

## ANOVA ANALYSIS

ANOVA analysis is a technique used to determine whether differences in two or more data sets are statistically significant. An example will be considered later on. A few more basics of statistics need to be covered before ANOVA can be discussed.

### STANDARD DEVIATION (OR SUM OF SQUARES DEVIATION OR ROOT MEAN SQUARE DEVIATION)

The standard deviation is a measurement of how much data deviates from an average value. Let's say an intern were to measure the temperature of water leaving a power plant multiple times to assure an accurate answer. The intern would then calculate the average and report the value. The average value does not say anything about the variation in the data. Compare the two data sets below; the average for both is 70 °F:

Set #1	72	72	70	67	70	69
Set #2	65	80	67	68	73	67

The first data set clearly has a lower variation in the data than the second despite identical averages. A way to account for this variation starts by considering the sum of the squared errors:

$$SSE = \sum_i (y_i - y_{ave})^2$$

In this case, the error is a deviation from the average value rather than a curve fit as was seen in the previous tutorial. Next the  $SSE$  can be divided by the number of data points ( $n$ ) minus one to obtain the variance denoted as  $\sigma^2$ :

$$\sigma^2 = \frac{SSE}{n - 1} = \frac{\sum_i (y_i - y_{ave})^2}{n - 1}$$

Then the square root can be taken of the variance to yield the standard deviation which is denoted by  $\sigma$ .

$$\sigma = \left( \frac{SSE}{n - 1} \right)^{1/2} = \left( \frac{\sum_i (y_i - y_{ave})^2}{n - 1} \right)^{1/2}$$

Another way to express the standard deviation is given below, where the average was not calculated as an intermediate step as would need to be done for the above equation:

$$\sigma = \left[ \frac{1}{n - 1} \left( \sum_i y_i^2 - \frac{(\sum_i y_i)^2}{n} \right) \right]^{1/2}$$

Either equation will work; the choice is up to the user. Students will notice that in other contexts the variance will be the *SSE* divided by  $n$  rather than  $n-1$ . The difference between the two is due to sample size. If  $n$  is small ( $<20$ ), the *SSE* should be divided by  $n-1$ . When  $n$  is sufficiently large, the *SSE* can be divided by  $n$  to obtain the variance. Usually engineers will have data sets that are small, so most engineers will simply use the  $n-1$  all of the time. When  $n$  is sufficiently large the error dividing by 99 instead of 100, for example, if  $n = 100$ , is quite small. As an engineer you should never mention this to a statistician who will probably overreact to your use of the variance and standard deviation.

The standard deviation is extremely helpful in calculating uncertainties in instruments. In laboratories, students will often notice that temperature measurements devices are within  $0.1^{\circ}\text{C}$  of the actual answer denoted as a plus/minus sign. One method to calculate this uncertainty is to take as many measurements as possible of the same thing with the instrument. For example, the company would have the intern measure the temperature of the water leaving the power plant a thousand times, ideally over a time period when the temperature does not change. The intern could calculate the standard deviation and find an interesting trend. For many sets of data, 68% of the measurements or 680 of the one thousand lie in between one standard deviation above and one standard deviation below the average value. This 68% is almost always exactly correct. The intern would then find that 95% of the data points or 950 lie within 2 standard deviations above and below the average value. Finally, an ambitious intern would find that 99% of the data points or 990 lie within 2.57 standard deviations of the average. It is very common for instrument manufacturers to use 2 or 2.57 times the standard deviation as the uncertainty in their instrument, because they include 95% and 99% of the data points respectively.

### ANOVA Analysis

Now that there is an understanding that the average by itself is not sufficient to truly describe a data set, other effects of the variation can be observed. ANOVA stands for the ANalysis Of VAriance. It can be used to see if a data set is significantly different from another or many others. Just knowing the average and standard deviation does not make it possible to determine significance. Given as a recipe, ANOVA involves roughly 6 steps:

- 1) Calculation of sum of squared errors
- 2) Calculation of degrees of freedom
- 3) Calculation of mean squares
- 4) Calculation of the F ratio
- 5) Determination of the critical F ratio
- 6) Determination of significance

For process engineers, ANOVA analysis is used to determine whether or not changing an input to a process really did have an effect, hopefully an improvement, on the process.

There are many possible real life examples. In class, the example of a beam failing a stress test was given based on extra curing time of the plastic wood. An alternative change in input could have been the temperature at which the plastic wood was cured. The only requirement for ANOVA is that there are sets of data for at least two different “treatments”. In the first step, the sum of the squared errors (*SSEs*) are calculated. The first *SSE* is for a combined pool of all of the data together denoted as  $SSE_{total}$ :

$$SSE_{total} = \sum_i y_{i,alldata}^2 - \frac{(\sum_i y_{i,alldata})^2}{N}$$

The  $N$  is the total number of data points if all of the data is pooled together. Then an *SSE* is calculated for the “treatment”. This is the amount of error or variance in the pooled data minus the error/variance from each data set individually. (Here error and variance can be considered interchangeable.)

$$SSE_{treatment} = SSE_{total} - SSE_{set\#1} - SSE_{set\#2}$$

If the equation for the *SSE* is plugged in for each of these the sum of  $y^2$  terms will cancel out leaving only the following:

$$SSE_{treatment} = \frac{(\sum_i y_{i,set\#1})^2}{n_{set\#1}} + \frac{(\sum_i y_{i,set\#2})^2}{n_{set\#2}} - \frac{(\sum_i y_{i,alldata})^2}{N}$$

This *SSE* accounts for the variance due to mixing the two data sets. Finally, an *SSE* is taken for the variance associated with the error in each data set individually which is called the “within”.

$$SSE_{within} = SSE_{total} - SSE_{treatment}$$

The next step is to calculate the degrees of freedom for the each of the three scenarios, “total”, “treatment”, and “within”. For the total, the degrees of freedom are equal to the number of total data points minus one.

$$dof_{total} = N - 1$$

The degrees of freedom for the “treatment” are equal to the number of different types of data sets ( $T$ ) minus one. For this class, scenarios with 2 treatments will be the only ones to be considered.

$$dof_{treatment} = T - 1 = 1$$

Finally the degrees of freedom for the “within” scenario are equal to the difference of the two above.

$$dof_{within} = dof_{total} - dof_{treatment}$$

The third step is to calculate the mean squares for the “treatment” and the “within”. They are given below:

$$MS_{treatment} = SS_{treatment}/dof_{treatment}$$

$$MS_{within} = SS_{within}/dof_{within}$$

The fourth step is to calculate the  $F$  ratio for the data.

$$F = MS_{treatment}/MS_{within}$$

The fifth step is to take the total degrees of freedom and the treatment degrees of freedom to the chart at the end of this tutorial. By lining up the column with the proper total degrees of freedom and the row with the proper treatment degrees of freedom, one can find the critical  $F$  ratio,  $F_{critical}$ .

Finally, if the  $F$  ratio calculated is smaller than the  $F_{critical}$ , then the two data sets are not significantly different. If the  $F$  ratio is larger than  $F_{critical}$ , then the two data sets are significantly different.

If the data sets are significantly different, then congratulations are deserved on devising a useful change to the process. If they are not significantly different, then either more data must be collected to confirm significance or the change is in fact not significant at all.

#### FINAL NOTE

Much of the derivation has been left out on purpose. In addition, the jargon such as “within” and “treatment” has been left with little explanation. The normal order of courses to learn this material thoroughly would be to take Calculus I, Calculus II, Calculus based Statistics, and then ANOVA would be covered in an advanced stats course. No attempt to cover all 4 courses will be made in ES 111. The ambitious student is encouraged to learn more about this analysis outside of this course.

Another reason to leave the detailed description in the shed is to prepare the student for “corporate culture”, which involves following recipes like the one given above. There may be other more appropriate techniques to determine significance and learning all of them may be a bit much to expect. However, a company’s culture may require using a recipe as the one given above to give data that is meaningful to upper management. It is a good idea to understand the method, but questioning its use should come from someone high up in the company and not an intern nor a new hire.

An example of MATLAB code is left out on purpose. The code required to carry out ANOVA is similar to the previous code used for statistics/linear regression. The student is all alone on this one in developing code.

The top row is the treatment degrees of freedom and the left hand column is the total degrees of freedom. The numbers in the chart are the critical F value.

df2\df1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.76	8.74	8.73	8.71	8.70	8.69	8.68	8.67	8.67	8.66
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.94	5.91	5.89	5.87	5.86	5.84	5.83	5.82	5.81	5.80
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.70	4.68	4.66	4.64	4.62	4.60	4.59	4.58	4.57	4.56
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.03	4.00	3.98	3.96	3.94	3.92	3.91	3.90	3.88	3.87
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.60	3.57	3.55	3.53	3.51	3.49	3.48	3.47	3.46	3.44
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.31	3.28	3.26	3.24	3.22	3.20	3.19	3.17	3.16	3.15
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.10	3.07	3.05	3.03	3.01	2.99	2.97	2.96	2.95	2.94
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.94	2.91	2.89	2.86	2.85	2.83	2.81	2.80	2.79	2.77
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.82	2.79	2.76	2.74	2.72	2.70	2.69	2.67	2.66	2.65
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.72	2.69	2.66	2.64	2.62	2.60	2.58	2.57	2.56	2.54
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.63	2.60	2.58	2.55	2.53	2.51	2.50	2.48	2.47	2.46
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.57	2.53	2.51	2.48	2.46	2.44	2.43	2.41	2.40	2.39
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.51	2.48	2.45	2.42	2.40	2.38	2.37	2.35	2.34	2.33
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.46	2.42	2.40	2.37	2.35	2.33	2.32	2.30	2.29	2.28
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.41	2.38	2.35	2.33	2.31	2.29	2.27	2.26	2.24	2.23
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.37	2.34	2.31	2.29	2.27	2.25	2.23	2.22	2.20	2.19
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.34	2.31	2.28	2.26	2.23	2.21	2.20	2.18	2.17	2.16
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.31	2.28	2.25	2.23	2.20	2.18	2.17	2.15	2.14	2.12
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.26	2.23	2.20	2.17	2.15	2.13	2.11	2.10	2.08	2.07
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.22	2.18	2.15	2.13	2.11	2.09	2.07	2.05	2.04	2.03
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.18	2.15	2.12	2.09	2.07	2.05	2.03	2.02	2.00	1.99
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.15	2.12	2.09	2.06	2.04	2.02	2.00	1.99	1.97	1.96
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.13	2.09	2.06	2.04	2.01	1.99	1.98	1.96	1.95	1.93
35	4.12	3.27	2.87	2.64	2.49	2.37	2.29	2.22	2.16	2.11	2.08	2.04	2.01	1.99	1.96	1.94	1.92	1.91	1.89	1.88
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.04	2.00	1.97	1.95	1.92	1.90	1.89	1.87	1.85	1.84
45	4.06	3.20	2.81	2.58	2.42	2.31	2.22	2.15	2.10	2.05	2.01	1.97	1.94	1.92	1.89	1.87	1.86	1.84	1.82	1.81
50	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07	2.03	1.99	1.95	1.92	1.89	1.87	1.85	1.83	1.81	1.80	1.78
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.95	1.92	1.89	1.86	1.84	1.82	1.80	1.78	1.76	1.75
70	3.98	3.13	2.74	2.50	2.35	2.23	2.14	2.07	2.02	1.97	1.93	1.89	1.86	1.84	1.81	1.79	1.77	1.75	1.74	1.72