SNRT Statistics Workshop
Polytechnic of Namibia
30 November 2009

Workshop leader:
Larkin Powell, University of Nebraska-Lincoln, http://snr.unl.edu/powell, lpowell3@unl.edu

Agenda
Review of data types
Review of statistical hypotheses and p-values
GenStat and Excel
Categorical data: Chi-square test
Continuous data: descriptive statistics
Continuous data: normal distribution test (discussion of need)
Continuous data: t-test
Continuous data: Mann-Whitney U (non-parametric)
Continuous data: ANOVA
Continuous data: correlations
Continuous data: linear regression

Directions for Excel and GenStat Analyses

Loading data into GenStat:

Click on the Desktop icon to run GenStat Discovery program.
Select to Start the Session using a BLANK SPREADSHEET.
Select Size of New Spreadsheet as the DEFAULT values (just click OK).

With you new, blank spreadsheet open (it will have "*" in each cell):
1. Go to your Excel spreadsheet (use our TTEST data set).
2. Highlight all of the data (not the column headings).
3. Click either EDIT—COPY or CNTRL-C to copy your data from the Excel spreadsheet.
4. Go to GenStat.
5. Select EDIT—PASTE or CNTRL-V to paste the data into the blank spreadsheet.
6. Trim the GenStat spreadsheet to the size of your data by clicking on SPREAD—DELETE—EMPTY ROW+COLUMNS. (you will see the other columns and rows disappear).
7. Transform your data into GenStat data file by clicking SPREAD—UPDATE—CHANGED DATA TO GENSTAT. (this creates a data set that GenStat can read and analyze)

Your data is now loaded and ready for analysis!
NOTE: if you look on GenStat's OUTPUT window, you will see a brief description of your data set. In our case, it provides:

```
  Identifier   Minimum      Mean   Maximum    Values   Missing
          C1     9.000     16.03     28.00        79         0

          C2     18.00     30.48     47.00        79         0
```

So, you've already done a simple characterization of your data without even trying…!

CHI-SQUARE TEST

Chi-square test in EXCEL:

Reminder: Null hypothesis is that the observed values do not differ from the expected values.

1. Go to the CHI_SQUARE_1D tab in the Excel file.
2. You should see observed values and expected values for colors of sheep following special breeding.
3. In an empty cell, type:

```
=CHITEST(a4:c4, a8:c8)
```

NOTE: You will then be prompted to provide the range of the observed values, and the range of the expected values—separated by a comma. In our case, the observed values are in cells A4, B4, and C4 (shorthand for these continuous cells is: A4:C4). And, the expected values are in cells A8, B8 and C8 (shorthand: A8:C8).

Excel will provide the P-value for the chi-square test. If P<0.05 you reject the null hypothesis. Note: Excel does not provide YOU the value for the calculated chi-square statistic or the degrees of freedom, but it uses them to obtain the P-value.

Note: you can do the same functions for 2-dimensional chi-square data sets. Just use the dimensions of the tables in the CHI_SQUARE_2D data set in the Excel file provided for the workshop.

Chi-square test in GenStat:

Reminder: Null hypothesis is that the observed values do not differ from the expected values.

The chi-square test in GenStat is normally done manually, without pre-loading the data into GenStat.

1. Click on STATS—STATISTICAL TESTS—CHI-SQUARE GOODNESS OF FIT.
2. You will be given a menu to tell GenStat where to find your observed and expected values.
3. To the right of "Observed Data" click on CREATE TABLE.
4. Tell GenStat to create the table using SPREADSHEET.
5. Provide GenStat with the size of our table. In this case, it is 1 row and 3 columns. Ignore its petty little reminder that a table usually has at least 2 rows. It thinks it knows everything…
6. When the empty table appears, go to your Excel spreadsheet, copy the data (data only—no headings) and paste it into this blank table. Leave the table (do not close it).
7. Repeat the process with "Expected Frequencies". Paste your expected values into the blank table provided.
8. Provide the degrees of freedom. In this case, it is [number of columns]-1 = 3-1 = 2.
9. Select the Pearson chi-square calculation.
10. Click OK.
11. Review your results on the Output Log. Note that GenStat provides the chi-square value, the df, and the P-value.

Note: you can do the same functions for 2-dimensional chi-square data sets. Just use the dimensions of the tables in the CHI_SQUARE_2D data set in the Excel file provided for the workshop.

DESCRIPTIVE STATISTICS

Descriptive Statistics in Excel:

Let's use the T-Test data…so go to that tab.

1. Scroll to the bottom of the data columns.
2. Under column A, select an empty cell and type the following:
   =average(a4:a82)   [this calculates the mean of your data in column A]

3. Now, try some other descriptive statistics:

Median:	=median(a4:a82)
Mode:	=mode(a4:a82)
Standard Deviation:	=stdev(a4:a82)

You can do the same for the B Column. You can actually copy and paste the directions from the A column to the same row in the B column, and Excel will update the equation for you.

NOTE: you can construct a 95% confidence interval around the mean by using the following:

Upper 95% CI limit = mean + [1.96 x SD]
Lower 95% CI limit = mean - [1.96 x SD]

Descriptive Statistics in GenStat:

We've already loaded our TTEST data in GenStat (if doing this outside the workshop, you'll have to re-load this data, following directions above).

1. Click on STATS—SUMMARY STATISTICS—SUMMARIZE CONTENTS OF VARIATES.
2. Notice that GenStat lists C1 and C2 as "Available Data" in the window provided.
3. Click on C1 and use the arrow button on the window to "move" C1 to be listed in the "Variates" window.
4. Do the same for C2.
5. You can select any summary statistics to be calculated. I suggest:
   Arithmetic mean, median, mode, standard deviation, No. of values, minimum, maximum
6. You can also select some plots to look at. I suggest the box plot.
7. Click OK.
8. Go to the output screen to view the summary statistics. They should match what Excel calculated?!
9. You can view the plots you created by going to the graphics windows created by GenStat. The box plot shows the median (central line), the 25 and 75% quartiles (meaning the central 50% of the data lies inside the box), and the range of the data (the whiskers).

Normal distribution test plot:

To my knowledge, there is no test for normality built into Excel. Automated calculations (using Excel) have been developed by others are available by googling "Excel test for normality".

In GenStat: To visually assess your data's fit to a normal distribution, click on STATS—DISTRIBUTIONS—PROBABILITY PLOTS.

1. Select which variable to plot. In our case, let's start with C1. (NOTE: this is still our TTEST data).
2. Double-click on C1 to move it to the Data Values field.
3. Do not change any of the default values. Be certain that "normal" is selected under the DISTRIBUTION option.
4. Click OK.
5. Inspect your plot. If the data points do not fall outside the 95% confidence interval bands, your data is approximately normal.

More info on creating and interpreting normal probability plots:
http://www.pathmaker.com/resources/tools/normal_test_plot.asp
http://www.math.hope.edu/swanson/statlabs/normal_pp.html
http://www.statit.com/support/quality_practice_tips/normal_probability_plot_interpre.shtml

Preparing for statistical tests in Excel:
Excel assumes that most people using Excel will NOT be doing statistical analyses, so it 'hides' the statistics tests from the general public. To make them available, follow these instructions (note that these must be modified slightly in Microsoft 2007 versions of Excel).

1. Click on Tools—Add-ins.
2. Put a 'check mark' next to BOTH of the Analysis Toolpak options. Click OK.
3. You are now ready to go. [Note: if you are using a version of Excel on your home computer, you may have opted (years ago) to not install these Add-Ins when you first installed Excel…you'll have to get your original Microsoft disks and install them!]

T-TESTS

T-Test in Excel:
There are many types of t-tests, including paired data, t-test of differences of a mean, and t-tests of two means. The data we will be using today invites us to see if the two means are different. So, we'll be doing two-tailed t-tests of two independent data sets.

1. Go to the TTEST tab in your Excel file to find the data. The data provides miles per gallon (sorry, Namibians…) for cars manufactured in the USA and Japan. There are 79 different models of cars from both countries to compare.
2. Click on TOOLS—DATA ANALYSIS. Select "t-test: two-sample assuming unequal variances"
3. We are then given a little wizard to tell Excel where to find our data. Click on the fields to show Excel where to find your two columns of data.
4. The hypothesized mean difference is 0. [our null hypothesis is that the means are the same for the two countries' cars]
5. Click on the button next to Output Range, and then tell Excel where to put the results of your test (somewhere below the data).
6. Click OK. You will find your t-value (use 2-tailed) and the P-value and degrees of freedom in a new table created by Excel.

T-Test in GenStat:
1. Click on STATS—STATISTICAL TESTS—one and two-sample t-tests.
2. Under the TEST option, select "two-sample unpaired".
3. Double-click on C1 and C2 to place them in the Data Set 1 and Data Set 2 fields.

4. Be sure that 2-sided test is selected.
5. Click OK.
6. Go to the Output log to see your results.

Notice that GenStat provides various descriptive statistics, in addition to the t-statistic and P-value. GenStat also does a preliminary test for equal variances of the two data sets before calculating the t-statistics.

NON-PARAMETRIC COMPARISON OF DATA SETS

Mann-Whitney U Test (also known as Wilcoxen Ranked Sums):

NOTE: Excel does not automatically do the Wilcoxen test. You can do it by hand in Excel, but you must have the formulas involved.

NOTE: I only recommend using non-parametric tests for ordinal data or in very extreme circumstances. This test does NOT test for the differences in means.

Null hypothesis: the two data sets are exactly the same.

1. Click on STATS—STATISTICAL TESTS—TWO-SAMPLE NON-PARAMETRIC TESTS
2. You have already brought in the data used in the TTESTS, so we will use it.
3. Select C1 and C2 as your data sets
4. Click OK.
5. Go to the Output log to find your results.

Single-factor ANOVA

1-factor ANOVA in Excel:

Null hypothesis: the means of prices of fuel are the same in all 6 cities in Australia.

1. *Use the data on the right side of the Excel spreadsheet (data in 6 columns).*
2. Click on TOOLS—DATA ANALYSIS—ANOVA SINGLE FACTOR
3. Provide the location of your data, grouped in columns (use the 6 columns provided)
4. Provide an output range for your results.
5. Click OK.
6. Your results will be placed in a new table created by Excel.

1-factor ANOVA in GenStat:

Null hypothesis: the means of prices of fuel are the same in all 6 cities in Australia.

1. Bring in your data from the Excel file – only data on CITY and PRICE. *NOTE: use the data structure on the left side of the Excel file—with data in two columns.*
2. Create a new spreadsheet in GenStat and copy the data.
3. Delete the extra columns and rows in the GenStat spreadsheet.
4. Highlight the column with your City data.
5. Click on SPREAD—FACTOR—CONVERT TO. [this tells GenStat that this column will be the 'factor' in our ANOVA analysis]
6. Convert the spreadsheet to a GenStat data set by clicking on SPREAD—UPDATE—CHANGED DATA TO GENSTAT.
7. Click on STATS—ANALYSIS OF VARIANCE.
8. In the ANOVA window, select your second column as the Y-VARIATE and your first column as the 'treatment'.
9. Change the INTERACTIONS option to "no interactions".
10. Click on OK.
11. Go to the output log to view your results. [Note that it provides the means of each city's prices, as well as the P-value which can be found under "F pr."

CORRELATION ANALYSIS

Correlation in Excel:

1. Go to the CORR tab in the Excel file provided. The data is to test for a relationship between hours of TV watching and grades.
2. Click on TOOLS—DATA ANALYSIS—CORRELATION
3. Provide the location of your data (no headings) to Excel.
4. Your data are grouped in columns.
5. Provide a location for the output range.
5. Click OK. Your results will be in a table created by Excel.

Correlation in GenStat:

1. Copy your data (no text headings) from Excel (CORR tab).
2. Create a new spreadsheet in GenStat.
3. Paste in your data.
4. Delete the extra cells.
5. "update GenStat to create a GenStat database for this data.
6. Click on STATS—SUMMARY STATISTICS—CORRELATIONS.
7. Use your two new columns for the two data sets.
8. Click OK.
9. Find your results on the Output Log. The correlation statistic (r) is -0.628.

LINEAR REGRESSION

Regression in Excel:

1. Go to the REGRESS tab in our Excel file. This data explores the question, "If Australia football teams spend more money, do they get more wins?"
2. Click on TOOLS—DATA ANALYSIS—REGRESSION.
3. We need to figure out which is our DEPENDENT and which is our INDEPENDENT variable. In this case, we hypothesize that the number of wins DEPENDS on the money spent. So, WINS is our DEPENDENT variable. That makes it the Y-variable.
4. Provide the range of your data for the X and Y variables (MONEY and WINS).
5. Provide an output range for your results.
6. Click OK.
7. Inspect your results in the table provided by Excel. Your slope ("x-variable") and intercept are key results, as well as the P-value for the slope (NULL H: slope is 0).

Regression in GenStat:

1. Bring in your data from Excel to GenStat.
2. Create a new spreadsheet in GenStat.
3. Paste your data from Excel into GenStat.
4. Delete the extra cells.
5. "Update" GenStat to create a GenStat database for this data.
6. Click on STATS—REGRESSION ANALYSIS--LINEAR MODELS.
7. Select "Simple Linear Regression"
8. Select "wins" (your second column) as the "RESPONSE VARIABLE.
9. Select "money" (your first column) as the "EXPLANATORY VARIABLE.
10. Go to the Output log to view your results.

For our data:

```
                  estimate         s.e.     t(14)  t pr.
Constant             -13.6         37.6     -0.36  0.724
C51                  1.168        0.644      1.81  0.091
```

The intercept is -13.6, and the slope is 1.168. The P-value for the slope is 0.091. Because P>0.05, we do not reject the null (NULL H: slope = 0). Thus, the money spent on football does not mean more wins. Note that the P-value is very close to 0.05, so there is SOME evidence, but not significant at the alpha=0.05 level.

SUMMARY:

There is much to explore in GenStat that we did not cover. Hopefully, the coverage of the basic statistics will show you the process to use in more complicated situations.

Feel free to contact me via email (lpowell3@unl.edu) if you have questions as you work with your data! Best wishes, Larkin Powell

NOTE: the data and other downloads (powerpoint and this instruction sheet) are available for download at:

http://snr.unl.edu/powell/PoN/stat_workshop.htm