9. Analysis of Variance (ANOVA)

Video Link:
https://www.youtube.com/watch?v=-
q0oEYccwkE&index=9&list=PL2fQHGEDK7Yyl1W9tgIo8wpYFTDumgc_j

Section 9.1: Categorical Explanatory Variable with 2 Levels
Section 9.2: Categorical Explanatory Variable with More than 2 Levels

Section 9.1: Categorical Explanatory Variable with 2 Levels

Now that you have a feel for analysis of variance we'll run the test using SPSS. We'll use an example first described in hypothesis testing.

Is major depression associated with a smoking quantity among current young adult smokers?
Or in hypothesis testing terms, "Are the mean number of cigarettes smoked per month equal or not equal for those individuals with and without major depression?"

The explanatory variable here is categorical with two levels, **MAJORDEPLIFE**, that is, the presence or absence of major depression. The response variable of **NUMCIGMO_EST**, smoking quantity, measured by the number of cigarettes smoked per month ranges from 1 to 2940.

1. Go through the appropriate steps to add **MAJORDEPLIFE** to your Variable Set. It is already a variable in the dataset. These steps are shown in 05. Data Management tutorial steps 5.3 #5 through #8.

2. Then label the dummy codes for **MAJORDEPLIFE**. These steps are shown in .04 Working with Data tutorial steps 4.5 #1 through #2. A dummy code of 0 represents Absence of Major Depression and a 1 represents Presence of Major Depression.

3. Go to **Analyze > Compare Means > One-Way ANOVA...**
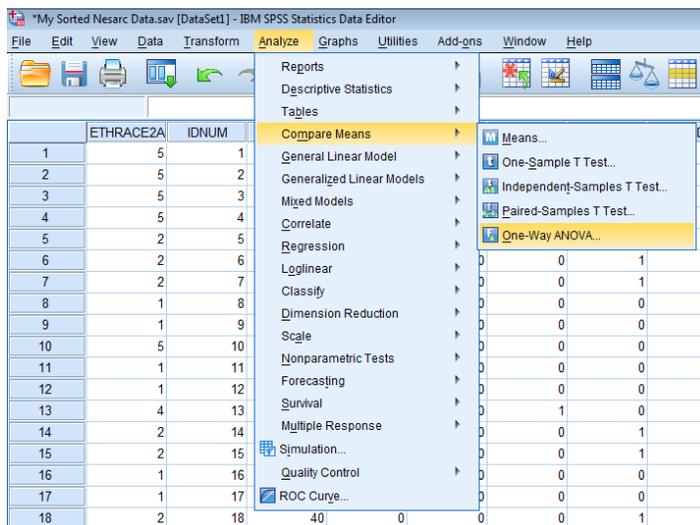
4. From the variables list on the left, use the arrows to put your categorical explanatory variable, **MAJORDEPLIFE,** in the **Factor:** window and your quantitative response variable, **NUMCIGMO_EST,** in the **Dependent List:** window.
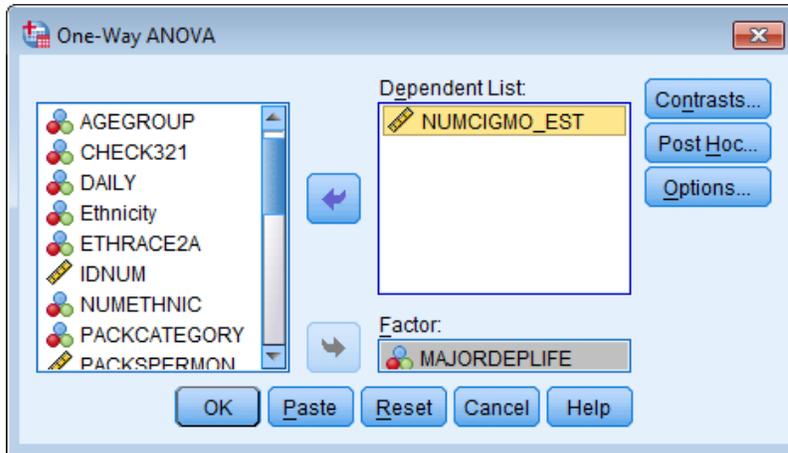
5. Click the **Options…** button and check the box by **Descriptive**. Click **Continue** then **OK**.
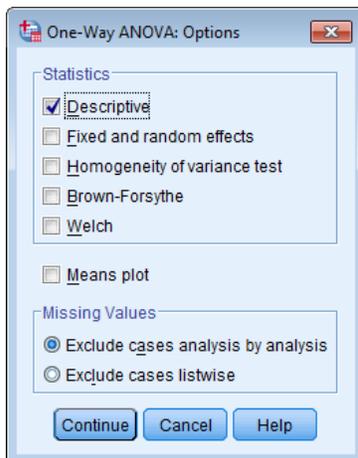
The first table in our output is the **Descriptives** table that shows us different values for each level of the categorical explanatory variable.

Descriptives

number of cigarettes smoked per month

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
Absence of Major Depression	1253	312.8380	269.00234	7.59942	297.9290	327.7470	1.00	2940.00
Presence of Major Depression	444	341.3750	288.49512	13.69137	314.4669	368.2831	1.00	2400.00
Total	1697	320.3044	274.43678	6.66195	307.2379	333.3709	1.00	2940.00

The first column, **N**, is the number of participants for each level and then total number of participants data that was analyzed with ANOVA. So here we see our categorical explanatory variable MAJORDEPLIFE has two levels with each categories label. We can see that 1697 observations were included in the analysis. We know from previous analyses that our total sample size is 1706 young adult smokers. Observations were excluded from the analysis because the participant did not have valid data for one or both of our variables.

The **Mean** column shows that young adult smokers without major depression, as indicated by a value of zero, smoke an average of 313 cigarettes per month, and that those with major depression, indicated by a value of 1, smoke on average 341 cigarettes per month.

The second table in our output is the **ANOVA** table. Directly below the table title you will see the variable label for your quantitative response variable **NUMCIGMO_EST**.

ANOVA

number of cigarettes smoked per month

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	266974.123	1	266974.123	3.550	.060
Within Groups	127468189.9	1695	75202.472		
Total	127735164.0	1696			

Our calculated F statistic found in the **F** column in this output, is 3.55. The significance, probability, or p value, associated with this F statistic, is labeled **Sig.**, and as you can see, the p value is .060, just over our p .05 cut point.

Because the p value is greater than 0.05, actually 0.06, we must accept the null hypothesis, and say that these means are statistically equal, and that there is no association between the presence or absence of major depression and the number of cigarettes smoked per month, among young adult smokers.

If I chose to reject the null hypothesis, I would be wrong 6 out of 100 times, and again, by normal scientific standards, this is not adequate certainty to reject the null hypothesis and say there is an association. Instead, we're going to accept the null hypothesis, and say that there is no association.

Had the p value been less than .05, I would know that there was a significant association. To interpret that as significant association, I would look at the means table. If p would have been less than .05, I can see that individuals with major depression smoke more than individuals without.

And again with a significant P value, I could have said that young adult smokers with major depression smoke significantly more cigarettes per month than young adult smokers without major depression.

So we've shown you the ropes in terms of a categorical explanatory variable that has 2 levels, as it did here with depression. For this interpretation, all we need to know is the p value and the means for each of the two groups.

Section 9.2: Categorical Explanatory Variable with More than 2 Levels

What if my categorical explanatory variable has more than two groups? In this example, I'm going to examine the association between ethnicity and smoking quantity. The explanatory variable, ethnicity, actually has five levels or groups. One is white, two is black, three is American Indian Alaskan native, four Asian native, Hawaiian Pacific Islander, and five, Hispanic or Latino.

So by running an analysis of variance, we're asking whether the number of cigarettes smoked differs for different ethnic groups.

1. Repeat steps 9.1 #3 through #5 but instead of using MAJORDEPLIFE for the **Factor:** use ETHRACE2A.

Descriptives

number of cigarettes smoked per month

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
1	1054	368.7865	281.43073	8.66864	351.7767	385.7963	1.00	2400.00
2	210	259.2738	278.67739	19.23057	221.3631	297.1846	1.00	2940.00
3	42	310.9881	260.11696	40.13692	229.9299	392.0462	2.00	1200.00
4	58	244.2586	195.07644	25.61479	192.9659	295.5514	2.50	1200.00
5	333	219.7583	220.85937	12.10302	195.9500	243.5665	1.00	1200.00
Total	1697	320.3044	274.43678	6.66195	307.2379	333.3709	1.00	2940.00

ANOVA

number of cigarettes smoked per month

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	6965167.919	4	1741291.980	24.396	.000
Within Groups	120769996.1	1692	71377.066		
Total	127735164.0	1696			

This time the F statistic is 24.4 with an associated p value is .000. Well, this tells me I can safely reject the null hypothesis, and say that there is an association between ethnicity and smoking quantity.

But all I know at this point is that not all the means are equal. I could eyeball each mean in the means table and make a guess as to which pairs are significantly different from one another.

For example, the ethnic group with the lowest mean number of cigarettes smoked per month among young adult smokers is ethnic group 5, Hispanic or Latino, and the group with the highest number of cigarettes smoked per month is ethnic group 1, White.

The F test and p value do not provide insight into why the null hypothesis can be rejected because there are multiple levels to my categorical explanatory variable. They do not tell us in what way the population means are not statistically equal.

Note that there are many ways for population means not to be all equal. Having each of them not equal to the other is just one of them. Another way could be that only two of the populations are not equal to one another.

In the case where the explanatory variable represents more than two groups a significant ANOVA does not tell us which groups are different from the others. To determine which groups are different from the others, we would need to perform a post hoc test. A post hoc test conducts "post hoc paired comparisons." Post hoc means after the fact. And these post hoc paired comparisons must be conducted in a particular way, in order to prevent excessive type 1 error. Type 1 error, as you'll recall, occurs when you make an incorrect decision about the null hypothesis, that is, you reject the null hypothesis when the null hypothesis is true.

Why can't we just perform multiple ANOVAs? That is, why can't we just subset our observations, and take 2 at a time. That is, compare White versus Black; White versus American Indian, Alaskan Native, etcetera, until all the paired comparisons have been made.

As you know, we accept significance and reject the null hypothesis at p less than or equal to 0.05, a 5% chance that we are wrong and have committed a type 1 error.

There's actually a 5% chance of making a type 1 error for each analysis of variance that we conduct on this question. Therefore, performing multiple tests means that our overall chance of committing type 1 error could be far greater than 5%.

Here is how it works out. Using the formula displayed under this table,

Family-Wise Type 1 Error Rate

# Tests	Comparison α	Family-wise α
1	.05	.05
3	.05	.14
6	.05	.26
10	.05	.40
15	.05	.54

$$\alpha_{FW} = 1 - (1 - \alpha_{PC})^c$$

Where c = # of comparisons, α=normal Type 1 Error (.05)

you can see that while one test has a type one error rate of .05, by the time we've conducted 10 tests on this question our chance of rejecting the null hypothesis when the null hypothesis is true is up to 40%. This increase in the type one error rate is called the Family-Wise error rate and is the error rate for the group of pair comparison.

Post hoc tests are designed to evaluate the difference between pairs of means while protecting against the inflation of type 1 error's and there are a lot of post hoc tests to choose from when it comes to analysis of variance. There's the Sidak, and the Holm T test, and Fisher's Least Significant Difference Test, Tukey's Honestly Significant Difference test, the Scheffe test, the Newman-Keuls test, Dunnett's Multiple Comparison test, the Duncan Multiple Range test, the Bonferroni Procedure.

It's enough to make your head swim!

While there are certainly differences in how conservative each test is in terms of protecting against type one error, in many cases it is far **less** important **which** post hoc test you conduct and far **more** important that you **do** conduct one.

In order to conduct post hoc paired comparisons in the context of my ANOVA examining the association between ethnicity and number of cigarettes smoked per month, I'm going to use the Duncan Multiple Range test.

2. Go to **Analyze > Compare Means > One-Way ANOVA...**
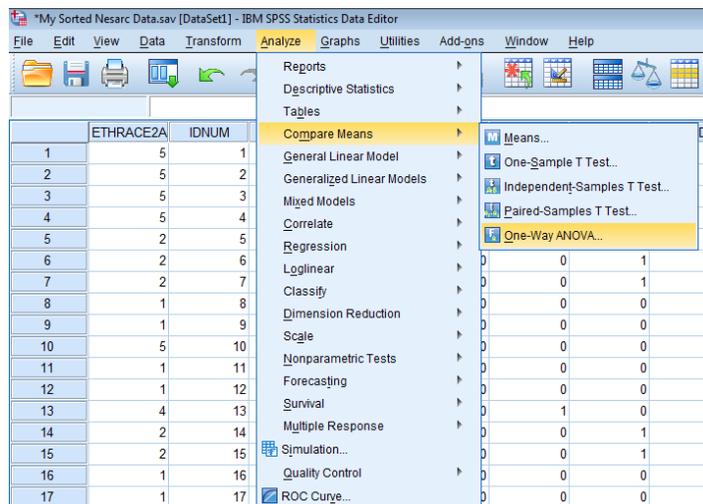
3. The variables that you previous used to determine if the overall ANOVA was significant or not will still be listed in the window for **Dependent List:** and **Factor:**
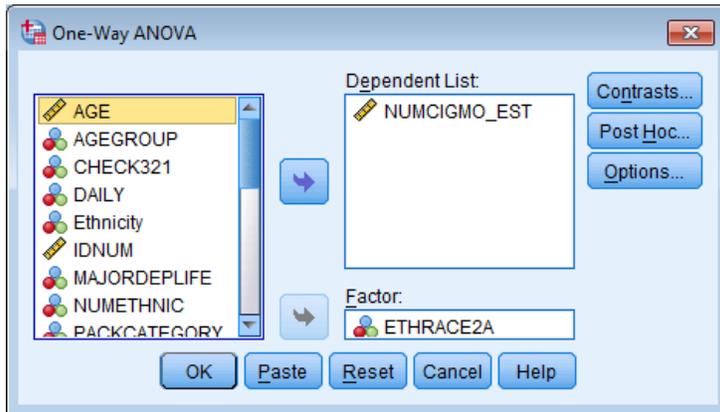
4. Click **Options…**, uncheck **Descriptive** as we already have this output, click **Continue**.
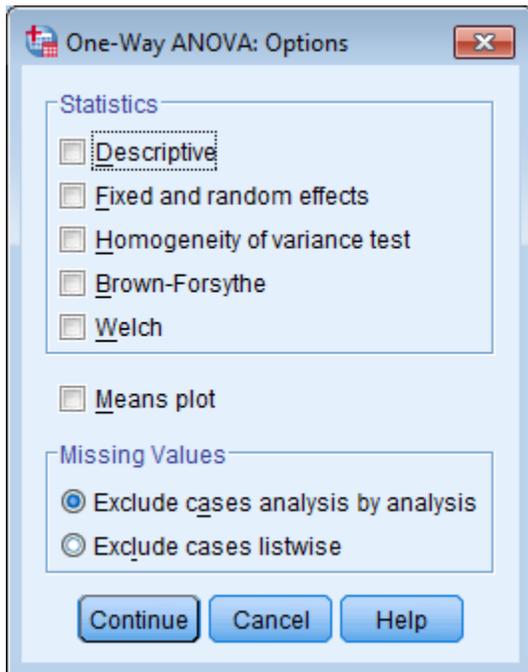
5. Then click **Post Hoc...**
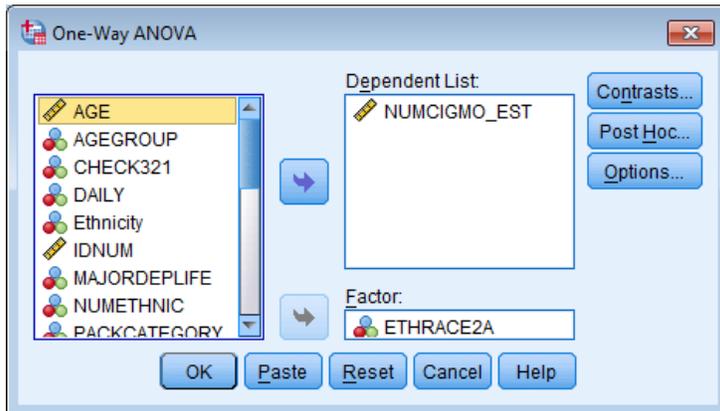
6. Check the box next to **Duncan**, click **Continue > OK**.
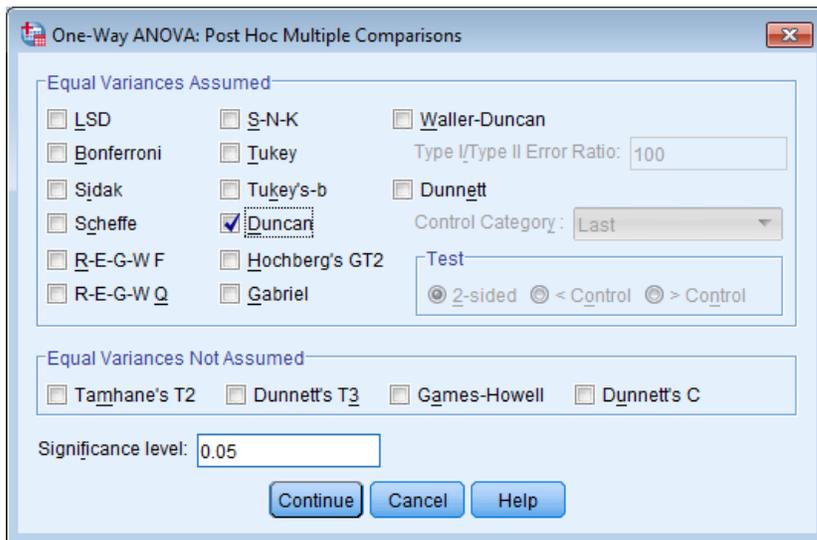
The following table will be generated in your output.

number of cigarettes smoked per month

Duncan[a,b]

ETHRACE2A	N	Subset for alpha = 0.05		
		1	2	3
5	333	219.7583		
4	58	244.2586	244.2586	
2	210	259.2738	259.2738	
3	42		310.9881	310.9881
1	1054			368.7865
Sig.		.327	.094	.125

Means for groups in homogeneous subsets are displayed.

a. Uses Harmonic Mean Sample Size = 100.473.

b. The group sizes are unequal. The harmonic mean of the group sizes is used. Type I error levels are not guaranteed.

Here is how the table should be interpreted. Basically, MEANS that appear in the same column together are not significantly different.

In column 1, groups 5, 4, and 2 appear together; therefore, groups 2 & 4, 2 & 5, and 4 & 5 are not significantly different.

In column 2, groups 4, 2, and 3 appear together; therefore, groups 2 & 4 (which we previously determined), 2 & 3, and 3 & 4 are not significantly different.

In column 3, groups 1 and 3 appear together; therefore, these are not significantly different.

So where are the significant differences? Group 1, which indicates White ethnicity, smokes significantly more cigarettes per month, than ethnic groups 2, 4 and 5 (i.e. Black, Asian, and Hispanic or Latino) because those group numbers never appear together in any of the 3 columns. Group 3, American Indian/Alaskan Native, smoke significantly more per month than group 5, Hispanic or Latino because those group numbers never appear together in any of the 3 columns.

Notice that some of the means appear in more than one column so you need to be careful and follow the rule that means that appear in the same column are not significantly different from one another. To help reduce your confusion write down each group number compared to every other group number prior to looking at the table.

1v2
1v3 2v3
1v4 2v4 3v4
1v5 2v5 3v5 4v5

Then as you see two group numbers appear in the same column mark that off your list as not significant. Whichever comparisons are unmarked once you have worked through the table are statistically significant.

We've covered a lot of information, so here's a quick summary of analysis of variance.

First, an analysis of variance is used when we have a categorical explanatory variable and a quantitative response variable and we want to examine differences in the mean response variable for each category of our explanatory variable.

The null hypothesis is that there is no relationship between the explanatory and response variable. In other words, that all means are equal. The alternate hypothesis is that not all means are equal.

The F statistic is calculated by comparing the variation among sample means to the variation within groups. If the variation among sample means wins out, the p value will be less than or equal to .05, and we have a significant finding. This would allow us to reject the null hypotheses, and say that the explanatory and response variables are associated.

If your explanatory variable has more than two levels or groups, you'll also need to conduct a Duncan's post hoc test.

Now you are ready to test a categorical to quantitative relationship. If your own research question does not include these types of variables, you might want to test the procedure with variables from your data set that do require an ANOVA. For example, you could look at mean age differences according to any categorical variable in the NESARC. Or, treating the grade level variable in ADDHEALTH as quantitative, you could look at mean differences in grade level, again, by any categorical variable that you choose. For both Gap-minder and the Mars craters data, there are many quantitative variables, so you might choose to categorize one of them for inclusion in an ANOVA.

Whatever types of variables you have, you will be able to test the association with the right tool.