10. Chi-Square

Video Link:
https://www.youtube.com/watch?v=h791-9JZK5E&list=PL2fQHGEDK7Yyl1W9tgIo8wpYFTDumgc_j&index=10

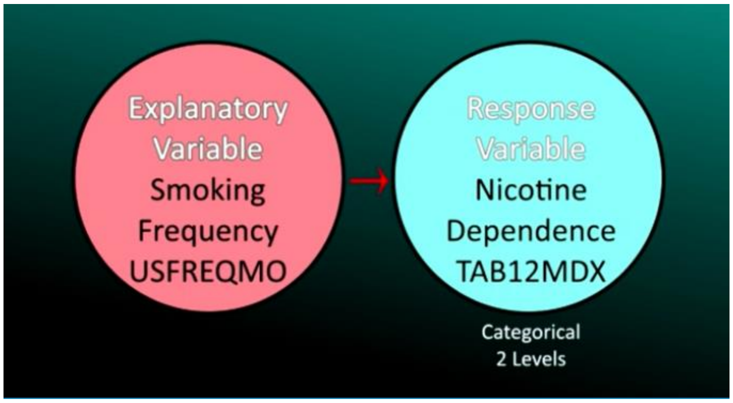
Section 10.1: Categorical Explanatory Variable and Categorical Response Variable
Section 10.2: Post Hoc tests for Categorical Explanatory with More than 2 Levels

Section 10.1: Categorical Explanatory Variable and Categorical Response Variable

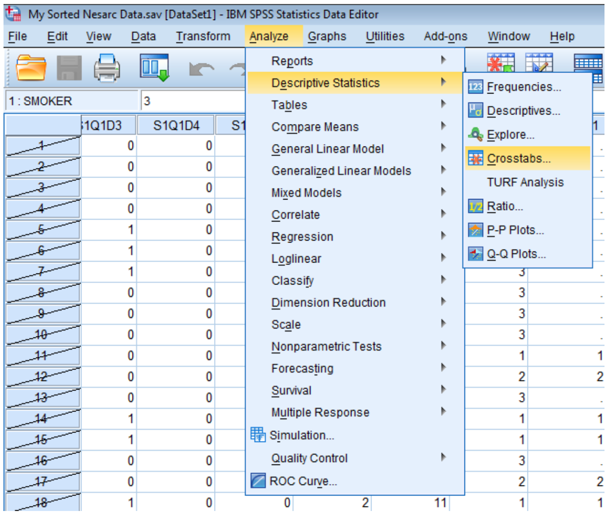
Now we're going to use the Chi Square test of Independence to test the hypothesis proposed about smoking frequency and nicotine dependence from working with NESARC data. Specifically, is how often a person smokes related to nicotine dependence among current young adult smokers? Or in hypothesis testing terms, is smoking frequency and nicotine dependence independent or dependent. That is, are the rates of nicotine dependence equal or not equal among individuals from my different smoking frequency categories?

For this analysis, we're going to use a categorical explanatory variable with 6 levels, the number of days smoked per month, which we called USFREQMO, with the following categorical values: smoking approximately 1 day/month, 2.5 days/ month, 5 days/month, 14 days/month, 22 days/month and 30 days/month.

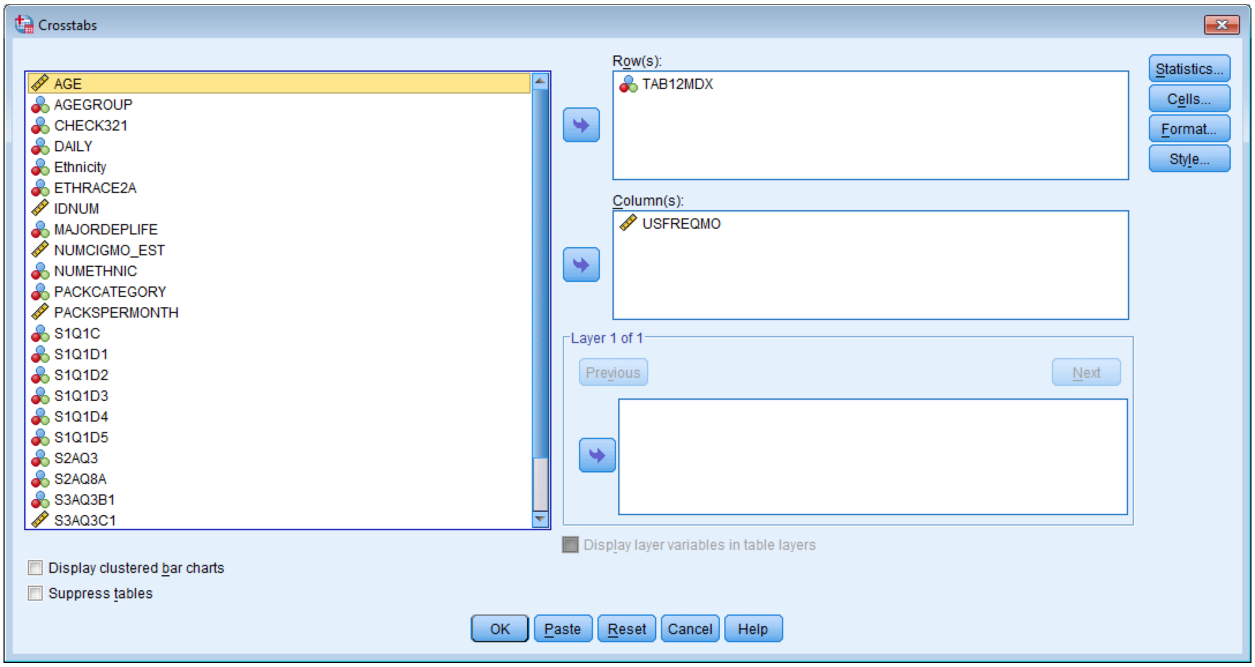
The response variable, called TAB12MDX, is categorical with 2 levels--the presence or absence of nicotine dependence in the past 12 months.



1. Click **Analyze > Descriptive Statistics > Crosstabs**.



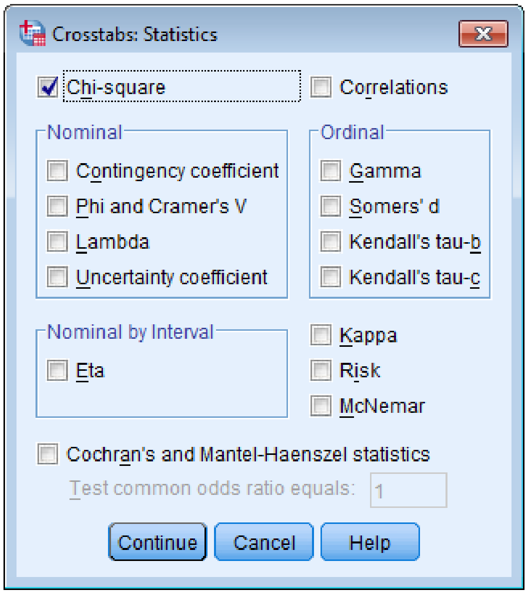
2. Using the arrows move your categorical explanatory variable to the window labeled **Column(s):** and your categorical response variable to the window labeled **Row(s):** then click **Statistics**.


Crosstabs

Row(s): TAB12MDX

Column(s): USFREQMO

3. Check **Chi-square** then click **Continue**.



Crosstabs: Statistics

☑ Chi-square ☐ Correlations

Nominal
☐ Contingency coefficient
☐ Phi and Cramer's V
☐ Lambda
☐ Uncertainty coefficient

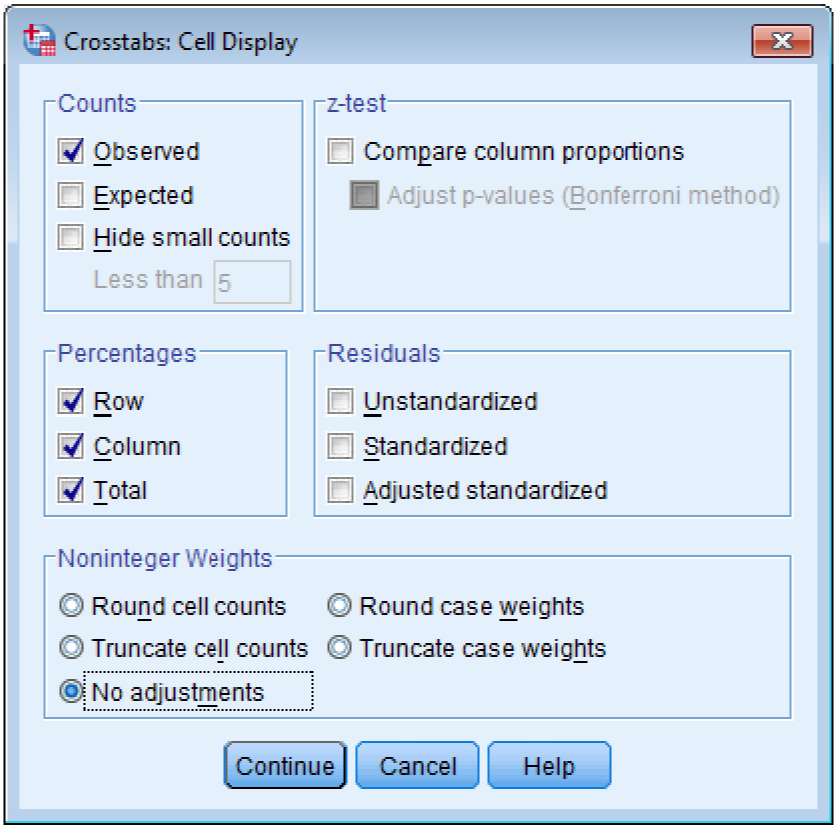
Ordinal
☐ Gamma
☐ Somers' d
☐ Kendall's tau-b
☐ Kendall's tau-c

Nominal by Interval
☐ Eta

☐ Kappa
☐ Risk
☐ McNemar

☐ Cochran's and Mantel-Haenszel statistics
Test common odds ratio equals: 1

4. Click **Cells…** Ensure that **Observed** is checked. Check **Row, Column**, and **Total** in the Percentages box. In the bottom box click **No adjustments**, then click **Continue > OK**.



Crosstabs: Cell Display

Counts
☑ Observed
☐ Expected
☐ Hide small counts
Less than 5

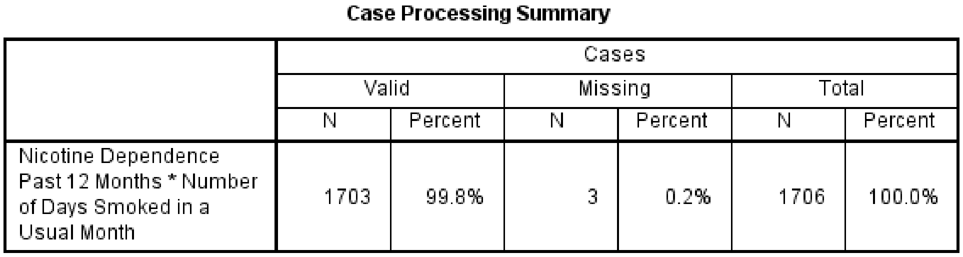
z-test
☐ Compare column proportions
☐ Adjust p-values (Bonferroni method)

Percentages
☑ Row
☑ Column
☑ Total

Residuals
☐ Unstandardized
☐ Standardized
☐ Adjusted standardized

Noninteger Weights
○ Round cell counts ○ Round case weights
○ Truncate cell counts ○ Truncate case weights
● No adjustments

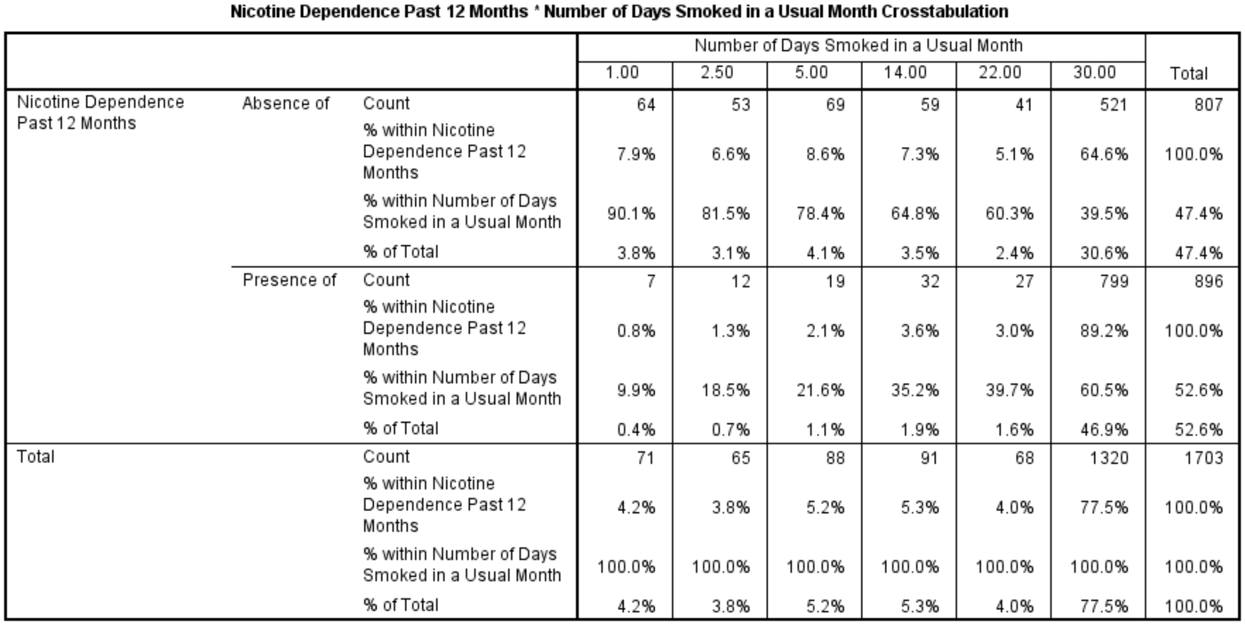
The first table, **Case Processing Summary** table, in the SPSS output shows you the number of participants used in the analysis in the **Valid** column, the number of participants with **Missing** data, therefore, not used in the analysis, and the **Total** number of participants.



Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Nicotine Dependence Past 12 Months * Number of Days Smoked in a Usual Month	1703	99.8%	3	0.2%	1706	100.0%

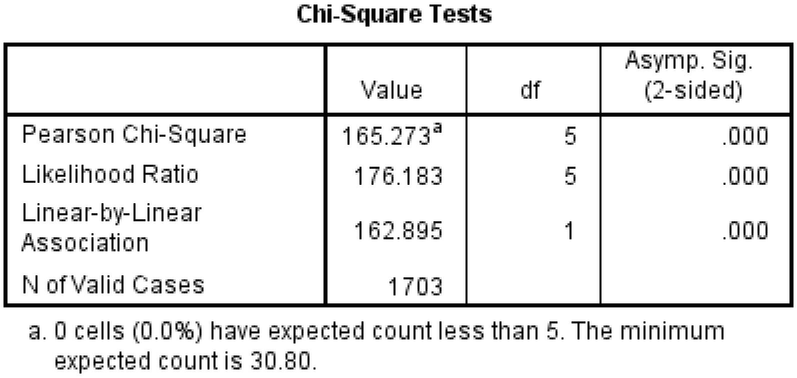
The table below is of the response variable by the explanatory variable. This table is known as the cross tabs or cross tabulation table where you can see a myriad of numbers and percentages with such labels as **Count** (i.e., frequency), **% within Nicotine Dependence** (i.e., row percentage for response variable), **% within Number of Days Smoked** (i.e., column percentage for explanatory variable)., and **% of Total**.



Nicotine Dependence Past 12 Months * Number of Days Smoked in a Usual Month Crosstabulation

			Number of Days Smoked in a Usual Month						Total
			1.00	2.50	5.00	14.00	22.00	30.00	
Nicotine Dependence Past 12 Months	Absence of	Count	64	53	69	59	41	521	807
		% within Nicotine Dependence Past 12 Months	7.9%	6.6%	8.6%	7.3%	5.1%	64.6%	100.0%
		% within Number of Days Smoked in a Usual Month	90.1%	81.5%	78.4%	64.8%	60.3%	39.5%	47.4%
		% of Total	3.8%	3.1%	4.1%	3.5%	2.4%	30.6%	47.4%
	Presence of	Count	7	12	19	32	27	799	896
		% within Nicotine Dependence Past 12 Months	0.8%	1.3%	2.1%	3.6%	3.0%	89.2%	100.0%
		% within Number of Days Smoked in a Usual Month	9.9%	18.5%	21.6%	35.2%	39.7%	60.5%	52.6%
		% of Total	0.4%	0.7%	1.1%	1.9%	1.6%	46.9%	52.6%
Total		Count	71	65	88	91	68	1320	1703
		% within Nicotine Dependence Past 12 Months	4.2%	3.8%	5.2%	5.3%	4.0%	77.5%	100.0%
		% within Number of Days Smoked in a Usual Month	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
		% of Total	4.2%	3.8%	5.2%	5.3%	4.0%	77.5%	100.0%

The **Chi-Square Tests** table below shows the calculation of the chi square statistic along with the associated p-value. You will only use the **Pearson Chi-Square** row in this table. Our p-value of .0001 clearly tells us that smoking and nicotine dependence are associated.



Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	165.273[a]	5	.000
Likelihood Ratio	176.183	5	.000
Linear-by-Linear Association	162.895	1	.000
N of Valid Cases	1703		

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 30.80.

A chi square table (i.e. cross tabulation table or cross tabs) can be very confusing on first examination. Before we try to interpret this output, let's look at 3 different tables that pull apart the different numbers represented in a cross tabs.

For our example, we're going to use percentages from a Chi -Square table examining the distribution of insured and uninsured individuals by geographic region.

Table A shows ROW percentages.

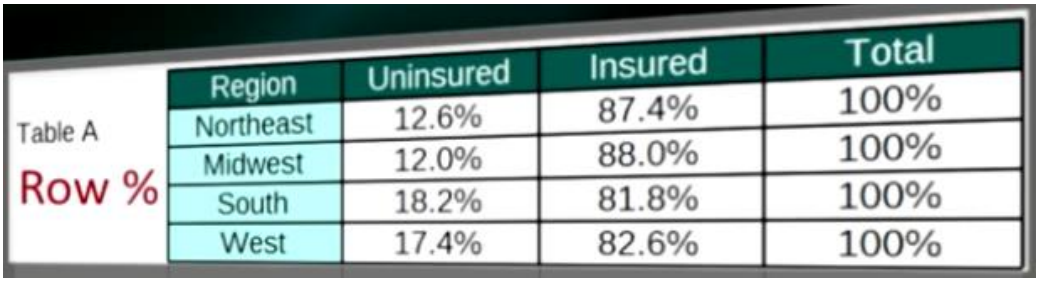


Table A

Row %

Region	Uninsured	Insured	Total
Northeast	12.6%	87.4%	100%
Midwest	12.0%	88.0%	100%
South	18.2%	81.8%	100%
West	17.4%	82.6%	100%

Each cell includes the percent of observations within each row i.e. within region Northeast, Midwest, South, and West that are either insured or uninsured. As you can see, adding across the rows gives us 100 percent of the observations within region.

Table B includes the total percent of observations in each cell. Here, the percentages in each row and column add up to a 100 percent.

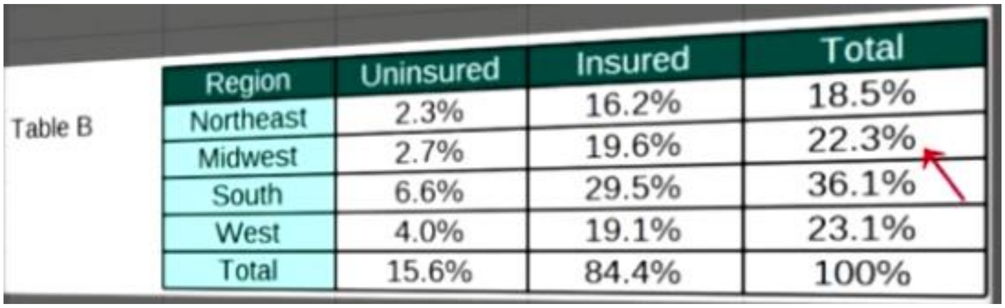
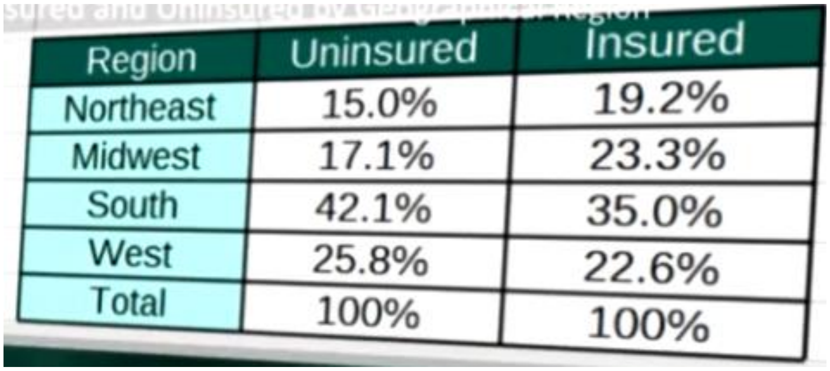


Table B

Region	Uninsured	Insured	Total
Northeast	2.3%	16.2%	18.5%
Midwest	2.7%	19.6%	22.3%
South	6.6%	29.5%	36.1%
West	4.0%	19.1%	23.1%
Total	15.6%	84.4%	100%

Finally, Table C shows column percentages. Each cell includes the percent of observations within each column that is within either the insured or uninsured group.

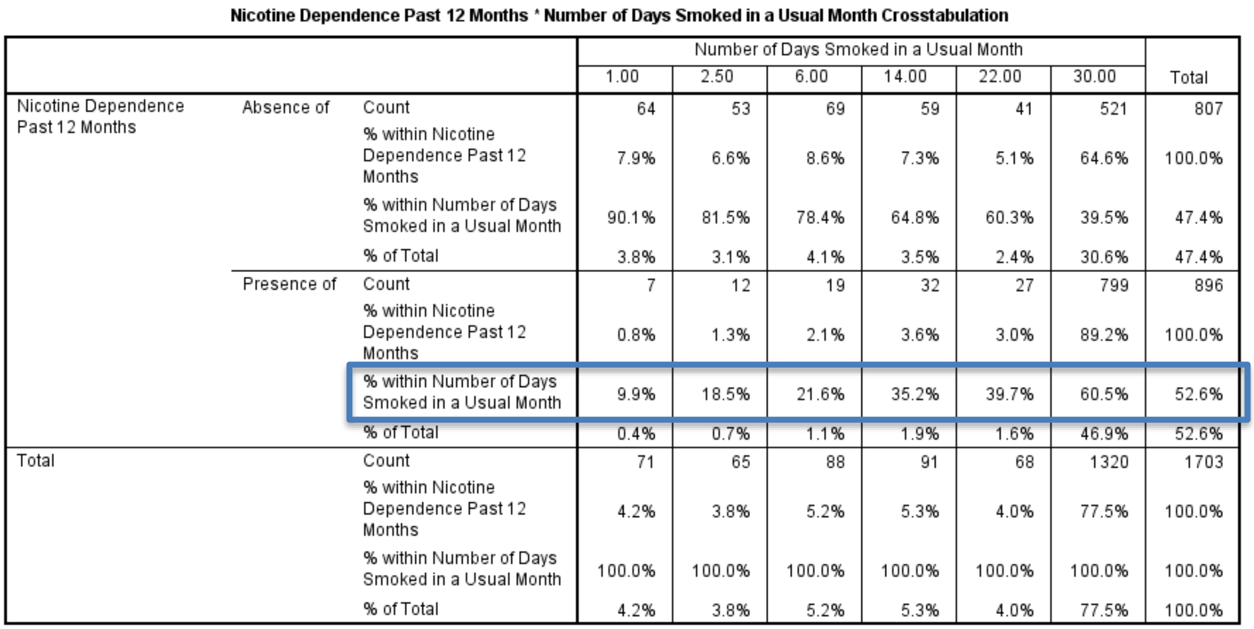


sured and Uninsured by Geographical Region

Region	Uninsured	Insured
Northeast	15.0%	19.2%
Midwest	17.1%	23.3%
South	42.1%	35.0%
West	25.8%	22.6%
Total	100%	100%

Adding down the columns gives us 100% of observation by insurance status.

So which of these percentage types should we examine when trying to interpret the Chi-Square results for smoking frequency and nicotine dependence?



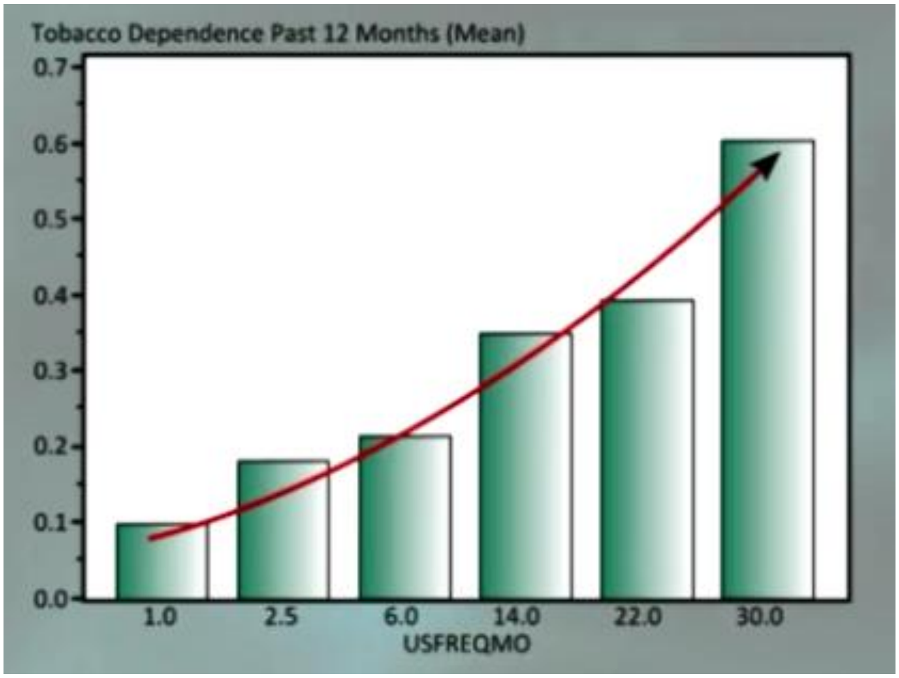
Nicotine Dependence Past 12 Months * Number of Days Smoked in a Usual Month Crosstabulation

			Number of Days Smoked in a Usual Month						Total
			1.00	2.50	6.00	14.00	22.00	30.00	
Nicotine Dependence Past 12 Months	Absence of	Count	64	53	69	59	41	521	807
		% within Nicotine Dependence Past 12 Months	7.9%	6.6%	8.6%	7.3%	5.1%	64.6%	100.0%
		% within Number of Days Smoked in a Usual Month	90.1%	81.5%	78.4%	64.8%	60.3%	39.5%	47.4%
		% of Total	3.8%	3.1%	4.1%	3.5%	2.4%	30.6%	47.4%
	Presence of	Count	7	12	19	32	27	799	896
		% within Nicotine Dependence Past 12 Months	0.8%	1.3%	2.1%	3.6%	3.0%	89.2%	100.0%
		% within Number of Days Smoked in a Usual Month	9.9%	18.5%	21.6%	35.2%	39.7%	60.5%	52.6%
		% of Total	0.4%	0.7%	1.1%	1.9%	1.6%	46.9%	52.6%
Total		Count	71	65	88	91	68	1320	1703
		% within Nicotine Dependence Past 12 Months	4.2%	3.8%	5.2%	5.3%	4.0%	77.5%	100.0%
		% within Number of Days Smoked in a Usual Month	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
		% of Total	4.2%	3.8%	5.2%	5.3%	4.0%	77.5%	100.0%

If the output is set with the explanatory variable categories across the top of the table and response variable categories down the side, it will be the column percent's that we want to interpret. In other words, we're interested in whether the rate of nicotine dependence differs according to which explanatory group the observations belong to i.e. which smoking frequency group.

Notice that we are not interested in the column percentages for those observations without nicotine dependence, indicated with a dummy code of 0 (i.e., Absence of). Instead, we're interested in describing the presence of nicotine dependence within the smoking frequency groups that is these column percentages (i.e., % within Number of Days Smoked) circled with blue.

5. If I use SPSS to generate a graph of the percent of young adult smokers with nicotine dependence within each smoking frequency category, I could visualize the association, and see that there seems to be a positive linear relationship. For a reminder of the steps return to SPSS tutorial 7. Bivariate Graphing section 7.2 steps #1 through #4.



We can see that the more days per month a young adult smokes, the more likely they are to have nicotine dependence. I know from looking at the significant p-value of .0001, that I will accept the alternate hypothesis that not all nicotine dependents rates are equal across smoking frequency categories. If my explanatory variable had only two levels, I could interpret the two corresponding column percentages and be able to say which group had a significantly higher rate of nicotine dependence. But my explanatory variable had six categories, so I know that not all are equal. I don't know which are different and which are not.

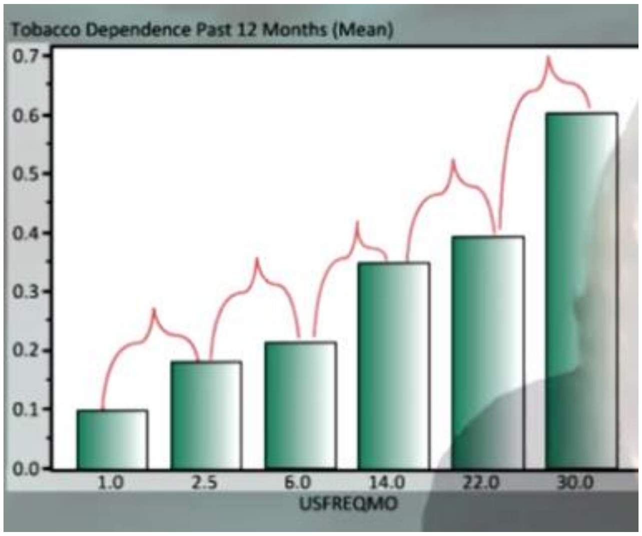
Section 10.2: Post Hoc tests for Categorical Explanatory with More than 2 Levels

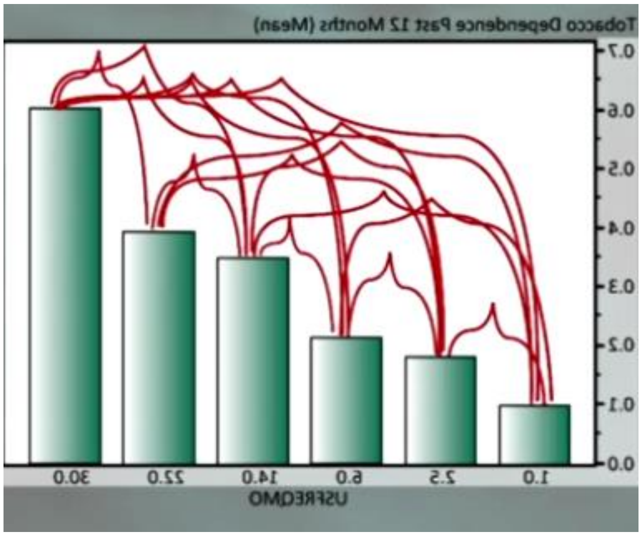
When the explanatory variable has more than two levels, the chi square statistic and associated p-value do not provide insight into why the null hypothesis can be rejected. It does not tell us what way the rates of nicotine dependence are not equal across the frequency categories. There are of course many ways for the rates to be unequal. Having each of them as unequal to the other is just one of them.

Maybe there are only two of the population rates that are not equal to one another. To determine which groups are different from the others we will again need to perform a post hoc test. By conducting post hoc comparisons between pairs of rates in a way that avoids excessive type 1 error--In other words, avoids rejecting the null hypothesis when the null hypothesis is true--We will be much better able to appropriately describe which population rates are different from the others.

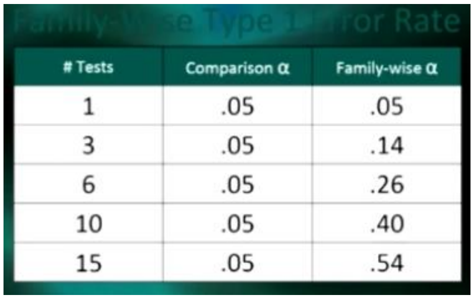
If we reject the null hypothesis, we need to perform comparisons for each pair of nicotine dependence rates across the six smoking frequency categories. In the case of six groups we actually need to perform 15 pair wise comparisons.

With these red brackets, I'm illustrating eight of the fifteen paired comparisons that we'll need to conduct. As you can see, there are so many it's actually difficult to illustrate this graphically.





If you will recall the family wise error rate for 15 different comparisons is .54.



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

This means that if we do not protect against type I error, we will be wrongly rejecting the null hypothesis and saying that there is an association over half the time.

Having about a 50/50 change of being right would obviously give us absolutely no confidence in our decisions.

So, to appropriately protect against type 1 error in the context of a chi square test we will use the post hoc approach known as the Bonferroni Adjustment. The goal of using the Bonferroni adjustment is to control the family wise error rate, also known as the maximum overall type 1 error rate, so that we can evaluate which pairs of nicotine dependence rates are different from one another.

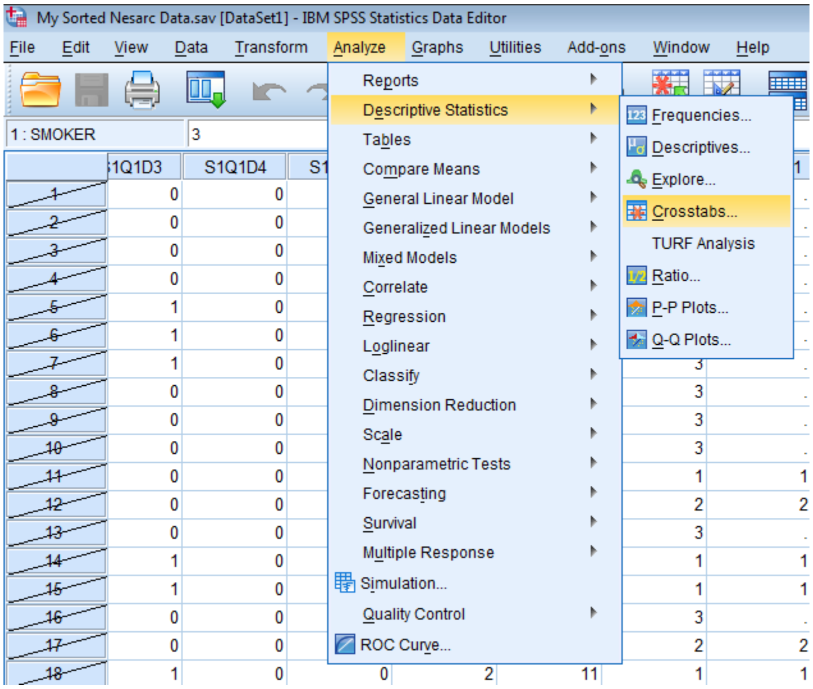
Briefly, the process would be to conduct each of the 15 paired comparisons, but rather than evaluating significance at the p.05 level, SPSS will automatically adjust the p-value to make it more difficult to reject the null hypothesis.

The adjusted p-value is calculated by dividing .05 by the number of comparisons that we plan to make. So if we make three comparisons, we would only reject the null hypothesis if the p-value were .017 or less.

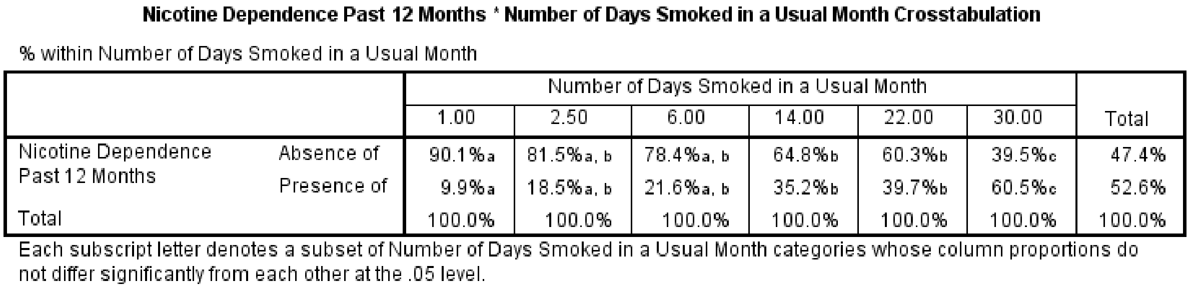
For the fifteen paired comparisons that we plan to make to better understand the association between smoking frequency and nicotine dependence, SPSS will adjust our p-value to .003.

For the actual post hoc testing, we need to run a chi-square test for each of the 15 paired comparisons.

1. Click **Analyze > Descriptive Statistics > Crosstabs**.



2. You will see that your explanatory and response variables are still in the correct spots. Click **Cells…** Uncheck **Observed**, **Row**, and **Total**. In the upper right click **Compare column proportions** and **Adjust p-values (Bonferroni method)** then click **Continue > OK**. You will get a duplicate of the Case Processing Summary and Chi-Square Tests table that we previously viewed. We are interested in the second table.



Nicotine Dependence Past 12 Months * Number of Days Smoked in a Usual Month Crosstabulation

% within Number of Days Smoked in a Usual Month

		Number of Days Smoked in a Usual Month						Total
		1.00	2.50	6.00	14.00	22.00	30.00	
Nicotine Dependence Past 12 Months	Absence of	90.1%$_{a}$	81.5%$_{a, b}$	78.4%$_{a, b}$	64.8%$_{b}$	60.3%$_{b}$	39.5%$_{c}$	47.4%
	Presence of	9.9%$_{a}$	18.5%$_{a, b}$	21.6%$_{a, b}$	35.2%$_{b}$	39.7%$_{b}$	60.5%$_{c}$	52.6%
Total		100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%

Each subscript letter denotes a subset of Number of Days Smoked in a Usual Month categories whose column proportions do not differ significantly from each other at the .05 level.

By unchecking the **Observed**, **Row**, and **Total** we get a table that shows just the column percentages. We want to focus on the **Presence of** row like we did previously.

3. Write out each of the 15 comparison that we are going to examine to ensure we do not overlook any of the comparisons.

1 v 2.5
1 v 6   2.5 v 6
1 v 14   2.5 v 14   6 v 14
1 v 22   2.5 v 22   6 v 22   14 v 22
1 v 30   2.5 v 30   6 v 30   14 v 30   22 v 30

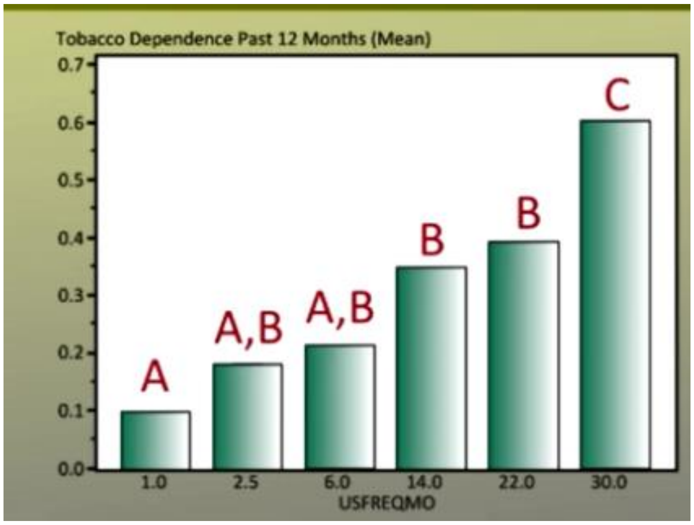
Look at the **Presence of row** in the table above. To the right of each percentage is a letter or letters. If one of the same letters appears next to both explanatory levels you are comparing that means those two levels are *not* statistically different from each other.

For example, for Number of Days Smoked in a Usual Month comparison of 1 v 2.50 we can see both have the letter 'a' next to the percentages of 9.9% and 18.5%, therefore, those two levels are not statistically different from each other. So I want to accept the null hypothesis since this probability value is not only NOT less than 0.05%, it is definitely NOT less than my Bonferroni adjusted p value of 0.003.

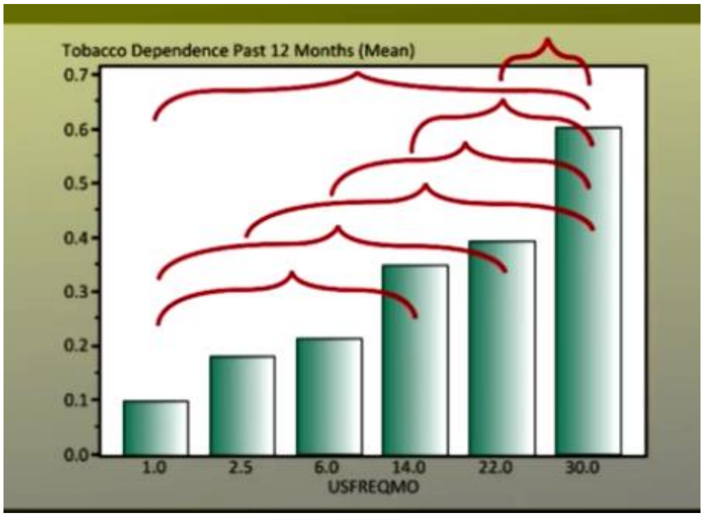
For comparison 1 v 30 we see column percentages of 9.9% and 60.5% and neither level has the same letter next to the percentage, therefore, those two explanatory levels are significantly different from each other.

Using the letter convention, in which nicotine dependence rates with the same letter are not significantly different, these post hoc findings can be pictured like this.

Here is another way we could picture the significant differences between rates. As you can see, the more differences there are, the more challenging the visualization can be to create.



Here is another way we could picture the significant differences between rates. As you can see, the more differences there are, the more challenging the visualization can be to create.

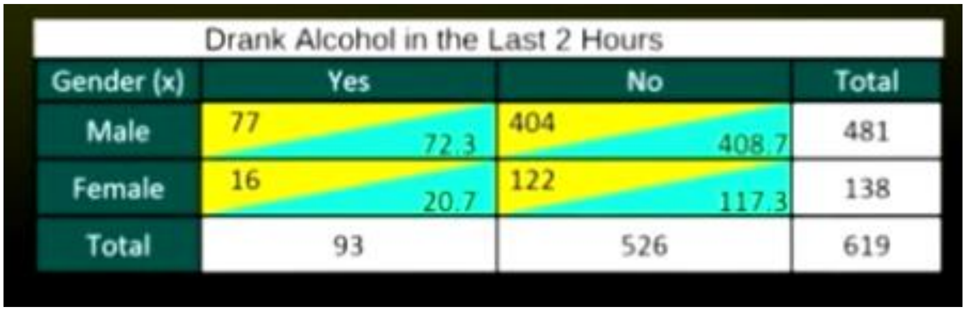


Let's quickly summarize Chi-square tests of independence.

First, a Chi square test of independence is used when we have a categorical explanatory variable and a categorical response variable.

The null hypothesis is that there is no relationship between the 2 categorical variables, they are independent. And the alternate hypothesis is that there is a relationship between the 2 categorical variables, and they are not independent.

The chi square statistic is calculated considering both the observed and expected counts in each of the table's cells.



Drank Alcohol in the Last 2 Hours			
Gender (x)	Yes	No	Total
Male	77 72.3	404 408.7	481
Female	16 20.7	122 117.3	138
Total	93	526	619

If your explanatory variable has more than two levels or groups, you'll also need to conduct a post-hoc test. We use the Bonferroni Adjustment to protect against type 1 error and then run the Chi square tests of independence for each paired comparison.

Now you are ready to test a categorical by categorical relationship. Also, if your own research question is a quantitative to categorical relationship, it's a good idea for you to categorize the quantitative explanatory variable and test the association with the chi square test of independence.