Sections 3.4, 3.5

Timothy Hanson

Department of Statistics, University of South Carolina

Stat 770: Categorical Data Analysis

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 $u_1 \leq u_2 \leq \cdots \leq u_I$ for X and $v_1 \leq v_2 \leq \cdots \leq v_J$ for Y. Define

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

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

r is the Pearson correlation

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 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 $H_0 : \rho = 0$.

Happiness and political ideology

Data (p. 83) from 2008 General Social Survey for subjects over 65 years old:

	Happiness		
Ideology	Not too happy	Pretty happy	Very happy
Liberal	13	29	15
Moderate	23	59	47
Conservative	14	67	54

SAS code

```
data table;
input Ideology$ Happiness$ count @@;
datalines;
Liberal      NotTooHappy 13 Liberal      PrettyHappy 29 Liberal      VeryHappy 15
Moderate     NotTooHappy 23 Moderate     PrettyHappy 59 Moderate     VeryHappy 47
Conservative NotTooHappy 14 Conservative PrettyHappy 67 Conservative VeryHappy 54
;
proc freq data=table order=data; weight count;
 tables Ideology*Happiness / chisq expected measures plcorr norow nocol;
run;
```

Recall that `chisq` gives tests of $H_0 : X \perp Y$. `measures` gives various measures of association, including r and $\hat{\gamma}$, as well as their (asymptotic) standard errors. `plcorr` gives the estimated polychoric correlation $\hat{\rho}_{pc}$.

SAS output

```
          Table of Ideology by Happiness

   Ideology     Happiness

   Frequency|
   Expected |
   Percent  |NotTooHa|PrettyHa|VeryHapp|  Total
   ---------+--------+--------+--------+
   Liberal  |     13 |     29 |     15 |     57
            | 8.8785 | 27.523 | 20.598 |
            |   4.05 |   9.03 |   4.67 |  17.76
   ---------+--------+--------+--------+
   Moderate |     23 |     59 |     47 |    129
            | 20.093 |  62.29 | 46.617 |
            |   7.17 |  18.38 |  14.64 |  40.19
   ---------+--------+--------+--------+
   Conserva |     14 |     67 |     54 |    135
            | 21.028 | 65.187 | 48.785 |
            |   4.36 |  20.87 |  16.82 |  42.06
   ---------+--------+--------+--------+
   Total          50      155      116      321
               15.58    48.29    36.14   100.00

  Statistics for Table of Ideology by Happiness

Statistic                     DF       Value      Prob
------------------------------------------------------
Chi-Square                     4      7.0681    0.1323
Likelihood Ratio Chi-Square    4      7.2666    0.1225
```

We *do not reject* H_0 : happiness is independent of ideology using X^2 or G^2.

SAS output

```
              Statistics for Table of Ideology by Happiness

Statistic                                      Value       ASE
------------------------------------------------------------------
Gamma                                         0.1849    0.0779
Pearson Correlation                           0.1352    0.0544
Polychoric Correlation                        0.1671    0.0690

                          Sample Size = 321
```

- 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, i.e. no evidence of a *monotone association*.
- $\hat{\gamma} = 0.185$. 95% CI given by $\hat{\gamma} \pm 1.96\text{se}(\hat{\gamma}) = 0.185 \pm 1.96(0.078) = (0.032, 0.338)$. We reject $H_0 : \gamma = 0$ at the 5% level! How to get p-value?
- $r = 0.135$ using default scores $u_i \in \{1, 2, 3\}$ and $v_i \in \{1, 2, 3\}$. Note that we reject $H_0 : \rho_P = 0$ at the 5% level. Focusing on the linear aspect of the scores helped refine our assessment of the relationship between ideology and happiness. Note that you cannot get M^2 directly in SAS, but rather r.

SAS output

Statistics for Table of Ideology by Happiness

Statistic	Value	ASE
Gamma	0.1849	0.0779
Pearson Correlation	0.1352	0.0544
Polychoric Correlation	0.1671	0.0690

Sample Size = 321

- $\hat{\rho}_{pc} = 0.167$ and we reject $H_0 : \rho_{pc} = 0$ as well at the 5% level. The underlying continuous 'happiness' and 'ideology' variables are significantly, positively associated.
- The general test of $H_0 : X \perp Y$ does not reject, but the correlation tests *do find an association* at the 5% level. More power by treating the data as ordinal rather than nominal!

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.

3.4.6 relationship between drinking during pregnancy & 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*.

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 2^{nd} 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. Midranks are used is SAS by specifying `scores=rank`.

3.5 & 16.5.2 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 $\boldsymbol{\pi} = (\pi_{+1}, \pi_{+2}, \ldots, \pi_{+J})$. 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 $\boldsymbol{\pi}$.

Conditioning on sufficient statistics

The marginal totals $(n_{+1}, \ldots, n_{+J})$ 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 $n_{+1}, \ldots, n_{+J}$ and $n_{1+}, \ldots, n_{I+}$ 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 $X_1^2, X_2^2, \ldots, X_{1000}^2$. 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.

Smoking and heart attacks

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

	Smoking level		
Outcome	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;
```

SAS output

```
           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_P (using scores 1,2 and 1,2,3), and polychoric ρ_{pc}. 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.

Fisher's exact test of $H_0 : \pi_1 = \pi_2$ for 2×2 tables

Example: A 7-year old child thinks that cats like gouda cheese more than dogs; she decides to try feeding cats and dogs gouda cheese and records whether they eat it. Her null hypothesis is that cats and dogs prefer gouda in the same proportions, $H_0 : \pi_c = \pi_d$. She wants to show the alternative $H_a : \pi_c > \pi_d$.

In her neighborhood there are 5 cats and 8 dogs nearby. Of the 5 cats, 2 eat the cheese; of the 8 dogs, 2 eat the cheese. We have $\hat{\pi}_c = 0.40$ and $\hat{\pi}_d = 0.25$ for the estimated proportions of cats and dogs that eat gouda cheese. There appears to be some evidence that cats like gouda more than dogs, but is it *significant*?

	eat cheese?		
animal	yes	no	total
cat	2	3	5
dog	2	6	8
total	4	9	13

P-value under $H_0 : \pi_c = \pi_d$

Under the null H_0 we cannot tell the difference between dogs and cats; we only "see" n_{+1} cheese eating animals and n_{+2} non-cheese eaters. If we pick out any $n_{1+} = 5$ animals without replacement, then the probability that there are exactly $n_{11} = k$ cheese eaters is hypergeometric:

$$P(n_{11} = k) = \frac{\begin{pmatrix} n_{+1} \\ k \end{pmatrix} \begin{pmatrix} n_{+2} \\ n_{1+} - k \end{pmatrix}}{\begin{pmatrix} n_{++} \\ n_{1+} \end{pmatrix}}.$$

Here, the sample size $n_{1+} = 5$ is fixed, as well as the number of cheese-eaters n_{+1}. Hence, all four marginal totals are fixed.

Restated: We draw n_{1+} balls without replacement from an urn that has n_{+1} white balls (cheese eaters) and n_{+2} black balls (non-cheese eaters). The number of white balls (cheese eaters) in this sample is $n_{11} = k$.

Fisher's exact test p-value

To compute the p-value, we find the probability of seeing sample $\hat{\pi}_c$ and $\hat{\pi}_d$ *at least* as far apart as what we observed. Fixing the row and column totals, there are three tables that give differences $\hat{\pi}_c - \hat{\pi}_d$ the same or greater than $\hat{\pi}_c - \hat{\pi}_d = 0.15$:

animal	eat cheese? yes	no	total
cat	2	3	5
dog	2	6	8
total	4	9	13

$\hat{\pi}_c = 0.40$, $\hat{\pi}_d = 0.25$

$$\frac{\binom{4}{2}\binom{9}{3}}{\binom{13}{5}} = 0.3916$$

animal	eat cheese? yes	no	total
cat	3	2	5
dog	1	7	8
total	4	9	13

$\hat{\pi}_c = 0.60$, $\hat{\pi}_d = 0.125$

$$\frac{\binom{4}{3}\binom{9}{2}}{\binom{13}{5}} = 0.1119$$

animal	eat cheese? yes	no	total
cat	4	1	5
dog	0	8	8
total	4	9	13

$\hat{\pi}_c = 0.80$, $\hat{\pi}_d = 0.00$

$$\frac{\binom{4}{4}\binom{9}{1}}{\binom{13}{5}} = 0.0070.$$

The p-value is $0.3916 + 0.1119 + 0.0070 = 0.5105$. We do not have evidence that there is an association between type of pet and whether they eat gouda.

SAS code & output

```
data cheese;
input animal$ eat$ count @@;
datalines;
cat yes 2 cat no 3
dog yes 2 dog no 6
;
proc freq order=data; weight count;
  tables animal*eat;
  exact fisher;
run;
------------------------------
    Fisher's Exact Test
Cell (1,1) Frequency (F)     2
Left-sided Pr <= F      0.8811
Right-sided Pr >= F     0.5105

Table Probability (P)   0.3916
Two-sided Pr <= P       1.0000
```

An especially nice feature of Fisher's exact test is that it is natural to have one-sided alternatives.

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. 101.
- Methods and ideas in this chapter can be recast in modeling framework explored in the rest of the book.