11/1/2022

ANOVA – Analysis of Variance

There are many situations in which an ANOVA test is appropriate, but we are going to start with the simplest case. One-way ANOVA is an extension, in some sense of the comparison of means for two samples, but now we are comparing three or more samples with each other.

One possible strategy for dealing with three or more samples is to compare them in pairs. In general, this would mean we'd have to conduct $\binom{N}{2}$ pairs of comparisons. For three samples, this is three pairs, but if there are 5 samples, this is 10 pairs and so on (N is the number of samples/populations). The ANOVA allows us to compare them all at once. We can test whether any of the means of the samples are different from each other. If they are not, there is no need to do the pairwise comparisons. If they are different, then it may be worth it to test further. After conducting an example test or two, we'll discuss Tukey's method for identifying which of the samples is different, or how they might group together.

Despite the name ANOVA, it is a test of means, and not variances, per se. The null hypothesis for the ANOVA is that all the means are the same. The alternative is that at least one mean is different (they do not all have to be different). We can express these hypotheses mathematically as follows:

$$H_0: \mu_i = \mu_j, \forall i \neq j$$
$$H_a: \mu_i \neq \mu_j \; for \; some \; i \neq j$$

Some terminology it is worth being aware of: the conditions under which the different samples are collected are sometimes referred to as treatments. In R, we will need to treat them as factors, even if they have numerical representations. It can also be useful to look at a comparative box plot of your samples to develop some intuition of what to expect.

We make certain assumptions when we perform ANOVA tests. The first is that the data is independent, similar to our independent two-sample t-test. We also assume that the data for each sample is normally distributed. And the variance of each group is the same. These are some of the features you can look for in your boxplot. Does the data for each group appear symmetric? Do the boxes for your boxplots appear to be about the same size? If these assumptions are violated, we may need to employ a different type of test (perhaps a non-parametric one) which we will look at in a future lecture.

An ANOVA test uses an F statistic. This is essentially a ratio of different types of variability.

$$F = \frac{MSTr}{MSE}$$

The MSTr is called the Mean Square for Treatments. It is the between sample variation. The MSE is the Mean Square Error and is the within sample variation.

The formula for MSTr is

$$MSTr = \frac{J}{I-1}\sum_{i}^{I}(\bar{X}_{i\cdot} - \bar{X}_{\cdot\cdot})^2$$

The variable I is the number of treatments, and J is the number of observations in each treatment (we assume here that the J is the same, but this need not be the case). The expression $\bar{x}_{i\cdot}$ is the mean within each ith sample. The variable $\bar{x}_{\cdot\cdot}$ is the mean of all the observations combined (referred to as the grand mean).

The formula for MSE is

$$MSE = \frac{1}{I}\sum_{i}^{I} S_i^2$$

The s_i^2 are just the variance for each treatment group.

These formulas are useful for seeing how the meaning is derived, but these calculations, like other variance calculations have shortcuts.

$$\text{SST} = \sum_{i=1}^{I}\sum_{j=1}^{J}(x_{ij} - \bar{x}_{\cdot\cdot})^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}x_{ij}^2 - \frac{1}{IJ}x_{\cdot\cdot}^2$$

$$\text{SSTr} = \sum_{i=1}^{I}\sum_{j=1}^{J}(\bar{x}_{i\cdot} - \bar{x}_{\cdot\cdot})^2 = \frac{1}{J}\sum_{i=1}^{I}x_{i\cdot}^2 - \frac{1}{IJ}x_{\cdot\cdot}^2$$

$$\text{SSE} = \sum_{i=1}^{I}\sum_{j=1}^{J}(x_{ij} - \bar{x}_{i\cdot})^2 \quad \text{where } x_{i\cdot} = \sum_{j=1}^{J}x_{ij} \quad x_{\cdot\cdot} = \sum_{i=1}^{I}\sum_{j=1}^{J}x_{ij}$$

When ANOVA is calculated, these formulas are usually used instead. The relationship to our F-statistic is below.

$$\text{MSTr} = \frac{\text{SSTr}}{I-1} \quad \text{MSE} = \frac{\text{SSE}}{I(J-1)} \quad F = \frac{\text{MSTr}}{\text{MSE}}$$

ANOVA output from statistical software typically also outputs these values, and the corresponding degrees of freedom, similar to the table shown below.

Source of Variation	df	Sum of Squares	Mean Square	*f*
Treatments	$I-1$	SSTr	MSTr = SSTr/$(I-1)$	MSTr/MSE
Error	$I(J-1)$	SSE	MSE = SSE/$[I(J-1)]$	
Total	$IJ-1$	SST		

Output sometimes will also include (off to the right) the F critical value for the significance level, and the P-value that corresponds to our F-statistic.

These calculations are quite complex, even for very small samples, so we will basically always use technology to perform these calculations.

We are going to walk through one small example just one time better understand the formulas.

Example.
Data was collected on three different methods of treating cotton fabric to prevent staining. The data collected is shown below.

Treatment 1	0.56	1.12	0.90	1.07	0.94
Treatment 2	0.72	0.69	0.87	0.78	0.91
Treatment 3	0.62	1.08	1.07	0.99	0.93

Our hypotheses are:

$$H_0: \mu_i = \mu_j, \forall i \neq j$$
$$H_a: \mu_i \neq \mu_j \; for \; some \; i \neq j$$

The means of our three treatments are, respectively, $0.918, 0.794, 0.938$. The sums of observations of each treatment are respectively $4.59, 3.97, 4.69$. The sum of all the observations is 13.25. We also have $I = 3, I - 1 = 2, J = 5, J - 1 = 4, IJ = 15, I(J - 1) = 12$.

Let's start with the $SST =$

$$\sum_{i=1}^{I}\sum_{j=1}^{J} x_{ij}^2 - \frac{1}{IJ}x_{\cdot\cdot}^2$$

$$\sum\sum x_{ij}^2 = (.56)^2 + (1.12)^2 + \cdots + (.93)^2 = 12.1351$$

$$\text{SST} = 12.1351 - (13.25)^2/15 = 12.1351 - 11.7042 = .4309$$

Then, we calculate $SSTr =$

$$\frac{1}{J}\sum_{i=1}^{I} x_{i\cdot}^2 - \frac{1}{IJ}x_{\cdot\cdot}^2$$

$$\text{SSTr} = \frac{1}{5}[(4.59)^2 + (3.97)^2 + (4.69)^2] - 11.7042$$
$$= 11.7650 - 11.7042 = .0608$$

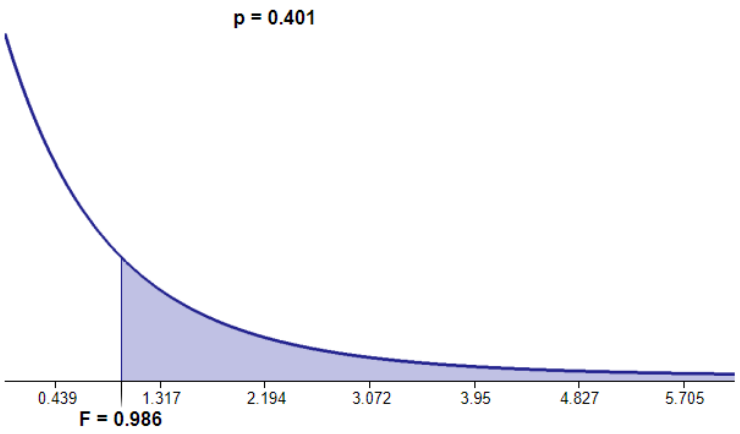
The SSE is the difference between the SST and SSTr:

$$SSE = .4309 - .0608 = .3701$$

We can complete our ANOVA table and complete the test.

Source of Variation	df	Sum of Squares	Mean Square	f
Treatments	$I - 1$	SSTr	MSTr = SSTr/$(I - 1)$	MSTr/MSE
Error	$I(J - 1)$	SSE	MSE = SSE/$[I(J - 1)]$	
Total	$IJ - 1$	SST		

SOURCE OF VARIATION	DF	SUM OF SQUARES	MEAN SQUARE	F
TREATMENTS	2	0.0608	0.0304	0.986
ERROR	12	0.3701	0.03084	
TOTAL	14	0.4309		

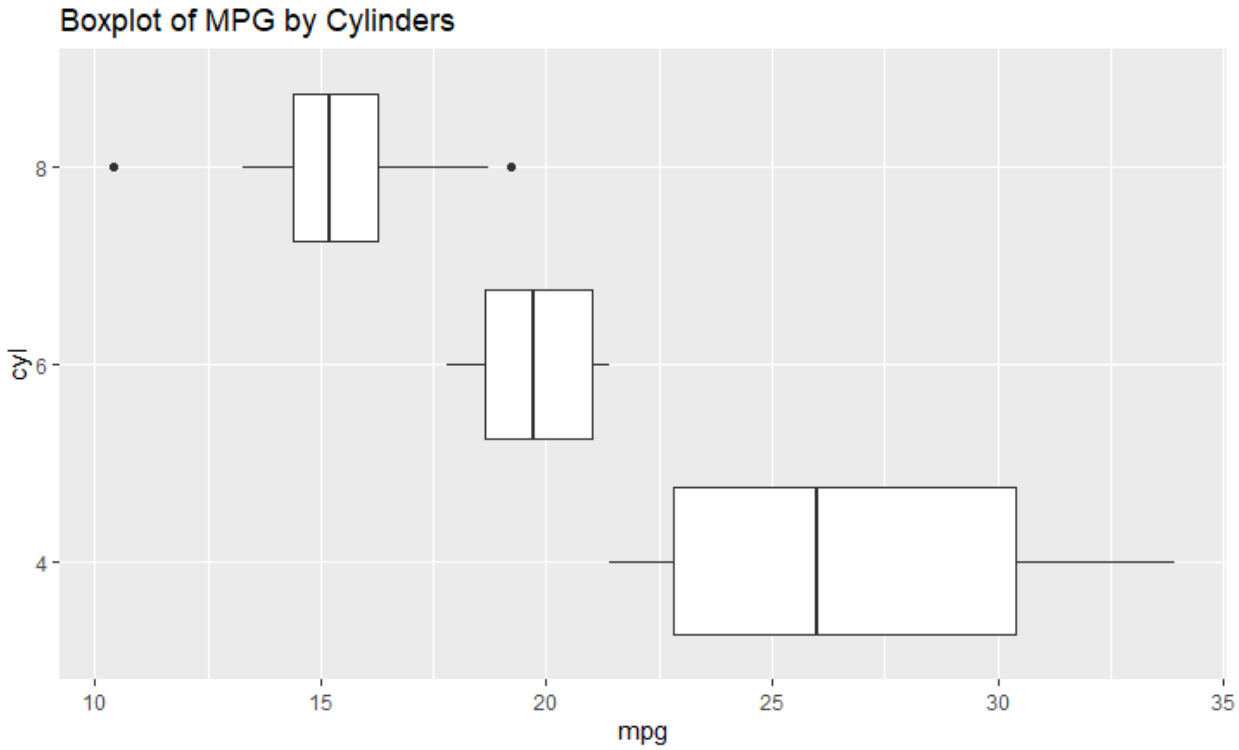


The P-value is about 0.401>>0.05 or any other common significance level. So, we fail to reject the null hypothesis. Thus, we don't have enough evidence to think that the means are different.

To apply the next method, to determine which groups are different, we are going to need to have an ANOVA test that detects a difference. Below, is the ANOVA table results from R, testing whether the number of cylinders in an engine impacts gas mileage (from the mtcars data).

```
            Df Sum Sq Mean Sq F value   Pr(>F)
cyl          2  824.8   412.4    39.7 4.98e-09 ***
Residuals   29  301.3    10.4
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can look at a boxplot to confirm these results.



The mpg for the vehicle groups do appear to be different. What we want to statistically establish now is how are these groups different. The two outside groups overlap somewhat with the middle group. Are we saying that all the groups are different from each other, or are we saying that the two outside groups are different but neither is different from the middle group? As noted before, we could do pairwise t-tests to perform the pairwise comparisons. However, we have another method, Tukey's method, that essentially constructs confidence intervals around each pair of comparisons, so that we can see how they overlap, if at all. We'll describe the method broadly below, and then we'll run the computation in R.

Tukey's method

When we construct confidence intervals to compare two means, we center the interval at the difference of two means. If 0 is included in the interval, there is no statistically significant difference, but if zero is not included, then the means are statistically significantly different.

To construct this confidence interval, however, Tukey does not use the t-distribution. Instead, this method uses what is called a Studentized range distribution, the notation for which is usually Q (or q). The distribution depends on the range for the set, the variance and the degrees of freedom.

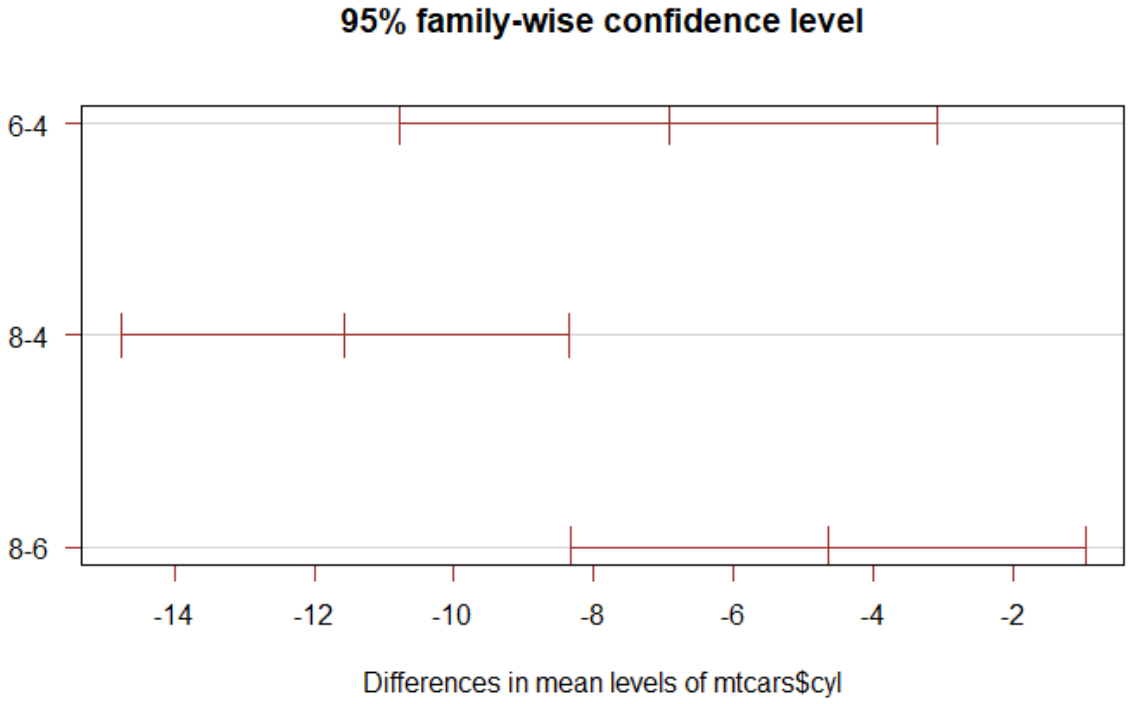
$$Q_{r,v} = \frac{range}{s}$$

Textbooks that cover Tukey's method usually provide a table of these values for various degrees of freedom and common confidence levels. This distribution is included in R. Once we find the appropriate value of Q for our data, we calculate the confidence intervals using the formula

$$\bar{y}_{i\bullet} - \bar{y}_{j\bullet} \pm \frac{1}{\sqrt{2}} q_{\alpha;\, r,\, N-r}\ \hat{\sigma}_{\epsilon} \sqrt{\frac{2}{n}} \quad i, j = 1, \ldots, r;\ i \neq j.$$

This formula assumes that sample sizes are equal.

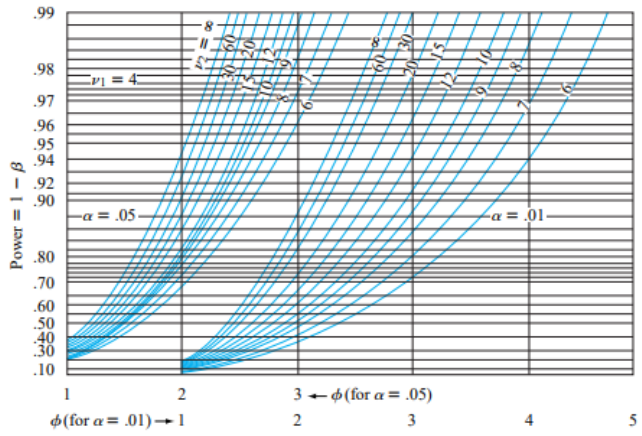
You won't be asked to calculate these intervals by hand. Instead, we can calculate them in R using the TukeyHSD() function. We can plot the intervals to see how they compare. A plot of the intervals for our ANOVA test are shown.



Notice that none of the intervals include 0, which tells us that all the samples are different from each other. The largest difference is between 4 cylinders and 8 cylinders, which agrees with our assessment of the boxplot.

Next semester we will discuss how ANOVA generates a model of the treatments, similar to applying dummy variables to the treatment levels, with some treatment effect from the overall mean.

Power calculations for ANOVA are generally built in to most statistical software because the calculations are quite difficult by hand (moreso than ANOVA itself). When technology is not available, statisticians have constructed power curves as shown in the graph below.



These curves differ for different degrees of freedom, different levels of alpha, and different sample sizes. For our purposes, it's worth knowing that these graphs exist, but we will depend on technology to perform these calculations for us rather than interpreting these types of graphs.

We looked at an example of calculating an ANOVA test with equal sample sizes. But it's not required that sample sizes must be equal. Shown below are the more general ANOVA formulas that can be used for unequal sample sizes in each treatment, and the corresponding Tukey's method formula. These are the formulas that are built into R.

$$\text{SST} = \sum_{i=1}^{I}\sum_{j=1}^{J_i}(X_{ij} - \overline{X}_{\cdot\cdot})^2 = \sum_{i=1}^{I}\sum_{j=1}^{J_i}X_{ij}^2 - \frac{1}{n}X_{\cdot\cdot}^2 \qquad \text{df} = n - 1$$

$$\text{SSTr} = \sum_{i=1}^{I}\sum_{j=1}^{J_i}(\overline{X}_{i\cdot} - \overline{X}_{\cdot\cdot})^2 = \sum_{i=1}^{I}\frac{1}{J_i}X_{i\cdot}^2 - \frac{1}{n}X_{\cdot\cdot}^2 \quad \text{df} = I - 1$$

$$\text{SSE} = \sum_{i=1}^{I}\sum_{j=1}^{J_i}(X_{ij} - \overline{X}_{i\cdot})^2 = \text{SST} - \text{SSTr} \quad \text{df} = \sum(J_i - 1) = n - I$$

Test statistic value:

$$f = \frac{\text{MSTr}}{\text{MSE}} \quad \text{where MSTr} = \frac{\text{SSTr}}{I - 1} \quad \text{MSE} = \frac{\text{SSE}}{n - I}$$

Rejection region: $f \geq F_{\alpha,I-1,n-I}$

$$w_{ij} = Q_{\alpha,I,n-I} \cdot \sqrt{\frac{\text{MSE}}{2}\left(\frac{1}{J_i} + \frac{1}{J_j}\right)}$$

Then the probability is *approximately* $1 - \alpha$ that

$$\overline{X}_{i\cdot} - \overline{X}_{j\cdot} - w_{ij} \leq \mu_i - \mu_j \leq \overline{X}_{i\cdot} - \overline{X}_{j\cdot} + w_{ij}$$

for every i and j ($i = 1, \ldots, I$ and $j = 1, \ldots, I$) with $i \neq j$.

Our main goal for ANOVAs, especially as they become more complex will be interpretation. You won't be expected to perform these calculations by hand.

In the next lecture, we'll talk about doing comparisons with more than one categorical variable.

References:

1. https://faculty.ksu.edu.sa/sites/default/files/probability_and_statistics_for_engineering_and_the_sciences.pdf
2. https://assets.openstax.org/oscms-prodcms/media/documents/IntroductoryStatistics-OP_i6tAI7e.pdf
3. http://www.statdistributions.com/f/
4. https://www.itl.nist.gov/div898/handbook/prc/section4/prc471.htm
5. https://r-graph-gallery.com/84-tukey-test.html
6. https://www.scribbr.com/statistics/anova-in-r/
7. https://stat.ethz.ch/R-manual/R-devel/library/stats/html/Tukey.html