Threshold models with fixed and random effects for ordered categorical data

Hans-Peter Piepho

Universität Hohenheim, Germany

Edith Kalka

Universität Kassel, Germany
1. Introduction
2. Case studies
3. Numerical issues
4. Simulations
5. Conclusion
1. Introduction

Analysis of ordered categorical data:

- ANOVA, LSD
- Transformation to normality
- Nonparametric methods
- Threshold model
Fig. 1.1: Observed frequencies for snack food scores (Best et al., 2000).
Consumers

Latent variable

\[ m \]

\[ θ_1, θ_2, θ_3, θ_4, θ_5, θ_6, θ_7, θ_8, θ_9 \]

\[ π_1, π_2, π_3, π_4, π_5, π_6, π_7, π_8, π_9 \]
Latent variables

Latent variable $\rightarrow$ observed score

$m$ (quantitative) $\quad y$ (ordered categorical)

"… increasing the energy of the physical stimulus, or the concentration or amount of food ingredient, should result in an increase in how strong something feels, looks, smells, or tastes. We increase the salt in a soup, and it tastes saltier."

(Lawless and Heymann, 1998, p. 208)
Multinomial probabilities

- Threshold values $\theta_1$, $\theta_2$, ...
- Multinomial probabilities $\pi_1$, $\pi_2$, ...

\[\pi_1 = \Phi(\theta_1 - \eta)\]
\[\pi_2 = \Phi(\theta_2 - \eta) - \Phi(\theta_1 - \eta)\]
\[\pi_3 = \Phi(\theta_3 - \eta) - \Phi(\theta_2 - \eta)\]
\[\vdots\]
\[\pi_{I-1} = \Phi(\theta_I - \eta) - \Phi(\theta_{I-1} - \eta)\]
\[\pi_I = 1 - \Phi(\theta_I - \eta)\]

$\Phi$ = standard normal c.d.f.
$\eta$ = mean on the latent scale
$I$ = number of categories
Linear modelling

Expected value on latent scale:

\[ \eta_k = \mu + \alpha_k \]

where

\[ \eta_k = \text{mean for the } k\text{-th panel} \]

\( (k = 1: \text{consumers} \ k = 2: \text{employees}) \)

\[ \alpha_k = \text{effect of } k\text{-th panel} \]

Random variable on latent scale:

\[ m_{jk} = \eta_k + e_{jk} \]

\[ e_{jk} \sim N(0,1) \]
Fig. 1.3: Fitted frequencies of scores for snack food scores.
Fig. 1.4: Fitted frequencies based on threshold model for snack food scores, assuming means of $\eta = 0$ and $\eta = -3$ for the latent variable.
Fig. 1.5: Histogram of observed overall acceptance data (Villanueva et al., 2000). Scores display significant variance heterogeneity by an LR-test, assuming normality (p = < 0.0001).
Fig. 1.6: Fitted histogram of overall acceptance data (Villanueva et al., 2000), assuming threshold model with homogeneous variance ($-2 \log L = 1104.0$).
Fig. 1.7: Fitted histogram of overall acceptance data (Villanueva et al., 2000), assuming threshold model with heterogeneous variance ($-2 \log L = 1097.0$). $\sigma_k = \text{standard deviation on latent scale for } k\text{-th product.}$
Unequal variance on latent scale

⇒ invalidates ANOVA of observed scores!

Example:

- Data generated by threshold model with three categories ($\theta_1 = -1$ and $\theta_2 = 1$)

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Group 1</th>
<th>Group 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Latent scale:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>0.9</td>
<td>1.0</td>
</tr>
<tr>
<td>Variance</td>
<td>1.0</td>
<td>4.0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Group 1</th>
<th>Group 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed scale:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>2.4</td>
<td>2.2</td>
</tr>
</tbody>
</table>
2. Case studies

Case study 1:

*Consumer preferences for on farm and industrially processed quarg samples*

- nine point hedonic scale
  
  1 = dislike extremely to
  
  9 = like extremely

- 152 consumers

- Seven quarg products
An initial model

\[ m_{ij} = \alpha_i + u_j + e_{ij} \]  \hspace{1cm} (2.1)

where

\[ m_{ij} = \text{latent variable for } i\text{-th product and } j\text{-th assessor} \]

\[ \alpha_i = \text{expected value of } i\text{-th product} \]

\[ u_j = \text{random effect of } j\text{-th assessor;} \]

\[ u_j \sim N(0, \sigma_u^2) \]

\[ e_{ij} = \text{random deviation of } j\text{-th assessor for } i\text{-th product;} \]

\[ e_{ij} \sim N(0, 1) \]

\[ \text{corr}(m_{ij}, m_{i',j'}) = \frac{\sigma_u^2}{\sigma_u^2 + 1} \]
Table 2.1: The Johnson System for $u_j$.  

<table>
<thead>
<tr>
<th>Distribution</th>
<th>Model for $u_j$</th>
<th>Support</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_L$</td>
<td>$u_j = \exp(w_j)$</td>
<td>$0 &lt; u_j \leq \infty$</td>
</tr>
<tr>
<td>$S_B$</td>
<td>$u_j = \phi \frac{\exp(w_j)}{1 + \exp(w_j)}$</td>
<td>$0 \leq u_j \leq \phi$</td>
</tr>
<tr>
<td>$S_U$</td>
<td>$u_j = \phi \sinh(w_j)$</td>
<td>$-\infty &lt; u_j \leq \infty$</td>
</tr>
</tbody>
</table>

$\S$: $w_j \sim N(\mu_w, \sigma_w^2)$
Factor-analytic models

\[ m_{ij} = \alpha_i + \lambda_i u_j + e_{ij} \]

\( \lambda_i \) = factor loading for \( i \)-th product
\( u_j \) = latent value for \( j \)-th assessor;
\( u_j \sim N(0,1) \)
\( e_{ij} \) = error; \( e_{ij} \sim N(0,1) \)

Model implies heterogeneity:

\[
\text{var}(m_{ij}) = \lambda_i^2 + 1
\]

\[
\text{corr}(m_{ij}, m_{i'j'}) = \frac{\lambda_i \lambda_{i'}}{\sqrt{(\lambda_i^2 + 1)(\lambda_{i'}^2 + 1)}}
\]
Model selection by AIC

Akaike Information Criterion = AIC:

\[ AIC = -2 \log L + 2p \]

where

\[ \log L = \text{maximized log-likelihood} \]

\[ p = \text{number of parameters} \]

"Smaller is better"
### Table 2.2: Model fits for quarg data.

<table>
<thead>
<tr>
<th>Model for latent variable</th>
<th>Distribution for random effects</th>
<th>AIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>$m_{ij} = \alpha_i + e_{ij}$</td>
<td>$u_j \sim N(0, \sigma_u^2)$</td>
<td>4464.7</td>
</tr>
<tr>
<td>$m_{ij} = \alpha_i + u_j + e_{ij}$</td>
<td>$u_j \sim S_L(\phi, \mu_w, \sigma_w^2)$</td>
<td>4442.8</td>
</tr>
<tr>
<td></td>
<td>$u_j \sim S_B(\phi, \mu_w, \sigma_w^2)$</td>
<td>4443.3</td>
</tr>
<tr>
<td></td>
<td>$u_j \sim S_U(\phi, \mu_w, \sigma_w^2)$</td>
<td>4445.0</td>
</tr>
<tr>
<td>$m_{ij} = \alpha_i + \lambda_i u_{j1} + \lambda_{i2} u_{j2} + e_{ij}$</td>
<td>$u_{jk} \sim N(0,1)$</td>
<td>4439.4</td>
</tr>
<tr>
<td>$m_{ij} = \alpha_i + \lambda_i u_{j1} + \lambda_{i2} u_{j2} + \lambda_{i3} u_{j3} + e_{ij}$</td>
<td>$u_{jk} \sim N(0,1)$</td>
<td>4407.3</td>
</tr>
<tr>
<td></td>
<td>$u_{jk} \sim N(0,1)$</td>
<td>4403.6</td>
</tr>
</tbody>
</table>
Factor-analytic threshold model (two multiplicative terms):

<table>
<thead>
<tr>
<th>Product</th>
<th>Estimate of $\alpha_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2.10 bc</td>
</tr>
<tr>
<td>2</td>
<td>2.20 c</td>
</tr>
<tr>
<td>3</td>
<td>1.57 d</td>
</tr>
<tr>
<td>4</td>
<td>1.96 be</td>
</tr>
<tr>
<td>5</td>
<td>1.70 de</td>
</tr>
<tr>
<td>6</td>
<td>2.11 bc</td>
</tr>
<tr>
<td>7</td>
<td>2.88 a</td>
</tr>
</tbody>
</table>

*SED* 0.132

Estimates in a column followed by the same letter are not significantly different.

ANOVA of scores:

<table>
<thead>
<tr>
<th>Product</th>
<th>Score mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5.43 b</td>
</tr>
<tr>
<td>2</td>
<td>5.60 b</td>
</tr>
<tr>
<td>3</td>
<td>4.45 c</td>
</tr>
<tr>
<td>4</td>
<td>5.17 b</td>
</tr>
<tr>
<td>5</td>
<td>4.63 c</td>
</tr>
<tr>
<td>6</td>
<td>5.41 b</td>
</tr>
<tr>
<td>7</td>
<td>6.67 a</td>
</tr>
</tbody>
</table>

LSD 0.445
Artificial example

- Four treatments (T1, T2, T3, T4)
- Comparison of T1 and T4 significant

Initial display:

\[
\begin{align*}
\text{T1} & \ a \\
\text{T2} & \ a \\
\text{T3} & \ a \ \Rightarrow \text{misses significance} \\
\text{T4} & \ a
\end{align*}
\]

of T1-T4 comparison

Insertion:

\[
\begin{align*}
\text{T1} & \ a \ b \\
\text{T2} & \ a \ b \\
\text{T3} & \ a \ b \\
\text{T4} & \ a \ b
\end{align*}
\]
Fig. 2.1: Plot of observed score means (ANOVA) versus estimated effects under threshold model for seven quarg products.
**Fig. 2.4**: Plot of score mean versus linear predictor under estimated threshold model for quarg data.
Case study 2:

Sensory descriptive analysis of on farm and industrial processed quarg samples by a trained panel

• trained panel of 15 assessors
• 8 quarg samples (B1-B8)
• 12 sensory attributes
• 12 sessions
• ordinal scale
  (0 = non-detectable to 5 = very strong)
Mixed linear model (per attribute):

$$m_{ijk} = u_j + \tau_i + \alpha_k + (\alpha \tau)_{ik} + e_{ijk}$$  \hspace{1cm} (2.6)

where

$u_j$ = effect of $j$-th assessor

$\tau_i$ = effect of $i$-th product

$\alpha_k$ = effect of $k$-th session (random)

$(\alpha \tau)_{ik}$ = interaction of $i$-th product and $k$-th session (random)
Principal component analysis

Two standardizations:

\[ \sigma = \sqrt{\frac{\sum_{i=1}^{I} (\tau_i - \bar{\tau})}{(I-1)}} \]  

(2.7a)

\[ \sigma = \sqrt{\sigma_{\alpha \tau}^2 + 1} \]  

(2.7b)

⇒ singular value decomposition of standardized values
**Fig. 2.5**: Biplot for PCA of 8 quarg samples and 12 sensory attributes based on standardization (2.7a). 83% of variance are explained. B1-B8: samples; letters: attributes)
Fig. 2.6: Biplot for PCA of 8 quarg samples and 12 sensory attributes based on standardization (2.7b). 91% of variance are explained. B1-B8: samples; letters: attributes).
Heterogeneity of variance

- We hypothesized that variance was larger for samples B1 through B4.

- Fit model with heterogeneity in variance of interaction \((\alpha \tau)_{ik}\).

\[ \sigma^2_{\alpha \tau(1)} : \text{variance for products B1 to B4} \]
\[ \sigma^2_{\alpha \tau(2)} : \text{variance for products B5 to B8} \]
<table>
<thead>
<tr>
<th>Attributes</th>
<th>Homogeneous</th>
<th>Heterogeneous</th>
<th>LR-test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\sigma_{\alpha_1}^2$</td>
<td>$\sigma_{\alpha_1}^2(1)$</td>
<td>$\sigma_{\alpha_1}^2(2)$</td>
</tr>
<tr>
<td>Granular</td>
<td>0.96</td>
<td>1.01</td>
<td>0.82</td>
</tr>
<tr>
<td>Separation of whey</td>
<td>0.71</td>
<td>1.24</td>
<td>0.29</td>
</tr>
<tr>
<td>Creamy yellow color</td>
<td>0.43</td>
<td>0.90</td>
<td>0.09</td>
</tr>
<tr>
<td>Bitter</td>
<td>0.04</td>
<td>0.01</td>
<td>0.07</td>
</tr>
<tr>
<td>Sour</td>
<td>0.01</td>
<td>0.00</td>
<td>0.07</td>
</tr>
<tr>
<td>Flavour of starter cultures</td>
<td>0.12</td>
<td>0.00</td>
<td>0.29</td>
</tr>
<tr>
<td>Flavour of yoghurt</td>
<td>0.15</td>
<td>0.00</td>
<td>0.37</td>
</tr>
<tr>
<td>Rancid</td>
<td>0.43</td>
<td>0.06</td>
<td>1.78</td>
</tr>
<tr>
<td>Other flavour</td>
<td>0.20</td>
<td>0.17</td>
<td>0.25</td>
</tr>
<tr>
<td>Firm</td>
<td>1.09</td>
<td>2.37</td>
<td>0.72</td>
</tr>
<tr>
<td>Gritty</td>
<td>1.40</td>
<td>2.03</td>
<td>0.21</td>
</tr>
<tr>
<td>Creamy</td>
<td>0.51</td>
<td>0.32</td>
<td>0.77</td>
</tr>
</tbody>
</table>
Inter-assessor agreement

- Assessed separately for each session

\[ m_{ij} = u_j + \tau_i + e_{ij} \]  \hspace{1cm} (2.8)

\( u_j \) = effect of \( j \)-th assessor

\( \tau_i \) = effect of \( i \)-th product

\( \tau_i \sim N(0, \sigma^2_\tau) \)

\( e_{ij} \) = random error; standard normal

Intra-class correlation:

\[ \rho_{IC} = \frac{\sigma^2_\tau}{1+\sigma^2_\tau} \]
Table 2.9: Estimates of the intra-class correlation $\rho_{IC}$ and Kendall's coefficient of concordance for quarg data.

<table>
<thead>
<tr>
<th>Session</th>
<th>granular</th>
<th>colour</th>
<th>sour</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kendall's measure of concordance:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.87</td>
<td>0.71</td>
<td>0.67</td>
</tr>
<tr>
<td>3</td>
<td>0.64</td>
<td>0.64</td>
<td>0.80</td>
</tr>
<tr>
<td>7</td>
<td>0.68</td>
<td>0.80</td>
<td>0.77</td>
</tr>
<tr>
<td>9</td>
<td>0.79</td>
<td>0.56</td>
<td>0.85</td>
</tr>
<tr>
<td>11</td>
<td>0.83</td>
<td>0.78</td>
<td>0.74</td>
</tr>
<tr>
<td>12</td>
<td>0.64</td>
<td>0.32</td>
<td>0.78</td>
</tr>
</tbody>
</table>

Intraclass correlation ($\rho_{IC}$):

<table>
<thead>
<tr>
<th>Session</th>
<th>granular</th>
<th>colour</th>
<th>sour</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.86</td>
<td>0.78</td>
<td>0.60</td>
</tr>
<tr>
<td>3</td>
<td>0.64</td>
<td>0.81</td>
<td>0.82</td>
</tr>
<tr>
<td>7</td>
<td>0.79</td>
<td>0.76</td>
<td>0.80</td>
</tr>
<tr>
<td>9</td>
<td>0.79</td>
<td>0.60</td>
<td>0.74</td>
</tr>
<tr>
<td>11</td>
<td>0.81</td>
<td>0.77</td>
<td>0.88</td>
</tr>
<tr>
<td>12</td>
<td>0.74</td>
<td>0.34</td>
<td>0.90</td>
</tr>
</tbody>
</table>
Case study 3

Consumer preferences for raspberry beverages

• Two regions (geographies)

• 28 beverages

• Panel 1 (123 assessors): 1 = dislike extremely, 9 = like extremely

• Panel 2 (227 assessors): 1 = bad, 5 = good

Questions:

• Difference in accuracy among scales
• Differences in ratings among regions
• Correlation among attributes
Difference in accuracy

Linear model:

\[ m_{ij} = \alpha_i + u_{jk} + e_{ijk} \] (2.9)

- \( \alpha_i \) = effect of \( i \)-th product
- \( u_{jk} \) = effect of \( j \)-th assessor on \( k \)-th scale
- \( e_{ijk} \) = error

\[ u_{jk} \sim N(0, \sigma^2_{u(k)}) \]
\[ e_{ijk} \sim N(0, \sigma^2_{e(k)}) \]

Separate set of thresholds for both scales!

Hypotheses of interest:

- \( H_{01} : \sigma^2_{u(1)} = \sigma^2_{u(2)} \)
- \( H_{02} : \sigma^2_{e(1)} = \sigma^2_{e(2)} \)
Table 2.9: Likelihood statistics and variance component estimates for liking, raspberry beverages.

**Geography 1:**

<table>
<thead>
<tr>
<th>Model restriction</th>
<th>$-2 \log L$</th>
<th>$\sigma^2_{e(1)}$</th>
<th>$\sigma^2_{e(2)}$</th>
<th>$\sigma^2_{u(1)}$</th>
<th>$\sigma^2_{u(2)}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>41767.6</td>
<td>1.06</td>
<td>1</td>
<td>0.43</td>
<td>0.21</td>
</tr>
<tr>
<td>$\sigma^2_{u(1)} = \sigma^2_{u(2)}$</td>
<td>41782.2</td>
<td>0.94</td>
<td>1</td>
<td>0.28</td>
<td>0.28</td>
</tr>
<tr>
<td>$\sigma^2_{e(1)} = \sigma^2_{e(2)}$</td>
<td>41768.5</td>
<td>1</td>
<td>1</td>
<td>0.39</td>
<td>0.21</td>
</tr>
<tr>
<td>$\sigma^2_{e(1)} = \sigma^2_{e(2)}; \quad \sigma^2_{u(1)} = \sigma^2_{u(2)}$</td>
<td>41783.9</td>
<td>1</td>
<td>1</td>
<td>0.30</td>
<td>0.30</td>
</tr>
</tbody>
</table>
Geography 2:

<table>
<thead>
<tr>
<th>Model restriction</th>
<th>$-2 \log L$</th>
<th>$\sigma_{e(1)}^2$</th>
<th>$\sigma_{e(2)}^2$</th>
<th>$\sigma_{u(1)}^2$</th>
<th>$\sigma_{u(2)}^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>23553.0</td>
<td>1.19</td>
<td>1</td>
<td>0.53</td>
<td>0.19</td>
</tr>
<tr>
<td>$\sigma_{u(1)}^2 = \sigma_{u(2)}^2$</td>
<td>23569.3</td>
<td>1.03</td>
<td>1</td>
<td>0.29</td>
<td>0.29</td>
</tr>
<tr>
<td>$\sigma_{e(1)}^2 = \sigma_{e(2)}^2$; $\sigma_{u(1)}^2 = \sigma_{u(2)}^2$</td>
<td>23569.5</td>
<td>1</td>
<td>1</td>
<td>0.38</td>
<td>0.19</td>
</tr>
</tbody>
</table>
Differences in ratings among regions

\[ m_{ijh} = \alpha_i + \gamma_{ih} + u_{jh} + e_{ijh} \]  

(2.11)

where

\( \alpha_i \) = main effect of \( i \)-th product

\( \gamma_{ih} \) = deviation of \( i \)-th product in \( h \)-th geography (\( h = 1, 2 \))

\( u_{jh} \) = effect of \( j \)-th assessor in \( h \)-th geography,

\( u_{jh} \sim N(0, \sigma_u^2) \)

\( e_{ijh} \) = error (standard normal)

Ratings the same?

\( H_0: \gamma_{i1} = \gamma_{i2} \) for every product \( i \)
Table 2.12: Comparison of mixed threshold models for testing
H₀: γ₁ᵢ = γ₂ᵢ (consumers behave identically in both regions).
Attribute: liking (case study 3).

<table>
<thead>
<tr>
<th>Scale</th>
<th>Model for ( m_{ijh} )</th>
<th>AIC</th>
<th>Test of H₀: ( \gamma_1 = \gamma_2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( \alpha_i + \gamma_{ih} + u_{jh} + e_{ijh} )</td>
<td>27360.4</td>
<td>Wald-( \chi^2 ) 2.20 p-value 0.0006</td>
</tr>
<tr>
<td></td>
<td>( \alpha_i + \gamma_{ih} + \lambda_i u_{jh} + e_{ijh} )</td>
<td>27309.0</td>
<td></td>
</tr>
<tr>
<td>9 point</td>
<td>( \alpha_i + \gamma_{ih} + u_{jh} + e_{ijh} )</td>
<td>37957.8</td>
<td>3.09 &lt;0.0001</td>
</tr>
<tr>
<td></td>
<td>( \alpha_i + \gamma_{ih} + \lambda_i u_{jh} + e_{ijh} )</td>
<td>37895.5</td>
<td>3.05 &lt;0.0001</td>
</tr>
</tbody>
</table>
Correlation among attributes

- liking and appearance
- geography 1
- five-point ordinal scale

Bivariate mixed model:

\[
\begin{align*}
    m_{ij1} &= \alpha_{i1} + u_{j1} + e_{ij1} \\
    m_{ij2} &= \alpha_{i2} + u_{j2} + e_{ij2}
\end{align*}
\]

where

- \(\alpha_{ih}\) = effect of \(i\)-th product for \(h\)-th attribute
- \(u_{jh}\) = effect of \(j\)-th assessor for \(h\)-th attribute
- \(e_{ijh}\) = residual corresponding to \(m_{ijh}\)
Distributional assumptions:

\[
\begin{pmatrix}
u_{j1} \\
u_{j2}
\end{pmatrix} \sim BN\left[\begin{pmatrix}0 \\ 0\end{pmatrix}, \begin{pmatrix}\sigma^2_{u1} & \rho_u \sigma_{u1} \sigma_{u2} \\ \rho_u \sigma_{u1} \sigma_{u2} & \sigma^2_{u2}\end{pmatrix}\right]
\]

\[
\begin{pmatrix}e_{ij1} \\ e_{ij2}\end{pmatrix} \sim BN\left[\begin{pmatrix}0 \\ 0\end{pmatrix}, \begin{pmatrix}1 & \rho_e \\ \rho_e & 1\end{pmatrix}\right]
\]

where

\(BN(.,.) = \text{bivariate normal distribution}\)

\(\rho_u = \text{between-product correlation}\)

\(\rho_e = \text{within-product correlation}\)
Bivariate multinomial probabilities of observed scores:

\[
\pi_{ab} = P(y_1 = a, y_2 = b) = \Phi(\theta_{1a} - \eta_1, \theta_{2b} - \eta_2, \rho_e) - \sum_{r=1}^{a} \sum_{s=1}^{b} \pi_{rs}
\]

\[a = 1, \ldots, 5; b = 1, \ldots, 5\]

with

\[y_1, y_2 = \text{observed scores for liking and appearance}\]
\[\Phi = \text{bivariate standard normal c.d.f}\]
\[\theta_{1a}, \theta_{2b} = \text{thresholds}; \theta_{15} = \theta_{25} = \infty\]
\[\eta_1, \eta_2 = \text{linear predictors (means) on latent scale}\]

**Result:** \[\hat{\rho}_e = 0.588 \quad (s.e. = 0.011)\] (within-product analysis)
Case study 4

Recovery of inter-assessor information

- Sensory panels of 33 assessors
- Three groups of cheese products
- Hedonic 9-point scale
- Tests in incomplete blocks: five or six products per assessor

<table>
<thead>
<tr>
<th>Type of analysis</th>
<th>Assumption about assessor effects</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intra-assessor analysis</td>
<td>fixed</td>
</tr>
<tr>
<td>Recovery of inter-assessor</td>
<td>random</td>
</tr>
<tr>
<td>information</td>
<td></td>
</tr>
</tbody>
</table>
Table 2.14: Average standard errors of a difference (s.e.d.).

<table>
<thead>
<tr>
<th>Use of inter-assessor Information</th>
<th>Hard cheese</th>
<th>Fresh cheese with additives</th>
<th>Fresh cheese</th>
</tr>
</thead>
<tbody>
<tr>
<td>yes</td>
<td>0.375</td>
<td>0.407</td>
<td>0.369</td>
</tr>
<tr>
<td>no</td>
<td>0.385</td>
<td>0.423</td>
<td>0.380</td>
</tr>
</tbody>
</table>
3. Numerical issues

$y$ = observed scores
$u$ = random effects

Joint density of $u$ and $y$:

$$\phi_{yu}(y, u) = \phi_{yu}(y|u)\phi_u(u) \quad (3.1)$$

where

$\phi_{yu}(y|u)$ = conditional density of $y$, given $u$ (multinomial)
$\phi_u(u)$ = marginal density of $u$ (normal)

Estimation $\Rightarrow$ maximize

$$\phi_y(y) = \int \phi_{yu}(y | v)\phi_u(v)dv \quad (3.2)$$
Numerical integration

• Approximations of the likelihood
  ⇒ poor performance for small samples

• Gaussian quadrature
  ⇒ Computationally demanding
  ⇒ Handles only one subject level

All methods rely on asymptotic theory!
4. Simulations

4.1 No random effects

Hypothesis:
Show that ANOVA may be problematic with ordinal data

Simulate according to:

\[ m_{ijk} = \alpha_i + u_j + e_{ijk} \]  \hspace{1cm} (4.1)

where

\( \alpha_i \) = product effect \( (i = 1, \ldots, I) \)
\( u_j \) = random assessor effect \( (j = 1, \ldots, J) \),
\( e_{ijk} \) = error (standard normal distribution)
\( (k = 1, \ldots, R) \)

Three (nine) categories
Simulations under:

• global null hypothesis
• departure from global null

Teste evaluated:

Global $H_0$:
• ANOVA F-test
• LR test for threshold model

$H_0$: $\alpha_1 = \alpha_2$:
• ANOVA-LSD
• paired t-test
• Wald-test for threshold model
Table 4.5: Simulation results for threshold model with fixed effects, Type I error at nominal significance level of 5%, Global $H_0$ false, nine categories ($\theta_1 = -0.5$, $\theta_2 = -0.4$, $\theta_3 = -0.3$, $\theta_4 = -0.1$, $\theta_5 = 0.1$, $\theta_6 = 0.2$, $\theta_7 = 0.3$, $\theta_8 = 0.5$, $\sigma_u = 0$). Test of $H_0$: $\alpha_1 = \alpha_2$.

<table>
<thead>
<tr>
<th>$R$</th>
<th>$J$</th>
<th>$I$</th>
<th>$\alpha_1$</th>
<th>$\alpha_2$</th>
<th>$\alpha_{i&gt;2}$</th>
<th>ANOVA</th>
<th>Threshold</th>
<th>Paired t-test</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>50</td>
<td>20</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0.420</td>
<td>0.082</td>
<td>0.061</td>
</tr>
<tr>
<td>1</td>
<td>50</td>
<td>20</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0.019</td>
<td>0.036</td>
</tr>
<tr>
<td>4</td>
<td>20</td>
<td>20</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0.410</td>
<td>0.068</td>
<td>0.058</td>
</tr>
<tr>
<td>4</td>
<td>20</td>
<td>20</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0.039</td>
<td>0.049</td>
</tr>
</tbody>
</table>
4.2 Pairwise comparison with heteroscedasticity on latent scale

Hypothesis:

Heteroscedasticity on latent scale of threshold model invalidates ANOVA

Simulated according to:

\[ m_{ij} = \alpha_i + e_{ij} \]  \hspace{1cm} (4.2)

where

\[ \alpha_i \] = product effect \((i = 1, 2)\)

\[ e_{ij} \] = error of \(j\)-th replicate for \(i\)-th product \((j = 1, \ldots, J_i)\)

\[ \sigma_1 \] = standard deviation of \(e_{1j}\) (product 1)

\[ \sigma_2 \] = standard deviation of \(e_{2j}\) (product 2)

Nine categories.
Tests evaluated:

- Simple t-test
- Satterthwaite t-test (heteroscedastic)
- Standard threshold model
- Threshold model with heteroscedastic errors
Table 4.6: Simulation results for heteroscedastic threshold model, Type I error at nominal significance level of 5%, $\alpha_1 = \alpha_2 = 0$. Test of $H_0: \alpha_1 = \alpha_2$.

<table>
<thead>
<tr>
<th>$\theta_1$</th>
<th>$\theta_2$</th>
<th>$\theta_3$</th>
<th>$\theta_4$</th>
<th>$\theta_5$</th>
<th>$\theta_6$</th>
<th>$\theta_7$</th>
<th>$\theta_8$</th>
<th>$\sigma_2$</th>
<th>t-test</th>
<th>$\sigma_1 = \sigma_2$</th>
<th>$\sigma_1 \neq \sigma_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$J_1 = 50$; $J_2 = 150$:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$-2.0$</td>
<td>$-1.8$</td>
<td>$1.0$</td>
<td>$1.2$</td>
<td>$1.4$</td>
<td>$1.6$</td>
<td>$1.8$</td>
<td>$2.0$</td>
<td>$2$</td>
<td>$0.624$</td>
<td>$0.023$</td>
<td>$0.063$</td>
</tr>
<tr>
<td>$-2.0$</td>
<td>$-1.7$</td>
<td>$-1.5$</td>
<td>$0.0$</td>
<td>$1.0$</td>
<td>$1.5$</td>
<td>$1.7$</td>
<td>$2.0$</td>
<td>$2$</td>
<td>$0.085$</td>
<td>$0.005$</td>
<td>$0.045$</td>
</tr>
<tr>
<td>$-2.0$</td>
<td>$-1.3$</td>
<td>$0.5$</td>
<td>$0.9$</td>
<td>$1.2$</td>
<td>$1.5$</td>
<td>$1.8$</td>
<td>$2.0$</td>
<td>$2$</td>
<td>$0.409$</td>
<td>$0.008$</td>
<td>$0.050$</td>
</tr>
<tr>
<td>$-2.0$</td>
<td>$-1.5$</td>
<td>$-1.0$</td>
<td>$0.0$</td>
<td>$0.5$</td>
<td>$1.0$</td>
<td>$1.5$</td>
<td>$2.0$</td>
<td>$2$</td>
<td>$0.065$</td>
<td>$0.007$</td>
<td>$0.038$</td>
</tr>
<tr>
<td>$-2.0$</td>
<td>$-1.8$</td>
<td>$1.0$</td>
<td>$1.2$</td>
<td>$1.4$</td>
<td>$1.6$</td>
<td>$1.8$</td>
<td>$2.0$</td>
<td>$1.4$</td>
<td>$0.260$</td>
<td>$0.018$</td>
<td>$0.057$</td>
</tr>
<tr>
<td>$-2.0$</td>
<td>$-1.7$</td>
<td>$-1.5$</td>
<td>$0.0$</td>
<td>$1.0$</td>
<td>$1.5$</td>
<td>$1.7$</td>
<td>$2.0$</td>
<td>$1.4$</td>
<td>$0.072$</td>
<td>$0.029$</td>
<td>$0.055$</td>
</tr>
<tr>
<td>$-2.0$</td>
<td>$-1.3$</td>
<td>$0.5$</td>
<td>$0.9$</td>
<td>$1.2$</td>
<td>$1.5$</td>
<td>$1.8$</td>
<td>$2.0$</td>
<td>$1.4$</td>
<td>$0.209$</td>
<td>$0.022$</td>
<td>$0.052$</td>
</tr>
<tr>
<td>$-2.0$</td>
<td>$-1.5$</td>
<td>$-1.0$</td>
<td>$0.0$</td>
<td>$0.5$</td>
<td>$1.0$</td>
<td>$1.5$</td>
<td>$2.0$</td>
<td>$1.4$</td>
<td>$0.058$</td>
<td>$0.024$</td>
<td>$0.057$</td>
</tr>
<tr>
<td>$\theta_1$</td>
<td>$\theta_2$</td>
<td>$\theta_3$</td>
<td>$\theta_4$</td>
<td>$\theta_5$</td>
<td>$\theta_6$</td>
<td>$\theta_7$</td>
<td>$\theta_8$</td>
<td>$\sigma_2$</td>
<td>t-test</td>
<td>$\sigma_1 = \sigma_2$</td>
<td>$\sigma_1 \neq \sigma_2$</td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>-2.0</td>
<td>-1.8</td>
<td>1.0</td>
<td>1.2</td>
<td>1.4</td>
<td>1.6</td>
<td>1.8</td>
<td>2.0</td>
<td>2</td>
<td>0.336</td>
<td>0.151</td>
<td>0.053</td>
</tr>
<tr>
<td>-2.0</td>
<td>-1.7</td>
<td>-1.5</td>
<td>0.0</td>
<td>1.0</td>
<td>1.5</td>
<td>1.7</td>
<td>2.0</td>
<td>2</td>
<td>0.071</td>
<td>0.132</td>
<td>0.068</td>
</tr>
<tr>
<td>-2.0</td>
<td>-1.3</td>
<td>0.5</td>
<td>0.9</td>
<td>1.2</td>
<td>1.5</td>
<td>1.8</td>
<td>2.0</td>
<td>2</td>
<td>0.242</td>
<td>0.107</td>
<td>0.048</td>
</tr>
<tr>
<td>-2.0</td>
<td>-1.5</td>
<td>-1.0</td>
<td>0.0</td>
<td>0.5</td>
<td>1.0</td>
<td>1.5</td>
<td>2.0</td>
<td>2</td>
<td>0.041</td>
<td>0.094</td>
<td>0.042</td>
</tr>
<tr>
<td>-2.0</td>
<td>-1.8</td>
<td>1.0</td>
<td>1.2</td>
<td>1.4</td>
<td>1.6</td>
<td>1.8</td>
<td>2.0</td>
<td>1.4</td>
<td>0.130</td>
<td>0.110</td>
<td>0.057</td>
</tr>
<tr>
<td>-2.0</td>
<td>-1.7</td>
<td>-1.5</td>
<td>0.0</td>
<td>1.0</td>
<td>1.5</td>
<td>1.7</td>
<td>2.0</td>
<td>1.4</td>
<td>0.056</td>
<td>0.099</td>
<td>0.054</td>
</tr>
<tr>
<td>-2.0</td>
<td>-1.3</td>
<td>0.5</td>
<td>0.9</td>
<td>1.2</td>
<td>1.5</td>
<td>1.8</td>
<td>2.0</td>
<td>1.4</td>
<td>0.092</td>
<td>0.088</td>
<td>0.059</td>
</tr>
<tr>
<td>-2.0</td>
<td>-1.5</td>
<td>-1.0</td>
<td>0.0</td>
<td>0.5</td>
<td>1.0</td>
<td>1.5</td>
<td>2.0</td>
<td>1.4</td>
<td>0.052</td>
<td>0.077</td>
<td>0.051</td>
</tr>
</tbody>
</table>

$J_1 = 150; \ J_2 = 50$: 
4.3 One random effect

Hypothesis:

Asymptotics are even more critical for mixed models

Simulated according to:

\[ m_{ijk} = \alpha_i + u_j + f_{ij} + e_{ijk} \]  

where

\( f_{ij} = \text{random assessor} \times \text{product interaction} \)
Table 4.7: Simulation results for threshold model with fixed and random effects (4.3), Type I errors at nominal significance level of 5%. Global $H_0$ true, $R = 4$ replicates, three categories. Test of $H_0: \alpha_1 = \alpha_2$. $\$: Wald-Test, §: LR-test.

<table>
<thead>
<tr>
<th>$\theta_1$</th>
<th>$\theta_2$</th>
<th>$\sigma_u$</th>
<th>$\sigma_f$</th>
<th>$J$</th>
<th>$I$</th>
<th>Overall ANOVA</th>
<th>Threshold model</th>
<th>Paired ANOVA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$\alpha_1 = \alpha_2$</td>
<td>$\alpha_1 = \alpha_2$</td>
<td>$\alpha_1 = \alpha_2$</td>
</tr>
<tr>
<td>-0.5</td>
<td>0.5</td>
<td>0</td>
<td>0.2</td>
<td>10</td>
<td>4</td>
<td>0.059</td>
<td>0.052</td>
<td>0.078</td>
</tr>
<tr>
<td>-0.5</td>
<td>0.5</td>
<td>0</td>
<td>0.2</td>
<td>10</td>
<td>8</td>
<td>0.051</td>
<td>0.045</td>
<td>0.064</td>
</tr>
<tr>
<td>-1.0</td>
<td>1.0</td>
<td>0</td>
<td>0.2</td>
<td>10</td>
<td>4</td>
<td>0.034</td>
<td>0.051</td>
<td>0.092</td>
</tr>
<tr>
<td>-1.0</td>
<td>1.0</td>
<td>0</td>
<td>0.2</td>
<td>10</td>
<td>8</td>
<td>0.049</td>
<td>0.053</td>
<td>0.072</td>
</tr>
<tr>
<td>-0.5</td>
<td>0.5</td>
<td>0</td>
<td>0.5</td>
<td>10</td>
<td>4</td>
<td>0.051</td>
<td>0.052</td>
<td>0.140</td>
</tr>
<tr>
<td>-0.5</td>
<td>0.5</td>
<td>0</td>
<td>0.5</td>
<td>10</td>
<td>8</td>
<td>0.064</td>
<td>0.051</td>
<td>0.122</td>
</tr>
<tr>
<td>-1.0</td>
<td>1.0</td>
<td>0</td>
<td>0.5</td>
<td>10</td>
<td>4</td>
<td>0.034</td>
<td>0.045</td>
<td>0.122</td>
</tr>
<tr>
<td>-1.0</td>
<td>1.0</td>
<td>0</td>
<td>0.5</td>
<td>10</td>
<td>8</td>
<td>0.045</td>
<td>0.040</td>
<td>-</td>
</tr>
</tbody>
</table>
5. Conclusion

- ANOVA can fail badly with ordinal data
- Threshold model well suited to analyse ordinal data
- Many experimental settings call for mixed models
- Threshold model can be extended to cover random effects
- Allows great flexibility in answering relevant research questions
- Maximum likelihood estimation computationally demanding
- Asymptotics work for large samples only
- Tests in particular need to be taken with a gain of salt
- More simulations needed