Statistical Tests for Parameterized Multinomial Models: Power Approximation and Optimization

Edgar Erdfelder
University of Mannheim
Germany
The Problem

• Typically, model applications start with a global goodness-of-fit test of a base model
  – Substantive theory predicts $H_0$ holds
  – Type-1 error probability is fixed at $\alpha = .05$ or $\alpha = .01$
  – Nothing is known about type-2 error probability $\beta$.

• If insignificant, special parameter tests follow
  – Substantive hypothesis predicts $H_1$ holds (one-tailed)
  – Type-1 error probability is fixed at $\alpha = .05$ or $\alpha = .01$
  – Nothing is known about type-2 error probability $\beta$.

• Conclusions:
  – No rational basis for accepting the model;
  – No rational basis for rejecting substantive hypotheses.
Overview

• An example: The storage-retrieval model
• Other parameterized multinomial models (PMMs)
• Goodness-of-fit tests for PMMs
• Power approximation for PMMs
• Critique of traditional power analysis methods
• Power as a function of the $H_1$ model parameters
• An illustrative example
• Results of a Monte Carlo Study
• Power optimization
• Summary and conclusions
An example:
The storage-retrieval model

• Goal: Measurement of probabilities
  – $c$: Storage as a cluster in episodic memory
  – $r$: Retrieval of a cluster from episodic memory
  – $u$: Storage and retrieval of a singleton

• Paradigm:
  – Free recall of
    • $N_1$ item pairs (lag between items controlled)
    • $N_2$ singletons
Data

- **Word pairs:**
  - $y_{1,1} = \text{freq}(C_{1,1})$: Number of word pairs recalled adjacenty
  - $y_{1,2} = \text{freq}(C_{1,2})$: Number of word pairs recalled not adjacenty
  - $y_{1,3} = \text{freq}(C_{1,3})$: Number of pairs with only one word recalled
  - $y_{1,4} = \text{freq}(C_{1,4})$: Number of pairs with none of the words recalled
  - $y_{1,1} + y_{1,2} + y_{1,3} + y_{1,4} = N_1$

- **Singletons:**
  - $y_{2,1} = \text{freq}(C_{2,1})$: Number of singletons recalled
  - $y_{2,2} = \text{freq}(C_{2,1})$: Number of singletons not recalled
  - $y_{2,1} + y_{2,2} = N_2$
Other Parameterized Multinomial Models (PMMs)

• Examples:
  – Log-linear models
  – Logit models
  – Ogive models
  – Latent-class models
  – Cultural consensus models
  – Signal detection models
  – ...
Assumptions and Notation

• Joint multinomial model:
  – For each population $k$, $k = 1, \ldots, K$, a multinomial model holds with $N_k$ observations and category probabilities $\pi_{k,j} = Pr(C_{k,j})$, $j = 1, \ldots, J_k$, $\pi_{k,1} + \pi_{k,2} + \ldots + \pi_{k,J_k} = 1$.
  – Observations are independent
  – $N = N_1 + \ldots + N_K$
  – $\boldsymbol{\pi} := (\pi_{1,1}, \ldots, \pi_{K,JK})$
  – $\boldsymbol{y} := (y_{1,1}, \ldots, y_{K,JK})$

• Parameterized multinomial model:
  – The probabilities $\pi_{k,j}$ are functions of $S$ real-valued latent parameters $\theta_s$, $s = 1, \ldots, S$, i.e.
  – $\pi_{k,j} = p_{k,j}(\theta_1, \ldots, \theta_S)$
  – $\boldsymbol{\theta} := (\theta_1, \ldots, \theta_S) \in \Omega$
Goodness-of-fit tests for PMMs

Hypothesis: \( H_0: \pi \in f(\Omega) \)

Tool: Power divergence statistic

\[
PD^\lambda(\theta; y) := \frac{2}{\lambda(\lambda + 1)} \cdot \sum_{k=1}^{K} \sum_{j=1}^{J_k} y_{k,j} \left[ \frac{y_{k,j}}{N_k \cdot p_{k,j}(\theta)} \right]^\lambda - 1, \quad -\infty < \lambda < \infty, \lambda \notin \{-1, 0\}
\]

with

\[
PD^{\lambda=0}(\theta; y) := \lim_{\lambda \to 0} PD^\lambda(\theta; y)
\]

\[
PD^{\lambda=-1}(\theta; y) := \lim_{\lambda \to -1} PD^\lambda(\theta; y).
\]
Asymptotic central distribution
(Read & Cressie, 1988, A6)

For any real-valued $\lambda$, the minimum $PD^\lambda$ statistic

$$PD^\lambda(\hat{\theta}_\lambda;y) := \min_{\theta \in \Omega} (PD^\lambda(\theta;y))$$

has an asymptotic central $\chi^2_{(df)}$ distribution under $H_0$ with

$$df = \sum_{k=1}^{K} (J_k - 1) - S$$

provided that Birch’s (1964) regularity conditions hold.
## Special cases

<table>
<thead>
<tr>
<th>$\lambda$</th>
<th>$\hat{\theta}(\lambda)$</th>
<th>$\text{PD}^\lambda(\hat{\theta}(\lambda);y)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Maximum-Likelihood estimator</td>
<td>Log-Likelihood-$\chi^2$ statistic ($G^2$)</td>
</tr>
<tr>
<td>2/3</td>
<td>Minimum $\text{PD}^{2/3}$ estimator</td>
<td>Cressie-Read statistic</td>
</tr>
<tr>
<td>1</td>
<td>Minimum chi-square estimator</td>
<td>Pearson’s $\chi^2$ statistic ($X^2$)</td>
</tr>
</tbody>
</table>
Asymptotic noncentral distribution 
(Mitra, 1958)

• Notation
  – $H_0$ probabilities: $\mathbf{p} = (p_1(\theta), ..., p_J(\theta))$
  – $H_1$ probabilities: $\mathbf{\pi} = (\pi_1, ..., \pi_J)$
  – $d_j = \pi_j - p_j$, $d_1 + ... + d_J = 1$

• Consider the class of local alternative hypotheses (Pitman alternatives)
  – $H_{1,N}$: $\mathbf{\pi} = \mathbf{p} + \mathbf{d} / \text{SQRT}(N)$
Theorem (Mitra, 1958; Read & Cressie, 1988, A8)

Under $H_{1,N}$ and Birch's regularity conditions, $PD^\lambda(\theta_{(\lambda)}; y)$ has an asymptotic noncentral $\chi^2(\gamma, df)$ distribution for any real-valued $\lambda$ and $N \to \infty$ with $df = J-1-S$ and noncentrality parameter (ncp)

$$\gamma = \sum_{j=1}^{J} \frac{d_j^2}{p_j(\theta)} = \sum_{j=1}^{J} \frac{\left(\pi_j - p_j(\theta)\right) \cdot \sqrt{N}}{p_j(\theta)}$$

$$= N \cdot \sum_{j=1}^{J} \frac{\left(\pi_j - p_j(\theta)\right)^2}{p_j(\theta)}$$

$$= PD^\lambda=1(\hat{\theta}_{(\lambda=1)}; e = N \cdot \pi) = X^2(e = N \cdot \pi).$$
Power approximation: Heuristics

Try $\chi^2(\gamma(\lambda), df)$, with

$$df = \sum_{k=1}^{K} (J_k - 1) - S$$
and

$$\gamma(\lambda) = PD^\lambda(\hat{\theta}(\lambda); e),$$

as an approximation to the noncentral distribution of

$PD^\lambda(\hat{\theta}(\lambda); y)$ also

- for finite $N$
- for both simple ($K=1$) and joint parameterized multinomial models ($K > 1$)
- for values of $\lambda$ in $\gamma(\lambda)$ different from $\lambda = 1$. 
Approach 1: Power as a function of sample size and effect size

Any noncentrality parameter $\gamma_{(\lambda)}$ can be written as a product of sample size and effect size:

$$\gamma_{(\lambda)} = N \cdot w_{(\lambda)}^2,$$

for $K = 1$:

$$w_{(\lambda)} = \sqrt{\frac{2}{\lambda(\lambda + 1)} \cdot \sum_{j=1}^{J} \pi_j \left[ \frac{\pi_j}{p_j(\theta)} \right]^{-1}}$$

for $K > 1$:

$$w_{(\lambda)} = \sqrt{\sum_{k=1}^{K} \tau_k \cdot \left( w_{(\lambda(k))} \right)^2}, \text{ with}$$

$$\tau_k = \frac{N_k}{N} \quad \text{and} \quad w_{(\lambda(k))} = \sqrt{\frac{2}{\lambda(\lambda + 1)} \cdot \sum_{j=1}^{J} \pi_{k,j} \left[ \frac{\pi_{k,j}}{p_{k,j}(\theta)} \right]^{-1}}$$
Cohen’s approach

• For $\lambda=1$, Cohens (1969, 1977, 1988) effect size measure $w$ is obtained as a special case:

$$\gamma = N \cdot w^2,$$

for $K = 1$: $w = \sqrt{\sum_{j=1}^{J} \frac{(\pi_j - p_j(\theta))^2}{p_j(\theta)}}$

for $K > 1$: $w = \sqrt{\sum_{k=1}^{K} \tau_k \cdot w_k^2}$, with

$$\tau_k = \frac{N_k}{N} \quad \text{and} \quad w_k = \sqrt{\sum_{j=1}^{J_k} \frac{(\pi_{k,j} - p_{k,j}(\theta))^2}{p_{k,j}(\theta)}}$$
Traditional forms of power analysis (Cohen, 1969, 1977, 1988)

• Effect size conventions
  – \( w = .10 \) (“small effect“)
  – \( w = .30 \) (“medium effect“)
  – \( w = .50 \) (“large effect“)

• Decide on effect size to detect.

• Compute the noncentrality parameter \( \gamma = N w^2 \)

• Types of power analysis
  – “Post hoc”: Compute 1-\( \beta \) as a function of \( w, \alpha, \) and \( N \)
  – “A priori”: Compute \( N \) as a function of \( w, \alpha, \) and 1-\( \beta \)
Problems of the traditional method

• Problem 1:
  – How does effect size translate into parameter values?

• Problem 2:
  – Same meaning of effect size labels in different models?

• Problem 3:
  – Role of the $\tau_k$ in joint PMMs is ignored.
Cohen (1988, p. 244)

On $w$ effect sizes conventions:

“Their use requires particular caution, since, apart from their possible inaptness in a particular substantive context, what is subjectively the same degree of departure or degree of correlation (...) may yield varying $w$, and conversely. The investigator is best advised to use the conventional definitions as a general frame of reference (...) and not to take them too literally.”
Approach 2: Power as a function of the model parameters under $H_1$

1) Specify your $H_0$ model
2) Specify your $H_1$ model with all parameter values fixed at „plausible values“
3) Choose $N_k$, $k = 1, \ldots, K$, and calculate expected frequencies under $H_1$
4) Fit the $H_0$ model to the $H_1$ expected frequencies by minimizing $PD^\lambda$ for some $\lambda$
5) Use the minimum $PD^\lambda$ value as the ncp $\gamma(\lambda)$
6) Compute $1-\beta(\pi) = Pr(\chi^2(\gamma(\lambda), \, df) \geq c_{(df, \, \alpha)}$)
An example: The storage retrieval model
Procedure and results  
(for $df = 1$, $\alpha = .05$)

- $H_0$: $u = a$
- $H_1$: $c = r = .50$,  
  $u = .40$, $a = .60$
- $N_1 = 600$, $N_2 = 300$
- $e_{1,1} = 150$
- $e_{1,2} = 48$
- $e_{1,3} = 144$
- $e_{1,4} = 258$
- $e_{2,1} = 180$
- $e_{2,1} = 120$

<table>
<thead>
<tr>
<th>$\lambda$</th>
<th>$\gamma(\lambda)$</th>
<th>$w(\lambda)$</th>
<th>$1-\beta(\pi)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1</td>
<td>18.66</td>
<td>.14</td>
<td>.99</td>
</tr>
<tr>
<td>-0.5</td>
<td>17.09</td>
<td>.14</td>
<td>.99</td>
</tr>
<tr>
<td>0</td>
<td>16.73</td>
<td>.14</td>
<td>.98</td>
</tr>
<tr>
<td>2/3</td>
<td>16.23</td>
<td>.13</td>
<td>.98</td>
</tr>
<tr>
<td>1</td>
<td>15.97</td>
<td>.13</td>
<td>.98</td>
</tr>
<tr>
<td>2</td>
<td>15.20</td>
<td>.13</td>
<td>.97</td>
</tr>
</tbody>
</table>
Open questions

• How good are these approximations for joint GPT models?
• How does approximation quality depend on the sample sizes?
• Does the approximation quality depend on the PD$^\lambda$ test statistic used?
• Does the approximation quality depend on the $\lambda$-value used to compute the ncp?
A Monte-Carlo study

- Storage-retrieval model ($\tau_1 = 2/3$, $\tau_2 = 1/3$)
- $H_0$: $u = a$, $df = 4-3 = 1$
- $H_1$: $c = r = .5$, $u = .5 - \Delta/2$, $a = .5 + \Delta/2$, for $\Delta = .00 / .10 / .20 / .30 / .40 / .50$
  $\quad (w = .00 / .07 / .13 / .19 / .24 / .28)$
- $N = 30 / 60 / 120 / 240 / 480 / 960$
- Type-1 errors $\alpha = .01 / .05 / .10$
- Statistics: $G^2$, $X^2$, Cressie-Read
- Noncentrality parameter: $\gamma(\lambda=0)$, $\gamma(\lambda=1)$, $\gamma(\lambda=2/3)$
- 1000 Monte Carlo samples per run
Results for $\alpha = .01$

- **Symbols**: Monte-Carlo estimated power for $G^2$ (top), Cressie-Read (middle), and $X^2$ (bottom).
- **Black lines**: Approximate power using same $\lambda$ in ncp $\gamma(\lambda)$ as in $PD^\lambda$ test statistic
- Sample sizes from top to bottom:
  - $N = 960$
  - $N = 480$
  - $N = 240$
  - $N = 120$
  - $N = 60$
  - $N = 30$
Results for $\alpha = .05$

- **Symbols**: Monte-Carlo estimated power for $G^2$ (top), Cressie-Read (middle), and $X^2$ (bottom).
- **Black lines**: Approximate power using same $\lambda$ in ncp $\gamma(\lambda)$ as in PD$^\lambda$ test statistic
- Sample sizes from top to bottom:
  - $N = 960$
  - $N = 480$
  - $N = 240$
  - $N = 120$
  - $N = 60$
  - $N = 30$
Absolute differences between Monte-Carlo power and approximate power formula ($\alpha = .05$)

<table>
<thead>
<tr>
<th>$N$</th>
<th>$G^2$ with $\gamma_{(0)}$</th>
<th>$C-R$ with $\gamma_{(2/3)}$</th>
<th>$X^2$ with $\gamma_{(1)}$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>mean</td>
<td>max.</td>
<td>mean</td>
</tr>
<tr>
<td>30</td>
<td>.065</td>
<td>.157</td>
<td>.075</td>
</tr>
<tr>
<td>60</td>
<td>.007</td>
<td>.019</td>
<td>.010</td>
</tr>
<tr>
<td>120</td>
<td>.021</td>
<td>.040</td>
<td>.027</td>
</tr>
<tr>
<td>240</td>
<td>.017</td>
<td>.031</td>
<td>.020</td>
</tr>
<tr>
<td>480</td>
<td>.004</td>
<td>.011</td>
<td>.005</td>
</tr>
<tr>
<td>960</td>
<td>.003</td>
<td>.014</td>
<td>.003</td>
</tr>
<tr>
<td>Mean</td>
<td>.020</td>
<td>.045</td>
<td>.023</td>
</tr>
</tbody>
</table>
Conclusions from the Monte-Carlo study

- In our example, the power approximation is acceptable for \( N > 50 \) and good for \( N > 200 \)
- Use the same \( \lambda \) parameter in the PD\( ^{\lambda} \) statistic and the noncentrality parameter \( \gamma(\lambda) \)
- Approximation accuracy appears to be worse for \( \alpha = .01 \) compared to \( \alpha = .05 \) and \( \alpha = .10 \)
- In general, the effect of \( \lambda \) on approximation accuracy appears to be small.
- However, approximation accuracy is slightly larger for \( G^2 \) compared to both the Cressie-Read statistic and Pearson’s \( X^2 \).
Power optimization

Problem:
Given a fixed total sample size and a fixed $\alpha$, is there any way to maximize the power?

Answer:
Yes, there are many!
1) \( \lambda \) Optimization

- Asymptotic results (see Read & Cressie, 1988):
  - Pitman efficiency (or Asymptotic Relative Efficiency, A.R.E.) of two different PD\( ^\lambda \) statistics is always 1.
  - Behadur efficiency is optimal for \( G^2 \)

- Approximation for finite \( N \):
  - One positive outlier --> positive \( \lambda \) is optimal
  - Several deviations of same size --> \( \lambda = 0 \) is optimal
  - One negative outlier --> negative \( \lambda \) is optimal

- Conclusions:
  - Effect of \( \lambda \) on power is small for \(-2 < \lambda < 2\).
  - \( G^2 \) appears to be a good default option
2) $\tau_k$ Optimization

- Power of $G^2$ as a function of $\tau_1$ for the storage retrieval model, given $\alpha = .10$ (top), .05 (middle) and .01 (bottom).
- $c = r = .50$, $u = .40$, $a = .60$
- $N = 480$
- Conclusions:
  - Strong effect of $\tau_1$ !!
  - Max. power for $\tau_1 = .652$
  - Thus, $\tau_1 = ... = \tau_k$ may be a bad default option!
3) $\theta$ Optimization

- Model parameters can be divided in $H_0$-relevant and $H_0$-irrelevant parameters. For the storage-retrieval model test:
  - $u$ and $a$ are $H_0$ relevant
  - $c$ and $r$ are $H_0$ irrelevant

- Problem:
  How to choose the values of the $H_0$-irrelevant parameters so as to maximize the power of the model test?
Approximate power of the $G^2$ test for the storage-retrieval model ($\alpha = .05$, $N_1 = 320$, $N_2 = 160$)

<table>
<thead>
<tr>
<th>Parameter values under $H_1$</th>
<th>ncp</th>
<th>approx</th>
</tr>
</thead>
<tbody>
<tr>
<td>$c$ $r$ $u$ $a$ $\gamma(0)$ $1-\beta(\pi)$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>.10</td>
<td>.80</td>
<td>.40</td>
</tr>
<tr>
<td>.10</td>
<td>.20</td>
<td>.40</td>
</tr>
<tr>
<td>.50</td>
<td>.80</td>
<td>.40</td>
</tr>
<tr>
<td>.50</td>
<td>.20</td>
<td>.40</td>
</tr>
<tr>
<td>.90</td>
<td>.80</td>
<td>.40</td>
</tr>
<tr>
<td>.90</td>
<td>.20</td>
<td>.40</td>
</tr>
</tbody>
</table>
4) Conditional versus unconditional tests

- Consider two nested models:
  - $M_0$ with parameter space $\Omega_0$
  - $M_1$ with parameter space $\Omega_1$
  - $\Omega_1 \subset \Omega_0$

- Problem:
  $M_1$ can be tested by an unconditional or a conditional $G^2$ test provided that $M_0$ holds.
  Which test is more powerful?
Example: Storage-retrieval model

- $N_1 = 160, N_2 = 80$
- $M_0: u = a$
- $M_1: u = a$ and $c = .30$
- Under $H_1: c = r = u = a = .50$ we obtain ($\alpha = .05$):
  - $G^2(M_1)$: $\text{df} = 4-2 = 2$, $\gamma_{(0)} = 8.62$, $1-\beta(\pi) = .75$
  - $G^2(M_1)-G^2(M_0)$: $\text{df} = 2-1 = 1$, $\gamma_{(0)} = 8.62$, $1-\beta(\pi) = .85$
- Therefore, use conditional $G^2$ difference tests whenever possible.
Summary and conclusions

• Do not ignore the power of model tests!
• The proposed approximation method works very well for joint PMMs with typical sample sizes
• Choose same $\lambda$ in the $\text{PD}^\lambda$ statistic and noncentrality parameter $\gamma(\lambda)$
• $G^2$ is a good default option for several reasons:
  – Approximation accuracy is optimal
  – Maximum power for “diffuse” noncentrality structures
  – Option of conditional tests
• Do not forget to optimize context conditions:
  – $\tau_k$ optimization
  – $\theta$ optimization
  – conditional tests whenever possible.