LEARNING OBJECTIVES

- Learn techniques for analyzing ratio outcome variables.
- Conduct a test for analysis of variance to determine whether within groups differences are less than between group differences.
- Conduct a simple linear regression to understand the concept of effect size.
- Rephrase findings using natural language

DIRECTIONS FOR ANOVA:

Analysis of Variance (ANOVA) is a method for determining if association exists between two variables when one variable is nominal and the second variable is ratio. In this example, we will examine three variables from the PEW Study to address the following question?

Does a person's age have some association with their perception of the social benefit of online environments?

TOGETHER (pial8a)	Have you ever experienced any of the following things online? Have you ever seen an online group come together to help a person or a community solve a problem? 1 Yes 2 No 8 Don't know 9 Refused
GOODFORSOCIETY (pial11)	Overall, when you add up all the advantages and disadvantages of the internet, would you say the internet has mostly been a GOOD thing or a BAD thing for society? 1 Good thing 2 Bad thing 3 **(VOL.)** Some of both
AGE	Record respondent's age in years

Typically, the ratio variable is the outcome variable, and most tests are structured as to whether or not group membership has an influence on the outcome variable. However, in this example, the ratio variable would be the independent variable.

1. How come AGE cannot logically serve as the dependent variable (outcome variable) in this example?

2. Write a null hypothesis and alternate hypothesis for **AGE** and **GOODFOR SOCIETY**.

3. Write a null hypothesis and alternate hypothesis for **AGE** and **TOGETHER**.

Open the workbook to the tab labeled ANOVA Sheet. Notice that that data is arranged differently than you see in the Working Sheet. For each variable, the indicators are arranged as columns. In each column are the responses to AGE for all the subjects for each indicator. In essence, the responses have been grouped according to the indicators for **GOOD FOR SOCIETY**, and then again for **TOGETHER**. Each indicator does not have the same level of subject in each column.

To calculate an ANOVA, go to the Data tab and select Data Analysis. Find the test for single factor ANOVA. Using the menu, enter the range for the data (be sure to indicate that your data has labels).

Find the box marked Alpha. This is referring the p value. In this test do you want to avoid Type I or Type II error? For this exercise, use a p value of 0.05.

Run the ANOVA test for **AGE** and **GOODFOR SOCIETY**.

4. Do you accept or reject the null hypothesis? Is there an association between these two variables? Express your finding in natural language.

Run the ANOVA test for **AGE** and **TOGETHER**.

5. Do you accept or reject the null hypothesis? Is there an association between these two variables? Express your finding in natural language.

Choose one other variable to test for an association with AGE. Copy the data from the working sheet, and configure the data in two columns so you can conduct your test.

6. Write your null and alternate hypotheses.
7. After doing the test, do you accept or reject the null hypothesis? Is there an association between these two variables? Express your finding in natural language.

DIRECTIONS FOR CORRELATION:

Correlation of two ration variables is a method for determining both association and effect size. The effect size is a measure of how much a change in one variable has a measurable effect on the change in another. How small or large is the change?

For example, if you wanted to learn the relationship between investment in staff training and how this might pay off in terms of sales, you could correlate the number of hours an employee is trained and the amount of sales that employee achieved in a given time period following the training. If there is an association, this would enable the researcher to make a statement such as, "For every hour spent in training, employees will increase their weekly sales by $1,000' (or whatever figure the data indicates).

Open the workbook to the tab labeled Correlation worksheet. The workbook contains data on 15 subjects. These are students who reported how many hours of sleep they had the night before, and an observation of how many minutes early or late they arrived to class (students were asked to report the time they arrived at class, and the number of minutes early or late was calculated from this).

In this example, we will be computing a common statistic used to test for correlation between two variables – Pearson's Correlation Coefficient – which is commonly designated by the letter *r*.

First, determine whether you want to avoid Type I or Type II error. Set your p value accordingly, and identify the corresponding critical value from the following table (Excel will not automatically calculate a p value for us, so you need to choose it manually in this example). The degrees of freedom are equal to the number of subjects minus the number of variables (n-2).

Critical values of r

two-tailed p value / degrees of freedom	0.100	0.050	0.010
1	0.988	0.997	1.000
2	0.900	0.950	0.990
3	0.805	0.878	0.959
4	0.729	0.811	0.917
5	0.669	0.754	0.875
6	0.621	0.707	0.834
7	0.582	0.666	0.798
8	0.549	0.632	0.765
9	0.521	0.602	0.735
10	0.497	0.576	0.708
11	0.476	0.553	0.684
12	0.458	0.532	0.661
13	0.441	0.514	0.641
14	0.426	0.497	0.623
15	0.412	0.482	0.606

8. Which critical value did you identify? Make a note of it in your workbook.
9. Write a null and alternate hypothesis.

Before you calculate the correlation coefficient, first calculate descriptive statistics for your ratio variables.

- Select Data Analysis on the Data tab and choose Descriptive Statistics.
- Click in the **Input Range** window, and select the area that contains the data (you don't need the USERID column). Make sure your input range includes names for the variables (so that the output is properly labeled). Check the **Summary Statistics** box
- For output options,. select the **New Worksheet Ply** button and type a name for the new sheet (e.g., Descriptive Statistics).

10. Examine the output. You should see two sets of statistics, one for the hours of sleep, and one for the minutes early or late to class. How would you characterize the distribution of these two variables? How large is the range? Is the distribution skewed or normal (hint: how much distance is there between the mean and the median)?

You can now examine the strength and direction of the association, if any, between the two variables by calculating the Pearson correlation coefficient.

- Select Data Analysis on the Data tab and choose Correlation.
- Click in the **Input Range** window, and select the area that contains the data
- For output options, select the **New Worksheet Ply** button and type a name for the new sheet.
- Examine the output. You should see a simple correlation matrix with four cells. Notice that each variable has a perfect correlation of 1 with itself. Also notice that there are no figures reported in the upper right-hand half of the matrix. This is because the figures in this part of the matrix would be identical with the figures in the lower left-hand part.
- The value of the Pearson Correlation Coefficient should appear in the lower left-hand box.

11. What value did you calculate? Is this higher or lower than your critical value? Do you accept or reject the null hypothesis?

It is also helpful to view the data graphically, so we will create a scatter-plot to display the data and calculate a linear relationship between the two variables. This is called a regression.

- Select the two columns on the data worksheet containing your data.
- On the Insert tab, choose a simple scatter chart (dots, no lines).

- The chart will appear on the Correlation Sheet. Select Move Chart from the Chart Tools Design tab, and place the chart on a new sheet, with an appropriate name.
- The default chart will be incorrectly labeled. Use the options on the Chart Tools Layout tab to add a correct title, and to label the x- and y-axes. Remove the legend.
- Finally, add a trend-line. Click on the Trend-line option on the Chart Tools Layout tab, and select more options. Choose the linear trend-line and check the options to display the Equation and R-squared value on the chart. (Double-check for yourself that the R-squared value is equal to the square of Pearson's r that you calculated earlier).

- Your chart now shows a trend-line, which is the line of best fit through the points, the regression equation and the square of the correlation coefficient. You may need to move the text-box holding the equation to display it clearly.

The slope of the line indicates the effect size of hours of sleep on minutes being early or late to class. For every hour of sleep, it will indicate a positive or negative number of minutes.

The R-squared value is the proportion of variance in the outcome variable that is explained by the value of the independent variable.

12. What can you conclude about the correlation between these two variables from this this analysis? Is there a positive association, negative association, or no association? How would you describe the effect size?