Lab 3:
The Sampling Distribution, Interval Estimation, Correlation and Covariance

Objectives:

1. Complete an example of a t-test, or two.
2. Still more practice with Excel: equations and functions.
3. ***Bivariate distributions*** – relationship between two variables.
4. Introduce the concepts of ***covariance*** and ***correlation*** – variables that are related to each other.
5. ***Descriptive measures*** of covariance and correlation.

Key Terms:

1. ***Hypothesis test*** – a ***t-test***.
2. ***Scatter Diagrams*** / X-Y Graphs.
3. ***Covariance*** and ***correlation***.

Data: Your Excel file from Lab 2, which continues our use of the Excel file: 2004 *Accord Prices.xls*.

Exercises:

- *t-tests - Tests of a True Null Hypothesis*
  1. A t-test is conducted when you have no idea about the true population standard deviation, which is the usual case in data analysis. Let's complete the following test using your two samples of size 30 drawn last week:
     - H_o: $\mu_Y = \$21{,}029$; H_a: $\mu_Y \neq \$21{,}029$.
     - *Choose α.* Let's choose $\alpha = 0.05$.
     - *Determine the critical value(s).* What kind of test are we doing? At this point, you can draw a lovely picture of your test. No results yet – just a picture of the test. We always do standardized tests in this class – they're a bit easier.
     - *Draw your sample and estimate; then, calculate the test statistic.* We always conduct tests as ***standardized tests*** so the test statistics will be a calculated t-value.
     - *Where does your test statistic fall? In which region? (You'll have two tests.)*
     - *What is your conclusion. (You'll have two conclusions.)*

  2. What should we expect for our experiment? Why? Explain.

- *Used Honda Accord Price-Mileage Relationships*

  1. We are now going to investigate relationships between two variables, bivariate relationships. We know that the Used Honda Accord Price (Y) was a random variable, but there maybe reasons for some of the differences in prices. One possibility that we all would expect is that price and mileage are related for these cars. There are several ways to see if there is such a relationship. First, we'll look at an **XY Scatter diagram**. Select the worksheet: **2004 Used Accord Data**. Go to the **Insert** ribbon (click the **Insert** tab). You'll see an icon for a **Scatter** diagram about in the middle of the ribbon. Click on **Scatter** and then choose a chart type. I chose the first one – Scatter with only Markers. Next, select a Layout: I like Layout 3 and Layout 9 for the Scatter Diagrams. Once you've made these basic choices, you need to enter the data. This screen is much the same as our data entry for the Histograms. Place Accord price on the vertical axis (the ***Y variable***) and mileage on the horizontal axis (the ***X variable***). What general relationship do you observe? Is this what you expected? You can then edit the title and axis labels (we must have excellent titles and labels) as well as the legend, text fonts, etc. Click on the data in the graph to get rid of the annoying lines and to reduce the marker size (**Format Data Series**, **Marker Options**, then **select Built-in**, which

will allow you to reduce the marker size from 7 to, say, 4. You can edit/modify anything in the graph upon which you can right-click.

2. Two summary measures for the relationship between two variables are covariance and correlation. Insert another worksheet and copy the prices and mileages for all 100 cars to that page. Create the columns necessary to calculate covariance and correlation, two summary measures. Look at the formulas below. **You need to create columns for the deviations of both X_i (mileage) and Y_i (price), the deviations squared and the product of the deviations**. The equations for covariance and correlations use the column sums:

$$\text{Covariance:} \quad \sigma_{XY} = \frac{\sum_{i=1}^{N}\left[(X_i - \mu_X)(Y_i - \mu_Y)\right]}{N}$$

$$\text{Correlation:} \quad \rho_{XY} = \frac{\left[\left(\sum (X_i - \mu_X)(Y_i - \mu_Y)\right)/N\right]}{\sqrt{\left(\sum (X_i - \mu_X)^2\right)/N}\sqrt{\left(\sum (Y_i - \mu_Y)^2\right)/N}} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

Notice that you also need the two population means (μ_X *and* μ_Y) and the two population standard deviations (σ_X and σ_Y). Use Excel functions to determine these. Place the results in columns to the right of the columns you've created for the deviations and products of deviations. The formulas above are both population measures. There are also sample statistics for covariance (s_{XY}) and correlation (r_{XY}) the denominators are (n – 1).

3. Now check your covariance and correlation results using Excel functions. The Excel function for covariance is "=COVAR(…)" and the function for correlation is "=CORREL(…)."

4. Check again using **Data Analysis** tools. Did you expect to find a difference between what you calculated and the value from Excel's data analysis tool?

5. **Interpretations**. Covariance really doesn't have a good interpretation. It only tells us if data are positively linearly associated, or negatively linearly associated. For that reason, we don't use covariance that often, but it does have applications to portfolio analysis. Correlation is preferred in most statistical work because it measures the ***strength of linear association***. If two variables are perfectly positively linearly associated, the correlation coefficient between those variables will equal 1. If they are perfectly negatively linearly associated, the correlation will be -1. Variables that are unrelated will have a correlation coefficient of about 0. To find out how closely associated variables are, we use the correlation coefficient. The closer to 1, the stronger the degree of positive association. The closer to -1, the stronger the degree of negative association.

6. We're getting really close to estimating a regression line – the line of "best fit" through these data. Excel will estimate and place a regression line through the data. To do that, click on one of the data markers in the scatter diagram. Once they are highlighted, click the right mouse button and choose **Add Trendline…** ; we want a **linear trendline**. You can also have Excel give the equation for the line by choosing the option **Display equation on chart**. After clicking Ok, you should see a line appear on the graph with the equation for that line. (You can move the equation anywhere on the graph, the line obviously has to stay put.) The equation provides the intercept and the slope for the XY relationship. This is the Ordinary Least Squares line, the best linear equation can fit to these data. We'll learn a lot more about these lines as we move on through the course.

7. Save your spreadsheet one last time before you leave.