CATEGORICAL DATA ANALYSIS

SPRING 2003

THREE WAY CONTINGENCY TABLES

We will start our discussion with combining the results of a series of 2×2 tables.
In considering the association of disease and exposure, it may be important to control for a third variable which is known to influence this association. Examples in epidemiology of confounding variables would be age and gender.

There are 3 issues to consider:

1. Is the odds ratio constant over the different strata
2. If the odds ratio is constant, is it equal to 1
3. If it is constant and not 1 then, what is the best estimate of its value

Notation: Let $i = 1, \ldots, I$ denote the levels of the stratification variable. For each value of i, considering disease (D) vs. exposure (E) we have the table:

	E	$\bar{E}$	
D	a_i	b_i	n_{1i}
$\bar{D}$	c_i	d_i	n_{2i}
	m_{1i}	m_{2i}	N_i

The tests for the above issues are developed simultaneously. One would think that the first test being the most important, would be conducted first but this isn't the case. First we will look at the Mantel-Haenszel-Cochran test of $H_0 : \psi_i = 1$ for all i. (where ψ_i represents the odds ratio for table i)

$$X^2_{MHC} = \frac{\left(\left| \sum a_i - \sum E_0 \left(a_i \mid \psi_i = 1 \right) \right| - 1/2 \right)^2}{\sum V_0 \left(a_i \mid \psi_i = 1 \right)} = \frac{\left(\left| \sum a_i - \sum \frac{n_{1i} m_{1i}}{n_i} \right| - 1/2 \right)^2}{\sum \frac{n_{1i} n_{21} m_{1i} m_{21}}{N_i^2 \left(N_i - 1 \right)}} \underset{H_0}{\sim} \chi^2 (1)$$

$$p = P\left\{ \chi^2 (1) \geq X^2_{MHC} \right\}$$

To determine if the data is "rich enough" to support the asymptotic use of the χ^2 distribution, we use a modification of the rule of 5. We compute the range of $\sum A_i(1) = \sum_{i=1}^{I} \frac{n_{1i} m_{1i}}{N_i}$ subject to the assumption of having fixed marginals..

Max = $\sum \min(m_{1i} n_{1i})$ and Min = $\sum \max(0, m_{1i} - n_{0i})$ if $\sum A_i(1)$ is 5 units away from both extremes, then we assume it is ok.

Tests that the odds ratio is homogeneous across all I strata are based on a comparison of the odds ratio in each stratum, $\hat{\psi}_i$, to an overall estimate. Hence we must consider estimates of the common value before we proceed further with tests and confidence intervals. That's why the progression isn't quite logical. Luckily, using computer packages, you get all the information presented to you at once. Two estimators we will consider are

1. The exact conditional estimator

   This is the value of ψ such that $\sum_{i=1}^{I} a_i = \sum E(a_i \mid n_{1i}, n_{2i}, m_{1i}, m_{2i}, \psi)$. This can be very computationally intensive and therefore alternatives were considered when the data was rich enough.

2. The Mantel-Haenszel estimator. This estimator is simple and seems to work well under almost any circumstances, i.e. it's pretty robust to "thin data".

$$\hat{\psi}_{MH} = \frac{\sum \frac{a_i d_i}{N_i}}{\sum \frac{b_i c_i}{N_i}} = \frac{\sum w_i \hat{\psi}_i}{\sum w_i} \qquad \text{where} \qquad \hat{\psi}_i = \frac{a_i d_i}{c_i b_i}, \quad w_i = \frac{b_i c_i}{N_i} \qquad \text{note } w_i \cong V(\hat{\psi}_i)^{-1}$$

   This is a weighted average of the individual odds ratios which is not affected by zero cells.

The test of homogeneity of the odds ratio over the I levels of the stratification variable is based on comparing each value of a_i to an estimate of its expected value under the null hypotheses:

$H_0 : \psi_i = \psi \neq 1$

$$X_H^2 = \sum_i \frac{\left(a_i - A_i(\hat{\psi})\right)^2}{V(a_i;\hat{\psi})} \underset{\text{large n's}}{\overset{H_0}{\sim}} \chi^2(I-1) \quad \text{and} \quad p = P\left(\chi^2(I-1) \geq X_H^2\right)$$

where $A_i(\hat{\psi})$ is defined implicitly by $\psi = \frac{A_i(n_{2i} - m_{1i} + A_i)}{(n_{1i} - A_i)(m_{1i} - A_i)}$. First you find the estimator of ψ and then you solve this equation for A_i.

A more specific test in the sense that the alternative is restricted to a linear trend in the odds ratio may be formed if the stratification variable is at least ordinal scaled. Let $x_1, x_2, \ldots x_I$ denote the I values of the stratification variable. The test for trend statistic is

$$X^2_{Trend} = \frac{\left[\sum x_i \left(a_i - A_i(\hat{\psi})\right)\right]^2}{\sum x_i^2 V(a_i;\hat{\psi}) - \frac{\left[\sum x_i V(a_i;\hat{\psi})\right]^2}{\sum V(a_i;\hat{\psi})}} \sim \chi^2(1)$$

This is conceptually a test of a regression coefficient for a_i regressed on x_i with weights $V(a_i;\hat{\psi})$