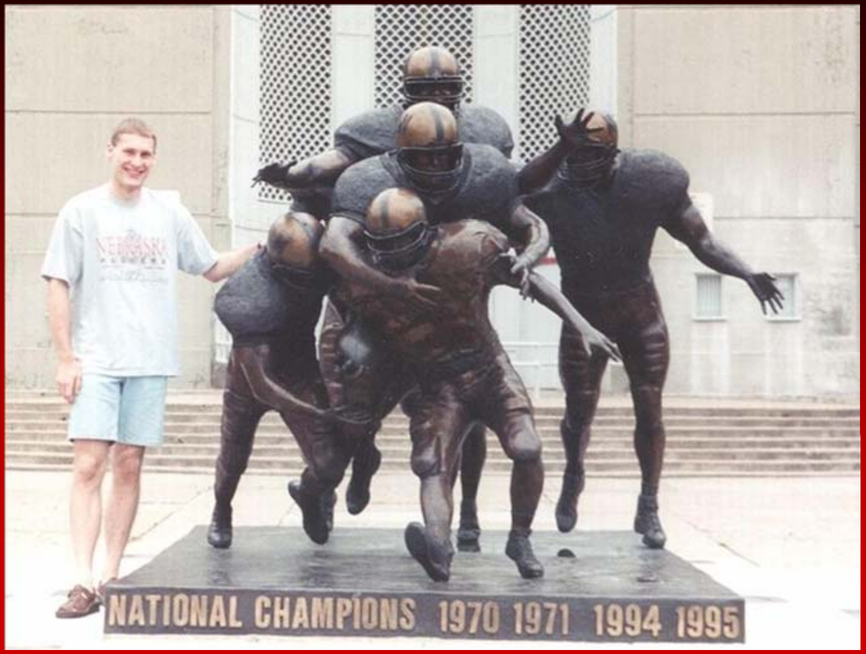
Testing for Marginal Independence Among Two or More Multiple Response Categorical Variables

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

Research supported by NSF grant SES-0207212

Helping to tackle a K-Stater



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

Waste Storage Method	Professional consultant	Veterinarian	State/local ext. service	Magazines	Feed comp. & rep.
	Sources of veterinary information				
Lagoon	34	54	50	63	41
Pit	17	33	34	43	37
Natural Drainage	6	23	30	49	34
Holding Tank	1	4	4	6	2

  - 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

Multiple-response categorical variables

- 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)

Multiple-response categorical variables

- 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
  - Loughin and Scherer (*Biometrics*, 1998)
  - Agresti and Liu (*Biometrics*, 1999)
  - Bilder, Loughin, and Nettleton (*Comm. Stat.: Comp & Sim.*, 2000)
  - 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,..,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)
- Y_j for j=1,…,c is similarly defined for the “column” items
- The set of subject responses is a vector of correlated binary responses
  - $(W_1,\ldots,W_r)'$ and $(Y_1,\ldots,Y_c)'$

Marginal independence – two variables (SPMI)

- Let $\pi_{ij} = P(W_i=1 \text{ and } Y_j=1)$
  $\pi_{i\bullet} = P(W_i=1)$
  $\pi_{\bullet j} = P(Y_j=1)$
- Hypothesis test for marginal independence between W and Y is
  - H_o: $\pi_{ij}=\pi_{i\bullet}\pi_{\bullet j}$ for i=1,…,r and j=1,…,c
    H_a: 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 rc 2×2 tables
  - Kansas farmer survey data

Waste Storage Method	Sources of veterinary information: Professional consultant	Veterinarian	State/local ext. service	Magazines	Feed comp. & rep.
Lagoon	34	54	50	63	41
Pit	17	33	34	43	37
Natural Drainage	6	23	30	49	34
Holding Tank	1	4	4	6	2

Lagoon	Veterinarian 1	Veterinarian 0	
1	54	89	
0	36	100	
			279

- 1=farmer picked item
  0=farmer did not pick item

Marginal independence – two variables (SPMI)

- Odds ratio form of SPMI
  - The W_i and Y_j 2×2 table

		Y_j = 1	Y_j = 0	
W_i	1	π_{ij}	$\pi_{i\bullet}-\pi_{ij}$	$\pi_{i\bullet}$
	0	$\pi_{\bullet j}-\pi_{ij}$	$1-\pi_{i\bullet}-\pi_{\bullet j}+\pi_{ij}$	$1-\pi_{i\bullet}$
		$\pi_{\bullet j}$	$1-\pi_{\bullet j}$	1

  - Let $OR_{WY,ij} = \dfrac{\pi_{ij}(1-\pi_{i\bullet}-\pi_{\bullet j}+\pi_{ij})}{(\pi_{i\bullet}-\pi_{ij})(\pi_{\bullet j}-\pi_{ij})}$
  - Hypotheses
    H_o: $OR_{WY,ij}$=1 for i=1,…,r and j=1,…,c
    H_a: 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

			Y_1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
			Y_2	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1
			Y_3	0	0	0	0	1	1	1	1	0	0	0	0	1	1	1	1	0	0	0	0	1	1	1	1	0	0	0	0	1	1	1	1
			Y_4	0	0	1	1	0	0	1	1	0	0	1	1	0	0	1	1	0	0	1	1	0	0	1	1	0	0	1	1	0	0	1	1
			Y_5	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1
0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	1	0	2	3	0	2	0	1	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	1	0	2	12	19	7	12	0	1	0	4	3	0	0	0	0	3	5	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	1	0	0	4	7	7	2	4	0	6	2	4	2	1	1	0	1	0	3	3	0	0	0	0	0	0	0	0	0	0	0	0	0	1	3
0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	1	1	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	0	0	11	9	22	0	10	3	2	2	13	1	1	3	2	0	3	3	15	0	0	0	1	0	0	0	1	0	0	0	1	0	2	3
1	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	1	0	0	0	2	0	0	0	0	0	1	0	0	0	0	0	1	2	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1
1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	1	0	0	0	2	1	0	2	0	1	0	2	4	0	0	0	0	0	2	1	0	1	0	0	0	0	0	1	0	0	0	0	0	0	4
1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
1	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
W_1	W_2	W_3	W_4																																

  - 434 of 2^9=512 cells contain a 0

Marginal independence – two variables (SPMI)

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

Marginal independence – two variables (SPMI)

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

			Y_1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
			Y_2	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1
			Y_3	0	0	0	0	1	1	1	1	0	0	0	0	1	1	1	1	0	0	0	0	1	1	1	1	0	0	0	0	1	1	1	1
			Y_4	0	0	1	1	0	0	1	1	0	0	1	1	0	0	1	1	0	0	1	1	0	0	1	1	0	0	1	1	0	0	1	1
			Y_5	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1
0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	1	0	2	3	0	2	0	1	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	1	0	2	12	19	7	12	0	1	0	4	3	0	0	0	0	3	5	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	1	0	0	4	7	7	2	4	0	6	2	4	2	1	1	0	1	0	3	3	0	0	0	0	0	0	0	0	0	0	0	0	0	1	3
0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	1	1	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	0	0	11	9	22	0	10	3	2	2	13	1	1	3	2	0	3	3	15	0	0	0	1	0	0	0	1	0	0	0	1	0	2	3
1	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	1	0	0	0	2	0	0	0	0	0	1	0	0	0	0	0	1	2	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1
1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	1	0	0	0	2	1	0	2	0	1	0	2	4	0	0	0	0	0	2	1	0	1	0	0	0	0	0	1	0	0	0	0	0	0	4
1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
1	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
W_1	W_2	W_3	W_4																																

		Sources of veterinary information				
		Professional consultant	Veterinarian	State/local ext. service	Magazines	Feed comp. & rep.
Waste Storage Method	Lagoon	34	54	50	63	41
	Pit	17	33	34	43	37
	Natural Drainage	6	23	30	49	34
	Holding Tank	1	4	4	6	2

Marginal independence – two variables (SPMI)

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

Y_1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Y_2	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1
Y_3	0	0	0	0	1	1	1	1	0	0	0	0	1	1	1	1	0	0	0	0	1	1	1	1	0	0	0	0	1	1	1	1
Y_4	0	0	1	1	0	0	1	1	0	0	1	1	0	0	1	1	0	0	1	1	0	0	1	1	0	0	1	1	0	0	1	1
Y_5	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1

- Let **G** be a r×2^r matrix containing all values of $(W_1,\ldots,W_r)'$
- Multinomial sampling in the joint table
  - Let τ_{gh} = probability of observing the g^{th} $(W_1,\ldots,W_r)'$ and h^{th} $(Y_1,\ldots,Y_c)'$
  - $\sum_g \sum_h \tau_{gh} = 1$
- Let $\pi = (\pi_{11},\ldots,\pi_{rc})'$ and $\tau = (\tau_{11},\ldots,\tau_{2^r2^c})'$
- Then $(\mathbf{G}\otimes\mathbf{H})\tau=\pi$
- If $\mathbf{g}_i'$ is the i^{th} row of **G** and $\mathbf{h}_j'$ is the j^{th} row of **H**, then $(\mathbf{g}_i'\otimes\mathbf{h}_j')\tau = \pi_{ij}$

Modified Pearson statistic

- Loughin (1998, *KSU tech. report*)
  - Let n be the sample size
    - $\hat{\pi}_{ij}$ = [# positive responses to W_i and Y_j]/n
    - $\hat{\pi}_{i\cdot}$ = [# positive responses to W_i]/n
    - $\hat{\pi}_{\cdot j}$ = [# positive responses to Y_j]/n
    - Positive = subject picks an item
  - Note that for the Kansas farmer data:

$$\hat{\pi}_{11} = 34/279 = 0.12$$
$$\hat{\pi}_{1\cdot} = (34+109)/279 = 0.51$$
$$\hat{\pi}_{\cdot 1} = (34+10)/279 = 0.16$$

		Professional consultant 1	Professional consultant 0	
Lagoon	1	34	109	
Lagoon	0	10	126	
				279

  - $X_M^2 = n\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{\left(\hat{\pi}_{ij}-\hat{\pi}_{i\cdot}\hat{\pi}_{\cdot j}\right)^2}{\hat{\pi}_{i\cdot}\hat{\pi}_{\cdot j}}$

Modified Pearson statistic

- Loughin (1998, *KSU tech. report*)
  - Problem: Not invariant to how "positive" responses are summarized
    - Switch definition: W_i=0 for positive, W_i=1 for negative
    - Positive could mean "do not" pick an item
    - X_M^2 can have 4 different values!!!!

$X_M^2 = 28.27$

Waste Stor. Method	Sources of veterinary information: Professional consultant	Veterinarian	State/local ext. service	Magazines	Feed comp. & rep.
Lagoon	34	54	50	63	41
Pit	17	33	34	43	37
Natural Drainage	6	23	30	49	34
Holding Tank	1	4	4	6	2

		Professional consultant 1	Professional consultant 0
Lagoon	1	34	109
Lagoon	0	10	126

$X_M^2 = 11.52$

Waste Stor. Method	Sources of veterinary information (not chosen): Professional consultant	Veterinarian	State/local ext. service	Magazines	Feed comp. & rep.
Lagoon	109	89	93	80	102
Pit	63	47	46	37	43
Natural Drainage	79	62	55	36	51
Holding Tank	12	9	9	7	11

$X_M^2 = 16.44$

Waste Stor. Method (not)	Sources of veterinary information: Professional consultant	Veterinarian	State/local ext. service	Magazines	Feed comp. & rep.
Lagoon	10	36	45	68	52
Pit	27	57	61	88	56
Natural Drainage	38	67	65	82	59
Holding Tank	43	86	91	125	91

$X_M^2 = 6.08$

Waste Stor. Method (not)	Sources of veterinary information (not chosen): Professional consultant	Veterinarian	State/local ext. service	Magazines	Feed comp. & rep.
Lagoon	126	100	91	68	84
Pit	172	142	138	111	143
Natural Drainage	156	127	129	112	135
Holding Tank	223	180	175	141	175

Modified Pearson statistic

- Proposed "modified" Pearson statistic
  - Sum the four different statistics to form an invariant statistic
  - 2×2 item response table

		Y_j 1	Y_j 0	
W_i	1	π_{ij}	$\pi_{i\bullet}-\pi_{ij}$	$\pi_{i\bullet}$
W_i	0	$\pi_{\bullet j}-\pi_{ij}$	$1-\pi_{i\bullet}-\pi_{\bullet j}+\pi_{ij}$	$1-\pi_{i\bullet}$
		$\pi_{\bullet j}$	$1-\pi_{\bullet j}$	1

$$X_S^2 = \overbrace{n\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(\hat{\pi}_{ij}-\hat{\pi}_{i\bullet}\hat{\pi}_{\bullet j})^2}{\hat{\pi}_{i\bullet}\hat{\pi}_{\bullet j}}}^{X_M^2} + n\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{[\hat{\pi}_{i\bullet}-\hat{\pi}_{ij}-\hat{\pi}_{i\bullet}(1-\hat{\pi}_{\bullet j})]^2}{\hat{\pi}_{i\bullet}(1-\hat{\pi}_{\bullet j})}$$
$$+ n\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{[\hat{\pi}_{\bullet j}-\hat{\pi}_{ij}-\hat{\pi}_{\bullet j}(1-\hat{\pi}_{i\bullet})]^2}{\hat{\pi}_{\bullet j}(1-\hat{\pi}_{i\bullet})} + n\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{[1-\hat{\pi}_{i\bullet}-\hat{\pi}_{\bullet j}+\hat{\pi}_{ij}-(1-\hat{\pi}_{i\bullet})(1-\hat{\pi}_{\bullet j})]^2}{(1-\hat{\pi}_{i\bullet})(1-\hat{\pi}_{\bullet j})}$$
$$= n\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(\hat{\pi}_{ij}-\hat{\pi}_{i\cdot}\hat{\pi}_{\cdot j})^2}{\hat{\pi}_{i\cdot}\hat{\pi}_{\cdot j}(1-\hat{\pi}_{i\cdot})(1-\hat{\pi}_{\cdot j})}$$

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	Professional consultant 0
Lagoon	1	34	109
Lagoon	0	10	126

		Veterinarian 1	Veterinarian 0
Lagoon	1	54	89
Lagoon	0	36	100

    - $X^2_S = \sum_{i=1}^{r}\sum_{j=1}^{c} X^2_{S,ij}$
  - If each $X^2_{S,ij}$ is naively treated as independent, X^2_S can be approximated by a χ^2_{rc} 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

- Proposed "modified" Pearson statistic
  - Asymptotic distribution of X^2_S under SPMI is a linear combination of independent χ^2_1
    - $X^2_S = n\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(\hat{\pi}_{ij}-\hat{\pi}_{i\bullet}\hat{\pi}_{\bullet j})^2}{\hat{\pi}_{i\bullet}\hat{\pi}_{\bullet j}(1-\hat{\pi}_{i\bullet})(1-\hat{\pi}_{\bullet j})} \xrightarrow{d} \sum_{i=1}^{rc}\lambda_i X_i^2$

where X_i^2 are independent χ^2_1

λ_i are the eigenvalues of $\mathbf{D}^{-1}\Sigma$

$\mathbf{D} = \text{Diag}[\pi_{i\bullet}\pi_{\bullet j}(1-\pi_{i\bullet})(1-\pi_{\bullet j})]$

Σ denote the asymptotic covariance matrix for

$$\sqrt{n}\begin{bmatrix} \hat{\pi}_{11}-\hat{\pi}_{1\bullet}\hat{\pi}_{\bullet 1} \\ \hat{\pi}_{12}-\hat{\pi}_{1\bullet}\hat{\pi}_{\bullet 2} \\ \vdots \\ \hat{\pi}_{rc}-\hat{\pi}_{r\bullet}\hat{\pi}_{\bullet c} \end{bmatrix}$$

Modified Pearson statistic

- Specific form of Σ
  - Note: $\sqrt{n}(\hat{\tau}-\tau) \xrightarrow{d} N(\mathbf{0}, \text{Diag}(\tau)-\tau\tau')$
  - Let $\pi^R = (\pi_{1\bullet},\ldots,\pi_{r\bullet})'$ and $\pi^C = (\pi_{\bullet 1},\ldots,\pi_{\bullet c})'$
  - $\Sigma = \mathbf{F}[\text{Diag}(\tau)-\tau\tau']\mathbf{F}'$ under SPMI where

$$\mathbf{F} = \mathbf{G}\otimes\mathbf{H} - \pi^R \otimes [\mathbf{H}(\mathbf{j}'_{2^r}\otimes \mathbf{I}_{2^c})] - [\mathbf{G}(\mathbf{I}_{2^r}\otimes \mathbf{j}'_{2^c})]\otimes \pi^C$$

$\mathbf{I}_a$ denotes an a×a identity matrix and $\mathbf{j}_a$ denotes an a×1 vector of 1's

  - Note that Σ will still depend on the τ_{gh} under the hypothesis of SPMI
    - For example, the (1,2) element of Σ when r=c=2 is

$$\text{AsCov}\left[\sqrt{n}(\hat{\pi}_{11}-\hat{\pi}_{1\bullet}\hat{\pi}_{\bullet 1}), \sqrt{n}(\hat{\pi}_{12}-\hat{\pi}_{1\bullet}\hat{\pi}_{\bullet 2})\right]$$
$$= (\pi_{1\bullet}-1)^2(\tau_{34}+\tau_{44}) + \pi_{1\bullet}^2(\tau_{14}+\tau_{24}) + \pi_{1\bullet}\pi_{\bullet 1}\pi_{\bullet 2}(\pi_{1\bullet}-1)$$

    - Remember sparseness in the joint table!

Modified Pearson statistic

- Notes about $X^2_S \xrightarrow{d} \sum_{i=1}^{rc}\lambda_i X_i^2$ where λ_i are the eigenvalues of $\mathbf{D}^{-1}\Sigma$ and X_i^2 are independent χ^2_1
  - $\mathbf{D}^{-1}\Sigma$ is generally not idempotent
  - λ_i generally are not 1
  - Generally should not use χ^2_{rc} approximation!
- Variety 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
  - Find δ such that $E\left[\delta\sum\lambda_i X_i^2\right] = rc$
  - $\delta = rc \Big/ \sum_{p=1}^{rc}\lambda_p$
  - $\sum_{p=1}^{rc}\lambda_p = \text{tr}(\mathbf{D}^{-1}\Sigma)$
  - Since $\mathbf{D} = \text{Diag}[\pi_{i\bullet}\pi_{\bullet j}(1-\pi_{i\bullet})(1-\pi_{\bullet j})]$ is a diagonal matrix, only the diagonal elements of Σ are needed!

Modified Pearson statistic

- First-order corrected statistic
  - Asymptotic variance of $\sqrt{n}(\hat{\pi}_{ij} - \hat{\pi}_{i\bullet}\hat{\pi}_{\bullet j})$ under SPMI
  - $\sqrt{n}(\pi_{ij} - \pi_{i\bullet}\pi_{\bullet j}) = f(\tau) = (\mathbf{g}_i' \otimes \mathbf{h}_j')\tau - [\mathbf{g}_i'(I_{2^r} \otimes \mathbf{j}_{2^c}')\tau][\mathbf{h}_j'(\mathbf{j}_{2^r}' \otimes I_{2^c})\tau]$
    - $\mathbf{g}_i'$ is the i^{th} row of **G** and $\mathbf{h}_j'$ is the j^{th} row of **H**
    - $(\mathbf{g}_i' \otimes \mathbf{h}_j')\tau = \pi_{ij}$
  - Asymptotic variance is $\dot{f}(\tau)\left[\text{Diag}(\tau) - \tau\tau'\right]\dot{f}(\tau)'$
    $= \left\{\mathbf{g}_i' \otimes \mathbf{h}_j' - \pi_{i\bullet}[\mathbf{h}_j'(\mathbf{j}_{2^r}' \otimes I_{2^c})] - \pi_{\bullet j}[\mathbf{g}_i'(I_{2^r} \otimes \mathbf{j}_{2^c}')]\right\}\left\{\text{Diag}(\tau) - \tau\tau'\right\}$
    $\left\{\mathbf{g}_i \otimes \mathbf{h}_j - \pi_{i\bullet}[(\mathbf{j}_{2^r} \otimes I_{2^c})\mathbf{h}_j] - \pi_{\bullet j}[(I_{2^r} \otimes \mathbf{j}_{2^c})\mathbf{g}_i]\right\}$
  - When the above expression is multiplied out, eighteen different terms result
  - Simplify using relationships between τ and π and incorporate SPMI
    - Obtain $\pi_{i\bullet}\pi_{\bullet j}(1-\pi_{i\bullet})(1-\pi_{\bullet j})$!

Modified Pearson statistic

- First-order corrected statistic
  - $\text{tr}(\mathbf{D}^{-1}\Sigma) = \sum_{i=1}^{r}\sum_{j=1}^{c}\left[\pi_{i\bullet}\pi_{\bullet j}(1-\pi_{i\bullet})(1-\pi_{\bullet j})\right]^{-1}\pi_{i\bullet}\pi_{\bullet j}(1-\pi_{i\bullet})(1-\pi_{\bullet j}) = rc$
    $\delta = rc \Big/ \sum_{p=1}^{rc}\lambda_p = rc/\text{tr}(\mathbf{D}^{-1}\Sigma) = 1$
  - Thus, X_S^2 is self-correcting!
- Second-order corrected statistic
  - Find a constant δ such that $\delta\sum_{i=1}^{rc}\lambda_i X_i^2 \Big/ E\left(\sum_{i=1}^{rc}\lambda_i X_i^2\right)$ has the same mean and variance as a χ_δ^2 random variable
  - $\delta = r^2c^2/\sum\lambda_i^2$
  - Corrected statistic is $rcX_S^2/\sum\hat{\lambda}_i^2$
  - Approximate by a χ^2 distribution with $r^2c^2/\sum\hat{\lambda}_i^2$ degrees of freedom
  - No nice simplification for $\sum\lambda_i^2$

Modified Pearson statistic

- Bootstrap X_S^2
  - Decompose the data into binary “item response” vectors for row and column MRCVs
    - **W**=$(W_1,\ldots,W_r)'$ and **Y**=$(Y_1,\ldots,Y_c)'$
    - (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, $X_{S,b}^{2*}$, for b=1,…,B
  - P-value $= \frac{1}{B}\sum_{b=1}^{B} I\left(X_{S,b}^{2*} > X_S^2\right)$

    where $I(A)=1$ if event A occurs, 0 otherwise

Modified Pearson statistic

- Bootstrap p-value combination methods
  - Combine the p-values from $X_{S,ij}^2$ (using a χ_1^2 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 $\max(X^2_{S,ij}) > \chi^2_{1-\alpha/rc}$
  - P-value = $P(X^2 > \max(X^2_{S,ij})) * rc$ where $X^2 \sim \chi^2_1$

Kansas farmer survey example

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

SPMI Testing Method	P-value
X^2_S using χ^2_{rc} app.	$3.11*10^{-6}$
2nd order corrected X^2_S	$3.07*10^{-5}$
Bootstrap X^2_S	<0.0001
Bootstrap prod. p-values	0.0001
Bootstrap min. p-values	0.0034
Bonferroni	0.0037

- Follow-up analysis
  - Determine why reject SPMI
  - Use a χ^2_1 approximation with each $X^2_{S,ij}$
    - Using a 0.05 significance level, the significant combinations are (W_1, Y_1), (W_1, Y_2), (W_2, Y_2), (W_2, Y_5), (W_3, Y_1), and (W_3, Y_4)
    - Bonferroni adjusted significance level of 0.05/20 produces (W_1, Y_1) = (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 τ
  - 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
  - Agresti and Liu (1998 tech report, 2001 *Soc. Meth & Res.*)
    - 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 assocations, 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 $\pi^R = (\pi_{1\cdot},\ldots,\pi_{r\cdot})'$ and $\pi^C = (\pi_{\cdot 1},\ldots,\pi_{\cdot c})'$
    - Specify odds ratios
      - Under SPMI: $OR_{WY,ij} = \dfrac{\pi_{ij}(1-\pi_{i\cdot}-\pi_{\cdot j}+\pi_{ij})}{(\pi_{i\cdot}-\pi_{ij})(\pi_{\cdot j}-\pi_{ij})} = 1$
      - Within W or Y

$$OR_{W,ii'} = \frac{P(W_i=1 \text{ and } W_{i'}=1)/P(W_i=1 \text{ and } W_{i'}=0)}{P(W_i=0 \text{ and } W_{i'}=1)/P(W_i=0 \text{ and } W_{i'}=0)}$$

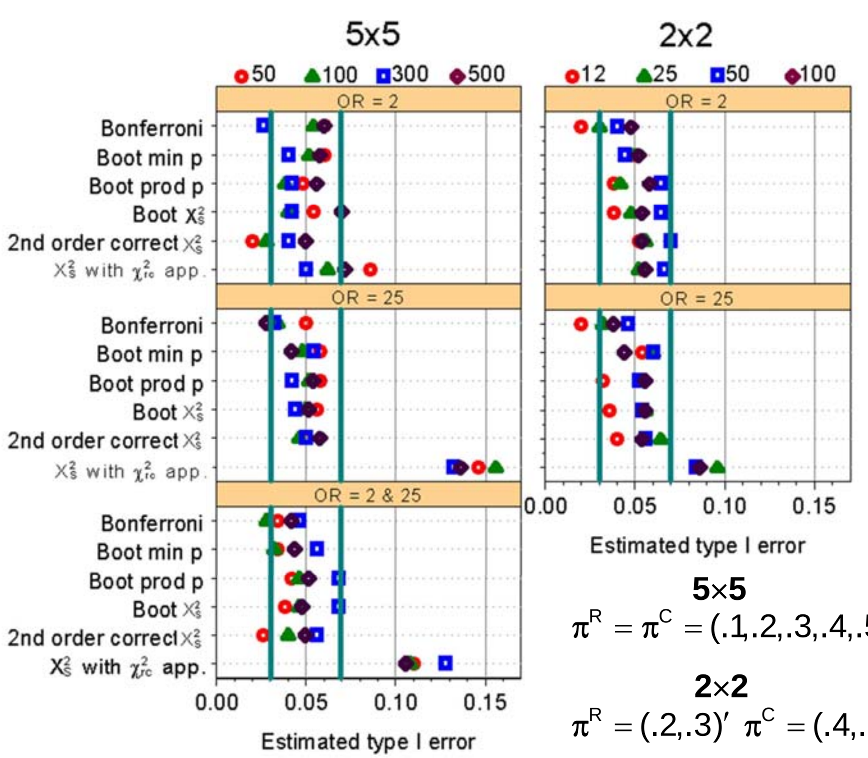
$$OR_{Y,jj'} = \frac{P(Y_j=1 \text{ and } Y_{j'}=1)/P(Y_j=1 \text{ and } Y_{j'}=0)}{P(Y_j=0 \text{ and } Y_{j'}=1)/P(Y_j=0 \text{ and } Y_{j'}=0)}$$

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



5×5

$\pi^R = \pi^C = (.1,.2,.3,.4,.5)'$

2×2

$\pi^R = (.2,.3)'$ $\pi^C = (.4,.5)'$

Simulations

- Type I error
  - X^2_S with a χ^2_{rc} approximation (first-order corrected) does not hold the correct size if there is strong pairwise association between items for W or items for Y.
  - Bonferroni can be a little conservative with 5×5 tables
  - Second-order corrected X^2_S can also be a little conservative with 5×5 tables
  - Bootstrap methods consistently hold the correct size

Simulations

- Power
  - Excluded X^2_S with a χ^2_{rc} 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

Simulations

- Power
  - 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_S

Recommendations

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

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
    - $\pi_{ij\bullet} = \pi_{i\bullet\bullet}\pi_{\bullet j\bullet}$, $\pi_{i\bullet k} = \pi_{i\bullet\bullet}\pi_{\bullet\bullet k}$, and $\pi_{\bullet jk} = \pi_{\bullet j\bullet}\pi_{\bullet\bullet k}$
  - Complete independence
    - $\pi_{ijk} = \pi_{i\bullet\bullet}\pi_{\bullet j\bullet}\pi_{\bullet\bullet k}$
  - 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

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

Go Big Red!

References

- Agresti, A. and Liu, I.-M. (1998). *Modeling responses to a categorical variable allowing arbitrarily many category choices.* Technical Report 575, University of Florida, Department of Statistics, Gainesville, FL.
- Agresti, A. and Liu, I.-M. (1999). Modeling a Categorical Variable Allowing Arbitrarily Many Category Choices. *Biometrics* 55, 936-943.
- Agresti, A. and Liu, I.-M. (2001). Strategies for modeling a categorical variable allowing multiple category choices. *Sociological Methods & Research* 29, 403-434.
- Bilder, C. R. and Loughin, T. M. (2001). On the First-order Rao-Scott Correction of the Umesh-Loughin-Scherer Statistic. *Biometrics* 57. 1253-1255.
- Bilder, C. R. and Loughin, T. M. (2002). Testing for Conditional Multiple Marginal Independence. *Biometrics*. 200-208.
- Bilder, C. R., Loughin, T. M., Nettleton, D. (2000). Multiple Marginal Independence Testing for Pick Any/c Variables. To appear in *Communications in Statistics: Simulation and Computation* 29(4).
- Coombs, C. H. (1964). *A theory of data*. New York: John Wiley & Sons, Inc.
- Coull, B. A. and Agresti, A. (2000). Random effects modeling of multiple binomial responses using the multivariate binomial logit-normal distribution. *Biometrics* 56, 73-80.
- Decady, Y. J. and Thomas, D. H. (2000). A simple test of association for contingency tables with multiple column responses. *Biometrics* 56, 893-896.
- Gange, S. J. (1995). Generating multivariate categorical variates using the iterative proportional fitting algorithm. *The American Statistician* 45, 134-138.
- Grizzle J. E., Starmer, C. F., and Koch, G. G. (1969). Analysis of categorical data by linear models. *Biometrics* 25. 489-504.
- Haber, M. (1986). Testing for pairwise independence. *Biometrics* 42, 429-435.
- Lang, J. and Agresti, A. (1994). Simultaneously modeling joint and marginal distributions of multivariate categorical responses. *Journal of the American Statistical Association* 89, 625-632.
- Loughin, T. M. (1998). *Testing for independence in contingency tables with multiple row and column responses.* Technical Report, Kansas State University, Department of Statistics, Manhattan, KS.
- Loughin, T. M. and Scherer, P. N. (1998). Testing for Association in Contingency Tables with Multiple Categorical Responses. *Biometrics* 54, 630-637.
- Pesarin, F. (1999). *Permutational testing of multiple hypotheses by nonparametric combinations of dependent tests.* Padova, Italy: Cleup Editrice.
- Rao, J. N. K. and Scott, A. J. (1981). The analysis of categorical data from complex surveys: chi-squared tests for goodness of fit and independence in two-way tables. *Journal of the American Statistical Association* 76, 221-230.
- Smith, W. R., Smith, D. R., and Noma, E. (1986). The multidimensionality of crime: a comparison of techniques for scaling delinquent careers. Journal of Quantitative Criminology 2, 329-353.