Statistics 101

Proudly built with R, Quarto and GitHub Copilot

Daniel Sánchez Pazmiño

Table of contents

A gentle look at analysis of variance

So far, we've talked about how to make statements about the population using a descriptive statistic, the mean and statistical inference. We have not made the difference between observational and inferential statistics. We could be using any kind of data to make statements about the population in the way we've done before.

ANOVA or analysis of variance is our first attempt at using statistical inference to make statements about the population using data from an experiment. The experiment is the key word here. We are going to use data from an experiment to make statements about the population, and this type of data alllows for the use of ANOVA. In some cases, observational data can also be analysed using ANOVA.

Experimental Design

There are several types of experiments, and different experiments have different ways of applying ANOVA to them. Consider the case where we have designed three special kinds of trainings to students. We say that these trainings will be the independent variable, because it is us, the researchers, who will control who gets the training and who doesn't. The type of training corresponds to the **factor**, and the levels of the factor are the different types of training. Often, the levels of the factor are called **treatments**. The dependent or response variable is the outcome of the experiment, and in this case, it is the score of the students on a test.

We call this experiment a "single factor experiment". This is because all that varies in the experiment is the type of training, which we could define as type A, type B and type C. The students are the same, the test is the same, the time of the test is the same, the place of the test is the same, etc. The only thing that changes is the type of training.

In a **completely randomised design**, we randomly assign the students to the different treatments. This is the most basic experimental design. Why randomise? This is because we want to be sure about the causal effect of the training on the student. If we were to, say, give the training based on a certain variable, such as the propensity to develop mental health issues, the design would no longer be randomised. By losing randomisation, when we analyse the response variables for all three treatments, we will not be able to say that the difference in the response variable is due to the training, because the propensity to develop mental health issues could also be a factor in the difference in the response variable.

If we are to then randomly assign students to each of three types of training, we will have three groups of students. Each group will have a certain number of participants, which we will denote as n_1, n_2 and n_3. The total number of participants will be $n_T = n_1 + n_2 + n_3$. Let's say that each group has 10 participants, so $n = 30$ and $n_1 = n_2 = n_3 = 10$

We can produce a table of the data, where the first column is the group, the second column is the type of training and the third column is the score of the student on the test. Below, we show an **extract** of the table which only shows the first two participants of each group, in long format. It is important that we keep the long format for the data, because it is the format that is required by statistical software.

Table 1: An extract of the example experiment

id	group	training	score
6	1	A	55.06
8	1	A	44.53
13	2	B	42.24
16	2	B	48.90
21	3	C	51.34

Table 2: Descriptive statistics of the score in the experiment

	training	Mean	SD	N	Median	Min	Max	Var
score	A	46.17	9.96	10	44.44	26.54	60.84	99.15
	B	48.82	10.67	10	45.06	40.02	74.16	113.90
	C	46.12	6.66	10	45.34	35.52	55.75	44.37

id	group	training	score
30	3	C	40.64

Using all of the data, we can produce descriptive statistics of the score, which is the response variable. You can see the table below.

Why do we want to use ANOVA? Well, remember that while the differences shown by the descriptive statistics might be interesting, we want to make statements about the population. The results that we get will vary across different selections of students, so we want to know if the differences between mean scores in the different groups are large enough to actually suggest that there would be differences in the population if we were to apply these trainings in the population.

Define the population means of the score for three groups as μ_1, μ_2 and μ_3. We want to know if the differences between these means are large enough to suggest that there would be differences in the population. We are effectively conducting a two sided test, where the null hypothesis is that the means are equal, and the alternative hypothesis is that the means are not equal. See below:

$$H_0 : \mu_1 = \mu_2 = \mu_3$$

$$H_1 : \text{Not all population means are equal}$$

Note that if we are to reject H_0, we cannot say which of the means are different. We can only say that at least one of the means is different. We will see later how we can find out which means are different.

ANOVA assumptions

ANOVA has several assumptions that need to be met in order for the results to be valid. We review them below.

1. The response variable is normally distributed in the population. This is the most important assumption. If the response variable is not normally distributed, then the results of the ANOVA will not be valid. We can check this assumption by looking at the histogram of the response variable.
2. The variance of the response variable is the same across all groups. This is a very classic assumption which is rarely met, but there are statistical adjustments which can be performed to the analysis to account for this. We will not cover those here (as they are quite technical) and there are easier ways of dealing with this using regression models. We can check this assumption by looking at the boxplot of the response variable.
3. Observations are independent. This means that the observations in one group are not related to the observations in another group. Likely this means that the observations are not repeated measures and that the participants of the experiment are not related to each other (they should be isolated).

ANOVA with completely randomised designs

As we've seen before, ANOVA is nothing more than a hypothesis test. However, it is a hypothesis test that is applied to more than two groups. The null hypothesis is that the means of the groups are equal, and the alternative hypothesis is that the means of the groups are not equal.

Here, we will define the concepts required to construct the hypothesis test, and the probability distribution that we will use to get the result of the hypothesis test.

It's important to define some notation. The number of levels in the factor is usually denoted as k in statistics textbooks. So, in the example that we had before, $k = 3$. When we want to refer to one of the different factor levels, we use the j index. So, in the example that we had before, $j = 1, 2, 3$. In a more general setting, we can say that $j = 1, 2, ..., k$.

Thus, the number of observations in the jth group is denoted as n_j. In the example that we had before, $n_1 = n_2 = n_3 = 10$. The total number of observations is denoted as n_T, and in the example that we had before, $n_T = 30$.

The j indexing, which is defined for the number of factor levels, can be used to index several other quantities. For example, the population mean of the jth group is denoted as μ_j. In the example that we had before, μ_1 is the mean of the first group, μ_2 is the mean of the second group and μ_3 is the mean of the third group. The sample mean of the jth group is denoted as $\bar{x}_j$. In the example that we had before, $\bar{x}_1$ is the sample mean of the first group, $\bar{x}_2$ is the sample mean of the second group and $\bar{x}_3$ is the sample mean of the third group.

The sample standard deviation will be denoted as s_j. We calculate the j-th descriptive statistics by calculating the descriptive statistics using only the observations in the jth group.

Variance estimation

Consider what would represent evidence that supports the alternative hypothesis, which states that the population means are not equal. If the population means are not equal, then the sample means will also not be equal. How much different must the sample means be in order to make a conclusion at the population level? This is what ANOVA will tell us. However, it is clear that more differences between sample means will be more convincing evidence that the population means are not equal.

ANOVA analyses the variability of the sample means at every different factor level. The more variance between the sample means, the more evidence that the population means are not equal, and thus the more likely we are to reject the null hypothesis, which says that the population means are equal.

The null hypothesis is saying that the population means are equal to each other because the people in the different groups come from the same distribution, and the differences between their sample means are **purely** due to chance, as they also come from the same sampling distribution. So, by performing the hypothesis test, we are essentially deciding whether or not there exist different sampling distributions.

How can we cleverly run this test? We will use the assumptions of the ANOVA method in order to produce a test statistic that will tell us how much evidence there is that the population means are not equal. We define an *estimate* of the population variance based on a calculation of the sampling variance. We can use the sample data to calculate the overall sample mean, which simply is the sample mean of all observations (regardless of treatment group). We can also calculate the sample mean of each group. We then produce an estimate of the sampling variance by calculating the standard deviation of all the sample means. If we multiply this number by n, we get the estimate of the population variance. This is called the between group variance.

If the null hypothesis is true, then the between group variance will be a good estimate of the sampling variance. If the null hypothesis is false, however, it is an overestimate of the sampling variance. ANOVA compares this estimate of the variance to the *within group variance*. We calculate the within group variance by calculating the variance of each group separately, and then averaging them.

The between group variance is usually called the mean square between groups, and is denoted as $MS_{between}$. The within group variance is usually called the mean square within groups, and is denoted as MS_{within}.

The F-test

As mentioned before, $MS_{between}$ will overestimate the sampling variance if the null hypothesis, which states that the population means are equal, is false. Thus, if the null hypothesis is false,

then $MS_{between}$ will be larger than MS_{within}. If the null hypothesis is true, then $MS_{between}$ will be smaller than MS_{within}.

We want to then compare the ratio of $MS_{between}$ and MS_{within}. If the null hypothesis is true, then this ratio will be close to 1. If the null hypothesis is false, then this ratio will be larger than 1. Statistical theory tells us that the ratio of two variances like these follow a probability distribution called the F, or Fisher, distribution. This distribution, like the normal and the t, has a table which you can use.

So, the test statistic for ANOVA is

$$F = \frac{MS_{between}}{MS_{within}}$$

which has degrees of freedom for both numerator and denominator. The degrees of freedom for the numerator is $k - 1$, and the degrees of freedom for the denominator is $n_T - k$.

The ANOVA table

The ANOVA table is a table that is used to organise the results of the ANOVA hypothesis test. It is a table that is organised in a very specific way. The table is shown below.

Source of variation	Sum of squares	Degrees of freedom	Mean square	F-statistic
Between groups				
Within groups				
Total				

The first column is the source of variation. The second column is the sum of squares. The third column is the degrees of freedom. The fourth column is the mean square. The fifth column is the F-statistic. Usually a sixth column is also included, which is the p-value.

How to calculate the mean square errors that we need for the ANOVA table is a bit complicated. There are specific formulas to calculate these by hand, it will be difficult to do and would require a fair amount of time and knowledge about notation. Statistical software like SPSS, R, SAS, etc. will do this for you very easily. It is more important to understand the concepts behind the ANOVA table than to be able to calculate it by hand. It is important though to note that the "Total" row is the sum of the "Between groups" and "Within groups" rows. Of course, we don't do any calculations for the total row in the F-statistic column. Everything will become clear when we do an example below.

As always, calculating the p-value is a bit complicated using only statistical tables, so we will use statistical software to calculate it for us and, for hand calculations, we will only calculate the F-statistic and compare to the critical value. To finalise the test, once we've calculated the

F-statistic, we can look up the critical value in the F table. It will ask you to give the degrees of freedom for the numerator and denominator. The degrees of freedom for the numerator is $k-1$, and the degrees of freedom for the denominator is $n_T - k$. The critical value will correspond to the value of F for a given significance level. Usually it is the right tail probability that is given in the table, which is the one that we will use for the hypothesis test. If the calculated F-statistic is larger than the critical value, then we reject the null hypothesis. If the calculated F-statistic is smaller than the critical value, then we fail to reject the null hypothesis.

The ANOVA table for the example of the training programs

The ANOVA table for the example that we had before is shown below.

```
            Df Sum Sq Mean Sq F value Pr(>F)
training     2   47.7   23.87   0.278  0.759
Residuals   27 2316.8   85.81
```

The leftmost column is $Pr(> F)$, which corresponds to the right tail probability of the F-statistic (the p-value). Because the p-value is greater than 0.05, we fail to reject the null hypothesis. In fact, this p-value is very large, which means that one would reject the hypothesis that the population means are not equal only if the significance level was very large, like 0.5 or 0.6. We never use such large significance levels, (the top is usually $\alpha = 0.1$) so we fail to reject the null hypothesis with a great deal of confidence. This means that we have no evidence that the population means are not equal, and thus we have no evidence that the training programs have a true effect on the scores.

Summary

- ANOVA is a hypothesis test that is used to test whether or not the population means of several groups are equal to each other.
- ANOVA requires a great deal of assumptions which really are only true in certain cases, and most importantly, when we have randomised the treatment groups. This is why ANOVA is usually used in experiments, but not with observational studies.
- ANOVA is a hypothesis test that is based on the F-distribution. The F-distribution is a probability distribution that is used to test whether or not the ratio of two variances is equal to 1.
- The ANOVA table is a table that is used to organise the results of the ANOVA hypothesis test. It is a table that is organised in a very specific way. It includes the source of variation, the sum of squares, the degrees of freedom, the mean square, and the F-statistic.

- The test statistic of ANOVA is the F-statistic, which is the ratio of the between group variance and the within group variance. If the null hypothesis is true, then this ratio will be close to 1. If the null hypothesis is false, then this ratio will be larger than 1.
- We use an F critical value table to find the critical value for the F-statistic. The degrees of freedom for the numerator is $k-1$, and the degrees of freedom for the denominator is $n_T - k$. We use the given level of significance to find the critical value. If the calculated F-statistic is larger than the critical value, then we reject the null hypothesis. If the calculated F-statistic is smaller than the critical value, then we fail to reject the null hypothesis.

Up next

- Regression
- Correlation
- Non parametric statistics
- SPSS I
- SPSS II