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 β.
- 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 β.
- 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

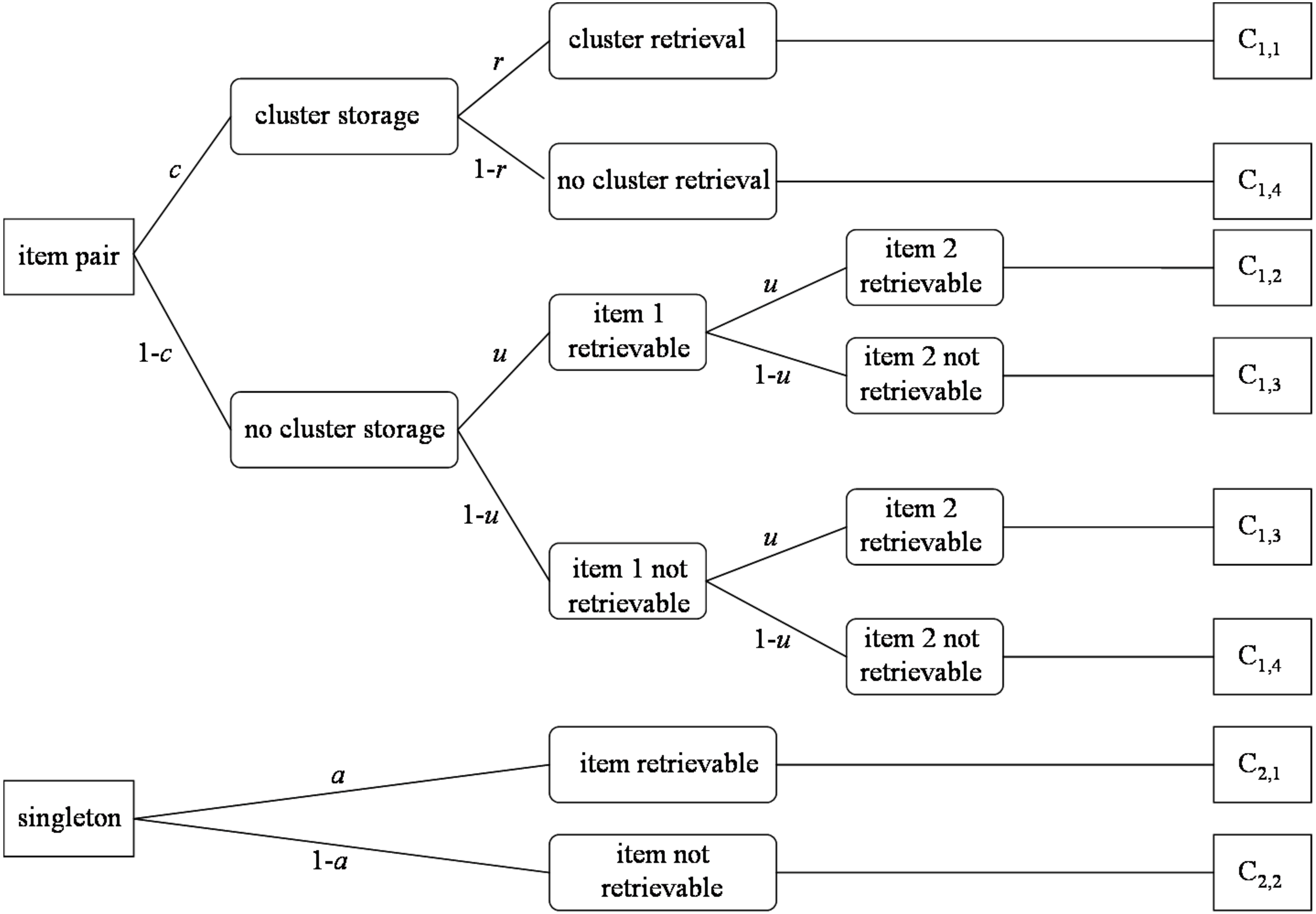
An example:
The storage-retrieval model

- 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:
  - $y_{1,1} = freq(\mathrm{C}_{1,1})$: Number of word pairs recalled adjacently
  - $y_{1,2} = freq(\mathrm{C}_{1,2})$: Number of word pairs recalled not adjacently
  - $y_{1,3} = freq(\mathrm{C}_{1,3})$: Number of pairs with only one word recalled
  - $y_{1,4} = freq(\mathrm{C}_{1,4})$: Number of pairs with none of the words recalled
  - $y_{1,1} + y_{1,2} + y_{1,3} + y_{1,4} = N_1$

- Singletons:
  - $y_{2,1} = freq(\mathrm{C}_{2,1})$: Number of singletons recalled
  - $y_{2,2} = freq(\mathrm{C}_{2,1})$: Number of singletons not recalled
  - $y_{2,1} + y_{2,2} = N_2$


item pair
c
1-c
cluster storage
r
1-r
cluster retrieval
no cluster retrieval
$C_{1,1}$
$C_{1,4}$
no cluster storage
u
1-u
item 1 retrievable
item 1 not retrievable
u
1-u
item 2 retrievable
item 2 not retrievable
$C_{1,2}$
$C_{1,3}$
u
1-u
item 2 retrievable
item 2 not retrievable
$C_{1,3}$
$C_{1,4}$
singleton
a
1-a
item retrievable
item not retrievable
$C_{2,1}$
$C_{2,2}$

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, $k = 1, ..., K$, a multinomial model holds with N_k observations and category probabilities $\pi_{k,j} = Pr(C_{k,j})$, $j = 1, ..., J_\mathrm{k}$, $\pi_{k,1} + \pi_{k,2} + ... + \pi_{k,Jk} = 1$.
  - Observations are independent
  - $N = N_1 + ... + N_K$
  - $\boldsymbol{\pi} := (\pi_{1,1}, ..., \pi_{K,JK})$
  - $\mathbf{y} := (y_{1,1}, ..., y_{K,JK})$
- Parameterized multinomial model:
  - The probabilities $\pi_{k,j}$ are functions of S real-valued latent parameters θ_s, $s = 1, ..., S$, i.e.
  - $\pi_{\mathrm{k,j}} = p_{\mathrm{k,j}}(\theta_1, ..., \theta_S)$
  - $\boldsymbol{\theta} := (\theta_1, ..., \theta_S) \in \Omega$

Goodness-of-fit tests for PMMs

Hypothesis: H_0: $\boldsymbol{\pi} \in f(\Omega)$

Tool: Power divergence statistic

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

with

$$\mathrm{PD}^{\boldsymbol{\lambda}=0}(\boldsymbol{\theta};\mathbf{y}) \;:= \lim_{\boldsymbol{\lambda}\to 0} \mathrm{PD}^{\boldsymbol{\lambda}}(\boldsymbol{\theta};\mathbf{y})$$

$$\mathrm{PD}^{\boldsymbol{\lambda}=-1}(\boldsymbol{\theta};\mathbf{y}) \;:= \lim_{\boldsymbol{\lambda}\to -1} \mathrm{PD}^{\boldsymbol{\lambda}}(\boldsymbol{\theta};\mathbf{y}).$$

Asymptotic central distribution (Read & Cressie, 1988, A6)

For any real-valued λ, the minimum PD^{λ} statistic

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

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

λ	$\hat{\theta}_{(\lambda)}$	$PD^{\lambda}(\hat{\theta}_{(\lambda)};\mathbf{y})$
0	Maximum-Likelihood estimator	Log-Likelihood-χ^2 statistic (G^2)
2/3	Minimum $PD^{\lambda=2/3}$ estimator	Cressie-Read statistic
1	Minimum chi-square estimator	Pearson´s χ2 statistic (X^2)

Asymptotic noncentral distribution (Mitra, 1958)

- Notation
  - H_0 probabilities: $\mathbf{p} = (p_1(\theta), ..., p_J(\theta))$
  - H_1 probabilities: $\boldsymbol{\pi} = (\pi_1, ..., \pi_J)$
  - $d_j = \pi_j - p_j$, $d_1 + ... + d_J = 1$
- Consider the class of local alternative hypotheses (Pitman alternatives)
  - $H_{1,N}$: $\boldsymbol{\pi} = \mathbf{p} + \mathbf{d}$ / SQRT(N)

Theorem
(Mitra, 1958; Read & Cressie, 1988, A8)

Under $H_{1,N}$ and Birch´s regularity conditions, $PD^{\lambda}(\hat{\theta}_{(\lambda)};\mathbf{y})$ has an asymptotic noncentral $\chi^2(\gamma, df)$ distribution for any real-valued λ and $N \rightarrow \infty$ with $df = J\text{-}1\text{-}S$ and noncentrality parameter (ncp)

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

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

$$= PD^{\lambda=1}(\hat{\theta}_{(\lambda=1)};\, \mathbf{e} = N \cdot \boldsymbol{\pi}) \;=\; X^2(\mathbf{e} = N \cdot \boldsymbol{\pi}).$$

Power approximation: Heuristics

Try $\chi^2(\gamma_{(\lambda)}, df)$, with

$$df = \sum_{k=1}^{K}(J_k - 1) - S \text{ and } \boldsymbol{\gamma}_{(\lambda)} = \mathrm{PD}^{\lambda}(\hat{\boldsymbol{\theta}}_{(\lambda)}; \mathbf{e}) \,,$$

as an approximation to the noncentral distribution of

$\mathrm{PD}^{\lambda}(\hat{\boldsymbol{\theta}}_{(\lambda)};\mathbf{y})$ also

- for finite N
- for both simple (K=1) and joint parameterized multinomial models ($K > 1$)
- for values of λ 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:

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

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

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

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

Cohen´s approach

- For λ=1, Cohens (1969, 1977, 1988) effect size measure *w* is obtained as a special case:

$$\boldsymbol{\gamma} = N \cdot w^2,$$

$$\text{for } K = 1: \; w = \sqrt{\sum_{j=1}^{J} \frac{\left(\boldsymbol{\pi}_j - p_j(\boldsymbol{\theta})\right)^2}{p_j(\boldsymbol{\theta})}}$$

$$\text{for } K > 1: \; w = \sqrt{\sum_{k=1}^{K} \boldsymbol{\tau}_k \cdot w_k^{\,2}} \text{, with}$$

$$\boldsymbol{\tau}_k = \frac{N_k}{N} \text{ and } w_k = \sqrt{\sum_{j=1}^{J_k} \frac{\left(\boldsymbol{\pi}_{k,j} - p_{k,j}(\boldsymbol{\theta})\right)^2}{p_{k,j}(\boldsymbol{\theta})}}$$

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-β as a function of w, α, and N
  - “A priori”: Compute N as a function of w, α, and 1-β

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 τ_k in joint PMMs is ignored.

Cohen (1988, p. 244)

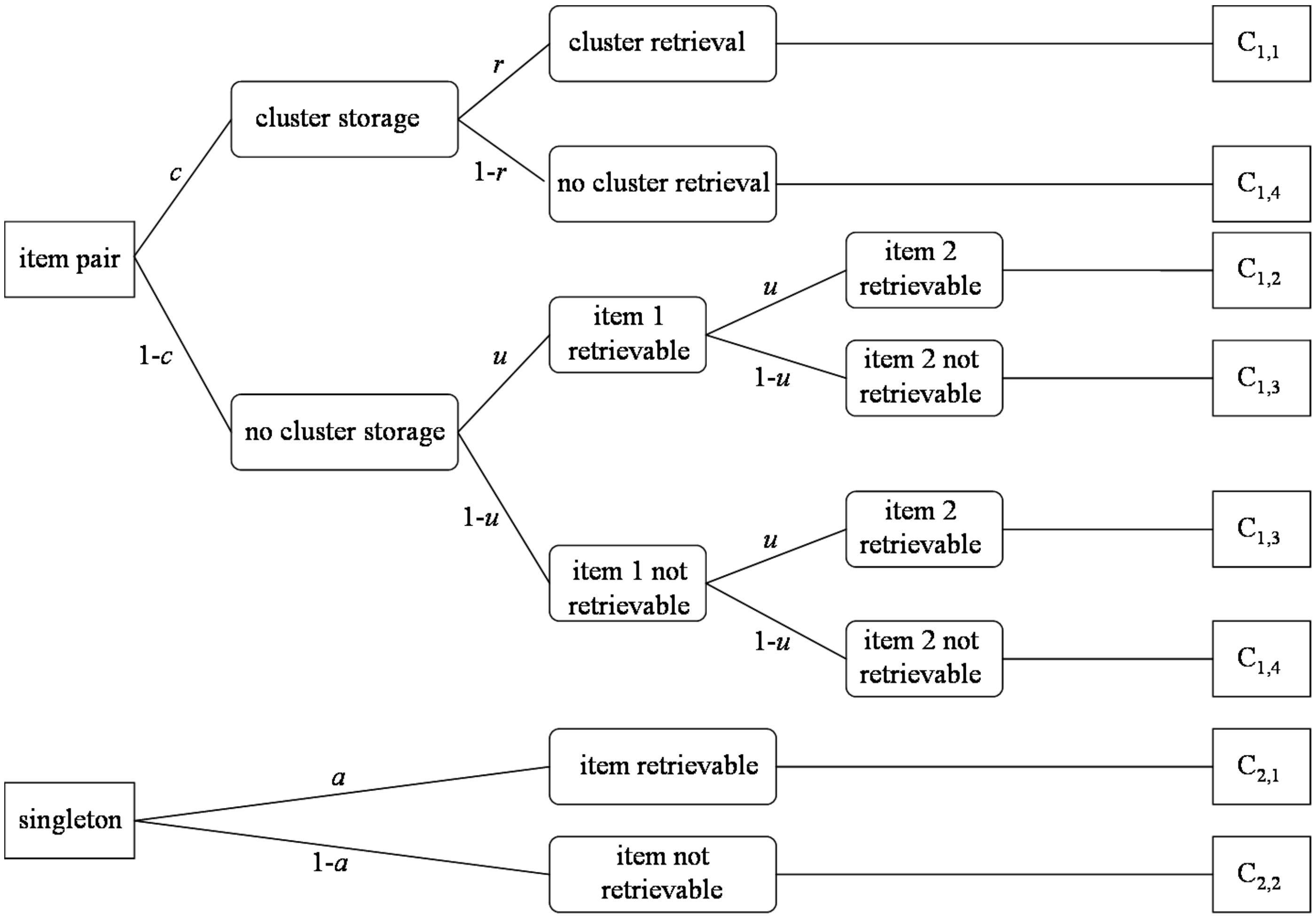
On *w* effect sizes conventions:

"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, $k = 1, ..., K$, and calculate expected frequencies under H_1
4) Fit the H_0 model to the H_1 expected frequencies by minimizing PD^λ for some λ
5) Use the minimum PD^λ value as the ncp $\gamma_{(\lambda)}$
6) Compute $1\text{-}\beta(\boldsymbol{\pi}) = Pr(\chi^2(\gamma_{(\lambda)}, df) \geq c_{(df,\, \alpha)})$

An example: The storage retrieval model



Procedure and results (for $df = 1$, $\alpha = .05$)

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

λ	$\gamma_{(\lambda)}$	$w_{(\lambda)}$	$1-\beta(\pi)$
-1	18.66	.14	.99
-0.5	17.09	.14	.99
0	16.73	.14	.98
2/3	16.23	.13	.98
1	15.97	.13	.98
2	15.20	.13	.97

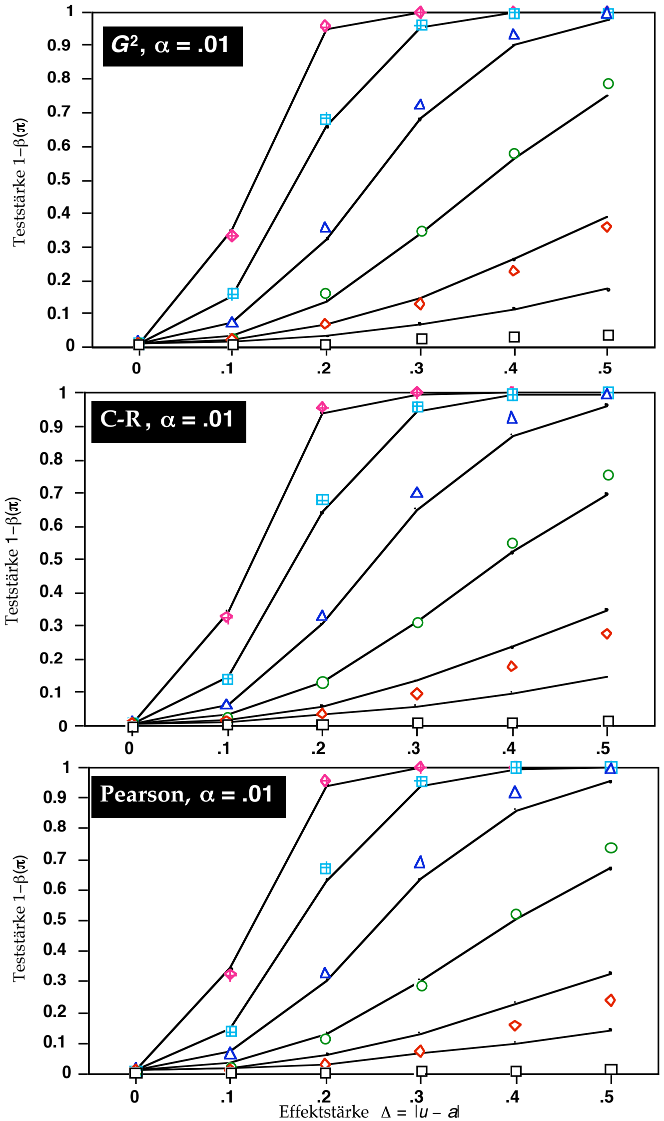
Open questions

- 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^{λ} test statistic used?
- Does the approximation quality depend on the λ-value used to compute the ncp?

A Monte-Carlo study

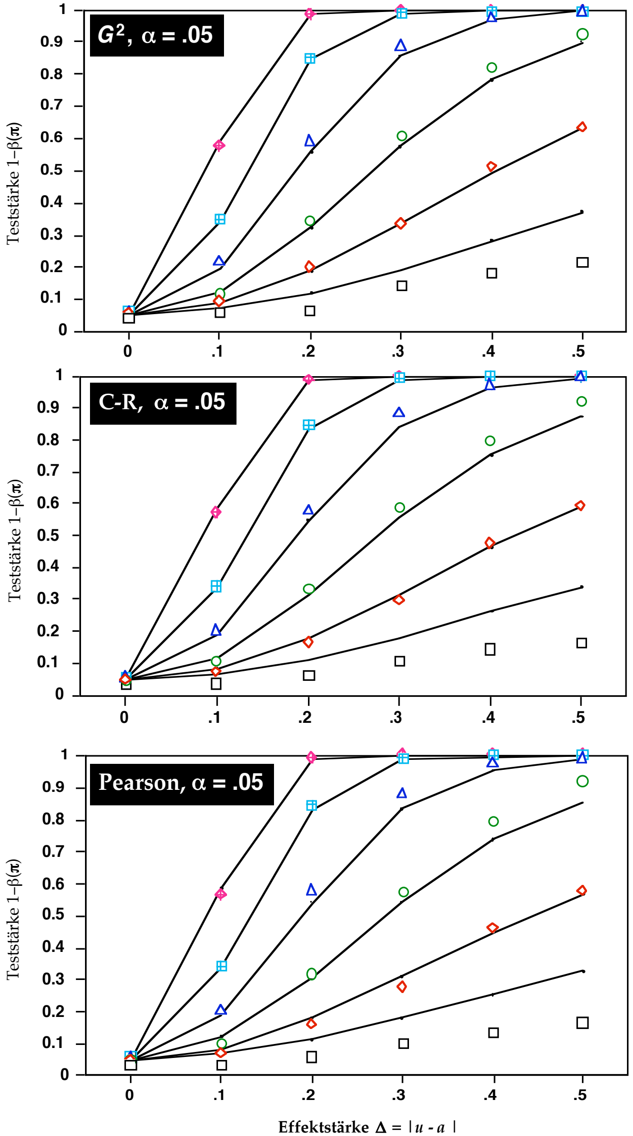
- Storage-retrieval model (τ_1= 2/3, τ_2= 1/3)
- H_0: $u = a$, $df = 4\text{-}3 = 1$
- H_1: $c = r = .5$, $u = .5 - \Delta/2$, $a = .5 + \Delta/2$, for
  $\Delta = .00 / .10 / .20 / .30 / .40 / .50$
  ($w = .00 / .07 / . 13 / .19 / .24 / .28$)
- N = 30 / 60 / 120 / 240 / 480 / 960
- Type-1 errors α = .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 λ in ncp $\gamma_{(\lambda)}$ as in PD^{λ} 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 λ in ncp $\gamma_{(\lambda)}$ as in PD^{λ} 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 ($\alpha = .05$)

	Test statistic and ncp used					
	G^2 with $\gamma_{(0)}$		C-R with $\gamma_{(2/3)}$		X^2 with $\gamma_{(1)}$	
N	mean	max.	mean	max.	mean	max.
30	.065	.157	.075	.173	.078	.162
60	.007	.019	.010	.020	.015	.032
120	.021	.040	.027	.049	.031	.066
240	.017	.031	.020	.041	.021	.048
480	.004	.011	.005	.010	.006	.014
960	.003	.014	.003	.014	.004	.016
Mean	**.020**	**.045**	**.023**	**.051**	**.026**	**.056**

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 λ parameter in the PD^{λ} 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 λ 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 α, is there any way to maximize the power?

Answer:

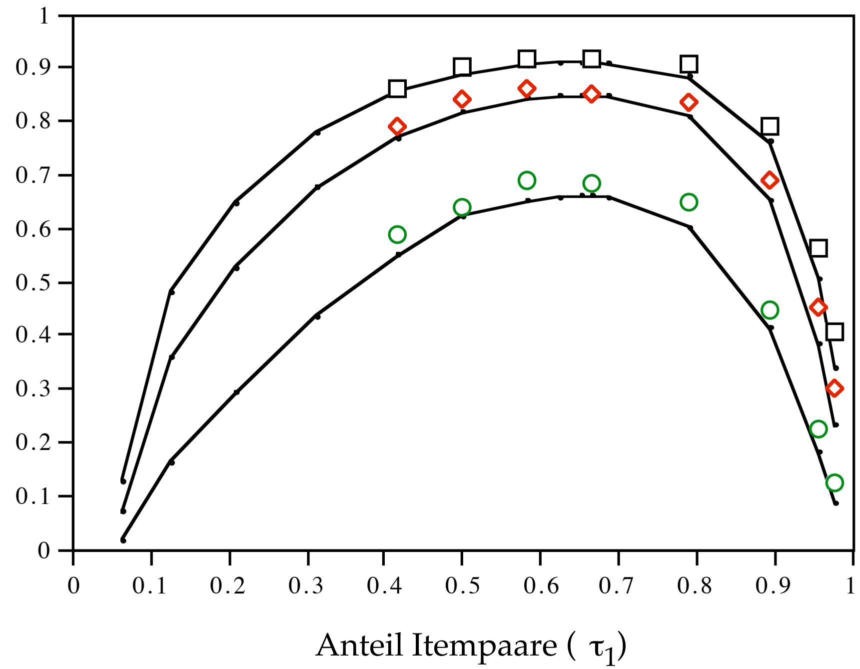
Yes, there are many!

1) λ Optimization

- Asymptotic results (see Read & Cressie, 1988):
  - Pitman efficiency (or Asymptotic Relative Efficiency, A.R.E.) of two different PD^{λ} statistics is always 1.
  - Behadur efficiency is optimal for G^2
- Approximation for finite N:
  - One positive outlier --> positive λ is optimal
  - Several deviations of same size --> $\lambda = 0$ is optimal
  - One negative outlier --> negative λ is optimal
- Conclusions:
  - Effect of λ on power is small for $-2 < \lambda < 2$.
  - G^2 appears to be a good default option

2) τ_k Optimization

- Power of G^2 as a function of τ_1 for the storage retrieval model, given α = .10 (top), .05 (middle) and .01 (bottom).
- $c = r = .50$, $u = .40$, $a = .60$
- $N = 480$
- Conclusions:
  - Strong effect of τ_1 !!
  - Max. power for $\tau_1 = .652$
  - Thus, $\tau_1 = \ldots = \tau_K$ may be a bad default option!



3) θ 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 ($\alpha = .05$, $N_1 = 320$, $N_2 = 160$)

Parameter values under H_1				ncp	approx
c	r	u	a	$\gamma_{(0)}$	$1\text{-}\beta(\pi)$
.10	.80	.40	.60	12.49	.94
.10	.20	.40	.60	12.49	.94
.50	.80	.40	.60	8.92	.85
.50	.20	.40	.60	8.92	.85
.90	.80	.40	.60	2.53	.36
.90	.20	.40	.60	2.53	.36

4) Conditional versus unconditional tests

- Consider two nested models:
  - M_0 with parameter space Ω_0
  - M_1 with parameter space Ω_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 ($\alpha = .05$):
  - $G^2(M_1)$: df = 4-2 = 2, $\gamma_{(0)} = 8.62$, $1\text{-}\beta(\pi) = .75$
  - $G^2(M_1)\text{-}G^2(M_0)$: df = 2-1 = 1, $\gamma_{(0)} = 8.62$, $1\text{-}\beta(\pi) = .85$
- 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 λ in the PD^{λ} 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:
  - τ_k optimization
  - θ optimization
  - conditional tests whenever possible.