How to organize and analyze research data

[[[BEGIN SIDEBAR AROUND HERE]]]

For this lesson, you'll need the following:

- **Microsoft Excel**
- **PASW (formerly SPSS) statistics software, or free trial download**

You should already be familiar with:

- **The basic functions and tools of Excel**
- **Descriptive statistics and inferential statistics**
- **Research design and how to identify independent variables and dependent variables**

[[[END SIDEBAR]]]

This short lesson is intended for graduate and advanced undergraduate students undertaking original research projects. After completing the 30-minute lesson, you will be able to to:

- Format raw research data in an Excel spreadsheet for efficient analysis
- Run basic descriptive statistics in Excel
- Prepare the document for statistical analysis in the software package PASW

Format data for analysis

Make sure the computer you are working on has Microsoft Excel installed. Open the document ToolBox_DataLesson.xls [[[NOTE: NEED LINK TO DOWNLOAD TUTORIAL EXCEL FILE]]] in Excel and follow along with the lesson, taking action as instructed.
In the Raw_Dataset spreadsheet, the name of each variable has been entered in the first row of each column.

1. Each variable name must be different from other variable names.
2. The first variable (in column A) is a **unique identifier.**
3. Variable names must start with a letter (not numbers or special characters), so change *4tutorial_types* to *tutorial_types*.
4. Name variables so that they are intuitive to you. Therefore, change *use* and *useful* to *perc_ease_of_use* and *perc_usefulness*, respectively.
5. Your spreadsheet should look like Figure 1, and it should now be readily apparent that Column F refers to *Perceived Usefulness* and Column G refers to *Perceived Ease of Use*.

[[[INSERT IMAGE AROUND HERE datalesson_fig3.png]]]
CAPTION: Tutorial type, gender, experience, and task categorical values have been formatted as dichotomous variables.

Data values, or rows, have been input for each subject in the experiment.

1. Decide on input conventions and stick to them. In *gender*, change the "female" value to "F".
2. Separate data into component values whenever possible by adding new columns. For example, *tasks_completed* values (Y/Y/N) can be broken up into three components. So, add columns *task1*, *task2*, and *task3*, and reformat values appropriately.
3. Double-check to ensure no data entry errors have been made, then delete the old *tasks_completed* column.
4. Your Raw_Dataset spreadsheet should now look like Figure 2.

[[[INSERT IMAGE AROUND HERE datalesson_fig2.png]]]
CAPTION: Values for each variable are entered in a consistent format.

Replace categorical data values with "0" or "1" (0=no, 1=yes) to indicate whether or not the value is represented for the given subject/item. This makes the categorical values **dichotomous**, which gives the researcher maximum flexibility in testing for relationships or correlations.

[[[BEGIN SIDEBAR AROUND HERE]]]

Dichotomous variables allow for the easy isolation of characteristics for analysis. For example, separating *experience* levels allows the researcher to test in PASW whether subjects with low experience have *perceived ease of use* scores significantly different from those with medium and high experience levels.

[[[END SIDEBAR]]]

1. Copy the Raw_Dataset values into the Sheet2 tab. Rename this sheet Formatted_Data.
2. In the Formatted_Data sheet, add new columns with **Insert > Columns** for each distinct value in categorical data columns (*tutorial_types*, *gender*, *exp_level*, *task1*, *task2*, and *task3*). You do not have to add a new column for *task1*, *task2*, and *task3* because their values are already dichotomous.
3. Rename columns according to each possible categorical value. Copy and paste the values from the original column.
4. Replace dichotomous data values with "0" or "1" (0=no, 1=yes) using the replace function: **Edit > Replace.**
5. Your Formatted_Data spreadsheet should now look like Figure 3.

[[[INSERT IMAGE AROUND HERE datalesson_fig3 (1).png]]]]
CAPTION: Tutorial type, gender, experience, and task categorical values have been formatted as dichotomous variables.

Run descriptive statistics

> **[[[BEGIN SIDEBAR AROUND HERE]]]**
> **Analyzing descriptive statistics may lead you to ask new questions. For example, *perc_enjoy* is much higher on average than *perc_usefulness*, which also has a large variance. This suggests that a segment of users may not have considered the system useful even though they enjoyed using it.**
> **[[[END SIDEBAR]]]**

Run descriptive (summary) statistics on the dataset and output results in a new spreadsheet. If necessary, rearrange columns so that data requiring summary, those with **continuous** values, are adjacent. (In the spreadsheet, the continuous values of *perc_enjoy*, *perc_usefulness*, and *perc_ease_of_use* are already adjacent.)
Write down the "input range" (i.e. H2:J26) of the data requiring summary, making sure not to include variable labels.

1. Go to **Tools > Add-Ins...** and make sure **Analysis ToolPak** is checked. Click OK.
2. Go to **Tools > Data Analysis** and select **Descriptive Statistics**. Click OK.
3. Enter the input range in the blank and check **Summary Statistics**. Click OK.

The descriptive statistics summary will output in a new sheet. Rename this spreadsheet Descriptive_Stats. Label the output according to the variables. For example, *Column 1* should be *perc_enjoy*. Your Descriptive_Stats spreadsheet should now look like Figure 4.
Analyzing descriptive statistics is a great way to start appraising a dataset before running inferential statistics. What insights about the dataset can you glean from the summary statistics for *perc_usefulness* and *perc_ease_of_use*?

[[[INSERT FIGURE ABOUT HERE datalesson_fig4.png]]]
CAPTION: Excel will not run descriptive statistics with non-numeric characters, so the variable labels must be re-entered.

Run inferential statistics

Open PASW and upload the spreadsheet. (IIT labs in Stuart Building, room 112, have PASW. You can also download a 30-day free trial.)

1. When PASW starts, it will prompt you for a data source. Select **Open an existing data source > More Files...**
2. Set the file type to Excel, then find and open the file.
3. From the list, select the Formatted_Data worksheet. Click OK. You should now see the dataset.

The most commonly used statistical methods and tests are found under the **Analyze** menu in the standard toolbar.

Figure 5 shows where to find some common statistical methods and tests. Under PASW's **Analyze** menu, see if you can also find the following:

1. Wilcoxon-Mann-Whitney test
2. Simple linear regression
3. Non-parametric correlation

[[[INSERT IMAGE ABOUT HERE datalesson_fig5.png]]]
CAPTION: Don't assume PASW doesn't run a given test. Many are located under broader statistical classifications.

Use your prior knowledge or the *Choosing the Correct Statistic* resource to determine which statistic to run to find whether there is a significant difference in *perc_usefulness* between *free_play* and other tutorial types.

[[[BEGIN SIDEBAR ABOUT HERE]]]

> **Consult UCLA's *Choosing the Correct Statistic* to ensure the statistical anlaysis is appropriate for your research question(s), variables of interest, and associated data types.**

[[[END SIDEBAR]]]

You should have determined that the Kruskal-Wallis one-way analysis of variance is the appropriate test. Go to **Analyze > Nonparametric Tests > K Independent Samples...** A new window will open prompting you to select the variables (by column name) to be tested.

1. Select the *perc_usefulness* variable from the scrollable list on the left, and click the top arrow to move it to the Test Variable List (i.e. dependent variable).
2. Select the *free_play* variable from the scrollable list, and click the bottom arrow to move it to the Grouping Variable section (i.e. independent variable).

3. Click **Define Range...**, and enter "0" in the Minimum field and "1" in the Maximum field. This specifies the two categorical variables to be tested. Your screen should now look like Figure 6. Click Continue.

[[[INSERT IMAGE ABOUT HERE datalesson_fig6.png]]]
CAPTION: For Kruskal-Wallis, an independent (free_play) and a dependent variable (perc_usefulness) are selected.

Click OK to display the statistical analysis results to new viewing window. Your results screen should look like Figure 7. Because statistical analysis output contains many numerical components (some of which must be cited in formal reports), it is a good idea to save it to a format independent of the PASW software. To save the results, go to **File > Export**. Change Document Type to the desired output format (Word, Excel, PDF, etc.), then click OK.

[[[INSERT IMAGE ABOUT HERE datalesson_fig7.png]]]
CAPTION: PASW allows users to export statistics to common document types.

Test your knowledge

Now that you've finished the lesson, please complete a short quiz to test your understanding of formatting research datasets for comprehensive statistical analysis.