Data Analysis and Statistical Methods
Statistics 651

http://www.stat.tamu.edu/~suhasini/teaching.html

Lecture 24 (MWF) Why ANOVA works and checking assumptions

Suhasini Subba Rao
Lecture 24 (MWF) Assumptions of ANOVA

Why does the ANOVA work?

• We see if the null is true, that is \( \mu_1 = \mu_2 = \mu_3 \), then \( \frac{SSB}{3-1} \) is an estimator of

\[
\frac{SSB}{2} \approx \sigma^2 + \frac{1}{3-1} \left( 4 \times (\mu_1 - \mu)^2 + 3 \times (\mu_2 - \mu)^2 + 4 \times (\mu_1 - \mu)^2 \right),
\]

where \( \mu = \frac{1}{11}(4\mu_1 + 3\mu_2 + 4\mu_3) = \mu \). Hence \( \frac{SSB}{3-1} \) is an estimator of \( \sigma^2 \).

Since \( \frac{SSW}{11-3} \) and \( \frac{SSB}{3-1} \) are both estimators of the variance. This means under the null the F-statistic

\[
F = \frac{SSB/2}{SSW/8}
\]

is close to 1 (\( F \) will be ‘small’).
• On the other hand, if the alternative is true, and at least one of the means are different, then the F-statistic is estimating

\[ 1 + \frac{1}{2\sigma^2} \left( 4 \times (\mu_1 - \mu)^2 + 3 \times (\mu_2 - \mu)^2 + 4 \times (\mu_1 - \mu)^2 \right), \]

where \( \mu = \frac{1}{11}(4\mu_1 + 3\mu_2 + 4\mu_3) \).

• In other words, this ratio will be much bigger than one. Moreover the bigger the difference \((\mu_k - \mu)^2\) the further \(\frac{SSB}{3-1}\) will be from \(\sigma^2\) and the bigger the ratio \(\frac{SSB}{3-1}/\frac{SSW}{11-3}\) will be.

• This is why when \(\frac{SSB}{3-1}/\frac{SSW}{11-3}\) is large we reject the null (since it difficult to get a large \(F\) by random chance - with all the means being the same). This pushes the F statistic to the right and makes it likely that the alternative can be detected.
The Assumptions of an ANOVA

• To do an ANOVA we need that that the observations are close normal (especially if the sample size is small, if the sample sizes are quite large this does not matter as much).

  *We explain how to check this below.*

• All the observations are completely independent of each other (just like the independent sample t-test).

  *Usually this is quite difficult to check. Typically, this can be determine done by considering how the data was collected.*

• The variations within each of the groups is roughly the same.

  *The Levene test can be used to determine if the variance between two populations is the same. It is more difficult for multiple populations. This assumptions is quite robust to differences.*
Checking the Normality assumption

• Analogous to the independent sample t-test we make a QQplot of the residuals.

• In order to define the notion of a residual, we need to understand that there is a ‘model’ behind what we are doing.

• In order to do an ANOVA. We write $X_{ij}$ as

$$X_{ij} = \mu_i + (X_{ij} - \mu_i)$$

$$= \mu_i + \varepsilon_{ij},$$

where $\varepsilon_{ij}$ is normally distributed with mean zero and variance $\sigma^2$ (same
variance for all observations). The ANOVA tests $\mu_i = \mu$ for all $i$ (mean is same).

Example: Recall the M&M example. We can write the number of peanut/peanut butter and plain M&Ms as:

\[
X_{\text{peanut}} = \mu_{\text{peanut}} + \varepsilon_{\text{peanut}}
\]
number of peanut M&Ms mean number of peanut M&Ms random variation between bags

\[
X_{\text{plain}} = \mu_{\text{plain}} + \varepsilon_{\text{plain}}
\]
number of plain M&Ms mean number of plain M&Ms random variation between bags

\[
X_{\text{peanut butter}} = \mu_{\text{peanut butter}} + \varepsilon_{\text{peanut butter}}
\]
number of peanut butter M&Ms mean peanut butter M&Ms random variation between bags
• To check the normality assumption we first estimate the residuals from the data:

$$\bar{\varepsilon}_{ij} = X_{ij} - \bar{X}_i,$$

this is the difference between what we observe and the average of the $i$th sample.

Example:

$$\bar{\varepsilon}_{\text{peanut}} = X_{\text{peanut}} - \bar{X}_{\text{peanut in sample}}$$

• The residuals, $\bar{\varepsilon}_{ij}$, have a zero mean. We **never** do an ANOVA on the residuals! But we can use the residuals to make a QQplot.

• We plot a QQ-plot for the residuals. That is a QQ plot of $\{\hat{\varepsilon}_{ij}; i = 1, \ldots, k, j = 1, \ldots, n_i\}$. 
• If the QQplot looks close to normal than we can use an ANOVA.

• However it is interesting to note that ANOVA is extremely robust to departures from normality. Only when the data is really, really far from normal or the variances are really, really different should we look for alternative tests (for example the Kruskal Wallis - see the end of these notes).
Dummy Example: calculating the residuals

- Recall the dummy data we looked at last time...

- We now use the sample means to calculate the residuals.

<table>
<thead>
<tr>
<th>Sample 1</th>
<th>Sample1 Res</th>
<th>Sample 2</th>
<th>Sample 2 Res</th>
<th>Sample 3</th>
<th>Sample 3 Res</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1</td>
<td>4.1-3.2=0.9</td>
<td>5.1</td>
<td>5.1-5.13=-0.03</td>
<td>6.6</td>
<td>-0.05</td>
</tr>
<tr>
<td>3.3</td>
<td>3.3-3.2=0.1</td>
<td>5.0</td>
<td>5.0-5.13=-0.13</td>
<td>6.2</td>
<td>-0.45</td>
</tr>
<tr>
<td>2.6</td>
<td>2.6-3.2=-0.6</td>
<td>5.3</td>
<td>5.3-5.13=0.27</td>
<td>7.3</td>
<td>0.65</td>
</tr>
<tr>
<td>2.8</td>
<td>2.8-3.2=-0.4</td>
<td>5.13</td>
<td></td>
<td>6.5</td>
<td>-0.15</td>
</tr>
</tbody>
</table>

- Make a QQplot of the residuals to check for normality.
The sample size is quite small, but there does not appear to be a large deviation from normality. Therefore the p-values obtained using an ANOVA will be quite accurate.
To obtain the residuals in JMP. Click on red triangle at the top left hand corner (the same to do the ANOVA), in the list of options click on save and then click on Save Residuals.
Example II: Tar in Cigarettes

The mean amount of tar in medium tar cigarettes is known to be 3mg. We want to see whether the mean amounts vary from brand to brand. 4 different brands were sampled, and for each brand the amount of tar was measured in a few cigarettes. The deviation from 3mgs is reported below.

<table>
<thead>
<tr>
<th>Brand</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>size</th>
<th>Mean</th>
<th>var</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>-0.307</td>
<td>-0.294</td>
<td>0.079</td>
<td>0.019</td>
<td>-0.136</td>
<td>5</td>
<td>-0.1278</td>
<td>0.031</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>-0.176</td>
<td>0.125</td>
<td>-0.013</td>
<td>0.082</td>
<td>0.091</td>
<td>0.459</td>
<td>6</td>
<td>0.0946</td>
<td>0.043</td>
</tr>
<tr>
<td>C</td>
<td>0.137</td>
<td>-0.063</td>
<td>0.24</td>
<td>-0.05</td>
<td>0.318</td>
<td>0.154</td>
<td>6</td>
<td>0.1266</td>
<td>0.023</td>
</tr>
<tr>
<td>D</td>
<td>-0.042</td>
<td>0.69</td>
<td>0.21</td>
<td>0.166</td>
<td>0.219</td>
<td>5</td>
<td>0.2486</td>
<td>0.072</td>
<td></td>
</tr>
</tbody>
</table>

the global mean using all the samples is 0.087

• Is there any evidence to suggest that the amount of tar varies between brands? Do the test at the 1% level.
The above is basically asking whether the mean amount of tar for each of the brand is the same or not. Therefore we are testing $H_0 : \mu_A = \mu_B = \mu_C = \mu_D$ against the alternative $H_A :$ at least one mean is different.

Since the p-value of 6% is greater than the significance level of 1%
we cannot reject the null. There is not sufficient evidence in the data to suggest the mean level of tar in the different brands is different.
Checking assumptions: QQplot of the tar residuals

The sample size is quite small, but there does not appear to be a large deviation from normality. Therefore the p-values obtained using an ANOVA will be quite accurate.
Example III

The nitrogen content of three different clover plants is given below.

<table>
<thead>
<tr>
<th>3DOK1</th>
<th>3DOK5</th>
<th>3DOK7</th>
</tr>
</thead>
<tbody>
<tr>
<td>19.4</td>
<td>18.2</td>
<td>20.7</td>
</tr>
<tr>
<td>32.6</td>
<td>24.6</td>
<td>21.0</td>
</tr>
<tr>
<td>27.0</td>
<td>25.5</td>
<td>20.5</td>
</tr>
<tr>
<td>32.1</td>
<td>19.4</td>
<td>18.8</td>
</tr>
<tr>
<td>33.0</td>
<td>21.7</td>
<td>18.6</td>
</tr>
<tr>
<td></td>
<td>20.8</td>
<td>20.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>21.3</td>
</tr>
</tbody>
</table>

- Use ANOVA and the $F$ distribution to test whether all three clovers have the same population mean.
The clover analysis in JMP

- The F-statistics under the null $H_0 : \mu_1 = \mu_2 = \mu_3$ has an F-distribution with 2, 15 degrees of freedom.
• The p-value for $P(F_{2,15} > 8.2) = 0.002$. 0.002 is quite small (smaller than $\alpha = 0.05$), so there is enough evidence to reject the null.

• Equivalently, look up the tables with $\alpha = 0.05$ we have $F_{2,15}(0.05) = 3.68$. The rejection region is any number bigger than 3.68. Since $3.68 < 9.7$, there is enough evidence to reject the null.

The ANOVA table is identical to the hand calculation, and we see that the p-value is 0.002 (0.2%).
The clover residuals

A screenshot of the residuals in the JMP output.
QQplot of residuals from the plant data

The residuals are calculated (look at the numbers of in the second table) and a QQplot is made using these numbers.

[Graph showing QQplot with summary statistics]
QQplot of residuals. There is some deviation from normality, hence we have to be cautious about the results of the test (in particular the reliability of the p-value). However, the p-value is very small 0.2%, therefore even if it is not too accurate it is unlikely to be much larger than 0.2% (to change the conclusion of the test).
Bonferonni corrections

Here we consider the multiple testing issue. You do not have to learn this (it is for those who do multiple tests in their research).

• If you have $k$ groups, and you want to t-tests on all these groups there will be $k(k - 1)/2$ different tests. Thus if you do each test at the 5% level, there is a $1 - 0.95^{k(k-1)/2}$ chance you will wrongly reject the null when there is no difference between the group means.

This can be quite large, even for moderately large number groups $k$.

• To prevent this and ensure that if the null is true there is 95% chance of rejecting any of these these tests, then you need to do each test at the $\alpha/m$-level where $m = k(k - 1)/2$ is the number of tests that you are doing.
• Example: This means if there are 4 groups, then \( k = 4 \times 3/2 = 6 \) tests need to be done. Each at the 5/6%-significance level.

Therefore, we can only reject the null if one of the p-values is less than 5/6%. This means that one of the p-values have to be extremely small in order reject the null.

• Using the Bonferonni correction is very conservative (can mean that is it is very difficult to reject the null when the alternative is true).
Aside: Post-hoc tests (optional)

- Suppose we reject the null that all the populations share the same mean. How to identify which population means are different?

- There are various methods for doing this, for example, Tukey’s method or Fisher’s Least Significant Difference.

- Example: Suppose we test $H_0 : \mu_1 = \mu_2 = \mu_3$ against the alternative that one of the means is different. There are 4 possibilities that this can be:

  $\mu_1 = \mu_2 \neq \mu_3$ or $\mu_1 = \mu_3 \neq \mu_2$

  $\mu_2 = \mu_3 \neq \mu_1$ or $\mu_1 \neq \mu_2 \neq \mu_3$.

- Here we describe the LSD test, which is used to identify which population seems to be the most likely.
You do not have to learn the details.

- The idea of the test is to do a type of $t$-test for every paired combination. But using a sample variance based on the all residuals (this increases the number of degrees of freedom in the t-distribution).

- We define the mean squared error as

$$s^2_W = \frac{1}{N} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_j)^2,$$

the above is the sample variance that we will use. The eagle-eyed may have noticed that this is simply $SSW/(N - k)$. Hence the sum of squares within groups divided by $(N - k)$. 

• We use \( s^2_W \) as the estimate of the common variance \( \sigma^2 \). Noting that we have assumed the variance of all the populations are the same.

• We use the \( t \)-distribution with \( N - k \) degrees of freedom and evaluate it at \( \alpha_k,0.05/2 \).

• For every pair \((a, b)\) we make something like a non-rejection region:

\[
- t_{\alpha/2}(N - k) \sqrt{s^2_W \left( \frac{1}{n_a} + \frac{1}{n_b} \right)} ,  
 s_W t_{\alpha/2}(N - k) \sqrt{s^2_W \left( \frac{1}{n_a} + \frac{1}{n_b} \right)}
\]

• If the difference \( \bar{X}_a - \bar{X}_b \) does not lie in this region we believe there is enough evidence to suggest the means are different.
• Remember there are situations where we use an ANOVA to reject the null that all the population means are the same, but when we do a LSD method on each pair there is not enough evidence to say that there is a difference in any of the pairs.

• We do the same procedure for every pair.

• For example if we reject the null that $\mu_1$, $\mu_2$ and $\mu_3$ are not all the same, then we considered the 3 pairs (do this pseudo t-test on them):

$$ (\mu_1, \mu_2), \ (\mu_1, \mu_3), \ (\mu_2, \mu_3). $$
Output of Post hoc Multiple comparisons table

• Look at the Post Hoc output. It gives the difference for every single difference.

• The p-value is basically the p-value for the t-test calculated using

\[
\frac{\bar{X}_a - \bar{X}_b}{\sqrt{s^2_W \left( \frac{1}{n_a} + \frac{1}{n_b} \right)}} \sim t_{N-k}.
\]

All you need to do is look for p-values which are less than \( \alpha/2 \). If it is less than \( \alpha/2 \) it indicates that the means are different.
• The Confidence Interval is:

\[
\left[ \bar{X}_a - \bar{X}_b - t_{\alpha/2}(N-k) \sqrt{\frac{s_W^2}{\left( \frac{1}{n_a} + \frac{1}{n_b} \right)}} \right],
\]

\[
\bar{X}_a - \bar{X}_b + t_{\alpha/2}(N-k) \sqrt{\frac{s_W^2}{\left( \frac{1}{n_a} + \frac{1}{n_b} \right)}}
\]

• **Important** A problem with this procedure is that even if all the means were the same (for the same reasons as the multiple testing problem) with a large probability we may declare at least two populations to have different means. See Chapter 9, Ott and Longnecker for other methods.
Nonparametric Tests: Kruskal-Wallis

• Suppose we make a QQ plot of the residuals and there seems to be a very large departure from normality, then the usual Analysis of Variance method using the F-test is not appropriate.

• We note that the Analysis of variance is quite robust to departures from normality, only when there seems to be a huge heavy tail on a massive skew do we need to look for an alternative test.

• The Kruskal-Wallis is a nonparametric procedure that one can use in the extreme non-normal cases. You don’t need to learn the procedure, but for those of you who are interested its given at the end of these notes.

• It is basically a generalisation of the Wilcoxon rank sum test (which we recall tests for equality of two populations).
Aside: Method: Kruskal-Wallis

- Because we have relaxed the assumption of normality, the null is $H_0 : k$ distributions are identical against $H_A$: Not all distributions are the same (they are a shift of each other).

- As in the Wilcoxon rank sum test rank all the observations and for each sample sum the ranks. For the $i$th sample we denote the sum as $T_i$.

- We define the test statistic as

$$H = \frac{12}{N(N + 1)} \sum_{i=1}^{k} \frac{T_i^2}{n_i} - 3(N + 1),$$

where $N = \sum_{i=1}^{k} n_i$. 
• Under the null, $H$ is distributed with what is called a $\chi^2$ distribution with $(k - 1)$ degrees of freedom ($\chi^2_{k-1}$). This is a positive distribution which is related to the $F$-distribution.

• I do not expect you know this test in detail, but it is worth bearing it in mind, and the reasons why it should be used.
Aside: Dummy Example: Kruskal-Wallis

<table>
<thead>
<tr>
<th>Sample 1</th>
<th>Rank</th>
<th>Sample 2</th>
<th>Rank Res</th>
<th>Sample 3</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1</td>
<td>4</td>
<td>5.1</td>
<td>6</td>
<td>6.6</td>
<td>10</td>
</tr>
<tr>
<td>3.3</td>
<td>3</td>
<td>5.0</td>
<td>5</td>
<td>6.2</td>
<td>8</td>
</tr>
<tr>
<td>2.6</td>
<td>1</td>
<td>5.3</td>
<td>7</td>
<td>7.3</td>
<td>11</td>
</tr>
<tr>
<td>2.8</td>
<td>2</td>
<td></td>
<td></td>
<td>6.5</td>
<td>9</td>
</tr>
<tr>
<td>$T_i$</td>
<td>10</td>
<td>18</td>
<td></td>
<td></td>
<td>38</td>
</tr>
</tbody>
</table>

- Calculate the test statistic:

$$H = \frac{12}{11(11 + 1)} \left( \frac{10^2}{4} + \frac{18^2}{3} + \frac{38^2}{4} \right) = 28.6.$$ 

- If we look up the $\chi^2$ (because it is $\chi^2$ with $3 - 1$ degrees of freedom)
tables for $\alpha = 0.05$ we see that $\chi^2_2(0.05) = 7.815$. Since $28.6 > 7.815$ we reject the null.