

3.4 $I \times J$ tables with ordinal outcomes

Tests that take advantage of ordinal data's structure can increase power and interpretability. We now assume both X and Y are ordinal.

3.4.1 Linear trend alternative to independence

If we are willing to replace the ordinal outcomes by numerical scores, we can compute something akin to a correlation between X and Y . Let $u_1 \leq u_2 \leq \dots \leq u_I$ for X and $v_1 \leq v_2 \leq \dots \leq v_J$ for Y . Define

$$r = \frac{\sum_{i=1}^I \sum_{j=1}^J n_{ij} (u_i - \bar{u}_i)(v_j - \bar{v}_j)}{\sqrt{\sum_{i=1}^I \sum_{j=1}^J n_{ij} (u_i - \bar{u}_i)^2 \sum_{i=1}^I \sum_{j=1}^J n_{ij} (v_j - \bar{v}_j)^2}},$$

where $\bar{u}_i = \sum_{j=1}^J n_{ij} u_i / n_{i+}$ and $\bar{v}_j = \sum_{i=1}^I n_{ij} v_j / n_{+j}$.

r is akin to a correlation between X and Y , and in fact *is* the sample correlation when each (X, Y) pair is replaced by its score (u, v) .

r is going to estimate something lurking underneath, a population parameter ρ . Testing $H_0 : \rho = 0$ is a test for linear association between X and Y .

Define the test statistic

$$M^2 = (n_{++} - 1)r^2.$$

$M^2 \overset{\bullet}{\sim} \chi_1^2$ when $H_0 : \rho = 0$.

For the job satisfaction table, $r = 0.2002$ and $M^2 = (96 - 1)0.2^2 = 3.81$. The p -value is $P(\chi_1^2 > 3.81) = 0.051$. If we instead test $H_0 : \rho \leq 0$, the p -value is half this, 0.026, and we accept a positive association, $\rho > 0$.

Note that $X^2 = 6.0$ and $G^2 = 6.8$ yielding p -values of 0.74 and 0.66 respectively for $H_0 : X \perp Y$.

Taking into account the ordinal nature of the data and focusing on linear trend helped refine our assessment of the relationship of X and Y . Note that you cannot get M^2 directly in SAS, but rather r . Useful for one-sided tests (positive or negative trend).

3.4.3 The γ statistic revisited

Recall that $\hat{\gamma}$ estimates γ , the probability of concordance minus the probability of discordance. When $H_0 : \gamma = 0$ is true, the probability of concordance is equal to the probability of discordance, and there is no evidence of a *monotone association*.

For the job satisfaction data, we obtain a 95% CI of $(-0.01, 0.45)$ for γ , obtained as $0.221 \pm 1.96(0.117)$, with $\hat{\gamma} = 0.221$ and $\text{se}(\hat{\gamma}) = 0.117$ obtained from SAS.

We cannot reject $H_0 : \gamma = 0$. However, we do reject $H_0 : \gamma \leq 0$ in favor of $H_1 : \gamma > 0$.

3.4.4 Using focused alternatives gives added power

- G^2 and X^2 test $H_0 : X \perp Y$. Does not take into account nature of ordinal data. $df = (I - 1)(J - 1)$ reflecting all possible ways data can be dependent.
- For ordinal data, $H_0 : \rho = 0$ and $H_0 : \gamma = 0$ (or one-sided versions) test no association versus *focused* alternatives that are a special case of dependence. These tests focus on one parameter that describes a specific, defined type of association (linear or monotone).
- Since the alternative is focused, there can be more power to detect an association. $df = 1$ instead of $df = (I - 1)(J - 1)$.

3.4.5 Choice of scores in computing r and M^2

The scores $u_1 \leq u_2 \leq \cdots \leq u_I$ for X and $v_1 \leq v_2 \leq \cdots \leq v_J$ for Y affect r and M^2 and therefore the p -value for $H_0 : \rho = 0$.

- A linear transformation of scores does not affect r or M^2 . For example, using $\{1, 2, 3, 4\}$ or $\{52, 53, 54, 55\}$ or $\{3, 6, 9, 12\}$ for X all yield the same r .
- For most data, different choices of scores tend to give roughly the same r and p -value.
- Highly unbalanced data will be more sensitive to the choice of scores.

Example: Study of relationship between drinking during pregnancy and congenital malformations.

Malformation	Drinks per day				
	0	< 1	1 – 2	3 – 5	≥ 6
Absent	17,066	14,464	788	126	37
Present	48	38	5	1	1

Let the scores for X be $\{1, 2\}$.

- For Y , $\{0, 0.5, 1.5, 4.0, 7.0\}$ yields $M^2 = 6.57$ with $p = 0.01$.
- For Y , $\{1, 2, 3, 4, 5\}$ yields $M^2 = 1.83$ with $p = 0.18$.

One solution to this discrepancy is to use scores suggested by the data: *midranks*.

For the alcohol variable, $17066 + 48 = 17144$ didn't drink during pregnancy. The midrank is $(1 + 17144)/2 = 8557.5$. The next category, those that averaged less than one drink per day, we start at 17145 and go up to $17144 + (14464 + 38) = 31646$. The midrank for the 2nd category is then $(17145 + 31647)/2 = 24395.5$ (book typo?). The midrank for the 1 – 2 category is $(31617 + 32409)/2 = 32013$, etc. Scores are $\{8557.5, 24395.5, 32013, 32473, 32555.5\}$.

Using these midranks yields $M^2 = 0.35$ and $p = 0.55$.

Here, inappropriate: treats 1 – 2 as being much closer to ≥ 6 than to 0 drinks. Probably best to use midranks when no obvious set(s) of scores exist.

The output for r is the Pearson correlation in `proc freq`. The default method for computing scores is to use numerical values provided (e.g. 1, 2, 3, ...) for factor levels. Midranks are used by specifying `SCORES=RANK`.

Polychoric correlation

For ordinal (X, Y) , we can envision underlying *latent* continuous variables (Z_1, Z_2) that determine (X, Y) according to cutoffs.

$$X = i \Leftrightarrow \alpha_{i-1} < Z_1 < \alpha_i,$$

and

$$Y = j \Leftrightarrow \beta_{j-1} < Z_2 < \beta_j,$$

where

$$-\infty = \alpha_0 < \alpha_1 < \cdots < \alpha_{I-1} < \alpha_I = \infty,$$

and

$$-\infty = \beta_0 < \beta_1 < \cdots < \beta_{J-1} < \beta_J = \infty.$$

A picture helps...

If we assume that (Z_1, Z_2) are bivariate normal $N_2(\mathbf{0}_2, \mathbf{\Sigma})$, where

$\mathbf{\Sigma} = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$, then there are $1 + (I - 1) + (J - 1)$ parameters to estimate: ρ , $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_{I-1})$, and $\boldsymbol{\beta} = (\beta_1, \dots, \beta_{J-1})$.

The parameter ρ is called the *polychoric correlation* between X and Y , and is estimated in SAS proc freq using the PLCORR option in the TABLES statement.

This idea is useful in regression modeling where each outcome (X_i, Y_i) has explanatory variables $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})$. Then we can specify $E(\mathbf{Z}_i) = \mathbf{B}\mathbf{x}_i$ and $\rho = \boldsymbol{\gamma}'\mathbf{x}_i$. (More later...)

3.5 Exact tests of independence

There's a lot of info in here (pp. 91-101, 10 pages). We'll focus on what's involved in obtaining exact p -values for X^2 and G^2 instead of asymptotic $\chi^2_{(I-1)(J-1)}$.

Instead of an asymptotic distribution, we need the *exact* distribution of cell counts under $H_0 : \pi_{ij} = \pi_{i+}\pi_{+j}$.

Under product multinomial sampling, the row totals are fixed at n_{i+} ahead of time. Under H_0 , the row counts are independent $\text{mult}(n_{i+}, \boldsymbol{\pi})$ where $\boldsymbol{\pi} = (\pi_{+1}, \pi_{+2}, \dots, \pi_{+J})$. There are $J - 1$ free, unknown parameters in the model under H_0 . These are *nuisance* parameters, since what we need to be able to do is find the distribution of cell counts assuming independence, not just for one particular value of $\boldsymbol{\pi}$. [Example?]

The marginal totals (n_{+1}, \dots, n_{+J}) carry all information for $\boldsymbol{\pi}$ – they are *sufficient* for $\boldsymbol{\pi}$. By conditioning on these sufficient statistics (which can lead to a UMP test), we end up with the pmf of the cell counts n_{ij} ,

$$p(n_{ij}) = \frac{\prod_{i=1}^I n_{i+}! \prod_{j=1}^J n_{+j}!}{n_{++}! \prod_{i=1}^I \prod_{j=1}^J n_{ij}!}.$$

This is the distribution of $\{n_{ij}\}$ from data having the same fixed marginals n_{+1}, \dots, n_{+J} and n_{1+}, \dots, n_{I+} as the observed data, assuming $H_0 : X \perp Y$ is true.

A simple way to approximate an exact p -value for an observed X_o^2 statistic is to simply randomly generate IJ cell counts $\{n_{ij}\}$ according to the above pmf, say 1000 times, and compute $X_1^2, X_2^2, \dots, X_{1000}^2$. The proportion of $\{X_m^2\}$ larger than the observed X_o^2 is the (Monte Carlo) exact p -value. The test is the same for multinomial sampling.

Example: a sparse table where the approximate $\chi^2_{(I-1)(J-1)}$ assumption is unreasonable.

Outcome	Smoking level		
	0 /day	1 – 24 / day	> 25 / day
Control (no heart attack)	25	25	12
Heart attack	0	1	3

```

data table;
  input Smoking$ Outcome$ count @@;
  datalines;
1 1 25 2 1 25 3 1 12 1 2 0 2 2 1 3 2 3
;
proc format;
  value $sc '1' = '0 / day' '2' = '1-24 / day' '3' = '>25 / day';
  value $oc '1' = 'No heart attack' '2' = 'Heart attack';
proc freq order=data; weight count;
  format Smoking $sc. Outcome $oc.;
  tables Smoking*Outcome / plcorr;
  exact chisq;
run;

```

Statistics for Table of Smoking by Outcome

Statistic	DF	Value	Prob
Chi-Square	2	6.9562	0.0309
Likelihood Ratio Chi-Square	2	6.6901	0.0353

WARNING: 50% of the cells have expected counts less than 5.
 (Asymptotic) Chi-Square may not be a valid test.

Pearson Chi-Square Test

Chi-Square	6.9562
DF	2
Asymptotic Pr > ChiSq	0.0309
Exact Pr >= ChiSq	0.0516

Likelihood Ratio Chi-Square Test

Chi-Square	6.6901
DF	2
Asymptotic Pr > ChiSq	0.0353
Exact Pr >= ChiSq	0.0724

Statistic	Value	ASE
Gamma	0.8717	0.1250
Pearson Correlation	0.2999	0.0973
Polychoric Correlation	0.6754	0.1924

Comments:

- SAS provides a warning on the small expected cell counts.
- Exact versus asymptotic tests provide different conclusions at the 5% level!
- Treating (X, Y) as ordinal shows a positive association between the number of cigarettes smoked and getting a heart attack using γ , Pearson ρ (using scores 1,2 and 1,2,3), and polychoric ρ . We would reject than any of these are zero.
- To get Monte Carlo estimate, specify MC with EXACT. Also possible to get exact CI for θ in 2×2 table with OR.
- The Pearson correlation is actually bounded away from -1 and 1 . Outside the scope of the class, but $r = 0.30$ may be “larger” than it appears.

3.7 Extensions...

- Ideas for testing independence, partitioning G^2 , std. Pearson residuals, etc. all generalize to threeway and higher dimensional tables.
- Often only interested in one outcome – i.e. one categorical variable is a natural Y . Logistic, Poisson, ordinal regression models useful here. Can also consider continuous predictors.
- If interested in types of conditional dependence in larger dimensional tables, log-linear models (and associated graph methods) useful.
- Often data are not given in the form of a table or counts; see p. 102.
- Methods and ideas in this chapter can be recast in modeling framework explored in the rest of the book.