

Lab 2 Part 1: Excel

Lab Objectives:

- ◆ Use a range of functions in Excel to calculate descriptive statistics, such as mean, median, and standard deviation.
- ◆ Use Excel to produce a properly formatted bar chart.
- ◆ Become familiar and comfortable with dragging formulas into new columns, and sorting values in Excel.
- ◆ Use Excel to produce a frequency table.

A sample of useful functions in Excel:

Open Microsoft Excel. Review the Excel functions provided in the following table. Try any functions that you are not familiar with, and make sure you understand how to use each function before proceeding.

Function	What it does	Example code
Add	Adds values	=A1+A2
Subtract	Subtracts values	=A1-A2
Multiply	Multiplies values	=A1xA2
Divide	Divides values	=A1/A2
Exponent	Raises a value to an exponent	=A1^6
Square root	Calculates the square root of a value	=SQRT(A1)
Sum	Calculates the sum of a dataset	=SUM(range)
Mean	Average of a dataset	=AVERAGE(range)
Median	Median (middle) of a dataset	=MEDIAN(range)
Standard deviation	Standard deviation of a dataset	=STDEV(range)
Counta	Counts the number of cells that contain a value	=COUNTA(range)
Countif	Counts the number of cells that meet a criteria	=COUNTIF(range,"criteria")
Min	Returns the minimum number in a list	=MIN(range)
Max	Returns the maximum number in a list	=MAX(range)
Rank	Returns the rank of a number in a list	=RANK(cell,range,1 or 2)

Exercise 1: Data Entry

Note that in Excel, each column has a letter and each row has a number. The specific coordinates of a cell are given by the column and letter where that cell is found.

- Enter the following data and labels into the “Sheet 1” worksheet as indicated below. Begin from the top left of the worksheet so that the word “student” is entered in cell A1.

Student	Midterm	Final	Labs
A	67	54	91
B	70	57	87
C	77	65	77
D	88	66	74
E	56	67	56
F	90	89	96
G	88	80	76

SUM

N

MEDIAN

MEAN

STDEV

SE

MIN

MAX

SUM/N

- Begin with the midterm column and type in the appropriate formulas for each calculation (sum, n, median, etc.). For example, to sum the numbers you should type =sum(B2:B8) in cell B10. Note that once you type the first open parentheses, you should use your mouse to highlight the range of numbers, rather than typing the range. Do this for each item that we want to calculate.
 - Consider what function you can use to have Excel count the N (number of samples) in the dataset. Use a formula rather than counting and typing this number yourself.
 - For Standard error of the mean, remember the equation we learned in class: $SE = \text{Standard deviation} / \text{square root of the sample size}$.
 - In the SUM/N cell, type your formula so that it is calculated by referencing the cell containing the sum, divided by the cell containing the N. If you do this correctly it should return the same value as the cell containing the mean.
- Once you have completed the formulas, highlight all of the cells containing formulas, and put your mouse over the black square contained in the lower right corner of the sum/N cell so that it changes to a small black + (see image below).

	B10		f _x	=SUM(B2:B8)
	A	B	C	D
1	Student	Midterm	Final	Labs
2	A	67	54	91
3	B	70	57	87
4	C	77	65	77
5	D	88	66	74
6	E	56	67	56
7	F	90	89	96
8	G	88	80	76
9				
10	SUM	536.0		
11	N	7.0		
12	MEDIAN	77.0		
13	MEAN	76.6		
14	STDEV	12.9		
15	SE	4.9		
16	MIN	56.0		
17	MAX	90.0		
18				
19	SUM/N	76.6		
20				

- Click on this square and then drag your mouse over two columns so that a rectangle is drawn around columns C and D. Then release your mouse. This drags the formulas you typed into the next two columns. You should see numbers fill in for all of your formulas. Excel will automatically use the values in each column (i.e. Final and Labs) for each formula that you dragged into that column.
 - Now you want to calculate each student's final grade. The midterm is worth 30%, the final is worth 45% and the labs are worth 25%. Type appropriate headings for the weighted marks beginning in cell F1. Begin with student A, and type a formula to convert each of his/her grades to the appropriate weighting. Once you have the appropriate formulas for student A, highlight the three cells (F2 to H2) and grab the lower right corner and drag the formulas down so that they are applied to the rest of the students.
 - In cell I1, add the title "Final mark". Type an appropriate formula for student A to calculate his/her final mark then drag the formula down the column to apply it to the rest of the students.
 - Now calculate the median, mean, etc. for the final mark. Go back and highlight all of the formulas in column B. Instead of dragging the formulas over, copy them (ctrl+c). Then click on cell I10 and paste (ctrl+v) to copy all of the formulas into that column.
 - Highlight ALL of the numbers in this sheet (cell B2 to I19). In the Excel ribbon "number" section, click on the drop down menu and select number.
 - In the same section of the ribbon, below the dropdown menu, click on the appropriate button to decrease the decimal places so that only one decimal place is showing.
- Go to Sakai and answer the first two questions about your data.

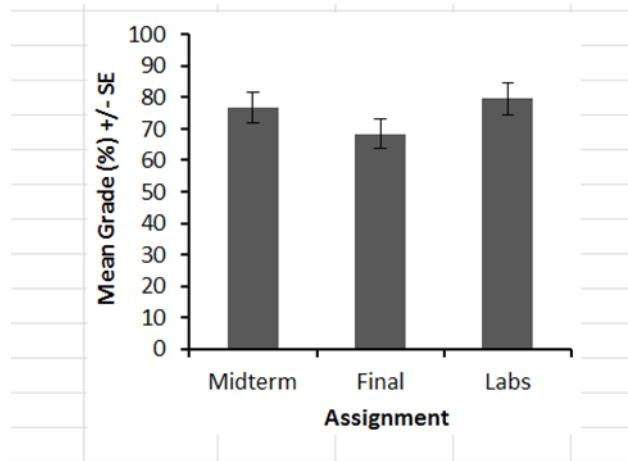
Exercise 2: Simple Graphing

Now that we have summarized our data, let's graph it. It's absolutely crucial that you always graph your data before doing any statistics. This allows you to visualize your data, and can make it easier to validate or choose the best statistical test to answer the question(s) you are asking.

Unfortunately Excel doesn't have advanced graphing options that can really "show all of the data" like we talked about in class. Instead we will create a bar chart that displays the mean mark and standard error for Midterm, Exam and Lab (out of 100).

- Highlight the column titles, press the ctrl button, then highlight the corresponding means. Click "insert" then click on the column chart. Remember that the default chart output is never the best way to show your data. Make the following modifications:
 - Click on the series 1 box and delete it.
 - Click on a bar so that all three bars are highlighted. Click on Chart tools – Format. Under shape fill, select a medium shade of gray to change the colour of your bars.
 - Click on the gridlines, and under Chart tools – Format – Shape outline, select "no outline".
 - Click on the numbers on your Y-axis, then right click and select "format axis". There are a few things we will change here. First, you should change the range so that it spans 0 to 100. Do this under "axis options" by changing the minimum and maximum. Now under "number", change the number of decimal places to zero. Then click on line colour, and select solid line, colour black. Then select on line style, and increase the width of the line to 1.25 pt.
 - Make the same changes to the line on the X-axis.
 - Select your chart as a whole by clicking somewhere near one of the edges. Under chart tools, select Layout and then Axis titles. Add both a horizontal axis title, and a vertical axis title (rotated). Give these titles appropriate names.
 - Select your chart, and click on chart tools – format. Resize the size of your chart so that it is 6 cm high and 7 cm wide. Adjust the positioning of your chart and the titles so that it fills the new space appropriately.
 - Add error bars that represent the standard error of the mean. Select chart tools – layout – error bars – more error bar options. You should always input your own calculations for the error bars, do not select any of the predefined numbers suggested by Excel. Select custom, then click on specify value. For the positive error values, click the button next to the box, and then select the three SE values from your sheet. Click the button again to go back to the first custom error bars box. Do the same for the negative error value (yes, select the same numbers twice). Click OK then close.
 - Finally, let's remove the border from the chart. Click on your chart, then right-click and select "format chart area". Select border colour, then select no line, and close.

- Check that your final graph looks like the one below. Have your TA check your graph to verify that it is completed correctly. If your graph does not look like this, please double-check that you made all of the appropriate modifications, or ask your TA for help.



Exercise 3: Preparing a table for variance

- Note that Excel contains several different worksheets for you to work with. Right-click on this first worksheet, select Rename, and give the sheet an appropriate name. Right-click on Sheet 2 and rename it "Variance". Use this sheet for the next exercise.
- Let's create a table showing the quantities needed to calculate the standard deviation and variance in snake undulation rate. Begin by adding the label "observations" to cell A1 and then inserting the following observations in column A.
 - 0.8, 0.9, 1.1, 1.1, 1.2, 1.4, 1.7, 1.8
- Use the appropriate formula to calculate the mean undulation rate in cell A11.
- Label cell B1 "deviations". In cell B2, type an appropriate formula to calculate the deviation from the mean, using a formula that references the first observation and the cell containing the mean (A11). In the formula, change the reference to the mean from A11 to $\$A\11 . This will tell excel to continue to continue to use the value in cell A11 when you pull down the formula, rather than using A12, A13, etc. Grab the lower right corner of the cell and pull the formula down so that it calculates the deviation for each observation. Check that you have done this correctly by using cell B11 to calculate the sum of the deviations. It should be 0.
- In column C, use the appropriate formula to calculate the squared deviations. In cell C11, calculate the sum of the squares. Use the sum of the squares and the formula learned in class to calculate the variance in undulation rate.
- Go to Sakai and answer the next question about variance.

Exercise 4: Sorting values

- Click on sheet 3, and rename it “sorting”. Label cell A1 “Observations” and enter the following data: 7, 36, 45, 9, 23, 25, 39, 12, 19, 21
 - Highlight the data. In the Excel ribbon, select “Sort and filter” then select “Sort smallest to largest”.
 - In cell B1 enter “Observation number”. Enter the number 1 into cell B2. Grab the lower right corner and pull down. You will notice Excel fills all of the cells with the value 1. Click on the box that pops up next to the lower right corner and select “fill series”. This should number each observation 1 to 10.
 - Use the appropriate formulas from class to calculate which observation (or average observations) corresponds to the median, first quartile, and third quartile.
- Go to Sakai and answer the next question about quartiles.

Exercise 5: Frequency table

- Click on the insert worksheet button to insert a new sheet. Rename it to “Frequency”.
- Determining mean and standard deviation is more complicated for a frequency table. Enter the following data into your worksheet, from a study examining rigor mortis in human bodies. Researchers recorded the number of bodies achieving rigor mortis in each hour after death, in one-hour intervals. Double-check that the data is entered correctly by entering a formula to calculate the total number of bodies (rather than typing it), and verifying that the sum is 114.

Hours	Number of bodies
1	0
2	2
3	14
4	31
5	14
6	20
7	11
8	7
9	4
10	7
11	1
12	1
13	2
Total	114

- To calculate the mean, we have to take into account how many bodies fall into each time group. Label cell D1 Hours x Bodies. Begin in cell D2 and multiply the number of hours by the number of

bodies. Pull this formula down so that it is calculated for all of the time points. In cell D15, calculate the sum these values. This is the $\sum Y_i$ value you will need to calculate the mean.

- In cell D17, calculate the mean time for rigor mortis, using the formula we used in class. Type the word mean in cell C17 to help you keep track.
 - To calculate standard deviation, we will use the shortcut method. To begin we need to know the mean squared $\times n$. Calculate mean squared in cell D18 using the appropriate formula. Calculate mean squared $\times n$ in cell D19. Label these cells in column C to keep track.
 - We must continue to take into account the number of bodies for each time point. Label cell E1 observations squared. Beginning in cell E2, type an appropriate formula so that the number of hours is squared and then multiplied by the number of bodies. Pull this formula down so that it is calculated for all of the time points. Sum these values. This is the sum of the observations squared.
 - In cell D21, calculate variance using the formula we learned in class. Use this to calculate standard deviation.
- Go to Sakai and answer the next question about this frequency table.

Additional Practice

NHL Births:

The file "*NHL births.xls*" contains information on the number of NHL players born in each month of the year, as well as the number of people in Canada born in each month, for a given time period. Use the appropriate Excel formulas to calculate the total number of births in Canada for this time period, and the total number of NHL players in the data set. Calculate the proportion of Canadian births that occur in each month of the year, and the proportion of NHL births that occur in each month of the year.

Vasopressin receptor expression in voles:

The gene for the vasopressin receptor V1a is expressed at higher levels in the brain of monogamous vole species than in promiscuous vole species. To test whether gene expression of V1a influences monogamy, Lim et al (2004) experimentally enhanced V1a expression in the forebrain of the meadow vole, a solitary promiscuous species. The percentage of time each male spent huddling with the female provided to him (an index of monogamy) was recorded, and was compared to control males that were left untreated. The data are provided in the Excel file "*Vasopressin Voles*". Use the appropriate Excel formulas to calculate mean, median, standard deviation, and standard error for the two groups.

- Answer the appropriate questions on Sakai about these two data sets.

Lab 2 Part 2: Introduction to SPSS

Lab Objectives:

- ◆ Understand the logic of SPSS data files
- ◆ Create data files and enter data
- ◆ Insert cases and variables
- ◆ Merge data files
- ◆ Import data from other sources into SPSS

Exercise 1: Understanding the Logic of SPSS Data Files

Open SPSS and follow along as you read the following information.

- Each SPSS data file has two main views that can be toggled using the tabs at the lower left. Begin by clicking on the **data view** tab. Note that in data view each row typically represents the data from **one case**, whether that is a person, animal, or object. Each column represents a different variable. A cell refers to the juncture of a specific row and column. For example, the first empty cell in the right hand corner would include the data for case 1, variable 1.
- To enter data, you could simply begin typing information into each cell. If you did this, SPSS would give each column a generic label such as var00001. Clearly this is not desirable, because you would have no way of identifying what var00001 meant later on. Instead, it is best to begin by specifying names for our variables. To do this, you can double click on any column header, which will take you to the Variable View. Alternatively, you can simply click on **Variable View** on the lower left hand corner of your screen.
- The first column of variable view is **Name**. Keep names short for ease of reading. SPSS has several rules for variable names. A variable name cannot contain spaces, and cannot begin with a number. Do not end a variable name in an underscore or period, although these can be used within the variable name. You also cannot use certain keywords as the full variable name, for example, ALL, AND, NOT, OR, etc. Error messages will appear if you have selected a name that is not allowed in SPSS.
- Next is the **Type** of variable. Left click on the empty cell, then left click on the small box with dots that appears. A dialog box appears. The most commonly used types of data for this course are numeric and string.
 - For **numeric** data, width and decimal places refer to the number of characters and decimal places that will be displayed in the Data Editor window. If you enter a value with more decimal places than are set in the display setting, SPSS will save the full value, but will only display a given number of decimal places in data view.
 - The **string** setting under variable type is used for any data that is entered as text. For example, you could type Male and Female if gender were a variable of interest. It is important to note that SPSS is case sensitive meaning that “female” and “Female” would not be viewed as the same category. Misspellings are also problematic with string data

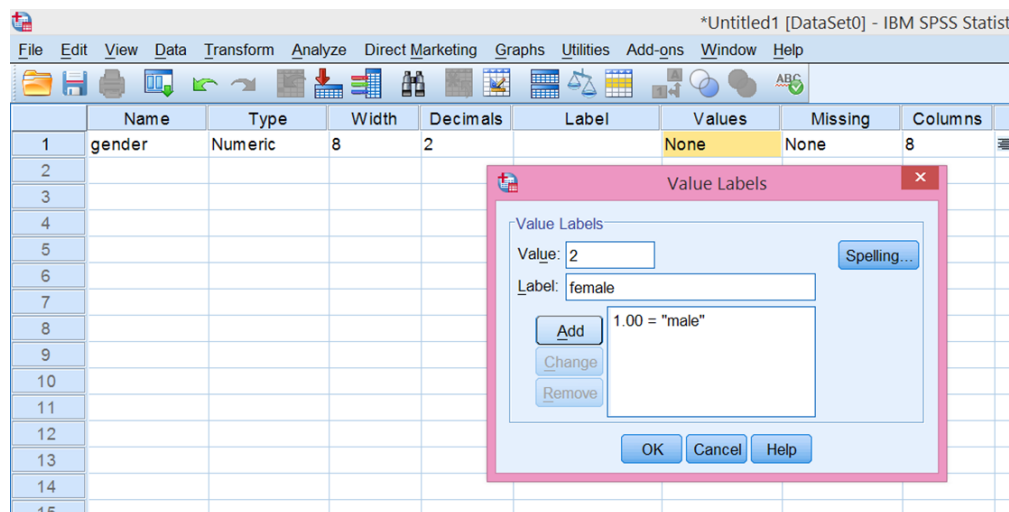
(e.g., “femal” would not be recognized as the intended “female”). For these reasons, it is often advantageous to use **numbers** to represent common categories, and then supply names for those labels, as will be discussed below.

→ **NOTE** that your selection under type does not refer to whether your variables are numerical or categorical, but rather whether the data that you input is stored as a number or text.

- The next columns in variable view are for **Width** and **Decimals**. You could have set this while specifying your variable type, or you can specify them in these columns. The default for width is 8 characters and the default for decimals is 2. To change this, left click the cell, and up and down arrows will appear. Left click the up arrow if you want to increase the number, click the down arrow to decrease the value. Alternatively, you can simply type the desired value in the cell.
- The next column is **Label**. This is a very nice feature that allows you to provide more information about the variable than you could fit in the 8 character variable name. For example, I might name a variable “depresin” and could type “Depression assessed at intake” as the full label. When you hold your cursor over a variable name in the Data View, the full label will appear. This is very useful when you need a quick reminder.
- The next column is **Values**. This allows you to assign variable labels. You will typically use this option for categorical variables. For example, we may want the number 1 to represent males and the number 2 to represent females when we enter data on gender. Let’s try this.

→ In variable view, type gender in the first **Name** column.

→ Scroll over to the **Values** column and left click. Then, left click on the ... box that appears on the right hand side of the cell. The Value Labels dialog box will appear.



→ Type 1 for value, and male for label. Click add. Repeat for females, using a value of 2. When you are done, click **Ok**.


- Skip the missing column for now. The **columns** column specifies the column width. The **align** column allows you to specify how the data will appear in the cells (justified left, right, or centre).
- The **measurement** column is where you specify whether your variables are categorical or numerical. Data with a limited number of distinct **categories** (e.g. gender, marital status, etc.) can be divided into two types. **Nominal** should be selected when there is no inherent order to the categories. **Ordinal** should be selected when there is a meaningful order of categories. For all **numerical** data (discrete and continuous) you should select the measure “**scale**”, which indicates that the data value indicates both the order of values and the distance between values.
- After you have completed specifying your variables, you can click on Data View and begin entering your data. Put your cursor on the cell in which you want to enter data. Type the value. Hit enter to move to the cell under the one you just filled, or use the arrow keys to move to the next cell in any given direction. Typically, you will either enter all of the values in one column by going down or you will enter all of the variables in a row going from left to right.

Exercise 2: Data Entry

- Open the file *depression scores.sav* and close the file with the gender variable. SPSS will ask you if you want to save the file. Click **No**. Navigate back to the *depression scores* data. As you can see, variables have been named and labeled, but the data have not been entered.
- Adjust the number of decimal places for the ID variable to zero.
- Check that the **measure** is correct for each variable indicated.
- Enter the data below into the Data View window. The data refer to depression scores at intake, then 1, 6, and 12 months post-intervention. Pay attention to your own preferences for data entry (i.e., using the arrows or enter, going across or down). Notice that there is no subject 10.
- When you are done, click **Save**, but do not close the file. We will continue to use it as an example.

ID	depressin	depress1	depress6	depress12
1	30.000	25.00	23.00	20.00
2	32.000	30.00	30.00	28.00
3	35.000	35.00	35.00	40.00
4	45.000	42.00	40.00	35.00
5	45.000	45.00	38.00	40.00
6	25.000	25.00	20.00	20.00
7	60.000	45.00	30.00	40.00
8	55.000	50.00	40.00	35.00
9	40.000	40.00	35.00	30.00
11	37.000	30.00	25.00	20.00
12	30.000	25.00	22.00	20.00

Inserting a variable:

- After specifying the types of variables for the depression data, you realize that you forgot to include a column for depress3 (depression at 3 months). A variable can be inserted in one of two ways.
 1. In Variable View, highlight the row for depress6 and then click **Insert Variable** under the **Edit** menu. This will place a new variable before the selected variable.
 2. In Data View, highlight the depress6 variable column and then click the **Insert Variable** icon . This will also place a new variable column before the selected variable.
- Use one of the approaches above to insert the new variable.
- **Name** the variable ID, and label it appropriately. Check that the **measure** is correct.
- Enter the data below.
- Click **Save**, and leave the file open.

ID	depress3
1	24.00
2	30.00
3	35.00
4	42.00
5	42.00
6	22.00
7	35.00
8	45.00
9	36.00
11	28.00
12	24.00

Inserting a Case:

- As you can see, the data for ID 10 is missing. Imagine that this missing data has been located and you want to enter it in the file. To keep the data in order, you want to insert a case between the person with ID 9 and ID 11. To do so, you can highlight the row for the case with ID 11, and either:
 1. Select **Edit** and then **Insert Case**, or
 2. Click on the **Insert Case** icon shown below. In either case, a blank row will appear before the highlighted case. Try it yourself.



- Insert a case for ID 10 using one of the above approaches.
- Enter the following data: 10, 38, 36, 35, 38, 38 for ID, depresin, depress1, depress3, depress6, and depres12 respectively.
- Check the accuracy of your data entry, then click **Save**.

Merging Files: Adding cases

Sometimes data that are related may be in different files that you would like to combine or merge. In this case, each file contains the same variables but different cases. Imagine that we want to add the information for four additional patients (currently saved in a separate file) to our depression scores file.

- With *depression scores.sav* open, click on the **Data** menu, then **Merge Files**, and **Add Cases**.
- Browse to the file *new patient scores.sav*. Select **Open**, then **Continue**. A dialog box will appear showing you which variables will appear in the new file. If all variables in the two files are identical, they will appear in the “variables in new active dataset” box. Click **OK**.
- View the new patient list, and ensure that the files were merged correctly.

Merging Files: Adding Variables

In other cases, you might have different data on the same cases or participants in different files. For example, the demographic information from the participants in the depression study may be saved in one file and the depression data in another file. You may want to combine these to see if demographic variables, like socioeconomic status or gender are related to depression. In this case, you need to be sure the variables on the same participants end up in the correct row so that they match up correctly. In this case, we will use ID to match cases. SPSS requires that the files you merge be in ascending order by the matching variable. So, in both files, ID must start at 1. If the cases are not in the correct order you can sort them, as described below.

- Keep *depression scores.sav* open.
- Check to see if the cases are in ascending order by ID.
- Now, open *depression demographics.sav*. These data are not in order by ID. To fix this, click **Sort Cases** under the **Data** menu.
- In the dialog box, select participant identification number and move it into the **Sort by** box by clicking the arrow. Make sure **Ascending** is selected for **Sort Order**. Then click **Ok**.
- While the demographic file is still open and active, click on **Merge Files** in the **Data** menu, and select **Add Variables**.

- The next dialog box will ask you to indicate which file the new variables are coming from. Select *depression scores.sav* and click **continue**.
- Select **Match cases on key variables** and **Cases are sorted in order of key variables in both datasets**. Then highlight id under excluded variables, and click the arrow next to **Key Variables** to move ID into that box. Click **Ok**.
- Check that the files have been merged.
- Click on **variable** view. Add values for gender so that we know that 1 is used for males, and 2 is used for females. Check that the “measures” for each variable are correctly indicated. There are some that should be changed.
- Return to the data view. Note that you can click on the following icon so that the value labels will be displayed, rather than values. If the labels are already displayed, this button will switch the display back to values.



- You may want to do a **Save As** and give the merged file a new name like *depression merged.sav* to help you remember what is in it.
- Go to Sakai and answer the questions about the merged depression data.

Exercise 3: Reading Data In From Other Sources

SPSS can also recognize data from several other sources. For example, you can open data from Microsoft Excel or CSV in SPSS. This is an attractive option, particularly if you do not have your own version of SPSS.

Opening data from EXCEL:

- Open the file *Spider amputation.xls* in Excel and look it over. You will see the file contains the data from the experiment we looked at in class measuring spider speed before and after pedipalp amputation. Close the file.
- In SPSS, select file, **open data**.
- A dialog box will appear. Under **Files of type**, select **Excel**. Navigate to *Spider amputation.xls*, select it, and click **Open**. A new dialog box will appear.
- Select **Read variable names from the first row of data**, because that is where the names appear in the Excel file. Then, click **Ok**.

- Check out your new file in SPSS. How does it look? Note that the variables are labeled a bit differently due to the spaces being removed. Click on variable view, and note that SPSS may have attempted to determine the measure for each variable, and that some labels are missing. You will always need to check that variables are identified correctly when importing from a spreadsheet.
- Go to Sakai and answer the question about the spider data.

Additional Practice

Bird Malaria

Import the data from the file “*Bird malaria*” into SPSS. This file contains the results from the study we looked at in class, in which the incidence of malaria was examined in birds that had eggs removed from their nest, compared to control birds. For the treatment variable, a value of 1 was used to identify control birds, and a value of 2 was used to identify birds who received the egg removal treatment. The response of the birds was recorded as 1 for no malaria, and 2 for malaria. This is important information to save in the data file. Add in this information using variable view, and check that it displays correctly in the data view.

Maternal mortality ratio & births attended by skilled health staff:

Import the data from the Excel files “*Maternal mortality 1998*” and “*Skilled health workers 1998*” into SPSS. Check that the variables are correctly labelled and that the data is in ascending order by country name. View 1 decimal place for skilled workers, and 0 decimal places for maternal mortality. Merge the two files in SPSS.

Visit the data section of the World Bank website, sorted by indicator:

<http://data.worldbank.org/indicator>

Click on Maternal mortality ratio (modeled estimate, per 100,000 live births). Download this data in Excel format. Return to the data section of the website and find the indicator Births attended by skilled health staff (% of total). Download this data in Excel format. Follow the appropriate steps to import the data from the year 2006 **only** for both variables into the same SPSS file as the 1998 data.

- Go to Sakai and answer the questions about the data.