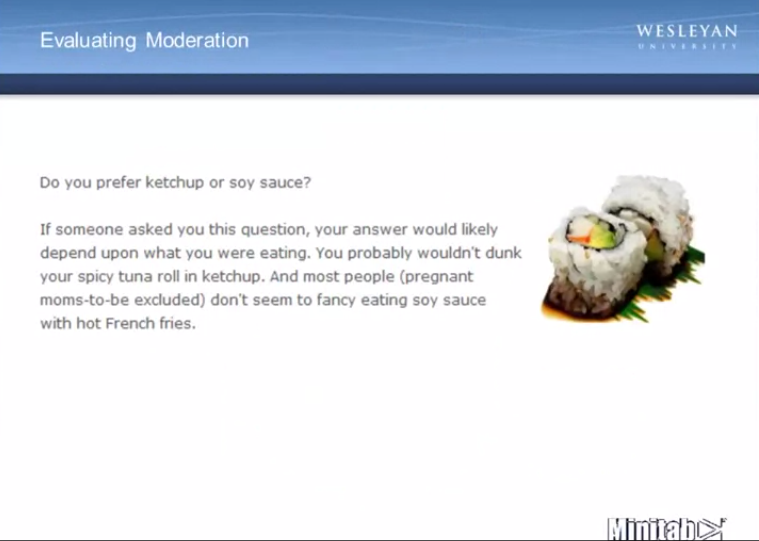
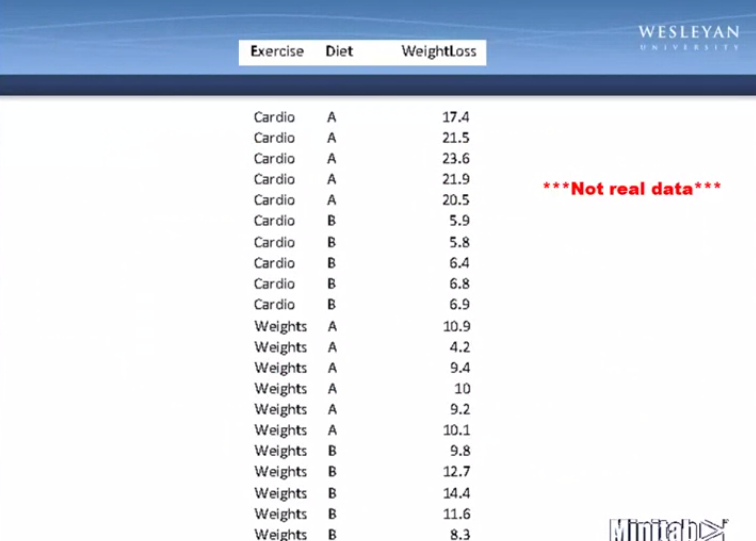
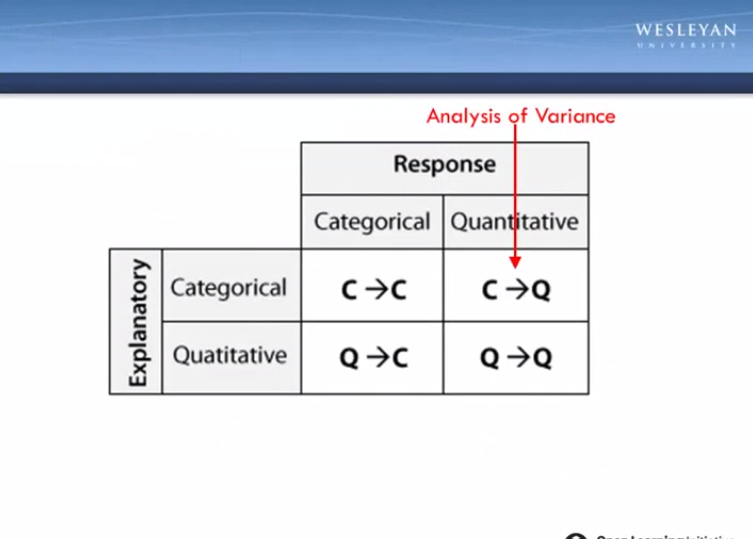
MODERATION



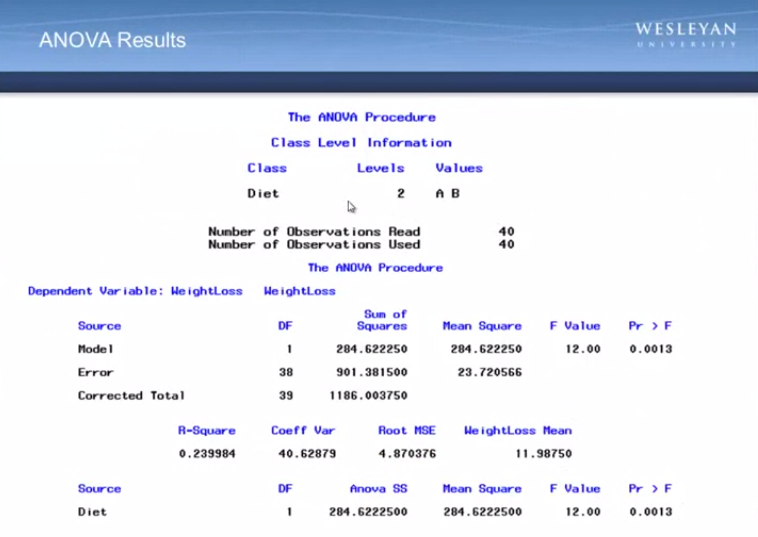
In this presentation, I want to talk to you about the concept of moderation. So far, in the models that we've been examining based on analysis of variance, chi-square, or Pearson coorelation. While our findings have suggested either statistical significance or non-significance, it is important to understand that these results represent what is true on average for the whole population. But what if our association of interest differs for different sub-groups of the population?



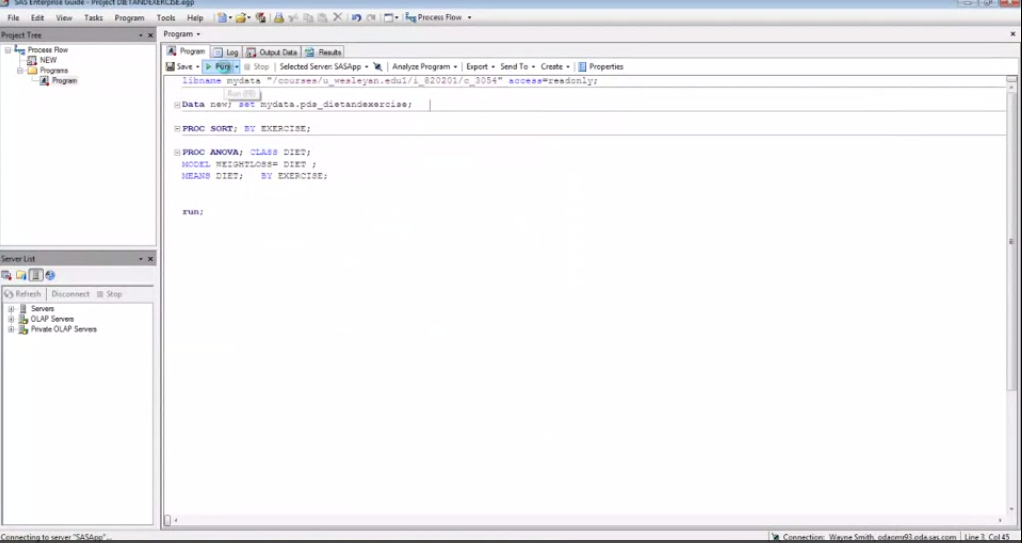
I like the examples that I found recently on a Minitab software blog. Think about this question. Do you prefer ketchup or soy sauce? Obviously, if someone asks you this question, your answer would probably depend upon what you're eating. Soy sauce may be preferred if you're eating sushi, and ketchup if you're eating French fries. So, the answer to questions differs based on a additionally categorical variable, that is what kind of food you're eating. So, let's just use a weight loss study example to illustrate how we can evaluate whether third variable may moderate the association between our two variables of interest. We'll be evaluating the association between diet plan A and B, our explanatory variable, and weight loss, our quantitative response variable. Here is a snapshot of the data.

As you can see, for each observation that is participant, we've measured and recorded which diet they were assigned to A or B and their weight loss one month later. We've also recorded a third variable, that is, which exercise program they followed, either cardio or weight training. Please know as we go through this example that this is not data drawn from an actual study, but only a hypothetical case to help us understand moderation.

Since we have a categorical explanatory variable, diet plan A or B, and a quantitative response variable, that is weight loss. We will of course need to use analysis of variance to evaluate the association. And this model SAS syntax should look familiar to you. Following PROC ANOVA, in our class statement we'll include our categorical explanatory variable and in our model statement. We will include the quantitative response variable equal to the categorical explanatory variable. Finally, the categorical explanatory variable will be included in our means statement. For our diet and weight loss example, the syntax would look like this.



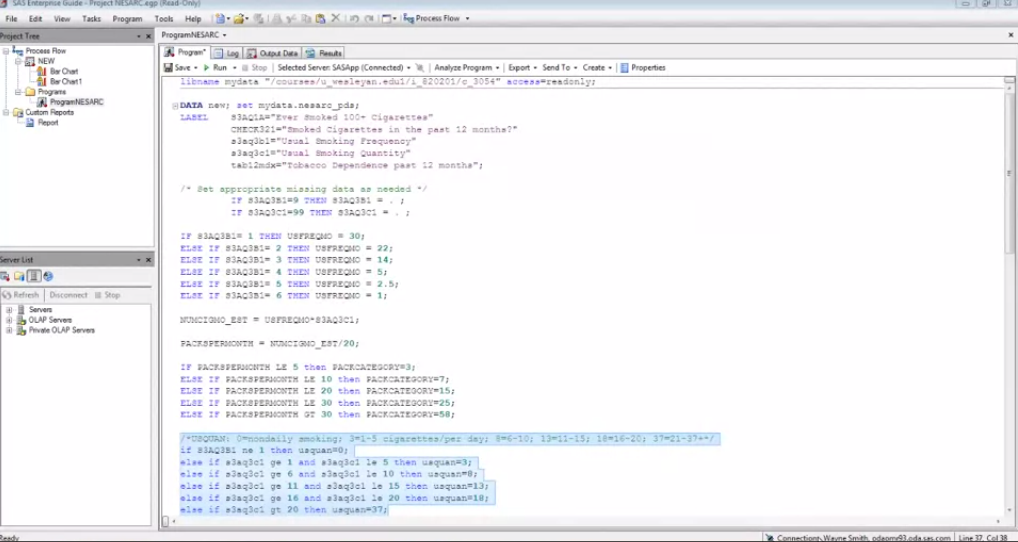
The resulting output for this analysis is shown here. As you can see, we're testing the association between diet A and B and weight loss. There are 40 observations in the data set. The F value is 12 and it's associated with a significant p-value. That is, a p-value less than 0.05. While this tells us that there is a significant association between diet type and weight loss to understand that association, we need to look at the output generated by the mean statement. Here we see that the average one month weight loss for diet A is about 14.7 pounds and that the average one month weight loss for diet B is about 9.3 pounds. So, in conjunction with the significant p-value, we can say that diet plan A is associated with significantly greater weight loss than diet plan B. Here I show the finding graphically as a bar chart with diet, the explanatory variable, on the x-axis and the mean weight loss, our response variable on the y-axis. Further, using advanced options associated with the SAS bar chart task, I can display not only the observed sample means but also the population standard error, which is the variability we would expect to see over multiple samples drawn from the population. But what about our third variable, exercise program? Would we get the same results in terms of the association between diet and weight loss for those participants using cardio and those participants using weight training? In statistics, moderation occurs when the relationship between two variables depends on a third variable. In this case, the third variable is referred to as the moderating variable, or simply the moderator. The effect of the moderating variable is often characterized statistically as an interaction. That is, a third variable that effects the direction and or strength of the relation between your explanatory and response variables. So, does type of exercise program affect the direction or strength of the relationship between diet and weight loss? Although the standard way of asking this question in the context of analysis of variance is to move to the use of a two way or two factor analysis of variance, rather than the one way or one factor ANOVA that we've been using. Instead, we're going to take a less-standard approach that can be consistently used across each of the inferential tools. That is ANOVA, chi-square, and Pearson correlation. In each of these contexts, we're actually going to be asking the question, is our explanatory variable associated with our response variable, for each population sub-group or each level of our third variable? That is, are diet type and weight loss associated for those doing the cardio exercise program? And are diet and weight loss associated for those using the weight-training program? To accomplish this, we are going to run two separate ANOVAs, one for each level of the third variable. That is, for each exercise program. Syntax to be added to the general PROC ANOVA code is circled here in red. We need to first sort the data according to the categorical third variable, then include a by statement. Telling SAS to run a variance for each level of the third variable. The specific syntax for a diet and exercise example is shown here.



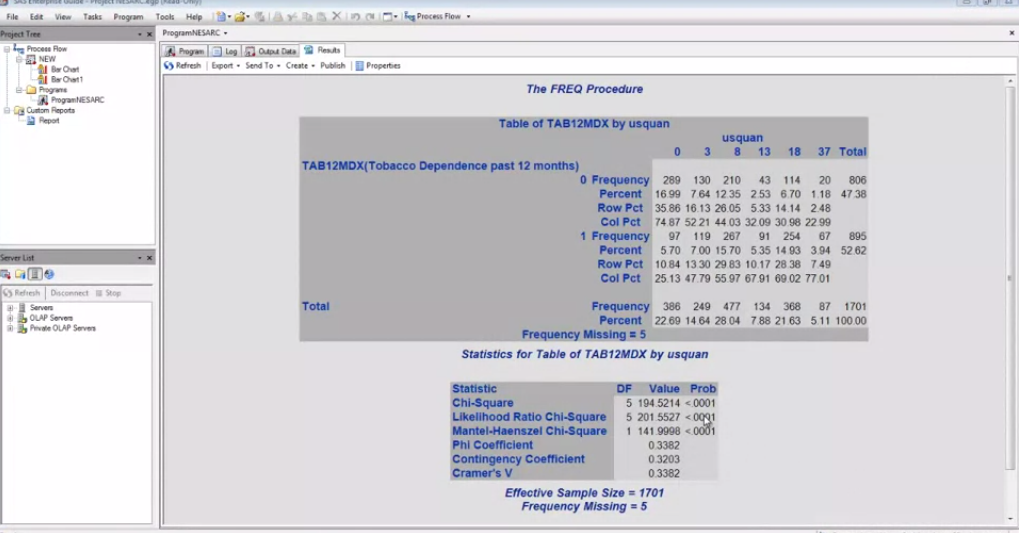
In SAS, what I call in this example diet and exercise data set and run the analysis to variance. Here are the results. The ANOVA table examining the relationship between diet and weight loss, for those in the cardio exercise group, shows a large f-value and a significant associated p-value. When examining the means table, we see that for those involved in the cardio exercise program, diet A is associated with greater weight loss, 20.5 pounds on average, than diet B, 7.1 pounds on average. The association between diet and weight loss for those involved in the weight training exercise program is also significant, with a large f-value. However, when examining the means, we find that the association is in the opposite direction. That is, for those involved in weight training, diet B is associated with greater weight loss, 11.5 pounds, compared to diet A, only 8.8 pounds.



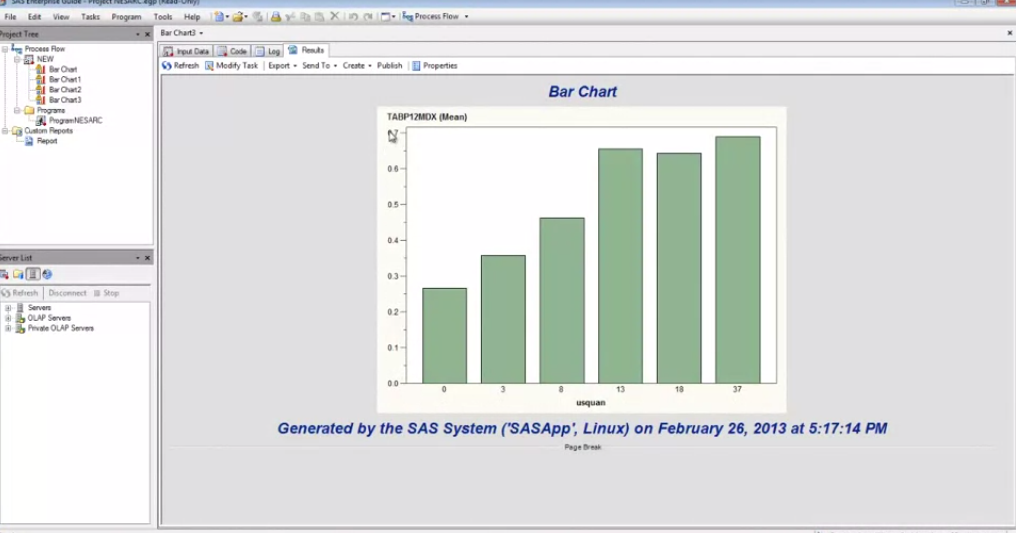
Here, these results are shown graphically. As you can see, the relationship between diet and weight loss depends on which exercise program is being used. When using cardio, diet A is significantly better for weight loss than diet B. When using weights, diet b is significantly better for weight loss than diet A. Thus, we can say the type of exercise moderates the relationship between diet and weight loss. Here I show the same graph, but using the advanced options associated with SAS bar charts, I have displayed the population standard error associated with each sample mean. That is, the variability we would expectto see over multiple samples drawn from the population. Remember, our goal in inference is not to only to describe the sample but also to draw conclusions about the population. Suppose that we did not evaluate exercise as a possible moderator and instead focused only on the association between diet and weight loss for the entire population. Based on this graph, obviously, we would've incorrectly concluded that diet A is better than diet B. As we now know, that is true only if we're looking at the cardio group.



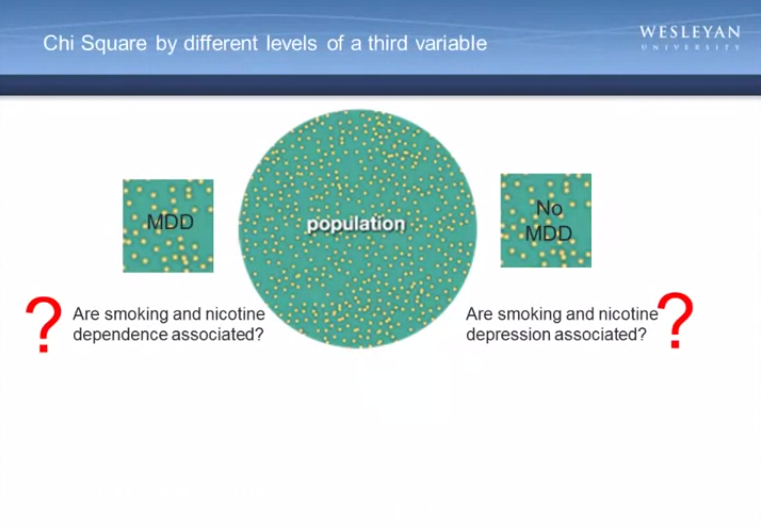
Now let's evaluate third variables as potential moderators in the context of chi-squared tests of independents. For this, I'm going to return to my original SAS program, using the NESARC data and asking the question, is smoking associated with nicotine dependence? I am going to create another smoking variable for this purpose reflecting how many cigarettes each young adult smoker smokes per day. Zero will indicate non-daily smokers. 3 indicates those smoking 1 to 5 cigarettes per day. 8, 6 to 10 cigarettes per day, 13, 11 to 15 cigarettes per day, 18, 16 to 20 cigarettes per day and 37, greater than 20 cigarettes per day. Now, I'm going to scroll down to the bottom of my program and request a chi-square test of the dependence, examining the association between nicotine dependence in the past 12 months, and this new smoking variable called USQUAN.



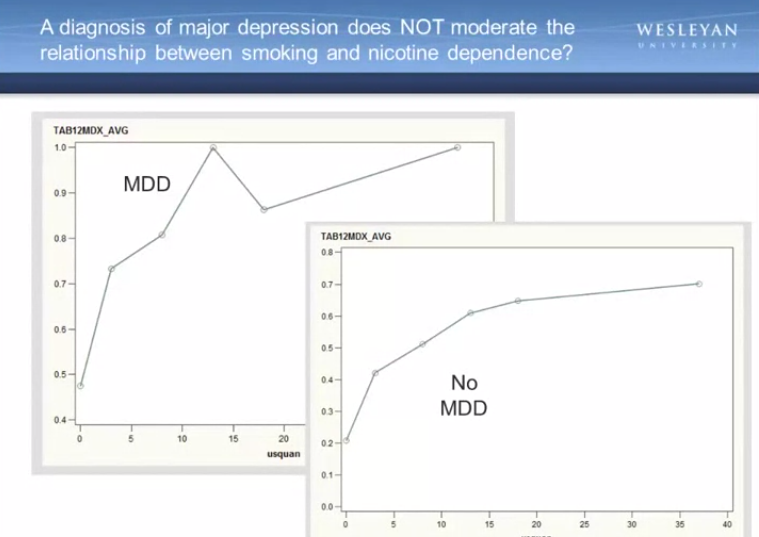
As we can see from the large chi-square value and significant p-value, smoking and nicotine dependence are significantly associated. In examining the column percents here of each smoking group with nicotine dependence, we see generally higher rates of nicotine dependence among groups that smoke more. So, among non-daily smokers, only 25% are nicotine dependent. Among those smoking 1 to 5 cigarettes per day, indicated by a code of 3, nearly 50% meet criteria for nicotine dependence. And again, these numbers increase with higher levels of smoking. To graph these proportions with the bar chart wizard, I would do the following. My explanatory variable, US Quantity, which is discrete or categorical. My response variable, nicotine dependence in the past 12 months, and I'll ask for the average, which will give me the proportion of individuals with nicotine dependence given that this variable is coded zero and one.



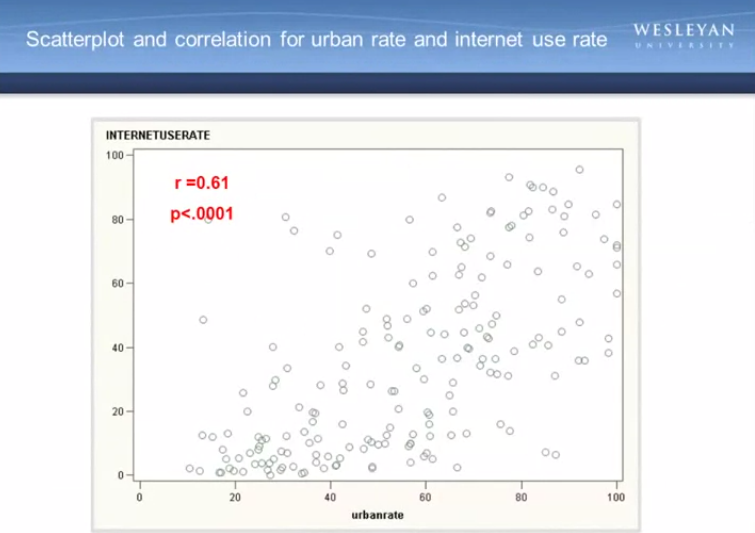
And this gives me the graphic representation of this positive linear relationship. As smoking quantity increases, so does the proportion of individuals with nicotine dependence. This finding is accurate with regard to the larger population of young adult smokers. But might a third variable moderate the relationship? I decide to evaluate major depressive disorder as my third variable. I am wondering if MDD affects either the strength or the direction of the relationship between smoking and nicotine dependence.



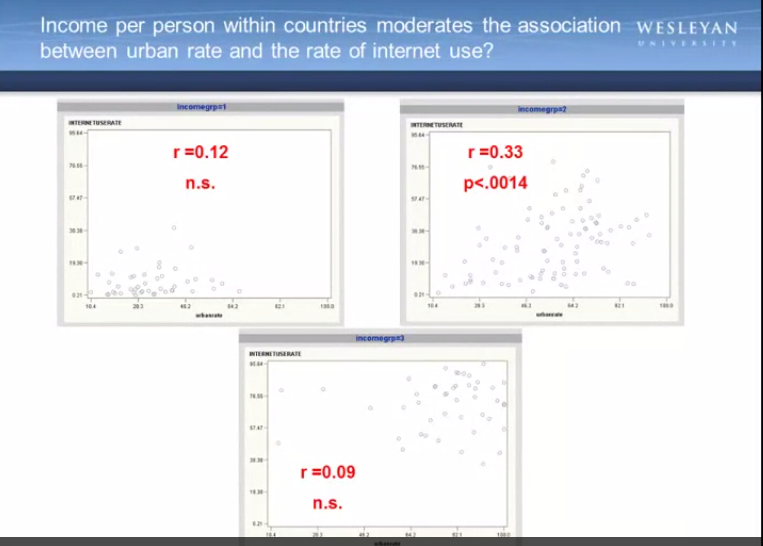
Is smoking related to nicotine dependence for each level of my third variable? That is, for those with major depression and for those without major depression. Similar to our ANOVA example, syntax to be added to the PROC FREQ code is circled here in red. We need to first sort the data according to the categorical third variable. Then, include a by statement telling SAS to run a chi-square for each level of the third variable separately. The specific syntax for my example is shown here. When I add this syntax to my SAS program, here are my results. You can see that I have a cross tabs or cross tabulation table looking at usual quantity by tobacco dependents in the past 12 months. First, for major depression equal to zero. That is, those without major depression. The chi-square value is large and again the p value is quite small, so I can say that this is a statistically significant relationship for those without major depression for those with major depression. Again, I find a large chi-square value and small p-value which is statistically significant. Looking at the column percents, I again detect what seems to be a positive linear relationship, with percentages of nicotine dependence increasing between lower levels of smoking and higher levels of smoking.



Using a line graph to examine the rates of nicotine dependence by different levels of smoking, it seems that both the direction and magnitude of the relationship is similar between smoking and nicotine dependence for those with major depression and for those without So, a diagnosis of major depression does not moderate the relationship between smoking and nicotine dependence. For both young adult smokers with major depression and those without, higher levels of smoking behavior is associated with higher rates of nicotine dependence.



So now, let's test for moderation within the context of our final inferential test, the correlation coefficient.You might remember this scatterplot and correlation based on the gap minder data between rate of urban dwellers in each country and the internet use rate. We found that this was a significant association, with a correlation of .61. But might this relationship, this correlation between urban rate and internet use rate, might it differ based on countries with different income levels? To explore this question, I created a categorical third variable called income group. In which the quantitative income per person variable was categorized as a high income country given a value of three, a moderate income country given a value of two and a low income country given a value of one. Similar to the adjustments that we made to our ANOVA syntax and our chi-square syntax when testing moderation, I do the same thing with my correlation coefficient. First, I sort the data by this new categorical third variable. Next, I run the correlation between urban rate and internet use, and then I include a by-statement telling SAS to calculate the correlation coefficient for each income group. When I examine the correlation coefficients between urban rate and internet use rate for each of the income groups, I find the following. For my low income group, the correlation between urban rate and internet use rate is 0.11 and my p-value is not significant for my moderate income countries. The association between Internet use rate and urban rate is 0.329, with a significant p-value at 0.0014. And finally, among high-income countries, the correlation code efficient is 0.089. Again, with a large p-value, suggesting that the association between urban rate and internet use rate is not significant for high income countries.



When I map these findings onto the associated scatterplots for each income group, I am able to better visualize the significant and non-significant relationships. Estimating a line of best fit within each scatterplot shows the positive association between urban rate and internet use rate among the moderate income countries. And almost no relationship between these two variables in both the low income and high income countries. So, while not difficult to do based on the skills that you have already gained in this course, asking questions about moderation can be an incredibly interesting way to explore your data and better understand your association of interest. Even without the use of multi-variance statistical technique, a more advanced topic than we cover here, we can still use bi-variate inferential tools of the ANOVA chi-squared and correlation to describe our sample, make inferences about the larger population and really begin to understand what relationships are significant and under what conditions. Or at what levels of our third variables these associations hold. But does a significant association imply causality? More on that topic in the next video.