Pentru a cita lucrarea / To cite the paper: Drane JW, Petrisor AI, Dragomirescu L (2003), A Quick Screen for Counties with High Incidence of Cancer, 1st MEDINF International Conference on Medical Informatics & Engineering, October 9-11, 2003, Craiova, Romania

A Quick Screen for Counties with High Incidence of Cancer

J. Wanzer DRANE, PhD, PE[1], **Alexandru-Ionut PETRISOR**, MSPH[2], and **Liviu DRAGOMIRESCU**, PhD[3]

[1] Department of Epidemiology and Biostatistics, Arnold School of Public Health, University of South Carolina, USA; [2] Department of Environmental Health Sciences, Arnold School of Public Health, University of South Carolina, USA
[3] Department of Ecology, Faculty of Biology, University of Bucharest, Romania

Introduction

- Cancer cases follow a Poisson distribution:

$$\hat{\theta} = \frac{\sum_i m_i}{\sum_i n_i}$$

where:
n_i, i=1, 2, ..., N is the number of people at risk in the i^{th} county;
m_i, i=1, 2, ..., N is the number of new cases in the i^{th} county

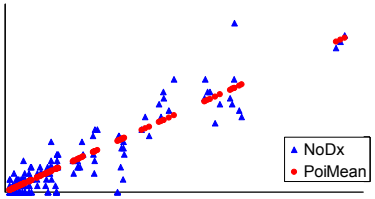
- Therefore, the estimated mean is: $\hat{E}(m_i) = n_i \cdot \hat{\theta}$
- Hyper-Poisson variation can be accommodated using linear regression of the actual number of cases observed in each area on the population at risk in those areas
- If $P(X \geq m_i \mid n_i \cdot \ ) \leq \alpha$, $\alpha = 0.01, 0.001$ etc., then the county is declared to have an epidemic
- Additional spatially indexed variables can be added to the regression model
- Regression must be forced through the origin

Data

- Data are found in a public file supplied by the South Carolina Central Cancer Registry for the period 1996-1998, aggregated by county and age groups.
- The variables are County (the actual county), YrDx (year), Race (race of subject: All_r- all races, B&O- African-American and others, and White- White), NoDx (number of newly diagnosed cases), and Npopn (population at risk of ovarian cancer). All_r is the sum of B&O and White.
- Cancer incidence and populations at risk were gathered and calculated by the SC Central Cancer Registry, Ms. Susan Bolick-Aldrich, MSPH, CTR, Director.

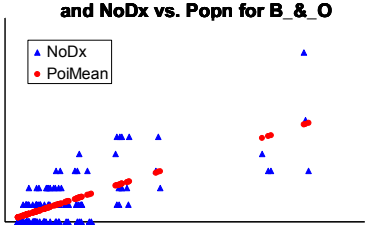
Results

Plot of Predicted Poisson Mean and NoDx vs. Popn for White



Linear regression of the number of cases in each area on the population at risk: white race

Plot of Predicted Poisson Mean and NoDx vs. Popn for B_&_O

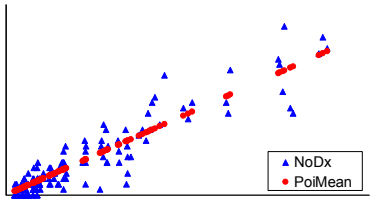


Linear regression of the number of cases in each area on the population at risk: black and other races

Results (continued)

County	YrDx	Race	Popn	NoDx	Poisson MEAN	PVALUE
Charleston	1996	All_r	306600	31	22.8926	0.06111
Charleston	1996	B&O	114475	4	4.92811	0.72474
Charleston	1996	White	192125	27	16.6302	0.01180
Charleston	1997	All_r	313478	16	23.406185	0.955922
Charleston	1997	B&O	117223	3	5.046413	0.879203
Charleston	1997	White	196255	13	16.987652	0.864353
Charleston	1998	All_r	316606	15	23.639740	0.976636
Charleston	1998	B&O	118364	3	5.095532	0.883167
Charleston	1998	White	198242	12	17.159645	0.920848

Plot of Predicted Poisson Mean and NoDx vs. Popn for All_r



Linear regression of the number of cases in each area on the population at risk: all races

Discussion

- In 1996 the p-value for all races is 0.0611, while for B&O is 0.7247 and for White is 0.0118. Any disparity of NoDx in excess of expectation is found in the White population. No county is statistically above expectation for either race.
- To continue this analysis, we must remember that -2 log (p-value) follows a chi-square distribution with DF = 2, and that the sum of independent chi-squares is a chi-square with the number of DF equal to the sum of the several DFs. Therefore,
  $\chi^2(6) = -2 \cdot [\log(0.6111) + \log(0.72474) + \log(0.01180)] = 10.508$
- The corresponding p-value is 0.89517. Therefore, even when one combines the three races for 1996 in Charleston, the p-value remains very high, and there is no cause for alarm.
- In terms of the linear regression, no further variables should be added. The regression analysis ends at this point. If there had been one or more counties with exceptionally high incidences (very small p-values), the regression model should be extended to include variables that might, a priori, be candidates as risk factors, or their respective surrogates, for ovarian cancer.
- The method is not limited by the population size, but works better when not too many NoDx are equal to 0.

Conclusion

- While least square analysis provides us with a line through the middle of all points, the variance of the Poisson remains equal to the mean. Therefore, the regression estimate of the variance is misleading, being used as a constant for all population sizes instead of being equal to the mean and estimated with the predicted mean.