\sum_{i=1}^{k}(\bar{y}_{i.} - \bar{y}_{..})^2 = 5\left[(84-86)^2+(85-86)^2+(89-86)^2+(86-86)^2\right] = 5\left[4+1+9+0\right] = 70.$

$SSB = k\sum_{j=1}^{b}(\bar{y}_{.j} - \bar{y}_{..})^2 = 4\left[(92-86)^2+(83-86)^2+(85-86)^2+(88-86)^2+(82-86)^2\right] =$

$4\left[36+9+1+4+16\right] = 264.$

$SST = \sum_{i=1}^{k}\sum_{j=1}^{b}(y_{ij} - \bar{y}_{..})^2 = \sum_{i=1}^{4}\sum_{j=1}^{5}(y_{ij} - \bar{y}_{..})^2 =$

$(89-86)^2 + (84-86)^2 + \ldots + (84-86)^2 + (88-86)^2 = 9+4+\ldots+4+4 = 560$

$SSE = SST - SSA - SSB = 560 - 70 - 264 = 226$

## ANOVA of CRD table

| (Source) | (SS) | (df) | (MS) | F |
|---|---|---|---|---|
| Methods (A) | 70 | 3 | 23.333 | 1.239 |
| blends (B) | 264 | 4 | 66.000 | 3.504 |
| Error | 226 | 12 | 18.833 | |
| Total | 560 | 19 | | |

1. Test the blend (blocks) effects

$$H_0 : \mu_{.1} = \mu_{.2} = \ldots = \mu_{.5}$$

$$H_1 : \text{at least one blend mean is different.}$$

$F = 3.504 > F_{0.05,4,12} = 3.26$, then we reject $H_0$, which means there do appear to be significant differences between blends.

2. Test the method effects

$$H_0 : \mu_{1.} = \mu_{2.} = \ldots = \mu_{.4}$$

$$H_1 : \text{at least one method mean is different.}$$

$F = 1.239 < F_{0.05,3,12} = 3.49$, then we fail to reject $H_0$, which means there is no significant differences between methods.