**BIOS 6244  Analysis of Categorical Data**
**October 12, 2005 Lecture**

Reading Assignment (will not be covered in class)
Sections 2.4.6 and 2.4.7, pp. 32-34.  These pages will be posted to the website for those
students who do not have a textbook.

Testing Independence for Ordinal Data (Sec. 2.5)

The $\chi^2$ and LR tests treat both classification variables in the I x J table as nominal.
However, when either the row or column variables are ordinal, statistical tests that are
more powerful are available.

Test for Linear Trend (Sect. 2.5.1)

When the row variable X and the column variable Y are both ordinal, it is usually of
interest to test for *trend*, i.e., as the level of X increases, does the level of Y tend to
increase (*positive trend*), decrease (*negative trend*), or stay the same (*no trend*). In the
most commonly used method, we assign a numerical value (or *score*) to each category of
both X and Y, and then measure the strength of linear association between the row and
column scores.

What test statistic do we use to perform the test for linear trend?  We know that the
Pearson correlation coefficient r measures the degree of linear association between any
two variables.  So, once we have assigned scores to the rows and columns, we just
calculate the correlation between them.

In the most general situation, suppose we have assigned scores $u_1 \le u_2 \le \dots \le u_I$ to the
rows and $v_1 \le v_2 \le \dots \le v_J$ to the columns.  Note that it is always possible to assign such
monotone scores to the rows and columns since we know that the rows and columns can
be ordered in some meaningful way.

The formula for Pearson's r calculated from the row and column scores is

$$r_{uv} = \frac{\sum_{j=1}^{J}\sum_{i=1}^{I} n_{ij} u_i v_j - \dfrac{\sum_{i=1}^{I} n_{i+} u_i \sum_{j=1}^{J} n_{+j} v_j}{n}}{\sqrt{\left[\sum_{i=1}^{I} n_{i+} u_i^2 - \dfrac{\left(\sum_{i=1}^{I} n_{i+} u_i\right)^2}{n}\right]\left[\sum_{j=1}^{J} n_{+j} v_j^2 - \dfrac{\left(\sum_{j=1}^{J} n_{+j} v_j\right)^2}{n}\right]}}.$$

This is equivalent to

$$r_{uv} = \frac{Cov(u, v)}{SD(u)SD(v)},$$

where *weighted* versions of the covariance and standard deviations are calculated. Note that the weights are determined from the observed cell frequencies and marginal totals.

With regard to scores, we usually just number the rows from 1 to I and the columns from 1 to J and then calculate $r_{uv}$. However, other scores may be more appropriate. (More on how to choose the row and column scores later.)

We know from the properties of Pearson's r that $-1 \le r_{uv} \le 1$. Values near 1 provide evidence of a positive linear trend, values near -1 provide evidence of a negative linear trend, and values near 0 provide evidence of no trend. If X & Y are independent then the population correlation is 0, so values of r near 0 also provide evidence that the rows and columns are independent.

We are interested in testing the null hypothesis $H_0$: X and Y are uncorrelated vs. the alternative $H_a$: X and Y are correlated. Here, "correlated" refers to linear association.

The test statistics is

$$M^2 = (n-1)r_{uv}^2.$$

Note that $M^2 \uparrow$ as both $r_{uv}^2 \uparrow$ and $n \uparrow$. For "large" n, $M^2$ has an approximate $\chi^2(1)$ distribution. Large values of $M^2$ indicate that the null hypothesis should be rejected, so we are interested only in an upper-tailed test. We can also use $r_{uv}$ (with the appropriate "+" or "-" sign attached, to indicate whether we have a positive or negative trend. We could perform the test for trend using $M = r_{uv}\sqrt{n-1}$ as the test statistic and calculate a lower or upper-tailed approximate p-value as appropriate, using the standard normal distribution.

It can be shown that the test for trend based on $M^2$ treats the row and column variables "symmetrically"; that is, if we interchange rows and columns (i.e., transpose the contingency table), the value of $M^2$ will be unchanged.

Alcohol and Infant Malformation Example (Sec. 2.5.2)

Table 2.7, p. 35, in our textbook, refers to a prospective study on the association between maternal drinking and congenital malformations. After the first 3 months of pregnancy, the women completed a questionnaire on their alcohol consumption during pregnancy. Following childbirth, observations were recorded on the presence or absence of congenital sex organ malformations.

**(NOTE:  Table 2.7 is reproduced on the following page.)**

**Table 2.7    Infant Malformation and Mother's Alcohol Consumption**

| Alcohol Consumption | Malformation | | Total | Percentage Present | Adjusted Residual |
|---|---|---|---|---|---|
| | Absent | Present | | | |
| 0 | 17,066 | 48 | 17,114 | 0.28 | −0.18 |
| < 1 | 14,464 | 38 | 14,502 | 0.26 | −0.71 |
| 1–2 | 788 | 5 | 793 | 0.63 | 1.84 |
| 3–5 | 126 | 1 | 127 | 0.79 | 1.06 |
| ≥ 6 | 37 | 1 | 38 | 2.63 | 2.71 |

*Source:* B. I. Graubard and E. L. Korn, *Biometrics 43:* 471–476 (1987).
Reprinted with permission of the Biometric Society.

Do these data indicate a trend?  From examining Table 2.7, we see that the % of infants with malformations decreases slightly when going from "0" to "< 1," but then increases monotonically for the remaining exposure categories, with a big jump from "3-5" to "≥ 6."  So, it appears just from looking at the data that there is a positive trend.

Alcohol consumption was measured as average # of drinks per day, so the rows obviously constitute an ordinal variable.  "Malformation," the column variable, is binary and binary variables can always be safely treated as ordinal.  Any choice of scores for a binary variable will yield the same value of $r_{uv}$ and M.  The standard way to score a variable whose levels represent "Absent" and "Present" is 0 = Absent and 1 = Present.

How do we score the rows in Table 2.7?  We want to choose scores that properly represent the distances between categories.  A reasonable way to score the rows in this example would be to choose the midpoint of the intervals  defining the grouped values for alcohol consumption.  In other words,

| Alcohol Consumption | Midpoint |
|---|---|
| 0 | 0.0 |
| < 1 | 0.5 |
| 1-2 | 1.5 |
| 3-5 | 4.0 |
| ≥ 6 | ? |

For the last category, Agesti uses 7.0 as the midpoint, i.e., he assumes the interval goes from 6-8.  This is somewhat arbitray, but not unreasonable.  Using SAS or SPSS to calculate Pearson's r for these data (more on how to do this later), we obtain

$$r_{uv} = .0142,$$

$$m = r_{uv} \sqrt{n-1} = .014\sqrt{32573} = 2.563 .$$

The approximate upper-tailed p-value calculated using the standard normal is .005. This provides strong evidence for a positive linear association.

For the general test of trend, $m^2 = (2.563)^2 = 6.569$. The approximate upper-tailed p-value using the $\chi^2(1)$ distribution is .010, also strong evidence of a linear trend (no direction specified).

These results lead us to reject the hypothesis of no correlation between X & Y and conclude that maternal alcohol consumption in the first trimester and congenital sex organ malformation in newborn infants are positively linearly related.

Note the extremely small value of $r_{uv}$ (.014), hardly indicative of a strong association. As Agresti points out, however, $r_{uv}$ is not a very good measure of association for tables like Table 2.7 that are highly unbalanced, with several small frequencies. However, the test for trend based on $r_{uv}$ is still valid. Other approaches that give more useful information will be covered in Chapter 4.

What happens if we apply the $\chi^2$ or LR tests to these data?

For the $\chi^2$ test:     $x^2 = 12.082$ (p = .017).

For the LR test:     $g^2 = 6.202$ (p = .185).

Thus, there is substantial disagreement between the 2 tests. This is understandable, given that the expected cell frequencies are so small: in cells (3,2), (4,2), and (5,2), the expected frequencies all less than 5, thereby limiting the applicability of the $\chi^2$ approximation for the test statistics.

**NOTE**: Applying the $\chi^2$ or LR tests to the data in Table 2.7 ignores the ordinality of the alcohol consumption categories. Therefore, their use is not recommended for contingency tables like Table 2.7, or for any contingency table in which either the row variable or column variable (or both) are ordinal.

Why Does Taking Account of Ordinality Increase Power? (Sec. 2.5.3)

In testing the independence of X & Y, the $\chi^2$ and LR tests are designed to detect the most general departure from independence – that of *any* type of statistical dependence. In achieving this level of generality, the tests sacrifice sensitivity.

When there truly is a positive or negative trend in the association between X & Y, the test for trend has a power advantage over the $\chi^2$ and LR tests because it is targeted toward detecting a specific type of statistical dependence; namely, a positive or negative linear

trend. Since the mean of a $\chi^2$ distribution is equal to the df, a relatively large value of $M^2$ will be further out in the tail of a $\chi^2(1)$ distribution than would a comparable value of $X^2$ or $G^2$ relative to a $\chi^2[(I-1)(J-1)]$ distribution. (See Figure 2.1.) Since the test statistic would be further out in the tail of a $\chi^2(1)$ than it would be in a $\chi^2[(I-1)(J-1)]$ distribution, the p-value would be lower, resulting in higher power. (When there truly is a linear trend in the association between X and Y, $M^2$ tends to have about the same value as $X^2$ or $G^2$. In the maternal alcohol example, $m^2 = 6.7$, $x^2 = 12.1$, and $g^2 = 6.2$.)

Another advantage of the test for linear trend vs. the $\chi^2$ and LR tests is that the $\chi^2$ approximation to the null distribution of the test statistics is better for smaller degrees of freedom. (So, it would be better if df = 1 than if df = 4, say.)

Choice of Scores (Sec. 2.5.4)

For most data sets, the choice of scores has little effect on the results. *Any* choice of monotone scores will usually yield about the same p-value for the test for trend. However, when the data are extremely unbalanced (i.e., when a few cells have many, many more observations than the remaining cells), the choice of scores can be very important.

Maternal Alcohol Example, cont.

In our analyses of these data, we scored the rows using the midpoint of the alcohol consumption categories. If we had used equally-spaced row scores (e.g., 1, 2, 3, 4, 5), the test statistic would have been only $m^2 = 1.828$ (p = .176). (Recall that $m^2 = 6.569$, p = .010, using the midpoint scores.) Note that any linear transformation of a set of row or column scores will have no effect on the value of $r_{uv}$ (or $m^2$) since Pearson's correlation coefficient is invariant with respect to linear transformations of the scores for either X or Y.) Thus, row scores of 1, 2, 3, 4, 5 will yield the same value of $m^2$ as 0, 1, 2, 3, 4 or 10, 20, 30, 40, 50.

The choice of row and column scores is arbitrary to some extent and should be justified in the context of the applied problem. If several choices of row and/or column scores seem appropriate, conduct a *sensitivity analysis*, i.e., rerun the analyses with each set of scores separately and then compare the results. If the same conclusion is reached regardless of the choice of scores, then we can be more confident in the validity of our conclusion. If the conclusion is score-dependent, then we should choose the set of scores that we can justify most easily to external reviewers of our work. Generally speaking, if the categories of X or Y are based on grouped data, then the midpoint of the interval corresponding to a given category is a good choice for the score for that category.

What If Both X & Y Are Ordinal?

In this case, basing the test for linear trend on the $\chi^2(1)$ approximation to the null distribution of $M^2$ is equivalent to using the normal approximation to the distribution of

Spearman's rho, which is a nonparametric alternative to Pearson's r. However, there are more accurate methods for calculating the p-value for Spearman's coefficient and should be used instead. (They are available in SAS and SPSS.) Other measures of association between ordinal variables been also been proposed (e.g., Kendall's tau, gamma), and these could be used instead of Spearman's rho. However, Spearman's rho is generally considered to be preferable to these other measures.

Trend Tests for I x 2 and 2 x J Tables (Sec. 2.5.5)

Assume now that the row variable X denotes "exposure" and the column variable denotes "disease." When X is binary, the resulting contingency table is 2 x J. Using scores of $u_1 = 0$ for "non-exposed," and $u_2 = 1$ for "exposed," it can be shown that the $\chi^2(1)$ test based on $M^2$ is equivalent to Student's t-test for comparing the mean column scores between the exposed and non-exposed groups.

An alternative way of assigning scores to the columns is to use the mid-rank of the observations in each column. To determine the midrank, rank all subjects from 1 to n in terms of the ordinal variable, and then average the ranks within each ordinal category.

As an example, consider the following 2x3 table constructed using the Toxicology and Trauma data:

|  | Alcohol Disorder | | | |
| --- | --- | --- | --- | --- |
| Cocaine Tox Screen Result | None | Abuse | Dependence | Withdrawal |
| Positive | 62 | 35 | 7 | 6 |
| Negative | 578 | 152 | 24 | 23 |
| Total | 640 | 187 | 31 | 29 |

The midranks for each level of "Alcohol Disorder" are determined as follows. The 640 subjects in the "None" category are all tied in terms of their alcohol "score" and share ranks 1-640. The midrank for this category is then the average of all numbers from 1 to 640, inclusive, which can be found by calculating $\frac{640+1}{2} = 320.5$. This mid-rank is then assigned as the score for the "None" category in the above table. In the "Abuse" category, all subjects are tied, and share ranks 641 through 827 (640 + 187 = 827). The midrank for this category is therefore $\frac{641+827}{2} = 734$. For "Dependence," the midrank

is $\dfrac{828+858}{2} = 843$ and for "Withdrawal," it is $\dfrac{859+887}{2} = 873$. So, the scores assigned to the "Alcohol Disorder" categories are 320.5, 734, 843, and 873.

Applying $M^2$ using this set of scores in a 2 x J table is equivalent to applying the Mann-Whitney-Wilcoxon nonparametric test using the normal approximation to the null distribution of the M-W-W test statistic. There are more accurate ways of determining the p-value for the M-W-W test and should be used instead. These are also available in SAS and SPSS.

If Y is binary (i.e., the contingency table is I x 2), we focus our attention on how the proportion of "diseased" subjects varies across the row categories. This formulation is appropriate for the maternal alcohol data presented in Table 2.7. If we use $M^2$ with scores of 1, 2, …, I assigned to the rows of the table, the resulting test is called the *Cochran-Armitage test for trend.* Using $M^2$ with other scores (perhaps based on the midpoints of the exposure categories) is equivalent to using the normal approximation to the null distribution of the *permutation test for trend.* Exact p-values for the Cochran-Armitage test and the permutation test can be obtained using StatXact. SAS and SPSS can also provide exact p-values in some situations.

Nominal-Ordinal Contingency Tables (Sec. 2.5.6)

The test for trend based on $M^2$ assumes that both X and Y are ordinal. If either X or Y is nominal with only 2 categories (i.e., binary), it can safely be treated as if it were ordinal. If either X or Y is nominal with more than 2 categories, the test for trend based on $M^2$ is inappropriate and should not be used. If X is nominal with more than 2 categories and Y is ordinal, an ANOVA-like procedure can be used to compare the mean response of the ordinal variable Y across the I rows. This statistics is covered in Section 7.3.6. If X is ordinal and Y is nominal with more than 2 categories, the $\chi^2$, LR, or the exact tests to be discussed in Section 2.6.5 should be used.