**Stat 401 – analysis of paired data, log transformations**

**Goals:** In this lab, we will:

explore two ways to run a paired t-test

look at the Wilcoxon signed rank test

look at how to transform data

show how to run a Wilcoxon rank sum test

We will use the case0202.txt data set for the first 3 parts. This is the hypocampus volume in twins, one with and one without schizophrenia, used as case study 2.2 in chapter 2.

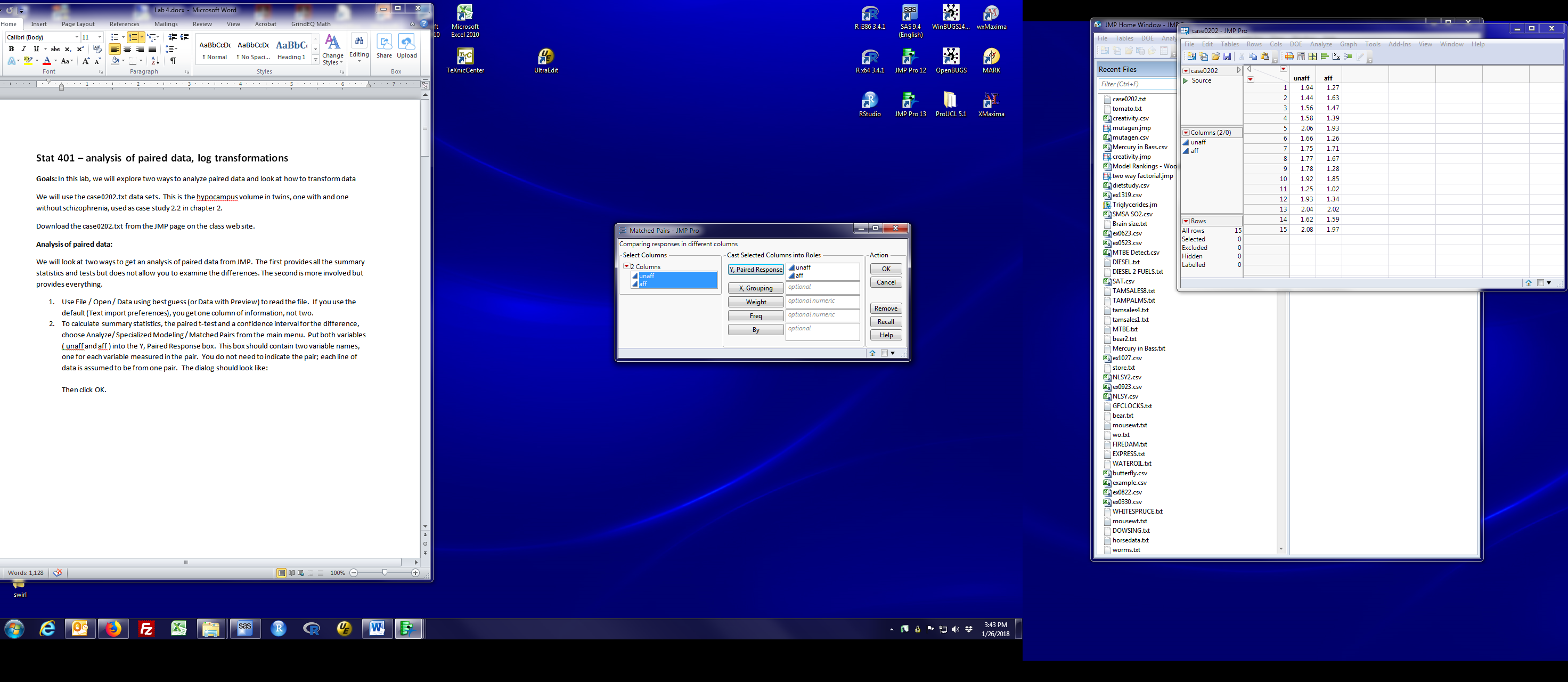
We will use the hamburger.csv data set for the last part

Download the case0202.txt and hamburger.csv from the JMP page on the class web site.

**Paired t-test:**

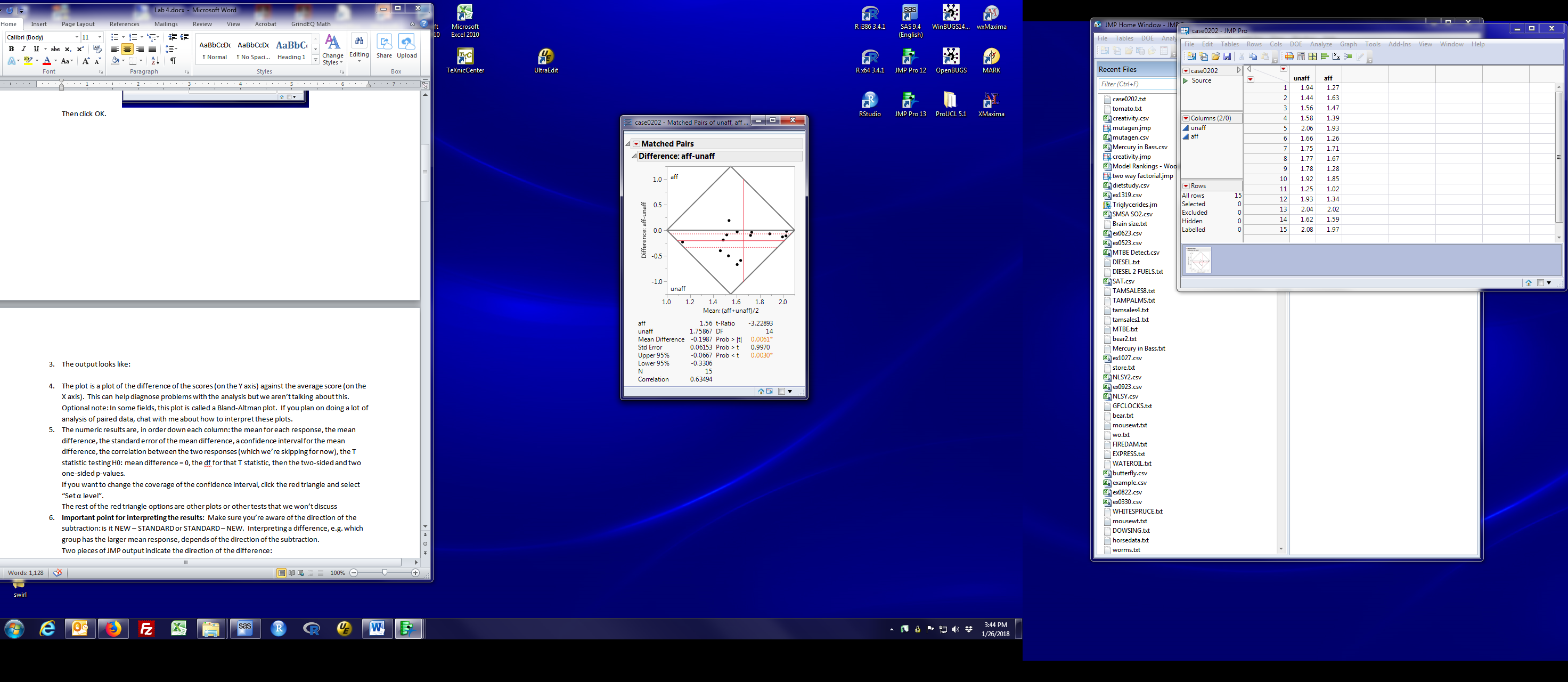
We will look at two ways to get a paired t-test from JMP. The first provides all the summary statistics and tests but does not allow you to examine the differences. The second is more involved but provides everything.

1. Use File / Open / Data using best guess (or Data with Preview) to read the file. If you use the default (Text import preferences), you get one column of information, not two.
2. To calculate summary statistics, the paired t-test and a confidence interval for the difference, choose Analyze/ Specialized Modeling / Matched Pairs from the main menu. Put both variables ( unaff and aff ) into the Y, Paired Response box. This box should contain two variable names, one for each variable measured in the pair. You do not need to indicate the pair; each line of data is assumed to be from one pair. The dialog should look like:



Then click OK.

1. The output looks like:



1. The black diamond provides an outline of possible outcomes. It is mathematically impossible for an observation to fall outside those black lines. I find them a distraction. If you want to get rid of them, click the red triangle by Matched Pairs and uncheck Reference Frame. Then all you see is the data and three horizontal lines. These are the mean difference and 95% confidence intervals for the mean difference.
2. The plot is a plot of the difference of the scores (on the Y axis) against the average score (on the X axis). This can help diagnose problems with the analysis but we aren’t talking about this.

Optional note: In some fields, this plot is called a Bland-Altman plot. If you plan on doing a lot of analysis of paired data, chat with me about how to interpret these plots.

1. The numeric results are, in order down the first column: the mean for each response, the mean difference, the standard error of the mean difference, a confidence interval for the mean difference, the number of pairs, the correlation between the two responses (which we’re skipping for now). Down the 2nd column, you have the T statistic testing H0: mean difference = 0, the df for that T statistic, then the two-sided and two one-sided p-values.

If you want to change the coverage of the confidence interval, click the red triangle and select “Set α level”.

The rest of the red triangle options are other plots or other tests that we won’t discuss or won’t discuss right now.

1. **Important point for interpreting the results**: Make sure you’re aware of the direction of the subtraction: is it unaff – aff or aff – unaff? Interpreting a difference, e.g. which group has the larger mean response, depends of the direction of the subtraction.

Two pieces of JMP output indicate the direction of the difference:

* 1. The label of the box: these results are for aff - unaff
  2. The relationship between the means for each group (first two pieces of numeric output)

To change the direction for the difference, swap the order of variables in the Y, Paired Response box (item 2).

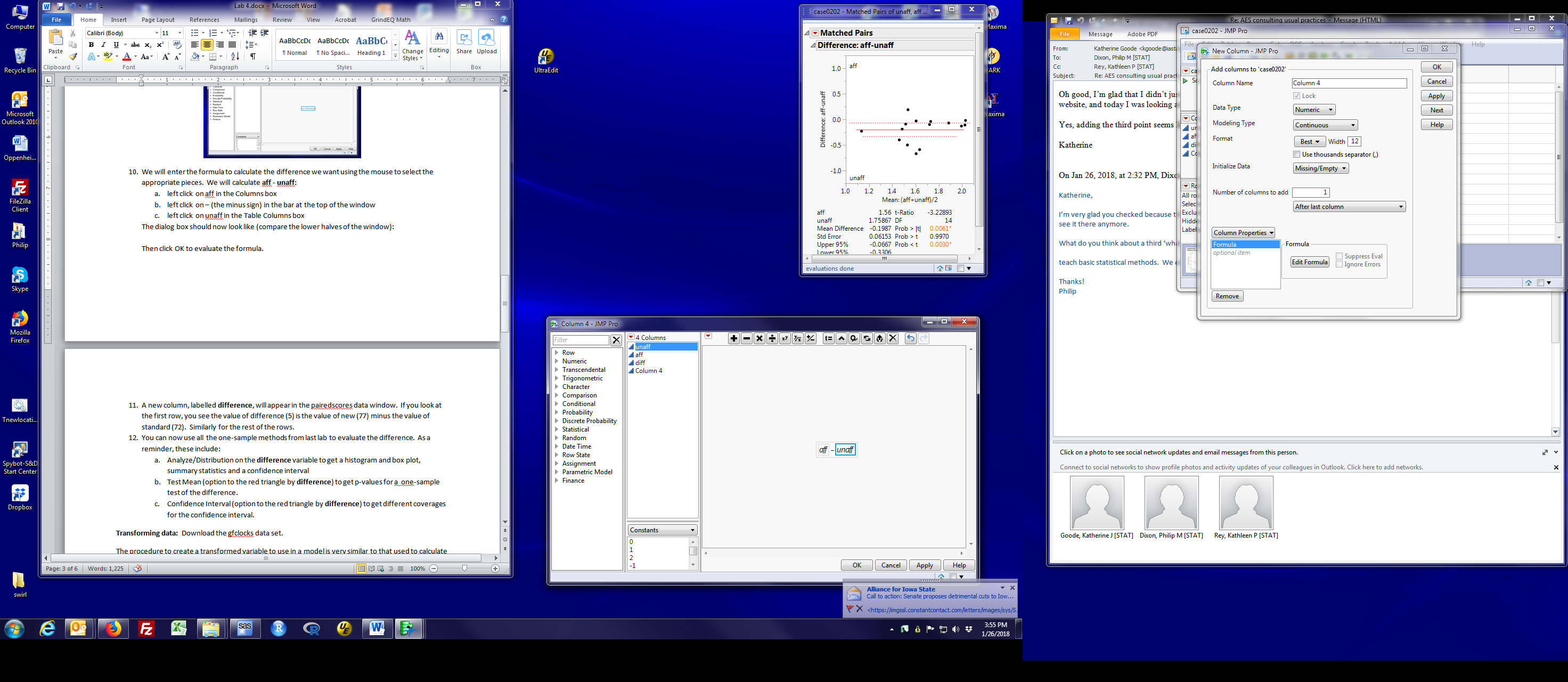
**Paired t-test (2nd approach):**

1. You may have noticed that book’s calculations start by calculating the difference within each pair but the previous method didn’t require that. JMP does that behind the scenes for you. If you want to examine the differences yourself, you need to explicitly calculate them. Doing that is the second approach.
2. Go back to the case0202 window (the one with the data) and right click on the empty column to the right of **aff**. Select ‘New Column’. Type the variable names in the Column Name box (replacing Column 3). Since this column will contain the difference, **diff**, is an informative name but the choice is yours. Then click the black triangle by Column Properties at the bottom of the window. A long pop-up menu will appear. Select Formula. A dialog box will appear looking like:



1. We will enter the formula to calculate the difference we want using the mouse to select the appropriate pieces. We will calculate **aff - unaff**:
   1. left click on aff in the Columns box
   2. left click on – (the minus sign) in the bar at the top of the window
   3. left click on unaff in the Table Columns box

The dialog box should now look like (compare the right-hand parts of the before and after windows):

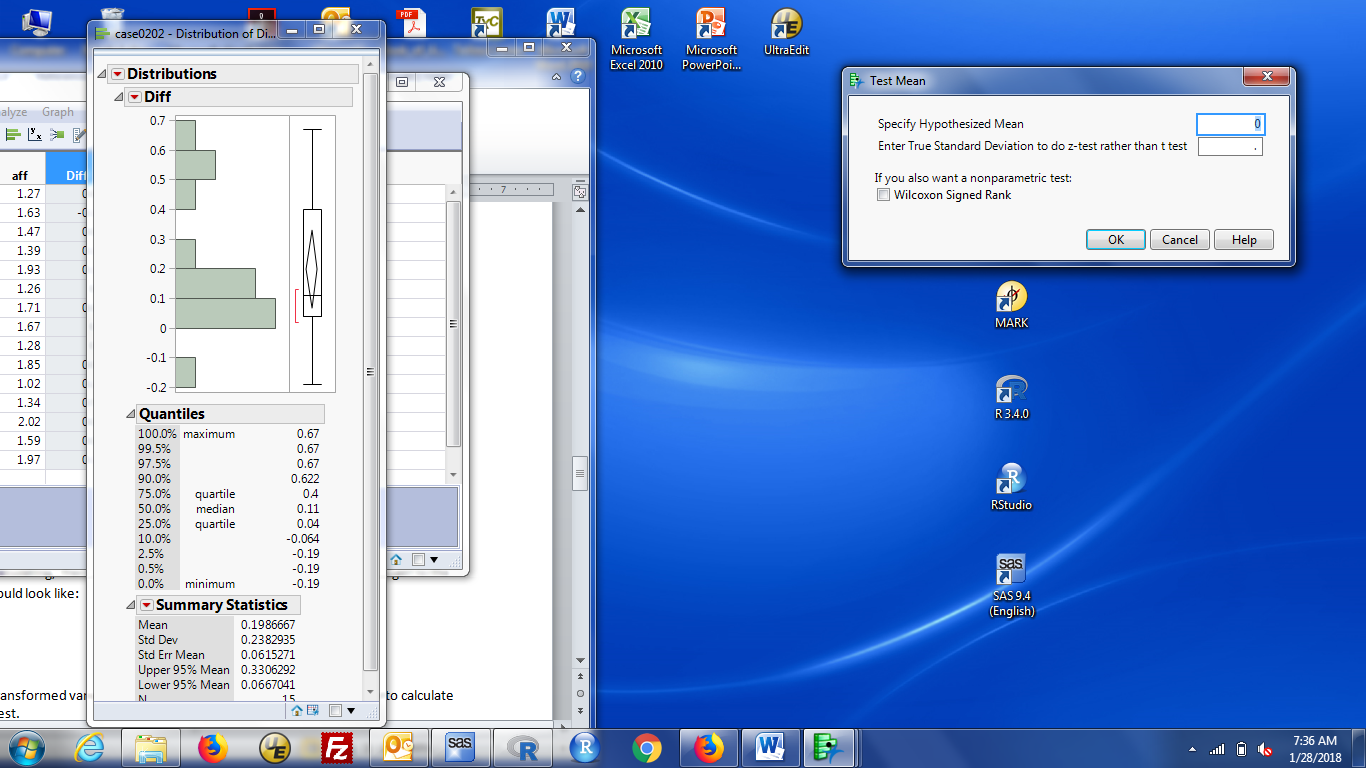


Then click OK to evaluate the formula and store the results in a column.

1. A new column, labelled **diff**, will appear in the case0202 data window. If you look at the first row, you see the value of difference (-0.67) is the value of aff (1,27) minus the value of unaff (1.94). Similarly for the rest of the rows.
2. You can now use all the one-sample methods from last lab to evaluate the difference. As a reminder, these include:
   1. Analyze/Distribution on the **diff** variable to get a histogram and box plot, summary statistics and a confidence interval
   2. Test Mean (option to the red triangle by **diff** after Analyze/Distribution) to get p-values for a one-sample test of the difference.
   3. Confidence Interval (option to the red triangle by **diff** after Analyze / Distribution) to get confidence intervals with the coverage you select.

**Wilcoxon signed rank test (nonparametric test for paired data):**

Follow the 2nd approach (calculating, then analyzing, a difference variable). Pause when you get to the Test Mean dialog box. It should look like:



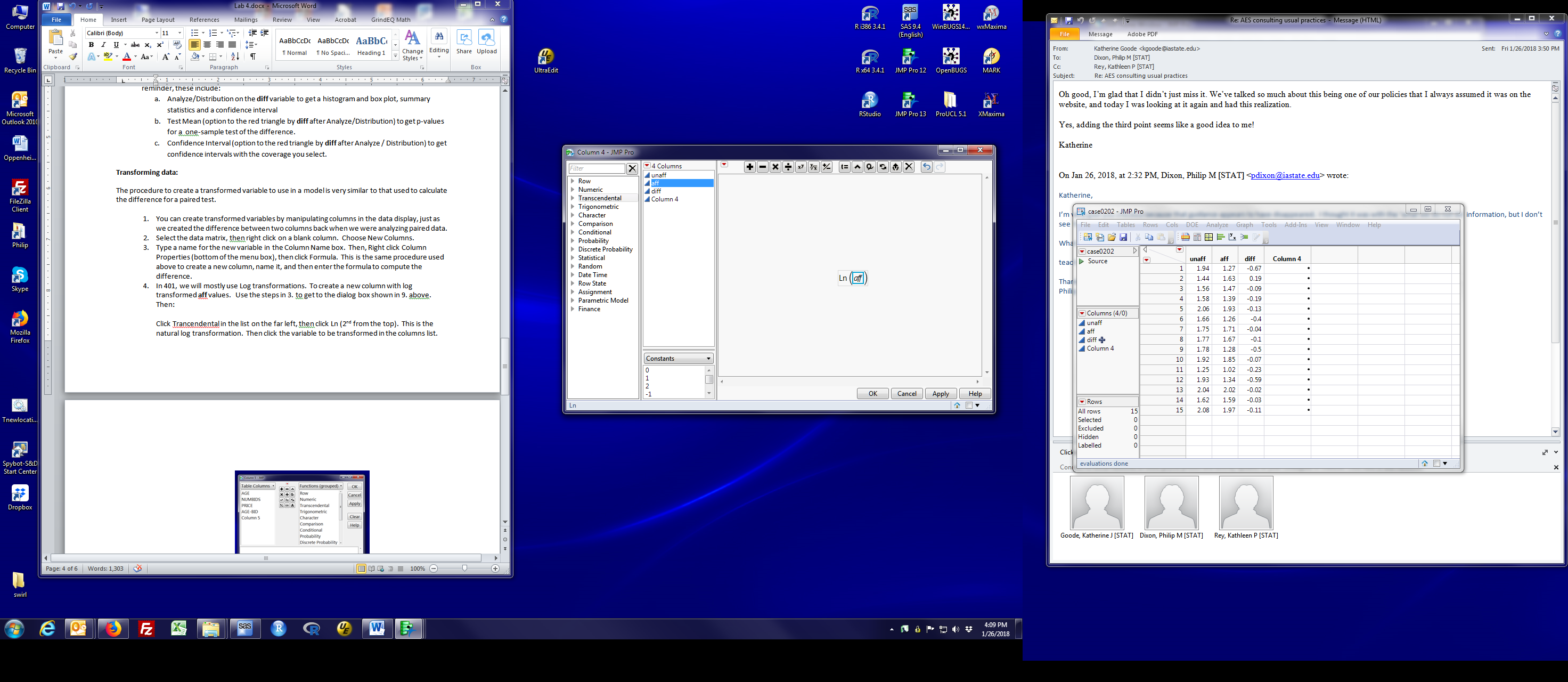
Check the box labeled Wilcoxon Signed Rank, then click OK. You see the Signed Rank test results next to the t-test results. Again, JMP provides three p-values. The two sided p-value is the row labeled Prob > |t|.

**Transforming data:**

The procedure to create a transformed variable to use in a model is very similar to that used to calculate the difference for a paired test.

1. You can create transformed variables by manipulating columns in the data display, just as we created the difference between two columns back when we were analyzing paired data.
2. Select the data matrix, then right click on a blank column. Choose New Columns.
3. Type a name for the new variable in the Column Name box. Then, Right click Column Properties (bottom of the menu box), then click Formula. This is the same procedure used above to create a new column, name it, and then enter the formula to compute the difference.
4. In 401, we will mostly use Log transformations. To create a new column with log transformed **aff** values. Use the steps in 3. to get to the dialog box shown in 9. above. Then:

Click Trancendental in the list on the far left, then click Ln (2nd from the top) or Log (3rd from the top). This is the natural log transformation. This is what I want when I talk about log transformations. Then click the variable to be transformed in the columns list. aff should appear in the formula box and you should see:



1. Click Ok and a new column will appear that contains log aff values.
2. If you do want log base 10, then choose log10 from the transendental list.
3. The square root transformation is obtained by clicking the button on the top row of buttons. Put aff in the x place and leave y blank.
4. If creating a transformed variable in JMP is opaque, you can always read the data set into Excel, create a new spreadsheet column with the transformed variable, save the new worksheet, then read the new data set into JMP.

**Wilcoxon rank sum test:** uses hamburger.csv.

Read in the data set. You will note that treatment is a nominal variable (categories, red bars) and cfu is a continuous variable (numbers, blue ramp) by default. If you read a data set and the treatment is a continuous variable, you need to change the modeling type for that variable (described in detail in an earlier lab).

Start the same way you would for a t-test: Analyze / Fit Y by X. Treatment in the X, Factor box and cfu in the Y, response box, then click OK.

Choose the analysis by clicking the red triangle, select Nonparametric. There are two options for the Wilcoxon test: Wilcoxon test and Exact Wilcoxon Test. The first uses the Normal approximation to the p-value (appropriate for large samples); the second does a permutation test on ranked values (best for small samples). To get the second, choose Wilcoxon Test, then choose Exact test from the Nonparametric menu, then Wilcoxon Exact test.

The Wilcoxon Test returns two p-values: The one labeled Chi-square approximation does not use the continuity correction. It is labeled a 1-way test (which will make more sense after our discussion of 1-way ANOVA); it is a two-sample test with a two-tailed p-value. The one labeled Normal Approximation uses the continuity correction.

If you additionally request the Wilcoxon Exact Test, the two-sided p-value is labelled Prob ≥|S - Mean|in the 2-sample: Exact Test box.