APPENDIX E

HOW TO USE A DATA SPREADSHEET

Excel

One does not necessarily have special statistical software to perform statistical analyses. Microsoft Office Excel can be used to run statistical procedures. Although in some respects Excel is not as preferable for data analyses as IBM SPSS, it is very user-friendly with simpler statistical procedures. This appendix describes how to use Excel to execute basic statistical calculations. Data from the 2004 version of the General Social Survey (GSS) is used for examples. This Appendix is based on Excel 2007 version, which differs in certain aspects from Excel's previous versions. The most notable change that affects the exercises presented in this appendix concerns the pivot table feature.

BASIC PROCEDURES

Starting Excel

To start Excel using Windows, click on the **Start** button at the bottom left corner of the screen. Under **Programs** locate and click the **Microsoft Excel** icon.

The layout of the Excel program has changed substantially for the Microsoft Office 2007 edition compared to its predecessors. Commands are now grouped in ribbons that are accessed by clicking on a specific tab. Thus, the **Home** tab grants access to a ribbon of several command groups: **Clipboard**, **Font**, **Alignment**, **Number**, **Styles**, **Cells**, and **Editing**.

Once the program is started you will see a Worksheet Area that consists of cells forming columns and rows. Rows are identified by numbers, and columns are identified by letters. Consequently, each cell has its own unique address—a combination of letters and numbers. For example, cell C6 is in column C, row 6. The dark rim around a cell means that the cell is highlighted or active. You can highlight a range of cells by clicking and dragging the cursor across several adjacent cells.

At the bottom left of the screen you will find worksheet tabs labeled Sheet 1, Sheet 2, etc. You can rename these worksheets, add additional ones, or delete ones you do not need.

Opening data

The data you need to use might be saved in a format other than Excel (file.xls). The data file we will be using in this tutorial is saved in SPSS format (file.sav). This is not a problem. Open the data file in SPSS. Click **File—Save As**—choose *Excel 97 and later (.xls)* in the **Save as Type** drop down menu. The file is ready to be open in Excel. To do so—either double-click

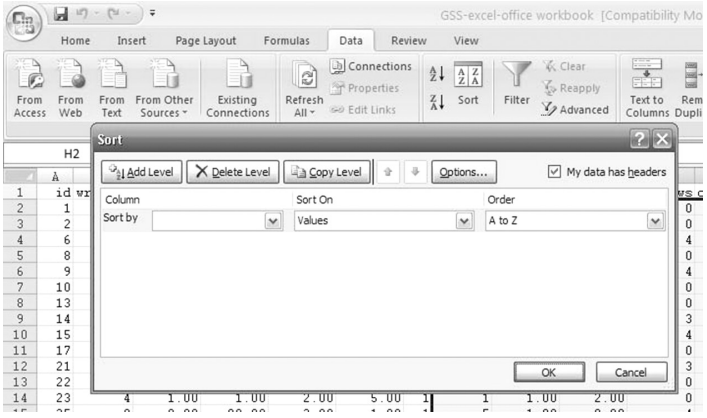
on the icon of the saved Excel file or, if Excel program is already open, click **Office** button at the top left of the screen—**Open**—locate the data file you need to work with.

Looking at the Data

Once the data is open you will discover that the view is somewhat different than what you saw in SPSS program. Variable names are displayed in the first row (row A). Unfortunately, there is no variable view worksheet in Excel, and transferring the file from SPSS to Excel results in a loss of variable labels and value labels. Keep the codebook for the data on hand!

Sorting Data

As you are exploring the data you might want to take advantage of the Sorting tool, which allows you to sort data by two or more variables in ascending or descending order. To sort your data, click the **Data** tab, find the **Sort and Filter** command group, and click on the **Sort** icon. A dialog window will open where you will specify the variables and order you want your data to be sorted by. If the names of variables (variable labels) are entered in the first row of your data make sure you check the "*My data has headers*" box at the top right of the dialog window. You can sort numeric data as well as text (in A to Z order). If you want to sort by more than one variable click **Add Level** button at the top of the dialog window. Click **OK**.

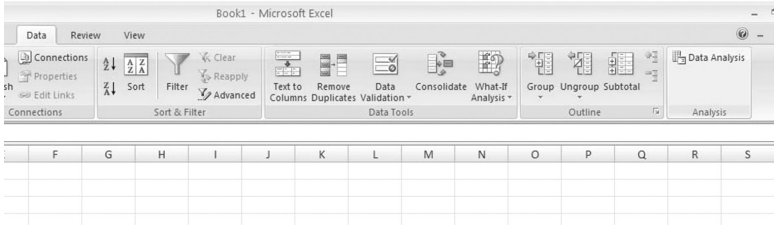


Missing Data

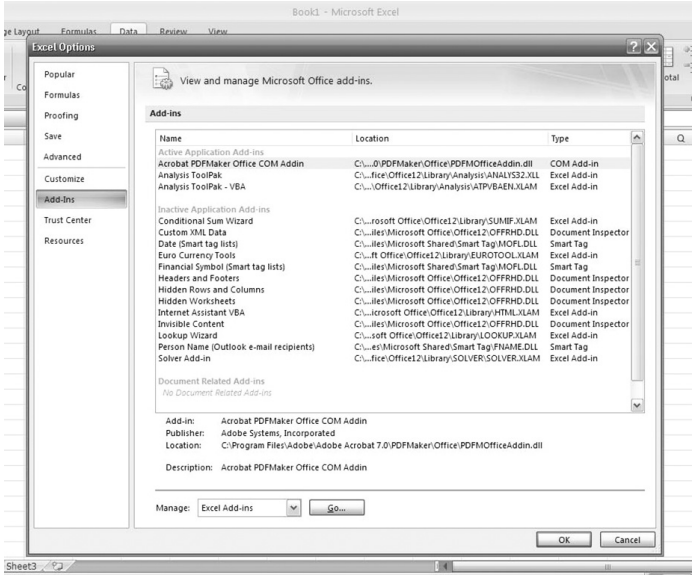
If you have converted an SPSS file to be used in Excel, be mindful of how missing data was coded in the original file. If missing data was entered as blanks, you have nothing to worry about. However, if missing data in the original data file was entered as a certain numerical expression (e.g. zeroes or 99) this might create problems for running certain statistical calculations in Excel (e.g. computing mean).

Loading Data Analysis Toolpak

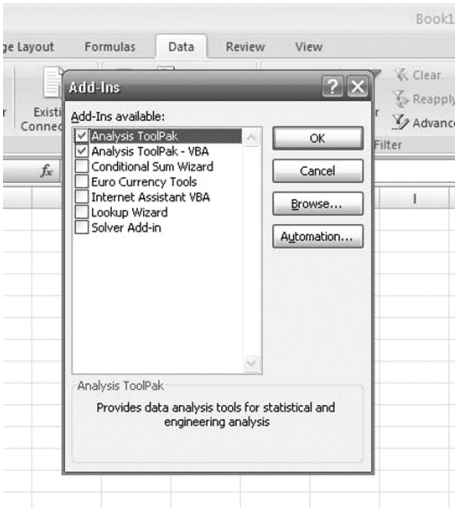
Before we start with running some basic analysis make sure that Excel's Data Analysis Toolpak is loaded on your computer when you start the program. The Data Analysis Toolpak comes as an Add-In to your Excel program and is found on the installation disk. To see if the Toolpak is loaded in Excel 2007, check if there is an Analysis icon under the Data tab.

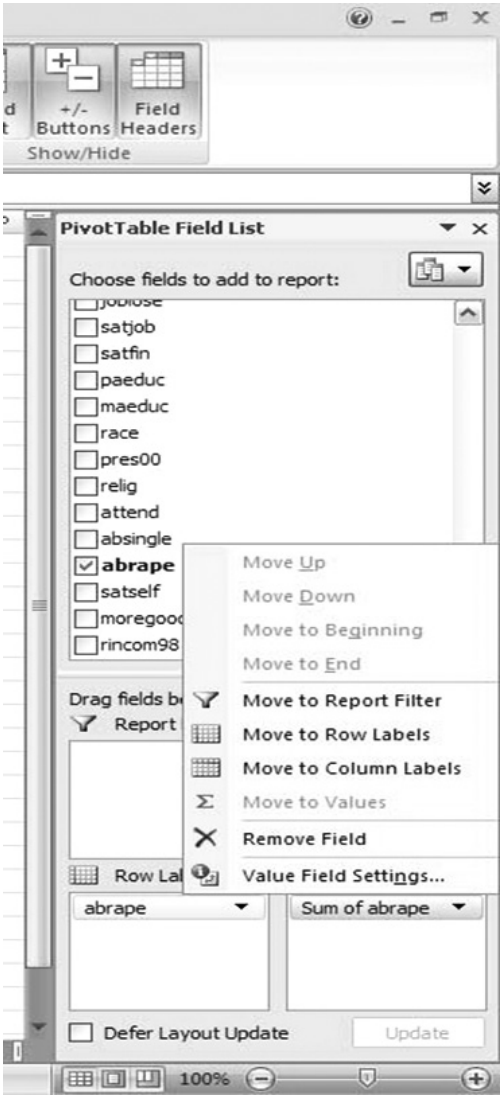


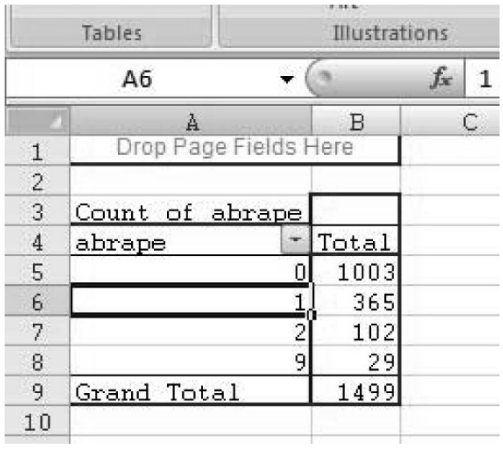
If you do not see one, click **Office** button—**Excel Options**—**Add-Ins**. In the dialog window that opens check if Analysis Toolpak and Analysis Toolpak-VBA are listed under *Active Application Add-ins*.



If they are listed as inactive select both of them and click **GO** at the bottom of the window. Next, check the boxes for both Add-Ins in the dialog box that opens. Click **OK**.







UNIVARIATE STATISTICS

Frequencies

The easiest way to create frequency distributions in Excel is by using the Pivot Table. To create a frequency distribution of the variable ABRAPE (pregnant as a result of rape) click on any cell in your data. Next click **Insert Tab - Pivot Table**. In the opened dialog window you will see a *Table/Range specified* that includes the call range for all of your data. It is also possible to specify the range manually by typing A1:AW1500. Next, lace your Pivot Table: Click **New Worksheet—OK**.

A new worksheet will open with a newly created Pivot Table. In a **Pivot Table Field List** on the right side choose and click the variable we are interested in – ABRAPE. Do not be alarmed that your newly created Pivot table displays the sum of all the ABRAPE values in the data. Since you are interested in the count or each of the values of the variable and not the sum of the values, click a dropdown menu located under the **Pivot Table Field List** and make sure that the **Fields Section and Areas Section Stacked** is selected. Next click on the ABRAP variable and drag it down to the **Row Labels** box. Next click the dropdown menu to the right of Sum of Abrape—**Value Field Settings**—change **Sum** to **Count—OK**. Now you have a table of frequency distributions for the variable ABRAPE. Unfortunately, Excel does not display value labels like SPSS does, so your table can look a little bit confusing with numbers in place of the actual values of the variable (e.g. "1" instead of "Yes"). To rectify this problem, you can either correct the codebook for the data or recode the variable prior to creating a Pivot table (substituting numerical values with text).

Excel does not offer an option to simultaneously display both the count and percentage within the same Pivot table. However, you can view the percentage distribution by clicking **Value Field Settings—Show values as**—choose **% of total**.

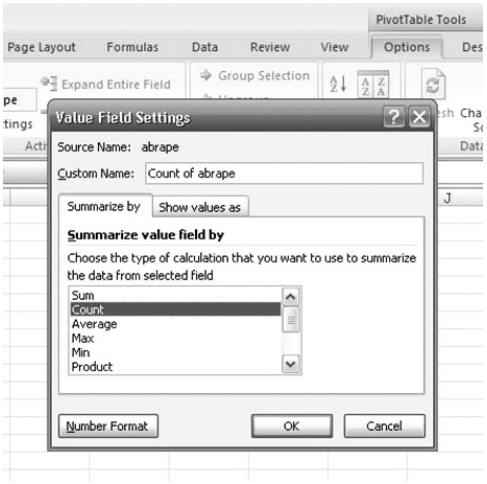
Descriptive Statistics

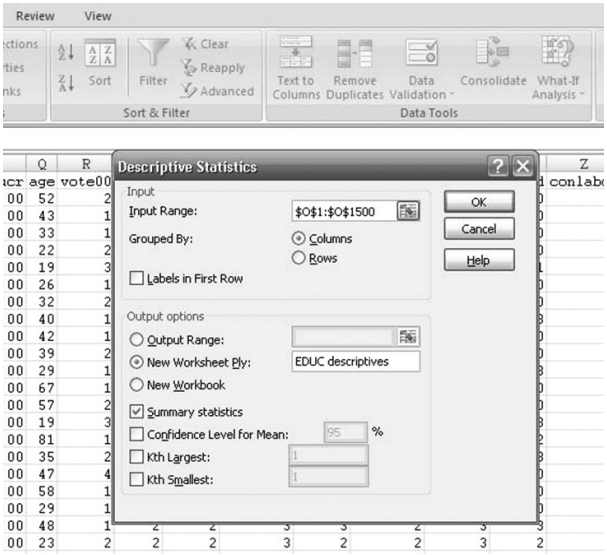
The easiest way to obtain descriptive measures on a variable in Excel is by using the Data Analysis Tool. Under the **Data** tab click **Data Analysis - Descriptive Statistics—OK**. In the descriptive statistics dialog window specify the range of the variable you are interested in. Variable EDUC occupies the range O1: O1500. You can either specify the range manually by typing it into the *Input Range* Dialog box or highlight the column that the variable occupies in the dataset.

Excel offers the option of grouping each variable in its own column or its own row, with former being the default option. Click *Labels in First Row* option, or you are risking receiving an error message "Input range contains non-numeric data" since the program will get confused with the name of the variable being the first cell in the column.

Next you can choose to either place the output into a new worksheet (default option), new workbook, or in the same worksheet as the data (Output Range). In the later case you have to specify the upper left cell "address" where you want your output to be pasted. Click the *Summary Statistics* box—**OK**. You will be presented with information on a range of statistical measures: mean, standard error, median, mode, standard deviation, sample variance, kurtosis, skewness, range, minimum, maximum, sum, count.

If you are interested in obtaining only a certain descriptive statistic you can take advantage of Excel's functions instead of using the Descriptive Statistics tool. For example, let's calculate the Mode of the EDUC variable. Either below or to the side of your data find some empty cells. In one of the cells type EDUC Mode. Then activate the cell immediately to the right, click **Formulas—Insert Function**. In the dialog window specify *Statistical* and then choose **MODE—OK**. In the *Function Arguments* window specify the range of the variable of interest (EDUC occupies O2:O1500). Click **OK.**





	A	B
1	Frequency distribution	
2	Count of abrape	
3	abrape	Total
4	YES	365
5	NO	102
6	N/A	29
7	NAP	1003
8	Grand Total	1499
9		
10		
11	Count of abrape	
12	abrape	Total
13	YES	24.35%
14	NO	6.80%
15	N/A	1.93%
16	NAP	66.91%
17	Grand Total	100.00%
18		

Graphing

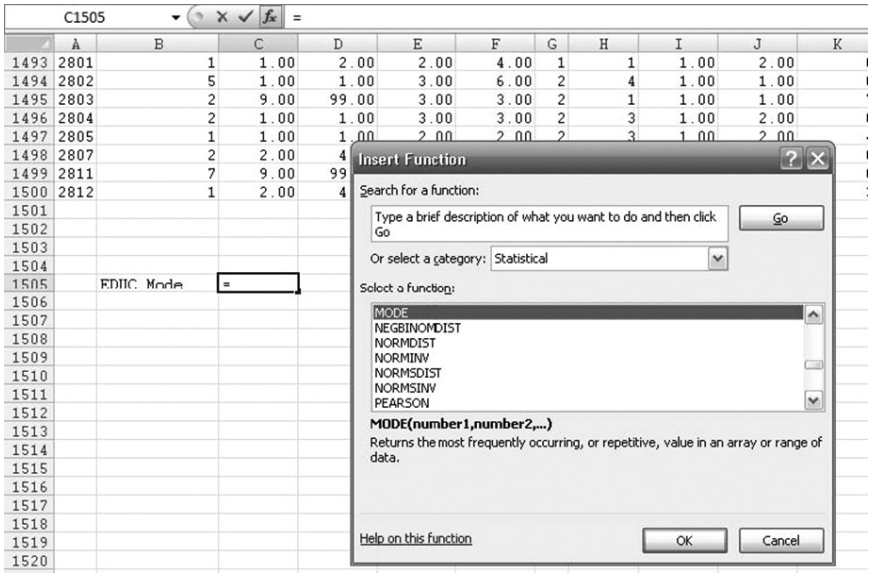
There are two ways to create graphical representation of your data in Excel: using the Pivot Table or the Histogram Analysis Tool. First, let's try using the Pivot Table option. Follow the steps to create a frequency distribution Pivot Table for the variable ABRAPE. It is very easy to create a bar chart from here. Click on any cell in the Pivot Table, then click **Insert tab—column—2D column**. You can format your chart using a **Chart Styles** section under the **Design Tab.**

Now let's practice creating a histogram using the Histogram Analysis Tool. In order to create a histogram, we have to specify "bin values" that represent the entire range of values of your variable. By specifying bins we are creating slots to gauge how many times a specific value appears in our data. Our histogram is supposed to convey the age distribution of our respondents. First, create a frequency distribution of the respondent's age variable. You will see that in our dataset the youngest respondent is 18 and the oldest is 89. Let's specify eight

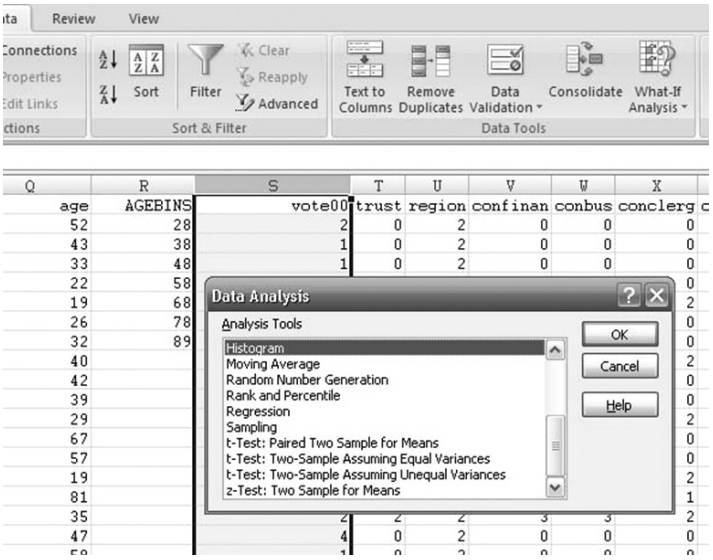
A1

	A	B
1	*educ*	
2		
3	Mean	13.71047
4	Standard	0.091856
5	Median	13
6	Mode	12
7	Standard	3.556369
8	Sample Va	12.64776
9	Kurtosis	220.2888
10	Skewness	9.109408
11	Range	99
12	Minimum	0
13	Maximum	99
14	Sum	20552
15	Count	1499
16		
17		

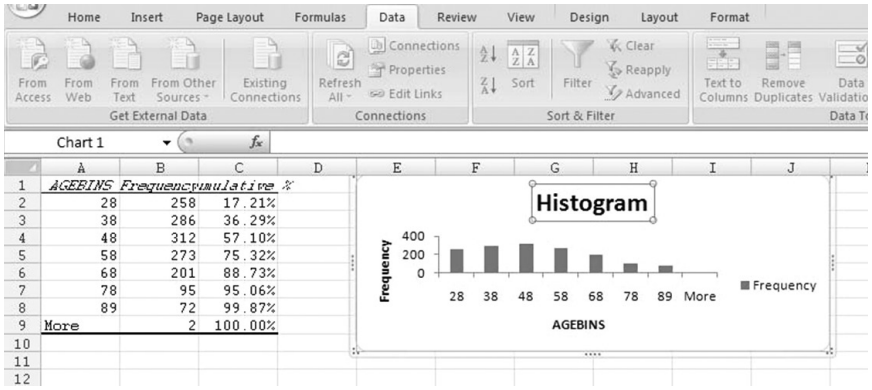
bins—28, 38, 48, 58, 68, 78, and 89—with each bin being the upper limit for a particular age group. Insert a new column in you data, title it AGEBINS, and enter the specified numbers.



Click **Data—Data Analysis—Histogram—OK**.



In the dialog window specify the *Input Range* (the range of the variable you are interested in) and the *Bin Range* (cell range where the bins values are specified). Make sure that the *Labels* box is checked. Choose the location of your output histogram (*New Worksheet Ply* is default). Check *Chart Output*. Click **OK**.



Your histogram will open in a new worksheet. It does not look perfect, but you can modify it to make it look better. First, make the histogram taller by clicking on it and dragging its lower border down. Second, change the axis titles and the title of the histogram. Third, you can remove the space between the bars: right-click on one of the bars and select **Format Data Series**, then drag the *Gap Width* cursor all the way to the left. The possibilities of formatting a chart in Excel are endless.

RECODING VARIABLES

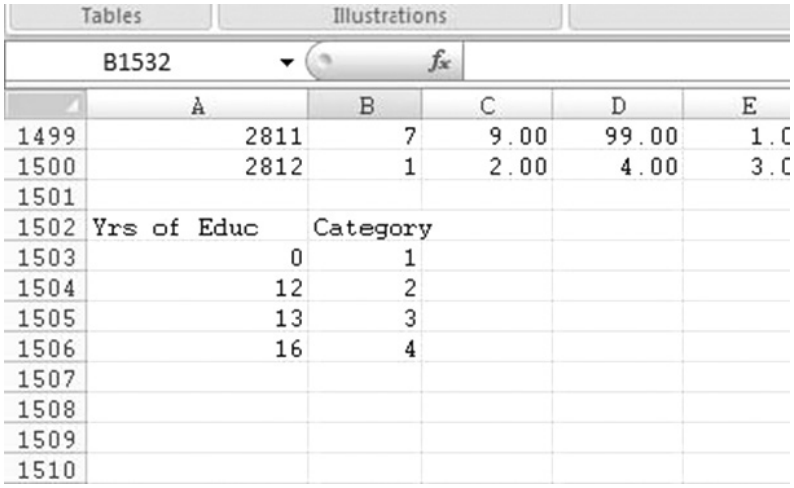
Recoding EDUC (highest level of education completed) variable. To recode EDUC into a new categorical variable you have to create a reference table that lists your new categories. Find a range of empty cells below your data and create a reference table where you specify:

Those who have 0-11 years of education are put in category 1

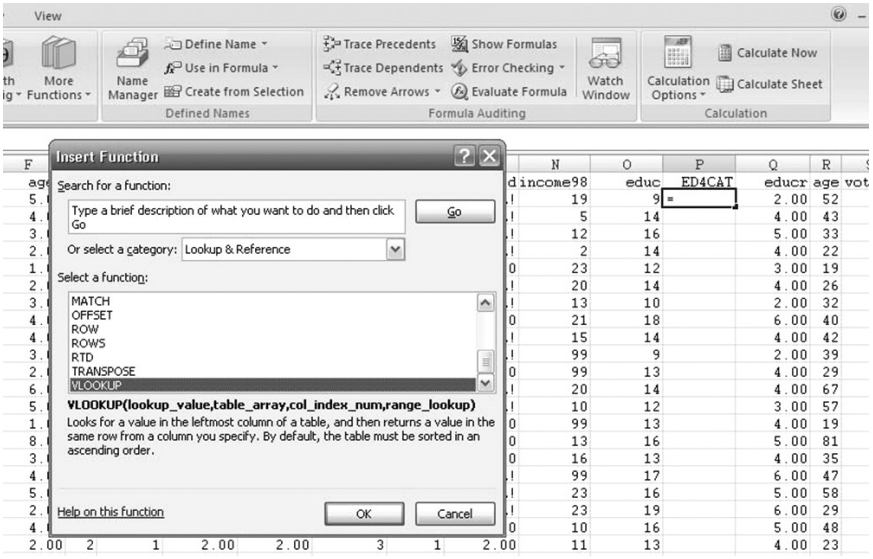
Those who have exactly 12 years of education put in category 2

Those who have 13 to 15 years of education are put in category 3

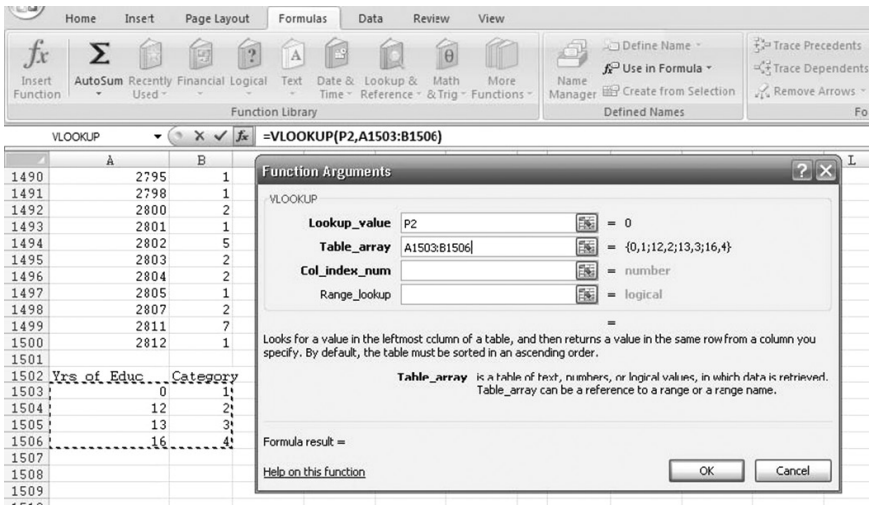
Those who have 16 or more years of education are put into category 4



Next insert a row next to the EDUC variable, this is where you will compute your new variable—ED4CAT. Highlight the cell right under the variable label. Click: **Formulas—Insert Functions**. Select *Lookup & Reference category* and choose **VLOOKUP** function in the menu below. Click **OK**.



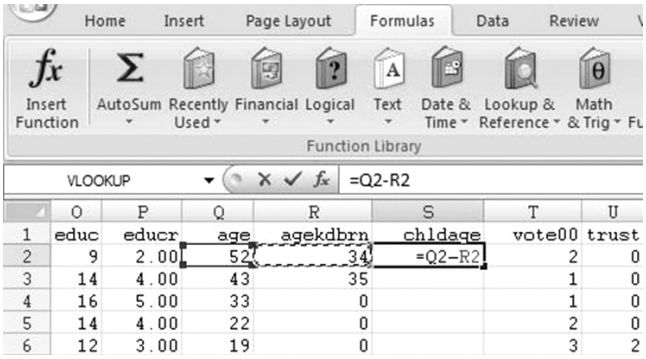
In the dialog window first specify the *Lookup_value*. The lookup value is the value of your original variable you want to look up and replace with some value of a new variable. *Table_array* is the location of the reference table we created to lookup values for the new variable. You can either highlight the location of the reference table with your mouse or type the cell range manually. *Col_index_num* refers to the position of the reference table column we want to look up the new variable values from. In our case the column that contains all the potential values of the new variable is column number 2 (column B). Click **OK**.



This gives you your first value of the new variable ED4CAT. To avoid repeating the above procedure for each case in the dataset, simply copy the formula from the cell just completed to other cells downward by clicking on the right bottom corner and dragging the mouse cursor down.

Computing a New Variable

It's quite easy to compute a new variable in Excel. All it takes is recollection of a few basic algebra rules. For example, to determine the age of each respondent's eldest child we need to compute the difference between the respondents' age and the age of the respondent when his or her first child was born from the respondent's. Let's start by inserting a new column anywhere in the worksheet and typing the name of our new variable (CHLDAGE) in the first row. In the empty cell right below type a formula for calculating the age of the respondent's eldest child. This formula will contain the "addresses" for cells that specify (1) the respondent's age (AGE) (Q2 in our example); and (2) respondent's age when his or her first child was born (AGEKDBRN) (R2). Our formula will be: **= Q2 – R2**. Press **Enter**, and there you have it the first respondent's eldest child is 18 years old. Copy the formula for the rest of the respondents.



	O	P	Q	R	S	T	U
1	educ	educr	age	agekdbrn	chldage	vote00	trust
2	9	2.00	52	34	=Q2-R2	2	0
3	14	4.00	43	35		1	0
4	16	5.00	33	0		1	0
5	14	4.00	22	0		2	0
6	12	3.00	19	0		3	2

BIVARIATE AND MULTIVARIATE STATISTICS

Crosstabulation

Exploring the relationship between two or more variables in Excel is done with the help of pivot tables. You will not find a Crosstab option like in SPSS. Let's say we are interested in the relationship between individual's level of education and whether or not he or she participates in elections. We will create a pivot table for two variables: EDUCR3 is a trichotomized measure of education and variable VOTE00, which assesses whether the respondent participated in 2000 Presidential elections.

Click **Insert** X **Pivot Table**. Select data range and the location for the Crosstab table (new worksheet). Click **OK**. On the right side of the new worksheet, in the *Pivot Table Field List* area first select **Field Section and Area Section Side-By-Side** option for more convenient display, then select (check the boxes) the two variables of interest VOTE00 and EDUCR3. Place VOTE00, the dependent variable, in the *Row Labels* box, and EDUCR3 in the *Column Labels*. Drag VOTE00 into the *Values* box, and change the **Value Field Settings** to ***Count*** instead of ***Sum***.

Your crosstab will look a little confusing due to presence of numbers instead of actual variable values. If you have the codebook for your data on hand, you can quickly change this manually. For example, we know that a value of 1 is assigned to the VOTE00 variable if the respondent voted in the 2000 Presidential elections. We can change the cell accordingly.

Excel does not offer an option to simultaneously display both the count and percentage within the same crosstab. You can observe percentages (of rows or columns) by clicking **Value Field Settings—% of column—OK**.

	A	B	C	D	E
1		Drop Page Fields Here			
2					
3	Count of vote00	educr3			
4	vote00	Grade School	High School	Some College	Grand Total
5	Voted	34.27%	63.68%	69.99%	63.22%
6	Did not vote	51.64%	29.35%	22.42%	28.44%
7	Ineligible	12.21%	5.72%	7.13%	7.48%
8	Refused to answer	0.47%	0.00%	0.11%	0.13%
9	Don't know/remember	1.41%	1.24%	0.34%	0.73%
10	Grand Total	100.00%	100.00%	100.00%	100.00%
11					
12					
13					
14					
15					
16	Count of vote00	educr3			
17	vote00	Grade School	High School	Some College	Grand Total
18	Voted	73	256	618	947
19	Did not vote	110	118	198	426
20	Ineligible	26	23	63	112
21	Refused to answer	1		1	2
22	Don't know/remember	3	5	3	11
23	Grand Total	213	402	883	1498
24					

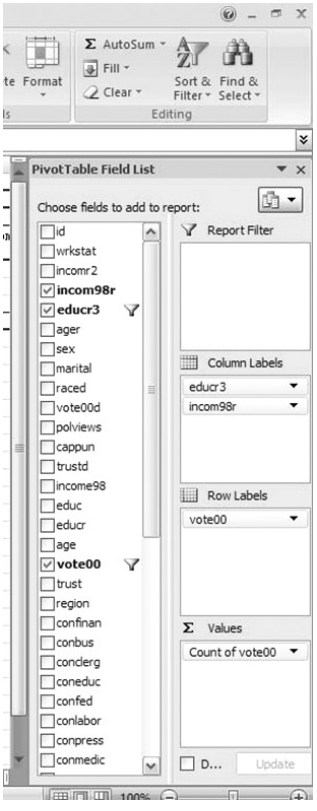
Since there were so few people that either refused to answer the question about voting or cannot recollect whether they voted, you have an option of dropping these two rows from the table (this will, of course, change the grand total counts, so you might not want to do this). Click the little arrow next to the VOTE00 cell in the table and uncheck these two rows. Click **OK**.

Three-Variable Crosstabs

It is very easy to convert the crosstab you have created into a three-variable crosstab.

The two-variable cross-tab we have created demonstrated that individuals with higher levels of education (those who have completed high school and/or some college) were more actively involved in voting in the 2000 Presidential election than individuals who have completed grade school. But perhaps education is not really a moving force behind

individuals' propensity to be politically active, and the relationship we observed is due to an extraneous factor. To explore the relationship between voting and another variable while controlling for education all we have to do is add a third variable to the pivot table (to the column labels).



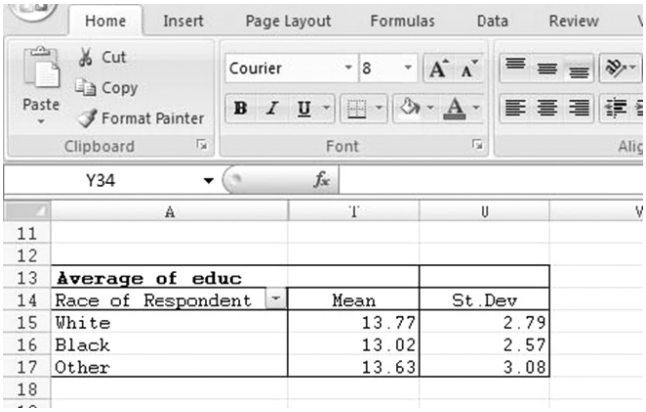
Comparing Means

Comparing means in Excel once again requires the use of Pivot tables. Let's examine whether individuals differ in terms of years of education by race. Click:

Insert—Pivot Table—[make sure the cell range for your data is specified in the *Table/Range* box]—**OK**.

In the *Pivot Table Field List* area check and drag the variables you are interested in (EDUC and RACE) into *Columns* and *Row Labels* boxes respectively. With your dependent variable (EDUC) also being in the *Values* box. Click **Value Field Settings—Average—OK**. This will produce a pivot table that shows mean years of education for each group within the race variable. Let's type the value labels of the race variable instead of numbers.

To accurately assess the difference between the mean years of education of Whites, Blacks, and those individuals who comprise the "Other" racial category, we need to calculate standard deviations by racial groups. This is also done in the pivot table. Let's copy the pivot table that we just created somewhere on the worksheet you have open. Now, all you have to do to obtain standard deviations by group is click **Value Field Settings** and choose **StdDev** under the *Summarize by* tab. In the figure below we have copied and pasted the Standard Deviations column adjacent to the Means column.



PRINTING

To print portions of Excel worksheets click the **Office** button—**Print**. Before printing, the **Page Setup** button under the **Page Layout** tab allows you to manipulate the orientation of the page (portrait or landscape), the margins of the page, rows and columns you want to print, as well as printing the output with or without gridlines.