Statistical and Thurstonian Models
Advances in Discrimination Testing

Rune Haubo B Christensen
with special thanks to
Prof. Per Bruun Brockhoff

DTU Informatics, IMM
Section for Statistics
Technical University of Denmark
rhbc@imm.dtu.dk

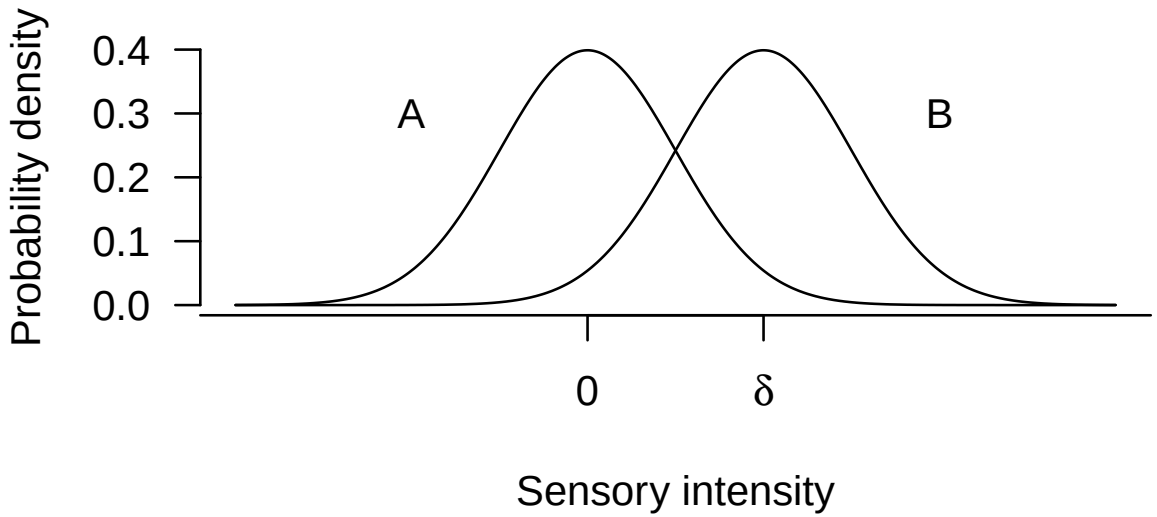
February 29th 2012

Outline

Outline

What is a Thurstonian Model?

1. A common scale for quantification of "Sensory Difference" $\rightarrow d'$ (or δ)
2. A psycho-physical model for the cognitive process
3. A stochastic model for the data-generating mechanism

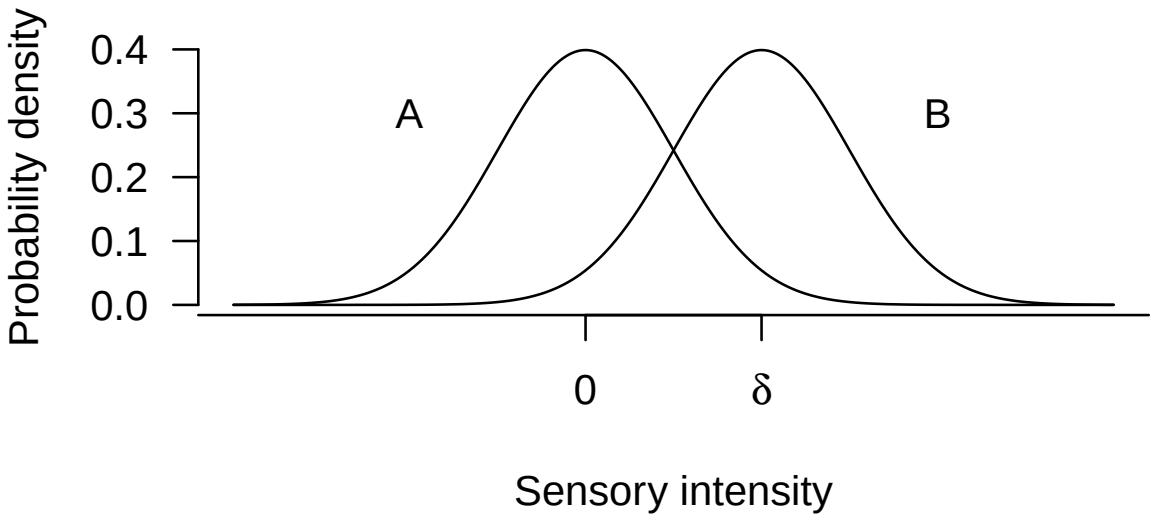


(Thurstone, 1927)

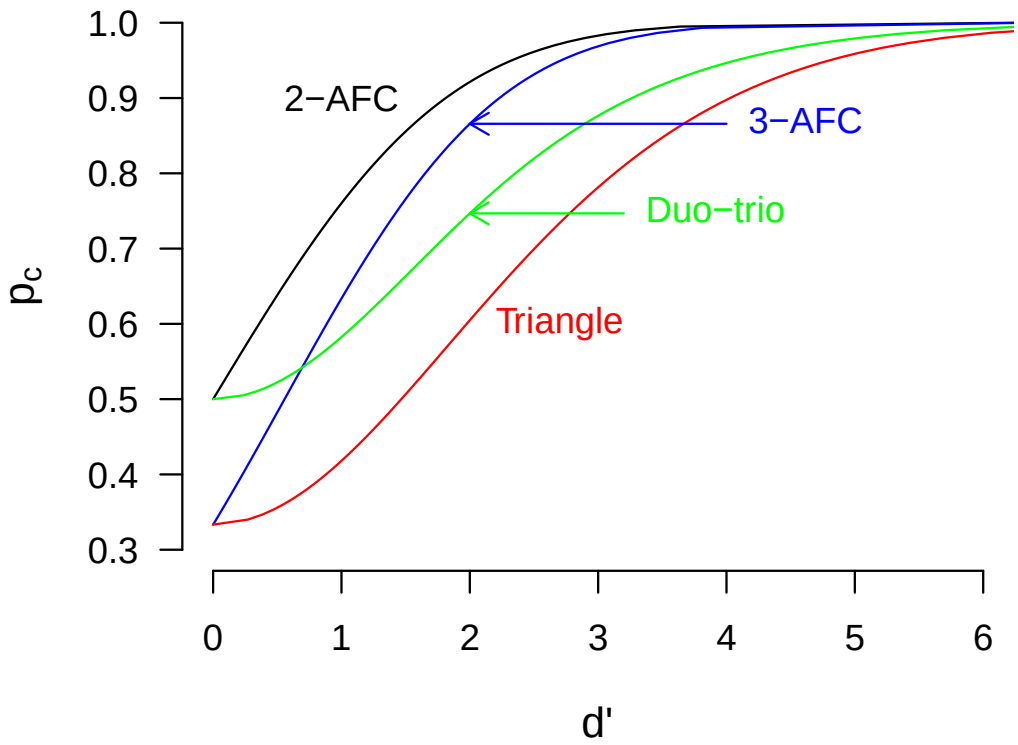
What is Thurstonian model really?

Assumptions:

- Perceptions are random and normally distributed (constant stimulus)
- Often: constant variance
- Decision rules are non-random (and given by the protocol)



Psychometric functions — linking p_c and d'



Only for protocols with a binomial outcome

$\rightarrow$ m-AFC, Triangle, Duo-Trio, Tetrads, 2-out-of-5, ...

Why Thurstonian Models?

There is a Thurstonian model for every sensory discrimination protocol

Other measures of sensory difference:

- Proportion correct (p_c)
- Proportion of discriminators (p_d)

Why d'?

- d' is universal — it can be estimated for all protocols.
- p_c and p_d does not exist for A-not A, same-different etc.
- p_c and p_d are protocol dependent when they do exist

Empirical evidence and "the paradox of discriminatory nondiscriminators"
(Byers and Abrams, 1953; Frijters, 1979)

Thurstonian and statistical models

Thurstonian models for some common protocols can be identified as well-known statistical models.

Protocol	Statistical model	Source
Triangle, m-AFC, ...	GLM with special links	(Brockhoff and Christensen, 2010)
A-not A	GLM with probit link	(Brockhoff and Christensen, 2010)
A-not A w. sureness	CLM	(Christensen et al., 2011)
Paired pref.	GLM with probit link	
Paired pref. (no-pref.)	CLM	(Christensen et al., 2012)

GLM: Generalized linear model (McCullagh and Nelder, 1989)
CLM: Cumulative link model (McCullagh, 1980)

Software for GLM: **R**-package **sensR**
Software for CLM: **R**-package **ordinal**

Thurstonian and statistical models

Why statistical models for sensory discrimination?

Advantages of identification of Thurstonian models as well-known statistical models:

1. Standard software for estimation, CI and tests
2. Regression extension of Thurstonian models
3. Ready extension to replicated situations via mixed effects models

Regression extension of Thurstonian models

Main idea:
Combine regression and ANOVA methods with Thurstonian models

- Control for experimental factors
- Joint model for several treatment effects
- Model order effects (order of servings)
- Detect and adjust for learning and fatigue effects
- Adjust for sessions and replicates

Outline

Sensory discrimination experiments — an example

Table: Triangle experiment with 80 men and 80 women.

	Men		Women	
Concentration	Correct	Total	Correct	Total
1	9	20	13	20
2	11	20	14	20
3	13	20	16	20
4	14	20	18	20

Objective:
What is the sensory difference between products?
How does d' depend on gender and concentration?

Analysis strategy — conventional 2-step approach

Step 1: Estimate all 8 d's
Step 2: Post-hoc comparisons

d' estimates:

	Concentration			
Gender	1	2	3	4
Men	1.19	1.72	2.23	2.50
Women	2.23	2.50	3.13	4.03
Total	1.72	2.10	2.65	3.13

χ^2 test for concentration effect: $p = 0.095$ (test proposed by Bi et al. (1997))

Remaining questions:

- How does "sensory difference" depend on gender and concentration?
- Is the effect of concentration different for men and women?
- Cumbersome, sub-optimal, silent about effect estimates

Analysis strategy — a regression approach

A regression/ANOVA approach:

$$\begin{aligned} \texttt{correct/total} &= \texttt{gender} + \texttt{conc} + \varepsilon \\ &= \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{E} \qquad E_i \sim N(0, \sigma^2) \end{aligned}$$

- Effects and interactions easy to formulate and test
- An invalid model
- Difficult interpretation of parameters

Analysis strategy — a logistic regression model

The statisticians approach — a logistic regression model:

$$\log\left(\frac{\pi_c}{1-\pi_c}\right) = \texttt{gender} + \texttt{conc}$$

$$g_{\text{logit}}(\pi_c) = \boldsymbol{X}\boldsymbol{\beta}$$

$$\texttt{correct}_i \sim \text{binom}(\pi_{ci}, \texttt{total}_i)$$

- A generalized linear model (GLM) (McCullagh and Nelder, 1989)
- A valid model
- Effects and interactions easy to formulate and test
- Difficult interpretation of parameters

Analysis strategy — a Thurstonian GLM

Our suggestion (Brockhoff and Christensen, 2010):
A Thurstonian GLM:

$$\begin{aligned} g_{\text{triangle}}(\pi_c) &= \texttt{gender} + \texttt{conc} \\ &= \boldsymbol{X}\boldsymbol{\beta} \\ \texttt{correct}_i &\sim \text{binom}(\pi_{ci}, \texttt{total}_i) \end{aligned}$$

- A valid model
- Effects and interactions easy to formulate and test
- Thurstonian interpretation of parameters!

Does d' depend on gender and concentration?

ANODE: Analysis of deviance — an extension of ANOVA

Source	df	deviance	p value
Total	7	13.05	
Model	2	12.64	0.0018
Gender	1	5.95	0.0147
Conc	1	7.00	0.0081
Residual	5	0.413	0.9950

Results:

- d' depends on gender and concentration
- Effect of concentration much stronger here (before $p = 0.095$)

Remaining questions:

- Is the dependence linear in concentration?
- Is the effect of concentration different for men and women?

Extended ANODE table

Including additional effects in the ANODE table:

Source	df	deviance	p value
Total	7	13.05	
Model	5	12.79	0.02542
Gender	1	5.95	0.0147
Conc(linear)	1	7.00	0.0081
Conc(remain)	2	0.045	0.9779
Gender:conc(linear)	1	0.120	0.7291
Residual	2	0.259	0.8784

Results:

- Effect of concentration is linear
- No interaction between concentration and gender

Main message: Complex hypotheses are straight forward to test!

What are the magnitudes of Conc and Gender effects?

Table: Parameter estimates

Effect	Estimate	Standard Error	z value	p value
Men	1.160	0.427	2.716	0.007
Women	2.188	0.401	5.455	< 0.001
conc	0.502	0.197	2.544	0.011

Effects are directly interpretable as d's

Using **sensR** (Christensen and Brockhoff, 2011) in **R**:

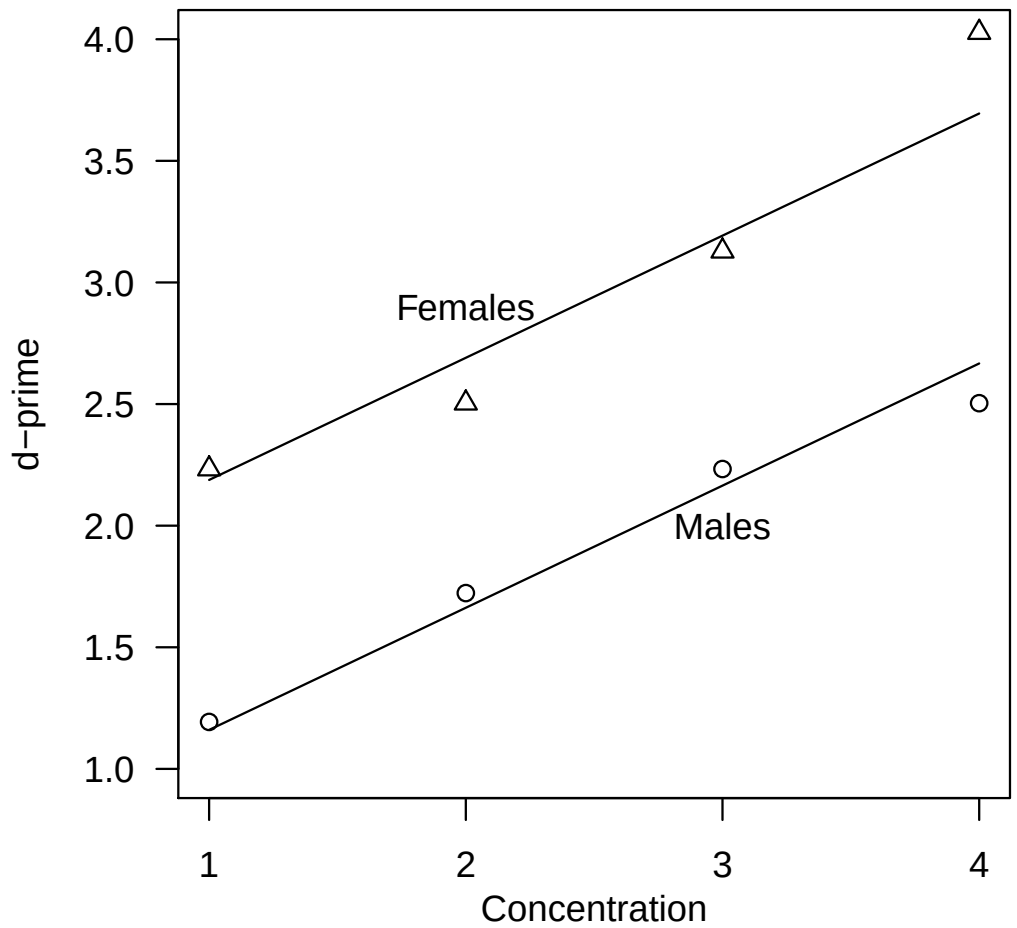
Fit model:

```
> model <- glm(y ~ gender + conc, family=triangle)
```

ANODE table:

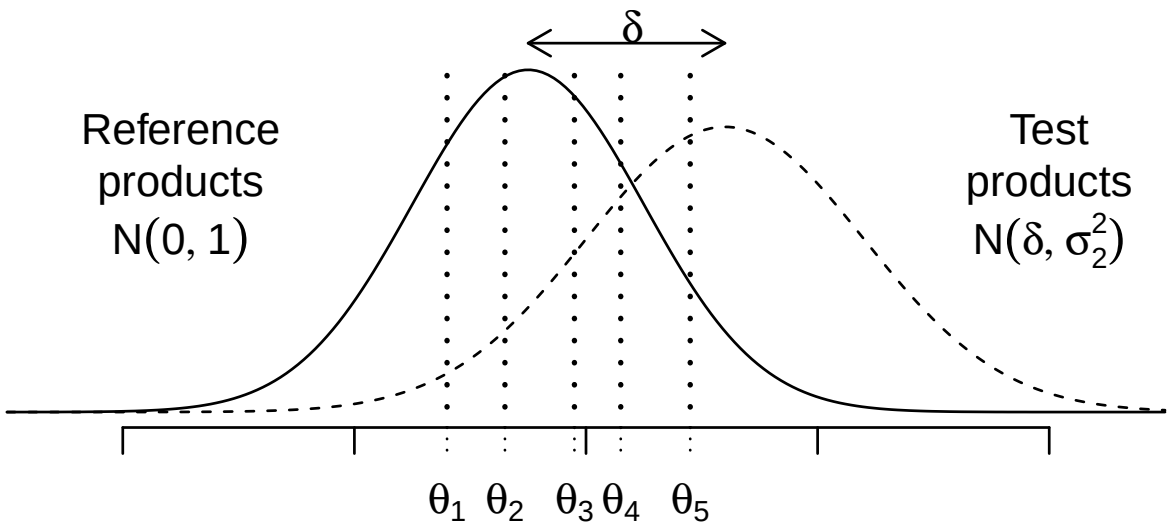
```
> drop1(model, test="Chisq")
```

Illustration of Concentration and Gender effects



Outline

Thurstonian model for the A-not A with sureness protocol



	'Reference'			'Not Reference'		
Product	Sure	Not Sure	Guess	Guess	Not Sure	Sure
Reference	132	161	65	41	121	219
Test	96	99	50	57	156	650

Table: Discrimination of packet soup (Christensen, Cleaver, and Brockhoff, 2011)

Thurstonian model as a Cumulative Probit Model

The bi-normal unequal-variances model:

$$P(S_i \leq \theta_j) = \Phi\left(\frac{\theta_j - \delta(\texttt{prod}_i)}{\sigma(\texttt{prod}_i)}\right)$$

- ML Estimation proposed by (Dorfman and Alf, 1969).
- Identified as a cumulative probit model (DeCarlo, 1998)

Table: Parameter estimates

Effect	Estimate	Standard Error	z value	p value
d'	0.827	0.0766	10.80	< 0.001
$\log \sigma$	0.217	0.0614	3.53	< 0.001
σ	1.242			

- Software: The **ordinal** package (Christensen, 2011)

A model for multiple test products

Table: Five test products were used.

	'Reference'			'Not Reference'		
Product	Sure	Not Sure	Guess	Guess	Not Sure	Sure
Reference	132	161	65	41	121	219
Test_1	36	42	22	19	58	192
Test_2	12	13	4	15	19	121
Test_3	19	23	10	14	24	95
Test_4	18	10	10	5	26	116
Test_5	11	11	4	4	29	126

Research questions:

- Is d' the same for all 5 test products?
- Do we have equal or unequal variances?
- Do all 5 test products have the same perceptual variance?

ANODE table for multiple test products model

- Extend the cumulative probit model to handle several products
- Much better than five separate models!

Source	df	deviance	p value
Total	25	225	
Model	10	197.51	< 0.001
d'	1	159.03	< 0.001
$d'_1, \ldots, d'_5$	4	25.14	< 0.001
σ	1	10.89	< 0.001
$\sigma_1, \ldots, \sigma_5$	4	2.45	0.653
Residual	15	27.50	0.025

- Efficient use of data — more insight
- More accurate estimates, stronger tests

d' estimates and ROC curves

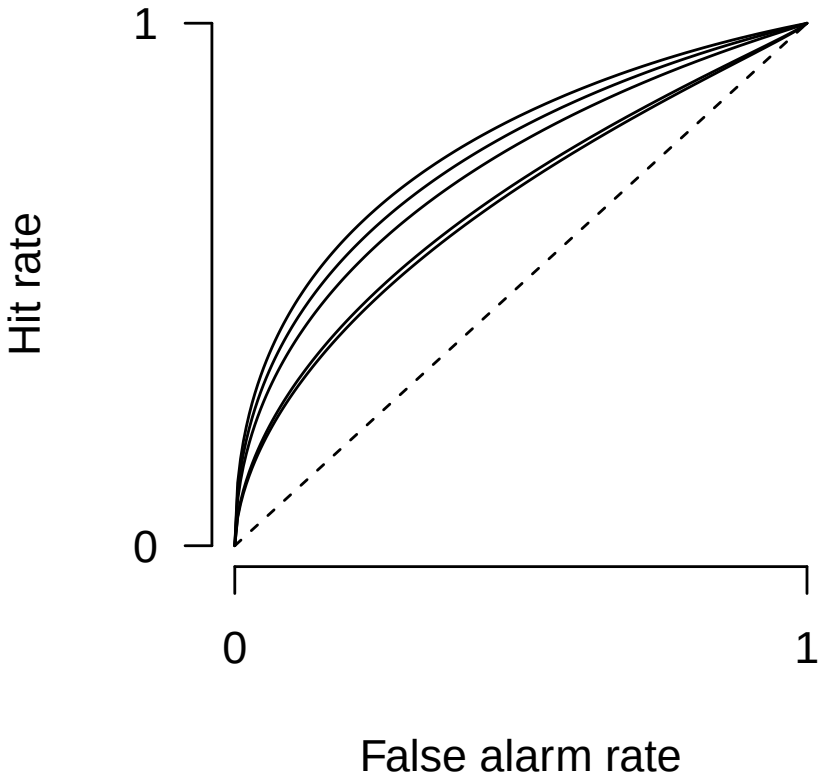


Figure: ROC curves for five test products

Table: Parameter estimates

Effect	Estimate	Std. Err.	p value
δ_1	0.642	0.091	< 0.001
δ_2	1.030	0.130	< 0.001
δ_3	0.601	0.115	< 0.001
δ_4	0.912	0.126	< 0.001
δ_5	1.138	0.135	< 0.001
$\log \sigma$	0.202	0.061	0.001
σ	1.224		0.001

Effects of explanatory variables

- Does d' differ between experimental sessions?
- Is d' higher for some consumers than others?

Type	Parameter	Estimate	Std. Error	p value
Location	δ_2	0.508	0.123	<0.001
	δ_3	0.909	0.135	<0.001
	δ_4	0.471	0.131	<0.001
	δ_5	0.782	0.141	<0.001
	δ_6	1.012	0.147	<0.001
	day: 2	-0.244	0.079	0.002
	soup.type: canned	-0.147	0.065	0.024
	soup.type: dry-mix	0.121	0.083	0.146
	prod: test, day: 2	0.260	0.126	0.039
log(Scale)	prod: test	0.198	0.061	0.001
Scale	prod: test	1.220		

Including assessor effects

Assumptions:

- Assessors do not use the response scale differently
- Assessors do not have different d's

Accommodate this with mixed model extensions:

- Allow normally distributed random effects for assessors

$$P(S_i \leq \theta_j) = \Phi\left(\theta_j - \delta(\texttt{prod}_i) - u(\texttt{assessor}_i)\right) \qquad u \sim N(0, \sigma_u^2)$$

Note: This is similar to assessor effects in models for sensory profiling!

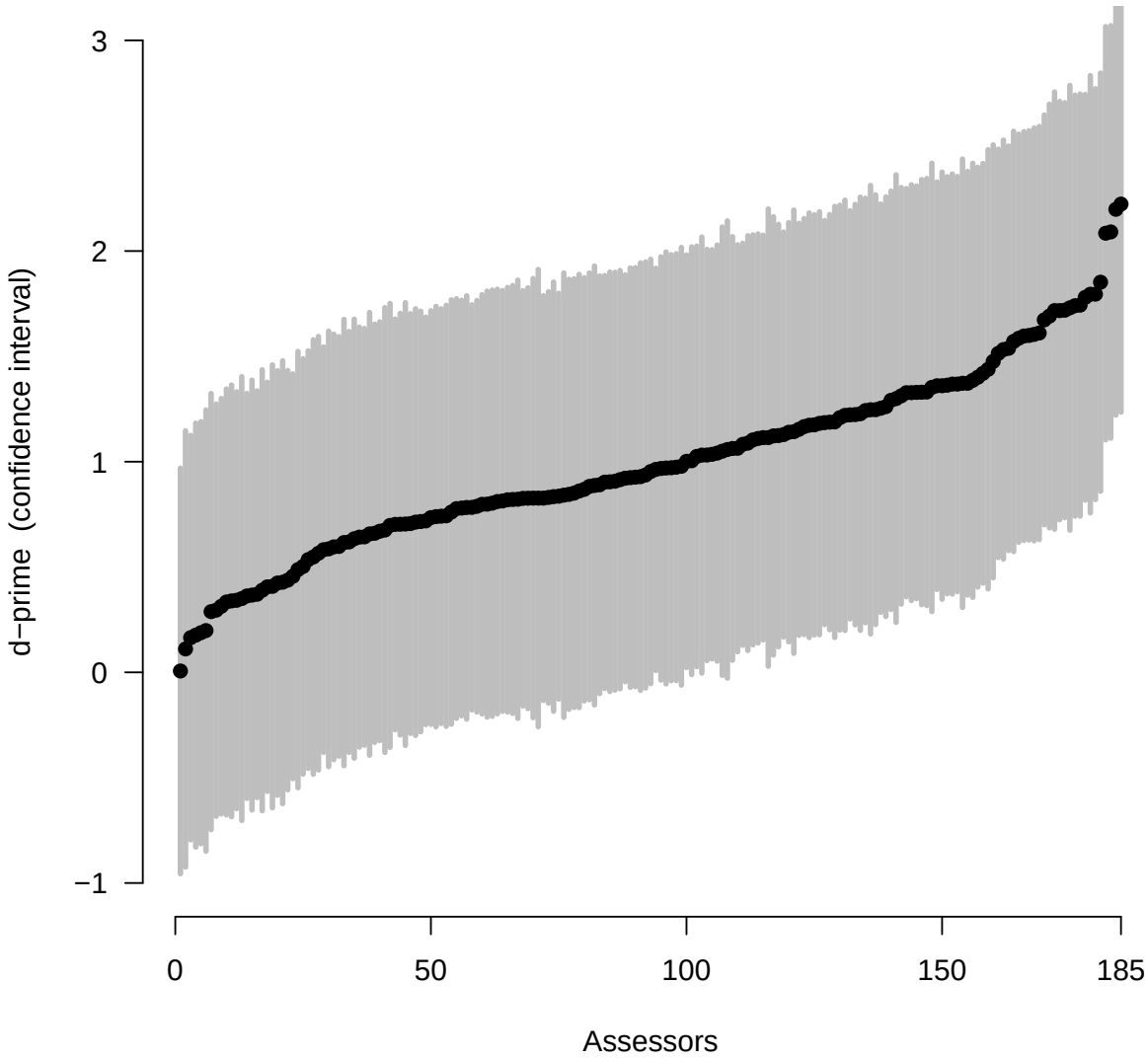
Which assessor effects are present?

- Do assessors use the response scale differently? – Var(θ) (assessor-specific response bias)
- Do assessors have different d's? Var(d')

Source	df	Deviance	p value
Assessor effect			
Var(θ)	1	40.18	< 0.001
Var(d')	1	58.83	< 0.001
Var(θ) + Var(d')	1	1.529	0.216

Conclusion:
Assessor-specific d's is the structure most supported by the data.

Inference for respondents — respondent-specific d's



Outline

The bitterness of white wines

Objective:
How does perceived bitterness depend on temperature and contact?

Table: The wine data (Randall, 1989), N=72

Variables	Type	Values
bitterness	response	1, 2, 3 ,4, 5
		less — more
temperature	predictor	cold, warm
contact	predictor	no, yes
judges	random	1, . . . , 9

Temperature and contact between juice and skins can be controlled when crushing grapes during wine production.

Data for the bitterness of white wines

Table: Ratings of the bitterness of some white wines. Data are adopted from Randall (1989).

			Judge								
Temperature	Contact	Bottle	1	2	3	4	5	6	7	8	9
cold	no	1	2	1	2	3	2	3	1	2	1
cold	no	2	3	2	3	2	3	2	1	2	2
cold	yes	3	3	1	3	3	4	3	2	2	3
cold	yes	4	4	3	2	2	3	2	2	3	2
warm	no	5	4	2	5	3	3	2	2	3	3
warm	no	6	4	3	5	2	3	4	3	3	2
warm	yes	7	5	5	4	5	3	5	2	3	4
warm	yes	8	5	4	4	3	3	4	3	4	4

Appropriate models for the Wine data

Ordinal data — **not continuous** data

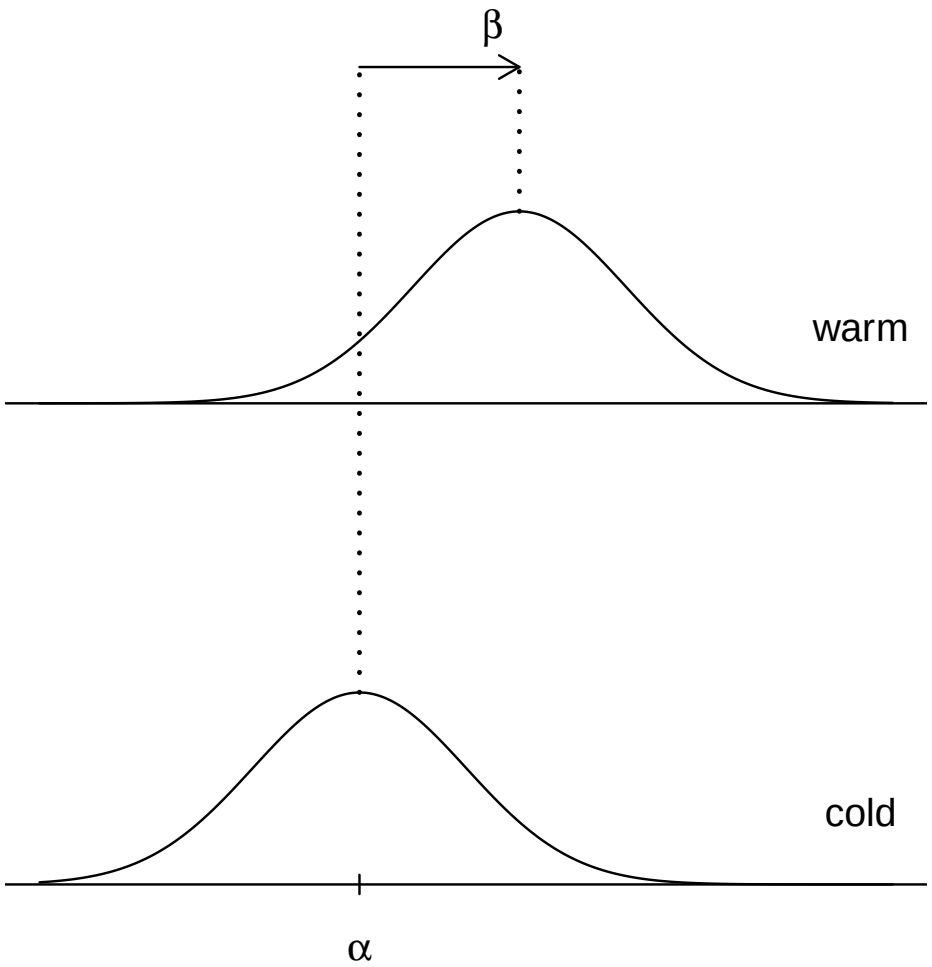
A linear regression model on the scores (1,. . . ,5)?
Breach of assumptions:

- The scores are **not** normally distributed
- A score of "4" is not twice as much as "2"
- Variance not likely to be constant

Our approach:
A cumulative link model (CLM)

- Only use information about ordering
- Intuitively: A linear model that respects the ordinal nature of the response

Thurstonian motivation of the cumulative link model

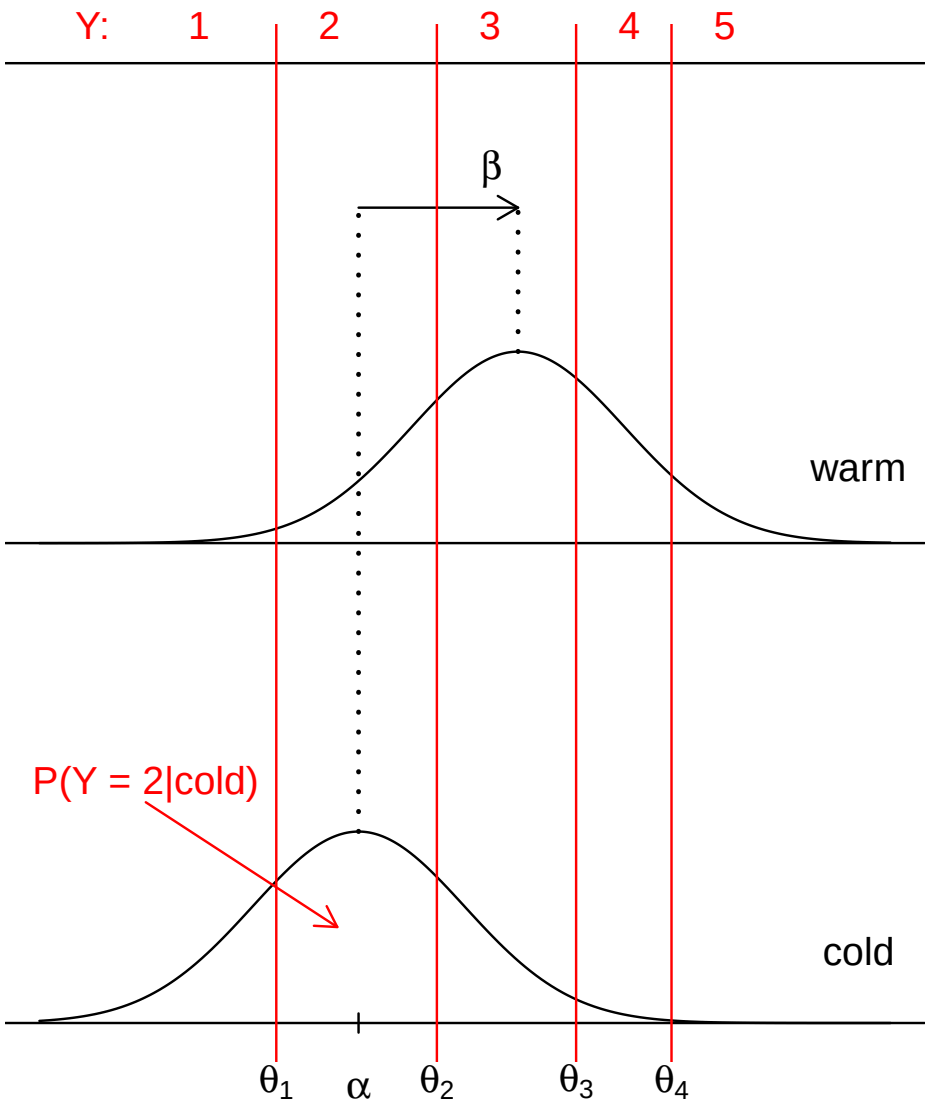


- *Latent* bitterness follows a linear model:

$$S_i = \alpha + \boldsymbol{x}_i^T \boldsymbol{\beta} + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2)$$
$$= \alpha + \beta(\texttt{temp}_i) + \varepsilon_i$$

- We only observe a grouped version of S_i:

Thurstonian motivation of the cumulative link model



- *Latent* bitterness follows a linear model:

$$S_i = \alpha + \boldsymbol{x}_i^T \boldsymbol{\beta} + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2)$$
$$= \alpha + \beta(\texttt{temp}_i) + \varepsilon_i$$

- We only observe a grouped version of S_i:
- $\theta_{j-1} \leq S_i < \theta_j \rightarrow Y = j$

$$P(Y_i \leq j) = \Phi(\theta_j - \boldsymbol{x}_i^T \boldsymbol{\beta})$$

A cumulative link model for the wine data

Additive effects for temperature and contact:

$$P(Y_i \leq j) = \Phi\left(\theta_j - \beta_1(\texttt{temp}_i) - \beta_2(\texttt{contact}_i)\right)$$

- Is there an interaction between `temp` and `contact`?

Table: ANODE table for the wine data.

Source	df	deviance	p value
Total	12	39.407	< 0.001
Treatment	3	34.606	< 0.001
Temperature, T	1	26.928	< 0.001
Contact, C	1	11.043	< 0.001
Interaction, $T \times C$	1	0.1514	0.6972
Residual	9	4.8012	0.8513

Allowing for differences between judges

Research questions:

- Are judges rating the wines differently?
- Are there differences between bottles?

Additive random effects for judges:

$$P(Y_i \leq j) = \Phi\left(\theta_j - \beta_1(\texttt{temp}_i) - \beta_2(\texttt{contact}_i) - u(\texttt{judge}_i)\right)$$

$$u(\texttt{judge}_i) \sim N(0, \sigma_u^2)$$

Additive random effects for judges and bottles:

$$P(Y_i \leq j) = \Phi\left(\theta_j - \beta_1(\texttt{temp}_i) - \beta_2(\texttt{contact}_i) - u(\texttt{judge}_i) - b(\texttt{bottle}_i)\right)$$

$$u(\texttt{judge}_i) \sim N(0, \sigma_u^2) \quad b(\texttt{bottle}_i) \sim N(0, \sigma_b^2)$$

ANODE for mixed effects CLM

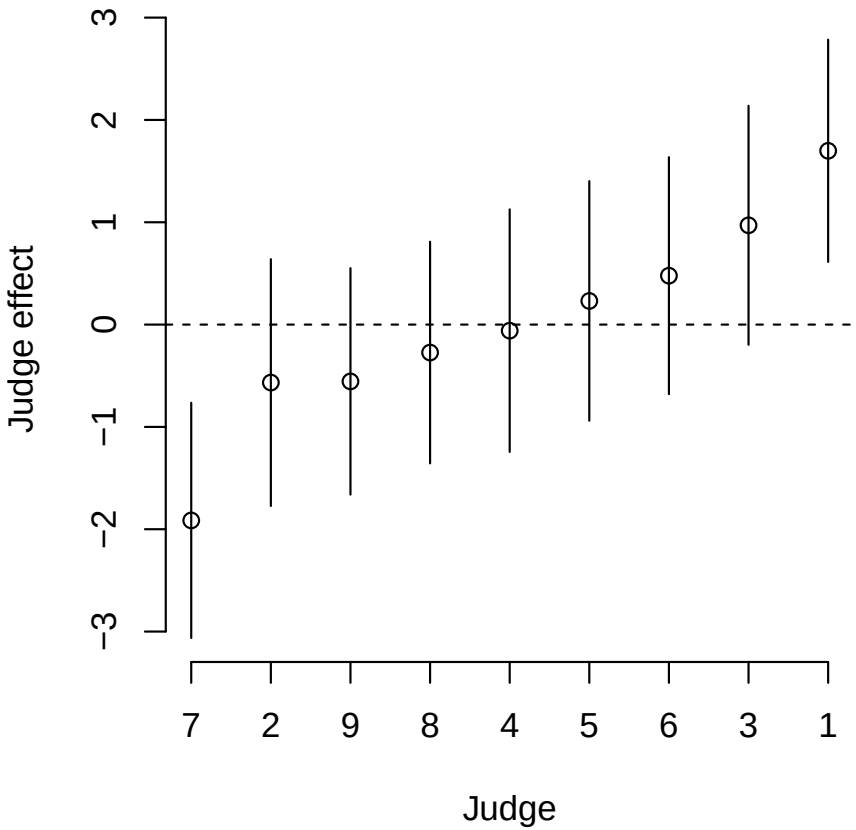
Table: ANODE table for the wine data with random effects.

Source	df	deviance	p value
Total	14	45.577	< 0.001
Var(Judge)	1	9.661	< 0.001
Var(Bottle)	1	0.001	0.998
Treatment	3	34.606	< 0.001
Temperature, T	1	25.384	< 0.001
Contact, C	1	14.238	< 0.001
Interaction, $T \times C$	1	0.1086	0.7417

Results:

- Bottles are probably not that different
- Judges do rate the wines differently

Panel inference — judge effects



Outline

Example: Preference for two commercial yoghurts

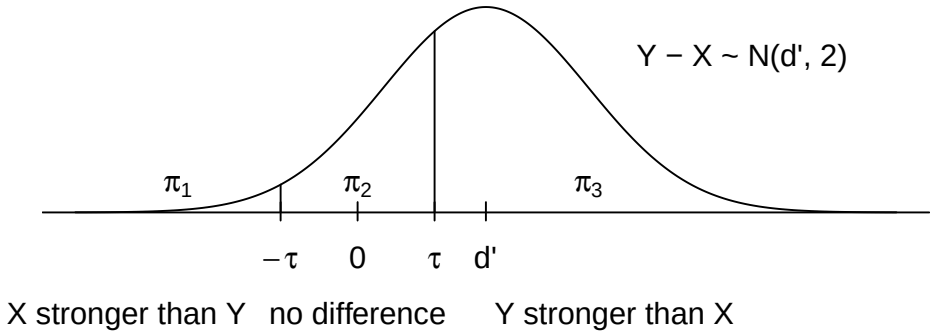
Table: 208 consumers with 4 replications (Christensen et al., 2012)

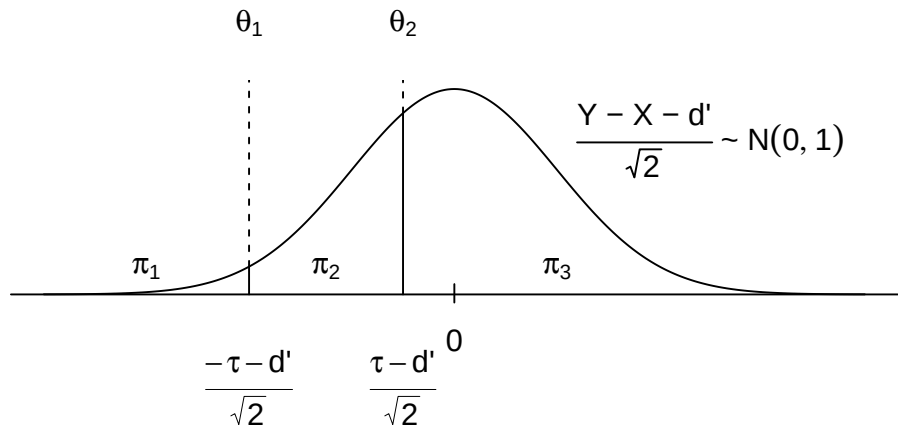
	Preference		
Condition	"prefer A"	"no-preference"	"Prefer B"
A	260	37	119
B	217	38	161

Research questions:

- Does the reference sample (A or B) in preceding duo-trio test affect preference?
- Are consumers differing in preference?

Thurstonian model for the 2-AC protocol





The Thurstonian model for the 2-AC protocol can be formulated as a cumulative link model:

$$\hat{\tau} = (\hat{\theta}_2 - \hat{\theta}_1)/\sqrt{2}$$

$$\hat{\delta} = (-\hat{\theta}_2 - \hat{\theta}_1)/\sqrt{2}$$

$$\text{se}(\hat{\tau}) = \sqrt{\{\text{var}(\theta_2) + \text{var}(\theta_1) - 2\text{cov}(\theta_2, \theta_1)\}/2}$$

$$\text{se}(\hat{\delta}) = \sqrt{\{\text{var}(\theta_2) + \text{var}(\theta_1) + 2\text{cov}(\theta_2, \theta_1)\}/2}$$

Parameter estimates

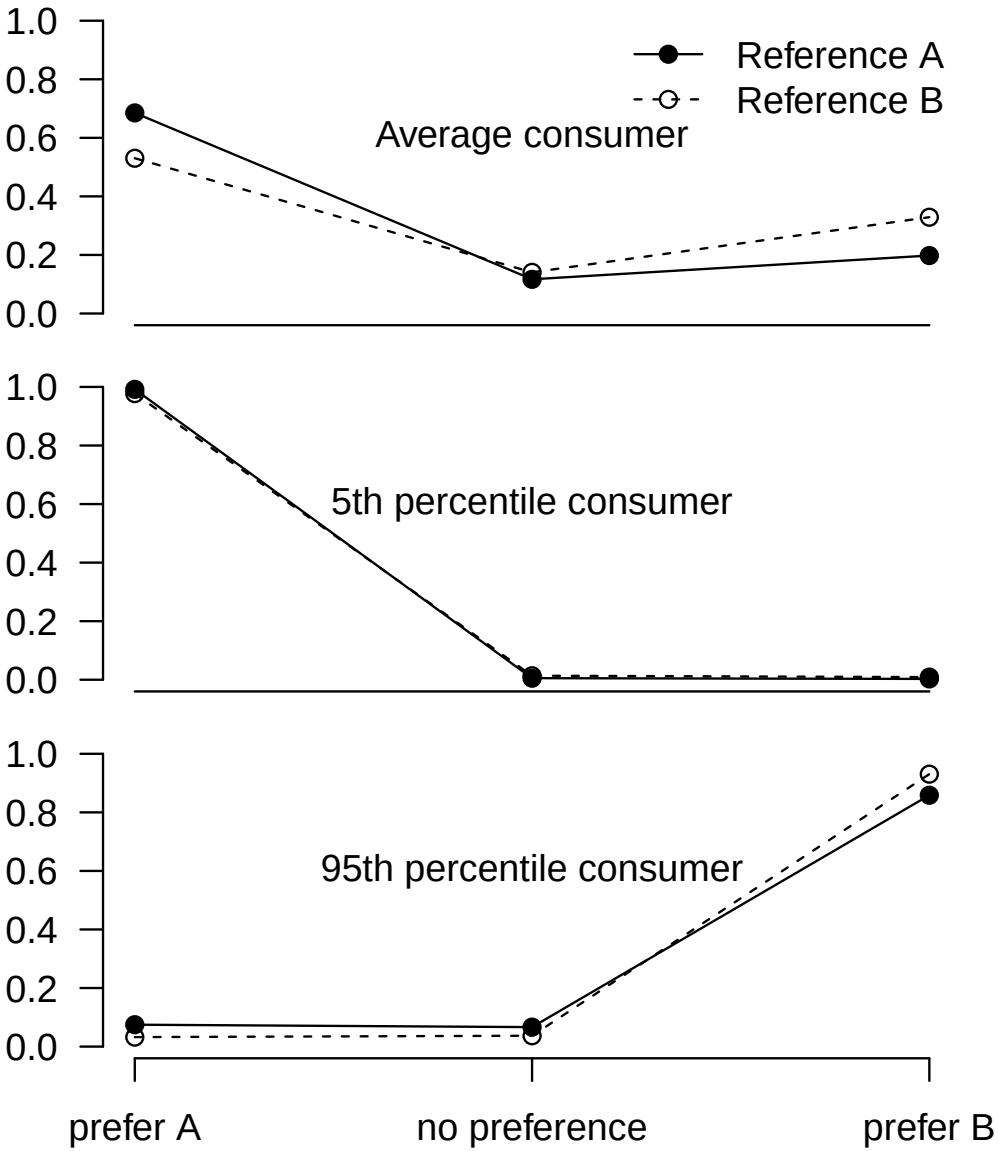
Table: Summary of a cumulative link mixed model fit to the yoghurt preference data

	Estimate	Std. Error	Lower	Upper	p-value
τ	0.259	0.029	0.202	0.316	
$d'_{\text{ref A}}$	-0.941	0.160	-1.254	-0.628	< 0.0001
$d'_{\text{ref B}}$	-0.367	0.154	-0.668	-0.066	0.0168
$\sigma_{d'}$	1.654		1.362	2.001	< 0.0001
log-likelihood	-668.9				

Identification as a well-known statistical model gave us:

- Easy tests and inference for important hypotheses (regression tools)
- Easy adjustment for replications
- Inference for the consumer population

Illustrating the model



- 95% of population within $\pm 1.96\sigma_{d'} = \pm 3.3$ (d' units)
- The largest effect is consumer differences: $\chi^2_1 = 153.6$, $p < 0.001$.
- Effect of reference in duo-trio test only for consumers with an average preference

Outline

Conclusions

- Many Thurstonian models for sensory discrimination protocols can be identified as well-known statistical models
- This facilitates modelling of for example:
  - Demographic differences between consumers
  - Effects of the experimental design
- Random effects for replications makes it possible to
  - Quantify population heterogeneity
  - Assess subject-specific performance
- Statistical results (e.g. asymptotic properties) for free
- Free software, **R** for estimation and CIs

Ongoing work and future challenges

Ongoing work:

- Derive regression framework for other protocols, e.g. same-different and degree-of-difference
- Compare different approaches to model replications
- Extend mixed-effects models to other protocols.

Open questions:

- How should we make similarity tests in replicated situations?
- How important is the normal assumption in conventional models for sensory profilling? $\rightarrow$ Comparison with cumulative link models

References

Bi, J., D. M. Ennis, and M. O'Mahony (1997). How to estimate and use the variance of d' from difference tests. *Journal of Sensory Studies 12*, 87–104.

Brockhoff, P. B. and R. H. B. Christensen (2010). Thurstonian models for sensory discrimination tests as generalized linear models. *Food Quality and Preference 21*, 330–338.

Byers, A. J. and D. Abrams (1953). A comparison of the triangular and two-sample taste test methods. *Food Technology 7*(185).

Christensen, R. H. B. (2011). ordinal—regression models for ordinal data. R package version 2010.12-15 http://www.cran.r-project.org/package=ordinal/.

Christensen, R. H. B. and P. B. Brockhoff (2011). sensR: An R-package for Thurstonian modelling of discrete sensory data. R package version 1.2.13 http://www.cran.r-project.org/package=sensR/.

Christensen, R. H. B., G. Cleaver, and P. B. Brockhoff (2011). Statistical and Thurstonian models for the A-not A protocol with and without sureness. *Food Quality and Preference 22*, 542–549.

Christensen, R. H. B., H.-S. Lee, and P. B. Brockhoff (2012). Estimation of the Thurstonian model for the 2-AC protocol. *Food Quality and Preference 4*, 119–128.

DeCarlo, L. T. (1998). Signal Detection Theory and Generalized Linear Models. *Psychological Methods 3*(2), 185–205.

Dorfman, D. D. and E. Alf (1969). Maximum-likelihood estimation of parameters of signal-detection theory and determination of confidence intervals – rating-method data. *Journal of Mathematical Psychology 6*, 487–496.

Frijters, J. E. R. (1979). The paradox of the discriminatory nondiscriminators resolved. *Chemical Senses and Flavour 4*(355).

McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society, Series B 42*, 109–142.

McCullagh, P. and J. Nelder (1989). *Generalized Linear Models* (Second ed.). Chapman & Hall/CRC.

Randall, J. (1989). The analysis of sensory data by generalised linear model. *Biometrical journal 7*, 781–793.

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review 34*, 273–286.