MDSC 643.01, Biostatistics I
Brant, Fall 2002

Aggregated Event Data - Analysis of Rates

The following data gives incidence of leukemia in males 0-14 during in a five year period in Birmingham, U.K. (source: Breslow and Day, Vol. I). The population figures are for the 4th year of that period.

Ages	Cases	Pop. (100k)	Year F.U.	Rate (per 100k)
<1	2	0.45	5	0.88
1 thru 4	47	1.82	5	5.15
5 thru 9	30	2.28	5	2.63
10 thru 14	13	2.03	5	1.28

The rates quoted above are per year, or more accurately per "person-year", as each of the members of the population contribute 5 years of information. It is assumed that the population is stable over the 5 years, so that the total denominator is obtained by multiplying the population figures by 5. For example, for the ≤ 1 category the rate is $\frac{2}{5\times .453} = .88$.

In general the incidence density (or informally, incidence rate) is calculated from X, the incidence frequency, and T, the total follow-up time as $\hat{\mathrm{ID}} = \frac{X}{T}$ (the estimate of the true incidence density, ID). Note that T can obtained by summing up individual follow-up periods for the population at risk, or approximated as above by the product of population size and "average" follow-up interval.

The random variability in the incidence frequence, X, can be usually be described by a Poisson distribution with parameter $\mu = \mathrm{ID}\times T$. Based on this the standard error for $\hat{\mathrm{ID}}$ is $\frac{\sqrt{X}}{T}$. When μ is small, (rule of thumb $\hat{\mu} = X < 10$) exact Poisson probabilities can be used to justify inference for ID. When $X \geq 10$, the normal approximation to the Poisson distribution justifies using $\hat{\mathrm{ID}} \pm \text{z-value} \frac{\sqrt{X}}{T}$ as an approximate confidence interval for the ID .

Comparing Rates

To compare two possibly different rates, $\mathrm{ID}_1, \mathrm{ID}_2$, we use the estimated rate ratio $\hat{\mathrm{RR}} = \frac{\hat{\mathrm{ID}}_1}{\hat{\mathrm{ID}}_2}$

As with previous experience, the most accurate large sample theory arises from applying natural logarithms , which leads to $\mathrm{SE}[\ln(\hat{\mathrm{RR}})] = \sqrt{\frac{1}{X_1} + \frac{1}{X_2}}$

where X_1, X_2 are the observed frequencies. The above can be used to obtain an approximate confidence interval $\hat{\mathrm{RR}}\, e^{\pm \text{z-value}\, SE}$ for the true RR , using the standard error as above. Exact calculations are also possible.