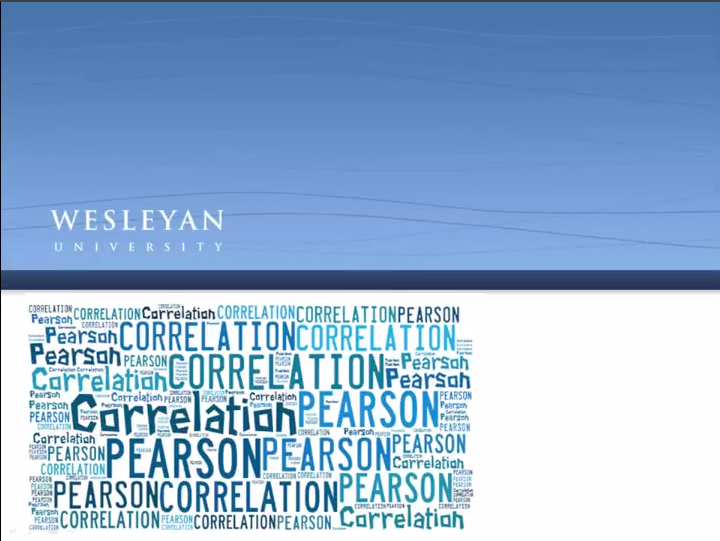
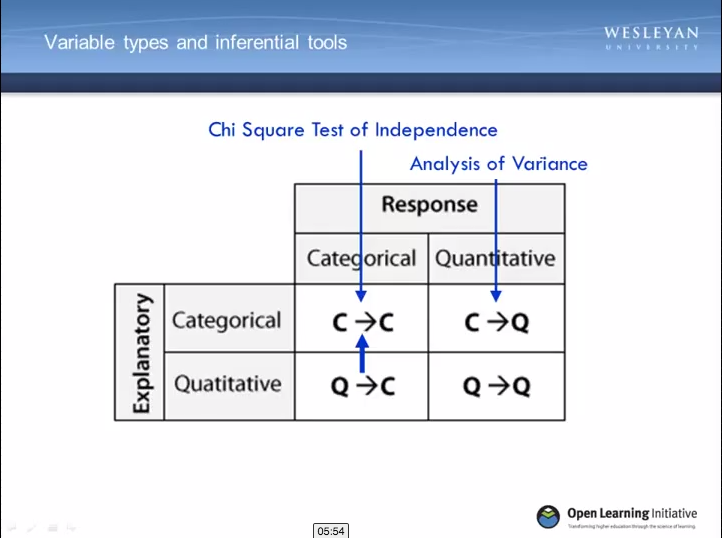
**Correlation**

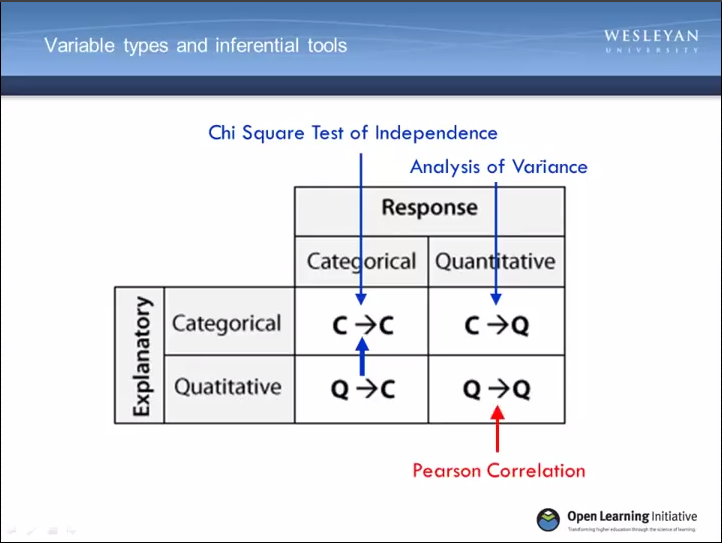
**Passion Driven Statistics: Weak 4, Video 4**



The topic of this presentation is the Pearson correlation coefficient. You now have the tools for examining the association between a categorical and quantitative variable, analysis of variance, and in the case of two categorical variables, the Chi-Square test of independence.

If you have a quantitative explanatory variable and a categorical response variable, for our purposes I'm encouraging you to categorize the quantitative explanatory variable and use the Chi-Square test of independence to examine the association.

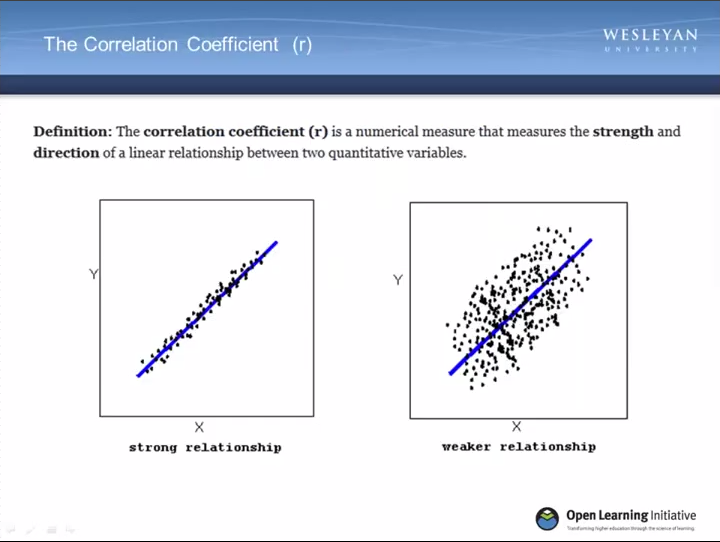




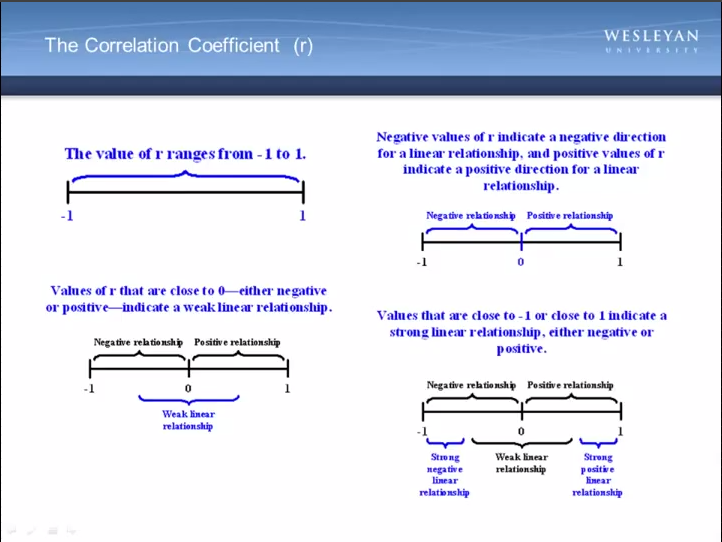
The final inferential tool for examining the association between two variables is

the Pearson correlation. The Pearson correlation is used to examine the association between two quantitative variables.

You might remember that when we discussed graphing two quantitative variables, we learned to examine this relationship looking at a scatter plot.



A discussion of the correlation coefficient and corresponding p-value will now allow us to evaluate whether or not the relationship that we observe in the sample holds for the larger population. The correlation coefficient, usually indicated with a lower-case r, is a numerical measure that assesses the strength of a linear relationship between two quantitative variables.

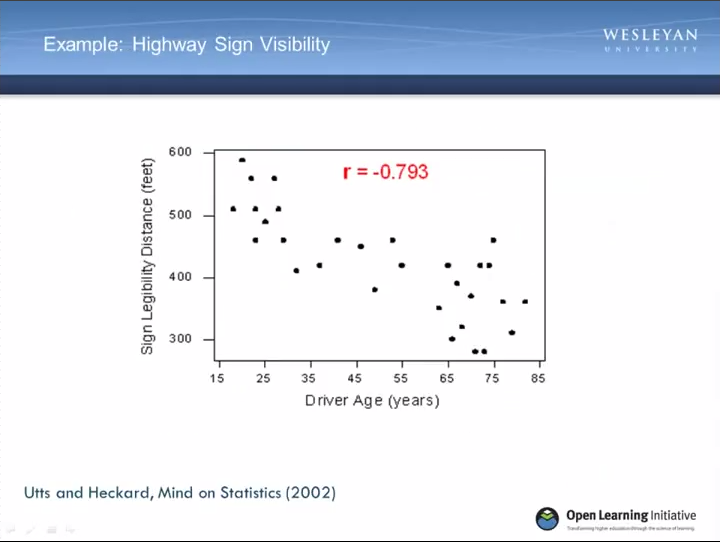


The value of r ranges from negative one to positive one.

And not surprisingly negative values of r indicated negative direction for a linear relationship between the two variables and positive values of r indicate a positive direction for the linear relationship. Values of r that are close to 0 whether they're negative or positive indicate a weak linear relationship.

And values that are close to negative 1 or close to positive 1 indicate a strong linear relationship, either negative or positive.

Let's look at an example:

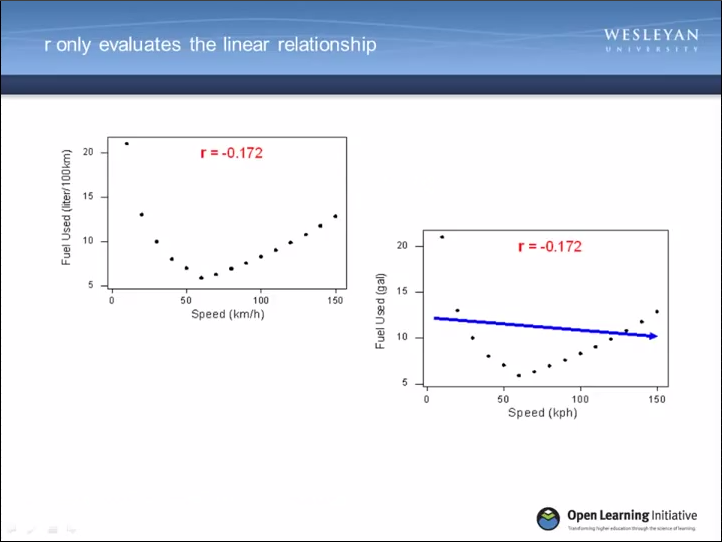


A Pennsylvania research firm conducted a study in which 30 drivers of ages 18 to 82

Years old were sampled. And for each one the maximum distance in feet at which he or she could read a newly designed sign was determined. The goal of this study was to explore the relationship between a driver's age and the maximum distance at which signs were legible.

The explanatory variable is age, and the response variable is distance. Here is a scatter plot for the data, which shows a negative linear relationship between the age of a driver and the maximum distance at which a highway sign was legible.

What about the strength of the relationship?



It turns out that the correlation between the 2 variables is r equal to negative .793. Since r is less than 0 it confirms that the direction of the relationship is negative.

Since r is relatively close to negative 1 it suggests that the relationship is moderately strong.

In context, the negative correlation confirms that the maximum distance at which a sign is legible generally decreases with age. The correlation only measures the strength of a linear relationship between two variables. It ignores any other type of relationship, no matter how strong it is. For example, consider the relationship between the average fuel usage of driving a fixed distance in a car and the speed at which the car drives.

Our data describes a fairly simple curvilinear relationship. The amount of fuel consumed decreases rapidly to a minimum for a car driving 60 kilometers per hour.

And then increases gradually for speeds exceeding 60 kilometers per hour. The relationship is very strong as the observations seem to perfectly fit the curve.

Although the relationship is strong the correlation r is equal to negative .172. This indicates a weak linear relationship.

This makes sense considering that the data fails to adhere closely to a linear form. The correlation is useless for assessing the strength of any type of relationship that is not linear, including relationships that are curvilinear, such as the one in our example.

Beware then of interpreting the fact that r is close to 0 as an indicator of a weak

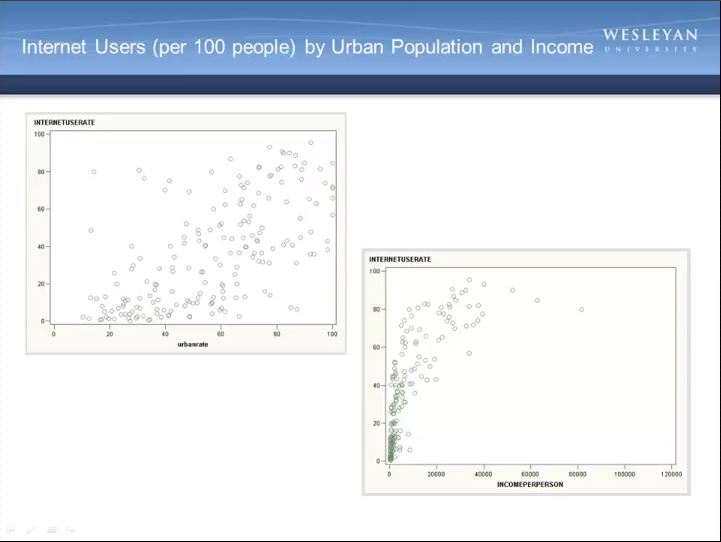
relationship, rather than a weak linear relationship.

This example also illustrates how important it is to always look at the data in the scatter plot, since as in our example, there might be a strong nonlinear relationship that r does not indicate. The same can be said for a large correlation coefficient.

So it's important to interpret both the scatter plot and the correlation together. As with the other inferential tools that we've discussed, an associated p value is also calculated for the correlation coefficient. And is interpreted as significant when it

is less than or equal to .05.

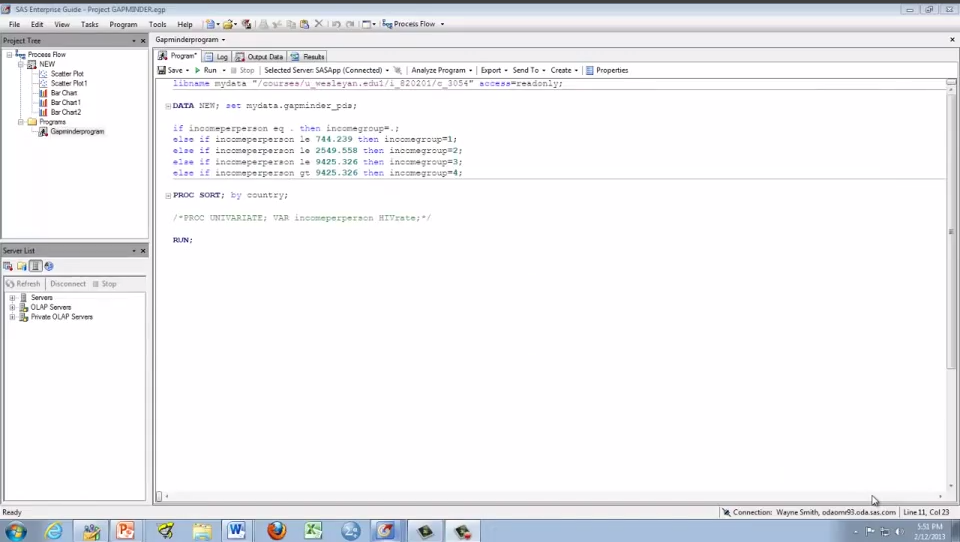
To demonstrate how to request a correlation coefficient in SAS, let’s go back to the scatter plugs that we created for few of the gap minder variables. You may remember these scatter plots that I presented to you when talking about visualizing the association between two quantitative variables.



The first shows the rate of internet users by the rate of the country's population living in urban settings. The second shows the rate of internet users by income per person.

From looking at the scatter plots we can guess that the associations are positive, that is a higher internet use rate is associated with both higher urban rates and greater income.

Now let's find the correlation coefficient:



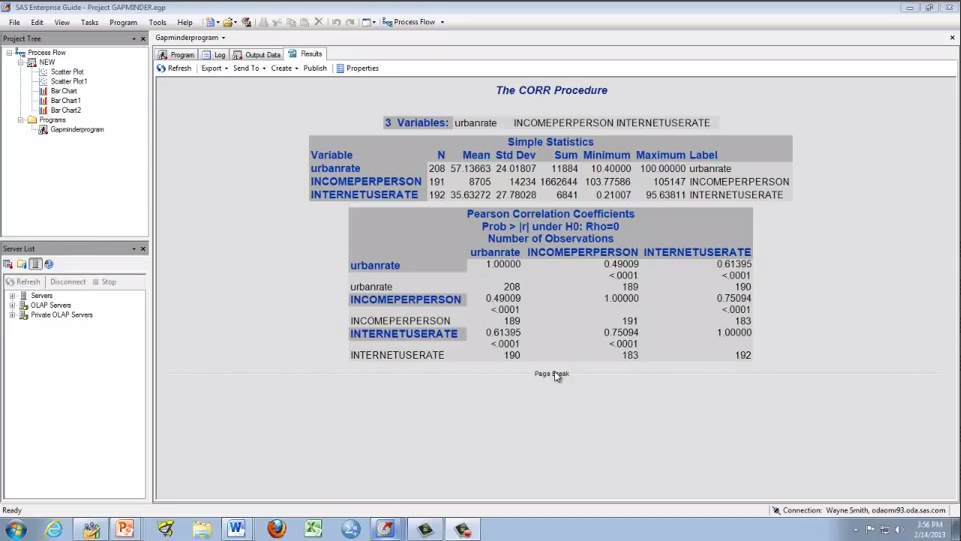
To do this in SAS, I only need to add one line of syntax to my program. I add PROC CORR, C, O, R, R, semicolon. The statement VAR, followed by the variables that I would like to examine in terms of their correlation with one another.

**Proc** **corr**; var QUAN\_IV QUAN\_DV;

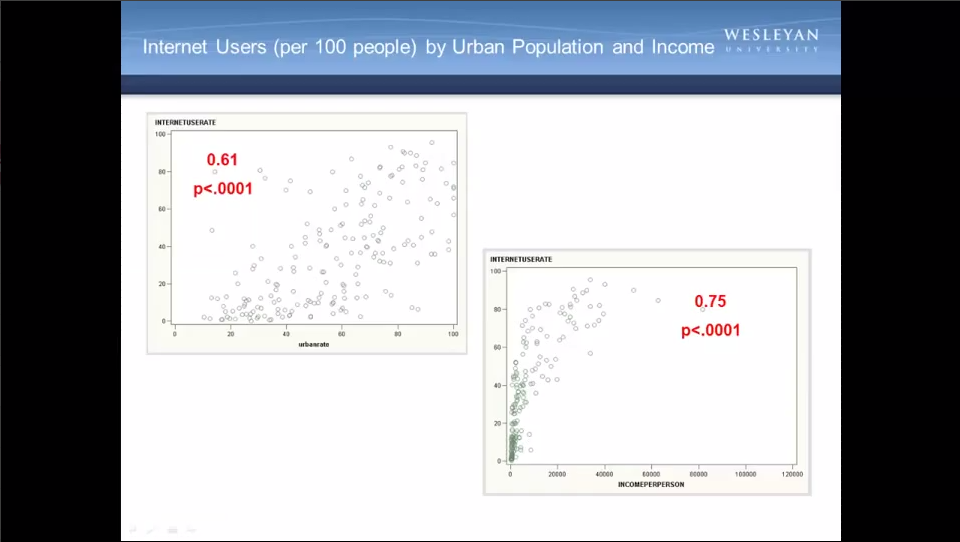
Even though I'm going to be looking at two different correlations, that between internet use and urban rate, and also internet use and income per person, I can actually list these together.

After ending the list with a semicolon I hit save and run.

And then SAS will give us the following output.



To locate the correlation coefficients of interest and the associated p values, we need to examine the Pearson Correlation Coefficient table here. And find the row and column where our two variables of interest intersect. So for the association between urban rate and Internet use rate…

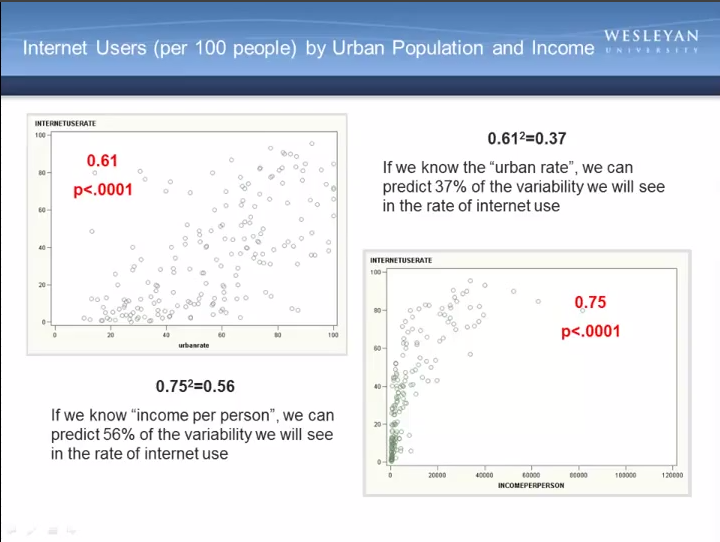


I see that the correlation coefficient is approximately .61 with a P value of .0001. This tells us that the relationship is statistically significant.

For the association between income per person and internet use rate, the correlation co-efficient is approximately .75 and again has a significant key value. So now, we can actually interpret the scatter plots and the coefficients together.

The association between internet use rate and income, the second scatter plot, is

fairly strong and also positive, as the scatter plot had already shown us.



The association between internet use rate and urban rate is also positive but

slightly more modest .61. Both are statistically significant, that is for both associations. It is highly unlikely that a relationship of this magnitude would be due to chance alone.



At this point, I am really happy to tell you that post-hoc tests do not come into play in the context of Pearson Correlation. Post-hoc tests are needed only when your research question includes a categorical explanatory variable with more than two

levels. Because our explanatory variable in the context of a correlation coefficient is

quantitative, there is never a need to perform a post op test.

So today I'm going to let you go a little early, but before I do there is just one last thing I want to say about the correlation coefficient.

If we square the correlation coefficient, that is, we multiply it by itself...

We get a value that is also helpful in understanding the association between two

quantitative variables.

Little r squared (r2) is the fraction of the variability of one variable that can be predicted by the other.

For the relationship between urban rate and internet use rate, if we square our correlation coefficient of 0.61, we get 0.37.

This could also be interpreted the following way:

If we know the urban rate, we can predict 37% of the variability we will see in the

rate of internet use. Of course, that also means that 63% of the variability is unaccounted for. If we square the correlation coefficient for income per person and internet use rate, we get a value of 0.56 suggesting that if we know income per person, we can predict 56% of the variability that we will see in the rate of internet use. This is a little bit more impressive. That is, we can predict over half the

variability.

Again, correlation coefficients are commonly denoted with a lowercase r, and

they're squared to determine the amount of variability that can be predicted. You might be wondering how much variability in internet use rates can be predicted if we consider both urban rate and income per person.

A multivariate inferential tool called multiple regression can be used to answer this question but we won't be getting into those details here.

For assignment eight, you will be running and interpreting correlation coefficients.

Enjoy!