Course Learning Outcomes for Unit V

Upon completion of this unit, students should be able to:

6. Differentiate between various research-based tools commonly used in businesses.
   6.1 Determine the most appropriate statistical procedure to use from among correlation, simple regression, and multiple regression to test hypotheses.

7. Test data for a business research project.
   7.1 Establish whether to accept or reject null and alternative hypotheses by using correlation, simple regression, and multiple regression.

Course/Unit Learning Outcomes	Learning Activity
6.1	Unit Lesson Video: *How to Find Correlation in Excel with the Data Analysis Toolpak* Video: *How to Use Excel-The PEARSON Function* Video: *Excel 2016 Correlation Analysis* Video: *How to Calculate a Correlation (and p value) in Microsoft Excel* Video: *Correlation Coefficient in Excel* Video: *How to Perform a Linear or Multiple Regression (Excel 2013)* Video: *Multiple Regression Interpretation in Excel* Unit V Scholarly Activity
7.1	Unit Lesson Video: *Excel 2016 Correlation Analysis* Video: *How to Calculate a Correlation (and p value) in Microsoft Excel* Video: *Correlation Coefficient in Excel* Video: *Multiple Regression Interpretation in Excel* Unit V Scholarly Activity

Required Unit Resources

In order to access the following resources, click the links below:

Glen, S. (2013, December 14). *How to find correlation in Excel with the Data Analysis Toolpak* [Video]. YouTube. https://www.youtube.com/watch?v=AjQA78tI39Q

A transcript of this video is available.

TheRMUoHP Biostatistics Resource Channel. (2014, November 6). *How to use Excel-The PEARSON Function* [Video]. YouTube. https://www.youtube.com/watch?v=JO-Gc5bEG70

A transcript of this video is available.

Porterfield, T. (2017, May 18). *Excel 2016 correlation analysis* [Video]. YouTube. https://www.youtube.com/watch?v=kr64tfZmiGA

A transcript of this video is available.

Quantitative Specialists. (2014, September 15). *How to calculate a correlation (and p-value) in Microsoft Excel* [Video]. YouTube. https://www.youtube.com/watch?v=vFcxExzLfZI

A transcript of this video is available.

MrSnyder88. (2009, November 8). *Correlation coefficient in Excel* [Video]. YouTube. https://www.youtube.com/watch?v=s2TVkYmmCAs

A transcript of this video is available.

economistician.com. (2015, May 15). *How to perform a linear or multiple regression (Excel 2013)* [Video]. YouTube. https://www.youtube.com/watch?v=wBocR96UdyY

A transcript of this video is available.

TheWoundedDoctor. (2013, May 6). *Multiple regression interpretation in Excel* [Video]. YouTube https://www.youtube.com/watch?v=tlbdkgYz7FM

A transcript of this video is available.

Unit Lesson

Data Analysis: Correlation and Regression

Unit IV discussed descriptive statistics and the importance of testing the data to ensure assumptions are met before using parametric statistical procedures. When using descriptive statistics, the data that are collected are described by the researcher both visually and statistically. The visual representation alone can reveal information about whether assumptions are met. Although all statistical tests have different assumptions, normality is universally shared and is relatively easy to observe through the use of histograms.

It is preferable to use parametric tests since they are more powerful than non-parametric tests, which have fewer assumptions that must be met. Regardless of the statistical procedure under consideration, the assumptions must be met if the researcher can have confidence in the validity of the results. Units V through VII will focus on inferential statistics, which include the parametric tests of correlation, regression, *t* test, and ANOVA.

Inferential Statistics

Unlike descriptive statistics, inferential statistics go beyond simply describing the data to making inferences, or predictions, about a population. The inferences are often based on the characteristics of a sample. Inferences, or predictions, are stated in the form of hypotheses. Results of statistical tests on samples are used to generalize those results to a population (Zikmund et al., 2013). Descriptive statistics and inferential statistics are not mutually exclusive. In fact, performing descriptive statistics should always be a precursor to inferential statistics for assumption testing for statistical procedures being considered.

Populations, Samples, and Generalization

Statistical procedures are used to answer questions about a population. A population can be people or things, such as a company's entire consumer base or the total units produced for a new product. A population can be very large or very small. For example, a company may collect productivity data on their 100 employees. They are interested in knowing if there is a relationship between the size of merit increases and job productivity. The 100 employees represent the entire population, which would be considered a census. Since data are collected from all 100 employees, the company can have certainty that the statistical results represent the entire population. In many instances, however, it is impractical and cost prohibitive to collect data from all participants in the population. In these scenarios, data is collected from a sample of the population. The statistical results from the sample are then used to generalize the findings to the population. Using the example above, now assume the company has a population of 200,000 employees. They decide to select a random sample of 100 employees to whom they have provided various merit increases. Like the example

above, their interest is to understand if there is a relationship between the size of merit increases and job productivity. If they determine that there is a statistically significant relationship between the size of merit increases and productivity, they can generalize those results to the population of 200,000 employees. This can inform their decision-making and planning regarding the size of raises to provide for the next fiscal year and the productivity increase they can forecast. This is the function of inferential statistics.

Relationships or Differences

Statistical analysis can be simplified as either looking for relationships (or associations) between variables or looking for differences between variables or groups. This unit considers statistical testing that looks for relationships between variables. The statistical procedures highlighted to test for relationships will be correlation, simple regression, and multiple regression. Correlation and regression analyses are parametric tests. Chi-square is a corresponding non-parametric test.

Correlation

Although many course concepts in research methods may be new and foreign, correlation may feel more familiar and comfortable. The concept of correlation makes intuitive sense to most people since relationships between variables (e.g., years of education and income, safety training hours and lost time hours, and hours of exercise and weight loss) occur frequently in daily life. Relationships naturally occurring between variables can be positive or negative. A positive or negative relationship between variables does not mean positive or negative in the context of making a value judgment of good or bad. A positive or negative relationship, in statistical terms, means the direction of the relationship.

An example of a positive relationship between variables is durable goods orders and the S&P 500 index. When durable goods orders decrease, there is a decrease in the S&P 500 index. When durable goods orders increase, there is an increase in the S&P 500 index. This is a positive relationship because both variables move in the same direction. As one variable increases, the other increases. Conversely, when one variable decreases, the other decreases.

An example of a negative relationship between variables is outdoor temperature and heating oil expenditures. When the outdoor temperature increases, heating oil expenditures decrease. When the outdoor temperature decreases, heating oil expenditures increase. This is a negative relationship because the variables move in opposite directions. As one variable increases, the other decreases. Conversely, when one variable decreases, the other increases.

Another important distinction that must be understood is the difference between correlation and causation. Even if a statistical test (e.g. Pearson's *r*) indicates a statistically significant relationship between variables, it must never be said that one variable causes the change in the other variable. For example, there is a positive correlation between ice cream sales and violent crime in New York City (both increase in the warmer months of the year, and both decrease in the cooler months). It would be absurd to say that ice cream causes violent crime—even though the relationship between variables does exist. This extreme example makes the point that correlation does not mean causation. Causation can only be statistically shown via experimental research designs, which have tight controls to manipulate variables.

Pearson Correlation Coefficient (*r*)

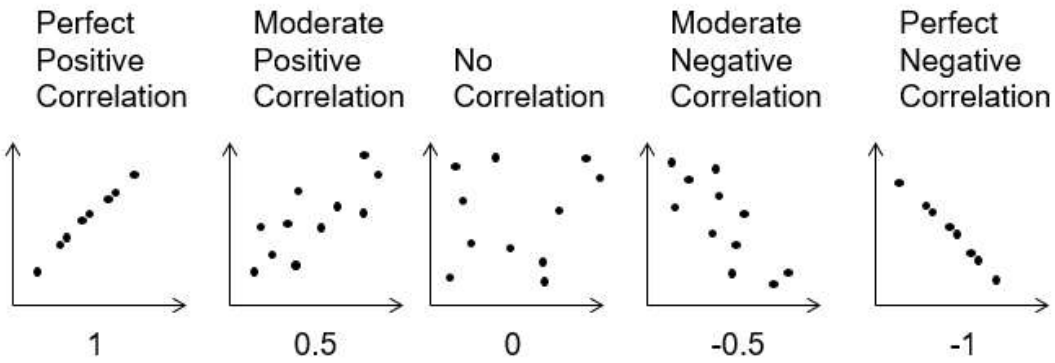
When conducting correlation analysis, the Pearson correlation coefficient (*r*) is the most commonly used parametric measure of association between two variables (Norusis, 2008). The Pearson statistic is represented by *r*, which is the standardized covariance between the variables, and measures the linear relationship between variables (Field, 2005). The Pearson correlation coefficient is sometimes represented by *R*, but this is normally used in the context of regression analysis. One can easily determine how to calculate *r* using long-hand by referring to a statistics textbook, but it is much easier and faster to use statistical software to quickly calculate the Pearson correlation coefficient. For the purposes of this course, it is most important to understand what Pearson's *r* is, what it measures, and how to interpret it, rather than how to calculate it by long-hand.

When using correlation analysis, a hypothesis is tested that there is no statistically significant relationship between variables. The null and alternative hypotheses would be stated like so.

Ho1: There is no statistically significant relationship between *X* and *Y*.

Ha1: There is a statistically significant relationship between *X* and *Y*.

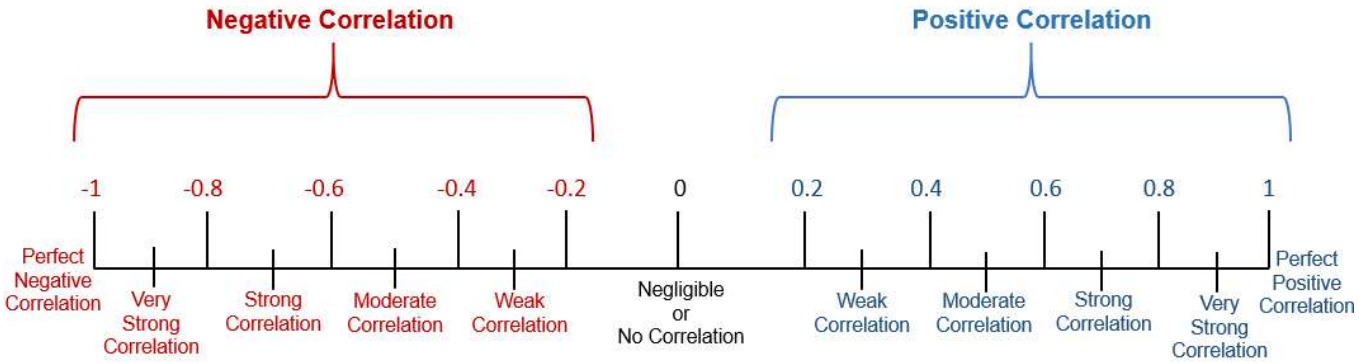
As mentioned above, the *r* statistic can indicate a positive relationship or a negative relationship between variables. The *r* statistic can also indicate no relationship at all between variables. An *r* of +1 indicates a perfect positive correlation, while an *r* of -1 indicates a perfect negative correlation (Field, 2005). The *r* statistic will always fall between +1 and -1. An *r* of 0 indicates no correlation exists between variables.



When reviewing the literature for research articles, it is very common to find *r* statistics less than .5. Given the fact that an *r* of 1 indicates a perfect correlation, a statistically significant *r* of .5 or less hardly seems large enough to get excited about; however, the American Psychological Association would disagree.

The American Psychological Association (as cited in Kerr et al., 2006) concluded that psychologists studying highly complex human behavior should be satisfied with correlations in the *r* = 0.10 to 0.20 range, and they should be generally pleased with correlations in the 0.25–0.35 area. The best new variables typically increase predictions, for instance, of job performance between 1% and 4%. A 10% contribution of emotional intelligence would be considered very large (Kerr et al., 2006).

Although there are no concrete guidelines for interpreting *r* and R^2, The following chart suggests some general guidelines that are fairly consistent with other rule-of-thumb published guidelines.



Adapted from *Guideline for Interpreting Correlation Coefficient* by I. Phanny, 2014. (https://www.slideshare.net/phannithrupp/guideline-for-interpreting-correlation-coefficient/2).

Coefficient of Determination (R^2)

The Pearson's *r* is useful itself, but the closely related coefficient of determination (R^2) is also very informative. Simply squaring *r* produces R^2, which indicates the amount of variability in one variable that is explained by the other variable (Field, 2005). According to the American Psychological Association (as cited in Kerr et al., 2006), a researcher should be generally pleased with a correlation of $r = .25$, which translates to a coefficient of determination $R^2 = .0625$. This means that the variable *x* explains 6.25% of the variability in the variable *y*. Most statistical software programs will calculate both *r* and R^2 for when running correlation analysis, so it is easy to see the strength of the association and the explained variance. Again, it is important not to confuse correlation with causation.

Examples of *r* and R^2:

$r = .10$, $R^2 = .01$ explains 1% of the total variance between the variables being tested

$r = .30$, $R^2 = .09$ explains 9% of the total variance between the variables being tested

$r = .50$, $R^2 = .25$ explains 25% of the total variance between the variables being tested

Interpreting Correlation Output Results

The following correlation analysis looked for a statistically significant relationship between the variables of height and weight. The results show that there is a moderately strong correlation $r = .6$ (Pearson's Correlation). It is also necessary to assess whether the correlation is statistically significant using an alpha of .05. The results indicate a *p value* of .023 < .05. Therefore, the null hypothesis is rejected, and the alternative hypothesis is accepted.

Reject *Ho1:* There is no statistically significant relationship between weight and height.

Accept *Ha1:* There is a statistically significant relationship between weight and height.

Correlations

		weight	height
weight	Pearson Correlation	1	.600*
	Sig. (2-tailed)		.023
	N	14	14
height	Pearson Correlation	.600*	1
	Sig. (2-tailed)	.023	
	N	14	14

*. Correlation is significant at the 0.05 level (2-tailed).

Although the information obtained through correlation analysis is revealing and useful, it is limited in that correlation analysis cannot be used to make predictions (Field, 2005). To be able to predict the value of a dependent variable (DV) from observations of the independent variable (IV), regression analysis must be used.

Regression Analysis

Relationships between variables can be useful for making predictions. Regression analysis is a concept that many students have heard of, even if they are not entirely comfortable with it. If the relationship between the variables *X* and *Y* are known, predictions can be made about how a change in *X* will relate to a change in *Y*. Remember that this is not stating that a change in *X* causes a change in *Y*. It is only possible to predict a change based on the relationship between variables. Regression analysis can be powerful, especially when multiple *X* variables are included (multiple regression) to make a prediction about a change in a single *Y* variable.

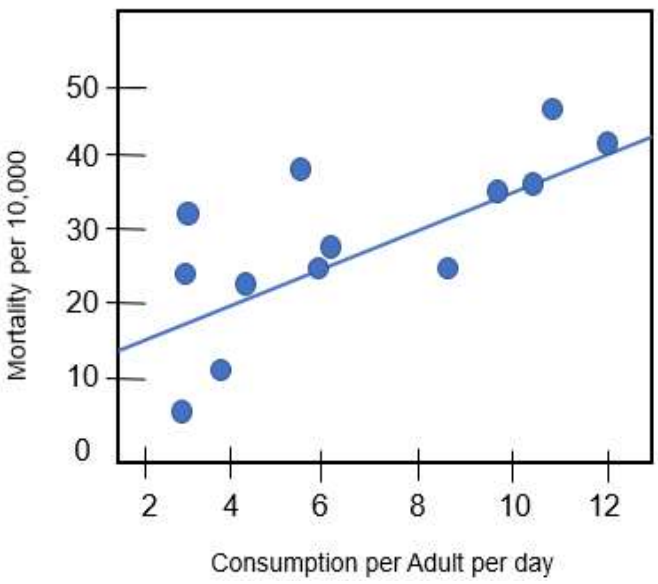
When using regression analysis, a hypothesis is tested that there is no statistically significant prediction of the dependent variable (i.e., *Y* or outcome variable) by one or more independent variables (*X*). If a single independent variable is used to predict *Y*, it is termed simple regression. If two or more independent *X* variables are used to predict *Y*, it is termed multiple regression.

The null and alternative hypotheses would be stated as follows.

Ho1: There is no statistically significant relationship to predict *Y* from X_1, $X_{2...}$and X_n.

Ha1: There is a statistically significant relationship to predict Y from X_1, X_2…and X_n.

Regression analysis uses a linear model to apply a line of best fit to the data. The line of best fit is the most optimal because it results in the smallest amount of difference between the observed data points and the line (Field, 2005). As the linear regression example below shows, a line of best fit is applied to the data for the variables mortality (DV) and cigarette consumption (IV). This is an example of simple linear regression because there is only one IV.



Adapted from images in *Multiple Linear Regression* by J. Neill, 2008 (https://www.slideshare.net/jtneill/multiple-linear-regression).

If all of the data points fell on a straight line, it would be a perfect linear relationship, which would allow us to make a perfect prediction of the Y axis variable by looking at the X axis variable (Norusis, 2008). A perfect linear relationship is rare, so we develop the regression model as $Y = a + b(X)$.

The resulting mathematical model is tested for statistical significance. If statistically significant, at a p value of less than .05, the IV data can be plugged into the model to be multiplied by the calculated coefficient, added to the calculated constant (Y-intercept or a_0), resulting in the predicted DV. The statistical software will calculate the model and values for a_0 and b_1, which will appear as the following equation:

$$Y = a_0 + b_1 (X)$$

or

$$DV = a_0 + b_1 (IV_1)$$

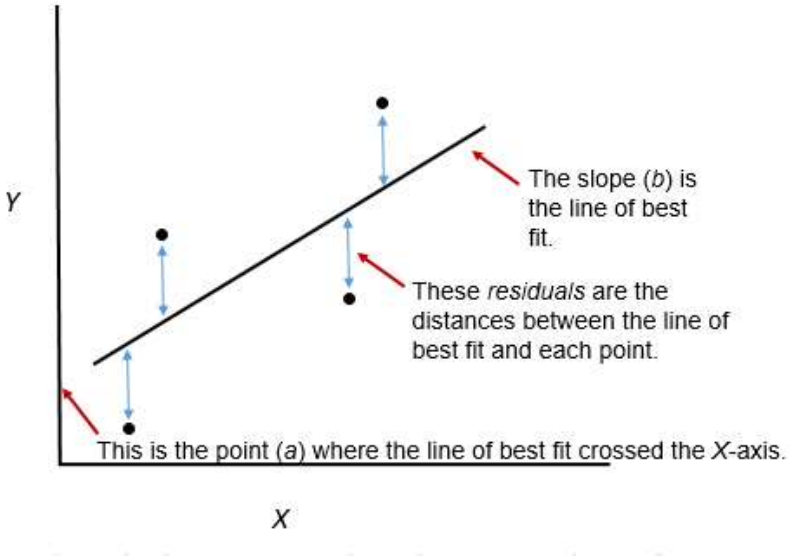
Simple regression creates the statistical model, shown above, with a single independent variable (IV), sometimes referred to as a predictor variable, and a single DV, sometimes referred to as the outcome or criterion variable. Multiple regression creates a statistical model with a single independent variable and two or more DVs. The multiple regression model is similar to the simple regression equation in that it still contains a Y-intercept, or a, but the multiple regression model contains multiple IVs and multiple corresponding coefficients, or b_x, as shown below.

$Y = a_0 + b_1X_1 + b_2X_2 + \ldots + b_nX_n$

or

$DV = a_0 + b_1(IV_1) + b_2(IV_2) + \ldots + b_n(IV_n)$

If the multiple regression model is statistically significant, at a *p* value of less than .05, the IV data can be plugged into the model to be multiplied by the calculated coefficients, added to the calculated constant (*Y*-intercept or a_0), resulting in the predicted DV.



Adapted from images in *Multiple Linear Regression* by J. Neill, 2008 (https://www.slideshare.net/jtneill/multiple-linear-regression).

Interpreting Regression Output Results

Interpreting simple and multiple regression output is similar. There are several key test statistics and *p* values that are returned in a regression analysis that must be evaluated to a) determine statistical significance and b) assess the strength of the linear regression model.

Multiple *R*: This is Pearson's *r*, as discussed in the correlation section. Regression often uses a capital *R* instead of *r*. This is simply the square root of r^2. Multiple *R* describes the strength of the correlation between the model and the dependent variable.

In the regression output below, the multiple *R* figure of 99.2% indicates a very strong positive correlation between the regression model and the dependent (output) variable.

***R* square (r^2):** This is the coefficient of determination as was discussed in the correlation section. Regression often uses a capital *R*. The square of *R* explains the amount of variation in the dependent (output) variable that is explained by the regression model.

In the regression output below, the *R* square (r^2) figure indicates that 98.3% of the variation in the dependent variable is explained by the regression model. This is a very high r^2.

ANOVA: This indicates whether the regression model is statistically significant in its ability to predict the dependent variable. ANOVA uses significance *F* for probability, and this is synonymous with the *p* value discussed previously in the course. A significance level of $F < .05$ indicates statistical significance.

In the regression output below, the significance level of $F = .000009 < .05$ would indicate that the null hypothesis should be rejected, and the alternative accepted that there is a statistically significant relationship between the regression model and the dependent variable.

The *t* Stat: This assesses the statistical significance of the individual predictor variable coefficients. A *p* value < .05 for any given *t* stat indicates statistical significance for the corresponding coefficient.

In the regression output below, the *p* values for the coefficients of variables 1, 2, and 3 are all <.05, which indicates that they each make a statistically significant contribution to the regression model (Field, 2005).

In multiple regression, the output will not always show statistical significance for all coefficients in the model. There is debate among researchers about how to treat non-significant coefficients. Some researchers recommend removing the non-significant variable(s) and rerunning the regression analysis with only the statistically significant predictor variable (Field, 2005). Other researchers recommend leaving all the coefficients in the regression model, even if some are not statistically significant. Still others recommend more complicated ways to treat the data. Regardless of these many conflicting suggestions, the important thing to

note is that non-statistically significant independent predictor variable coefficients do not contribute significantly to the regression model and, therefore, have no influence on predicting the dependent variable, even if they are left in the model. It is only the statistically significant coefficients that have predictive power in the regression model.

Multiple Regression and Correlation Analysis – Excel Output for 3 Independent Variables

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.9916677
R Square	0.9834049
Adjusted R Square	0.9751073
Standard Error	0.2861281
Observations	10

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	29.1087841	9.702928	118.51727	9.93505E-06
Residual	6	0.491215904	0.081869		
Total	9	29.6			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-45.796348	4.877650787	-9.38902	8.288E-05	-57.73152916	--3.8612
X Variable 1	0.5969718	0.081124288	7.358731	0.0003225	0.398467818	0.795476
X Variable 2	1.1768377	0.084074178	13.99761	8.29E-06	0.971115623	1.38256
X Variable 3	0.4051086	0.042233591	9.592096	7.341E-05	0.301766772	0.508451

After evaluating the regression output and determining that the model explains a great deal of variability in the dependent variable and is strongly correlated with the dependent variable and that there is a statistically significant relationship between the regression model and the dependent variable, the model can be expressed as a predictive equation.

DV = -45.796348 + 0.5969718(Variable 1) + 1.1768377(Variable 2) + 0.4051086(Variable 3)

The model can then be used to predict the DV by plugging in arbitrary values for Variable 1, Variable 2, and Variable 3.

To interpret the output for simple regression, the same steps are followed as in multiple regression. The only difference in the process is that there will only be one predictor variable and coefficient that will be analyzed for statistical significance, rather than multiple predictor variables.

References

Field, A. (2005). *Discovering stats using SPSS* (2nd ed.). SAGE.

Kerr, R., Garvin, J., Heaton, N., & Boyle, E. (2006). Emotional intelligence and leadership effectiveness. *Leadership & Organization Development Journal*, *27*(4), 265–279.

Neill, J. (2008). *Multiple linear regression* [PowerPoint slides]. https://www.slideshare.net/jtneill/multiple-linear-regression

Norusis, M. J. (2008). *SPSS 16.0 guide to data analysis* (2nd ed.). Upper Saddle River, NJ: Prentice Hall.

Phanny, I. (2014). *Guideline for interpreting correlation coefficient* [PowerPoint slides]. SlideShare. https://www.slideshare.net/phannithrupp/guideline-for-interpreting-correlation-coefficient/2

Zikmund, W. G., Babin, B. J., Carr, J. C., & Griffin, M. (2013). *Business research methods* (9th ed.). Cengage.

Suggested Unit Resources

The following video shows how to calculate Pearson's correlation coefficient (Pearson's *r*) by long-hand. It provides a great appreciation for software.

statslectures. (2010, October 3). *Pearson's r correlation* [Video]. YouTube. https://www.youtube.com/watch?v=2B_UW-RweSE

A transcript of this video is available.