Analysis of Variance (ANOVA)

Module 13

Statistics 251: Statistical Methods

Updated 2019

Analysis of variance (ANOVA or AOV)

The methods learned in Modules 7-10 only dealt with comparisons of two means or proportions. The question is, why not just do several 2-sample tests if we have at least two means? The reason is the Type I error, α, $\alpha = P(Reject\ H_0|H_0\ true)$ (rejecting a true null hypothesis). By doing several 2-sample t-tests simultaneously, since they would not be wholly independent, it increases the Type I error rate.

As an example, the number of 2-sample comparisons is the number of factor (treatment) groups choose 2 (as in a combination), $\binom{k}{2}$ where k is the number of factor groups and 2 because we are doing 2-sample comparisons. So if we had say $k = 4$ groups, then the number of comparisons to do in that case would be $\binom{4}{2} = \frac{4!}{2!(4-2)!} = 6$, *each having their own Type I error rate of 5%*, meaning that the overall Type I error rate for the entire experiment would be $6(0.05) = 0.3$. The ANOVA procedure protects the Type I error rate from inflating by doing multiple tests.

Hypotheses

The hypotheses for a (1-way) ANOVA for CRD (completely randomized design). The hypotheses only state that there are (or are not) differences among the factor group means **but does not indicate *where* the differences are, just if there are some**

$$H_0 : \mu_1 = \mu 2 = \cdots = \mu_k$$

$$H_a : H_0 \text{ not true (or at least one } \mu_i \text{ differs)}$$

Anova terms I

The results of the analysis is displayed in a table. Shown below is the generic version of the table and the following slides will define and give formulas for the values of the ANOVA output table.

Source	df	SS	MS	F	`Pr>F`
Factor (or Treatment)	k-1	SST	MST	$= \frac{MST}{MSE}$	$P(F > F_{calc})$
Error	n-k	SSE	MSE		
Total	n-1	TSS			

Anova terms II

Most of the calculations involve figuring out the variation (variances) between groups, within groups, and the total variation. Formulas on next slide.

(1) Sources of variation
   (a) Factor (between)
   (b) Error (within, residuals)
   (c) Total

(2) Sums of squares (basically numerators of variances)

(a) Factor or Treatment ($SS(Factor)$ or SST): sum of squared distances between each factor mean ($\overline{y}_i$) and the overall (grand) mean ($\overline{y}..$ or $\overline{\overline{y}}$)
(b) Error ($SS(Error)$ or SSE): sum of squared distances between each individual observation (y_{ij}) and their corresponding factor mean ($\overline{y}_i$)
(c) Total($SS(Total)$ or TSS): sum of squared distances between each individual observation (y_{ij}) and the grand mean ($\overline{y}..$)

(3) Degrees of freedom (df)
   (a) Factor: $df_{factor} = k - 1$ where k is the number of factor groups
   (b) Error: $df_{error} = n - k$ where n is the total number of observations in the experiment
   (c) Total: $df_{total} = n - 1$
(4) Mean squares (basically variances)
   (a) Factor ($MS(Factor)$): variance for factor is sum of squares for factor divided by the factor degrees of freedom (df_{factor})
   (b) Factor ($MS(Error)$): variance for error is sum of squares for error divided by the error degrees of freedom (df_{error}); also computed by the sum of each group variance multiplied by each group sample size minus 1
   (c) Total: could be calculated in the same manner but is not usually calculated nor used

Anova terms III

The main goal of anova is to calculate the sums of squares (SS), mean square (MS), and the test statistic. The following are the calculations for all the values needed for the hypothesis test.

$$SS(Factor) = \sum n_i(\overline{y}_i - \overline{y}..)^2$$

$$SS(Error) = \sum (y_i - \overline{y}_i)^2 = \sum s_i^2(n_i - 1)$$

$$SS(Total) = \sum (y_i - \overline{y}..)^2 = SST + SSE$$

Anova Terms IV

Mean squares

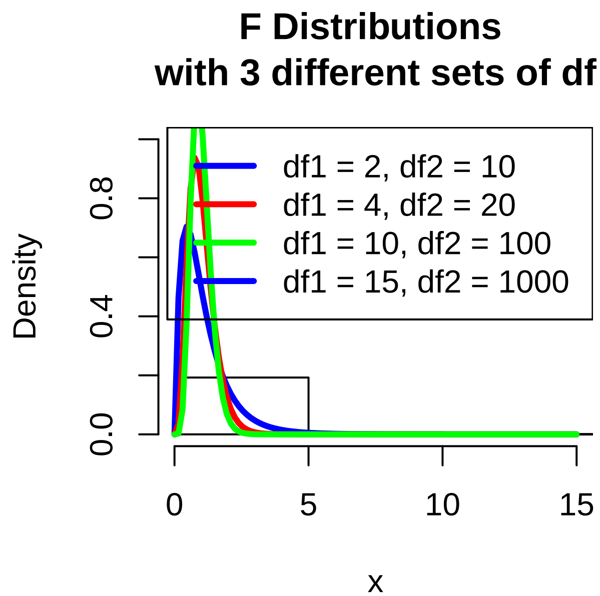
$$MST = \frac{SST}{df_{factor}} = \frac{SST}{k-1}$$

$$MSE = \frac{SSE}{df_{error}} = \frac{SSE}{n-k}$$

There is no calculation of (nor use of) the Total mean square.

Anova Terms V

Test Statistic: called an F statistic from the F probability distribution. Like the χ^2 distribution, F changes shape as df (2 of them) vary. The first df is df_{factor} and the second is df_{error} from the ANOVA output table. The rows of the distribution table are df_{factor} and the columns are df_{error}.



Anova Terms VI

$$F = \frac{SST/(k-1)}{SSE/(n-k)} = \frac{MST}{MSE}$$

$$pvalue = P(F > F_{calc,df_1,df_2})$$

$$reject\ H_0\ if\ pvalue \leq \alpha$$

The F distribution has two degrees of freedom, df_1 and df_2. $df_1 = k-1$ (df for treatment) and $df_2 = n-k$ (df for error). The *pvalue* for the test is included in software output. When computing by hand, we will use the critical value approach and

$$reject\ H_0\ if\ F_{calc} \geq F_{\alpha,df_1,df_2}$$

The F table to find critical values

We will use either the table that is in your statistical tables packet (last pages of packet) or the website statdistributions.com. Examples below are using statdistributions.com. The given α, F, and df values will be in the form: F_{α,df_1,df_2}.

There are 4 different distributions on this website. In the upper right, click on the link for the F table F distribution.

(1) $F_{0.05,3,10}$. On the website, input 0.05 in the box labeled `p-value`, input 3 in the box labeled `numerator d.f.`, and input 10 in the box labeled `denominator d.f.`. The last thing to double check is that it should have the option `right tail` selected (if it is not, select `right tail`). $F_{0.05,3,10} = 3.708$

(2) $F_{0.1,3,10}$. On the website, input 0.10 in the box labeled `p-value`, input 3 in the box labeled `numerator d.f.`, and input 10 in the box labeled `denominator d.f.`. The last thing to double check is that it should have the option `right tail` selected (if it is not, select `right tail`). $F_{0.1,3,10} = 2.728$

The F table to find *pvalues*

The website will also calculate a *pvalue* for you if you have the value of your test statistic and dfs

(1) $F_{calc} = 5.097, df_1 = 3, df_2 = 17$. On the website, input 5.097 in the box labeled `F-value`, input 3 in the box labeled `numerator d.f.`, and input 17 in the box labeled `denominator d.f.`. The last thing to double check is that it should have the option `right tail` selected (if it is not, select `right tail`). $F_{3,10} = 5.097$ with $pvalue = 0.011$. The results would be since $pvalue = 0.011 \leq \alpha(0.05)$, H_0 is rejected.

(2) $F_{calc} = 2.134, df_1 = 3, df_2 = 8$. On the website, input 2.134 in the box labeled `F-value`, input 3 in the box labeled `numerator d.f.`, and input 8 in the box labeled `denominator d.f.`. The last thing to double check is that it should have the option `right tail` selected (if it is not, select `right tail`). $F_{3,8} = 2.134$ with $pvalue = 0.174$. The results would be since $pvalue = 0.174 \nleq \alpha(0.05)$, H_0 is not rejected.

Assumptions of ANOVA

(1) $E(\epsilon_i) = 0$; the mean of the residuals should be approximately 0
(2) $V(\epsilon_i) = \sigma^2_\epsilon$; the variance of the residuals should be constant for all values of the response
(3) $Cov(\epsilon_i, \epsilon_i') = 0$; independence of residuals
(4) $\epsilon_i \sim N(0, \sigma^2_\epsilon)$; residuals should have an approximate normal distribution with mean 0 and constant variance

We will not check the assumptions but you will be responsible for knowing them and being able to list them either in mathematical terms or plain words.

Incomplete ANOVA table

One thing intro stat textbooks love to do to students is to give them an exercise in completing an ANOVA table using the basic formulas. So I could not let you miss out.

Source	df	SS	MS	F
Treatment	4	26.3		
Error		52.8		
Total	29			

To find df_{error}, know it is $n - k$. $df_{treatment} = k - 1$ so $k = 4 + 1 = 5$. $df_{total} = n - 1 = 29$ so $n = 30$. Now $df_{error} = n - k = 30 - 5 = 25$. As a check, this means that $df_{total} = df_{treatment} + df_{error}$.

To find TSS, $TSS = SST + SSE = 26.3 + 52.8 = 79.1$

The MS (mean squares) calculations are the individual SS divided by their corresponding df. So $MST = \frac{SST}{df_{treatment}} = \frac{26.3}{4} = 6.575$, $MSE = \frac{SSE}{df_{error}} = \frac{52.8}{25} = 2.112$. No Total mean square is calculated.

F is the test statistic (called F_{calc} in these notes). $F = \frac{MST}{MSE} = \frac{6.575}{2.112} = 3.1131629$.

Completion of the incomplete ANOVA table

Source	df	SS	MS	F
Treatment	4	26.3	6.575	3.113
Error	25	52.8	2.112	
Total	29	79.1		

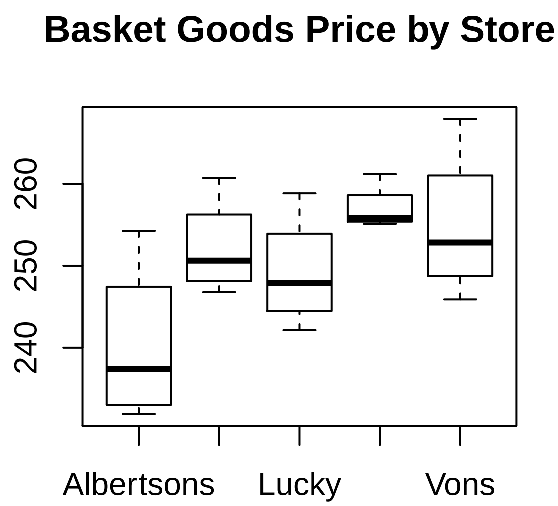
Example I (no software, except graph)

A design to study the prices of a fixed basket of goods from different grocery stores in California was conducted. The top 5 major grocery stores were used and prices (in US dollars) over a total of 4 weekly visits per store to check on the fixed basket were collected.

Store/Week	1	2	3	4	n_i	$\overline{y}_i$	s_i^2
Albertsons	254.26	240.62	231.90	234.13	4	240.23	101.197
Ralphs	256.03	255.65	255.12	261.18	4	256.99	7.923
Vons	267.92	251.55	245.89	254.12	4	254.87	87.509
Alpha Beta	260.71	251.80	246.77	249.45	4	252.18	36.542
Lucky	258.84	242.14	246.80	248.99	4	249.19	49.526

Example I (no software, except graph)

Graph of fixed basket prices by store.



Example I (no software, except graph)

The goal is to create the ANOVA table. One value not given in the table that is required is the overall grand mean, $\overline{y}_{..}$, which is the mean of all observations of the experiment (regardless of which treatment group they are in); it is also the overall mean of the group means. There are 5 stores, so $k = 5$ and 4 observations per store, so $n = 4(5) = 20$. Is there sufficient evidence that there is at least one store that is different in the price of basket goods?

$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$ and $H_a :$ H_0 *not true*

$\overline{y}_{..} = \frac{\sum y_i}{n} = \frac{\sum \overline{y}_i}{k} = \frac{240.23+256.99+254.87+252.18+249.19}{5} = 250.6935$

$SST = \sum n_i(\overline{y}_i - \overline{y}_{..})^2 = 4[(240.23 - 250.69)^2 + (256.99 - 250.69)^2 + (254.87 - 250.69)^2 + (252.18 - 250.69)^2 + (249.19 - 250.69)^2] = 684.63733$

$SSE = \sum (y_i - \overline{y}_i)^2 = \sum s_i^2(n_i - 1) = (4 - 1)(101.197 + 7.923 + 87.509 + 36.542 + 49.526) = 848.093125$

$TSS = \sum (y_i - \overline{y}..)^2 = SST + SSE = 684.63733 + 848.093125 = 1532.730455$

$df_{treatment} = df_1 = k - 1 = 5 - 1 = 4$, $df_{error} = n - k = 20 - 5 = 15$, $df_{total} = n - 1 = 20 - 1 = 19$
$MST = \frac{SST}{df_1} = \frac{684.63733}{4} = 171.1593325$

$MSE = \frac{SSE}{df_2} = \frac{684.63733}{15} = 56.5395417$

No Total mean square calculated.

$F = \frac{MST}{MSE} = \frac{171.1593325}{56.5395417} = 3.0272501$

Example I

The ANOVA table:

Source	df	SS	MS	F
Treatment	4	684.64	171.16	3.0273
Error	15	848.09	56.54	
Total	19	1532.73		

Example I

Now to figure out if we will reject the null hypothesis.

We can reject H_0 if $pvalue \leq \alpha$ ($pvalue$ approach) or if $F_{calc} \geq F_{\alpha,df_1,df_2}$ (critical value approach). Since we computed by hand, we will use the second method (critical value method).

Using statdistributions.com, we need $F_{\alpha,df_1,df_2} = F_{0.05,4,15}$. Input '$p - value:$ '0.05, '$numeratord.f.:$ '4, and '$denominatord.f.:$ '4. $F_{0.05,4,15} = 3.056$ F with alpha=5%, df1=4,df2=15

Since $3.0273 \ngeq 3.056$ (barely!), we cannot reject H_0. There are no significant differences in the fixed basket prices by store.

Example II (with software output)

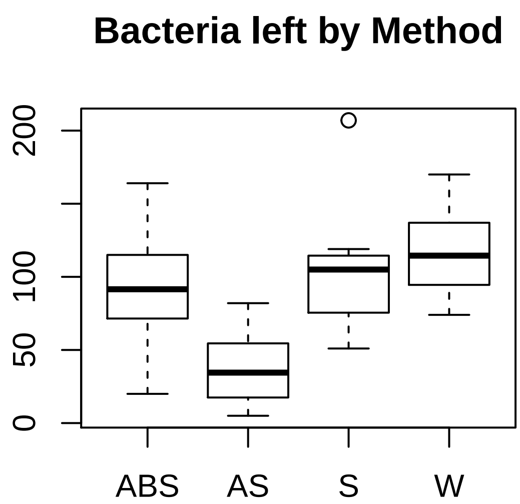
Learning to read the output from a statistical software program. The following example will use output from the program R.

Washing hands is supposed to remove potentially harmful (and definitely gross) bacteria from your hands, thus minimizing the spread of illness and other random goobers (not the goofy kind). A completely randomized design was used to study different hand-washing methods to determine if there are differences in the amount of bacteria left on hands based on method. A total of 32 subjects were randomly assigned to one of 4 methods: water only (W), regular soap (S), antibacterial soap (ABS), and alcohol spray (AS). Is there sufficient evidence that at least one hand-washing method differs in the amount of bacteria left on the hand?

Example II (with software output)

$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$, $H_a:$ H_0 not true

```r
boxplot(Bacteria~Method,data=Hw,main='Bacteria left by Method')
```




```
fit=lm(Bacteria~Method,data=Hw)
anova(fit)
```

```
Analysis of Variance Table

Response: Bacteria
          Df Sum Sq Mean Sq F value   Pr(>F)
Method     3  29882  9960.7  7.0636 0.001111 **
Residuals 28  39484  1410.1
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Notice all the df, SS, MS, F, and *pvalue* is all input into a table of output. The main things of interest are the F value and *pvalue*. $F = 7.0636$ with $pvalue = 0.001111$. Since $pvalue = 0.001111 \leq \alpha(0.05)$, H_0 is rejected. There is at least one hand-washing method is better at removing bacteria from the hands.

Multiple Comparisons

Multiple comparisons are only to be done if and only if (iff) the null hypothesis of ANOVA is rejected. (If the null is not rejected, you are saying there are no differences, so why would you try and find where the non-existent differences are?!?)

So now that we have seen an example of rejecting the null hypothesis of an ANOVA problem, we can just look and see if there are differences, right? Nope! That would be too easy, wouldn't it? Remember an earlier slide from this lecture that stated the Type I error rate would increase, depending on how many 2-sample comparisons we do? That is why. The hand-washing example with $k = 4$ would require $\binom{k}{2} = \binom{4}{2} = 6$ 2-sample comparisons, and the larger k is, the more comparisons to do and the larger Type I error without a modified procedure to execute the comparisons.

There are many different multiple comparisons, we will learn one of the more commonly used ones called Tukey's Honest Significant Difference (Tukey's HSD).

Tukey's (not turkey) HSD

This is a modified 2-sample CI that uses a different statistical distribution called the Studentized Range distribution, denoted as $q_\alpha(k, df_2)$. You will not have to use the distribution, just interpret the output from the comparison.

Any pair of means will be determined to be significantly different if the magnitude of their difference is greater than the cutoff value, which is in essence a bound (margin of error). That is if,

$$|\overline{y}_i - \overline{y}_j| \geq HSD \; where \; HSD = q_\alpha(k, df_2)\sqrt{\frac{MSE}{n_i}}$$

Let's wash some hands! Now that we rejected the null hypothesis, a multiple comparison, specifically Tukey's HSD, is appropriate.

Toward the bottom of the following output, there is an section with a header that reads `Treatments with the same letter are not significantly different.`, the treatment means are listed in order (largest to smallest) and there is a column called `groups`. The letters tell you which groups are statistically different. The groups that have the *same* groups letter are statistically the *same*. Different groups letters are statistically *different*. There is also a value close to the groups that says `Minimum Significant Difference`. The value of the HSD is what the absolute value of the difference between any 2 means needs to be greater than if we wanted to look at the comparison in CI-type formatting (we will not here but something for future classes use of statistics).

```
Study: fit ~ "Method"

HSD Test for Bacteria

Mean Square Error:  1410.143

Method,  means

    Bacteria      std r Min Max
ABS     92.5 41.96257 8  20 164
AS      37.5 26.55991 8   5  82
S      106.0 46.95895 8  51 207
W      117.0 31.13106 8  74 170

Alpha: 0.05 ; DF Error: 28
Critical Value of Studentized Range: 3.861244

Minimun Significant Difference: 51.26415

Treatments with the same letter are not significantly different.

    Bacteria groups
W      117.0      a
S      106.0      a
ABS     92.5      a
AS      37.5      b
```

`Minimum Significant Difference: 51.26415`. The value 51.26415 is the HSD value that the absolute value of the difference between any 2 means needs to be greater than.

The groups lettering indicates that AS (alcohol spray) has the only different letter, `b`, and is significantly different than the other methods (all other methods share the letter `a` so they are all the same).