Hypothesis Test Notes

Analysis of Variance (ANOVA)

Recall that the goodness of fit categorical data test can be used when comparing a percentage in 3 or more groups. What if we have quantitative data from 3 or more groups and want to compare the mean averages?

The answer to this is the ANOVA test.

ANOVA stands for “Analysis of Variance”

It is a favorite of statisticians because it is very versatile and can be used for comparing the means of quantitative data sets.

ANOVA Null and Alternative Hypothesis

The “one-way” ANOVA hypothesis test is used to compare 1 mean average between several groups. If you want to compare more than one mean from several groups, that is called a “Two-way ANOVA”. (We will only cover one-way ANOVA)

Example 1-Mean Average Salaries for people living in five states in Australia.

Suppose we want to compare the mean average weekly salary for people living in 5 states in Australia. (Northern Territory, New South Wales, Queensland, Victoria, and Tasmania) We think they are different. As with all multiple population hypothesis tests, you should label the populations.

\[ \mu_1 : \text{Northern Territory} \]
\[ \mu_2 : \text{New South Wales} \]
\[ \mu_3 : \text{Queensland} \]
\[ \mu_4 : \text{Victoria} \]
\[ \mu_5 : \text{Tasmania} \]

Here is the null and alternative hypothesis for the ANOVA test. Remember an ANOVA is a multiple \( \mu \) test for 3 or more groups.

\[ H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 \]
\[ H_A : \text{at least one is } \neq \text{ (claim)} \]
When doing an ANOVA test, it is good to find the sample size (n), the sample mean of each group, and the standard deviation or variance for each group. Here is the sample data from StatCrunch.

**Summary statistics:**

<table>
<thead>
<tr>
<th>Column</th>
<th>n</th>
<th>Mean</th>
<th>Std. dev.</th>
<th>Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>North Territory</td>
<td>35</td>
<td>1534.5395</td>
<td>701.52474</td>
<td>492136.96</td>
</tr>
<tr>
<td>New South Wales</td>
<td>35</td>
<td>1536.8228</td>
<td>677.14095</td>
<td>458519.87</td>
</tr>
<tr>
<td>Queensland</td>
<td>35</td>
<td>1368.2912</td>
<td>536.31969</td>
<td>287638.81</td>
</tr>
<tr>
<td>Victoria</td>
<td>35</td>
<td>1149.0504</td>
<td>516.55309</td>
<td>266827.09</td>
</tr>
<tr>
<td>Tasmania</td>
<td>35</td>
<td>898.69512</td>
<td>386.35397</td>
<td>149269.39</td>
</tr>
</tbody>
</table>

Note: **The key question:** Are these sample means different because of sampling variability (random chance) OR are they different because at least one of the populations really is different?

To answer this, we need a test statistic and a P-value.

**ANOVA Test Statistic – F distribution**

The T test statistic can only be used to compare two things, either a sample mean to a population mean or the mean averages from two populations. Either way, the T-test statistic cannot handle 3 or more groups.

F-test statistic to the rescue

The F-test statistic uses variance to measure how different the sample means are. It relies on two specific variances. Remember Variance (standard deviation squared) is a measure of spread that determines how far values are from the mean.

- Variance between the groups (How far the sample means for each group are from the overall mean of all the groups combined)
- Variance within the groups (How far each sample value is from its own sample mean.)

The F-test statistic

\[ F = \frac{\text{Variance between the groups}}{\text{Variance within the groups}} \]

**F-test statistic sentence:** The ratio of the variance between the groups to the variance within the groups.
Now let’s watch the 3 ANOVA videos on Kahn Academy to see how the F is calculated.

Notes about the F-test statistic

- In a fraction, when the numerator is significantly larger than the denominator, the overall fraction is large. So if the variance between the groups is much larger than the variance within the groups, this will give a large F-test statistic (small P-value) and indicates that the sample means are significantly different. (Unlikely to happen by random chance, reject the null hypothesis)

- In a fraction, when the numerator is the same or smaller than the denominator, the overall fraction is small. So if the variance between the groups is much smaller than the variance within the groups, this will give a small F-test statistic (large P-value) and indicates that the sample means are not significantly different. (Could have happen by random chance, fail to reject the null hypothesis)

- There are three degrees of freedom in an ANOVA: df within the groups, df between the groups, and df total.

Here are the degrees of freedom in our last example of Australia weekly salaries

- df between = # groups – 1 = 5 – 1 = 4
- total df = total # of data values from all groups combined – 1 = 35x5 – 1 = 175 – 1 = 174
- df within each group = (35 – 1) + (35 – 1) + (35 – 1) + (35 – 1) + (35 – 1) = 34 + 34 + 34 + 34 + 34 = 170

Notice df between (4) + df within (170) = df total (174)

- Do not calculate the F test statistic by hand with a calculator. It is a really difficult calculation. Calculate the F-test statistic with a computer software like StatCrunch or StatKey.

- Like Chi-Squared, the F-distribution is always skewed right and the ANOVA test is always a right tailed test.
How to do an ANOVA test with StatCrunch

Copy and Paste your raw quantitative data from each group into some columns of StatCrunch.

To check assumptions, you will want to create histograms of all your data sets to check shape. Go to “graph” menu and click on “histogram”. (Or you can check the dot plot option in the one-way ANOVA menu.) A side by side boxplot is also a nice summary of center and spread.

To calculate the F-test statistic and P-value, Go to the “stat” menu, then “ANOVA”, then “one way”.

Stat ➔ ANOVA ➔ One Way

Hold the control key down to select the columns where your data is and push compute. Here is the printout we got.

**Analysis of Variance results:**
Data stored in separate columns.

<table>
<thead>
<tr>
<th>Column</th>
<th>n</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Std. Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>North Territory</td>
<td>35</td>
<td>1534.5395</td>
<td>701.52474</td>
<td>118.57932</td>
</tr>
<tr>
<td>New South Wales</td>
<td>35</td>
<td>1536.8228</td>
<td>677.14095</td>
<td>114.45771</td>
</tr>
<tr>
<td>Queensland</td>
<td>35</td>
<td>1368.2912</td>
<td>536.31969</td>
<td>90.654574</td>
</tr>
<tr>
<td>Victoria</td>
<td>35</td>
<td>1149.0504</td>
<td>516.55309</td>
<td>87.313408</td>
</tr>
<tr>
<td>Tasmania</td>
<td>35</td>
<td>898.69512</td>
<td>386.35397</td>
<td>65.305741</td>
</tr>
</tbody>
</table>

**ANOVA table**

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
<th>F-Stat</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Columns</td>
<td>4</td>
<td>10484499</td>
<td>2621124.8</td>
<td>7.9217156</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Error</td>
<td>170</td>
<td>56249332</td>
<td>330878.43</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>174</td>
<td>66733832</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Let’s see if we understand what we are seeing. Notice the MS (mean sum of squares) is the sum of squares (SS) divided by degrees of freedom (df).

MS (columns) is the variance between the groups (2621124.8)

MS (Error) is the variance within the groups (330878.43)

So the F-test statistic is calculated by the formula

\[ F = \frac{\text{Variance between the groups}}{\text{Variance within the groups}} = \frac{2621124.8}{330878.43} = 7.9217 \]
So the variance between the groups is almost 8 times greater than the variance within the groups. Is this significantly large for an F?

Again, when unsure about a test statistic refer to a simulation or the P-value. If the sample data is in the tail of the simulation or if the P-value is close to zero it is significant.

Notice in our printout from StatCrunch we got the following P-value: “<0.0001”. This means the actual P-value is very close to 0.

P-value ≈ 0

From our study of P-values, we know this is very significant. So the F test statistic is significantly large and the variance between the groups is significantly greater than the variance within the groups. This is highly unlikely to happen by random chance.

Reject the null hypothesis $H_0$

Conclusion?
Recall the claim was that at least one state was different than the others (alternative hypothesis). Since we rejected the null, we support this claim. Our P-value is very small and our test statistic very large, so we have significant sample evidence.

Conclusion: There is significant sample evidence to support the claim that the mean average salaries of people in Northern Territory, New South Wales, Queensland, Victoria, and Tasmania are different.
Assumptions for ANOVA

- Random
- Sample sizes at least 30 or bell shaped
- Groups Independent of each other
- Equal population variances (no group has a standard deviation more than twice as big as any other group)

Check the assumptions for the Australian states problem.

**Random?** The data was a random sample of people living in these states in Australia.

**30 or bell shaped?** Some histograms were bell shaped and some a little skewed, but since the sample sizes were 35 (over 30), it does pass this assumption.

**Independent groups?** The data may fail this one. Salaries from state to state all may be related due to their reliance on the overall economy and unemployment rates of Australia.

**Equal Population Variances?** Looking at the standard deviations listed in the ANOVA printout. The largest standard deviation is 701.5 and the smallest was 386.4, so no standard deviation was more than twice as large as any other. So it passes this assumption.

**Simulating the F-distribution**

We can also determine if the F-test statistic is sufficiently large by simulation. Go to [www.lock5stat.com](http://www.lock5stat.com) and click on “StatKey”. Click on “ANOVA for difference in means”.

In the top right corner, change the “ants” problem to “Fish Gill, Gill rates by Calcium”. This is exploring if the amount of calcium is related to how well the gills of a fish work.

Here is the null and alternative hypothesis. Remember when the sample means are equal that means calcium is not related to how well a fish’s gills function. When sample means are different, this means that calcium is related.

\[ H_0 : \mu_1 = \mu_2 = \mu_3 \]
\[ H_A : \text{at least one is } \neq \text{ (claim)} \]

Notice the F-test statistic has already been calculated.
Let’s see if the $F = 4.648$ is significant by simulating $F$ test statistic assuming the null hypothesis is true and the population means are equal.

2000 simulations gave the following. Notice we are looking for the probability that the sample data (original $F$ test statistic) or more extreme happened by random chance. This is the $P$-value.

Notice the original sample $F$ test statistic did fall in the tail. It also has a very small $P$-value (0.012). Both of these things tell us that the $F$ test statistic of 4.648 was significant.

Reject the null hypothesis.

There is sufficient sample evidence to support the claim that calcium is related to the function of a fish’s gills (the population means are different).