BDK08 Public Health Exercises

KEY

Use the data set BDK08\_CancerMapping to complete the following exercises. This data set contains information about the number of cases of cancer in 13,823 locations in New York State between 2005 and 2009. The variables are:

|  |  |
| --- | --- |
| geoid10 | A 12 digit geographic identifier. The first two digits represent state (all are 36 for New York), the next 3 represent the county, the next 6 represent the tract, and the final digit is a block group. Each block represents 1000-1500 people. In some cases, blocks were combined when there were very few cancer cases in order to preserve confidentiality. These ids include the letters DOH and a 4 digit identifier. |
| observed\_[cancer type] | Number of diagnoses cases of [cancer type]. observed\_total indicates the total number of all cancer cases in the block group. |
| expected\_[cancer type] | Number of expected cases of [cancer type]. expected\_total indicates the total expected number of all cancer cases in the |

1. Calculate the different between observed and expected total cancer cases for each block group (different = observed – expected). Which three regions had the greatest positive difference? Which three had the greatest negative difference? Does it look like any of the predictions were accurate?

Greatest positive difference (many more observed cases than expected cases):

360811551011: 72.47

360470974001: 52.28

360850187022: 52.12

Greatest negative difference (many fewer observed cases than expected cases):

36059DOH0012: -101.15

360610238011: -91.46

360811551023: -89.78

1. For each of the regions with the biggest differences in observed and expected cancer cases, what type(s) of cancer contributed to these differences?

360811551011: 72.47 lung (24.21), other (23.90), pancreas (9.08)

360470974001: 52.28 lung (13.89), other (7.70), prostate (7.67)

360850187022: 52.12 colorectal (12.10), lung (9.88), NHL (7.28)

36059DOH0012: -101.15 lung (-19.02), other (-16.11), prostate (-14.08)

360610238011: -91.46 prostate (-16.28), lung (-15.96), breast (-11.95)

360811551023: -89.78 other (-14.27), lung (-12.69), prostate (-11.84)

For positive differences (more observed cases than expected), lung cancer contributed the most. For negative differences (fewer observed cases than expected), lung and prostate cancer contributed the most. This might imply that the quantity of lung cancer cases is very difficult to predict, and/or has a lot of variability. Additionally, for the years 2005-9, prostate cancer was under-predicted in these areas.

1. If you were the researcher who generated this data set (and therefore had information about where each of the block groups is located), what would your next steps be in the study of cancer cases in New York? What additional information might you want to gather to prevent future cancer cases (or at least improve your ability to predict it)?

Possibilities:

Find out if any of the locations above are near each other.

Find out how many cancer cases developed in each year.

Create a visualization of the data, such as a map of the cancer cases. Additionally, filter it by year to see if trends can be discovered.

Look for variables/events/circumstances that led to the particularly large or small amounts of cancer occurrences. If these could be identified, figure out ways to prevent the things that cause the large number of cancer cases. Conversely, see if the variables that decrease the number of cancer cases can be implemented in other areas.