ANOVA

Donna Dietz

American University

dietz@american.edu

STAT 202 - Spring 2020

ANOVA

ANOVA

ANOVA stands for ANalysis Of VAriance (or VAriation).

It can also be called *Fisher's analysis of variation*.

Ronald Fisher

Ronald Fisher introduced the term variance and proposed its formal analysis in a 1918 article *The Correlation Between Relatives on the Supposition of Mendelian Inheritance*. His first application of the analysis of variance was published in 1921. Analysis of variance became widely known after being included in Fisher's 1925 book *Statistical Methods for Research Workers*. (Wikipedia)

Fisher

Portrait of Ronald Fisher as a young man
17 February 1890 – 29 July 1962

In his honor, we call the relevant curve family the *F distribution*. (Don't forget George Snedecor's contributions though!)

Think of it as a massive t-test

Since we've already covered Mr. Student (Gosset) and his tables, you can almost think of ANOVA as running all the possible t-tests on several groups at once, to see if there are any groups that differ. However, you should avoid doing all those t-tests anyhow.

There is no harm in doing ANOVA on two groups, and you should end up with the same p-value anyhow, so why bother with t-tests?
(I honestly haven't the foggiest! I wouldn't use them!)

Reminder

Test	**Data Type**	**Data Type**
Linear Regression	Numerical	Numerical
Chi-Square	Categorical	Categorical
ANOVA	Categorical	Numerical

Recall...

When we found the standard deviation by hand, we used a similar process.

Use for SS_{Total} and SS_{WG}.

- Calculate mean of data
- Find all deviations from the mean
- Square them all
- Add them all
- (Stop here this time!)

Find SS_{Total} and SS_{WG}

We are going to trace through a very simplistic example of ANOVA by hand and again on StatCrunch.

Group X: $\{1, 3, 5, 7, 9\}$

Group Y: $\{6, 7, 8, 9, 10\}$

Combined Groups: $\{1, 3, 5, 7, 9, 6, 7, 8, 9, 10\}$

SS_{WG}

Group X: $\{1, 3, 5, 7, 9\}$ has $\mu = 5$.

x	1	3	5	7	9
$x - \mu$	-4	-2	0	2	4
$(x - \mu)^2$	16	4	0	4	16

All of those add up to 40.

Group Y: $\{6, 7, 8, 9, 10\}$ has $\mu = 8$.

y	6	7	8	9	10
$y - \mu$	-2	-1	0	1	2
$(y - \mu)^2$	4	1	0	1	4

All of those add up to 10.

Overall, that adds up to 50.

$$SS_{WG} = 50.$$

StatCrunch calls this *Error*.

SS_{Total}

Overall, the 10 items have a mean of 6.5. This is sometimes called *GM* or the *Grand Mean*.

1	3	5	7	9	6	7	8	9	10
-5.5	-3.5	-1.5	.5	2.5	-.5	.5	1.5	2.5	3.5
30.25	12.25	2.25	.25	6.25	.25	.25	2.25	6.25	12.25

The sum of these values is 72.5.

$$SS_{Total} = 72.5.$$

StatCrunch calls this *Total*.

SS_{BG}

What are those subscripts anyhow?
WG means "Within Groups"
BG means "Between Groups"
Total means overall, as if you didn't have groups at all.

For each group, we square the difference between the group mean and the Grand Mean, and multiply that by the number of items in the group. Then, add the components from each group.

Group X: $5 \cdot (6.5 - 5)^2 = 5 \cdot (1.5)^2 = 11.25$.

Group Y: $5 \cdot (6.5 - 8)^2 = 5 \cdot (1.5)^2 = 11.25$.

These have to be the same because we only have two groups and they are of the same size, so this should not be a surprise. Overall, we get 22.5.

$$SS_{BG} = 22.5.$$

StatCrunch calls this *Columns*.

Shocking craziness!

x	y
1	6
3	7
5	8
7	9
9	10

Analysis of Variance results:
Data stored in separate columns.

Column statistics

Column	n	Mean	Std. Dev.	Std. Error
x	5	5	3.1622777	1.4142136
y	5	8	1.5811388	0.70710678

ANOVA table

Source	DF	SS	MS	F-Stat	P-value
Columns	1	22.5	22.5	3.6	0.0943
Error	8	50	6.25		
Total	9	72.5			

Tukey HSD results (95% level)
x subtracted from

	Difference	Lower	Upper	P-value
y	3	-0.64611269	6.6461127	0.0943

$$SS_{Total} = SS_{WG} + SS_{BG}.$$

$$72.5 = 50 + 22.5$$

THIS IS INSANE!!! Of course, you get a p-value and confidence interval as always.

BUT WHAT WAS THAT?

$$SS_{Total} = SS_{WG} + SS_{BG}$$

This looks very much like $\sigma^2_{A+B} = \sigma^2_A + \sigma^2_B$.
It also reminds us of $A^2 + B^2 = C^2$.
What is going on!?!? I don't see any right triangles here!

Pythagoras strikes again!

Please understand that it is **not** the case that this is true point-by-point. In other words, you cannot take just one point from the data set and expect this relationship to hold. It only holds in the aggregate!

Example: For the y-data point of 9, you would get a squared deviation from GM as $(9 - 6.5)^2$, or 6.25. Your deviation from your group would be 1, so squaring it would still give you 1. The component from the group to the GM is 1.5 which becomes 2.25. Notice that 6.25 is **NOT equal** to 3.75! But when you do this for all the data points and add up all the components, everything always balances out perfectly!

Getting to the F-Statistic

Now that we've done the hard work, finding the F-Statistic is the easy part.

$$n = number \;\; of \;\; points$$

$$k = number \;\; of \;\; groups$$

$$MS_{BG} = \frac{SS_{BG}}{k-1}$$

$$MS_{WG} = \frac{SS_{WG}}{n-k}$$

$$F = \frac{MS_{BG}}{MS_{WG}}$$

Your degrees of freedom are split between the numerator (k-1) and denominator (n-k). You go to your F-tables and look up your p-value. (Those tables are huge, though, so I'm providing them only on Blackboard. You can of course use StatCrunch also.)

Back to the example

$$n = 10$$

$$k = 2$$

$$MS_{BG} = \frac{22.5}{1} = 22.5$$

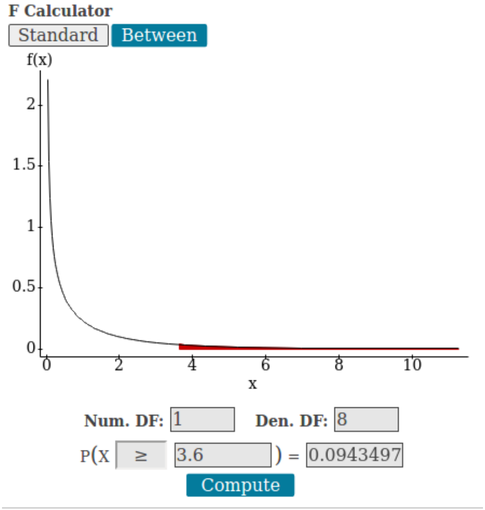
$$MS_{WG} = \frac{50}{8} = 6.25$$

$$F = \frac{22.5}{6.25} = 3.6$$

For $df = 1$ in the numerator and $df = 8$ in the denominator, the F-statistic 3.6 gives a p-value on StatCrunch of about 0.0943.

Verifying...

Stat > Calculators > F



And you should just include the Tukey test in case you do reject the Null.

John Tukey

“Two-key” not 'turkey' please!

John Tukey

John Wilder Tukey (June 16, 1915 – July 26, 2000) was an American mathematician best known for development of the Fast Fourier Transform (FFT) algorithm and box plot. The Tukey range test, the Tukey lambda distribution, the Tukey test of additivity, and the Teichmüller–Tukey lemma all bear his name. He is also credited with coining the term 'bit'. (Wikipedia)

Example 2

To build some intuition, let's have two groups with the same mean.

Group X: $\{1, 3, 5, 7, 9\}$ still has $\mu = 5$

Group Y: $\{2, 3, 4, 5, 6, 7, 8\}$ also has $\mu = 5$

Combined Groups: $\{1, 3, 5, 7, 9, 2, 3, 4, 5, 6, 7, 8\}$ of course has $\mu = 5$.

What now?

x	y
1	2
3	3
5	4
7	5
9	6
	7
	8

Analysis of Variance results:
Data stored in separate columns.

Column statistics

Column	n	Mean	Std. Dev.	Std. Error
x	5	5	3.1622777	1.4142136
y	7	5	2.1602469	0.81649658

ANOVA table

Source	DF	SS	MS	F-Stat	P-value
Columns	1	0	0	0	1
Error	10	68	6.8		
Total	11	68			

The SS_{WG} component from Group X is the same as before, 40.

2	3	4	5	6	7	8
-3	-2	-1	0	1	2	3
9	4	1	0	1	4	9

The sum for Group Y is 28.

$$SS_{WG} = 40 + 28 = 68.$$

Example 2

But it's obvious that this is identical to SS_{Total} because all the subtractions were from 5.

It is also obvious in this one case, that the between group component has to be zero, because their means are identical!

Since the $SS_{BG} = 0$, the F-statistic will be zero. This gives us a p-value of 1, which seems reasonable. There is nothing less extreme, so the chances of getting this result or something more extreme is all cases possible! That's 100%.

x	y
1	2
3	3
5	4
7	5
9	6
	7
	8

Analysis of Variance results:
Data stored in separate columns.

Column statistics

Column	n	Mean	Std. Dev.	Std. Error
x	5	5	3.1622777	1.4142136
y	7	5	2.1602469	0.81649658

ANOVA table

Source	DF	SS	MS	F-Stat	P-value
Columns	1	0	0	0	1
Error	10	68	6.8		
Total	11	68			

Example 3

Again, to build intuition, let's consider two groups that are identical but shifted by a large amount.

x	y
1	101
3	103
5	105
7	107
9	109

Analysis of Variance results:
Data stored in separate columns.

Column statistics

Column	n	Mean	Std. Dev.	Std. Error
x	5	5	3.1622777	1.4142136
y	5	105	3.1622777	1.4142136

ANOVA table

Source	DF	SS	MS	F-Stat	P-value
Columns	1	25000	25000	2500	<0.0001
Error	8	80	10		
Total	9	25080			

We already know about Group X, so the SS_{WG} is twice that for just Group X or 40+40=80. The two groups are perfectly separated by 100 units, for an average of 50. The SS_{BG} will then be ten times the square of 50, or 25000 as shown in the table. Finding SS_{Total} by hand is annoying, so let's skip that.

Example 3

This time, SS_{BG} and therefore MS_{BG} is huge compared to MS_{WG} which is tiny. The F-statistic will be huge and highly significant!

x	y
1	101
3	103
5	105
7	107
9	109

Analysis of Variance results:
Data stored in separate columns.

Column statistics

Column	n	Mean	Std. Dev.	Std. Error
x	5	5	3.1622777	1.4142136
y	5	105	3.1622777	1.4142136

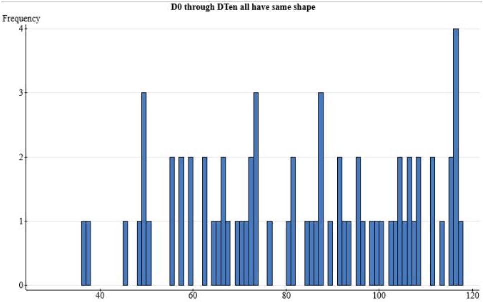
ANOVA table

Source	DF	SS	MS	F-Stat	P-value
Columns	1	25000	25000	2500	<0.0001
Error	8	80	10		
Total	9	25080			

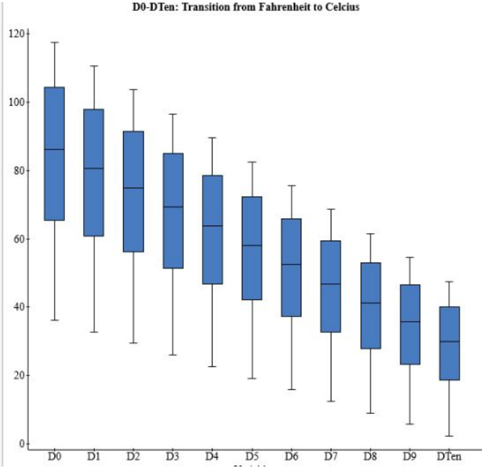
Lots of groups

ANOVA is great for looking at lots of groups, so I'm going to do another experiment! I'm starting with some bogus temperature data in Fahrenheit, and I'm going to slowly slide the values and squish them together, so they become Celcius values. There is no deep reason for doing this, but I just want to see when ANOVA tells me I have different distributions!

Underlying data



Boxplots


D0-DTen: Transition from Fahrenheit to Celcius
120
100
80
60
40
20
0
D0
D1
D2
D3
D4
D5
D6
D7
D8
D9
DTen

ANOVA Results

Analysis of Variance results:
Data stored in separate columns.

Column statistics

Column	n	Mean	Std. Dev.	Std. Error
D0	70	83.346504	22.815327	2.7269532
D1	70	77.864437	21.801313	2.6057553
D2	70	72.38237	20.787298	2.4845573
D3	70	66.900303	19.773284	2.3633594
D4	70	61.418236	18.759269	2.2421615
D5	70	55.93617	17.745254	2.1209636
D6	70	50.454103	16.73124	1.9997657
D7	70	44.972036	15.717225	1.8785677
D8	70	39.489969	14.703211	1.7573698
D9	70	34.007902	13.689196	1.6361719
DTen	70	28.525835	12.675182	1.514974

ANOVA table

Source	DF	SS	MS	F-Stat	P-value
Columns	10	231408.54	23140.854	71.164021	<0.0001
Error	759	246808.82	325.17631		
Total	769	478217.36			

The ANOVA results give $p < 0.0001$ so it's highly significant. But since we have more than 2 groups, we don't know whom to blame!

Tukey Analysis

Tukey HSD results (95% level)
D0 subtracted from

	Difference	Lower	Upper	P-value
D1	-5.4820668	-15.32297	4.3588361	0.781
D2	-10.964134	-20.805037	-1.1232307	0.0151
D3	-16.446201	-26.287103	-6.6052976	<0.0001
D4	-21.928267	-31.76917	-12.087364	<0.0001
D5	-27.410334	-37.251237	-17.569431	<0.0001
D6	-32.892401	-42.733304	-23.051498	<0.0001
D7	-38.374468	-48.215371	-28.533565	<0.0001
D8	-43.856535	-53.697438	-34.015632	<0.0001
D9	-49.338602	-59.179504	-39.497699	<0.0001
DTen	-54.820668	-64.661571	-44.979765	<0.0001

Here, we see that the distribution D0 is not significantly different from D1, and maybe not even from D2, but for D3 and beyond, it's very very significant.

Tukey Analysis

D1 subtracted from

	Difference	Lower	Upper	P-value
D2	-5.4820668	-15.32297	4.3588361	0.781
D3	-10.964134	-20.805037	-1.1232307	0.0151
D4	-16.446201	-26.287103	-6.6052976	<0.0001
D5	-21.928267	-31.76917	-12.087364	<0.0001
D6	-27.410334	-37.251237	-17.569431	<0.0001
D7	-32.892401	-42.733304	-23.051498	<0.0001
D8	-38.374468	-48.215371	-28.533565	<0.0001
D9	-43.856535	-53.697438	-34.015632	<0.0001
DTen	-49.338602	-59.179504	-39.497699	<0.0001

D2 subtracted from

	Difference	Lower	Upper	P-value
D3	-5.4820668	-15.32297	4.3588361	0.781
D4	-10.964134	-20.805037	-1.1232307	0.0151
D5	-16.446201	-26.287103	-6.6052976	<0.0001
D6	-21.928267	-31.76917	-12.087364	<0.0001
D7	-27.410334	-37.251237	-17.569431	<0.0001
D8	-32.892401	-42.733304	-23.051498	<0.0001
D9	-38.374468	-48.215371	-28.533565	<0.0001
DTen	-43.856535	-53.697438	-34.015632	<0.0001

Here, we see that the distribution D1 is not significantly different from D2, and maybe not even from D3, but for D4 and beyond, it's very very significant. Since all these distributions were made in a systematic way relative to each other, these values are going to keep repeating!

Tukey Analysis

D3 subtracted from

	Difference	Lower	Upper	P-value
D4	-5.4820668	-15.32297	4.3588361	0.781
D5	-10.964134	-20.805037	-1.1232307	0.0151
D6	-16.446201	-26.287103	-6.6052976	<0.0001
D7	-21.928267	-31.76917	-12.087364	<0.0001
D8	-27.410334	-37.251237	-17.569431	<0.0001
D9	-32.892401	-42.733304	-23.051498	<0.0001
DTen	-38.374468	-48.215371	-28.533565	<0.0001

D4 subtracted from

	Difference	Lower	Upper	P-value
D5	-5.4820668	-15.32297	4.3588361	0.781
D6	-10.964134	-20.805037	-1.1232307	0.0151
D7	-16.446201	-26.287103	-6.6052976	<0.0001
D8	-21.928267	-31.76917	-12.087364	<0.0001
D9	-27.410334	-37.251237	-17.569431	<0.0001
DTen	-32.892401	-42.733304	-23.051498	<0.0001

Again...

Tukey Analysis

D5 subtracted from

	Difference	Lower	Upper	P-value
D6	-5.4820668	-15.32297	4.3588361	0.781
D7	-10.964134	-20.805037	-1.1232307	0.0151
D8	-16.446201	-26.287103	-6.6052976	<0.0001
D9	-21.928267	-31.76917	-12.087364	<0.0001
DTen	-27.410334	-37.251237	-17.569431	<0.0001

D6 subtracted from

	Difference	Lower	Upper	P-value
D7	-5.4820668	-15.32297	4.3588361	0.781
D8	-10.964134	-20.805037	-1.1232307	0.0151
D9	-16.446201	-26.287103	-6.6052976	<0.0001
DTen	-21.928267	-31.76917	-12.087364	<0.0001

D7 subtracted from

	Difference	Lower	Upper	P-value
D8	-5.4820668	-15.32297	4.3588361	0.781
D9	-10.964134	-20.805037	-1.1232307	0.0151
DTen	-16.446201	-26.287103	-6.6052976	<0.0001

D8 subtracted from

	Difference	Lower	Upper	P-value
D9	-5.4820668	-15.32297	4.3588361	0.781
DTen	-10.964134	-20.805037	-1.1232307	0.0151

And yet again...

Tukey Analysis

D7 subtracted from

	Difference	Lower	Upper	P-value
D8	-5.4820668	-15.32297	4.3588361	0.781
D9	-10.964134	-20.805037	-1.1232307	0.0151
DTen	-16.446201	-26.287103	-6.6052976	<0.0001

D8 subtracted from

	Difference	Lower	Upper	P-value
D9	-5.4820668	-15.32297	4.3588361	0.781
DTen	-10.964134	-20.805037	-1.1232307	0.0151

D9 subtracted from

	Difference	Lower	Upper	P-value
DTen	-5.4820668	-15.32297	4.3588361	0.781

And finally...

What can you do?

Now that you've spent a semester in STAT 202, what can you do that you couldn't do before?

Not knowing what you could do before, I'm guessing many of you can now do these things that are new:

- Figure out a good sample size for a poll you might wish to do.
- Come up with a null and alternative hypothesis in a research setting.
- With software, use LSR and make meaningful statements about correlation
- Add uncorrelated random variables correctly
- Add correlated random variables correctly (or know to avoid it)
- With software, use χ^2 analysis correctly for categorical variables
- With software, use ANOVA correctly to tell if groups differ.

What matters

What matters most is not what you can recall or what you forget.

What matters most is what's left after you forget the precise content.

You have a better understanding of statistical reasoning. You're less likely to be led astray by bogus arguments. You're not intimidated by statistical language.

You have a vague memory of everything you once knew, and you can look it up again later if you need to know the details again.

Progression 11

Contingency table results:
Rows: A
Columns: A

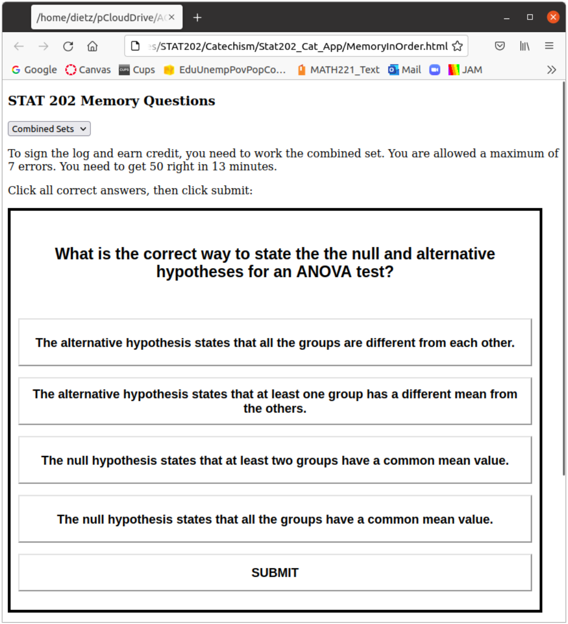
Cell format
Count (Expected count)

	false	true	Total
false	36 (12.96)	0 (23.04)	36
true	0 (23.04)	64 (40.96)	64
Total	36	64	100

Chi-Square test:

Statistic	DF	Value	P-value
Chi-square	1	100	<0.0001

THE LAST MEMORY QUESTION


/home/dietz/pCloudDrive/A
s/STAT202/Catechism/Stat202_Cat_App/MemoryInOrder.html
Google
Canvas
Cups
EduUnempPovPopCo...
MATH221_Text
Mail
JAM
STAT 202 Memory Questions
Combined Sets
To sign the log and earn credit, you need to work the combined set. You are allowed a maximum of 7 errors. You need to get 50 right in 13 minutes.
Click all correct answers, then click submit:
What is the correct way to state the the null and alternative hypotheses for an ANOVA test?
The alternative hypothesis states that all the groups are different from each other.
The alternative hypothesis states that at least one group has a different mean from the others.
The null hypothesis states that at least two groups have a common mean value.
The null hypothesis states that all the groups have a common mean value.
SUBMIT

STAT 202 Memory Questions

Combined Sets

To sign the log and earn credit, you need to work the combined set. You are allowed a maximum of 7 errors. You need to get 50 right in 13 minutes.

Click all correct answers, then click submit:

What is the correct way to state the the null and alternative hypotheses for an ANOVA test?

The alternative hypothesis states that all the groups are different from each other.

The alternative hypothesis states that at least one group has a different mean from the others.

The null hypothesis states that at least two groups have a common mean value.

The null hypothesis states that all the groups have a common mean value.

SUBMIT