

Appendix B

Statistical formulae and special terminology

Most formulae used in this report are described in Appendix 2 of the *All-Ireland cancer statistics report 1994-96*, March 2001. Two exceptions are described below:

**Confidence intervals:** The confidence intervals in this report are calculated using the gamma distribution as described in these formulae:

$$\text{LowerLimit} = \frac{v}{2y} (\chi^2)^{-1}_{\frac{2y^2}{v}}(\alpha/2)$$

$$\text{UpperLimit} = \frac{v + wM^2}{2(y + wM)} (\chi^2)^{-1}_{\frac{2(y + wM)^2}{v + wM^2}}(1 - \alpha/2)$$

where y is the age-adjusted rate, v is the variance as calculated in the equation,

$$v = \sum_{i=1}^m d_i (s_i / P_i)^2$$

wM is the maximum of the weights  $s_i/P_i$ ,  $1-\alpha$  is the confidence level desired (i.e. if 95% confidence intervals are needed, use  $\alpha = 0.05$ ), and  $(\chi^2)^{-1}_x$  is the inverse of the  $\chi^2$  distribution with x degrees of freedom.

**Spatial Scan Statistic:** The spatial scan statistic is described by Kulldorff (2002) as the following process:

The purely spatial scan statistic imposes a circular window on the map. The window is in turn centred on each of several possible grid points positioned throughout the study region. For each grid point, the radius of the window varies continuously in size from zero to some upper limit. In this way, the circular window is flexible both in location and size. In total, the method creates an infinite number of distinct geographical circles with different sets of neighbouring data locations within them. Each circle is a possible candidate for a cluster.

For each location and size of the scanning window, the alternative hypothesis is that there is an elevated rate within the window as compared to outside. Under the Poisson assumption, the likelihood function for a specific window is then proportional to:

$$(c/n)^c ([C-c]/[C-n])^{(C-c)} I()$$

where C is the total number of cases over the whole area, c is the number of cases within the window, and n is the covariate adjusted expected number of cases within the window under the null-hypothesis.

I() is an indicator function. When SaTScan is set to scan only for clusters with high rates, I() is equal to 1 when the window has more cases than expected under the null-

hypothesis, and 0 otherwise. The opposite is true when SaTScan is set to scan only for clusters with low rates. When the program scans for clusters with either high or low rates, then  $I()=1$  for all windows.

The likelihood function is maximized over all window locations and sizes, and the one with the maximum likelihood constitutes the most likely cluster. This is the cluster that is least likely to have occurred by chance. The likelihood ratio for this window constitutes the maximum likelihood ratio test statistic. Its distribution under the null-hypothesis is obtained by repeating the same analytic exercise on a large number of random replications of the data set generated under the null hypothesis. The p-value is obtained through Monte Carlo hypothesis testing, by comparing the rank of the maximum likelihood from the real data set with the maximum likelihoods from the random data sets. If this rank is  $R$ , then  $p = R / ( 1 + \#simulation )$ . In order for  $p$  to be a 'nice looking' number, the number of simulations is restricted to 999 or some other number ending in 999 such as 1999, 9999 and 29999. That way it is always possible to reject or not reject the null hypothesis for typical cut-off values such as 0.05, 0.01 and 0.001. Additional information and the software is available at: <http://www.satscan.org/>

Kulldorff M and Information Management Services, Inc (2002). *SaTScan v.3.05: Software for the spatial and space-time scan statistic*  
Bethesda, MD USA: National Cancer Institute