EXCEL
Microsoft Excel
Microsoft Corporation

Exercise 4. Analyzing Census Data in Excel

Purpose: Because Census data typically is provided in a spreadsheet form and usually in *Excel* format this exercise will cover some of the basic operations necessary to use *Excel* to analyze Census data..

Examining an Ancestry

1. Open the ancestry table (*STF3 PCT16*) that you created earlier in Ex. 2.

If you do not have the table, go to the Census web site and use *American Factfinder* to select *Summary File 3* and then table *PCT16* for all California counties.

B1 GEO_ID2

	A	B	C	D	E	F	G
1	GEO_ID	GEO_ID2	SUMLEVEL	GEO_NAME	PCT016001	PCT016002	PCT016003
2	Geography Identifier	Geography Identifier	Geographic Summary Level	Geography	Total population: Total	Total population: First ancestry reported	Total population: First ancestry reported; Acadian/Cajun
3	05000US06001	06001	050	California	1443741	1261632	79
4	05000US06003	06003	050	California	1208	1058	2
5	05000US06005	06005	050	California	35100	25683	5

2. When you open the data table which will probably be labeled *dt_dec_2000_sf3_u_data1.xls* unless you changed it, note the various columns. Under GEO_NAME many of the county labels are hidden because of the narrow column width. Expand the column width to see all the names.

3. Scan across the columns and find an ancestry group you would like to examine in more detail. Avoid very small groups like *Acadian* since their low numbers are difficult to assess. Also, some categories like *Arab* are summaries of individual groups that follow.

When you have located a group, copy *GEO_ID2, GEO_NAME, PCT016001*, and your group to a new spreadsheet.

4. Compute the percent of the total population for each county represented by your group.

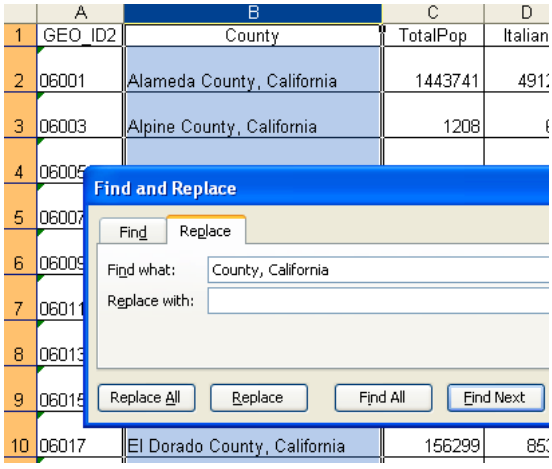
5. Change the column labels to eight character variable names and then delete the lengthy names in row 2. To delete the row, click on the number 2 to highlight the entire row. Then select *Edit > Delete*.

	A	B	C	D	E
1	GEO_ID2	County	TotalPop	Italian	PctItal
2	06001	Alameda County, California	1443741	49120	3.4
3	06003	Alpine County, California	1208	60	5.0
4	06005	Amador County, California	35100	2262	6.4
5	06007	Butte County, California	203171	8941	4.4
6	06009	Calaveras County, California	40554	2338	5.8

In the figure above persons of Italian ancestry have been selected. Note in Column *B* the repetitious label of *County, California*. Since the file contains only counties we will delete the excessive wording.

6. Click on the top of Column *B* so that all of its cells are selected.

7. Select *Edit > Replace* and the window below will open. Enter *County, California* in the *Find what*: space. Click *Replace All.*



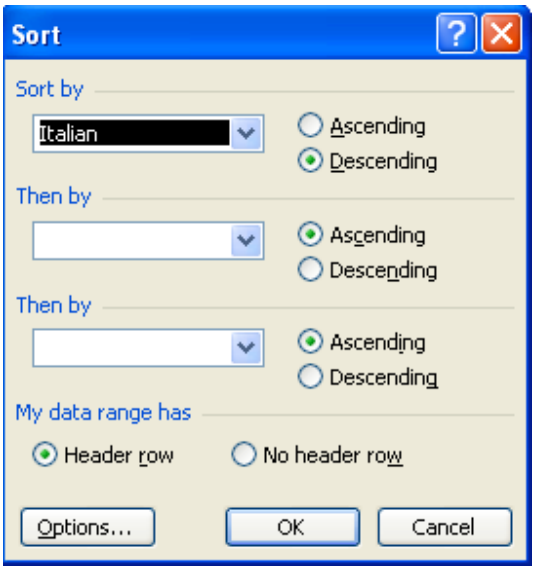
Now your labels are much more practical.

	A	B	C	D	E
1	GEO_ID2	County	TotalPop	Italian	PctItal
2	06001	Alameda	1443741	49120	3.4
3	06003	Alpine	1208	60	5.0
4	06005	Amador	35100	2262	6.4
5	06007	Butte	203171	8941	4.4
6	06009	Calaveras	40554	2338	5.8

These mass edits are frequently useful in creating useable labels when you have many values to process.

8. Select *Data* > *Sort*.

When the *Sort* window opens locate your group population in the *Sort by* window and then select the *Descending* button. Click *OK*.



This is a very useful function for isolating very high and low numbers in a distribution. Note that you can sort on more than one variable so, for example, you could sort by state and then by county within a state.

	A	B	C	D	E
1	GEO_ID2	County	TotalPop	Italian	PctItal
2	06037	Los Angeles	9519338	214671	2.3
3	06059	Orange	2846289	107878	3.8
4	06073	San Diego	2813833	105667	3.8
5	06085	Santa Clara	1682585	71793	4.3
6	06013	Contra Costa	948816	53527	5.6

9. What are the leading counties with your ethnic population?

10. What are the counties with the fewest number of your ethnic group?

While it is interesting to note the numbers of ethnic persons it soon becomes apparent that in many cases the counties with the highest numbers ethnic also have the most people. Thus, we might want to control for population differences to see where the proportion of ethnic persons is greatest and least.

11. Select *Data* > *Sort*.
This time sort by the percent ethnic column.

	A	B	C	D	E
1		County	TotalPop	Italian	PctItal
2	06097	Sonoma	458614	34632	7.6
3	06041	Marin	247289	17654	7.1
4	06081	San Mateo	707161	46616	6.6
5	06005	Amador	35100	2262	6.4
6	06091	Sierra	3555	218	6.1

Sorting data this way is more interesting since it reveals something about the proportion of ethnic settlement. It seems that Italians are concentrated near San Francisco Bay and in nearby counties associated with wine grape growing.

One county, Sierra, has a relatively high percent Italian, but the numbers involved are quite low. Thus, in addition to examining the percent ethnic, it is important to also consider the magnitude of the numbers behind the percents.

12. Look over the percentages for your ethnic group. Do you know of any reasons why the percentages are high in some counties? You could check the Internet or your library for the history of settlement for your group.

For even more detail you can prepare a graph of your distribution. Look over the next section to see about making a graph.

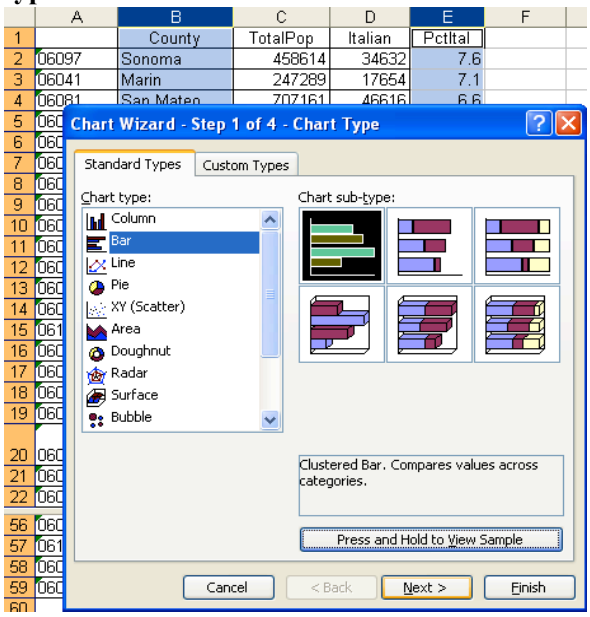
13. Save your table for use later.

Graphing an Ancestry

While sorting data does give us access to specific information, graphing and mapping the data helps us get a sense of the general trends in value and location of the counties.

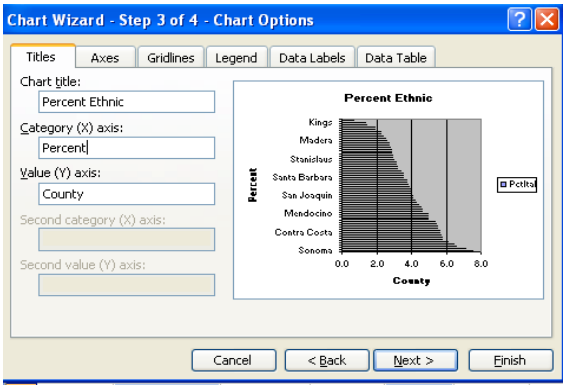
1. Click on Column *B* (*County*) and while holding the *control* key, click on Column *E* (*Pct Ethnic*) in your ethnic table.

2. Select *Insert* > *Chart*. This will begin a "wizard" with four screens to help you design the chart. The first screen is shown below. From the first window select the *Bar* type and click *Next*.



3. From the second screen select Next.

4. In the third step enter *Percent* in the *Y-axis* category and enter *County* in the *X-axis* category. (This seems the reverse of the display, but it may be a carry over from the *Column* chart design.)

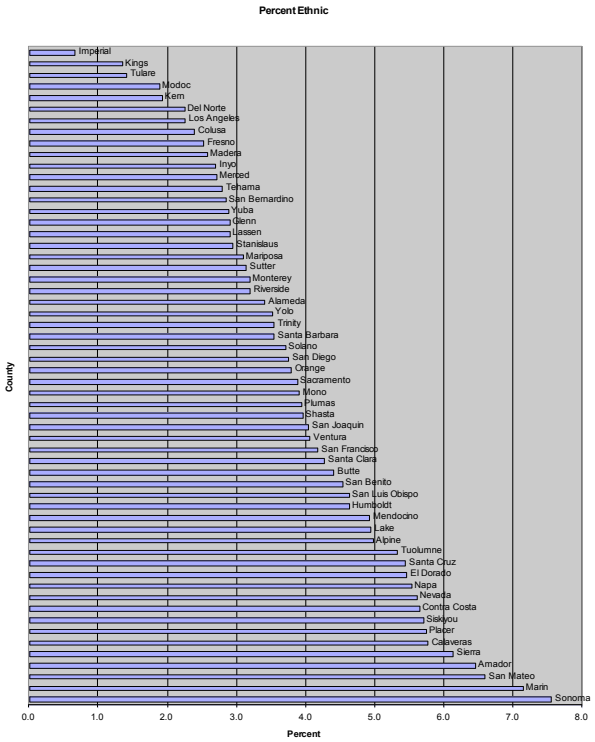


Select the *Legend* tab and deselect the *Show Legend* button. It is not needed for just one variable.

Select the *Data Labels* tab and click the *Category Name* button.

Select the *Axes* tab and deselect the *Category (X) axis* button. Then click *Next*.

5. In the final step click on the *As New Sheet* button and then *Finish*. Your chart will appear as a new screen.



If your page layout is set to *Landscape,* select *File* > *Page Layout* and click the *Portrait* button to better visualize the distribution.

You can make various cosmetic changes to the look of your graph by double-clicking on the axes, lines, bars, or titles.

Note that a new tab, *Chart 1*, has been added to the bottom of your *Excel* window.

Look over your chart. Are there any extreme or unusual values? Any sudden changes in trends?

A Frequency Graph of Ancestry

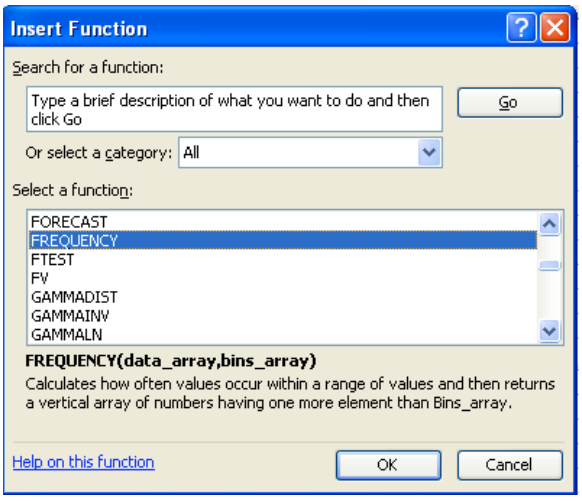
Unfortunately *Excel* does not provide a tool to create a frequency graph. This graph shows the number of observations (counties) that fall within various categories that extend from the lowest values to the highest. The size of the categories (called *bins* in *Excel*) depends on the range of the data and the size of the paper the graph will appear on. The number of bins does have some impact on the look of the frequency graph, and so you might want to generate more than one version of the frequency graph.

1. Open your percent ethnic table if it is not still open. Make a copy of your percent ethnic column and paste it to the right of the last column.

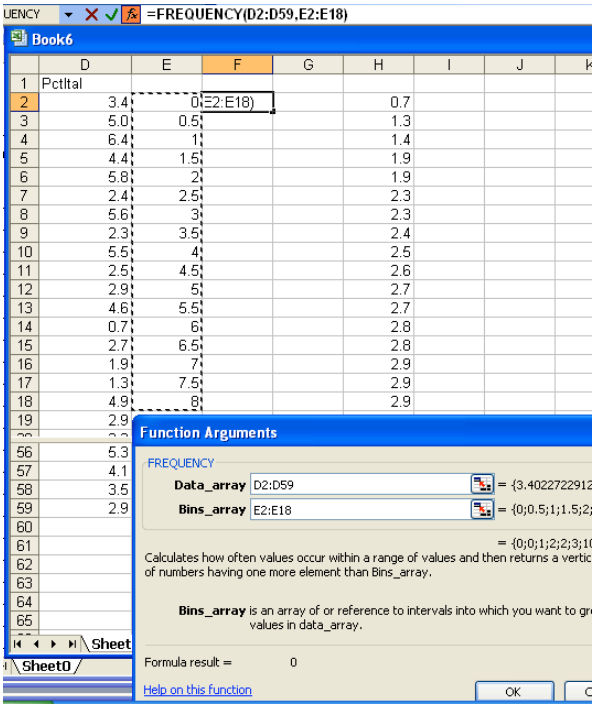
2. Sort the new column in ascending order to determine the range of the percentages. Fortunately, the orders of magnitude are much less than when dealing with raw numbers and this helps in setting the category interval. For the Italian data shown here the range of values extends from 0.7 to 7.6 percent. Thus we can set a category (bin) size of 0.5 to get 14 categories.

3. To carry out a *Frequency* function we need to set a series of category break values. **In the column to the right of the percent column and next to the first data value enter a 0. Below that enter the equation =d2 + 0.5**
Then fill down 16 rows so that the last value will be 8.0. Note the function is just a quick way to create a list of numbers. You could type them all in if you wish or if you choose to create irregular bin values.

4. Click in the column to the right of the bin values. Select *Insert* > *Function* and the Insert *Function* window will open. Under the *Or select a category:* window select *All*. Then under *Select a function* window scroll down to *FREQUENCY*. Note the explanation and example at the bottom of the window. For this function you will need your data array and the array of bins you just created.



5. Click OK.



6. In the *Function Arguments* window click on the *Data_array* window. Then in your spreadsheet click the first cell containing ethnic percent values. Then hold the *shift* key and click the last cell in that column. Here there are 58 cells, one for each California county.

7. Click on the *Bins_array* window and in your spreadsheet click on the first cell containing a bin value which is *0* here. Then hold the *shift* key and click the last bin value.

Then click *OK*. A single value will appear in the spreadsheet.

8. Select the cell holding the new frequency value and then hold the *shift* key and click to the right of the last bin value so all cells next to the bin values in the frequency column are highlighted.

9. Hit the *F2* key at the top of the keyboard.

10. Hold the *control* and *shift* keys and click the *enter* key. The frequencies for the remaining bins will be calculated.

F2 {=FREQUENCY(D2:D59,E

Book6

	D	E	F
1	PctItal	Bins	Frequency
2	3.4	0	0
3	5.0	0.5	0
4	6.4	1	1
5	4.4	1.5	2
6	5.8	2	2
7	2.4	2.5	3
8	5.6	3	10
9	2.3	3.5	5
10	5.5	4	10
11	2.5	4.5	5
12	2.9	5	6
13	4.6	5.5	3
14	0.7	6	6
15	2.7	6.5	2
16	1.9	7	1
17	1.3	7.5	1
18	4.9	8	1
19	2.9		

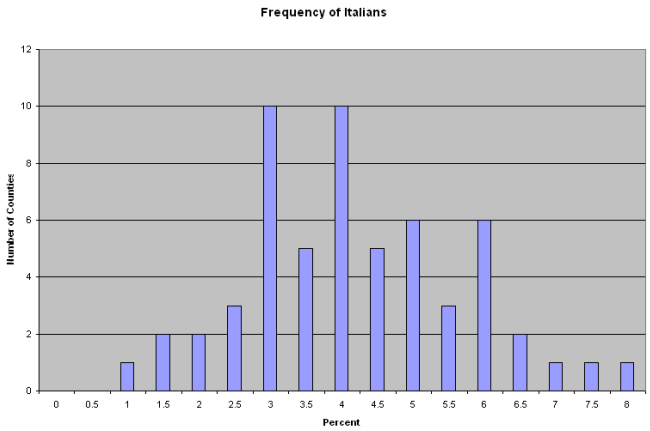
11. To be able to use the Bin values as labels you will need to convert the numbers to text. **Click on the bin and frequency columns so all cells are highlighted in them. Select *Edit* > *Copy*. Then select *Edit* > *Paste Special* and be sure to hit the *Values* button. Click *OK*.** This will remove the equations from the cells.

12. Now highlight only the bins column. Select *Data* > *Text to columns.*

In the *Convert Text to Columns Wizard* hit the *Next* button to get to *Step 3 of 3*. In that window click the *Text* button under the *Column data format* section. Then hit *Finish*. The numbers will shift leftward and now be considered as text.

13. Highlight the data values in both the *Bins* and *Frequency* columns. Then select *Insert* > *Chart*. At *Step 1* accept the vertical bar style and click *Next*. At *Step 2* click *Next*.

14. At Step 3 under *Titles* enter *Frequency of* [your group], enter *Percent* in the *X-axis* window, and *Number of Counties* under the *Y-axis* window. Under the *Legend* tab deselect the *Show legend* button. Then click *Next*. At *Step 4* click the As a new sheet: button and click *Finish*. The frequency graph will appear.



Look over the frequency of counties across the categories. For the percent Italian shown above, the percents are quite normal looking with some potential breakpoints at 3.5 and 5.5

Exercises

1. Download from SF3 tables P159B, P159C, P159D, P159H, and P159I. These are Poverty Status in 1999 by Age for selected ethnic groups. Create a composite table for persons only below the poverty level from each of the ethnic groups. The table should be Race as the column variable by Age as the row variable. Each cell shows the number of persons below the poverty level.

Compute the percent of the total population that is in poverty for each group.

Compute the percent of the ethnic total in poverty for each age category.

Compare the race groups as to their poverty levels both for the totals and for the age categories.

2. Download from SF3 tables P145B, P145C, P145D, P145H, and P145I. These are Sex by Age for different ethnic groups.

Create a composite table of the total population for each of the ethnic groups by age. Collapse the age categories into larger categories of 0 – 17, 18 – 44, 45 – 64, and 65 and higher.

Compare the age structures for each of the ethnic groups. What groups have large numbers of young persons? Discuss some implications of this.