

Stat 342 - Wk 9: Hypothesis Tests and Analysis

Crash course on ANOVA,
proc glm

Crash Course: ANOVA

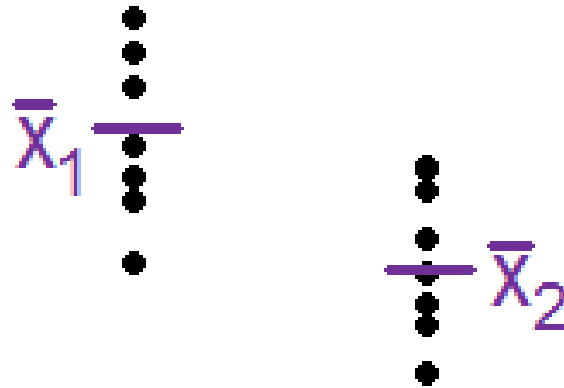
AnOVa stands for Analysis Of Variance.

Sometimes it's called ANOVA, and sometimes AOV.

ANOVA is the natural extension of the two-sample t-test.

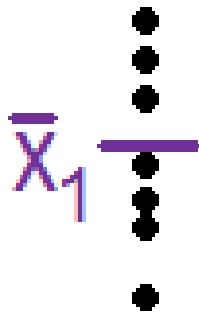
Two-sample T-test: Are the means of these two groups the same?

ANOVA: Are the means of ALL of these groups the same?

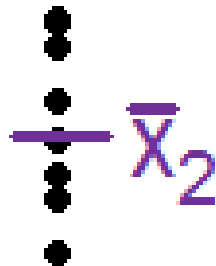


With the two-sample t-test, we did this by taking a sample mean from each group and comparing the difference to the ***standard error of the difference.***

The t-test determines whether this difference was significant (in other words, testing the null hypothesis that the difference between the true means was **zero**) was the t-score.

$$\bar{X}_1$$


A diagram showing a horizontal purple line representing the mean \bar{X}_1 . Above and below the line are five black dots each, representing individual data points in a sample.

$$\bar{X}_2$$


A diagram showing a horizontal purple line representing the mean \bar{X}_2 . Above and below the line are five black dots each, representing individual data points in a sample.

$$t = \frac{\bar{X}_1 - \bar{X}_2}{SE}$$

It was always $t = (\text{difference}) / \text{Standard Error}$.

The definition of standard error depended on the details (paired/independent, pooled/non-pooled standard deviation)

If the difference was bigger, the t-score was bigger and we more often rejected the null hypothesis.

$$\bar{X}_1 \frac{\bullet}{\bullet} \frac{\bullet}{\bullet} \frac{\bullet}{\bullet} \frac{\bullet}{\bullet} \frac{\bullet}{\bullet}$$

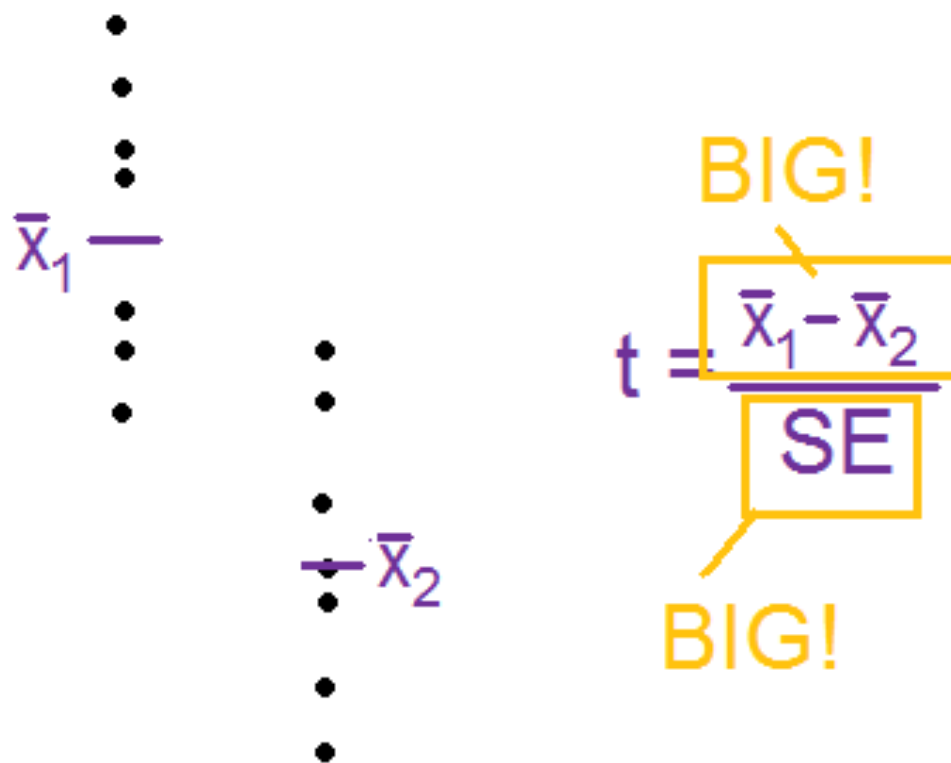
BIG!

$$t = \frac{\bar{X}_1 - \bar{X}_2}{SE}$$

$$\frac{\bullet}{\bullet} \frac{\bullet}{\bullet} \frac{\bullet}{\bullet} \frac{\bullet}{\bullet} \frac{\bullet}{\bullet} \bar{X}_2$$

It's easier to say a difference is real when the sample mean difference is larger. (Easier to detect larger effects)

If there was more scatter between the points within a group, the standard error got bigger, and we more often failed to reject the null hypothesis.



Standard error also gets smaller when there are more data points.

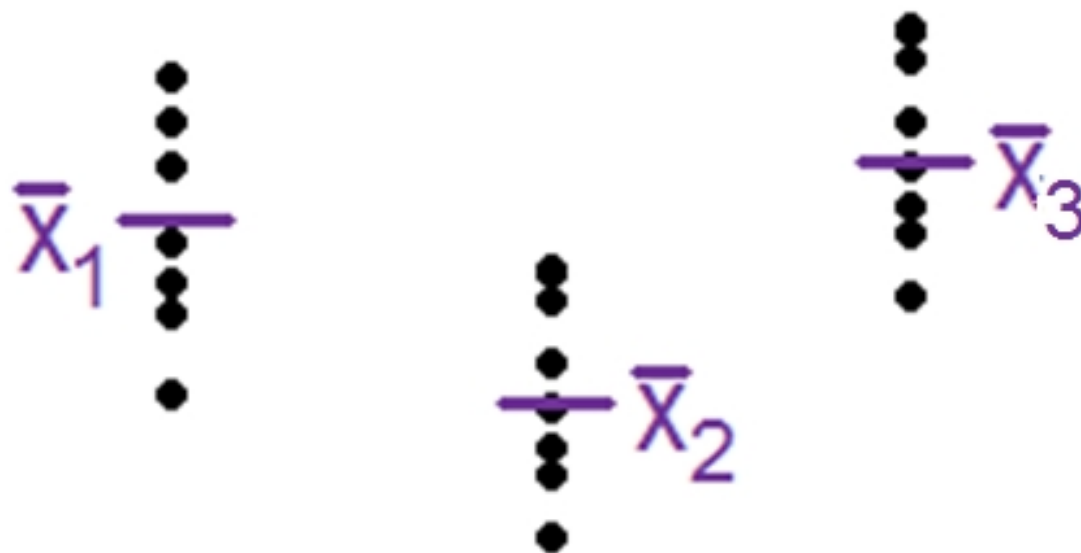
The diagram shows two vertical columns of dots representing data points. The left column has a horizontal line at the top labeled \bar{x}_1 . The right column has a horizontal line at the bottom labeled \bar{x}_2 . To the right of these columns is the t-test formula: $t = \frac{\bar{x}_1 - \bar{x}_2}{SE}$. The numerator $\bar{x}_1 - \bar{x}_2$ is enclosed in a yellow box with the word "BIG!" written above it in yellow. The denominator SE is enclosed in a yellow box with the word "BIG!" written below it in yellow.

$$t = \frac{\bar{x}_1 - \bar{x}_2}{SE}$$

In every case that the t-test is used, you're ultimately just answering one question over and over again:

Are the differences *between* the two groups large compared to the differences *within* each group?

Can we use the t-test to determine if there are differences between any of the three means from three samples?



We can't do this all as a single t-test, because the t-test is only a comparison between **two** sample means. We have three.

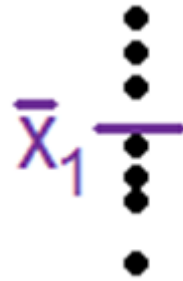
We could test each pair of groups and look for differences, but that introduces issue of multiple testing (the more tests you to, more likely you are to commit an error like falsely rejecting the null)

A much cleaner solution is the ***F-test*** of ANOVA.

$$F = \frac{MS_{\text{between}}}{MS_{\text{within}}}$$



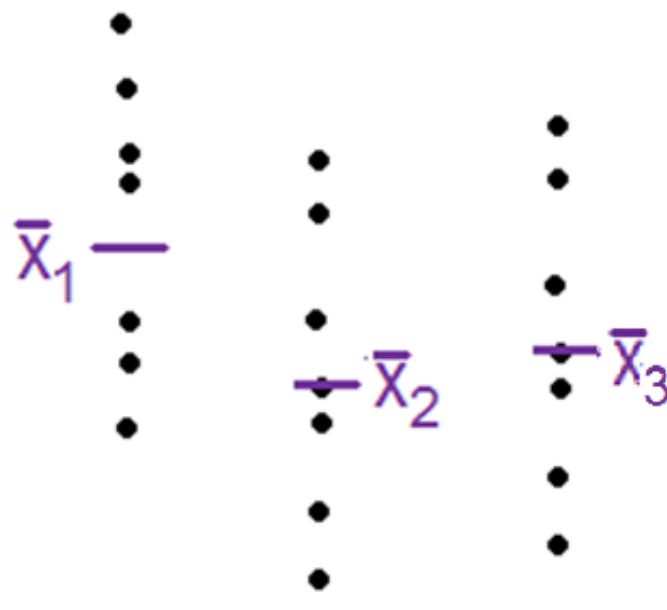
Please axolotl questions if you are confused.



MS stands for ***mean square***, and MS_{within} is the average squared difference from a data point to the average for the that group. It's the mean squared WITHIN a group

If we were just looking at a single group, this average squared distance would be the standard deviation squared, or the ***variance***.

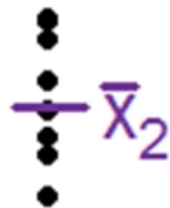
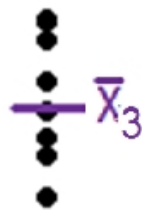
MS_{within} is large when the spread within the samples is large.



$$F = \frac{MS_{\text{between}}^{\text{small}}}{MS_{\text{within}}^{\text{large}}}$$

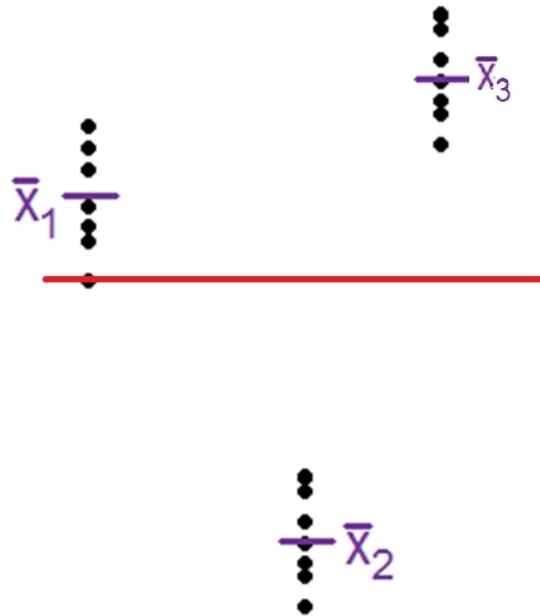
Spread/variance within a sample makes it hard to detect differences between the samples, and so the F-statistics gets smaller, just like the t-statistic.

MS_{between} is large when there are large differences between the sample means. MS_{between} stands for the differences between means, instead of within them.



$$F = \frac{MS_{\text{between}}^{\text{large}}}{MS_{\text{within}}^{\text{small}}}$$

Here, the average (squared) difference from a group mean to the **grand mean**, the average of data points from all the groups put together, is much larger than the differences between each point and its group mean.



$$F = \frac{MS_{\text{between}}^{\text{large}}}{MS_{\text{within}}^{\text{small}}}$$

F will be large and there is strong evidence that there is some difference in the true means between the groups.



ANOVA is a big-picture tool.

If the ANOVA F-test yields a small p-value, that means the sample means are far enough to reject the hypothesis that the difference between true means is zero.

It also means that some of the variance is explained by groups.

In correlation, the closer values get to a straight line, the more variance is explained (r^2 gets closer to 1)

In ANOVA, the closer values get to their group means, the more variance is explained (again, proportion explained gets closer to 1)

Just as when X has nothing to do with Y in correlation/regression $r^2=0$, if the group has nothing to do with the measured values, none of the variance is explained.

Yarn Breakage Example

Consider the dataset called “warpbreaks”.

The responses are counts of number of breaks in yarn (breaks) of two different kinds of wool (wool) under three different levels in tension (tension).

The number of breaks is the response (Y), and the levels of tension are the explanatory variable (X).

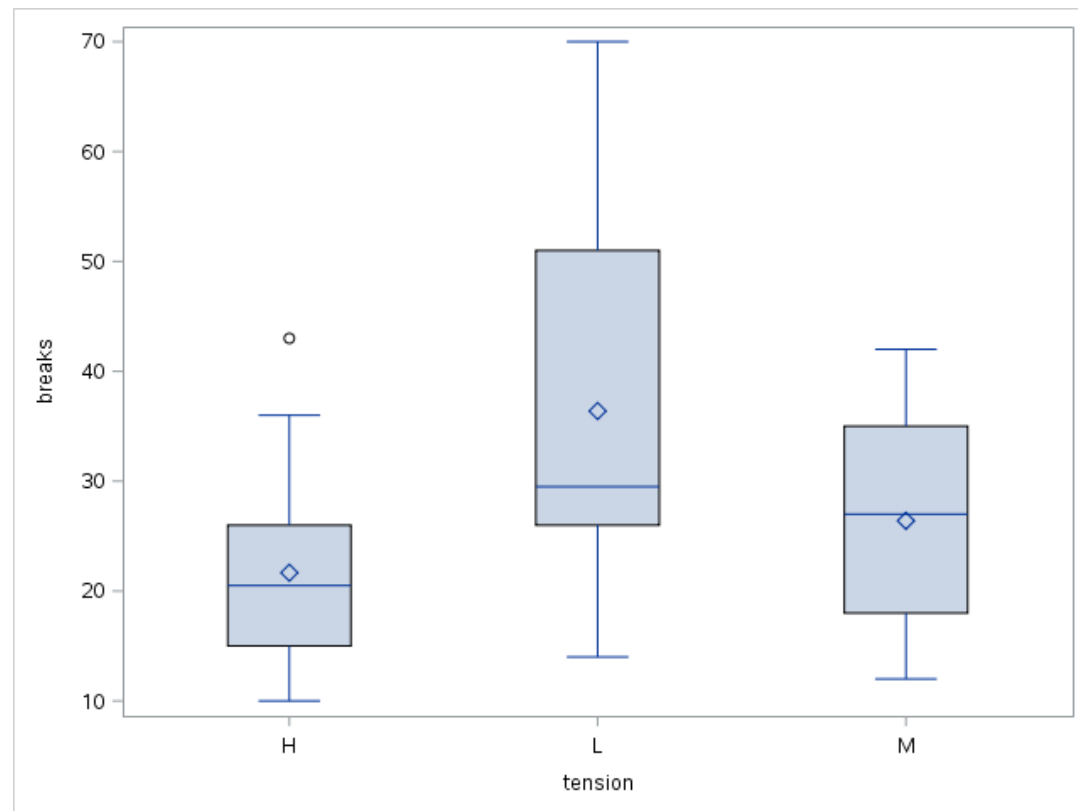
The basic syntax using proc glm for this is...

```
proc glm data=warpbreaks;  
  class tension;  
    /*Tells SAS that 'tension' is categorical */  
  model breaks = tension;  
    /*Set up the explanatory, response vars */  
run;
```

In ANOVA, the variance within each group is assumed to be the same. We can check this assumption informally with a side-by-side boxplot.

If a box is much larger than the others, ANOVA won't work as well.


```
proc sgplot data=warpbreaks;  
    vbox breaks / category=tension;  
run;
```



More formally, we can do either a **Levene test** or a **Bartlett test** on the data, a test of equal variance.

For three groups (for both tests), this is:

$$H_o: \sigma_1 = \sigma_2 = \sigma_3$$

H_A : At least one pair of variances is unequal.

```

proc glm data=warpbreaks;
    class tension;
    model breaks = tension;
    means tension / hovtest=levene;
run;

```

Levene's Test for Homogeneity of breaks Variance ANOVA of Squared Deviations from Group Means					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
tension	2	404326	202163	5.28	0.0082
Error	51	1952401	38282.4		

```
proc glm data=warpbreaks;
  class tension;
  model breaks = tension;
  means tension / hovtest=bartlett;
run;
```

Bartlett's Test for Homogeneity of breaks Variance			
Source	DF	Chi-Square	Pr > ChiSq
tension	2	9.6571	0.0080

Both tests detected unequal variance. They will typically agree very closely, especially when the sample sizes are equal, so choosing just one is fine.

ANOVA is typically robust to unequal variance, but to be safe, we're going to use the WELCH ANOVA as well as the standard ANOVA and see if there's a difference in the results.

We do this by adding the 'welch' option to the 'means' statement in proc glm.

The following will produce standard anova output, the test for equal variance, and the alternative welsh anova.

```
proc glm data=warpbreaks;  
    class tension;  
    model breaks = tension;  
    means tension / hovtest=bartlett welch;  
run;
```

The standard ANOVA output looks like this:

The GLM Procedure					
Dependent Variable: breaks					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	2034.259259	1017.129630	7.21	0.0018
Error	51	7198.555556	141.148148		
Corrected Total	53	9232.814815			

R-Square	Coeff Var	Root MSE	breaks Mean
0.220329	42.20732	11.88058	28.14815

The Welch ANOVA output looks like this:

Welch's ANOVA for breaks			
Source	DF	F Value	Pr > F
tension	2.0000	5.80	0.0070
Error	32.3200		

What's different between these two ANOVAs?

- The detail. Sum of squares, and hence R-squared are not meaningful terms in Welch's ANOVA, so they don't show.
- The DF, or Degrees of Freedom. Welch's ANOVA doesn't have DF in the typical sense, but has an 'equivalent DF'. This will always be less than the standard DF, and is based on how unequal the variances are.
- The P-value. By making a less strict assumption, Welch's p-value is larger, which is the trade-off for discarding the assumption of equal variance.

What's the same between these two ANOVAs?

- Both the ANOVAs give a small p-value, < 0.01 . This indicates strong evidence that the means are different, DESPITE any additional confusion introduced by unequal variance.

Recommendation: Since the unequal variances are not problematic, report the standard ANOVA.

The most important parts are in yellow.

	Df	Sum Sq	Mean Sq	F-value	Pr(>F)
tension	2	2034	1017	7.20	.001753
Residuals	51	7199	141		

The p-value is especially important because it shows what you really want to know:

Are the group means different or not?

In this case, yes.

We can introduce a second variable by adding it to the explanatory variables in the model. Make sure to also add it to the list of categorical variables in the 'class' statement.

```
proc glm data=warpbreaks;  
    class tension wool;  
        /*Both of these are categorical */  
    model breaks = tension wool;  
        /*Use two explanatory variables together */  
run;
```

The output you want is separated into three tables.

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	2484.925926	828.308642	6.14	0.0012
Error	50	6747.888889	134.957778		
Corrected Total	53	9232.814815			

R-Square	Coeff Var	Root MSE	breaks Mean
0.269141	41.27139	11.61713	28.14815

Source	DF	Type I SS	Mean Square	F Value	Pr > F
tension	2	2034.259259	1017.129630	7.54	0.0014
wool	1	450.666667	450.666667	3.34	0.0736

We can also introduce an interaction by adding it to the model. Interactions are marked with an asterisk *. This is true even if one or both the variables involved are numeric.

Note that we don't need to specify that the interaction is categorical; it's implied.

```
proc glm data=warpbreaks;  
    class tension wool;  
    model breaks = tension wool tension*wool;  
run;
```

The output of the two-way ANOVA with an interaction is very similar.

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	3487.703704	697.540741	5.83	0.0003
Error	48	5745.111111	119.689815		
Corrected Total	53	9232.814815			

R-Square	Coeff Var	Root MSE	breaks Mean
0.377751	38.86680	10.94028	28.14815

Source	DF	Type I SS	Mean Square	F Value	Pr > F
tension	2	2034.259259	1017.129630	8.50	0.0007
wool	1	450.666667	450.666667	3.77	0.0582
tension*wool	2	1002.777778	501.388889	4.19	0.0210

One drawback of ANOVAs is that we can't, by default, test between individual category means. ANOVA looks each categorical variable holistically and determines if there are any differences between the means from there. It does NOT however, tell you which means are different.

We can investigate individual group means with either a 'means' statement or an 'lsmeans' statement. Each of these statements has different capabilities.

With the 'means' statement, we can compare individual means, while making multiple comparison adjustments.

Multiple comparison adjustments are necessary because without them, your chance of falsely finding a significant difference explodes with the number of groups that you are comparing.

With 3 groups, we are making 3 comparisons,

With 4 groups, we are making 6 comparisons,

With 8 groups, we are making 28 comparisons.

There are three multiple comparisons we're going to consider.

The BONFERRONI comparison just takes the p-value of each comparison between two means from t-test and multiplies the p-value by the number of comparisons M (or alternatively, divides the pairwise alpha by M)

$\alpha_{\text{test}} = 0.05 / 3 = 0.0167$ with 3 groups,

$\alpha_{\text{test}} = 0.05 / 10 = 0.005$ with 5 groups, and

$\alpha_{\text{test}} = 0.05 / 28 = 0.0018$ with 8 groups.

The SIDAK option also adjusts the p-value, but accounts for the possibility of multiple false positives. It also assumes that each comparison is an independent trial.

The Bonferroni adjustment makes no assumption about the relationship between comparisons, but it also tends to be overly conservative.

We can test for both using separate 'means' statements.

```
proc glm data=warpbreaks;  
    class tension wool;  
    model breaks = tension wool  
tension*wool;  
    means wool / t;  
    means tension / bon;  
    means tension / sidak;  
  
run;
```

The grouping letters tell us with pairs were found to be significantly different or not.

Tension L is in a grouping all on its own, and
Tensions M and H have a grouping together.

Means with the same letter are not significantly different.			
Bon Grouping	Mean	N	tension
A	36.389	18	L
B	26.389	18	M
B			
B	21.667	18	H

Tension L (Low) and tension M (medium) are significantly different.

tension L and tension H (medium) are significantly different.

But

tension M and tension H are NOT significantly different.

Means with the same letter are not significantly different.			
Bon Grouping	Mean	N	tension
A	36.389	18	L
B	26.389	18	M
B			
B	21.667	18	H

Under the Sidak adjustment, we come to the same conclusions (typically this is the case, although Sidak may find a few significant differences that Bonferroni does not)

Means with the same letter are not significantly different.			
Sidak Grouping	Mean	N	tension
A	36.389	18	L
B	26.389	18	M
B			
B	21.667	18	H

The other table that each of these adjustments provides tells you the difference in the means that would be necessary to determine a significant difference.

Bonferroni (Dunn) t Tests for breaks

Alpha	0.05
Error Degrees of Freedom	48
Error Mean Square	119.6898
Critical Value of t	2.48078
Minimum Significant Difference	9.0468

Sidak t Tests for breaks

Alpha	0.05
Error Degrees of Freedom	48
Error Mean Square	119.6898
Critical Value of t	2.47392
Minimum Significant Difference	9.0218

When there are only two groups, only one comparison is made, so we can use the t-test.

For the two types of wool, no difference in means is found.

Means with the same letter are not significantly different.			
t Grouping	Mean	N	wool
A	31.037	27	A
A			
A	25.259	27	B

As with other procs, you can specify alpha. However, it's done in the means statement and not the setting.

```
proc glm data=warpbreaks;  
    class tension wool;  
    model breaks = tension wool  
tension*wool;  
    means wool / t (alpha=0.01) ;  
    means tension / bon(alpha=0.01) ;  
    means tension / sidak(alpha=0.01) ;  
  
run;
```

Bonferroni and Sidak work as adjustments to the alpha, so they work for ANY set of multiple comparisons, including those for regression. For ANOVA specifically, there is Tukey's Honestly Significant Difference.

```
proc glm data=warpbreaks;  
    class tension wool;  
    model breaks = tension wool  
tension*wool;  
    means tension / tukey;  
    means tension*wool / tukey;  
run;
```

Notice that Tukey's is even more efficient (requires a smaller minimum difference) than either Bonferroni or Sidak.

Tukey's Studentized Range (HSD) Test for breaks

Alpha	0.05
Error Degrees of Freedom	48
Error Mean Square	119.6898
Critical Value of Studentized Range	3.42021
Minimum Significant Difference	8.8195

Means with the same letter are not significantly different.

Tukey Grouping	Mean	N	tension
A	36.389	18	L
B	26.389	18	M
B			
B	21.667	18	H

However, it doesn't give the full results for the interaction.
This is an oversight.

Level of tension	Level of wool	N	breaks	
			Mean	Std Dev
H	A	9	24.5555556	10.2726714
H	B	9	18.7777778	4.8933061
L	A	9	44.5555556	18.0977285
L	B	9	28.2222222	9.8587243
M	A	9	24.0000000	8.6602540
M	B	9	28.7777778	9.4310362



Tukey:

More than just a turkey and a toucan.

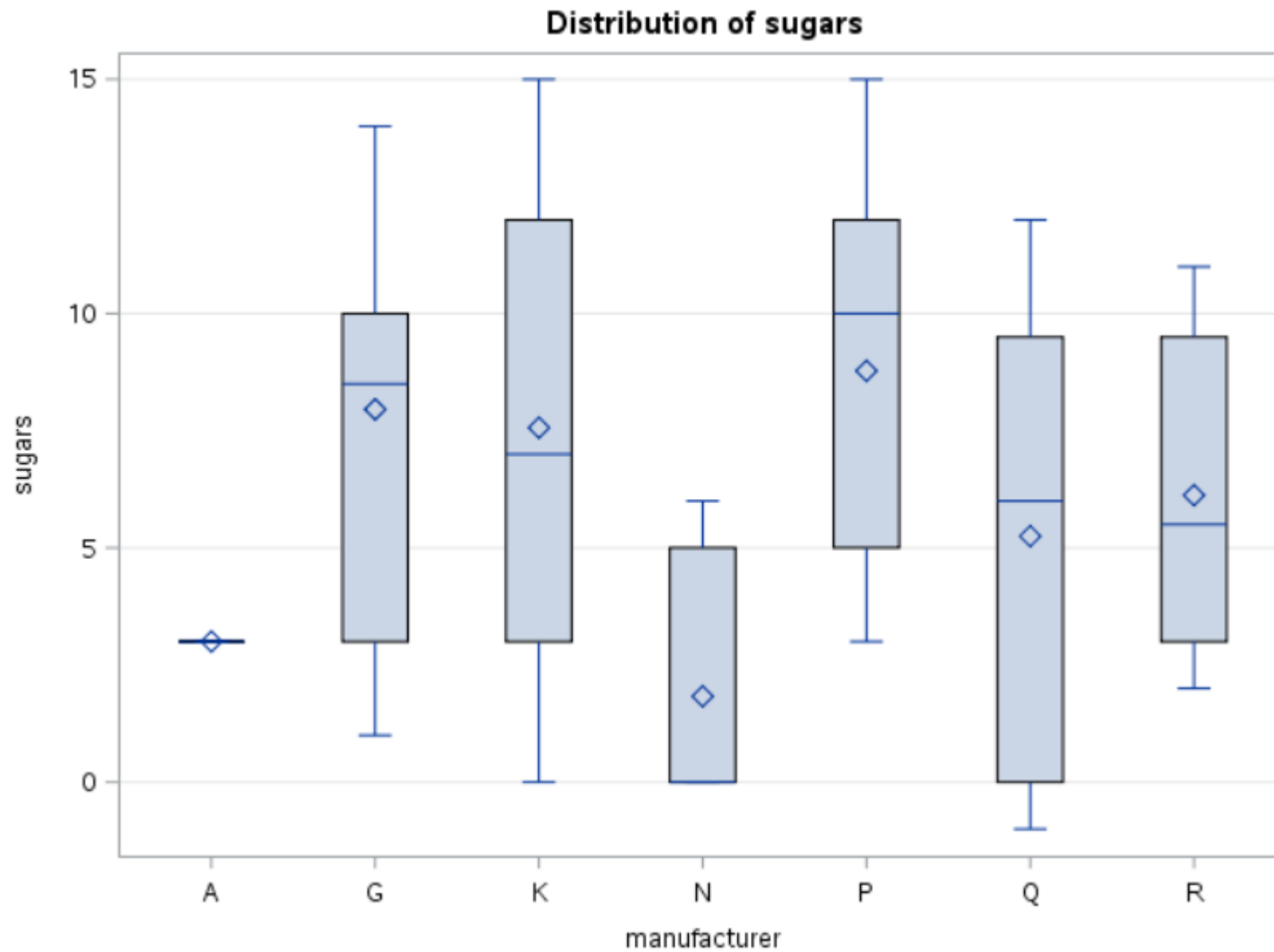
Preview for next time, we will use proc GLM run models that use both continuous and categorical variables.

```
proc glm data = cereal;  
    class manufacturer;  
    model sugars= manufacturer protein fat  
sodium;  
run;
```

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	9	575.109798	63.901089	4.62	<.0001
Error	67	926.422670	13.827204		
Corrected Total	76	1501.532468			

R-Square	Coeff Var	Root MSE	sugars Mean
0.383015	53.71934	3.718495	6.922078

Source	DF	Type I SS	Mean Square	F Value	Pr > F
manufacturer	6	262.1618593	43.6936432	3.16	0.0086
protein	1	117.7968587	117.7968587	8.52	0.0048
fat	1	186.4820380	186.4820380	13.49	0.0005
sodium	1	8.6690417	8.6690417	0.63	0.4313



Alpha	0.05
Error Degrees of Freedom	67
Error Mean Square	13.8272
Critical Value of Studentized Range	4.29875

Comparisons significant at the 0.05 level are indicated by ***.				
manufacturer Comparison	Difference Between Means	Simultaneous 95% Confidence Limits		
P - G	0.823	-3.649	5.296	
P - K	1.213	-3.232	5.657	
P - R	2.653	-2.840	8.145	
P - Q	3.528	-1.965	9.020	
P - A	5.778	-6.137	17.692	
P - N	6.944	0.987	12.902	***
G - P	-0.823	-5.296	3.649	
G - K	0.389	-2.981	3.760	
G - R	1.830	-2.837	6.496	
G - Q	2.705	-1.962	7.371	
G - A	4.955	-6.603	16.512	
G - N	6.121	0.915	11.327	***
K - P	-1.213	-5.657	3.232	