

Using the *Health Indicators* database to help students research Canadian health issues

Joel Yan, Statistics Canada, joel.yan@statcan.ca, 1-800-465-1222
With input from Brenda Wannell, Health Statistics, Statistics Canada

Assignment

Health issues are of great interest to students. Everyone can use the *Health Indicators* database on the Statistics Canada site, www.statcan.ca to look for possible relationships between demographic factors, and health outcomes, or to look for patterns in health data by subprovincial region. The *Health Indicators* database contains a wealth of free data, as well as related articles from Statistics Canada and the Canadian Institute for Health Information. Two major surveys providing much of the data are the National Population Health Survey and the Canadian Community Health Survey. Health related data are provided for Canada, the provinces, and over 150 health regions across Canada. By correlating variables across these areas, we can assess the likelihood that there is a relationship among them. The assignment below is intended to serve as a demonstration of the wealth of data available for student analysis on the Health Indicators database and how the data can be imported into Fathom for analysis.

Question for analysis: Is smoking in Canada correlated to level of education?

Related Expectations for the Ontario Mathematics of Data Management (MDM4U) Course (with the corresponding MDM4U course unit and curriculum document page numbers shown in brackets):

- Solve problems involving complex relationships with the aid of diagrams. (ODV.02, Overall expectations – page 49)
- Locate data to answer questions of significance of personal interest by searching well-organized databases. (OD1.01, Organizing data – page 49)
- Describe the relationship between two variables by use the use of scatter graphs and interpreting the correlation coefficient. (ST4.01, and ST4.02, Statistics – page 52)

Procedure

Access Health Indicators Data

1. On the Statistics Canada website home page, www.statcan.ca, click on '**Our products and services**' (in the top blue bar).
 2. Choose '**Free**' under 'Browse our Internet publications'
 3. Scroll down and click on the subject '**Health**' then choose '**Health Indicators**'.
 4. Click on the latest release - currently '**Volume 2003**'.
- Scan the page and answer:

Question: How many health indicators does this product provide at the health region, province/territory, and Canada level? _____ (Answer: over 80)

5. Now you have many choices for selecting data. Click '**Data tables**' on the left side bar.

Note: At another time, you could go to the **Profiles section** if you want to extract multiple variables for a single health region.

6. Data tables are organized here according to the Health Indicator framework:

[Health status](#)

[Non-medical determinants of health](#)

[Health system performance](#)

[Community and health system characteristics](#)

In this case we are looking for smokers by health region and want to see if there is any correlation with average educational attainment.

7. Click 'Non-medical determinants of Health' to go to the large menu of tables available.
8. Select 'Smoking', the first determinant that we will analyze.
9. This brings us to a menu with several datasets available. Select 'Smoking status household population, aged 12 and over, 2000/2001', since these data are available by health region.
10. For output format, click the blue 'CANSIM' button at the right of the Smoking status data line. Working with the data in CANSIM format will enable us to combine datasets and output the results to a spreadsheet.
11. This opens up Table 105-0027, the specific CANSIM table with detailed smoking data by health region. Under Geography, scroll down and select all the health regions in one province of interest. You select areas by highlighting them on the list. In the sample graphs shown later, we selected all the 37 Public Health Units in Ontario. Note: Public Health Units are defined in Ontario only and roll up to larger District Health Councils.

Question: As indicated in brackets after the word **Geography**, for how many different geographic areas do we have the data on smoking rates? _____ (Answer: 199).

12. Under **Age group** select 'Total, 12 years and over'.
13. Under **Sex** select 'Both sexes'.
14. Under **Smoking**, select 'Current daily or occasional smoker'.
15. Under **Characteristics**, select 'Percent'.

16. Click 'Continue' at the bottom of the screen.
17. Now select Option 1: Table.
18. On the next screen, scroll down past the list of selected series and check that these are the areas you had requested. If yes, click 'Continue'. If not, click 'Modify request'. This will take you back to the previous screen and let you make changes to your selection.
19. On the next screen under **output format**, click on the drop-down box and select 'CSV file for spreadsheet use, time as columns'. Click 'Go' at the bottom of the screen to export the data.
20. Click 'Open' to view the table within your spreadsheet program. Answer the questions below.
How many columns are used in this spreadsheet? ____ (Answer: 6)
What does each column contain? _____
How many columns of actual numeric data are there? ____ (Answer: 1)
How many rows of actual data are there? _____ (Answer: 37)
Note: Each row corresponds to the smoking rate data for one region in Canada.
21. Next we label the column of data in the spreadsheet. In the first row above the column for the data on smokers, enter "Percent_smokers". This will help us interpret this attribute and will become the default attribute name later on when we import the data into Fathom.
22. Now delete the columns we do not need. Delete all the columns except the first (the area name) and last (the percent smokers) column.

Extracting an Explanatory Factor into a Spreadsheet

23. Save the spreadsheet to disk with a file name "Smokers". We will return to this file later to append possible explanatory or related factors.
24. Re-open the *Health Indicators* Data Tables window by clicking on '**Data Tables**' in the left side bar. Note: It should be still open.
25. Click on or scroll down to the 'Non-medical determinants of health'.
26. Under 'Living and working conditions' select 'Education', a factor that we want to explore in relation to smoking rates.
27. This brings us to a menu with 2 different data options available. Since we want to combine the data with the smoking data to look for a correlation, make sure to select a dataset with the same year and geography (e.g. health regions) used for the

(smoking) data we have already selected. In this case both datasets have the matching year and level of geography.

28. Scroll to 'Postsecondary graduates, 2001'. Click on this attribute name to see a new screen which explains the meaning of this term. For output format, again click on the blue 'CANSIM' button, as we will again output the results to a spreadsheet.
29. This opens up CANSIM table 109-0200 with the detailed education data by health region. To be compatible with the smoking data already extracted, on the CANSIM screen for the selected table, under **Geography** scroll down and select exactly the same regions that you selected above (e.g. Public Health Units for Ontario) .
30. Under **Census profile** select 'Post-secondary graduates aged 25 to 54, proportion of population aged 25 to 54 (Percent)'.
31. **Questions:** How many characteristics are included within the Census Profile? ____
Note: Each of these could be examined as a possible explanatory variable for the smoking rates
Name one characteristic on this list that you feel might be correlated to smoking rates by health region? _____
32. Click 'Continue' at the bottom of the screen.
33. Now select Option 1: Table.
34. On the next screen, you need to scroll down and click 'Continue'.
35. On the next screen under **output format**, click on the drop-down box and select 'CSV file for spreadsheet use, time as columns'. Click 'Go' to export the data.
36. Click 'Open' to view the table within your spreadsheet program.
37. Next we again delete the columns we do not need, all those except the first (the area name) and last (the percent with postsecondary education) column.
38. Label the column containing the education data. In the first row above this column, enter "Percent_Postsecondary_Education". This will become the default attribute name when we import the data into Fathom.

Merging the two Spreadsheet Files using EXCEL

39. Position your cursor on the row where you have the attribute names, just above the first observation. Using your mouse click and drag to highlight the (two) columns and all the rows of numeric data that you want to merge with the smoking data already extracted. Once you have selected the domain, select Copy from the 'Edit' pull-down menu. Or click the Copy icon from the top icon bar.

Note: You do not need to include the footnotes and source notes in the domain to be copied.

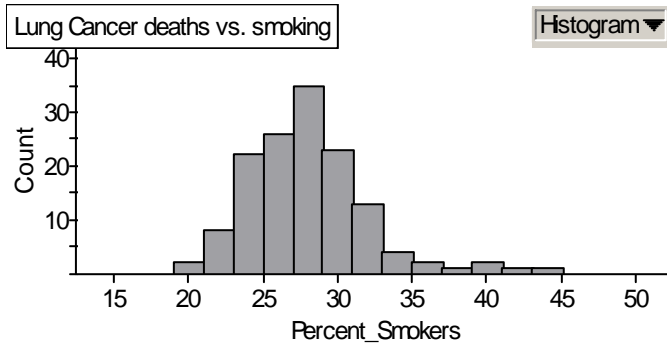
40. Open your previous Excel file (“Smoking”), if it is not already open.
41. Position your cursor on the row where you have the attribute names in the row just above the first observation (this should be the data for the first selected health unit) and on the first blank column available. This should be just to the right of the smoking rates. Then Paste in the new data from the other spreadsheet. Tip: Use the Paste icon or select Paste from the ‘Edit’ pull-down menu.
42. Now we need to check that the geographic areas from the two files match. Scroll down the spreadsheet and verify that the two area names on each row (from the two different files) are properly aligned and correspond. If the first few names are not aligned properly, undo the paste that you just did, carefully position the cursor and redo the previous steps.
43. Once we have verified that all the areas match, we can delete one of the two columns containing the area name.

Importing the Health Data into Fathom

44. Highlight the rows and columns of data in the spreadsheet and click the Copy icon. Choose **Copy** from the **Edit** menu.
45. Switch to **Fathom**. (If Fathom isn’t already running, you will need to launch it)
46. In a new document, make a new empty collection.
47. With the collection selected, chose **Paste Cases** from the **Edit** menu. This will import the health data from the spreadsheet.
48. Double click on the collection box. Change the name of the collection to a meaningful name, such as ‘Smoking vs. Education Level’.
49. Make a case table for the collection (for example by choosing **Case table** from the **Insert** menu)
50. You may need to edit the attribute names. If so, double click on each name in turn and change the names of the attributes as appropriate to Health-Region, Percent_Smokers, and Percent_Postsecondary_Education.
51. If the first case does not have numeric values for the two numeric attributes, delete this case. Scroll quickly through the case table. If there are any cases for which both numeric values are suppressed, delete these cases, by highlighting them, pulling down the Edit menu, and then selecting ‘delete case’.
52. Save the Fathom document by choosing **Save** from the **File** menu.

Graphing the Data

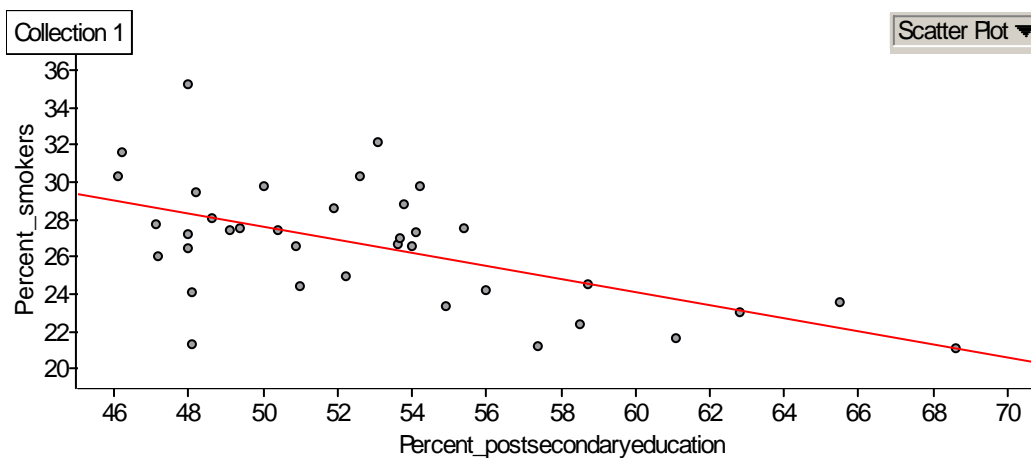
53. Bring down a graph. Drop the Percent current or occasional smokers attribute on the x-axis. Hold down the Ctl key as you do this, to exclude any of the non-numeric values from appearing on the graph. Using the pull-down menu above the graph, change the graph type to a histogram.



54. Write a description of the pattern you see on this graph. Is the distribution approximately normal? _____ Tip: Viewing the graph as a normal quantile plot will help you with this analysis.

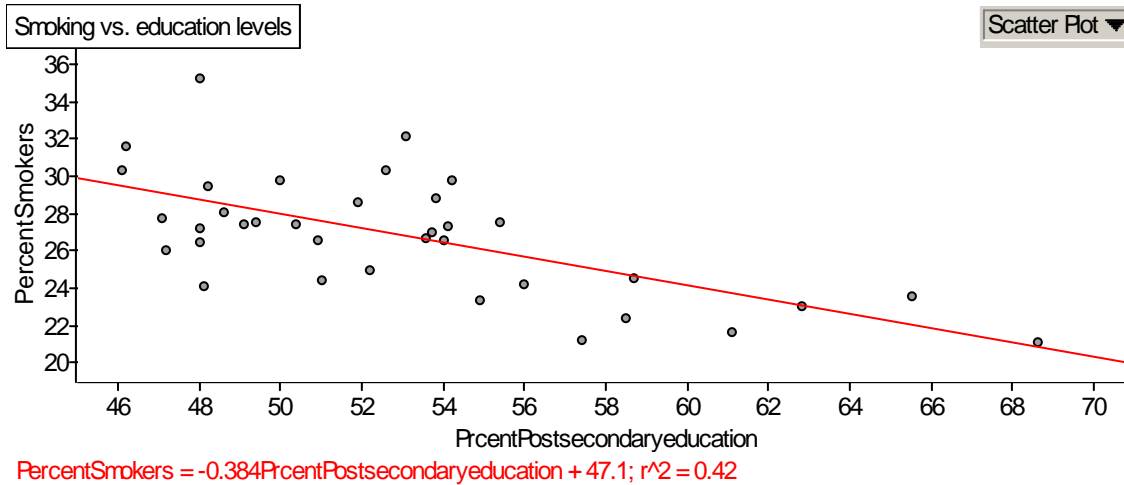
55. Bring down another graph. Drag the educational attainment attribute to the x-axis and the Percent_Smoking attribute to the y-axis

56. Right click on the graph and select the 'Least-squares Line' option. This will overlay in red the least- square line, as well as provide its equation and R squared value, as shown below.



57. Analyze the outliers, those values that are furthest from the line of best fit and have the biggest impact. Highlight the farthest outlier and identify its location (this appears on the lower left of the screen).

58. **Questions:** Why might this area be an outlier? _____ Now click this outlier point, and delete this case. What is the impact on the r^2 value and on the slope of the line? _____ (Answer: r^2 changes from 0.33 to 0.42 and the slope reduces from $-.351$ to $-.384$) After you have seen the impact on the equation, click Edit Undo to restore that point.



Further Analysis

Repeat the steps above to select other variables related to Health status and Non-medical determinants of health. The **Health Indicators** product contains many such variables. Again the variables can be combined and imported into Fathom or other analytical software for further analysis. Some relationships you may consider exploring, to start with:

- life expectancy against smoking, drinking, or life stress;
- lung cancer mortality against exposure to second hand smoke;
- infant mortality;
- circulatory disease death against level of physical activity, or against BMI;
- cancer deaths against dietary practices;
- deaths due to specific conditions against disease prevention measures such as flu shots, mammography, pap smears;
- level of depression versus income
- chronic conditions such as diabetes, asthma, high blood pressure against any of the non- medical determinants of health.

Caution: For whichever of the more than 80 health indicators you extract for your analysis, remember to select exactly the same health regions (under Geography) and similar time periods. Once the data have been merged in a single spreadsheet, check that all the geographic areas match.

Added codes Feb. 1, 2004