

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

**Objectives:**

1. Still more practice with Excel: equations and functions.
2. ***Bivariate distributions*** – relationship between two variables.
3. Introduce the concepts of ***covariance*** and ***correlation*** – variables that are related to each other.
4. ***Descriptive measures*** of covariance and correlation.

**Key Terms:**

1. ***Scatter Diagrams*** / X-Y Graphs.
2. ***Covariance*** and ***correlation***.

**Data:** Your Excel file from Lab 2, which continues our use of the Excel file: AmherstHomeSales2006.xlsx. And an additional data set on home prices in Belchertown and Amherst.

**Exercises:**

◆ ***Two-Sample Hypothesis Test of Means***

1. The question we ask is: ‘Are the population mean house prices the same in Amherst and Belchertown?’ Thus, our hypothesis can be stated:  $H_0: \mu_A = \mu_B$ ;  $H_A: \mu_A \neq \mu_B$ .
2. If the sample means for both Amherst and Belchertown homes are normally distributed (CLT here), then the difference is also normal, and we can apply the Z-test, or t-test if we don’t know the true standard deviations.
3. We presume no knowledge of the true population standard deviations. Thus, our safest choice for test statistics is a formulation that assumes unequal variances:

$$t_{calc} = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\left(\frac{s_1^2}{n_1}\right) + \left(\frac{s_2^2}{n_2}\right)}} , \text{ the degrees of freedom calculation for this statistic is a bit}$$

involved and will fall somewhere between the sample sizes for the two samples. Luckily, as we’ll see, Excel takes care of this for us.

4. We could easily calculate that in Excel, but if we go to **Data** and **Data Analysis**, we’ll find **t-test: Two-Sample Assuming Unequal Variances**. Click OK. The next window asks for the data range for each of the “two variables.” Here, we’ll treat the Amherst prices and Belchertown prices as different variables. Whoa! What do we need to do first before trying this test??
5. Right, once you’ve done that, highlight the Amherst prices for Variable 1 and the Belchertown prices for Variable 2. What is the **Hypothesized Mean Difference**? You also get to choose  $\alpha$ . Then select the **Output Range**: a cell with lots of space below and to the right. Now click OK and Bingo! Your two-sample t-test is complete.

◆ ***Amherst and Belchertown Home Price and Characteristic Relationships***

1. Let’s investigate associations between home prices (sale prices) and the characteristics of the homes. These relationships between two variables are called bivariate relationships. We would consider that a home’s price depends upon its characteristics such as size, number of rooms, age,

acreage, etc. We would thus say that price is the dependent (Y) variable and the other characteristics are independent (X) variables.

2. **XY Scatter diagrams** can help us see (visually) any relationships/associations between variables. Open the *Amherst and Belchertown Home Prices.xlsx* spreadsheet available in the “L” folder or on the course website. Click on an empty cell away from the data.
3. Go to the **Insert** ribbon and click the icon for a scatter diagram – it’s called, appropriately, **Scatter**. You’ll see some choices, let’s just go with the first one, “Scatter with only markers.”
4. We next have to choose a layout – let’s go with the first one in the list – markers and labels.
5. Next, we click **Select Data and Add**.
6. Because we have good reason to pick **Price** as the dependent variable, it will go on the vertical or Y axis. For the X variable, you’re free to make your own choices. Which of the variables provided do you think will give the strongest relationship? Create a few XY Scatter diagrams and see if you can determine which variable seems most closely associated with price.
7. You can then edit the titles 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.

◆ Covariance and Correlation – numeric measures of bivariate association.

1. Two summary measures for the relationship between two variables are covariance and correlation. Insert another worksheet and copy the data for Price and Age to that spreadsheet. Be sure you keep the proper data together! The easiest way to do this is probably to create a copy of the worksheet, then simply delete the other columns. 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  $X_i$  (age) and  $Y_i$  (price), the deviations squared and the product of the deviations.** The equations for covariance and correlations use the column sums for the products of deviations and the deviations squared:

$$\text{Covariance: } s_{XY} = \frac{\sum_{i=1}^N [(X_i - \bar{X})(Y_i - \bar{Y})]}{n - 1}$$

$$\text{Correlation: } r_{XY} = \frac{[(\sum (X_i - \bar{X})(Y_i - \bar{Y})) / (n - 1)]}{\sqrt{(\sum (X_i - \bar{X})^2) / (n - 1)} \sqrt{(\sum (Y_i - \bar{Y})^2) / (n - 1)}} = \frac{s_{XY}}{s_X s_Y}$$

Notice that you also need the two means ( $\bar{X}$  and  $\bar{Y}$ ) and the standard deviations ( $s_X$  and  $s_Y$ ). 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.

2. Now check your covariance and correlation results using Excel functions. The Excel function for covariance is “=COVAR(…)” and the function for correlation is “=CORREL(…)”
3. Check again using **Data Analysis** tools. Choose Covariance and highlight a number of variables all at once. Be sure to highlight the labels – they’ll be needed in the table that appears. Did you expect to find a difference between what you calculated and the value from Excel’s data analysis tool? Do the same for Correlation – Excel will create a table for all variables selected.

4. **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.
5. 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.

◆ *Amherst Single Family Assessed Value – Sales Price Relationships*

1. We know that the Amherst single family home sale price (Y) was a random variable, and when relationships with home characteristics are considered, we would place price on the vertical axis. But for local assessors, they are to create assessed values based on market prices. Let's check to see if it seems the assessors are assigning assessed values that are closely associated with market prices. Here we would say that **Assessed Value** (Y) depends upon market prices. In this Scatter Diagram, let's place **Assessed Value** on the vertical (Y) axis and **Sale Price** on the horizontal (X) axis. What general relationship do you observe? Is this what you expected? Put a trendline on the graph as well.
2. Determine the *correlation coefficient* for Assessed Value and Sale Price. How strong is the linear association?