

Statistical analysis of research data



Statistical Analysis of Research Data: Using Microsoft Excel

Instructor:

Paul W. Thurman

Columbia University
Joseph L. Mailman School of Public Health
School of International and Public Affairs
Graduate School of Business

Paul.Thurman@Columbia.edu
(917) 647-9090

Host:

Dr. Terry Moody

MoodyT@mail.nih.gov
(240) 276-7785

June 8, 2016

ST agenda

Statistical Analysis of Research Data: Using Microsoft Excel

Building 10, Room 2-3750
National Institutes of Health
Bethesda, Maryland

Sponsored by the NCI Center for Cancer Research, Office of Training and Education

The instructor is Paul Thurman of Columbia University (Paul.Thurman@Columbia.edu).

AGENDA

June 8, 2016: Basic Continuous Data Analysis Using Microsoft Excel

- 1:00 pm: • Excel setup; confirm that everyone has loads tools and copies files
- 1:15 pm: • Descriptive Statistics (Tutorial #1)
- 1:30 pm • Hypothesis Testing of Two Sample Means (Tutorial #2)
- 2:15 pm: BREAK
- 2:30 pm: • Correlation and Linear Regression (Tutorial #3)
- 3:15 pm: • One-Way Analysis of Variance (Tutorial #4)
- 3:30 pm: • Chi-Squared Tests for Randomized Response (Tutorial #5)
- 4:00 pm: ADJOURN

For additional information contact the Course Coordinator, Dr. Terry Moody (moodyt@mail.nih.gov). Phone: 240-276-7785.

Microsoft excel

Statistical Analysis of Research Data: Using Microsoft Excel

To explore ways to use simple Microsoft Excel tools to conduct the various continuous data analyses outlined in the SARD course, we will use some sample data (from published medical, drug, and nutrition studies). The workbook entitled, *SARD Tutorial Data.xls*, comprises seven different Microsoft Excel spreadsheets, which contain data we will use during today's session:

- “Desc Stats”: Concentrations of serum IgM in 298 children aged 6 months to 6 years
- “Two Samp Large”: Mean dietary intake over 10 pre-menstrual and 10 post-menstrual days (paired samples), and 24 hour total energy expenditure (MJ/day) in groups of lean and obese women (independent samples)
- “Two Samp Small”: Same data as in “Two Samp Large” but with smaller sample sizes
- “Sing Var Reg”: Data from 40 type 1 diabetic patients
- “Multi Var Reg”: Data for 50 patients being treated with digoxin for congestive heart failure
- “ANOVA”: Red cell folate levels (microgram/liter) in three groups of cardiac bypass patients given different levels of nitrous oxide ventilation
- “Chi Squared”: Observed frequencies of patients with various potassium levels and various medical conditions (e.g., hypertension)

Below are some analyses that we will perform—in real time with Microsoft Excel and the datasets, above—during our afternoon, hands-on working session. Solving all of these will require many of the Data Analysis Toolkit functions in Microsoft Excel¹:

1. Descriptive Statistics and Histograms
 - a. Describe serum levels using both descriptive statistics and a histogram.
 - b. Are there any outliers?
 - c. Which statistic is best to use to summarize central tendency?
2. Two-Sample Hypothesis Tests of Means – Large Samples
 - a. For the dietary intake dataset, is there a statistically significant difference in mean intake before vs. after menstruation? What type of test is required?
 - b. For the energy expenditure dataset, is there a difference in mean energy expended between lean and obese patients? What type of test is required, here?

¹ Tutorials for the required Excel functions are included later in this course packet.

Microsoft excel

3. Two-Sample Hypothesis Tests of Means – Small Samples
 - a. Repeat the same analyses as in question 2.
 - b. Be sure to state any/all assumptions required in order to perform the requisite hypothesis tests of two means.
4. Single-Variable Linear Regression
 - a. Construct a linear model that relates blood glucose (input) to shortening velocity (output).
 - b. Is this model significant; that is, does the input significantly predict the output?
 - c. How much variance in velocity does glucose explain using this model?
 - d. Examine the residuals—do you see any possible violations or heteroscedasticity?
 - e. Construct a 95% prediction interval for velocity given a glucose level of 14.2 mmol/l.
 - f. Construct a similar prediction interval for velocity given a glucose level of 20.2 mmol/l.
5. Multiple-Variable Linear Regression
 - a. Which independent variable—creatinine clearance or urine flow—best correlates with digoxin clearance?
 - b. Build a single-variable model to predict digoxin clearance using this best correlate. How good is this model in terms of fit, predictive power, etc.?
 - c. Now, build a multiple-input model using both independent variables to predict digoxin clearance. Comment on regression statistics regarding explained variance, predictive power, and parameter-specific significance (or not). Also examine residuals for each independent variable and comment on any patterns or possible assumption violations.
 - d. Colleagues have looked at this data and stated that digoxin clearance is highly correlated with creatinine clearance but largely independent of urine flow. Is this true? If so, what does this mean in simple terms?
6. One-Way Analysis of Variance
 - a. Do all three treatment groups have the same red folate levels?
 - b. That is, are the red folate levels independent of the treatment option? If not, why not?
7. Chi-Squared Test for Randomized Response
 - a. Is there a relationship between potassium level and hypertension, or are potassium levels largely independent of potassium level? Said another way, are the there potassium levels just randomly/proportionally scattered throughout hypertensive patients, or is the distribution of them not random?
 - b. (OPTIONAL) What about a relationship between potassium level and gender?
 - c. (OPTIONAL) Note that the hyperkalemic group is quite small. Re-run your analyses without this group and see if your conclusions change. Be sure to state your hypotheses, statistical tests, and conclusions.

Managerial Statistics

Managerial Statistics Columbia Executive MBA Program

Excel Tutorial #1: Descriptive Statistics¹

In this tutorial, we discuss how to use the “Data Analysis” add-in in Excel to compute various descriptive statistics for data ranges.

Adding the “Add-Ins” to the Tools Menu

In order to use the “Data Analysis” toolkit, you’ll need to add them to your Tools menu:

| Step | Action | Result |
|------|---|--|
| 1 | Run Excel and select the Tools menu from the menu bar | Tools menu displays starting with “Spelling...” “Auditing,” etc. |
| 2 | See if the option “Data Analysis...” appears at the bottom of the menu (after “Options...”) | If you see the “Data Analysis...” menu option, skip to the next section, <i>Computing....</i> If not, continue with Step 3 |
| 3 | Select the “Add-Ins...” option in this menu (about half-way down the list) | After several seconds, a new window will appear entitled, “Add-Ins” with several checkbox entries in a list |
| 4 | Check the boxes next to the “Analysis Toolpak” and “Analysis Toolpak - VBA” add-ins | Boxes next to these add-ins will be checked with a checkmark |
| 5 | Also, it’s a good idea to check the “AutoSave” add-in so that Excel prompts you to save your work periodically. | Box next to the “AutoSave” add-in will be checked |
| 6 | Push the “OK” button to the right of the list of add-ins | Excel will now “add-in” the functionality of these toolpaks. This may take a few seconds |
| 7 | Repeat steps 1 and 2 in this list | You should now see the “Data Analysis...” option in the “Tools” menu |

Now, we’re ready to use this toolpak to do some statistical analyses!

Computing Descriptive Statistics for Spreadsheet Data

Let’s start by entering some data into a blank spreadsheet to use for this (and future) tutorials. Enter the information below into “Sheet1” in Excel:

| | A | B | C | D | E | F |
|---|--------------------------------|---------|---------|---------|---------|----------|
| 1 | Monthly Rates of Return | | | | | |
| 2 | | | | | | |
| 3 | Date | S&P | Viacom | AT&T | GM | Coke |
| 4 | 01/30/98 | 0.8799 | 0.7541 | 2.1407 | -4.6296 | -18.8406 |
| 5 | 02/27/98 | 7.5187 | 14.9701 | -2.5948 | 18.9860 | 6.6964 |
| 6 | 03/31/98 | 5.5580 | 11.9792 | 7.7869 | -1.7226 | -3.3473 |
| 7 | 04/30/98 | 1.3716 | 7.9070 | -8.5551 | -0.5535 | 5.8442 |
| 8 | 05/29/98 | -1.6289 | -5.1724 | 1.2474 | 6.6790 | 1.9427 |
| 9 | 06/27/98 | 2.4171 | 3.4091 | 0.8214 | 1.8261 | 2.1063 |

¹ The steps and results shown assume a Windows 95 installation. Similar results occur under Windows 97.

Managerial statistics

Managerial Statistics Columbia Executive MBA Program

Excel Tutorial #1: Descriptive Statistics (Continued)...

Note that these are real monthly rates of return for the stocks shown in the first half of 1998. Now, let's compute some simple descriptive statistics for these data; e.g., maximum, minimum, range, mean, median, variance, and standard deviation. To compute these, we could use the Function Wizard and insert formulae for these statistics below the columns of data like we did in math camp.

However, let's use our new-found "Data Analysis..." toolkit to help us here. Let's look at some descriptive statistics for the S&P 500 data in the B column of the data. To do this, we follow the steps below:

| Step | Action | Result |
|------|--|---|
| 1 | Select the "Tools" menu and then the "Data Analysis..." menu item (NOTE: You do not need to select any data range, yet) | A new window appears entitled, "Data Analysis" with several options listed including "Anova: Single Factor," etc. |
| 2 | Select the "Descriptive Statistics" option and either double-click or press the "OK" button on the upper-right | A new window appears entitled, "Descriptive Statistics" with several fields and checkboxes |
| 3 | Now, we select the data range to analyze. Select the range, B3:B9 on the original "Sheet1". Note that this includes the column label, "S&P" | The "Input Range" field in the "Descriptive Statistics" window now contains the range, B3:B9 |
| 4 | Next, since the S&P data are in a column, select the "Grouped by Columns" option. Also, since we have included labels with our data, check the "Labels in First Row" selection | "Grouped by Columns" option should be bulleted, and the "Labels in First Row" option should be checked |
| 5 | Finally, select the "Summary Statistics" box at the bottom of the window so that the toolpak will give us descriptive statistics. Also, if you want the results in a particular sheet (with a name), select the "New Worksheet Ply:" option and provide a name in the blank to the side; e.g., "S&P Stats" | "Summary Statistics" box is checked, and the "New Worksheet Ply:" option is selected (with a name for the new worksheet, if applicable) |
| 6 | Select no other options or features, and press the OK button in the upper-right | Descriptive statistics should appear in a new worksheet! |

Note that mean, (sample) variance, standard deviation, range, minimum, and maximum are shown in the output. Other statistics, such as the kurtosis, skewness, and standard error are also shown, but we will discuss these later in the term.

Finally, note that we could have performed the steps above on all of our stock data at the same time. That is, if we had selected all of our data—i.e., the range B3:F9—Excel would have created summary statistics for all of the data as shown on the following page:

Managerial Statistics

Managerial Statistics Columbia Executive MBA Program

Excel Tutorial #1: Descriptive Statistics (Continued)...

| <i>S&P</i> | | <i>Viacom</i> | | <i>AT&T</i> | | <i>GM</i> | | <i>Coke</i> | |
|-------------------------|--------------|-------------------------|--------------|-------------------------|--------------|-------------------------|-------------|-------------------------|-------------|
| Mean | 2.686066667 | Mean | 5.641183333 | Mean | 0.141083333 | Mean | 3.4309 | Mean | -0.93305 |
| Standard Error | 1.357493856 | Standard Error | 3.044844991 | Standard Error | 2.215493992 | Standard Error | 3.476072202 | Standard Error | 3.865040584 |
| Median | 1.89435 | Median | 5.65805 | Median | 1.0344 | Median | 0.6363 | Median | 2.0245 |
| Mode | #N/A | Mode | #N/A | Mode | #N/A | Mode | #N/A | Mode | #N/A |
| Standard Deviation | 3.325167276 | Standard Deviation | 7.458316575 | Standard Deviation | 5.426829808 | Standard Deviation | 8.514603203 | Standard Deviation | 9.467377265 |
| Sample Variance | 11.05673741 | Sample Variance | 55.62648613 | Sample Variance | 29.45048177 | Sample Variance | 72.4984677 | Sample Variance | 89.63123228 |
| Kurtosis | -0.674386392 | Kurtosis | -0.937691407 | Kurtosis | 1.183192445 | Kurtosis | 2.252482256 | Kurtosis | 3.205729052 |
| Skewness | 0.391898632 | Skewness | -0.227195286 | Skewness | -0.415749762 | Skewness | 1.492009576 | Skewness | -1.74979474 |
| Range | 9.1476 | Range | 20.1425 | Range | 16.342 | Range | 23.6156 | Range | 25.537 |
| Minimum | -1.6289 | Minimum | -5.1724 | Minimum | -8.5551 | Minimum | -4.6296 | Minimum | -18.8406 |
| Maximum | 7.5187 | Maximum | 14.9701 | Maximum | 7.7869 | Maximum | 18.986 | Maximum | 6.6964 |
| Sum | 16.1164 | Sum | 33.8471 | Sum | 0.8465 | Sum | 20.5854 | Sum | -5.5983 |
| Count | 6 | Count | 6 | Count | 6 | Count | 6 | Count | 6 |
| Confidence Level(95.0%) | 3.489543346 | Confidence Level(95.0%) | 7.827010437 | Confidence Level(95.0%) | 5.695099306 | Confidence Level(95.0%) | 8.93551346 | Confidence Level(95.0%) | 9.935386884 |

Excel tutorial 2

Managerial Statistics Columbia Executive MBA Program

Excel Tutorial #2: z- and t-tests of Means and of Mean Differences¹

In this tutorial, we discuss how to use the “Data Analysis” add-in in Excel to test hypotheses of means and of mean differences. In order to complete this tutorial, you will need to have the “Data Analysis” add-in in your “Tools” menu. If you do not, please refer to *Excel Tutorial #1: Descriptive Statistics* for help on installing and accessing this toolkit.

In addition, you should have the stock data shown below (from the first tutorial) available for use:

| | A | B | C | D | E | F |
|---|--------------------------------|---------|---------|---------|---------|----------|
| 1 | Monthly Rates of Return | | | | | |
| 2 | | | | | | |
| 3 | Date | S&P | Viacom | AT&T | GM | Coke |
| 4 | 01/30/98 | 0.8799 | 0.7541 | 2.1407 | -4.6296 | -18.8406 |
| 5 | 02/27/98 | 7.5187 | 14.9701 | -2.5948 | 18.9860 | 6.6964 |
| 6 | 03/31/98 | 5.5580 | 11.9792 | 7.7869 | -1.7226 | -3.3473 |
| 7 | 04/30/98 | 1.3716 | 7.9070 | -8.5551 | -0.5535 | 5.8442 |
| 8 | 05/29/98 | -1.6289 | -5.1724 | 1.2474 | 6.6790 | 1.9427 |
| 9 | 06/27/98 | 2.4171 | 3.4091 | 0.8214 | 1.8261 | 2.1063 |

Testing Mean and Mean Difference Hypotheses

Hypothesis testing is one of the core topics of this course. One of the most common hypothesis tests performed is one of comparing average values or means of two sets of sample data. For example, if we have product defect information on two similar/comparable products—one that we use extensively today and another that we are considering adding to our inventory—it would be nice to know if, statistically, the mean number of failures of the two products are, based on a sample, the “same” or if they differ by a predetermined amount (subject, of course, to some experimental tolerance or *Type I/alpha error*).

Thus, if one product, on average, fails as much as the other, we may be led to not switch suppliers (if one product is currently our preferred choice). However, if the new product that we are examining appears to fail less often than our current one, we may be led to switch suppliers. If Product A is our current choice, Product B is the new product that we are examining, and Mean Number of Failures_A is the mean number of failures for Product A (and similarly for Product B), then the hypothesis that we are testing is one of *mean differences*:

H_0 : Mean Number of Failures_A = Mean Number of Failures_B (our null assumption)

H_a : Mean Number of Failures_A \neq Mean Number of Failures_B (our alternative)

Note here that our null or “going in” assumption is that the products are similar and, thus, fail at about the same average rate. Our alternative hypothesis is that they fail at different rates. Note

¹ The steps and results shown assume a Windows 95 installation. Similar results occur under Windows 97.

Excel tutorial 2

Managerial Statistics Columbia Executive MBA Program

Excel Tutorial #2: z- and t-tests of Means and of Mean Differences (Continued)...

that this is a two-tail hypothesis test. If we wanted to test $\text{Failures}_A \geq \text{Failures}_B$ we could use a one-tail test, instead.

If we re-write these hypotheses, we can see that we are testing *mean differences*:

$$H_0: (\text{Mean Number of Failures}_A) - (\text{Mean Number of Failures}_B) = 0$$

$$H_a: (\text{Mean Number of Failures}_A) - (\text{Mean Number of Failures}_B) \neq 0$$

Fortunately, we can use Excel to “do the math” with respect to these hypothesis tests of mean differences and help us understand if we should reserve judgment (not reject the null hypothesis) or reject the null hypothesis. The “Data Analysis” toolkit has four functions devoted to helping us test mean difference hypotheses; however, care should be taken to ensure that the proper one is used:

| Data Analysis Option | Use When You Have... ² |
|---|--|
| <ul style="list-style-type: none"> z-test: Two Sample for Means | <ul style="list-style-type: none"> Independent samples/outcomes; i.e., “unpaired” observations (such as market returns from two different portfolios/managers) Large sample sizes (>30 items in both samples/lists) Known <i>population</i> variances of both data sets (computed via the =VARP(...) function) |
| <ul style="list-style-type: none"> t-test: Two-Sample Assuming Equal Variances | <ul style="list-style-type: none"> Independent samples/outcomes; i.e., “unpaired” observations Small sample sizes (<30 items in both samples/lists) An assumption that the <i>population</i> variances are equal (or near equal). (NOTE: The add-in will compute the variances of the <i>samples</i> for you. If they are very different, you may want to use the “t-test: Two-Sample Assuming Unequal Variances” below) |
| <ul style="list-style-type: none"> t-test: Two-Sample Assuming Unequal Variances | <ul style="list-style-type: none"> Independent samples/outcomes; i.e., “unpaired” observations Small sample sizes (<30 items in both samples/lists) An assumption that the <i>population</i> variances are unequal. (NOTE: The add-in will compute the variances of the <i>samples</i> for you. If they are close/near-equal, you may want to use the “t-test: Two-Sample Assuming Equal Variances” above) |
| <ul style="list-style-type: none"> t-test: Paired Two Sample for Means (OPTIONAL - we may not cover this in class) | <ul style="list-style-type: none"> Dependent samples/outcomes; i.e., “paired” observations (such as height and weight) Small sample sizes (<30 items in both samples) No assumption about population variance is required as Excel will assume that covariance exists between the data lists and will account for them via changes in the degrees of freedom used to calculate the t-statistic |

We will now discuss how to use the “Data Analysis” add-in to apply each of these tests in turn. Note that we will show the general mechanics and sample output here based on our stock return

² Note that in each case, a Type I or alpha error threshold will be required by the add-in; e.g., Alpha = 5%. Thus, care should be taken when specifying an alpha level based on whether your test is one- or two-tailed.

Excel tutorial 2

Managerial Statistics
Columbia Executive MBA Program

Excel Tutorial #2: z- and t-tests of Means and of Mean Differences (Continued)...

data, but we will not go into the details of interpreting the output. This will be left for class discussion.

z-test: Two Sample for Means

To perform a z-test of two sample means, we need to know the information below. Note that the critical elements needed here are the *population variances* for each data set. If you do not know the population statistics for the data sets, you CANNOT use this test!:

- **Variable 1 variance (known):** The *population* variance of data set 1, which is usually the data set with the larger (sample) mean. This value will need to be computed and actually typed in to the add-in field. *The add-in will NOT compute it for you nor will you be able to reference a cell containing the variance*
- **Variable 2 variance (known):** The *population* variance of data set 2, which is usually the data set with the *smaller* (sample) mean. *Same caveat applies; you must compute and hand-enter this value into the add-in*
- **Hypothesized mean difference:** The difference between the two sample means (positive number) that we believe exists. If we are testing pure equality/inequality, then this value should be zero. If we believe the (sample) means differ by at least a value x , then this value should be used. *This value must also be entered (i.e., cannot be referred to in a cell)*
- **Alpha:** The Type I tolerance. Again, be careful here that you use the right number depending on what type of test you're using. For example, if the hypothesized mean difference is zero, then we are testing whether the means are strictly equal; thus, we are using a two-tailed test. An alpha of 5% would give us 2.5% in each tail. However, if we are testing whether the mean difference is greater than x , then we are using a one-tailed test. An alpha level of 5% would give us 5% in the (right-hand) tail. *This value must be entered as a decimal, not as a percent; i.e., "0.05," not "5"*

To demonstrate this test, we will use our familiar stock return data and test the mean returns from GM and from Viacom. Here, we want to test the null hypothesis that they are the same (or that the mean difference is zero) against the alternative that they are (statistically) different (mean difference is non-zero). Note that we can use the z-test since these monthly returns are assumed to be independent:

$$H_0: \text{Average Viacom Return} - \text{Average GM Return} = 0$$
$$H_a: \text{Average Viacom Return} - \text{Average GM Return} \neq 0$$

However, we need to know the *population* variances for these data sets. For the sake of this exercise, let's assume that the population variances equal the sample variances. (IN GENERAL, HOWEVER, THIS IS NOT THE CASE!!!) The sample variances can be computed in the

Excel tutorial 2

Managerial Statistics Columbia Executive MBA Program

Excel Tutorial #2: z- and t-tests of Means and of Mean Differences (Continued)...

spreadsheet with the data via the =VAR(...) function or via the “Descriptive Statistics” tool (see *Excel Tutorial #1:...*). Below, we show some descriptive statistics in the spreadsheet:

| | A | B | C | D | E | F |
|----|--------------------------------|---------|---------|---------|---------|----------|
| 1 | Monthly Rates of Return | | | | | |
| 2 | | | | | | |
| 3 | Date | S&P | Viacom | AT&T | GM | Coke |
| 4 | 01/30/98 | 0.8799 | 0.7541 | 2.1407 | -4.6296 | -18.8406 |
| 5 | 02/27/98 | 7.5187 | 14.9701 | -2.5948 | 18.9860 | 6.6964 |
| 6 | 03/31/98 | 5.5580 | 11.9792 | 7.7869 | -1.7226 | -3.3473 |
| 7 | 04/30/98 | 1.3716 | 7.9070 | -8.5551 | -0.5535 | 5.8442 |
| 8 | 05/29/98 | -1.6289 | -5.1724 | 1.2474 | 6.6790 | 1.9427 |
| 9 | 06/27/98 | 2.4171 | 3.4091 | 0.8214 | 1.8261 | 2.1063 |
| 10 | | | | | | |
| 11 | Variance | 11.0567 | 55.6265 | 29.4505 | 72.4985 | 89.6312 |
| 12 | Std. Dev. | 3.3252 | 7.4583 | 5.4268 | 8.5146 | 9.4674 |
| 13 | Mean | 2.6861 | 5.6412 | 0.1411 | 3.4309 | -0.9331 |

Note here that the Viacom and GM (sample) average returns appear close—5.64 vs. 3.43—and that the variances, although large, are also of the same relative “size.” But is this difference in average returns due to our sampling? Is it due to the (population) variances of each being so large? Are the returns, generally speaking, the “same” for the population based on this sample data? This is what we want Excel to help us test. (Note that when we phrased our hypotheses earlier, we used the Viacom mean first since it is larger than GM’s.) Here are the steps to perform to complete this hypothesis test:

| Step | Action | Result |
|------|---|---|
| 1 | Select the “Data Analysis” option from the “Tools” menu | “Data Analysis” window appears with various options |
| 2 | Select the “z-test: Two Sample for Means” option and press the OK button | A new window entitled, “z-test: Two Sample for Means” appears |
| 3 | Now fill in the data ranges. First, fill in the “Variable 1 Range” with Viacom’s returns (and label) (C3:C9). Next, fill in “Variable 2 Range” with GM’s (E3:E9). Next, enter “0” for the “Hypothesized Mean Difference” since we are testing strict equality. Next, enter the computed variances for Viacom and GM, respectively, in the “Variable 1 <and 2> Variance (known)” fields. Again, note that you have to actually type these in (55.6265 and 72.4985, respectively). Finally, enter an “Alpha” level of 0.05 (NOT 5!) | The “Variable 1 Range,” “Variable 2 Range,” “Hypothesized Mean Difference,” “Variable 1 Variance (known),” “Variable 2 Variance (known),” and the “Alpha” fields should be filled in with the appropriate values. |
| 4 | Since we have included labels in the ranges entered in Step 3, we need to check the “Labels” checkbox | “Labels” checkbox should contain a checkmark |
| 5 | Let’s put our output in a new worksheet ply titled, “z-test” | “New Worksheet Ply:” option is selected with the name “z-test” entered |
| 6 | Now, let’s get the results! Press the OK button and watch Excel work its magic. | Hypothesis/z-test results appear in a worksheet ply named, “z-test” (see next page) |

Excel tutorial 2

Managerial Statistics
Columbia Executive MBA Program

Excel Tutorial #2: z- and t-tests of Means and of Mean Differences (Continued)...

z-Test: Two Sample for Means

| | <i>Viacom</i> | <i>GM</i> |
|------------------------------|---------------|-----------|
| Mean | 5.641183333 | 3.4309 |
| Known Variance | 55.6265 | 72.4985 |
| Observations | 6 | 6 |
| Hypothesized Mean Difference | 0 | |
| z | 0.478306888 | |
| P(Z<=z) one-tail | 0.316215903 | |
| z Critical one-tail | 1.644853 | |
| P(Z<=z) two-tail | 0.632431806 | |
| z Critical two-tail | 1.959961082 | |

Note here that the z-score (lowercase z) for our test statistic, zero, is 0.4783. Since this is not in the critical region—i.e., greater than the two-tail critical value of 1.95996—we will *not* reject the null hypothesis. That is, based on the sample data we have seen, the average returns of these two stocks are, 95% of the time, statistically equivalent. We are not led to believe that they are different (or that Viacom's average return is strictly greater than GM's → one-tail test) based on our analysis. More on how to analyze these results in class.

Just note that the important things to check here are that (1) the *population* variances you entered are correct (IF YOU ONLY HAVE SAMPLE VARIANCES, YOU CAN'T USE THIS TEST!), (2) the number of observations is correct, (3) the hypothesized mean difference value matches the statement in your hypothesis test (H_a), and (4) whether the z-value is greater than the critical values (one- or two-tail, depending on how you set up your hypothesis test).

CAUTION: We used a z-test here with only six data points and no population variance data. In general, as we outlined above, the t-tests would be appropriate here. However, we use the small data set and sample variances here with the z-test just to show you how to use the Excel/add-in tool. In general, with this few data points and/or no population variance data, you will want to perform t-tests on similar hypotheses.

Now, we describe the t-tests. Note that these follow pretty much the same setup and data entry as before (without the need to input computed sample variances).

t-test: Two-Sample Assuming Equal Variances

To perform a t-test (small sample size) when you believe the population variances are equal, perform this test in the “Data Analysis” toolpak. Note that the items required here are fewer in number than those required for the z-test; you do not need to know any population statistics/variances. (See **bolded** text for changes from instructions in prior section):

Excel tutorial 2

Managerial Statistics
Columbia Executive MBA Program

Excel Tutorial #2: z- and t-tests of Means and of Mean Differences (Continued)...

| Step | Action | Result |
|------|---|---|
| 1 | ... | ... |
| 2 | Select the “t-test: Two-Sample Assuming Equal Variances” option and press the OK button | A new window entitled, “t-test: Two-Sample Assuming Equal Variances” appears |
| 3 | Now fill in the data ranges. First, fill in the “Variable 1 Range” with Viacom’s returns (and label) (C3:C9). Next, fill in “Variable 2 Range” with GM’s (E3:E9). Next, enter “0” for the “Hypothesized Mean Difference” since we are testing strict equality. Finally, enter an “Alpha” level of 0.05 (NOT 5!) | The “Variable 1 Range,” “Variable 2 Range,” “Hypothesized Mean Difference,” and the “Alpha” fields should be filled in with the appropriate values. |
| 4 | ... | ... |
| 5 | Let’s put our output in a new worksheet ply titled, “t-test, Equal Vars” | “New Worksheet Ply:” option is selected with the name “t-test, Equal Vars” entered |
| 6 | Now, let’s get the results! Press the OK button and watch Excel work its magic. | Hypothesis/t-test results appear in a worksheet ply named, “t-test, Equal Vars” (see below) |

Here’s the output from this test:

t-Test: Two-Sample Assuming Equal Variances

| | <i>Viacom</i> | <i>GM</i> |
|------------------------------|---------------|------------|
| Mean | 5.641183333 | 3.4309 |
| Variance | 55.62648613 | 72.4984677 |
| Observations | 6 | 6 |
| Pooled Variance | 64.06247691 | |
| Hypothesized Mean Difference | 0 | |
| df | 10 | |
| t Stat | 0.478306974 | |
| P(T<=t) one-tail | 0.321358277 | |
| t Critical one-tail | 1.812461505 | |
| P(T<=t) two-tail | 0.642716555 | |
| t Critical two-tail | 2.228139238 | |

Note again here that the sample variances differ by almost 20. Thus, we may want to repeat this test but use the “t-test: Two Sample Assuming Unequal Variances.” However, after performing this test, we see that the “t Stat” of 0.4783 is not greater than the “t Critical two-tail” value of 2.2281. Thus, we will *not* reject the null hypothesis; we will reserve judgment and conclude that these two means are statistically identical (95% of the time).

Excel tutorial 2

Managerial Statistics Columbia Executive MBA Program

Excel Tutorial #2: z- and t-tests of Means and of Mean Differences (Continued)...

t-test: Two-Sample Assuming Unequal Variances

To perform the same test as before but with the assumption of *unequal* population variances, we use the “t-test: Two-Sample Assuming Unequal Variances” test. Note that the items and steps required here are essentially the same as with the ...Equal Variances test. The difference is in the way Excel computes the variance of the two data sets (i.e., not “pooled” as in the previous test). (See **bolded** text for changes from instructions in prior section):

| Step | Action | Result |
|------|---|---|
| 1 | ... | ... |
| 2 | Select the “ t-test: Two-Sample Assuming Unequal Variances ” option and press the OK button | A new window entitled, “ t-test: Two-Sample Assuming Unequal Variances ” appears |
| 3 | Now fill in the data ranges. First, fill in the “Variable 1 Range” with Viacom’s returns (and label) (C3:C9). Next, fill in “Variable 2 Range” with GM’s (E3:E9). Next, enter “0” for the “Hypothesized Mean Difference” since we are testing strict equality. Finally, enter an “Alpha” level of 0.05 (NOT 5!) | The “Variable 1 Range,” “Variable 2 Range,” “Hypothesized Mean Difference,” and the “Alpha” fields should be filled in with the appropriate values. |
| 4 | ... | ... |
| 5 | Let’s put our output in a new worksheet ply titled, “ t-test, Unequal Vars ” | “New Worksheet Ply:” option is selected with the name “ t-test, Unequal Vars ” entered |
| 6 | Now, let’s get the results! Press the OK button and watch Excel work its magic. | Hypothesis/t-test results appear in a worksheet ply named, “ t-test, Unequal Vars ” (see below) |

Here’s the output from this test:

t-Test: Two-Sample Assuming Unequal Variances

| | <i>Viacom</i> | <i>GM</i> |
|------------------------------|---------------|------------|
| Mean | 5.641183333 | 3.4309 |
| Variance | 55.62648613 | 72.4984677 |
| Observations | 6 | 6 |
| Hypothesized Mean Difference | 0 | |
| df | 10 | |
| t Stat | 0.478306974 | |
| P(T<=t) one-tail | 0.321358277 | |
| t Critical one-tail | 1.812461505 | |
| P(T<=t) two-tail | 0.642716555 | |
| t Critical two-tail | 2.228139238 | |

Excel tutorial 2

Managerial Statistics
Columbia Executive MBA Program

Excel Tutorial #2: z- and t-tests of Means and of Mean Differences (Continued)...

Again, as expected, the “t Stat”—the test statistic of 0.4783 based on our hypothesized mean difference—is not in the critical two-tail region (i.e., greater than the “t Critical two-tail” value of 2.2281). Thus, we will *not* reject the null hypothesis.

OPTIONAL: t-test: Paired Two Sample for Means

In this test, we assume that the two data sets are dependent or *paired* in some way. For example, height and weight data are highly correlated; thus, if we have these data from a sample of people, and we want to compare means, we may use this “paired sample” test.

The setup and steps are exactly the same as the other two t-tests. However, the output is different because the observations are assumed to be paired. Thus, the degrees of freedom used to calculate the t-statistic are less than the degrees of freedom that would be used for non-paired data:

t-Test: Paired Two Sample for Means

| | <i>Viacom</i> | <i>GM</i> |
|------------------------------|---------------|------------|
| Mean | 5.641183333 | 3.4309 |
| Variance | 55.62648613 | 72.4984677 |
| Observations | 6 | 6 |
| Pearson Correlation | 0.350437967 | |
| Hypothesized Mean Difference | 0 | |
| df | 5 | |
| t Stat | 0.592077573 | |
| P(T<=t) one-tail | 0.28977785 | |
| t Critical one-tail | 2.015049176 | |
| P(T<=t) two-tail | 0.5795557 | |
| t Critical two-tail | 2.570577635 | |

Note here that the degrees of freedom is “5” instead of the “10” that we got in the previous two t-tests. This changes both the test statistics—the “t stat” value—as well as the critical values. We will discuss this test, when to use it, and interpretations that can be made from it in class if we have time.

Final Thoughts

The hypothesis tests that we’ve outlined are relatively simple yet very powerful. For the most part, you will use the middle two tests: t-tests assuming either equal or unequal variances. The z-test, while helpful, is only used with large sample data sets where population variances are known (which is rare, in general).

Note that other hypothesis tests—e.g., tests of regression coefficients, F-tests for equal variances of two samples, etc.—are also available via the “Data Analysis” toolkit.

Excel tutorial 3

Managerial Statistics
Columbia Executive MBA Program

Excel Tutorial #3: Simple and Multiple Linear Regression Analysis¹

In this tutorial, we discuss how to use the “Data Analysis” add-in in Excel to create both simple and multiple (variable) linear regression models. In order to complete this tutorial, you will need to have the “Data Analysis” add-in in your “Tools” menu. If you do not, please refer to *Excel Tutorial #1: Descriptive Statistics* for help on accessing this toolkit.

In addition, you should have the stock data shown below (from the first tutorial) available for use:

| | A | B | C | D | E | F |
|---|--------------------------------|---------|---------|---------|---------|----------|
| 1 | Monthly Rates of Return | | | | | |
| 2 | | | | | | |
| 3 | Date | S&P | Viacom | AT&T | GM | Coke |
| 4 | 01/30/98 | 0.8799 | 0.7541 | 2.1407 | -4.6296 | -18.8406 |
| 5 | 02/27/98 | 7.5187 | 14.9701 | -2.5948 | 18.9860 | 6.6964 |
| 6 | 03/31/98 | 5.5580 | 11.9792 | 7.7869 | -1.7226 | -3.3473 |
| 7 | 04/30/98 | 1.3716 | 7.9070 | -8.5551 | -0.5535 | 5.8442 |
| 8 | 05/29/98 | -1.6289 | -5.1724 | 1.2474 | 6.6790 | 1.9427 |
| 9 | 06/27/98 | 2.4171 | 3.4091 | 0.8214 | 1.8261 | 2.1063 |

Simple Linear Regression Models

Regression models are helpful because they give us a way to use historical data to predict future behavior (with appropriate caveats for error estimations). The simplest regression model is given by the one-variable equation of the line:

$$Y = mX + b,$$

Where X is the *independent* or *benchmark* variable, Y is the *dependent* variable, and m and b are the slope and intercept of the line, respectively. Also note that m is the *coefficient of X* in this equation.

Now, let's assume that we would like to predict or determine Viacom's monthly return based on the S&P's return in the same month. That is, if we know the S&P return in a given month (*independent* variable), can we predict (with some level of certainty) what Viacom's return (*dependent* variable) will be in that month?

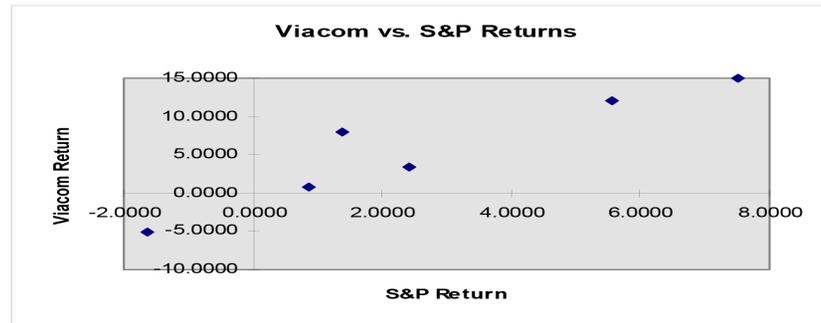
Regression analysis helps us do this assuming, of course, that there is some simple, linear relationship (with non-zero slope) between these two return variables. To see this, we can graph a scatter plot of S&P returns and Viacom returns:

¹ The steps and results shown assume a Windows 95 installation. Similar results occur under Windows 97.

Excel tutorial 3

Managerial Statistics
Columbia Executive MBA Program

Excel Tutorial #3: Simple and Multiple Linear Regression Analysis (Continued)...



Note that if we had graphed AT&T vs. S&P, no obvious (non-zero slope) relationship would have been seen; thus, a simple linear regression may not be useful. Our goal, now, is to determine the line that best “fits” these Viacom-S&P points and that will allow us, for a given S&P return, predict the Viacom return. To use Excel to help us determine the equation of the regression line—i.e., the slope and intercept of it—we will use the “Regression” function in the “Data Analysis” toolpak:

| Step | Action | Result |
|------|---|--|
| 1 | Select the “Data Analysis” option from the “Tools” menu | “Data Analysis” window appears with various options |
| 2 | Select the “Regression” option and press the OK button | A new window entitled, “Regression” appears |
| 3 | Now fill in the data ranges. First, fill in the <i>dependent</i> variable range in the “Input Y Range:” field. Here, we are trying to predict Viacom returns; thus, the range (and label) of Viacom’s returns should be entered / selected (C3:C9). For the <i>independent</i> variable, enter the range (and label) including the S&P data (B3:B9) in the “Input X Range:” field | The “Input Y Range:” field should contain the Viacom return range (C3:C9), and the “Input X Range:” field should contain the S&P return range (B3:B9) |
| 4 | Since we have included labels in the ranges entered in Step 3, we need to check the “Labels” checkbox | “Labels” checkbox should contain a checkmark |
| 5 | Let’s put our output in a new worksheet ply titled, “Simple Regression.” To do this, select the “New Worksheet Ply:” option and enter the worksheet ply name in the field to the right | “New Worksheet Ply:” option is selected with the name “Simple Regression” entered in the field to the right. (NOTE: no other options or checkboxes should be selected) |
| 6 | Now, let’s get the results! Press the OK button and watch Excel work its magic. | Regression results appear in a worksheet ply named, “Simple Regression” (see next page) |

Excel tutorial 3

Managerial Statistics
Columbia Executive MBA Program

Excel Tutorial #3: Simple and Multiple Linear Regression Analysis (Continued)...

SUMMARY OUTPUT

| <i>Regression Statistics</i> | |
|------------------------------|-------------|
| Multiple R | 0.938661647 |
| R Square | 0.881085687 |
| Adjusted R Square | 0.851357109 |
| Standard Error | 2.875496778 |
| Observations | 6 |

ANOVA

| | <i>df</i> | <i>SS</i> | <i>MS</i> | <i>F</i> | <i>Significance F</i> |
|------------|-----------|-------------|-------------|-------------|-----------------------|
| Regression | 1 | 245.0585038 | 245.0585038 | 29.63766651 | 0.005528201 |
| Residual | 4 | 33.07392688 | 8.268481719 | | |
| Total | 5 | 278.1324306 | | | |

| | <i>Coefficients</i> | <i>Standard Error</i> | <i>t Stat</i> | <i>P-value</i> | <i>Lower 95%</i> | <i>Upper 95%</i> | <i>Lower 95.0%</i> | <i>Upper 95.0%</i> |
|-----------|---------------------|-----------------------|---------------|----------------|------------------|------------------|--------------------|--------------------|
| Intercept | -0.014084478 | 1.567540082 | -0.008985083 | 0.993261301 | -4.36628248 | 4.338113524 | -4.36628248 | 4.338113524 |
| S&P | 2.105408582 | 0.386735808 | 5.444048724 | 0.005528201 | 1.031655617 | 3.179161547 | 1.031655617 | 3.179161547 |

Excel tutorial 3

Managerial Statistics
Columbia Executive MBA Program

Excel Tutorial #3: Simple and Multiple Linear Regression Analysis (Continued)...

This is a lot of data to analyze! However, we will focus only on a couple of pieces of this output for now: the “R Square” term and the “Coefficients.” The “R Square” term, or *coefficient of determination*, tells us how good of a “fit” our regression line is to the data points. In this case, R square is 0.881085687. One way of interpreting this is as *explained variance*. That is, this regression equation predict or *explains* roughly 88% of the variance in the data. Thus, roughly 88% of the Viacom return prediction is explained by the independent (S&P) variable/return.

The “Coefficients” section at the bottom actually gives us the two pieces of information we were searching for: the slope and intercept of the regression line. In this case, the “Intercept” is -0.014084478. The slope, or coefficient of the *independent* variable (S&P return) is 2.105408582. Thus, putting all of this information together, we see that a regression equation of:

$$\text{Viacom Return} = (2.105408582) \times (\text{S\&P Return}) + (-0.014084478)$$

gives us a prediction that explains approximately 88% of the variance in Viacom returns (based on S&P returns). (NOTE: The other 12% of the variance is known as *unexplained variance*. Our goal with regression analysis is to minimize this unexplained variance while not putting too many variables in the model to make it cumbersome or expensive to use. More on these notions in class...)

Great! Now we have a way to predict returns. Note that the regression equation is not perfect--i.e., if we input one of the known returns for the S&P, say 2.4171 for 6/27/98—the resulting predicted Viacom return is, by plugging this into the regression equation, 5.0749. This is not 3.4091 as we see in the dataset. (This is because this data point is not on the regression line.)

Multiple Linear Regression Modeling

But what if we want to predict a (dependent) variable with more than one (independent) variable? For example, what if we turn the problem above around and want to predict S&P returns based on Viacom, AT&T, and GM returns²? That is, what if we want to use *multiple* independent variables to predict another (dependent) variable?

The answer: use the same tool but with more independent variables! In general, we can extend our one-variable model above to include more variables. Generally, speaking, we can create a simple (still linear) multiple regression model of the form:

$$Y = m_1X_1 + m_2X_2 + m_3X_3 + \dots + m_nX_n + b,$$

Where Y is still the predicted/dependent variable, $m_1 \dots m_n$ are the coefficients of the independent variables/observations $X_1 \dots X_n$, and b is the constant or “intercept” of this regression “line.”

² Note that this analysis is inappropriate. We would not consider the S&P benchmark as an independent variable here. However, for teaching purposes, we use this model as an illustration of multiple regression.

Excel tutorial 3

Managerial Statistics
Columbia Executive MBA Program

Excel Tutorial #3: Simple and Multiple Linear Regression Analysis (Continued)...

In the case of this problem, let's assume we want to predict the S&P monthly return based on monthly returns of three stocks: Viacom, AT&T, and GM. Mathematically, we want to create a simple *multiple regression model*:

$$\text{S\&P Return} = m_1 \times (\text{Viacom return}) + m_2 \times (\text{AT\&T return}) + m_3 \times (\text{GM return}) + b$$

Our goal, again, is simple: determine the coefficients or "slopes" m_1 , m_2 , and m_3 , and the constant or "intercept" b . To do this, we follow the same steps as with single-variable regression, but we slightly modify Steps 3, 5, and 6 above (see **bolded** text below for changes):

| Step | Action | Result |
|------|--|--|
| 1, 2 | ... | ... |
| 3 | Now fill in the data ranges. First, fill in the <i>dependent</i> variable range in the "Input Y Range:" field. Here, we are trying to predict S&P returns; thus, the range (and label) of S&P returns should be entered / selected (B3:B9). For the <i>independent</i> variable, enter the range (and label) including the Viacom, AT&T, and GM data (C3:E9) in the "Input X Range:" field | The "Input Y Range:" field should contain the S&P return range (B3:B9) , and the "Input X Range:" field should contain the Viacom, AT&T, and GM return range (C3:E9) |
| 4 | ... | ... |
| 5 | Let's put our output in a new worksheet ply titled, " Multiple Regression." To do this, select the "New Worksheet Ply:" option and enter the worksheet ply name in the field to the right | "New Worksheet Ply:" option is selected with the name " Multiple Regression" entered in the field to the right. (NOTE: no other options or checkboxes should be selected) |
| 6 | ... | Regression results appear in a worksheet ply named, " Multiple Regression" (see next page) |

The results of this regression analysis appear on the following page. Note that from this, we can make two observations. First, by using Viacom, AT&T, and GM returns, we can explain almost 98% of the variance in S&P returns using this simple, linear, multi-variable regression model. This should not be surprising since these three stocks are used to compute the S&P return! However, notice that with just these three stocks and their returns—instead of all 500—you can make a pretty accurate prediction (based on historical data) of the S&P return.

Second, by looking at the coefficients, we see that the regression equation that we can use to predict S&P returns:

$$\text{S\&P Return} = (0.3942 \times \text{Viacom}) + (0.1701 \times \text{AT\&T}) + (0.0913 \times \text{GM}) + 0.1251$$

Notice also from this equation that Viacom has the most "influence" of the three stocks on the S&P return since its coefficient is largest. A large swing in Viacom's return, therefore, may affect the S&P more than a large swing in GM's return.

Excel tutorial 3

Managerial Statistics
Columbia Executive MBA Program

Excel Tutorial #3: Simple and Multiple Linear Regression Analysis (Continued)...

SUMMARY OUTPUT

| <i>Regression Statistics</i> | |
|------------------------------|-------------|
| Multiple R | 0.987732311 |
| R Square | 0.975615119 |
| Adjusted R Square | 0.939037796 |
| Standard Error | 0.821001266 |
| Observations | 6 |

ANOVA

| | <i>df</i> | <i>SS</i> | <i>MS</i> | <i>F</i> | <i>Significance F</i> |
|------------|-----------|-------------|-------------|-------------|-----------------------|
| Regression | 3 | 53.9356009 | 17.97853363 | 26.67267746 | 0.036353424 |
| Residual | 2 | 1.348086157 | 0.674043079 | | |
| Total | 5 | 55.28368705 | | | |

| | <i>Coefficients</i> | <i>Standard Error</i> | <i>t Stat</i> | <i>P-value</i> | <i>Lower 95%</i> | <i>Upper 95%</i> | <i>Lower 95.0%</i> | <i>Upper 95.0%</i> |
|-----------|---------------------|-----------------------|---------------|----------------|------------------|------------------|--------------------|--------------------|
| Intercept | 0.1250621 | 0.44169756 | 0.283139667 | 0.803685895 | -1.775410434 | 2.025534635 | -1.775410434 | 2.025534635 |
| Viacom | 0.394220806 | 0.052563186 | 7.49994124 | 0.017317591 | 0.168059513 | 0.620382098 | 0.168059513 | 0.620382098 |
| AT&T | 0.170135064 | 0.070141633 | 2.425593151 | 0.136110167 | -0.131660234 | 0.471930362 | -0.131660234 | 0.471930362 |
| GM | 0.091267454 | 0.047497875 | 1.921506033 | 0.194617419 | -0.113099551 | 0.295634458 | -0.113099551 | 0.295634458 |

Excel tutorial 3

Managerial Statistics
Columbia Executive MBA Program

Excel Tutorial #3: Simple and Multiple Linear Regression Analysis (Continued)...

Some Caveats

A few caveats should be mentioned regarding the use—and often *abuse*—of linear regression analysis:

- Not all relationships are linear! Some relationships involve variables of higher powers and/or other functions; e.g., $\log x$. This is why it is CRITICAL to PLOT THE DATA FIRST! In this way, you can see if a linear relationship will be helpful or not.
- Not all data are related! If the slope of the (linear) regression equation is near zero, then a linear regression analysis may not be that helpful. Again, GRAPH THE DATA TO SEE WHAT'S GOING ON! (Remember, a line with zero slope does not vary regardless of what happens to the independent variable!)
- Adding more variables doesn't always help! The model with the most variables is not always the best one. In fact, if there are costs associated with collecting additional data/variables, a simpler model may make more sense. Start with simple models and then construct ones with more than one variable only if substantially more variance is explained.
- Variables can be “multicollinear!” Sometimes independent variables in a regression model will depend on each other or will highly (and falsely) “correlate” with each other; e.g., height and weight. This can often lead to erroneous results and conclusions, and can lead you to believe that you have a better (linear) model than is really the case.

In class we will discuss these and more caveats and cautions that you should employ when performing regression analysis.

Appendix: Correlation Analysis

A useful technique to employ when beginning a regression analysis is to analyze the *correlations* among the potential independent variables and the dependent variable. The correlation between two variables is a unitless number between -1 and 1 that indicates the relative relationship of movement of the variables. For example, if variable Y moves in the same direction and amount as variable X , the X and Y are *perfectly correlated*; i.e., their correlation coefficient is 1 . (Think of a line with slope = 1 ; X and Y move together, positively.)

However, if Y moves the same amount as X but in the *opposite* direction, then X and Y are *negatively correlated*; i.e., their correlation coefficient is -1 . (Think of a demand curve, which relates price and quantity; i.e., a line with a slope of negative 1 .) Finally, if the correlation of two variables is zero (or near zero), then a change in one variable is not indicated by a change in the other; that is, there is *no correlation* between them. (Think of the line $y = 4$; any change in x causes no change in y .)

Excel tutorial 3

Managerial Statistics Columbia Executive MBA Program

Excel Tutorial #3: Simple and Multiple Linear Regression Analysis (Continued)...

The reason that correlations among variables is important in the development of regression models is that independent variables that are highly correlated with dependent variables tend to produce better, more predictive regression models. This should make intuitive sense because if a dependent variable, say SALES, is more highly correlated with the independent variable AGE_OF_SALESPERSON then it is with YEARS_WITH_COMPANY, then we would expect to be more accurate in predicting sales using the salesperson's age—as opposed to years of service--as an input. Thus, we would *reduce uncertainty* by using a more highly correlated input/independent variable.³

Excel's Data Analysis tools provides a way for us to calculate correlations among multiple variables. To create a *correlation matrix*, which shows the correlations of one variable with all others, we perform the following steps:

| Step | Action | Result |
|------|--|--|
| 1 | Select the "Data Analysis" option from the "Tools" menu | "Data Analysis" window appears with various options |
| 2 | Select the "Correlation" option and press the OK button | A new window entitled, "Correlation" appears |
| 3 | Now fill in the data ranges. First, fill in the input range in the "Input Range:" field. Here, we want to correlate each variable with all other variables; thus, the entire data range (with labels) should be entered / selected (B3:F9) | The "Input Range:" field should contain the entire data range--all returns (B3:F9) |
| 4 | Since our data are organized by columns and since we have included labels in the ranges entered in Step 3, we need to select the "Columns" option and check the "Labels" checkbox | "Columns" option should be selected, and the "Labels" checkbox should contain a checkmark |
| 5 | Let's put our output in a new worksheet ply titled, "Correlations." To do this, select the "New Worksheet Ply:" option and enter the worksheet ply name in the field to the right | "New Worksheet Ply:" option is selected with the name "Correlations" entered in the field to the right |
| 6 | Now, let's get the results! Press the OK button and observe the results. | A correlation matrix appears in a worksheet ply named, "Correlations" (see next page) |

³ Note, however, that we have to be careful when examining highly correlated independent and dependent variables. A condition known as *multicollinearity* can often give apparently highly predictive/accurate models when, in fact, the variables move together for other reasons (and thus the model is relatively worthless). For example, using both height and weight to predict age may be "overkill;" since height and weight naturally move together (and are highly correlated), using both to predict age will produce a seemingly awesome model when, in fact, these results are misleading. Using both variables to predict age is probably as predictive as just using one.

Excel tutorial 3

Managerial Statistics
Columbia Executive MBA Program

Excel Tutorial #3: Simple and Multiple Linear Regression Analysis (Continued)...

| | <i>S&P</i> | <i>Viacom</i> | <i>AT&T</i> | <i>GM</i> | <i>Coke</i> |
|-----------------|----------------|---------------|-----------------|-----------|-------------|
| <i>S&P</i> | 1 | | | | |
| <i>Viacom</i> | 0.938662 | 1 | | | |
| <i>AT&T</i> | 0.128558 | -0.098933 | 1 | | |
| <i>GM</i> | 0.470349 | 0.350438 | -0.263711 | 1 | |
| <i>Coke</i> | 0.255053 | 0.342337 | -0.50149 | 0.627514 | 1 |

Now, by examining these correlations, we can develop a good first guess as to which independent variable (Viacom, AT&T, GM, or Coke return) might best predict the dependent variable, S&P return. In this case, we see that Viacom returns correlate at the 0.9387 level with S&P returns. (AT&T returns, on the other hand, only correlate with S&P returns at the 0.1286 level.) Thus, Viacom returns might be a good input variable to use to predict the response variable, or S&P returns.

Note also that highly negatively correlated variables also make good predictors; that is, the *magnitude* of the correlation is often more important than the *sign* of it. We will examine correlations more closely in class, but realize that they are often helpful when starting down the path of building regression models. However, you must be aware of underlying data relationships. If data naturally move together—e.g., weight and height—a high correlation between them may falsely lead you to believe that one predicts the other with a high degree of confidence.

Excel tutorial 4

Managerial Statistics
Columbia Executive MBA Program

Excel Tutorial #4: Analysis of Variance (ANOVA)¹

In this tutorial, we discuss how to use the “Data Analysis” add-in in Excel to perform analyses of variance (ANOVA) on multiple sets of data. In order to complete this tutorial, you will need to have the “Data Analysis” add-in in your “Tools” menu. If you do not, please refer to *Excel Tutorial #1: Descriptive Statistics* for help on accessing this toolkit.

In addition, you should have the stock data shown below (from previous tutorials) available for use:

| | A | B | C | D | E | F |
|---|--------------------------------|----------------|---------------|-----------------|-----------|-------------|
| 1 | Monthly Rates of Return | | | | | |
| 2 | | | | | | |
| 3 | Date | S&P | Viacom | AT&T | GM | Coke |
| 4 | 01/30/98 | 0.8799 | 0.7541 | 2.1407 | -4.6296 | -18.8406 |
| 5 | 02/27/98 | 7.5187 | 14.9701 | -2.5948 | 18.9860 | 6.6964 |
| 6 | 03/31/98 | 5.5580 | 11.9792 | 7.7869 | -1.7226 | -3.3473 |
| 7 | 04/30/98 | 1.3716 | 7.9070 | -8.5551 | -0.5535 | 5.8442 |
| 8 | 05/29/98 | -1.6289 | -5.1724 | 1.2474 | 6.6790 | 1.9427 |
| 9 | 06/27/98 | 2.4171 | 3.4091 | 0.8214 | 1.8261 | 2.1063 |

Simple (Single Factor) Analysis of Variance

ANalysis Of VAriance (ANOVA) is simply an extension of the tests of means that we performed in *Excel Tutorial #2: z- and t-tests of Means and of Mean Differences*. Recall that in class and in *Tutorial #2* that we tested the hypothesis that two (sample) means are equal; for example, we tested the (null) hypothesis that the average Viacom return equaled the average GM return:

$$H_0: \text{Average Viacom Return} = \text{Average GM Return}$$

$$H_a: \text{Average Viacom Return} \neq \text{Average GM Return}$$

Here, we assume that the returns are equal but we examine the alternative hypothesis that they are not. Another way of stating these hypotheses is via *mean differences*; that is, we can rearrange the equations above and test the difference of the average Viacom and GM returns against the value 0:

$$H_0: \text{Average Viacom Return} - \text{Average GM Return} = 0$$

$$H_a: \text{Average Viacom Return} - \text{Average GM Return} \neq 0$$

Recall that we can use the Excel mean testing tools to evaluate these hypotheses subject to an alpha (Type I) error level.

¹ The steps and results shown assume a Windows 95 installation. Similar results occur under Windows 97.

Excel tutorial 4

Managerial Statistics Columbia Executive MBA Program

Excel Tutorial #4: Analysis of Variance (ANOVA) (Continued)...

But what if we want to test/compare several means at once? For example, what if we want to test the (null) hypothesis that all company returns are, on average, equal. That is:

$$H_0: \text{Avg. Viacom Return} = \text{Avg. AT\&T Return} = \text{Avg. GM Return} = \text{Avg. Coke Return}$$
$$H_a: \text{A least one avg. stock return differs from others; e.g., AT\&T Return} \neq \text{GM Return}$$

Notice, however, that if we end up rejecting the null hypothesis, we will only know that (at least) one average stock return is different; we *will not* be able to determine which one it is statistically!

The Excel/Data Analysis ANOVA tools allow us to test such hypotheses; that is, that means from two or more samples are equal. In the simple or “single factor” case, we assume further that the samples are drawn from populations with the same mean. **THIS IS A CRITICAL ASSUMPTION; IF THIS IS NOT TRUE, THEN THE SIMPLE/SINGLE-FACTOR ANOVA TOOL CANNOT BE USED!** We’ll make this assumption for the purposes of teaching this tool with the data above.

To perform a simple, single-factor ANOVA test such as the one above, perform the following steps:

| Step | Action | Result |
|------|--|---|
| 1 | Select the “Data Analysis” option from the “Tools” menu | “Data Analysis” window appears with various options |
| 2 | Select the “Anova: Single Factor” option and press the OK button | A new window entitled, “Anova: Single Factor” appears |
| 3 | Now fill in the data ranges. First, fill in the “Input Range” with the range of all sample data that you want to test. In this case, since we are testing four average stock returns, we’ll select the range C3:F9. Next, since these data are grouped by columns, select the “Grouped by Columns” option. Finally, enter an “Alpha” level of 0.05 (NOT 5!) | The “Input Range” and the “Alpha” fields should be filled in with the appropriate values. Also, the “Columns” option should be selected. |
| 4 | Since we have included labels in the ranges entered in Step 3, we need to check the “Labels” checkbox | “Labels” checkbox should contain a checkmark |
| 5 | Let’s put our output in a new worksheet ply titled, “ANOVA (Single Factor)” | “New Worksheet Ply:” option is selected with the name “ANOVA (Single Factor)” entered |
| 6 | Now, let’s get the results! Press the OK button and watch Excel work its magic. | The ANOVA results appear in a worksheet ply named, “ANOVA (Single Factor)” (see next page) |

Excel tutorial 4

Managerial Statistics Columbia Executive MBA Program

Excel Tutorial #4: Analysis of Variance (ANOVA) (Continued)...

Anova: Single Factor

SUMMARY

| Groups | Count | Sum | Average | Variance |
|--------|-------|---------|-------------|-------------|
| Viacom | 6 | 33.8471 | 5.641183333 | 55.62648613 |
| AT&T | 6 | 0.8465 | 0.141083333 | 29.45048177 |
| GM | 6 | 20.5854 | 3.4309 | 72.4984677 |
| Coke | 6 | -5.5983 | -0.93305 | 89.63123228 |

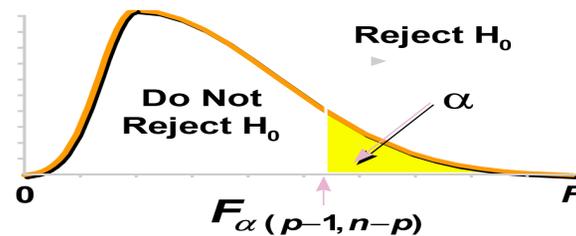
ANOVA

| Source of Variation | SS | df | MS | F | P-value | F crit |
|---------------------|-------------|----|-------------|-------------|-------------|-------------|
| Between Groups | 164.0665681 | 3 | 54.68885603 | 0.884909076 | 0.465765627 | 3.098392654 |
| Within Groups | 1236.033339 | 20 | 61.80166697 | | | |
| Total | 1400.099907 | 23 | | | | |

Note that the top part of the output provides us with some summary statistics of the individual stock return data; i.e., number of data items (count), sum, average, and variance. The bottom part, labeled “ANOVA,” contains the analysis of variance both between and within these groups of data.

Remember, we are testing a hypothesis here—the hypothesis that all the mean stock returns from the four stocks are the same. Thus, we need to examine a “critical value” and determine whether our “test statistic” falls in the “rejection region.” If the test statistic does fall in the critical region, we would reject the null hypothesis that the mean returns for the four stocks are the same. If not, we will reserve judgment and not reject the null hypothesis.

But what test statistic, critical value, and rejection region are we talking about here? Notice that unlike the two-mean tests in *Tutorial #2*, no z- or t-statistics are shown here nor are any probability statements. The key in this test is the use of a new statistic—one that is helpful when comparing variances of data sets: the F statistic. Like the t- and the z-statistics, the F statistic has a distribution that can be seen in your textbook. Like the t-statistic, degrees of freedom is a parameter that drives selection of F-statistics as is the alpha tolerance level and the sample size:



Excel tutorial 4

Managerial Statistics
Columbia Executive MBA Program

Excel Tutorial #4: Analysis of Variance (ANOVA) (Continued)...

However, the same hypothesis testing mentality and approach apply here. But before we move on, a word about sources of variation.

Sources of Variation

When comparing means of multiple data sets, variation in these means can come from two sources as shown in the ANOVA output earlier: “between groups” and “within groups.” That is, variation of mean stock returns can either be caused by sampling variability—“between groups” errors caused by sampling Coke, AT&T, etc. stock returns—or by errors “within groups” caused by biases, variances, etc. within a specific stock’s return data; e.g., GM.

If means of multiple data sets are equal, we would expect low “between groups” errors compared to “within groups;” that is, we would expect most variation to be explained by the data themselves and not by the sampling performed. However, if the “between groups” errors are large compared to “within groups,” we might be led to conclude that means are different since sampling causes more errors than the data themselves. A good pictorial description of these sources of variation can be found in the McClave text.

The F-statistic helps us examine these two sources of variation by looking at a ratio of them. Specifically, the F-statistic is the ratio of “between groups” mean square errors and “within groups” mean square errors. If this ratio is near one, then we can conclude that the sources of variation “balance each other out,” and we would NOT reject our null hypothesis. However, large test F-statistics would lead us to conclude that the variation “between groups” is large compared to variation “within groups;” this would lead us to conclude that (at least) one sample mean differs from the others. We’ll discuss this more in class. Now, back to the results.

Interpreting the Excel Output

Now that we understand how the F-statistic works, we can apply our tried-and-true hypothesis testing techniques to see if we should accept or reject the null hypothesis in this case.

First, look at the three columns labeled “F,” “P-value,” and “F crit” in the “Between Groups” variation line. These three pieces of information give us all we need to know to test the hypothesis mentioned earlier. Here, “F” is the F-test statistic; that is, the ratio of the variations mentioned earlier that come from the samples we have. “F crit,” on the other hand, is the critical value—which defines the rejection region—determined by the degrees of freedom and the alpha level specified (5% in this case).

As usual and as seen on the chart on the previous page, we need to determine whether the “F” value is to the left or right of the “F crit” value. If “F” is greater than “F crit,” then the test value lies in the rejection region → reject the null hypothesis. Alternatively, we can also test this hypothesis by using P-values. If the “P-value” given in the results is less than the alpha level specified, we know the test statistic lies in the critical region → reject the null.

Excel tutorial 4

Managerial Statistics
Columbia Executive MBA Program

Excel Tutorial #4: Analysis of Variance (ANOVA) (Continued)...

As shown in the results, the “F” value—the test statistic—of 0.885 is much less than the “F crit” value of 3.098. Thus, we would NOT reject the null hypothesis and conclude that all of the mean returns are equivalent. Alternatively, we can see that the “P-value” of the F-test statistic is 0.466, much greater than our alpha tolerance of 0.05. Thus, again, we are led to not reject the null hypothesis.

Final Thoughts

A few final thoughts about using the simple, single-factor ANOVA tool:

- In general, analysis of variance (ANOVA) tools only apply if three conditions are met:
 - The probability distributions of the *populations* of all samples must be normal
 - The probability distributions of the *populations* of all samples must have equal variances
 - The *samples* must be random and independent
- The single-factor ANOVA tool can only be used if we assume in addition that the *population* means of all sample data are equivalent
- The ANOVA tool tests the hypothesis that all means are equal or alternatively, that at least one mean is different. However, the ANOVA tool *does not* tell us which mean is different (if the null hypothesis is rejected)
- *Post hoc* analysis can be performed to determine which mean(s) is/are different. Methods include Bonferroni, Scheffe, and Tukey
- Finally, if you perform a single-factor ANOVA test on just two samples (means), you will get the same results as if you had performed the t-test of means assuming equal variances. Try it! Thus, when comparing two means with unknown population variances, you can use the single-factor ANOVA tool, instead

Excel tutorial 5

Managerial Statistics Columbia Executive MBA Program

Excel Tutorial #5: The Chi-Squared Distribution and Tests for Independence¹

In this tutorial, we discuss how to use the three Chi-Squared Distribution worksheet functions in Excel to test observed vs. expected values for *preference* or *independence*. NOTE: The functions discussed in this tutorial are NOT available in the Data Analysis Toolkit; they must be used directly in spreadsheet cells through function notation; e.g., “=AVERAGE(A4:A9).”

A Test for Independence (vs. Dependence or Preference)

Thus far, we have learned to test hypotheses of sample data and how statistics of sample data relate to each other. For example, we can test hypotheses for sample means and for differences of means; we can test one or two at a time (simple z- and t-tests) or multiple means (ANOVA). Also, we have learned how to use sample data to estimate or to predict results via regression modeling. But hypothesis testing plays a part in regression modeling, too, either at the model level (ANOVA/F-test for ratio of variances) or at the individual coefficient level (t-test for significance to the model).

We can also use correlation analysis (and other descriptive statistics) to see how data “move” with each other. Correlation analysis plays a large role in regression modeling as we want highly correlated variables as independent ones in our model because they (usually) explain a lot of variance for us. This leads to higher coefficients of determination, and thus, better model “fits.”

But what about testing data for independence? Remember (long, long ago) when we studied probability theory and we discussed the notion of independence? If a conditional probability, say $P(B|A)$ equals the simple, marginal probability, $P(B)$, then we say that events A and B are *independent*; that is, knowing that A occurred doesn't help us with respect to the probability of B. But can we test this with “real world” data...especially when we only have sample data?

Yes! To show this, let's look at some tabular data. (NOTE: You may want to enter this data into Excel as we'll be using it to test some hypotheses later.):

| | A | B | C | D | E | F |
|---|--------------------------|-------------|-------------|-------------|------------|--------|
| 1 | Survey Results (ACTUALS) | | | | | |
| 2 | CLIENT'S INCOME | | | | | |
| 3 | | | Under \$30K | \$30K-\$60K | Over \$60K | TOTALS |
| 4 | BROKER'S RATING | Outstanding | 48 | 64 | 41 | 153 |
| 5 | | Average | 98 | 120 | 50 | 268 |
| 6 | | Poor | 30 | 33 | 16 | 79 |
| 7 | TOTALS | | 176 | 217 | 107 | 500 |

Here, we show a broker's “satisfaction” rating based on income levels of the people the broker serves². Now, we can generate all kinds of probabilities with this table; e.g., the probability that the broker gets an outstanding rating is 153/500. The (conditional) probability that the broker

¹ The steps and results shown assume a Windows 95 installation. Similar results occur under Windows 97.

² A table showing data in this way is called a *contingency table*.

Excel tutorial 5

Managerial Statistics
Columbia Executive MBA Program

Excel Tutorial #5: The Chi-Squared Distribution and Tests for Independence (Continued)...

gets an outstanding rating given that the person surveyed makes over \$60K is 41/107. Now, it is easy enough to test conditional vs. marginal probabilities to determine if the events “client’s income” and “broker’s rating” are independent—remember that midterm problem??--but how can we apply hypothesis testing theory to this problem?

The Answer: The Chi-Squared (χ^2) Distribution

The Chi-Squared distribution is a special type of probability distribution that shows how repeated samples are distributed³. For example, if we repeatedly sample data from a population, we would like to know if we are either (1) getting completely “random” (and thus independent) data or (2) getting dependent or preferential data patterns (and thus dependent). In terms of hypothesis testing, we are assuming (null case) that our results or *observed values* are exactly what we expect. Thus, our alternative view is that this is not the case; that is, that some preference or data dependence exists:

H₀: “Observed values” are as “expected;” that is, no preference/dependence is evident
H_a: Preferences/data dependence exists; that is, observed values differ from expected

For example, using the data rating/income data above, we would like to test the (null) assumption that a broker’s rating *does NOT depend* on the survey respondent’s income. But, someone may claim that the more a client makes, the better ratings he/she gives his/her broker. (NOTE: this would be the alternative hypothesis; i.e., that rating *depends on* client income.)

But how do we test such hypotheses? We have the “actuals” given to us—the actual survey results. But how do we know what the “expected” values are? For example, we know that 48 people (almost 10% of the 500 people surveyed) had incomes of less than \$30K but thought their broker was “outstanding.” What should we have expected this number to be?

The answer—look at the contingency table. Let’s look at the <\$30K - Outstanding intersection. How many people would we have *expected* to answer in this way. From the totals and from basic probability theory, we see that the marginal probability of a respondent making less than \$30K and replying “outstanding” (assuming H₀: independence) is: $(176*153)/500 = 53.856$. That is, it’s the (row total * column total) / sample size.

³ The Chi-Squared distribution comes for a special class of probability distributions—Gamma distributions—which we will not cover here. This distribution shows how the H-statistic, the Kruskal-Wallis test for completely randomized data, is distributed (see Chapter 15). The shape of the distribution depends on the degrees of freedom (see textbook pages 911-913) Bottom line: this distribution helps us test whether data are “random” (and thus independent) or follow some pattern or preference (and thus dependent).

Excel tutorial 5

Managerial Statistics
Columbia Executive MBA Program

Excel Tutorial #5: The Chi-Squared Distribution and Tests for Independence (Continued)...

Thus, we would have *expected* almost 54 people to respond in this way—assuming the two events of income and rating are independent—but only 48 people *actually* did. If we compute the same expected values for the rest of the cells, we get the table on the next page (NOTE: please create this table/output below the survey data for use later, too):

| | A | B | C | D | E | F |
|----|---|--------------------|--------------------|--------------------|-------------------|---------------|
| 11 | Hypothesized/Derived Values (EXPECTED) | | | | | |
| 12 | CLIENT'S INCOME | | | | | |
| 13 | | | Under \$30K | \$30K-\$60K | Over \$60K | TOTALS |
| 14 | BROKER'S | Outstanding | 53.856 | 66.402 | 32.742 | 153 |
| 15 | RATING | Average | 94.336 | 116.312 | 57.352 | 268 |
| 16 | | Poor | 27.808 | 34.286 | 16.906 | 79 |
| 17 | | TOTALS | 176 | 217 | 107 | 500 |

But what explains these *expected vs. actual* differences? Simple sampling error or something else...data dependence? Is it possible that the events income and rating are dependent and cause our expected values (under the null hypothesis of independence) to differ significantly from the actual/surveyed values? The chi-squared distribution will help us test our hypotheses.

Using Excel to Perform Chi-Squared Tests

Three statistical functions in Excel can help us with this task. Although these functions are not available in the Data Analysis Toolkit, we can still use them directly in the spreadsheet. The three functions that we will use are:

- CHITEST(actual_range, expected_range)
- CHIINV(probability, degrees_freedom)
- CHIDIST(x, degrees_freedom)

Recall that in order to use functions in Excel cells, you must put an “=” sign in front of the function; e.g., =SUM(cell_range). Alternatively, you can use the “ f_x ” button on the toolbar, which is the function wizard. Also note that each function takes certain parameters. We will now discuss how to load the parameters into the function and how to interpret the results.

Typically, when using the Chi-Squared distribution/test, only the first function above is used, CHITEST. CHITEST uses the actual and expected values and returns a p-value; that is, the value to the right of the test statistic. You can then compare this p-value to your prescribed alpha level to determine whether the null hypothesis should be rejected. If the p-value is less than alpha, then the test statistic must lie to the right of the critical value → reject the null. However, if the p-value is greater than alpha, then the test statistic must lie to the left of the critical value → do NOT reject the null. NOTE: Chi-Squared tests are always one-tailed.

Excel tutorial 5

Managerial Statistics Columbia Executive MBA Program

Excel Tutorial #5: The Chi-Squared Distribution and Tests for Independence (Continued)...

Now, back to our data. To test our hypotheses—remember our null assumption is that the events are independent; our alternative is that they are not—we now use the CHITEST function to evaluate our data. Using the function wizard—that f_x button on the menu bar—we can apply the CHITEST function to test the actual vs. expected values in the two tables above. First, type the word “CHITEST” into cell A20. Next, put the cursor/mouse in cell B20. Now, click on the wizard button, choose “Statistical” functions on the left and then “CHITEST” on the right. Push the “Next” button to move to the next dialog box.

Now, you are prompted to enter/select both the actual_range and expected_range; i.e., the actual and expected values. According to the row/column labels above, we want range C4:E6 in the actual_range field and range C14:E16 in the expected_range field. Note that after entering/selecting these cells, the resulting value appears in the upper right-hand corner of the window. However, push the “Finish” button. The value 0.369725177 appears in cell B20.

Now, we have the p-value for the test statistic. If we were testing our hypotheses with an alpha-level of 0.05 (95% level), we would *not reject* the null hypothesis. Since the p-value is greater than alpha, we can infer that the test statistic is to the left of the critical value → do NOT reject the null. Thus, the a respondent’s income does NOT appear to influence/affect his feelings about his broker’s performance. That is, based on this sample and confidence level, the events “income” and “rating” appear to be independent. There is no “preference” at higher income levels to be more favorable to the broker’s rating.

However, if our alpha level had been 40% (only at the 60% level), we would *reject* the null. Here, the p-value is less than alpha → test statistic is greater than the critical value → reject the null. In this case, we would say (at a fairly weak confidence level) that there is a preference/data dependence here. That is, the events “income” and “rating” appear to be dependent (our alternative hypothesis). However, an alpha of 40% is very, very weak. We point this case out here just for comparison purposes.

Thus, it appears that our data are not dependent and that knowing a client’s income does not help us guess/determine his/her rating his/her broker’s performance. Note that this test can be applied in a number of situations, now. For example, if you believe that people will prefer one (or two, or three) products equally, you can test your hypothesis by conducting some surveys and running a CHITEST. For example, if you think half of your customers will prefer product A, 25% will prefer product B, and the remaining 25% will prefer product C, you can test your assumptions in this way (as long as you design the experiment properly)⁴.

⁴ Note that the validity of a chi-squared test depends primarily on the experiment being multinomial (an extension of the binomial experiments we discussed earlier in the term). That is, we assume independent trials, mutual exclusivity of outcomes, constant probabilities throughout the experiment, etc. If an experiment is not set up in this way, a chi-squared test will likely be inappropriate.

Excel tutorial 5

Managerial Statistics Columbia Executive MBA Program

Excel Tutorial #5: The Chi-Squared Distribution and Tests for Independence (Continued)...

But what do the other two functions tell us? The CHIINV function can be used to compute the actual test-statistic from the p-value given by CHITEST; that is, it gives us the inverse of the CHITEST (or CHIDIST) function (thus the name, CHIINV). You give the function a probability (p-value), CHIINV returns a point or (test-) statistic. You can then compare the test statistic to the critical value (from a table, book, or from Excel) to make a decision regarding your hypotheses. To see how this works, put the word “CHIINV-Test” in cell A21 and move your mouse/cursor to cell B21. Now, in order to use the CHIINV (and the CHIDIST) function, we need to know the degrees of freedom. In this case, degrees of freedom is given by:

$$df = (\text{number of rows} - 1) \times (\text{number of columns} - 1)$$

In our case, we have three columns of data and three rows; thus, the degrees of freedom is $(3 - 1) \times (3 - 1) = 2 \times 2 = 4$. Now, let's use the Excel function wizard to help us. With the cursor in cell B21, push the wizard button, choose statistical functions, select the CHIINV function, and push the “Next” button.

Now, the probability that we want to supply is the result of the CHITEST in cell B20. So, put or select the cell B20 in the probability field. Finally, put 4 in the degrees of freedom field, and press the “Finish” button. The test statistic value of 4.277705068 appears in cell B21. This is the test statistic along the x-axis of the distribution. We know from CHITEST that the area to the right of this point is approximately 0.3697. But what do we compare this value from CHIINV (4.2777) to in order to evaluate our hypothesis test?

We can use CHIINV to help us here, too. By using our alpha level as our probability, we can determine where the critical value is. Put the word “CHIINV-Crit” in cell A22. Move the cursor to cell B22, apply the function wizard, pick the CHIINV function, and use the alpha level of 0.05 as the probability/p-value. We get the result 9.487728465. Now we can compare the CHIINV-Test to CHIINV-Crit. Note that the test statistic (4.2777) is less than the critical value (9.4877); thus, we fail to reject the null hypothesis; that is, the events/data appear to be independent. By using CHIINV, we can compare test and critical values instead of p- and alpha values.

Finally, the CHIDIST function can be used to find areas to the right of points along the x-axis. This function is usually not used in hypothesis testing, but it is nice to know that we can get areas to the right of points (i.e., p-values) using this function. For example, if you want to know the p-value for the point 4.5, use the CHIDIST formula (or wizard) to retrieve it. In this case, with four degrees of freedom, =CHIDIST(4.5,4) returns the p-value of 0.34254748.

For More Information...

For more reading on the Chi-Squared distribution and tests for independence, please refer to the course text. The following page shows the Excel worksheet used for the analysis. All output described in this tutorial is also shown.

Excel tutorial 5

Managerial Statistics
Columbia Executive MBA Program

Excel Tutorial #5: The Chi-Squared Distribution and Tests for Independence (Continued)...

| | A | B | C | D | E | F |
|----|---|--------------------|---|--------------------|-------------------|---------------|
| 1 | Survey Results (ACTUALS) | | | | | |
| 2 | CLIENT'S INCOME | | | | | |
| 3 | | | Under \$30K | \$30K-\$60K | Over \$60K | TOTALS |
| 4 | BROKER'S | Outstanding | 48 | 64 | 41 | 153 |
| 5 | RATING | Average | 98 | 120 | 50 | 268 |
| 6 | | Poor | 30 | 33 | 16 | 79 |
| 7 | | TOTALS | 176 | 217 | 107 | 500 |
| 8 | | | | | | |
| 9 | | | | | | |
| 10 | | | | | | |
| 11 | Hypothesized/Derived Values (EXPECTED) | | | | | |
| 12 | CLIENT'S INCOME | | | | | |
| 13 | | | Under \$30K | \$30K-\$60K | Over \$60K | TOTALS |
| 14 | BROKER'S | Outstanding | 53.856 | 66.402 | 32.742 | 153 |
| 15 | RATING | Average | 94.336 | 116.312 | 57.352 | 268 |
| 16 | | Poor | 27.808 | 34.286 | 16.906 | 79 |
| 17 | | TOTALS | 176 | 217 | 107 | 500 |
| 18 | | | | | | |
| 19 | Function | Result | Parameters | | | |
| 20 | CHITEST | 0.369725177 | actual_range = C4:E6; expected_range = C14:E16 | | | |
| 21 | CHIINV-Test | 4.277705068 | probability = B20; degrees_freedom = B25 = 4 | | | |
| 22 | CHIDIST-Crit | 9.487728465 | probability=alpha=B24 = 0.05; degrees_freedom=B25=4 | | | |
| 23 | | | | | | |
| 24 | Alpha | 0.05 | Given | | | |
| 25 | df | 4 | $(r - 1) * (c - 1)$ | | | |