

**Constructing chi-square goodness-of-fit tests for multinomial data  
that are more powerful than Pearson's  $X^2$**

Harry Joe

Department of Statistics, University of British Columbia,

Alberto Maydeu-Olivares

Faculty of Psychology. University of Barcelona

Durham, NH

July 2, 2008

## Background

- Consider the hypotheses
  - Simple:  $H_0 : \boldsymbol{\pi} = \boldsymbol{\pi}_0$  vs.  $H_1 : \boldsymbol{\pi} \neq \boldsymbol{\pi}_0$ .
  - Composite:  $H_0 : \boldsymbol{\pi} = \boldsymbol{\pi}(\boldsymbol{\theta})$  vs.  $H_1 : \boldsymbol{\pi} \neq \boldsymbol{\pi}(\boldsymbol{\theta})$ , where  $\boldsymbol{\theta}$  is a  $q$  parameter vector to be estimated from the data.
- Testing these hypotheses has been in most cases impossible as  $C$  (number of cells) grows large
  - Major problem in multivariate categorical data analysis

## What's the source of the problem?

- Asymptotic  $p$ -values for Pearson's, and likelihood ratio test statistics incorrect when some cell probabilities are small
- But as model get larger, some cell probabilities must be small!!

## Solution: limited information testing

- For simple nulls use: 
$$L_r = N (\mathbf{p}_r - \boldsymbol{\pi}_r)' \boldsymbol{\Xi}_r (\mathbf{p}_r - \boldsymbol{\pi}_r)$$
- For composite nulls use 
$$M_r = N (\mathbf{p}_r - \hat{\boldsymbol{\pi}}_r)' \hat{\mathbf{C}}_r (\mathbf{p}_r - \hat{\boldsymbol{\pi}}_r)$$
- $\boldsymbol{\pi}_r$  is an  $s$ -dimensional vector of joint moments up to order  $r$
- $$\mathbf{C}_r = \boldsymbol{\Xi}_r^{-1} - \boldsymbol{\Xi}_r^{-1} \boldsymbol{\Delta}_r \left( \boldsymbol{\Delta}_r' \boldsymbol{\Xi}_r^{-1} \boldsymbol{\Delta}_r \right)^{-1} \boldsymbol{\Delta}_r' \boldsymbol{\Xi}_r^{-1} = \boldsymbol{\Delta}_r^{(c)} \left( \boldsymbol{\Delta}_r^{(c)'} \boldsymbol{\Xi}_r \boldsymbol{\Delta}_r^{(c)} \right)^{-1} \boldsymbol{\Delta}_r^{(c)'}$$
- $$\mathbf{C}_r = \mathbf{C}_r \boldsymbol{\Sigma}_r \mathbf{C}_r, N \text{ Acov}(\dot{\mathbf{p}}_r - \dot{\boldsymbol{\pi}}_r) = \boldsymbol{\Xi}_r, N \text{ Acov}(\dot{\mathbf{p}}_r - \dot{\hat{\boldsymbol{\pi}}}_r) = \boldsymbol{\Sigma}_r$$
- $$\boldsymbol{\Delta}_r = \frac{\partial \boldsymbol{\pi}_r(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'},$$
- For simple nulls: 
$$L_r \xrightarrow{d} \chi_s^2$$
- For composite nulls, if model is identified from  $\boldsymbol{\pi}_r$ , 
$$M_r \xrightarrow{d} \chi_{s-q}^2$$

## Limited information testing

- Pearson's  $X^2$  and  $\hat{X}^2$  are special cases of these families
- If marginal probabilities are not small, accurate  $p$ -values are obtained even in gigantic models
  - The less information we use the better the asymptotic approximation under the null (and also under sequences of local alternatives)
- Often, higher power is obtained the less information is used
- Recommendation: For composite nulls, use the smallest  $r$  for which the model is identified (usually  $r = 2$ )

## Open questions

- What to do in ultra-gigantic models? (e.g.,  $7^{100}$ )
  - Can we use summary statistics other than marginal probabilities?
- Why do we obtain sometimes more power when discarding information?

## In this paper we give

- Conditions for quadratic forms in  $\boldsymbol{\kappa} = \mathbf{T}_{\boldsymbol{\kappa}} \boldsymbol{\pi}$ , a linear summary statistic of  $\boldsymbol{\pi}$ , to be asymptotically chi-square under simple and composite nulls
- Results involving the asymptotic power of two summary statistics  $\boldsymbol{\kappa}_1$  and  $\boldsymbol{\kappa}_2$ , with  $\boldsymbol{\kappa}_2 = \mathbf{T}_{21} \boldsymbol{\kappa}_1$  being a further reduction of  $\boldsymbol{\kappa}_1$

## Simple null hypotheses

- With  $\boldsymbol{\kappa} = \mathbf{T}_\kappa \boldsymbol{\pi}$ ,  $s$ -dimensional;  $\boldsymbol{\Xi}_\kappa := N \text{Acov}(\hat{\boldsymbol{\kappa}} - \boldsymbol{\kappa}) = N\mathbf{T}_\kappa \text{Acov}(\mathbf{p} - \hat{\boldsymbol{\pi}}) \mathbf{T}_\kappa'$

$$L_\kappa = N(\hat{\boldsymbol{\kappa}} - \boldsymbol{\kappa})' \boldsymbol{\Xi}_\kappa^{-1} (\hat{\boldsymbol{\kappa}} - \boldsymbol{\kappa}) \xrightarrow{d} \chi_s^2$$

provided condition T is satisfied

$$\text{rank}(\mathbf{T}_\kappa) = s \neq \text{rank} \begin{pmatrix} \mathbf{1}_C' \\ \mathbf{T}_\kappa \end{pmatrix}$$

i.e.,  $\mathbf{T}_\kappa$  is of full row rank and  $\mathbf{1}_C'$  is not in its row span

## Composite null hypotheses

- With  $\Delta_{\kappa} = \frac{\partial \kappa(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'} = \mathbf{T}_{\kappa} \frac{\partial \boldsymbol{\pi}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}$ ,

$$\mathbf{C}_{\kappa} = \boldsymbol{\Xi}_{\kappa}^{-1} - \boldsymbol{\Xi}_{\kappa}^{-1} \Delta_{\kappa} \left( \Delta_{\kappa}' \boldsymbol{\Xi}_{\kappa}^{-1} \Delta_{\kappa} \right)^{-1} \Delta_{\kappa}' \boldsymbol{\Xi}_{\kappa}^{-1} = \Delta_{\kappa}^{(c)} \left( \Delta_{\kappa}^{(c)'} \boldsymbol{\Xi}_{\kappa} \Delta_{\kappa}^{(c)} \right)^{-1} \Delta_{\kappa}^{(c)'}$$

$$M_{\kappa} = N \left( \hat{\boldsymbol{\kappa}} - \boldsymbol{\kappa}(\hat{\boldsymbol{\theta}}) \right)' \hat{\mathbf{C}} \left( \hat{\boldsymbol{\kappa}} - \boldsymbol{\kappa}(\hat{\boldsymbol{\theta}}) \right) \xrightarrow{d} \chi_{s-q}^2$$

provided conditions  $T$  and  $D$  are satisfied

- *Condition D:*  $\Delta_{\kappa}$  is of full column rank  $q$  and  $s > q$ 
  - (i.e., the model is identified from  $\boldsymbol{\kappa}$  and  $df = s - q > 0$ )

## Some algebraic results for composite nulls

- Pearson's  $\hat{X}^2$  is a member of  $M_{\kappa}$
- Let  $\kappa_1$  and  $\kappa_2$  be two statistics satisfying conditions  $T$  and  $D$ 
  - If there's a one-to-one linear relationship between both (they're of the same dimension),  $\kappa_2 = \beta + \mathbf{B}\kappa_1$  with  $\mathbf{B}$  invertible,  $M_{\kappa_1} = M_{\kappa_2}$
  - If  $\kappa_2$  is a further reduction of  $\kappa_1$ , with  $\mathbf{M}_{21} = \mathbf{C}_{\kappa_1} - \mathbf{T}_{21}'\mathbf{C}_{\kappa_2}\mathbf{T}_{21}$

$$M_{\kappa_1} - M_{\kappa_2} = \left( \hat{\kappa}_1 - \kappa_1(\hat{\theta}) \right)' \hat{\mathbf{M}}_{21} \left( \hat{\kappa}_1 - \kappa_1(\hat{\theta}) \right) \geq 0$$

equality occurs if  $\hat{\mathbf{M}}_{21} \left( \hat{\kappa}_1 - \kappa_1(\hat{\theta}) \right) = \mathbf{0}$ .

- For any statistic  $\kappa$  satisfying conditions  $T$  and  $D$ ,  $M_{\kappa} \leq \hat{X}^2$

## Asymptotic theory for composite nulls

- We assume a sequence of local alternatives  $\boldsymbol{\pi}(\boldsymbol{\theta}_N)$  such that  $\hat{\boldsymbol{\theta}}_N$  minimizes the Kullback-Liebler distance  $L(\boldsymbol{\theta}_N) = \boldsymbol{\pi}(\boldsymbol{\theta}_N)' \ln(\boldsymbol{\pi}(\boldsymbol{\theta}_N)/\boldsymbol{\pi}(\boldsymbol{\theta}_0))$
- The local direction of the sequence of alternatives is  $\boldsymbol{\delta} = \lim_{N \rightarrow \infty} \sqrt{N} \left( \boldsymbol{\pi}(\boldsymbol{\theta}_N) - \boldsymbol{\pi}(\hat{\boldsymbol{\theta}}_N) \right)$ ;  $\boldsymbol{\delta}$  satisfies  $\boldsymbol{\delta} \mathbf{1}_C' = \mathbf{0}$  and  $\boldsymbol{\Delta}_0 \mathbf{D}_0^{-1} \boldsymbol{\delta} = \mathbf{0}$
- The noncentrality parameter is  $\text{npc}(M_{\kappa}) = \boldsymbol{\delta}'_{\kappa} \mathbf{C}_{\kappa} \boldsymbol{\delta}_{\kappa} = \boldsymbol{\delta}'_{\kappa} \mathbf{T}'_{\kappa} \mathbf{C}_{\kappa} \mathbf{T}_{\kappa} \boldsymbol{\delta}$

## Asymptotic power results for comparing two summary statistics

- Let  $\kappa_1$  and  $\kappa_2$  be two statistics satisfying conditions  $T$  and  $D$ , with  $\kappa_2$  is a further reduction of  $\kappa_1$ , and  $\mathbf{M}_{21} = \mathbf{C}_{\kappa_1} - \mathbf{T}_{21}' \mathbf{C}_{\kappa_2} \mathbf{T}_{21}$
- The difference of two noncentrality parameters,  $\text{ncp}(M_{\kappa_1}) - \text{ncp}(M_{\kappa_2})$ , for  $M_{\kappa_1}$  versus  $M_{\kappa_2}$  is  $\boldsymbol{\delta}_{\kappa_1}' \mathbf{M}_{21} \boldsymbol{\delta}_{\kappa_1}$
- $\text{ncp}(M_{\kappa_1}) - \text{ncp}(M_{\kappa_2}) \geq 0$  with equality possible if  $\mathbf{M}_{21} \boldsymbol{\delta}_{\kappa_1} = \mathbf{0}$ .
- Implications: For a direction of local alternatives  $\boldsymbol{\delta}$  such that
  - $\text{ncp}(M_{\kappa_2}) / \text{ncp}(M_{\kappa_1}) = 1$ ,  $M_{\kappa_2}$  will be more powerful (there are fewer  $df$ )
  - $\text{ncp}(M_{\kappa_2}) / \text{ncp}(M_{\kappa_1})$  is large (.9?),  $M_{\kappa_2}$  may be more powerful
- For some directions  $\boldsymbol{\delta}$ , power may be zero

## Simplest IRT example we could think of

- Null model: Tjur's loglinear version of Rasch model with constant "intercept"
  - 2 parameter model: constant "intercept"  $\gamma$  and constant "slope"  $\sigma$
- True model: Tjur's loglinear version of Rasch model (unconstrained intercepts)
- In discrete exponential family or loglinear models,

$$\Delta = \frac{\partial \pi(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'} = \boldsymbol{\Gamma} \mathbf{X}', \quad \boldsymbol{\Gamma} = \text{diag}(\boldsymbol{\pi}) - \boldsymbol{\pi} \boldsymbol{\pi}'$$

where  $\mathbf{X}'$  is a  $C \times q$  matrix of constants

- $\boldsymbol{\kappa}_x = \mathbf{X} \boldsymbol{\pi}$  provides a minimal set of statistics that identify the model

## Choosing the statistics $\kappa$

- Unfeasible choices:
  - $\kappa_x$  may not be used because  $df = 0$
  - Maydeu-Olivares & Joe's statistics based on moments can not be used because this null model is not identified from moments
  
- Feasible choices
  - 1) Select simple summaries that are not redundant to  $\kappa_x$  (verify condition  $T$ )
  - 2) Compute a basis for the null space of  $\mathbf{X}$ , say  $\mathbf{N}_x$ , which satisfies  $\mathbf{N}_x \mathbf{X}' = \mathbf{0}$ .  
Then, by construction, any subsets of rows of  $\mathbf{N}_x$  can be added to  $\mathbf{X}$  to yield a  $\kappa$  such that conditions  $T$  and  $D$  are satisfied.

## Statistics being compared

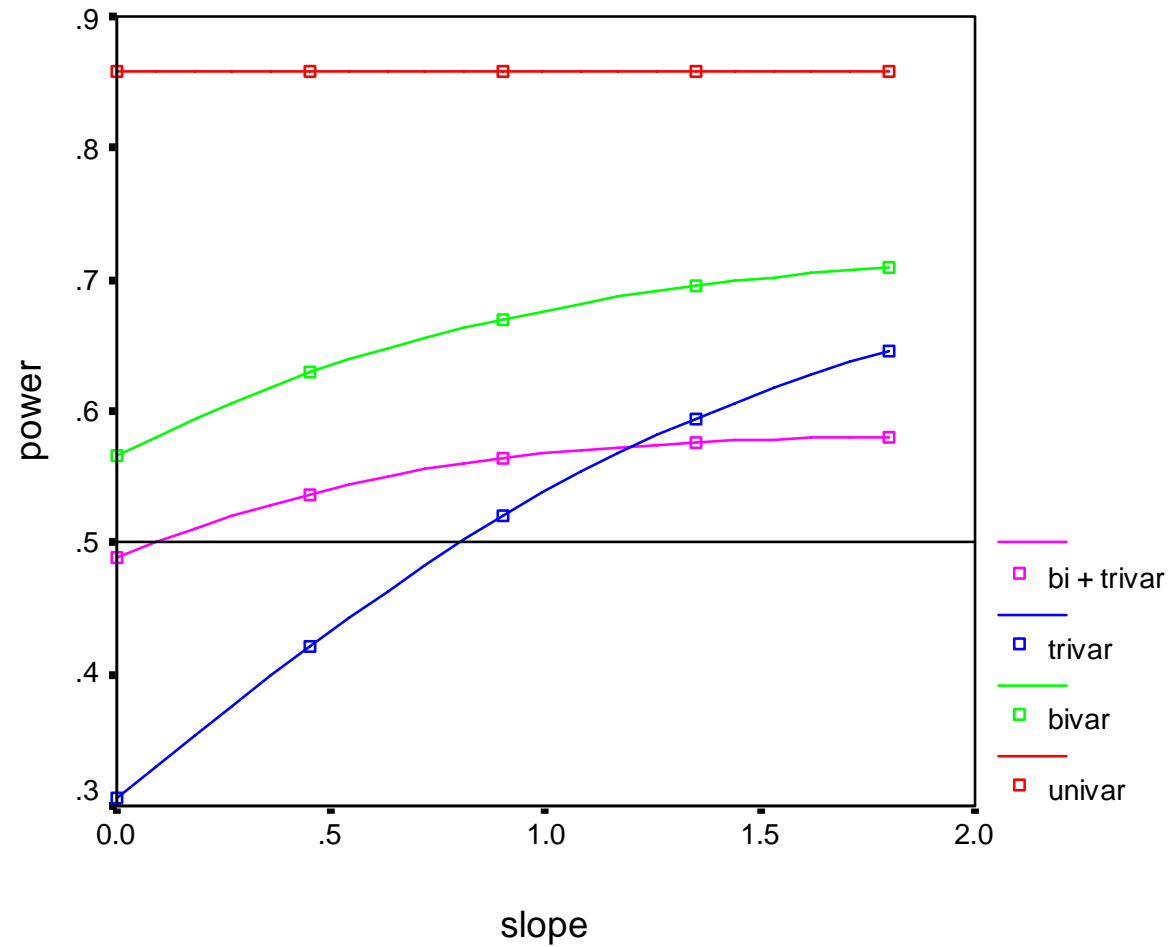
- 1) set of  $E(Y_j = 1) = \Pr(Y_j = 1)$  except for item 1
- 2) set of  $E(Y_j Y_k)$  except for the pair involving items 1 and 2
- 3) set of  $E(Y_j Y_k Y_l)$  except for the triplet involving items 1, 2 and 3
- 4) = 2) + 3)

## Power of $M_{\kappa}$ relative to power of $X^2$ (fixed at .5)

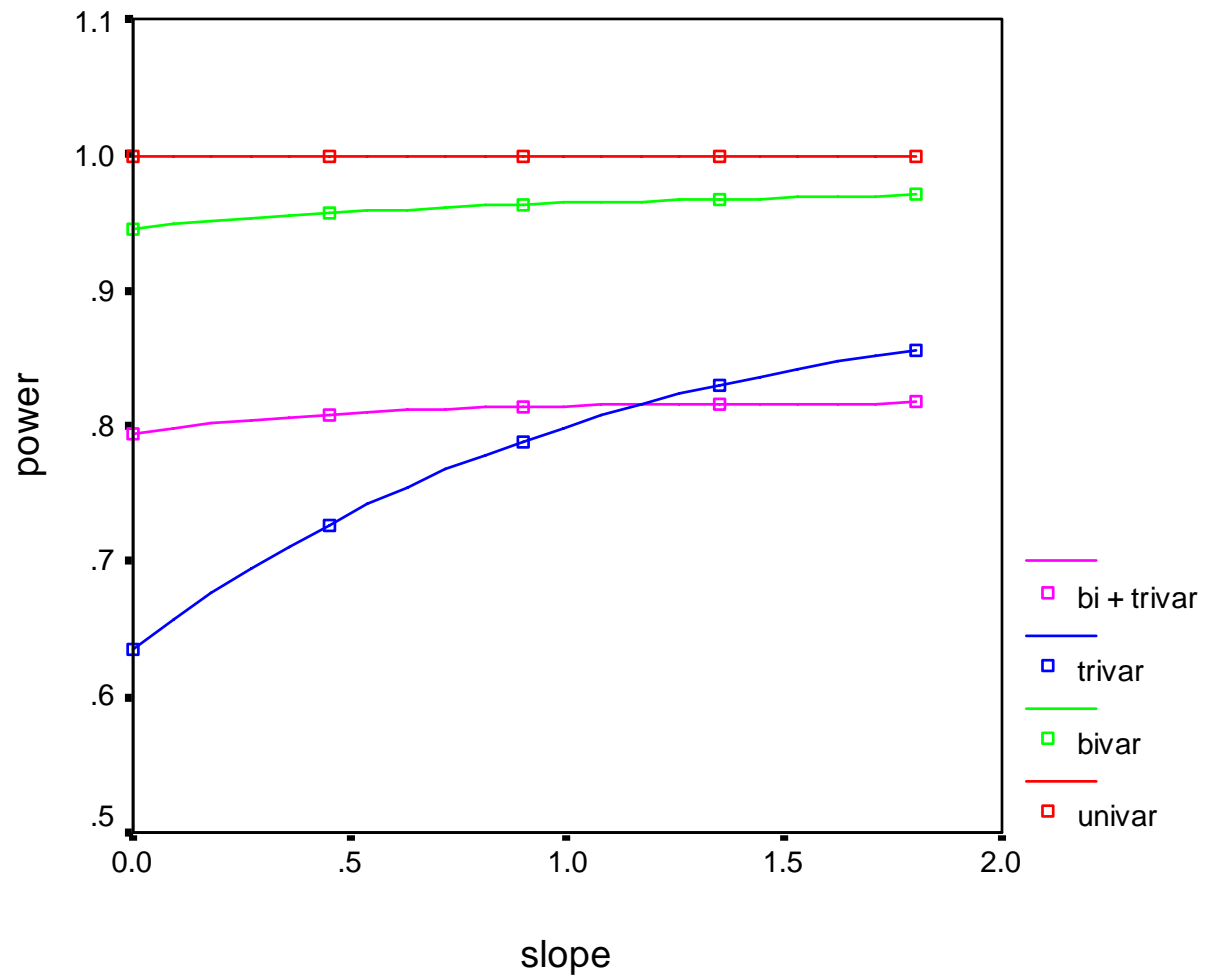
$C = 32$  and  $256$  for  $n = 5$  and  $8$

$n$ ( $s$ )		5 (9)	5 (14)	5 (14)	5 (23)	8 (15)	8 (35)	8 (63)	8 (90)
	$\sigma_0$	stat <sub>1</sub>	stat <sub>2</sub>	stat <sub>3</sub>	stat <sub>4</sub>	stat <sub>1</sub>	stat <sub>2</sub>	stat <sub>3</sub>	stat <sub>4</sub>
ncps $M_{\kappa}/X^2$	0.00	1.000	0.732	0.396	0.835	1.000	0.875	0.624	0.961
	0.45	1.000	0.824	0.540	0.917	1.000	0.918	0.724	0.983
	0.90	1.000	0.888	0.670	0.962	1.000	0.945	0.803	0.993
	1.35	1.000	0.930	0.773	0.984	1.000	0.963	0.864	0.997
	1.80	1.000	0.957	0.849	0.993	1.000	0.976	0.909	0.999
power of $M_{\kappa}$	0.00	0.859	0.566	0.306	0.489	0.999	0.946	0.635	0.795
	0.45	0.859	0.629	0.421	0.537	0.999	0.957	0.727	0.808
	0.90	0.859	0.670	0.521	0.564	0.999	0.963	0.789	0.814
	1.35	0.859	0.695	0.595	0.576	0.999	0.967	0.830	0.816
	1.80	0.859	0.710	0.645	0.581	0.999	0.970	0.856	0.817

## Power of $M_\kappa$ , $n = 5$



# Power of $M_{\kappa}$ , $n = 8$



## Concluding remarks

- It is not surprising to find  $M_\kappa$  statistics that are more powerful than  $\hat{X}^2$  over a variety of local directions
- The new results enable us to develop
  - limited information tests for unidimensional multinomials as well (example for Poisson versus zero-inflation and/or overdispersion)
  - tests for hyper-gigantic models (if it can be estimated, it can be tested)
- **Bottom line:** select the statistic whose distribution under the null is best approximated by asymptotic methods and with highest power for alternatives of interest