

20. Analysis of Variance

In our previous problems we had one sample and tested whether our sample differed from a known population mean. In the analysis of variance we will test multiple samples – and hence multiple treatment groups – against each other. The basic idea used by the **AN**alysis **O**f **V**ariance (**ANOVA**) is that the total variability within the pooled sample has three components:

- the variability within each treatment group;
- the variability between the groups; and
- the residual variability—everything else.

ANOVA tests to see of the "between groups" differences are large enough to conclude that the groups are really different.

For motivation, consider a couple of examples.

20.1. Example.

Suppose that there are three different diet plans each enrolling five clients. At the end of one month the following losses are recorded:

Plan A	Plan B	Plan C
5	15	10
5	15	10
5	15	10
5	15	10
5	15	10

The average loss in each plan is

$$\bar{x}_A = 5 \quad \bar{x}_B = 15 \quad \text{and} \quad \bar{x}_C = 10$$

Remember that the **population variance** involved calculating how much each observation differed from the mean, $x_i - \bar{x}$, squaring it $(x_i - \bar{x})^2$, and then averaging the squared differences

$$\frac{1}{n} \sum_i (x_i - \bar{x})^2$$

ANOVA involves **comparing sums of squared differences** like the above. The differences

$$(x_i - \bar{x})$$

represent the **error** between the observed value and the average value. With the diet plan data, the **sum of the squared differences** over the entire sample

$$\sum_{i=1}^{15} (x_i - \bar{x})^2 = 250$$

is the numerator in the variance

$$\sigma_{Total}^2 = \left(\frac{\sum_{i=1}^{15} (x_i - \bar{x})^2}{15} \right) = 16.667$$

But there's another way to think about the variability. We could instead calculate the variability **between the treatments**, i.e., between the **plans**. Since the average loss in the three plans is

$$\frac{1}{3}(5 + 15 + 10) = 10,$$

the variability in the column averages is

$$\sigma_{plans}^2 = \frac{(5 - 10)^2 + (15 - 10)^2 + (10 - 10)^2}{3} = \frac{50}{3} = 16.667.$$

In the above example, **the two numbers, σ_{Total}^2 and σ_{plans}^2 are the same.**

This isn't surprising since

In this example, *all* of the variability in the sample is due to the **differences in treatments** (column effects) and *none* of the variability is due to **individual differences in the clients**

In principle, there are three sources of variability in our diet plan example:

- the variability within each treatment group;
- the variability between the groups; and
- the residual variability—all other sources of variability having nothing to do with the groups.

In our idealized example, the only variability is between the groups.

In the real world you will never get such perfect data. There will always be some variability due to influences other than the treatments (column effects). In the diet plan, initial weight, gender, age, exercise and other uncontrolled variables (including measurement error) will result in other sources of variability. More realistic data might be similar to that given in the next example.

20.2. Example.

Suppose that the data from the three diet plans had been:

	Plan A	Plan B	Plan C
	5	10	5
	5	15	5
	5	15	10
	10	15	10
	15	15	10

Using the two methods above, find the overall variance and the variance due

to the *treatment (column) effects*.

Solution. We still have the same fifteen numbers as in the first example, except now those numbers are distributed differently between the plans. However, this won't change the total variance, so

$$\sigma^2_{Total} = 16.667$$

as before.
The sample means are now

$$\bar{x}_A = 8 \quad \bar{x}_B = 14 \quad \text{and} \quad \bar{x}_C = 8$$

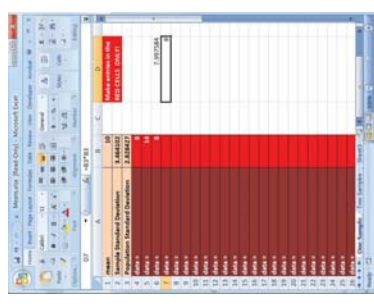
and so we can find the variance due to the column effects using the MEANS spreadsheet:

- enter the column means 8, 14 and 8 into the spreadsheet;
- find the population variance by squaring the population standard deviation. Since the spreadsheet calculates the value of the population standard deviation in cell B3, you can do this by entering

$$=b3*b3$$
into a cell on the spreadsheet and pressing enter. This gives you a value of

$$\sigma^2_{Plans} = 8.$$

Thus about 50% ($\frac{8}{16.667}$) of the total variability in the population is due to the differences in treatments. The remaining variability is due to differences within the treatment groups and to residual effects not attributable to the diet plans.



ANOVA lets you decide whether or not the kinds of differences observed in the previous example are statistically significant. In particular, ANOVA will test

$$H_0 : \mu_A = \mu_B = \mu_C \quad \text{against}$$

H_A : the means are not all the same

Another way of saying the alternative hypothesis is:

$$H_0 : \mu_A = \mu_B = \mu_C \quad \text{against}$$

H_A : at least 2 differ one from the other

since if the means are not all the same, at least two must be different, one from the other.

The test will be accomplished by comparing the *variance* of the overall sample against the *variance* of the sample means. In practice, however, programs like the `analyzeThis` spreadsheet calculate sums of squared differences, as we did above. As we shall see, the report generated by

are theoretically equivalent to ANOVA and computationally simpler. If you are given the actual data, the `analyzeThis` spreadsheet will do all the necessary calculations and perform the ANOVA test for you. If you are just given summary information—for example on the final exam—you will use the `Formulas` spreadsheet and will be given the following information:

- Sample means for each treatment group.
- Overall sample standard deviation s_{Total} .
- Sample size n for each treatment group.
- Significance level.

the ANOVA tab of `AnalyzeThis` includes sums of squared differences and averages—mean square differences—for each of the sources of variability mentioned above.

The basic assumptions of ANOVA are

- Each group is drawn from a normally distributed population;
- All populations have a common variance;
- all samples are independent;
- within each group sample, all observations are random and independent.

The test is fairly robust against violations of the normality assumption, provided that the sample sizes are equal, sufficiently large, and symmetrical. The test is less robust against violations of the second assumption, that all the group populations have the same variance. It's also possible to test for violations of the above assumptions, although we won't cover that in this course.

Remark. Usually ANOVA is only done for three or more treatment groups. When you have just two groups, hypothesis tests for means

20.3. Example.

An attorney is interested in whether knowledge of prior bad acts—criminal convictions—will influence potential juror's view of the guilt or innocence of a person accused of a crime. She randomly samples potential jurors and then divides her sample into three groups. Group I is told that the accused has a criminal record, Group II is told that the accused has no criminal record, and Group III is given no information on prior bad acts. Each subject then fills out a questionnaire measuring how likely it is, in the view of the subject, that the accused is guilty.

Group	I	II	III
mean	8	3.8	5
sample size	5	5	5

The pooled standard deviation was 2.8. At the 1% significance level do the data show that information on prior bad acts influences juror perceptions?

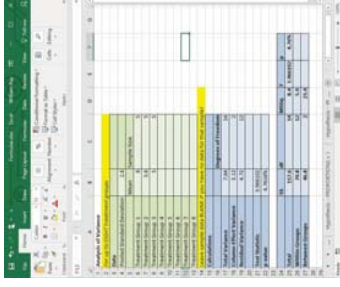
Solution.

Step 1. The above table, together with the pooled standard deviation of 2.8 provides the dictionary needed for the spreadsheet.

Step 2. Now enter the summary data and the pooled standard deviation into the spreadsheet FORMULAS.XLSX, found in the resources section for this course on LEARN.OU.EDU. Note that you will need to select the tab at the bottom labeled ANOVA.

Step 3. Read the significance level from the spreadsheet. If it's less than the pre-set target, then we reject the null hypothesis, otherwise we accept the null hypothesis.

In this case, the pre-set significance level was 1% and the *p*-value is 4.76%, so we do NOT reject the null hypothesis. If we believe that the



Step 2. Now enter the summary data and the pooled standard deviation into the spreadsheet FORMULAS.XLSX, found in the resources section for this course on LEARN.OU.EDU. Note that you will need to select the tab at the bottom labeled ANOVA.

Step 3. Read the significance level from the spreadsheet. If it's less than the pre-set target, then we reject the null hypothesis, otherwise we accept the null hypothesis. ANOVA is always a right-tailed test



End of Solution Template

The spreadsheet reports some other calculations in addition to the *p*-value.

means are not all the same, then the chance of a Type I Error is 4.76% Another way of stating this concluding is that the results are not highly significant but are significant since $1\% < p < 5\%$.

Remark. Note that an ANOVA is *always* a right-tailed test.

Solution Template

Step 1. Build your dictionary if it's not already in the problem:

	pooled standard deviation	σ
Mean for treatment group 1	\bar{x}_1	
Mean for treatment group 2	\bar{x}_2	
Mean for treatment group 3	\bar{x}_3	

The spreadsheet permits up to eight treatment groups, although ANOVA in general works for any finite number of treatment groups.

First, it calculates the **total variance**, which is the square of the pooled standard deviation. Second, it calculates the **column effect variance**, which is the amount of variability due to the treatments. The spreadsheet uses the following formula (in the case of four treatment groups):

$$V_C = \frac{n_A \bar{x}_A^2 + n_B \bar{x}_B^2 + n_C \bar{x}_C^2 + n_D \bar{x}_D^2}{n_A + n_B + n_C + n_D} - \bar{x}^2$$

where \bar{x} is the pooled average.

In the example with the lawyer and the potential jurors, all three groups had the same sample size—a **balanced sample**. This is not necessary for ANOVA—the groups can have different sizes.

Third, the spreadsheet calculates the **residual variance**, which is the variability due to all other sources than the treatments. This is the column variance subtracted from the pooled variance.

In addition to the above items, it calculates the **degrees of freedom** for each quantity. For the total variance, this is the size of the pooled

sample minus one. For the column effects, it is the number of treatment groups minus one. For the residual variance, it's the difference between these two numbers.

The above all needed to calculate the test statistic for ANOVA:

$$F = \frac{v_C / (\text{degrees of freedom for } v_C)}{v_R / (\text{degrees of freedom for } v_R)}$$

Similar to our initial motivating example, the F statistic compares two sources of variation in the data:

- Variation due to differences in the columns; and
- Variation due to residual effects.

Since the spreadsheet does all of these calculations for you, they are included here just for completeness.

Note that the spreadsheet also has a second table using "Sum of Squares" and "Mean Sum of Squares" statistic. The values for SS_{Total} and SS_{Between} are calculated as before, while

$$SS_{\text{Within}} = SS_{\text{Total}} - SS_{\text{Between}}$$

The degrees of freedom are

SS_{Total}	total sample size - 1
SS_{Between}	number of groups - 1
SS_{Within}	difference of the above

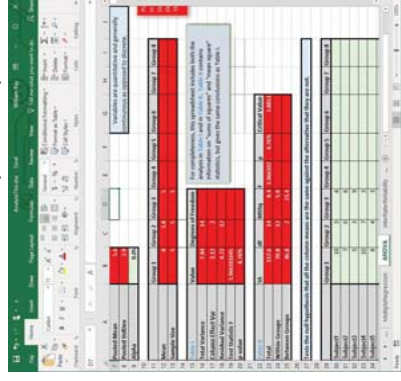
while the Mean Sum of Squares column $MSSq$ is the Sum of Squares SS divided by the corresponding degrees of Freedom df . The F Statistics F is

$$\frac{MSSq_{\text{Between Groups}}}{MSSq_{\text{Within Groups}}}$$

It turns out that the two approaches to calculating the F statistic are equivalent. The "sum of squares" approach takes somewhat better advantage of the precision of the computer and is slightly faster on large data sets, which is why it's the preferred method. Thus, the F statistic, the p -value, and the conclusions are the same in this both tables.

The good news is that the spreadsheet does all of these calculations for you. You need to understand conceptually that the test is comparing sources of variability in the data and be able to interpret the p -value.

Finally, the AnalyzeThis spreadsheet produces the same results, but starts with the raw scores. The final numbers may be slightly different since AnalyzeThis uses the full precision of the computer while we rounded the test averages and pooled standard deviation in creating the Formulas version.



20.4. Example.

A researcher is interested whether the presence of pets will influence the socialization of tenants in long-term care facilities. She selects four group homes and chooses a random sample of tenants from each home as follows:

Group 1	10 tenants	group areas include dogs
Group 2	7 tenants	group areas in include cats
Group 3	15 tenants	both cats and dogs
Group 4	13 tenants	no pets

After four weeks, the researcher scores each subject on a standardized socialization scale and obtains the following results.

	Group 1	Group 2	Group 3	Group 4
	67	69	80	64
	78	65	78	81
	82	83	62	57
	79	66	67	65
	84	64	74	84
	100	69	71	80
	83	71	87	78
	75	78	78	66
	73	70	70	71
	75	86	86	74
			72	72
			76	69
			59	58
			81	
			71	

Can the researcher conclude that there is a difference in the socialization between the three groups?

Solution.

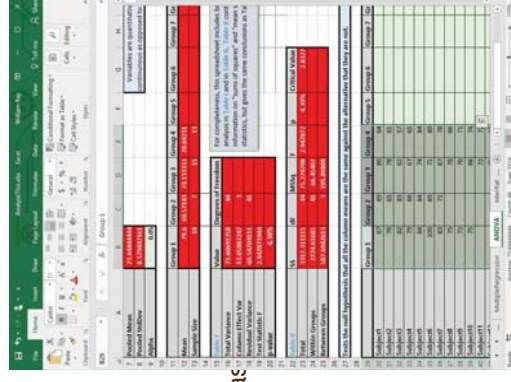
20.5. Example.

Researchers are interested in the relationship, if any, between alcohol consumption and blood pressure. A group of 32 non-smoking females, aged 20-65, were divided into four groups as follows

- Group A: no alcohol consumption for two weeks;
- Group B: 13 ml of red wine daily for two weeks;
- Group C: 13 ml of non-alcoholic red wine daily for two weeks;
- Group D: 1125 ml of beer daily for two weeks.

	Group A	Group B	Group C	Group D
change in blood pressure	-0.1	1.9 mm	1.9 mm	2.9 mm
# of subjects	8	8	9	9

For this sample, the pooled standard deviation was 1.21 mm. At the 5% significance level, do the above data show that there is a difference in the blood pressure response times between groups?



Notice that in this problem we have the raw data instead of summary data. Using the [AnalyzeThis](#) spreadsheet makes this easy—all we have to do is enter the data. See also the spreadsheet [ANOVAExample.XLSX](#), which has the data pre-loaded. Since the p -value is 4.39%, we have significant but not highly significant evidence that the treatment groups have different socialization levels.

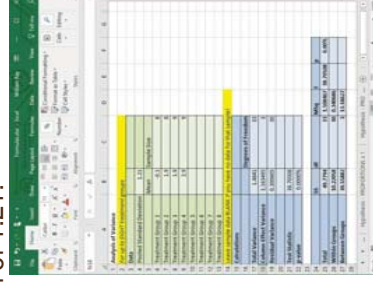
Solution.

Step 1. The above table provides the summary data for four treatment groups and for a pooled standard deviation of 1.21.

Step 2. Now enter the summary data and the pooled standard deviation into the spreadsheet [FORMULAS.XLSX](#), found in the resources section for this course on [LEARN.OU.EDU](#). Note that you will need to select the tab at the bottom labeled ANOVA.

Step 3. Read the significance level from the spreadsheet. If it's less than the pre-set target, then we reject the null hypothesis, otherwise we accept the null hypothesis.

In this case, the pre-set significance level was 5% and the p -value is .0000%, so we reject the null hypothesis. The chance of a Type I Error



is less than 0.0000%.

