TESTING MEANS AND DIFFERENCES IN MEANS
EXCEL LAB #6

BUSN/ECON/FIN 130: Applied Statistics
Department of Economics and Business
Lake Forest College
Lake Forest, IL 60045
Copyright, 2013

Overview

This lab is written for Excel 2010, which is available to students in the library. The notation => can be read as "go to" or "click on." This notation will most often be used when navigating the menu or toolbars in Excel. To indicate a command or icon that you might click on or search for in Excel, **bold** will be used. Likewise, anything that you are to type into Excel will be **bolded** in the instructions. Do not enter such text as bolded text unless the instructions ask you to do so.

Tutorial

Start Excel and open **2007 MLB Attendance Data.xlsx**. This file contains the attendance numbers at all home games during the 2007 MLB season for the St. Louis Cardinals, Chicago Cubs, Milwaukee Brewers, Chicago White Sox, Cleveland Indians, and Minnesota Twins.

As each team plays 81 home games each year, there should be 81 attendance figures for each team. Scroll to the bottom of the data, however, and you see that there are 81 (row 82) attendance figures for only five of the six teams. The Cleveland Indians only have 77 (row 78) attendance figures. This is because in April of 2007 Cleveland experienced a week of snow storms. This caused Cleveland to cancel one game against the Seattle Mariners that was actually made up in Seattle and to play three games against the Los Angeles Angels in Milwaukee. Therefore, Cleveland only had 77 "home games" that were played in Cleveland in 2007.

1. Insert two blank columns to the left of the Cardinals data (hint: **left click on column A and drag through column B** => **right click in the shaded area** => **Insert**).

2. Set columns A through L to Arial 10 point for for columns A through L. Right justify columns A through K. Left justify column L.

3. Set the width of column A to 20. Set the width of columns B, E, H, and K to 3, the width of columns C, D, F, G, I, J, M, and N to 10, and the width of column L to 25.

4. Insert blank rows at the top of the worksheet so that the data begin in row 20 (hint: **left click on row 2 and drag through row 19** => **right click in the shaded region** => **Insert**). The data for each team should now begin in row 20 (and end in row 100, except for the Indians which ends at row 96).

5. In cell A2 enter **Average (xbar)**. In cell A3 enter **St. Dev (s)**. In cell A4 enter **N**. In cell A5 enter **μ0**. This is the Greek letter mu that can be found under **Symbols**.

6. Next we will enter formulas to calculate these four objects for each team.

    1. In cell C2 enter **=AVERAGE(C20:C100)**.
    2. In cell C3 enter **=STDEV(C20:C100)**.
    3. In cell C4 enter **=COUNT(C20:C100)**.

    Have all three cells report numbers with no decimals and with a comma separating hundreds and thousands.

7. Copy these three formula cells to the other columns of data.

8. The average attendance for the Cardinals is 43,854 and for the Cubs it is 40,154. In cells C5 and D5 enter **40,000**. Similarly, average attendance for the Brewers is 35,421 and for the White Sox is 33,141. In cells F5 and G5 enter **35,000**. Average attendance for the Indians is 28,880 and for the Twins is 28,227. In cells I5 and J5 enter **28,000**.

9. In cell A7 center **Critical Value (95%)**. In cell A8 enter **Margin of Error**. In cell A9 enter **95% CI Lower Bound**. In cell A10 enter **95% CI Upper Bound**.

10. In cell C7 we want to enter the critical value associated with a 95% confidence level, using a large sample (N = 81 is large enough for us). We already know from class that the critical value is 1.96, but we should let Excel calculate it for us. As all of our tests are two-sided, the 95% confidence level puts 2.5% of the probability above the critical value (and 2.5% will be below the negative critical value). Therefore, in cell C7 enter **=NORMSINV(0.975)**. Make cell C7 report to 3 decimal places. Cell C7 now reports **1.960** as the critical value.

11. In cell C8 we want to enter the margin of error, which equals the critical value * the standard deviation of the sample ÷ the square root of the sample size. That is:

    $$\text{Margin of Error} = z_{\alpha/2} \cdot \frac{s_x}{\sqrt{n}} .$$

    Thus, in cell C8 enter **=C7*C3/sqrt(C4)**. As attendance numbers are large, we don’t need decimal points for the margin of error. Therefore, have cell C8 report no decimal points, but include a comma before the thousands digit if it exists.

12. In cells C9 and C10 enter the lower and upper bound on the 95% confidence interval respectively. As two-sided confidence intervals add and subtract the margin of error from the average value (x-bar), in cell C9 enter **=C2-C8** and in cell c10 enter **=C2+C8**. Change the number format so cells C9 and C10 report no decimals but include commas. You should have that the 95% confidence interval for the Cardinals is from 43,563 up to 44,146.

13. Copy and paste cells C7 and C10 to the appropriate cells for the other five teams.

14. In cell A12 enter **t-statistic** and in cell A13 enter **p-value**.

15. The test we have in mind is testing whether the team's population mean equals the proposed mean, μ_0. In this case, the test statistic is

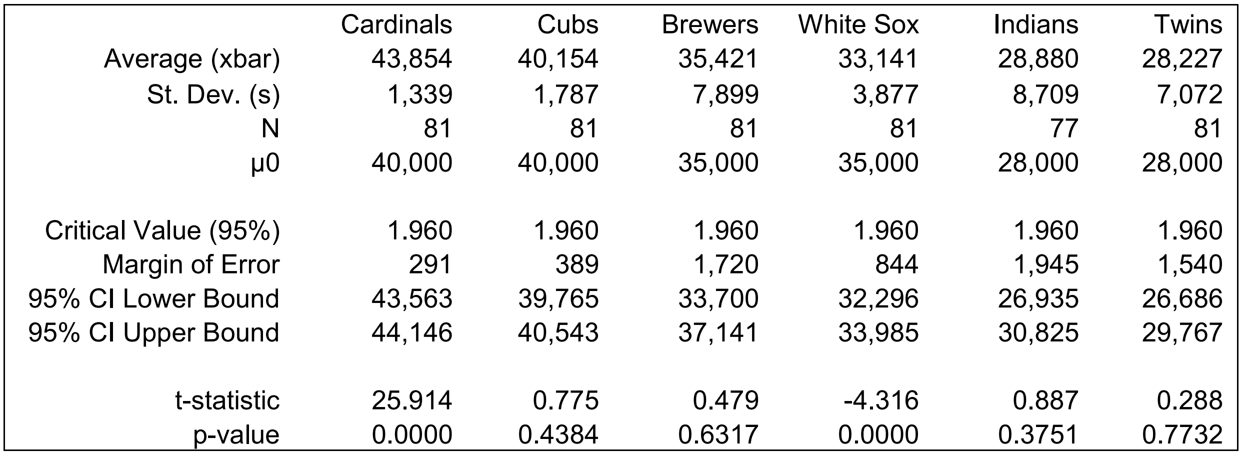
$$\text{t-stat} = \frac{\overline{x} - \mu_0}{\left(s_x \big/ \sqrt{n} \right)}.$$

Thus, in cell C12 enter **=(C2-C5)/(C3/sqrt(C4))**. As test statistics are usually reported to three decimal places, change the number format to reflect this. The t-statistic for the Cardinals is 25.914.

16. As we are conducting a two-sided test with a large sample, the *p*-value is twice the probability that lies above the t-statistic. Excel's **NORMSDIST()** function returns the probability from a N(0,1) distribution that is below the value in parentheses. Thus, **1-NORMSDIST()** will return the probability above the value in parentheses, which then must be doubled. Lastly, note that for this formula to work for all test statistics, the test statistic must be positive, and therefore we also need to use Excel's absolute value function **ABS()**. Therefore, enter **=2*(1-NORMSDIST(ABS(C12)))** into cell C13. As *p*-values are usually reported to four decimal places, change the number format to reflect this. The *p*-value for the Cardinals is 0.0000 (as it should be since the t-stat of over 25 is huge!).

17. Copy and paste cells C12 and C13 to the corresponding cells for the other five teams.

18. Your worksheet should now look like the following.



	Cardinals	Cubs	Brewers	White Sox	Indians	Twins
Average (xbar)	43,854	40,154	35,421	33,141	28,880	28,227
St. Dev. (s)	1,339	1,787	7,899	3,877	8,709	7,072
N	81	81	81	81	77	81
µ0	40,000	40,000	35,000	35,000	28,000	28,000
Critical Value (95%)	1.960	1.960	1.960	1.960	1.960	1.960
Margin of Error	291	389	1,720	844	1,945	1,540
95% CI Lower Bound	43,563	39,765	33,700	32,296	26,935	26,686
95% CI Upper Bound	44,146	40,543	37,141	33,985	30,825	29,767
t-statistic	25.914	0.775	0.479	-4.316	0.887	0.288
p-value	0.0000	0.4384	0.6317	0.0000	0.3751	0.7732

19. The last part of the tutorial will test whether the sample means are statistically different across populations. In cell A15 enter **Test Across Pops**. In cell A16 enter **s_x1_x2**. In cell A17 enter **t-statistic**. In cell A18 enter **p-value**.

20. For a data set that includes two unmatched samples (as these data are), the sample standard deviation of the difference in means is

$$s_{\bar{x}_1-\bar{x}_2} = \sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}} .$$

Therefore, in cell C16 enter **=SQRT((C3^2/C4)+(D3^2/D4))**. This produces the sample standard deviation of the difference in attendance numbers between the Cardinals and the Cubs. Change the number format of cell C16 to not report decimals.

21. You should have that the sample standard deviation for testing Cardinals' attendance against the Cubs' attendance is 248. Copy and paste cell C16 to cell F16 to compare the Brewers to the White Sox and to cell I16 to compare the Indians to the Twins.

22. To calculate the **t-statistic** associated with testing that there is no difference in means, recall that:

$$\text{test statistic} = \frac{(\bar{x}_1 - \bar{x}_2) - d}{s_{\bar{x}_1-\bar{x}_2}},$$

where *d* = 0 in this case. Thus, in cell C17, enter **=(C2-D2)/C16**. Change cell C17 to report to three decimal places. The t-statistic testing that the average Cardinals attendance equals the average Cubs attendance is 14.916. Copy and paste cell C17 to cells F17 and I17.

23. As before, we are conducting a two-sided test with large samples. In cell C18 enter **=2*(1-NORMSDIST(ABS(C17)))**. Change the number format of cell C18 to have four decimal places.

The p-value for the test that the Cardinals' average attendance equals the Cubs' average attendance is 0.0000 (as it should be since the t-stat is almost 15). Thus, we reject at all standard significance levels the claim that average attendance at Cardinals' games equals average attendance at Cubs' games.

24. Copy and paste cell C18 to cells F18 and I18.

Notice the claim that the Brewers' average attendance equals the White Sox average attendance is associated with a *p*-value of 1.97%. At the 1% significance level, therefore, we would fail to reject there is a difference in average attendance numbers, but at the 2% significance level (or higher) we would reject the hypothesis that there is no difference in average attendance. There is no statistical evidence, however, that the Indians and Twins have different average attendance numbers, as the *p*-value associated with that difference is 60.55%.

25. ***t-Tests Using the Data Analysis Toolpak***.

It is worth knowing how to have Excel conduct these tests for you. The only difference with having Excel do the work is that Excel will use the appropriate *t* distribution according to both sample sizes rather than assuming sample sizes of 81 for both variables (or even 77 for Cleveland's data) is large enough to use a N(0,1). Thus, Excel's results will produce exactly the same test statistics but slightly different (always slightly higher) and statistically more precise *p*-values.

1. In cell L1 enter **Cardinals vs. Cubs**.
2. Load Excel's Data Analysis ToolPak: **File** tab => **Options** => **Ad Ins** => **highlight Data Analysis ToolPak VBA** => **Go** => **check Data Analysis ToolPak** => **OK**.
3. To use the Data Analysis ToolPak to test the difference in means: **Data** tab => **Analysis** box => **Data Analysis** => **t-test: Two Sample Assuming Unequal Variances** => **OK** => **C20:C100** (Variable 1 range) => **D20:D100** (Variable 2 range) => **0** (Hypothesized Mean Difference) => **check Output Range** => **L2** (Output range) => **OK** (be sure that **labels** in not checked before clicking OK).

Cells M5 and N5 report the average attendance for the Cardinals and Cubs respectively (which are identical to cells C2 and D2). More importantly for our use, cell M10 reports the test statistic associated with the test, **14.916**, which equals what we had previously found in cell C17. The *p*-value of the test is located in cell M13, which is essentially zero.

4. In cell L17 enter **Brewers vs. White Sox**.
5. Use Excel's Data Analysis ToolPak to test the null hypothesis that there is no difference in average attendance figures between the Brewers and the White Sox: **Data** tab => **Analysis** box => **Data Analysis** => **t-test: Two Sample Assuming Unequal Variances** => **OK** => **F20:F100** (Variable 1 range) => **G20:G100** (Variable 2 range) => **0** (Hypothesized Mean Difference) => **check Output Range** => **L18** (Output range) => **OK** (be sure that **labels** in not checked before clicking OK).

Cells M21 and N21 report the average attendance for the Brewers and White Sox respectively, (which are identical to cells F2 and G2). More important for our use, cell M26 reports the test statistic associated with the test, **2.332**, which equals what we had previously found in cell F17. The *p*-value of the test is located in cell M29, which is 0.0214. This *p*-value, which uses a *t*-distribution rather than a N(0,1), is only slightly higher than the *p*-value of 0.0197 we calculated earlier in cell F18, and it is statistically more precise.

6. In cell L33 enter **Indians vs. Twins**.
7. Use Excel's Data Analysis ToolPak to test the null hypothesis that there is no difference in average attendance figures between the Indians and the Twins: **Data** tab => **Analysis** box => **Data Analysis** => **t-test: Two Sample Assuming Unequal Variances** => **OK** => **I20:I100** (Variable 1 range) => **J20:J100** (Variable 2 range) => **0** (Hypothesized Mean Difference) => **check Output Range** => **L34** (Output range) => **OK** (be sure that **labels** in not checked before clicking OK).

Cells M37 and N37 report the average attendance for the Indians and Twins respectively, equaling cells I2 and J2. More important for our use, cell M42 reports the test statistic associated with the test, **0.516**, which equals what we had previously found in cell I17. The *p*-value of the test is located in cell M45, which is 0.6063. This *p*-value, which uses a *t*-distribution rather than a N(0,1), is only slightly higher than the *p*-value of 0.6055 we calculated earlier in cell I18, and it is statistically more precise.

26. Save this file YourName_Lab6_Tutorial.xls.

Exercises

1. **Male-Female Wages and SAT Scores.xlsx** contains annual income figures and SAT scores for 103 randomly chosen men and 89 randomly chosen women. Use Excel formulas to calculate the average annual income and average SAT scores for both men and women, the sample standard deviation for all four variables, and the sample sizes of each.

2. Under the proposal that average annual income for both sexes is $80,000, calculate the critical value and margin of error associated with a 97% confidence level.

3. Provide the lower and upper bound on the 97% confidence interval for average annual income for both men and women.

4. Calculate the t-statistic and *p*-value associated with the tests that average annual income of men equals $80,000 and that the average annual income of women equals $80,000.

5. In preparation to test if average annual income for men equals that of women, calculate the sample standard deviation for the difference in average income.

6. Calculate the t-statistic and *p*-value of the test that average annual income of men equals average annual income of women.

7. Repeat step 6 using Excel's Data Analysis Toolpak's **t-test: Two Sample Assuming Unequal Variances**.

8. Repeat steps 2 through 7 above testing that average SAT scores equal 800, separately for both men and women.

9. Save your Excel work as YourName_Lab6_Exercises.

10. Fill in the answer sheet for the lab.

Turning in your work

Email both YourName_Lab6_Tutorial.xlsx and YourName_Lab6_Exercise.xlsx to your professor as file attachments to a single email with the subject heading Excel Lab 6: Your Name. Turn in your filled-in answer sheet during class.

Answer Sheet for Lab #6: Testing Means and Differences in Means

Name: ______________________

1. What is the margin of error when testing average annual income for men = $80,000?

2. What is the test-statistic when testing average annual income for women = $80,000?

3. What is the sample standard deviation in the difference in men and women average annual income?

4. What *p*-value do you calculate when testing that there is no difference in average annual incomes between men and women?

5. What *p*-value does Excel produce (using Data Analysis) when testing that there is no difference in average annual incomes between men and women?

6. What accounts for the difference in *p*-values between 4 and 5?

7. Would you reject or fail to reject the claim that there is no difference in average annual incomes between men and women? Explain.

8. What is the 97% confidence interval for the average SAT score for men?

9. What is the *p*-value associated with the test that average SAT scores for women = 800?

10. Would you reject or fail to reject the claim that average SAT scores for women equal 800? Explain.

11. What is the average SAT score for men? What is the average SAT score for women? Is the difference statistically significant? Explain.