Testing for Marginal Independence Among Two or More Multiple Response Categorical Variables

Christopher R. Bilder
Department of Statistics
Oklahoma State University
www.chrisbilder.com
bilder@okstate.edu

Research supported by NSF grant SES-0207212

Multiple-response categorical variables
- Purpose: Analyze survey data that arises from questions that ask “Choose all that apply” or “pick any” from a set of c predefined items
  - Multiple-response categorical variables (MRCVs)
  - Pick any/c variables – Coombs (1964)
- Survey of 279 Kansas farmers conducted by the Department of Animal Sciences at Kansas State University
  - What are your primary sources of veterinary information? Pick all that apply:
    - Professional consultant
    - Veterinarian
    - State or local extension service
    - Magazines
    - Feed companies and representatives

Multiple-response categorical variables
- Survey of 279 Kansas farmers
  - What swine waste disposal methods do you use? Pick all that apply:
    - Lagoon
    - Pit
    - Natural drainage
    - Holding tank
Multiple-response categorical variables

Survey of 279 Kansas farmers

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Lagoon</td>
<td>34</td>
<td>54</td>
<td>50</td>
<td>63</td>
<td>41</td>
</tr>
<tr>
<td>Pit</td>
<td>17</td>
<td>33</td>
<td>34</td>
<td>43</td>
<td>37</td>
</tr>
<tr>
<td>Natural Drainage</td>
<td>6</td>
<td>23</td>
<td>30</td>
<td>49</td>
<td>34</td>
</tr>
<tr>
<td>Holding Tank</td>
<td>1</td>
<td>4</td>
<td>4</td>
<td>6</td>
<td>2</td>
</tr>
</tbody>
</table>

- Farmers can be represented in more than one cell of the table.
- Marginal table
- Are the sources of veterinary information and waste storage methods independent?
  - The “usual” Pearson chi-square test for independence should not be used!
- Main focus of this talk is to develop procedures to test for independence between two MRCVs

Other questions in the survey

- What methods of waste disposal do you use?
  - Injection of liquid swine waste, surface spreading, lagoon oxidation-breakdown, diversion terraces, dirt lots
  - Which of the following do you test your swine waste for?
    - Nitrogen, phosphorus, salt
- Test for independence among more than two multiple–response categorical variables!
- “Pick any” questions are not just limited to swine waste!
  - Ethnicity – 2000 census allowed more than one
  - Soft drinks (Holbrook, Moore, and Winer, 1982)
  - Reasons for supporting or opposing death penalty (Gallup Org., 2000)
  - Contraceptives (Foxman et al., 1997)

Goals of NSF grant research is to parallel similar models and tests typically performed in categorical data analysis

- What types of hypotheses would be of interest?
- What does independence between MRCVs mean?
- What types of models to use?

Past research

- Only one multiple-response categorical variable
- Test for multiple marginal independence (MMI)
  - Test for marginal independence between one multiple-response and one single-response categorical variable
  - Agresti and Liu (*Biometrics*, 1999)
  - Thomas and Decady (*Biometrics*, 2000)
  - Bilder and Loughin (*Biometrics*, 2001)
- Test for conditional multiple marginal independence (CMMI)
  - Test for MMI within strata
  - Similar to a Cochran-Mantel-Haenszel test
  - Bilder and Loughin (*Biometrics*, 2002)
Marginal independence – two variables (SPMI)

- Marginal independence testing between two MRCVs
  - Let $W$ and $Y$ denote the multiple response categorical variables
    - $W =$ swine waste storage method
    - $Y =$ sources of veterinary information
  - Let $W_i$ for $i=1,\ldots,r$ denote the “row” variable items
    - Item refers to a level of the multiple-response categorical variable
    - $W_1$ is lagoon, $W_2$ is pit, …
    - $W_i = 1$ if subject picks item (positive response)
    - $W_i = 0$ if subject does not pick item (negative response)
  - Let $Y_j$ for $j=1,\ldots,c$ be similarly defined for the “column” items
  - The set of subject responses is a vector of correlated binary responses
    - $(W_1,\ldots,W_r)'$ and $(Y_1,\ldots,Y_c)'$

- Let $\pi_{ij} = P(W_i=1 \text{ and } Y_j=1)$
  - $\pi_i = P(W_i=1)$
  - $\pi_j = P(Y_j=1)$

Hypothesis test for marginal independence between $W$ and $Y$

- $H_0$: $\pi_{ij} = \pi_i \cdot \pi_j$ for $i=1,\ldots,r$ and $j=1,\ldots,c$
- $H_1$: At least one of the equalities does not hold
  - “Marginal” since only concerned about $W_i$ and $Y_j$

Marginal independence – two variables (SPMI)

- Agresti and Liu (Biometrics, 1999) first called this a test for “simultaneous pairwise marginal independence” (SPMI)
  - Independence is simultaneously being tested in $r \times c$ tables
  - Kansas farmer survey data

<table>
<thead>
<tr>
<th>Waste Storage Method</th>
<th>Sources of Veterinary Information</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lagoon</td>
<td>Professional consultant</td>
</tr>
<tr>
<td></td>
<td>Veterinarian</td>
</tr>
<tr>
<td></td>
<td>State/local ext. service</td>
</tr>
<tr>
<td></td>
<td>Magazines</td>
</tr>
<tr>
<td></td>
<td>Feed comp. &amp; rep.</td>
</tr>
<tr>
<td>Lagoon</td>
<td>1</td>
</tr>
<tr>
<td>Pit</td>
<td>4</td>
</tr>
<tr>
<td>Natural Draining</td>
<td>6</td>
</tr>
<tr>
<td>Holding Tank</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Waste Storage Method</th>
<th>Sources of Veterinary Information</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lagoon</td>
<td>Professional consultant</td>
</tr>
<tr>
<td></td>
<td>Veterinarian</td>
</tr>
<tr>
<td></td>
<td>State/local ext. service</td>
</tr>
<tr>
<td></td>
<td>Magazines</td>
</tr>
<tr>
<td></td>
<td>Feed comp. &amp; rep.</td>
</tr>
<tr>
<td>Lagoon</td>
<td>54</td>
</tr>
<tr>
<td>Pit</td>
<td>54</td>
</tr>
<tr>
<td>Natural Draining</td>
<td>34</td>
</tr>
<tr>
<td>Holding Tank</td>
<td>1</td>
</tr>
</tbody>
</table>

Marginal independence – two variables (SPMI)

- Odds ratio form of SPMI
  - The $W_i$ and $Y_j$ $2 \times 2$ table

<table>
<thead>
<tr>
<th>$W_i$</th>
<th>$Y_j$</th>
<th>$\pi_{ij}$</th>
<th>$\pi_i$</th>
<th>$\pi_j$</th>
<th>$1-\pi_i$</th>
<th>$1-\pi_j$</th>
<th>$1-\pi_{ij}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>$\pi_{ij}$</td>
<td>$\pi_i-\pi_{ij}$</td>
<td>$\pi_j$</td>
<td>$1-\pi_i$</td>
<td>$1-\pi_j$</td>
<td>$1-\pi_{ij}$</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>$\pi_{ij}$</td>
<td>$\pi_i-\pi_{ij}$</td>
<td>$\pi_j$</td>
<td>$1-\pi_i$</td>
<td>$1-\pi_j$</td>
<td>$1-\pi_{ij}$</td>
</tr>
</tbody>
</table>

- Let $\text{OR}_{WY,ij} = \frac{\pi_{ij}(1-\pi_i - \pi_j + \pi_{ij})}{(\pi_i - \pi_{ij})(\pi_j - \pi_{ij})}$
- Hypotheses
  - $H_0: \text{OR}_{WY,ij} = 1$ for $i=1,\ldots,r$ and $j=1,\ldots,c$
  - $H_1$: At least one of the equalities does not hold
Marginal independence – two variables (SPMI)

- Joint table
  - 1 – farmer picks item; 0 farmer does not pick item

- Why not just test for independence in the joint table?
  - Joint independence \( \Rightarrow \) SPMI (marginal independence)
  - Joint independence \( \not\Rightarrow \) SPMI (marginal independence)
  - Number of parameters under independence
    - \( r + c \) for SPMI
    - \( 2r + 2c \) for joint independence
  - Sparse joint table is the norm

Let \( G \) be a \( r \times 2^c \) matrix containing all possible values of \( (Y_1, \ldots, Y_c)' \)
- Column headers in the joint table
- Kansas farmer example

Let \( H \) be a \( c \times 2^c \) matrix containing all possible values of \( (Y_1, \ldots, Y_c)' \)

Multinomial sampling in the joint table
- Let \( \tau_{gh} = \text{probability of observing the } g^{th} (W_1, \ldots, W_r)' \text{ and } h^{th} (Y_1, \ldots, Y_c)' \)
- \( \sum_g \sum_h \tau_{gh} = 1 \)

Let \( \pi = (\pi_{11}, \ldots, \pi_{rc})' \) and \( \tau = (\tau_{11}, \ldots, \tau_{2^c2^c})' \)

Then \( (G \otimes H)\tau = \pi \)

If \( g' \) is the \( i^{th} \) row of \( G \) and \( h' \) is the \( j^{th} \) row of \( H \), then \( (g' \otimes h')\tau = \pi_{ij} \)
Modified Pearson statistic

  - Let \( n \) be the sample size
    \[ \hat{\pi}_i = \frac{\text{# positive responses to } W_i \text{ and } Y_j}{n} \]
    \[ \hat{\pi}_j = \frac{\text{# positive responses to } Y_j}{n} \]
  - Positive = subject picks an item
  - Note that for the Kansas farmer data:
    \[ \hat{\pi}_{11} = \frac{34}{279} = 0.12 \]
    \[ \hat{\pi}_i = \frac{34 + 109}{279} = 0.51 \]
    \[ \hat{\pi}_j = \frac{34 + 10}{279} = 0.16 \]
  - \( X_M^2 = n \sum_{i=1}^{r} \sum_{j=1}^{c} \left( \hat{\pi}_{ij} - \hat{\pi}_i \hat{\pi}_j \right)^2 \)

\[
X_M^2 = 28.27
\]

\[
X_M^2 = 11.52
\]

\[
X_M^2 = 16.44
\]

\[
X_M^2 = 6.08
\]
Modified Pearson statistic

- Proposed “modified” Pearson statistic
  - If the “usual” Pearson statistics for each of the rc 2×2 tables, say \( X^2_{s,ij} \), are summed, the same statistic results!
  - Example tables:
    
    | Professional consultant | 1 | 0 | 10 |
    |--------------------------|---|---|----|
    | Lagoon                   | 34| 109|
    | 0                       | 10| 126|

- If each \( X^2_{s,ij} \) is naively treated as independent, \( X^2_s \) can be approximated by a \( \chi^2 \) random variable.
- Reject SPMI if \( X^2_s > \chi^2_{rc,1-\alpha} \)
- In most cases, each \( X^2_{s,ij} \) is NOT independent.

Modified Pearson statistic

- Specific form of \( \Sigma \)
  - Note: \( \sqrt{n}(\hat{\tau} - \tau) \xrightarrow{d} N(0, \text{Diag}(\tau - \tau')) \)
  - Let \( \pi^R = (\pi_1, \ldots, \pi_r)' \) and \( \pi^C = (\pi_{c1}, \ldots, \pi_{rc})' \)
  - \( \Sigma = F[\text{Diag}(\tau - \tau')]F' \) under SPMI

\[
F = G \otimes H - \pi^R \otimes [H(j^c_2 \otimes I^c_2) - [G(I^c_2 \otimes j^c_2)] \otimes \pi^C
\]

| I_2 denotes an a×a identity matrix and j_2 denotes an a×1 vector of 1's |

- Note that \( \Sigma \) will still depend on the \( \tau_{gh} \) under the hypothesis of SPMI
- For example, the (1,2) element of \( \Sigma \) when \( r = c = 2 \) is
  
  \[
  \text{AsCov}
  \left[
  \sqrt{n} \left( \hat{\pi}_{11} - \hat{\pi}_{1c} \hat{\pi}_{c1} \right), \sqrt{n} \left( \hat{\pi}_{12} - \hat{\pi}_{1c} \hat{\pi}_{c2} \right)
  \right]
  = (\pi_{1c} - 1)^2 (\tau_{3d} + \tau_{4d}) + \pi_{1c}^2 (\tau_{1d} + \tau_{2d}) + \pi_{1c} \pi_{c2} (\pi_{1c} - 1)
  \]
- Remember sparseness in the joint table!

Modified Pearson statistic

- Proposed “modified” Pearson statistic
  - Asymptotic distribution of \( X^2_s \) under SPMI is a linear combination of independent \( \chi^2_1 \)
  - \( X^2_s = \sum_{i=1}^{rc} \sum_{j=1}^{c} (\hat{\pi}_{ij} - \hat{\pi}_{i.} \hat{\pi}_{.j})^2 \)
  
  \[
  d \rightarrow \sum_{i=1}^{rc} \lambda_i X^2_i
  \]

where \( X_i^2 \) are independent \( \chi^2_1 \)

- \( \lambda_i \) are the eigenvalues of \( D^{-1} \Sigma \)

\[
D = \text{Diag}[\pi_{1a}(1-\pi_{1a})(1-\pi_{a})]
\]

\( \Sigma \) denote the asymptotic covariance matrix for

\[
\sqrt{n}
\begin{bmatrix}
\hat{\pi}_{11} - \hat{\pi}_{1c} \hat{\pi}_{c1} \\
\hat{\pi}_{12} - \hat{\pi}_{1c} \hat{\pi}_{c2} \\
\vdots \\
\hat{\pi}_{rc} - \hat{\pi}_{r.} \hat{\pi}_{.c}
\end{bmatrix}
\]

- Varieties of ways to proceed!

First-order corrected statistic

- Similar to what Rao and Scott (1981, JASA) did for Pearson chi-square statistics in complex sampling designs

\[
\text{Find } \delta \text{ such that } E\left[ \delta \sum \lambda_i X^2_i \right] = rc
\]

\[
\delta = rc \sum_{p=1}^{rc} \lambda_p
\]

\( \sum_{p=1}^{rc} \lambda_p = \text{tr}(D^{-1} \Sigma) \)

- Since \( D = \text{Diag}[\pi_{1a}(1-\pi_{1a})(1-\pi_{a})] \) is a diagonal matrix, only the diagonal elements of \( \Sigma \) are needed!
Modified Pearson statistic

- **First-order corrected statistic**
  - Asymptotic variance of \( \sqrt{n}(\hat{\pi}_i - \hat{\pi}_i\hat{\pi}_i) \) under SPMI
  - \( \sqrt{n}(\hat{\pi}_i - \hat{\pi}_i\hat{\pi}_i) = f(\tau) = (g' \otimes h')\tau - [g(I_z \otimes I_z')]\tau[h'(j_z \otimes I_z')\tau] \)
  - \( \hat{g}' \) is the \( i \)th row of \( G \) and \( \hat{h}' \) is the \( j \)th row of \( H \)
  - Asymptotic variance is
  - When the above expression is multiplied out, eighteen different terms result
  - Simplify using relationships between \( \tau \) and \( \pi \) and incorporate SPMI
  - Obtain \( \pi_i\pi_j(1-\pi_i)(1-\pi_j)! \)

- **Second-order corrected statistic**

Modified Pearson statistic

- **Bootstrap \( X^2_b \)**
  - Decompose the data into binary “item response” vectors for row and column MRCVs
  - \( W=(W_1,...,W_r)' \) and \( Y=(Y_1,...,Y_c)' \)
  - \( (1,0,1,0) \) means item 1 and item 3 were picked
  - Take B resamples of size n by randomly selecting \( W \) and \( Y \) independently
  - Resampling under the special case of null hypothesis
  - For each resample, calculate the test statistic, \( X^2_{b,i,j} \) for \( b=1,...,B \)
  - **P-value** = \( \frac{1}{B} \sum_{b=1}^{B} I(X^2_{b,i,j} > X^2_b) \)
  - where \( I(A)=1 \) if event A occurs, 0 otherwise

Modified Pearson statistic

- **Bootstrap p-value combination methods**
  - Combine the p-values from \( X^2_{b,i,j} \) (using a \( \chi^2 \) app.) for \( i=1,...,r \) and \( j=1,...,c \) to form a “new” test statistic
  - Product of the p-values or minimum p-value - \( \tilde{p} \)
  - P-values are likely to be correlated
    - Usual p-value combination methods based on independence are not appropriate
    - Combine p-values of correlated tests - Pesarin (1999)
  - **Algorithm**
    - Resample in the same manner as before
    - Calculate \( \tilde{p}_b \) for each resample
    - **P-value** = \( \frac{1}{B} \sum_{b=1}^{B} I(\tilde{p}_b < \tilde{p}) \)
**Modified Pearson statistic**

- Bonferroni
  - Reject SPMI if \( \max(X_{S,ij}^2) > \chi_{1-\alpha/rc}^2 \)
  - P-value = \( P(X^2 > \max(X_{S,ij}^2)) \cdot rc \) where \( X^2 \sim \chi_i^2 \)

**Kansas farmer survey example**

- Evidence against marginal independence (SPMI)
  - 10,000 resamples for bootstrap methods
  - Use covariance matrix without SPMI restriction

**Follow-up analysis**

- Determine why reject SPMI
- Use a \( \chi^2 \) approximation with each \( X_{S,ij}^2 \)
  - Using a 0.05 significance level, the significant combinations are (W1, Y1), (W1, Y2), (W2, Y2), (W2, Y5), (W3, Y1), and (W3, Y4)
  - Bonferroni adjusted significance level of 0.05/20 produces (W1, Y1) = (Lagoon, Professional consultant)

**Model-based approaches summary**

- Why?
  - Model may give a nice way to interpret deviations from SPMI
- Generalized loglinear models
  - Lang and Agresti (1994, JASA) – MLE of \( \tau \)
  - Haber (1986, Biometrics) – WLS
- Random effect models
  - Agresti and Liu (1998, FL tech report)
    - Found the models to can produce a poor fit for MMI
    - Suggest using multivariate binomial logit-normal models (Coull and Agresti, Biometrics 2000)
    - \( r+c \) dimension numerical integration needed

**Model-based approaches summary**

- GEE
  - Since examining the pairwise associations, need to specify the marginal and pairwise expectations of \( W_i \) and \( Y_j \)
  - Alternating logistic regression procedure of Carey, Zeger, and Diggle (1993, Biometrika)
  - Need large \( n \) for Wald test of SPMI to hold the correct size
Simulations

Type I error
- Estimated type I error rate: Proportion of data sets in which SPMI is incorrectly rejected
- Data generated under SPMI using an algorithm by Gange (1995)
  - Specify $\pi^R = (\pi_{1r}, \ldots, \pi_{rr})'$ and $\pi^C = (\pi_{1c}, \ldots, \pi_{cc})'$
  - Specify odds ratios
    - Under SPMI: $\text{OR}_{W,j} = \frac{\pi_{ij}(1 - \pi_{ij} + \pi_{ji})}{(\pi_{ij} - \pi_{ji})(\pi_{ii} - \pi_{ij})}$
    - Within W or Y
      $\text{OR}_{W,j} = \frac{P(W_i = 1 \text{ and } W_i = 1) / P(W_i = 1 \text{ and } W_i = 0)}{P(W_i = 0 \text{ and } W_i = 1) / P(W_i = 0 \text{ and } W_i = 0)}$
      $\text{OR}_{Y,j} = \frac{P(Y_i = 1 \text{ and } Y_i = 1) / P(Y_i = 1 \text{ and } Y_i = 0)}{P(Y_i = 0 \text{ and } Y_i = 1) / P(Y_i = 0 \text{ and } Y_i = 0)}$

Simulations

Type I error
- Settings held constant for each simulation
  - Nominal type I error rate=0.05
  - 500 data sets generated
  - 1,000 resamples for bootstrap methods
  - Expected range of estimated type I error rates for methods holding the nominal level:
    $0.05 \pm 2 \sqrt{\frac{(0.05)(1-0.05)}{500}} = 0.05 \pm 0.0195$
- Trellis plot on next slide shows estimated type I error rates
  - Includes only some of the cases examined
  - Results generalize to other cases

Simulations

Type I error
- Bonferroni can be a little conservative with 5x5 tables
- Second-order corrected $X^2$ can also be a little conservative with 5x5 tables
- Bootstrap methods consistently hold the correct size
Simulations

Power
- Excluded $X^2$ with a $\chi^2$ approximation
- Proportion of data sets in which SPMI is correctly rejected
- Data generated same way as in the type I error simulation study except that $OR_{WY,ij} \neq 1$
- Conclusions:
  - There is not one best procedure

Conclusions:
- Some p-value combination methods are better at detecting certain types of alternative hypotheses
- Deviation from SPMI for only a few $OR_{WY,ij}$; higher power:
  - Minimum p-value has higher power
  - Bonferroni
- Deviation from SPMI for most $OR_{WY,ij}$ by the same degree; higher power:
  - Product of p-values
  - Bootstrap $X^2$

Recommendations
- Use the bootstrap methods
- Bonferroni and 2nd order corrected $X^2$ work well also

Simulations

Recommendations

More than two MRCVs

What types of hypotheses would be of interest?
- Consider 3 multiple response categorical variable case
  - Let $\mathbf{V} = (V_1, V_2, \ldots, V_k)'$
  - $\pi_{ijk} = P(W_i=1, Y_j=1, V_k=1)$
- Pairwise independence
  - $\pi_{ij} = \pi_{i\cdot \cdot \cdot j\cdot \cdot \cdot}$, $\pi_{i\cdot \cdot \cdot k\cdot \cdot \cdot} = \pi_{i\cdot \cdot \cdot \cdot \cdot k\cdot \cdot \cdot}$, and $\pi_{\cdot \cdot \cdot j\cdot \cdot \cdot k\cdot \cdot \cdot} = \pi_{\cdot \cdot \cdot j\cdot \cdot \cdot \cdot \cdot k\cdot \cdot \cdot}$
- Complete independence
  - $\pi_{ijk} = \pi_{i\cdot \cdot \cdot j\cdot \cdot \cdot k\cdot \cdot \cdot}$
  - Extend modified Pearson statistic
  - Model based approaches?
Further Work

- Estimation and model based approaches
- Complex sampling designs
- Randomized response
  - Sensitive questions – ask two ways with known probability
    - What drugs do you use?
    - What drugs do you not use?
  - Observe response without knowing which question was asked
    - Protects identity of subject
- Include ordinal single response categorical variables
  - Ordered alternative hypothesis

References