Analyzing Product and Individual Differences in Sensory Discrimination Testing by Thurstonian and Statistical models

Linander, Christine Borgen

Publication date: 2018

Document Version
Publisher's PDF, also known as Version of record

Link back to DTU Orbit

Citation (APA):
Analyzing Product and Individual Differences in Sensory Discrimination Testing by Thurstonian and Statistical models

Christine Borgen Linander

Kongens Lyngby 2018
PhD-2018-480
Summary

Sensory discrimination tests are used to gain information about products by using the human senses to evaluate the samples. More specifically, a sensory discrimination study is conducted when the aim is to investigate whether products are perceptibly different. Such studies are often considered for food, beverages as well as personal care products. An example is when a company gets a new supplier of an ingredient in one of their products. It is of high importance to investigate how this change of the ingredient affects the product. Even though the chemical composition of the product changes, it does not necessarily mean that people can detect the difference. These days, people become more and more interested in how to improve their health. This is also reflected in the companies’ desire to make their products healthier without changing how the product is perceived by their customers. Therefore, it is important to conduct sensory discrimination tests when ingredients are changed. This thesis is concerned with the analysis of product and individual differences in sensory discrimination testing.

Sensory discrimination tests become more and more advanced raising a need for new types of analysis of sensory discrimination data. This thesis contributes with the development of Thurstonian models and how these can be aligned with well-known statistical models. Generalized linear mixed models are used in many applications. However, it is not common to consider such complicated models when considering sensory discrimination tests. Actually, sensory discrimination tests are often analyzed by too simplistic methods, ignoring important variables, such as individuals, that affect the results of the analysis. One focus of this project is to propose a way to incorporate such effects in the models when analyzing data from sensory discrimination studies. These mod-
els, including random effects, are called Thurstonian mixed models. Considering generalized linear mixed models for sensory discrimination studies opens up for many possibilities. It becomes possible to gain information about the individuals, the so-called assessors, as well as making more proper conclusions regarding the products. Moreover, the estimates of product and individual differences are obtained on the d-prime scale.

Often multiple sensory attributes are considered in a discrimination study. These can be analyzed individually by the Thurstonian mixed models we are introducing. This thesis is presenting a multivariate analysis to gain knowledge about the product and individual differences across the sensory attributes. This is achieved by analyzing the product and individual differences, on the d-prime scale, by principal component analysis.

Sensory discrimination tests are sometimes conducted to investigate the performance of sensory panels or to compare different laboratories. In such tests, multiple d-prime values can be obtained. For sensory discrimination tests, which lead to binomially distributed responses, we propose a new test statistic for the comparison of multiple d-prime values. The test statistic we suggest is an improved way of analyzing multiple d-prime values compared to a previous suggested test statistic.
Resumé


Sensoriske diskriminationstests bliver mere og mere avancerede hvilket øger behovet for nye typer af analyser af data fra sensoriske diskriminationstests. Denne afhandling bidrager med udviklingen af Thurstonske modeller og hvordan disse kan kombineres med velkendte statistiske modeller. Generaliserede lineære mixtede modeller bliver brugt i mange anvendelser. Imidlertid er det ikke almindeligt at betragte sådanne komplicerede modeller når data fra sensoriske diskriminationstests betragtes. Faktisk bliver data fra sensoriske diskriminationstests ofte analyseret med for simple modeller som ignorerer vigtige variable, som individer, hvilket påvirker resultaterne af analysen. Et fokus for dette projekt er at foreslå en måde at indkorporere sådanne effekter i modellen når data fra sensoriske diskriminationsstudier bliver analyseret. Disse modeller, som medtager

Ofte bliver mange sensoriske egenskaber betragtet i et diskriminationsstudie. Disse kan analyseres enkeltvis ved brug af de Thurstonske mixede modeller vi introducerer. Denne afhandling præsenterer en multivariat analyse for at få viden om produktforskelle samt individforskelle på tværs af de sensoriske egenskaber. Dette opnås ved at analysere produktforskelle og individforskelle, på ’d-prime’ skalaen, ved ’principal component analysis’.

Sensoriske diskriminationstests bliver indimellem udført for at undersøge præstationen af sensoriske paneler eller for at sammenligne forskellige laboratorier. I sådanne tests er det muligt at få mange ’d-prime’ værdier. For sensoriske diskriminationstests, som giver binomialfordelte responsvariable, foreslår vi en ny teststørrelse til at sammenligne adskillelige ’d-prime’ værdier. Teststørrelsen vi foreslår er en forbedret måde at analysere mange ’d-prime’ værdier på sammenlignet med en tidligere foreslået teststørrelse.
This thesis was prepared at Technical University of Denmark, Department of Applied Mathematics and Computer Science, Statistics and Data Analysis section, in partial fulfillment of the requirements for acquiring the Ph.D. degree in Applied Statistics. The project was funded by the Technical University of Denmark and Unilever U.K. Central Resources Limited. The project was supervised by Professor Per Bruun Brockhoff. Occasionally, Rune Haubo Bojesen Christensen has been a co-supervisor.

The thesis deals with the analysis of product and individual differences in sensory discrimination testing. Sensory discrimination testing is a type of testing used in sensory science, where people are used as the measurement instruments. The main focus is developing methods aligning Thurstonian methods with statistical models.

The thesis consists of three research papers and a book chapter. An introductory part gives an overview of the thesis. Background and aspects that were not considered in the papers are considered in the thesis.

Lyngby, 01-July-2018
List of contributions

This thesis is based on the following research papers and a book chapter


The following talk was presenting some of the work included in this thesis

The following posters have been presenting some of the contents of the work

1. The Comparison And Analysis Of Multiple d-primes From Varying Test Protocols. *The 9th Pangborn Conference, Toronto, Canada, September 4-8 2011*

Acknowledgments

Foremost, I would like to thank my supervisor Per Bruun Brockhoff for the numerous conversations we have had during these years. You have always been very enthusiastic and committed. Additionally, I would like to thank Per for his continued belief in me, throughout the many challenges I have encountered during my time as a Ph.D. student. This has given me hope and motivation to continue during the tough times.

A huge thanks to my co-supervisor Rune Haubo Bojesen Christensen. You have always brought new and valuable perspective on the matters. The many discussions we have had have been very rewarding leading to many new insights for me.

I would also like to thank Unilever for making this Ph.D. possible. I would especially like to thank Graham Cleaver for sharing his expertise regarding sensory science as well as sensometrics. It has always been a pleasure to work with Graham, and I have learned much from the many discussions we have had. Also a special thanks to Rebecca (Becky) Evans for always making me feel welcome visiting Unilever. Furthermore, I would like to thank Becky for the many chats we have had regarding sensory science.

Thanks to all of my fellow Ph.D.-students and colleagues for creating a nice working atmosphere. A special thanks goes to Federica Belmonte for the many coffee dates we have had. We have shared many frustrations and joys.
I would also like to thank my family and friends for bearing with me during all of the ups and downs throughout the making of this Ph.D. You have all been amazing.

Furthermore, I would like to thank my daughter Freya for all her kisses and smiles. You have been my sunlight on a raining day.

Lastly, and above all, I would like to thank my loving and caring husband Toke. Thank you for all your tremendous love and support. Your encouragement has meant everything to me - I could not have finished this Ph.D. without you!
Contents

Summary i

Resumé iii

Preface v

List of contributions vii

Acknowledgments ix

1 Introduction 1
   1.1 Overview of the Thesis 3

2 Sensory Discrimination Testing 5
   2.1 Sensory discrimination tests 5
   2.2 simple-binomial test protocols 6
   2.3 Analyzing sensory discrimination studies with simple-binomial tests 7
   2.4 Thurstonian modelling 8
   2.5 Analyzing sensory discrimination studies using sensR 10

3 Thurstonian Mixed Models 13
   3.1 Generalized Linear Mixed Models 13
   3.2 The binomial distribution 14
   3.3 Thurstonian framework 15
   3.4 Thurstonian Models 17
   3.5 Simplification of Thurstonian models 20
   3.6 Estimation of parameters 21
   3.7 Estimation of d-prime values 22
This thesis deals with Thurstonian and statistical models of data obtained from sensory discrimination studies. Such studies are used to gain knowledge about products by using the human senses to evaluate the samples. More specifically, a sensory discrimination study is conducted when the aim is to investigate whether products are perceptibly different. Thus, sensory discrimination tests are conducted for products that are very similar with only subtle differences. Such studies are often considered for food, beverages as well as personal care products. Examples of the use of sensory discrimination tests are when a company gets a new supplier of an ingredient used in one of their products or if an ingredient is changed due to new health initiatives in a company. It is of high importance to investigate how this change of ingredient(s) affects the product. Even though the chemical composition of the product changes, it does not necessarily mean that people can detect the difference. However, it would be catastrophic for a company if a new product was put in the market replacing an old product in the believe that these two products are perceived to be identical when in fact they are not. Therefore, it is important to conduct sensory discrimination tests when ingredients are changed.

Sensory discrimination tests become more and more advanced, raising a need for new types of analysis of sensory discrimination data. One aim of the project is to align Thurstonian modelling with modern statistical models by considering how
the individuals can be accounted for. For normally distributed data, individuals are typically accounted for by including them as a random effect in the so-called linear mixed models. However, when considering sensory discrimination tests, the data are not normally distributed. For a group of sensory discrimination tests, the so-called simple-binomial tests, data are binomially distributed. For non-normal data the equivalence of the linear mixed models are the generalized linear mixed models. These are used in many applications. However, it is not common to consider such complicated models when considering sensory discrimination tests. Actually, sensory discrimination tests are often analyzed by too simplistic methods, ignoring important variables, such as individuals, that affect the results of the analysis, potentially leading to improper conclusions regarding the products. In this thesis, generalized linear mixed models are considered for the binary paired comparison where potential individual differences are modelled. These models, including random effects, are called Thurstonian mixed models. Considering generalized linear mixed models for sensory discrimination studies opens up for many possibilities. It becomes possible to gain information about the individuals, the so-called assessors, as well as making more proper conclusions regarding the products. Moreover, the estimates of product and individual differences are obtained on the d-prime scale.

A d-prime value is the estimate of an underlying sensory difference between two products. This is dating back to [Thurstone (1927)]. It is ongoing work to align Thurstonian models with modern statistical models. It is an area of research that has been considered in many years and for several different scenarios. Considering the analysis of sensory data as generalized linear models has been looked into in several settings. It is an area of research back to at least 1989 with the paper by [Randall (1989)]. In [Critchlow and Fligner (1991)] it is considered how a paired comparison can be seen as a generalized linear model. Recent advances has been considered in [Brockhoff and Christensen (2010)] where discrimination tests are recognized as generalized linear models. Other contributions, not necessarily for discrimination testing, in the work of aligning Thurstonian models with well-known statistical models are [Christensen and Brockhoff (2009); Christensen et al. (2011); Christensen and Brockhoff (2013); Christensen et al. (2012)].

Often multiple sensory attributes are considered in a discrimination study. These can be analyzed individually by the Thurstonian mixed models we are introducing. This thesis is presenting a multivariate analysis to gain knowledge about the product and individual differences across the sensory attributes. This is achieved by analyzing the product and individual differences, on the d-prime scale, by principal component analysis.
1.1 Overview of the Thesis

Sensory discrimination tests are sometimes conducted to investigate the performance of sensory panels or to compare different laboratories. Generally, sensory discrimination tests become more and more advanced. Examples of such sensory discrimination studies are: a study comparing different tests (Dessirier and O’Mahony (1999)); and comparing the same test in different locations (Sauvageot et al. (2012)); and comparison of different versions of the same test (Lee and Kim (2008)). In such tests, multiple d-prime values can be obtained. For sensory discrimination tests, which lead to binomially distributed responses, we propose a new test statistic for the comparison of multiple d-prime values. The test statistic we suggest is an improved way of analyzing multiple d-prime values compared to a previous suggested test statistic.

1.1 Overview of the Thesis

This thesis consists of seven chapters followed by three research papers as well as one book chapter included in the appendices. Throughout the thesis, as well as in the papers in Appendix A and B, an existing discrimination study is used as an example. This discrimination study is explained in Linander et al. (2018a) (Appendix A).

1.1.1 Main Chapters

This thesis consists of seven chapters, each introducing or addressing relevant topics for the papers and the book chapter presented in the appendices in the thesis.

Chapter 2 gives an introduction to sensory discrimination testing. The primary focus is to provide the reader unfamiliar with sensory discrimination tests knowledge about aspects of sensory discrimination testing that are used in the thesis.

In Chapter 3, an introduction to the relevant aspects of generalized linear models as well as generalized linear mixed model is provided. The aim of this chapter is to get the non-statistical reader more familiar with relevant aspects to better understand the contents of Linander et al. (2018a). Furthermore, aspects of a Thurstonian mixed model is considered in more details than the scope of Linander et al. (2018a).
Chapter 4 investigates how trustworthy the Thurstonian mixed models introduced in Chapter 3 and Linander et al. (2018a) are.

In Chapter 5 different aspects of principal component analysis using d-prime values are investigated, which are only briefly investigated in Linander et al. (2018b).

Chapter 6 is analyzing data which are obtained as an additional question in the binary paired comparison.

Chapter 7 is introducing and investigating aspects of comparison of multiple d-prime values which are considered in Linander et al. (2018).

1.1.2 Journal Papers

Three journal papers are included in the appendices. Furthermore, a book chapter is also included in the appendices.

The first paper included in Appendix A is written for Food, Quality and Preference. The paper is regarding the analysis of data obtained from a sensory discrimination study using the binary paired comparison protocol.

The paper included in Appendix B is written for Food, Quality and Preference. The paper is considering Principal Component Analysis of d-prime values.

The paper included in Appendix C is written for Journal of Sensory Studies. The paper is considering a way to compare and analyze multiple d-prime values obtained from potential different sensory discrimination test protocols.

The book chapter in Appendix D is a part of the book Rogers (2017). The book chapter is giving a thorough introduction to how sensR can be used to analyze data from sensory discrimination tests.
Sensory discrimination testing is a type of test used in sensory science. It is used when the aim is to investigate whether products are perceptibly different (Lawless and Heymann 2010). An example of such a situation is if a company gets a new supplier of an ingredient and it is desired to investigate if this affects how the product is perceived. Another example is when a company makes a change of an ingredient due to e.g. health initiatives and would like information about the influence this change has for the product.

Sensory discrimination testing is a topic of ongoing research and many papers and books exist introducing this topic (Rogers 2017; Bi 2006; Lawless and Heymann 2010; Ennis et al. 2014).

2.1 Sensory discrimination tests

Sensory discrimination tests are different in many ways. One difference is regarding the cognitive task associated with the test. Overall, two strategies
associated with the tests are considered: the skimming strategy and the comparison of difference. Put differently, two different types of discrimination tests exist: with or without a nature of the difference between the samples. Tests with a nature of the difference are e.g. n-AFC methods where the question is of the type "Which sample is the most bitter?", and the A-not A method where the question is "Is the sample A or not-A?". This type of test is using the skimming strategy. Tests without a nature of the difference are methods comparing the distance of differences e.g. the Duo-trio method where the question asked is "Which sample is the same as the control sample?" and the Same-Diff with "Are the samples the same or different?" see e.g. Bi (2006) and O'Mahony et al. (1994).

Another difference is whether the test has response bias or not. For tests with a response bias psychological factors will influence each assessor’s understanding of the choices e.g. whether two samples are "same" or "different" in a same-diff test. For tests with response bias it is inappropriate with a guessing probability (see e.g. Rogers (2017)). For tests without response bias the guessing probability, denoted by $p_g$, is the probability of choosing the correct sample by chance. This depends on the number of possible answers in the test and is given as:

$$p_g = \frac{\# \text{ correct combinations}}{\# \text{ combinations}} \quad (2.1)$$

### 2.2 simple-binomial test protocols

In this thesis the focus is on the so-called simple-binomial tests, which as the name implies lead to binary data. These consist of the 2-Alternative Forced Choice (AFC) method, the 3-AFC, triangle, duo-trio and tetrad. The latter has become very popular in recent advances (Ennis, 2012; Rogers, 2017).

In the 2-AFC test an assessor is comparing two samples, one of each product, and is asked to choose the sample with the strongest (or weakest) sensory intensity of the sensory attribute in question e.g. sweetness or saltiness. The 3-AFC test is equivalent to the 2-AFC test except for the number of samples being compared. In the 3-AFC an assessor is comparing three samples where two are from one product and the third is from another. The 2-AFC and the 3-AFC tests are specified tests since the difference is known.

The triangle test involves three samples, two from one product and one from
another product, for which an assessor is asked to identify the odd sample. The triangle test is an unspecified test since the difference between the products is unknown.

For the duo-trio test an assessor is receiving three samples where one is marked as the reference. Of the two remaining samples one is from the same product as the reference product and one is from another product. The assessor must choose the sample, of the remaining two samples, that is the same as the reference sample. As for the triangle test, the difference in a duo-trio test is unspecified.

The tetrad method is using four samples, where two samples are from one product and two are from another product. The assessor’s task is to group the samples such that the two samples from the same product are grouped as one group. The tetrad test is also an unspecified test since no information about the differences of the products is given.

2.3 Analyzing sensory discrimination studies with simple-binomial tests

Let $X_1, \ldots, X_n$ be observations from $n$ independent simple-binomial tests. Then the total number of correct answers is binomially distributed:

$$X = \sum_{i=1}^{n} X_i \sim \text{Binomial}(p_c, n)$$  \hspace{1cm} (2.2)

since $X_i \sim \text{Binomial}(p_c, 1)$ for all $i$. Furthermore, $p_c = P(X_i = 1)$ is the probability of a correct answer.

When analyzing such data, three different levels of analysis exist (Næs et al., 2010). The three levels are using the proportion of correct answers, the proportion of discriminators or the underlying sensory difference $\delta$. The proportion of correct answers is the proportion of times the correct sample was chosen. The proportion of discriminators is the proportion of individuals that would detect the product difference. The underlying sensory difference is explained in Section 2.4.
There exist a unique relation between the proportion of correct answers and the proportion of discriminators given by:

\[ p_c = p_g + p_d(1 - p_g) \]

where \( p_d \) is the proportion of discriminators (Næs et al., 2010). Furthermore, there exist a function relating the proportion of correct answers with the underlying sensory difference, the so-called *psychometric* function. Thus, it is possible to transform between the three levels of analysis.

The expected proportion of correct answers and discriminators depend on the test used (Næs et al., 2010; Rogers, 2017). Thus, it is not possible to compare the proportion of correct answers or the discriminators for different tests. For the same sensory difference between products, it is expected that different tests lead to different proportions of correct answers and therefore also different proportion of discriminators. The reason for the discrepancy between the proportions is due to the fact that the cognitive task of some tests is more difficult than for other tests.

Gridgeman’s paradox was introduced in Gridgeman (1970). It is known for showing that the triangle test and the 2-AFC test of the same stimulus lead to assessors that answered wrongly in the triangle test but correctly in the 2-AFC test (Frijters, 1979). Thus, it appeared that nondiscriminators were able to discriminate. It was shown in Frijters (1979) that this discrepancy was caused by the use of the proportion of discriminators as the measure for the difference between the products. Using \( \delta \) as the measure of the sensory difference between the products, the triangle and the 2-AFC lead to the same estimate of \( \delta \) and the Gridgeman’s paradox was resolved. Several authors have pointed out that the proportion of discriminators is a bad measure of product differences and \( \delta \) is to be preferred (Frijters, 1979; Ennis, 1993; Ennis and Jesionka, 2011; Jesionka et al., 2014).

2.4 Thurstonian modelling

Thurstonian modelling is used as a way to quantify the sensory differences between products. It is based on that differences between samples of a product can occur as well as differences in how the samples are perceived by the human exist. Thus, Thurstonian modelling is based on the fact that the sensory intensity of a product not will be constant but will vary. It is assumed that such sensory intensities are normally distributed with equal variances (see e.g. O’Mahony and
The Thurstonian measure of underlying sensory differences is denoted by $\delta$ and it is defined as the difference between the means divided by the standard deviation. The bigger $\delta$ is the more distinguishable the products are.

### 2.4.1 Psychometric functions

The psychometric function is the function relating the proportion of correct answers to $\delta$. In this section the psychometric functions for the simple-binomial tests are given.

The psychometric functions for the 2-AFC, 3-AFC, triangle and duo-trio have been written and defined many times in the literature; see Brockhoff and Christensen (2010) and Ennis (1993) as well as references therein. The psychometric function for the unspecified tetrad has also been introduced in the literature; Ennis et al. (1998).

The psychometric function for the 2-AFC reads:

$$f_{2\text{AFC}}(\delta) = \Phi\left(\frac{\delta}{\sqrt{2}}\right) = p_c$$  \hspace{1cm} (2.3)

The psychometric function for the 3-AFC reads:

$$f_{3\text{AFC}}(\delta) = \int_{-\infty}^{\infty} \varphi(z-\delta)\Phi^2(z)dz = p_c$$  \hspace{1cm} (2.4)

The psychometric function for the duo-trio reads:

$$f_{d-t}(\delta) = -\Phi\left(\frac{\delta}{\sqrt{2}}\right) - \Phi\left(\frac{\delta}{\sqrt{6}}\right) + 2\Phi\left(\frac{\delta}{\sqrt{2}}\right)\Phi\left(\frac{\delta}{\sqrt{6}}\right) = p_c$$  \hspace{1cm} (2.5)

The psychometric function for the triangle reads:

$$f_{\text{tri}}(\delta) = 2\int_{0}^{\infty} \left(\Phi\left(-z\sqrt{3} + \delta\sqrt{2/3}\right) + \Phi\left(-z\sqrt{3} - \delta\sqrt{2/3}\right)\right) \varphi(z)dz = p_c$$  \hspace{1cm} (2.6)

The psychometric function for the unspecified tetrad reads:

$$f_{\text{tetrad}}(\delta) = 1 - 2\int_{-\infty}^{\infty} \varphi(x)\left(2\Phi(x)\Phi(x-\delta) - (\Phi(x-\delta))^2\right)dx = p_c$$  \hspace{1cm} (2.7)
2.4.2 Estimation

The maximum likelihood estimate of the probability of a correct answer reads:

$$
\hat{p}_c = \begin{cases} 
    x/n & \text{if } x/n \geq p_g \\
    p_g & \text{if } x/n < p_g
\end{cases}
$$

(2.8)

where $\hat{p}_c$ cannot be lower than the guessing probability.

Due to the invariance property of the maximum likelihood estimate \cite{Pawitan2001} the maximum likelihood estimate of $\delta$ is given by:

$$
d' = f_{psy}(\hat{p}_c)
$$

(2.9)

When the products are not different $\hat{p}_c = p_g$ and $d' = f_{psy}(p_g) = 0$. Therefore, the valid values for the parameters are:

$$
p_c \in [p_g, 1] \quad \text{and} \quad \delta \in [0, \infty)
$$

(2.10)

2.5 Analyzing sensory discrimination studies using sensR

In this section, a brief introduction to analyzing data from sensory discrimination studies in sensR is given. For a more detailed description of the analysis using sensR see \cite{Brockhoff2017}. Besides the so-called basic simple-binomial test protocols, many other test protocols exist. The sensR package provides a way to analyze data from many of these. Furthermore, the package ordinal enables analysis of test protocols that are analyzed by the so-called cumulative link models \cite{Christensen2018}.

An overview of the functionality in the sensR package is given in Table 2.1 (which is the same as Table 15.1 in \cite{Brockhoff2017} with minor changes).

The sensR package is making many analyses and transformations available in R for sensory discrimination tests. It is possible to do difference testing as well as similarity testing. Furthermore, it is possible to estimate the three different parameters $p_c, p_d$ and $\delta$. Power and sample size calculations are also possible. Moreover, it is possible to do the replicated analysis based on the beta-binomial model as well as the corrected beta-binomial model. Additionally, it is possible to simulate replicated sensory protocol data. Furthermore, it is possible to
Table 2.1: Overview of the functionality in the **sensR** package (together with the **ordinal** package). The parentheses indicate that no specific functions for similarity testing are available, but through valid likelihood-based confidence intervals it can be done.

<table>
<thead>
<tr>
<th></th>
<th>d-prime estimation</th>
<th>difference hypothesis test</th>
<th>similarity hypothesis test</th>
<th>power</th>
<th>sample size</th>
<th>simulation</th>
<th>likelihood confidence interval</th>
<th>replicated analysis</th>
<th>regression/anova-glm</th>
<th>d-prime comparisons</th>
</tr>
</thead>
<tbody>
<tr>
<td>duo-trio</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>triangle</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>tetrad</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-AFC</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>3-AFC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Double triangle</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Double duo-trio</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Double 2-AFC</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Double 3-AFC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unspecified</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>2-out-of-5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unspecified</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>2-out-of-5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>with forgiveness</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unspecified</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Hexad test</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A-not A</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Same-Different</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-AC</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Degree of</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Difference</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(DOD)</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>A-not A</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>with sureness</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
transform values of one of the three parameters $p_c, p_d$ and $\delta$ to any of the other. In addition, it is possible to plot Thurstonian distributions.

Difference and similarity testing is carried out by the `discrim` function:

```r
 discrim(correct, total, d.prime0, pd0, conf.level = 0.95,
          method = c("duotrio", "tetrad", "threeAFC", "twoAFC",
                     "triangle", "hexad", "twofive", "twofiveF"),
          double = FALSE,
          statistic = c("exact", "likelihood", "score", "Wald"),
          test = c("difference", "similarity"), ...)
```

Transformation of $p_c, p_d$ and $\delta$ to one of the other is possible by using the `rescale` function:

```r
 rescale(pc, pd, d.prime, std.err,
         method = c("duotrio", "tetrad", "threeAFC", "twoAFC",
                     "triangle", "hexad", "twofive", "twofiveF"),
         double = FALSE)
```

The analysis of replicated data is carried out by the `betabin` function:

```r
 betabin(data, start = c(.5,.5),
         method = c("duotrio", "tetrad", "threeAFC", "twoAFC",
                     "triangle", "hexad", "twofive", "twofiveF"),
         vcov = TRUE, corrected = TRUE, gradTol = 1e-4, ...)
When data follow a normal distribution, linear models provide a broad range of analyses of these data. However, when data do not follow a normal distribution an alternative class of models, the so-called Generalized Linear Models (GLMs) and Generalized Linear Mixed Models (GLMMs), is used (McCullagh and Nelder 1989; Agresti 2013; McCulloch et al. 2008; Agresti 2015).

In this chapter these models are embedded into a Thurstonian framework, leading to the possibility to obtain the estimates of the effects on the d-prime scale. In Brockhoff and Christensen (2010) it was established that a sensory discrimination test can be considered as a generalized linear model. In this chapter, as well as in Linander et al. (2018a), the models introduced in Brockhoff and Christensen (2010) are extended such that random effects are included as the explanatory variables that are explaining the response variable.

### 3.1 Generalized Linear Mixed Models

Generalized linear (mixed) models are the equivalence of linear (mixed) models, in situations where data do not follow a normal distribution. The aim is to gain knowledge about which, if any, explanatory variables that are important for the
response variable. The Generalized Linear Model as well as the Generalized Linear Mixed Model are described in this section. The general way a GLM or a GLMM is written is by:

\[ g(\cdot) = \eta \]  

(3.1)

\( g \) is the so-called link function, and \( \eta \) is the so-called linear predictor. The linear predictor is expressing which explanatory variables that are affecting the mean value of the response variable. When only fixed effects are included in \( \eta \) the model is a generalized linear model. When at least one random effect is included in \( \eta \) the model is a generalized linear mixed model. The exact way the model in (3.1) is defined depends on whether random effects are included or not. When no random effects are included the model is a generalized linear model in which the link function \( g \) is linking how the mean value of the response variable is related to the explanatory variables. Hence the name, link function.

Let \( Y = (Y_1, \ldots, Y_I)^T \) be the response variable. A generalized linear model is of the form:

\[ g(\mathbb{E}(Y)) = X\beta \]  

(3.2)

where \( X \) is the \( I \) by \( J \) design-matrix containing the \( J \) explanatory variables for the \( I \) observations. \( X \) can include main effects as well as interactions that are of interest. Furthermore, \( \beta = (\beta_1, \ldots, \beta_J)^T \) is the vector of the parameters.

When random effects are included, the specification of model (3.1) gets more complicated than in (3.2). When random effects are included the link function \( g \) is linking the conditional mean of the response variable given the random effects to the explanatory variables:

\[ g(\mathbb{E}(Y|u)) = X\beta + Zu \]  

(3.3)

where \( u \sim N(0, \Sigma) \) and \( Z \) is the design matrix for the random effects. Throughout this thesis, the distribution for the response variable is the binary distribution. Thus, the models considered in this chapter will be for a binary response.

### 3.2 The binomial distribution

The binomial distribution is used in a wide range of applications. Sensory discrimination testing is one area where the binomial distribution plays an important role.
Let $Y_1, \ldots, Y_n$ be independent binomially distributed random variables with probability parameter $p = P(Y_i = 1)$ for $i = 1, \ldots, n$ and count parameter 1. Now:

$$Y = \sum_{i=1}^{n} Y_i$$ (3.4)

is binomially distributed with parameters $n$ and $p$:

$$Y \sim \text{binomial}(p, n)$$ (3.5)

The probability mass function for the binomial distribution reads:

$$p(y) = \binom{n}{y} p^y (1-p)^{n-y}$$ (3.6)

A requirement for generalized linear models is that the distribution must be a member of the natural exponential family. It can be realized that the binomial distribution is in fact a member of the natural exponential family by:

$$p(y) = \binom{n}{y} p^y (1-p)^{n-y} = \binom{n}{y} e^{y \log(p)} (1-p)^{n-y} = \binom{n}{y} e^{y \log(p) + (n-y) \log(1-p)}$$

see e.g. Jørgensen (1997) for further details.

### 3.3 Thurstonian framework

The Thurstonian way of modelling, which was introduced in Section 2.4, is a way to quantify sensory differences. The sensory difference, $\delta$, is defined as the difference in means, relatively to the standard deviation, for the normal distributions explaining the sensory intensity. There is a unique relation between $\delta$ and the probability of a product being chosen. This relation is the so-called psychometric function.
This type of model will be considering data obtained from binary paired comparisons, which can be considered as unbounded 2-AFC tests. Unbounded in the sense that no correct answer exists and therefore the probability of getting a correct answer is considered as the probability of choosing one product. In this thesis, as well as the papers in the appendices, we refer to the products as a control product and a test product. Thus, the probability of choosing the test product does not have the restriction introduced in Section 2.4.2.

When developing the Thurstonian framework, assumptions are made regarding the sensory intensities of the test products as well as the control product. As in Linander et al. (2018a) let

\[ C \sim N(\mu_c, \sigma^2) \quad \text{and} \quad T \sim N(\mu_t, \sigma^2) \]  

be the sensory intensities for the control and a test product respectively. Moreover, the Thurstonian underlying relative difference is defined as

\[ \delta = \frac{\mu_t - \mu_c}{\sigma} \]  

The psychometric function, denoted by \( f_{\text{pair}} \), can for this setting be defined as the probability that the test product is chosen. This is the probability that the test product have a larger sensory intensity than the control:

\[ f_{\text{pair}}(\delta) = P(T > C) = \Phi\left( \frac{\delta}{\sqrt{2}} \right) = p \]  

where \( \Phi \) is the cumulative distribution function for the standard normal distribution and \( p \) is the probability that the test product is chosen over the control product. The derivation of (3.9) is shown in Linander et al. (2018a).

The Thurstonian framework is defined for one test product at a time. Thus, underlying sensory differences are considered; \( \delta_1, \ldots, \delta_I \), where \( I \) is the number of test products considered.

When allowing for differences for the assessors the assumptions regarding the sensory intensities are that each assessor has its own sets. More specifically, let

\[ C_j \sim N(\mu_{c_j}, \sigma^2) \quad \text{and} \quad T_j \sim N(\mu_{t_j}, \sigma^2) \]  

be the sensory intensities for the control and a test product respectively for the \( j \)-th assessor, where \( j = 1, \ldots, J \). This can be interpreted in one of two ways regarding the distribution for the sensory intensity of the control. The assessors have the same distribution for the control, meaning that \( C_j \sim N(\mu_c, \sigma^2) \) for all \( j \), or that the distribution of the sensory intensity of the control product is different.
for the different assessors. In both situations, the sensory intensity for the test product is different for the assessors. The Thurstonian framework is defined for one test product at a time for each assessor. Thus, $IJ$ underlying sensory differences are considered; $\delta_{11}, \ldots, \delta_{1J}, \delta_{21}, \ldots, \delta_{IJ}$, where $I$ is the number of test products considered and $J$ is the number of assessors. Furthermore, $\delta_{ij}$ is the sensory difference between the $i$th test product and the control for the $j$th assessor.

### 3.4 Thurstonian Models

The expected value of a binomially distributed variable equals the probability times the count parameter. Thus, for $Y \sim \text{binomial}(p, n)$ the expected value reads:

$$E[Y] = np \quad (3.11)$$

with the special case of a Bernoulli distributed variable with $n = 1$:

$$E[Y] = p \quad (3.12)$$

Thus, when considering the generalized linear models for the binomially distributed data, the probabilities will be used in the model.

When considering the probability in the binary distribution it is possible to impose a linear structure on it by modelling it in a generalized linear model. This is what is considered in [Brockhoff and Christensen (2010)](https://doi.org/10.1007/978-3-642-25565-1_3).

When the probability of choosing a test product is modelled such that it is affected by test products, the model reads:

$$g(p_i) = \mu + \alpha_i = \eta_i \quad (3.13)$$

where $i = 1, \ldots, I$ represents test products, $\mu$ is the overall average difference between test products and the control and $\alpha_i$ is the difference for the $i$th test product to the average product-difference $\mu$. $p_i = P(Y_{ijk} = 1)$ is the probability that the $i$th test product is chosen for the $j$th assessor in the $k$th session. Furthermore,

$$Y_{ijk} \sim \text{Binomial}(p_{ij}, 1) \quad (3.14)$$
Thus, the probability mass function for $Y_{ijk}$ is given by:

\[
p(y_{ijk}) = \binom{1}{y_{ijk}} p^{y_{ijk}} (1 - p)^{1 - y_{ijk}}
\]

\[
= p^{y_{ijk}} (1 - p)^{1 - y_{ijk}}
\]

It is possible to add random effects to the linear predictor in the right-hand side of equation (3.13). When the probability of choosing a test product is modelled such that it is affected by the main effects of assessors as well as test products, the model reads:

\[
g(p_{ij}) = \mu + \alpha_i + b_j = \eta_{ij}
\]  

(3.15)

where $j = 1, \ldots, J$ represents the assessors and $b_j$ is the random effect of the $j$th assessor. The random effects are assumed to be independent and identically distributed:

\[
B_j \sim N(0, \sigma^2_b)
\]  

(3.16)

where $b_j$ is a realization of the random variable $B_j$. Furthermore, the remaining parameters are given as for (3.13). When a random effect is added to the linear predictor the probability is no longer the unconditional probability. Therefore, the probability in (3.15) is the conditional probability given the random effects:

\[
p_{ij} = P(Y_{ijk} = 1 | B_j = b_j)
\]

Furthermore, the binary data are binomially distributed conditional on the realized values of the random variable $B_j = b_j$:

\[
Y_{ijk} | B_j = b_j \sim \text{Binomial}(p_{ij}, 1)
\]  

(3.17)

where

\[
p_{ij} = f_{\text{pair}}(\mu + \alpha_i + b_j)
\]

The probability mass function for the conditional distribution given in (3.17) reads:

\[
p(y_{ijk}|b_j) = \binom{1}{y_{ijk}} p_{ij}^{y_{ijk}} (1 - p_{ij})^{1 - y_{ijk}}
\]

\[
= p_{ij}^{y_{ijk}} (1 - p_{ij})^{1 - y_{ijk}}
\]

\[
= f_{\text{pair}}(\mu + \alpha_i + b_j)^{y_{ijk}} (1 - f_{\text{pair}}(\mu + \alpha_i + b_j))^{1 - y_{ijk}}
\]

The density for the random effects is given by:

\[
p(b_j) = \frac{1}{\sqrt{2\pi\sigma^2_b}} \exp \left\{ -\frac{1}{2\sigma^2_b} b_j^2 \right\}
\]
The simultaneous distribution of $Y_{ijk}$ and $B_j$ is given by the probability mass function:

$$p(y_{ijk}, b_j) = p(b_j)p(y_{ijk}|b_j)$$ (3.18)

and the marginal distribution of $Y_{ijk}$ is given by the probability mass function:

$$p(y_{ijk}) = \int_{-\infty}^{\infty} p(y_{ijk}, b_j)db_j$$

$$= \int_{-\infty}^{\infty} p(b_j)p(y_{ijk}|b_j)db_j$$

$$= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2_b}} \exp \left\{ -\frac{1}{2\sigma^2_b} b_j^2 \right\} \cdot f_{\text{pair}}(\mu + \alpha_i + b_j)^{1-y_{ijk}}(1 - f_{\text{pair}}(\mu + \alpha_i + b_j))^{y_{ijk}}db_j$$

It is possible to extend the model in (3.15) by adding the assessor-by-product interaction:

$$g(p_{ij}) = \mu + \alpha_i + b_j + d_{ij}$$ (3.19)

where $d_{ij}$ is the random effect of the interaction for the $j$th assessor and $i$th test product. The random effects are assumed to be independent and identically distributed:

$$D_{ij} \sim N(0, \sigma^2_d)$$ (3.20)

where $d_{ij}$ is a realization of the random variable $D_{ij}$. It is assumed that $D_{ij}$ is independent of $B_j$. Furthermore, the probability in (3.19) is the conditional probability given the random effects:

$$p_{ij} = P(Y_{ijk} = 1|B_j = b_j, D_{ij} = d_{ij})$$

In addition, the binary data are binomially distributed conditional on the realized values of the random variables $B_j = b_j$ and $D_{ij} = d_{ij}$:

$$Y_{ijk}|B_j = b_j, D_{ij} = d_{ij} \sim \text{Binomial}(p_{ij}, 1)$$ (3.21)

where

$$p_{ij} = f_{\text{pair}}(\mu + \alpha_i + b_j + d_{ij})$$

In the remaining parts of the thesis, including the papers in the appendices, $b_j$ and $d_{ij}$ are used as the realizations as well as referring to the random variables $B_j$ and $D_{ij}$. It will be clear from the context whether they refer to a realization or the random variable.
3.5 Simplification of Thurstonian models

An important aspect to consider is whether it is possible to simplify the model. The possible simplification will be investigated by likelihood ratio testing.

3.5.1 Likelihood function

For simplicity the likelihood function is written for model (3.15). The likelihood function of the marginal distribution of $Y_{ijk}$ is given as:

$$L(\beta, b_j; Y) = \prod_{j=1}^{J} \int_{-\infty}^{\infty} \prod_{i=1}^{I} \prod_{k=1}^{K} p(y_{ijk}|b_j)p(b_j) db_j$$

where $\beta = (\mu, \alpha_1, \ldots, \alpha_I)^T$ is the vector of the parameters for the fixed effects and $Y = (Y_{111}, \ldots, Y_{IJK})$ is the vector of the observations. The log-likelihood function becomes:

$$\ell(\beta, b_j; Y) = \log L(\beta, b_j; Y) = \sum_{j=1}^{J} \log \left( \int_{-\infty}^{\infty} \sum_{i=1}^{I} \sum_{k=1}^{K} p(y_{ijk}|b_j)p(b_j) db_j \right)$$

3.5.2 Hypothesis testing

As for linear mixed models, the approach when trying to simplify a model is by considering the hypothesis test for the interactions. The first effect to consider is the highest-order interaction, which for model (3.19) is the two-way interaction between assessors and products. The hypothesis test for the assessor-by-product interaction reads:

$$H_0 : \sigma_d^2 = 0 \quad \text{versus} \quad H_1 : \sigma_d^2 > 0$$

(3.22)

The alternative hypothesis is one-sided due to the variance being non-negative. Therefore the model under the null hypothesis reads:

$$g(p_{ij}) = \mu + \alpha_i + b_j$$

The test that is considered is the likelihood ratio test, which is given by minus twice the log-likelihood function under the alternative hypothesis minus the log-likelihood function under the null hypothesis:

$$X_{\text{LRT}} = -2 \log Q = -2(\ell_0 - \ell_1) = 2(\ell_1 - \ell_0)$$
where $\ell_0$ and $\ell_1$ are the log-likelihood functions under the null and alternative hypothesis respectively. The distribution of $X_{LRT}$ can be considered in two ways. Either by the usual asymptotic theory for which $X_{LRT} \sim \chi^2_1$ or by taking into account that the situation under the null hypothesis is on the boundary of the parameter space leading to $X_{LRT}$ following a mixture of the two $\chi^2$ distributions with 0 degrees of freedom and 1 degree of freedom respectively.

When the hypothesis test for the assessor-by-product interaction results in a non-significant interaction, the hypothesis tests for the main effects of assessor and product are well-defined and interpretable and defined as in Linander et al. (2018a). However, in the situation where the assessor-by-product interaction is significant the hypothesis test for the main effects become more complicated.

When considering the hypothesis tests for the main effects, in the situation of a non-significant assessor-by-product interaction, the hypotheses read:

$$H_0 : \sigma^2_b = 0 \text{ versus } H_1 : \sigma^2_b > 0 \quad (3.23)$$

and

$$H_0 : \alpha_i = 0 \text{ for all } i \text{ versus } H_1 : \alpha_i \neq 0 \text{ for at least one } i \quad (3.24)$$

In the situation of a significant assessor-by-product interaction one must consider how to interpret the hypotheses in (3.23) and (3.24). What would it mean say if $\sigma^2_b = 0$ when $\sigma^2_d > 0$? Would there be a meaningful way to interpret that the "main" effect of assessor is non-significant whereas the assessor-by-product interaction is significant? What would the assessor-by-product interaction really express? One interpretation is that the assessor-by-product interaction is expressing that there are assessor dependent differences between the products. When considering the test of product, when the assessor-by-product interaction is significant, could mean that the differences between the products depend on the assessors, that no differences purely between the products exist.

### 3.6 Estimation of parameters

This section covers how estimation using the GLM and the GLMM is done. Two different approaches are used depending of the definition of the variable; one for fixed effects and another for random effects.
3.6.1 Estimation of fixed effects

This section describes how fixed effects are estimated using GLMs and GLMMs. The fixed effects are estimated by maximum likelihood estimation. For generalized linear models as well as generalized linear mixed models, the maximum likelihood estimates are found as solutions to the score equations. The score equations are given as

\[
\frac{\partial}{\partial \beta} \log L(\beta, b_j; Y) = \frac{\partial}{\partial \beta} \ell(\beta, b_j; Y) = 0
\]  

(3.25)

The maximum likelihood estimate \( \hat{\beta} \) is the solution to (3.25).

3.6.2 Prediction of random effects

Random effects are realizations of unknown random variables. It can be of high importance to get information about the value of their realizations. This section explains how the predictions of the random components from a GLMM are defined.

The prediction of a random effect in a generalized linear mixed model is given as the mode in the conditional distribution of the random effect given data (Jiang et al., 2001; Christensen and Brockhoff, 2012; Bates et al., 2015). The mode of a conditional distribution is the value that maximizes the conditional density. Therefore, \( \hat{b}_j \); the prediction for the \( j \)th assessor is maximizing \( p(b_j|y_j) \) where \( y_j \) is the vector of observations for the \( j \)th assessor.

3.7 Estimation of d-prime values

Due to the definition of the psychometric function for the binary paired comparison, given in (3.9), the parameters in the linear predictor \( \eta_{ij} \) in (3.15), is on the d-prime scale. Thus, \( \delta_{ij} \); the sensory difference between the \( i \)th test product and the control for the \( j \)th assessor, is given as:

\[
\delta_{ij} = \mu + \alpha_i + b_j
\]  

(3.26)

It is possible to obtain product, as well as assessor specific sensory differences from (3.26).

The product specific sensory differences are the differences between the products
3.8 Fitting the models using \texttt{R}

and the control for an average assessor. Since \( b_j \sim N(0, \sigma^2_b) \) an average assessor corresponds to \( \text{E}[b_j] = 0 \). Therefore, an average assessor is when \( b_j = 0 \) and the product specific d-prime values are given as:

\[ \delta_i = \mu + \alpha_i \]

There are two different assessor specific sensory differences. The first type is \( b_j \) which is the difference, on the d-prime scale, from the average product-difference \( \mu \) for the \( j \)th assessor. The second type is the difference between an average test product and the control for the \( j \)th assessor. This is for an average test product corresponding to \( \alpha_i = 0 \). Thus,

\[ \delta_j = \mu + b_j \]

3.8 Fitting the models using \texttt{R}

It is possible to fit Thurstonian models as well as Thurstonian mixed models in \texttt{R}. When having random effects in the model, giving a Thurstonian mixed model, such models can be considered in \texttt{R} by using the \texttt{lme4} package (Bates et al. (2015)).

Let \( \text{dat} \) be a data frame with a row for each observation \( Y_{ijk} \). Moreover, let the columns of \( \text{dat} \) be the response variable as well as explanatory variables. More specifically, let \texttt{Attribute} be the response variable and let \texttt{Assessor}, \texttt{Product} and \texttt{Session} be the explanatory variables. \texttt{Session} is not used in the model since we are considering models with effects of assessors and products. The model in (3.15) is fitted by:

\[
\begin{align*}
\text{fm} & \leftarrow \text{glmer(} \text{Attribute} \sim \text{Product} + (1|\text{Assessor}) + (1|\text{Assessor:Product}), \\
& \quad \text{data} = \text{dat}, \\
& \quad \text{family} = \text{binomial(probit)}, \\
& \quad \text{contrasts} = \text{list(}"\text{Product}"=\text{contr.sum}, \\
& \quad \text{control} = \text{glmerControl(}\text{optimizer}="bobyqa"))
\end{align*}
\]

Here, the \texttt{family} option is set to be \texttt{binomial(probit)} which means that the inbuilt link function \texttt{probit} is used for binomially distributed data. The psychometric function defined in (3.9) reads:

\[
p_{ij} = f_{\text{pair}}(\delta_{ij}) \\
= \Phi\left( \frac{\delta_{ij}}{\sqrt{2}} \right)
\]
Now, the link function becomes

\[
\delta_{ij} = f_{\text{pair}}^{-1}(p_{ij})
= \Phi^{-1}(p_{ij})\sqrt{2}
\]

which is the probit link multiplied by the square-root of 2.

The parameters in the linear predictor \( \eta_{ij} \) in (3.15) are on the d-prime scale. To get the parameters from the fitted model on the d-prime scale the estimates from \( \text{fm} \) must be multiplied by \( \sqrt{2} \). Thus, the product specific d-prime values are obtained by:

\[
\text{alphas} \leftarrow \text{fixef}()[-1]\times\text{sqrt}(2)
\text{alphas} \leftarrow c(\text{alphas}, 0-\text{sum}(\text{alphas}))
\]

where the last value of \( \alpha_i \) is found using the restriction that the parameters must sum to zero.

The assessor specific d-prime values, \( \hat{\beta}_j \), are obtained by:

\[
\text{ranef()}\text{"Assessor"}\times\text{sqrt}(2)
\]

To be able to do the hypothesis test of a significant assessor-by-product interaction, the model without the assessor-by-product interaction must be fitted:

\[
\text{fm2} \leftarrow
\text{glmer()}\text{Attribute} \sim \text{Product} + (1|\text{Assessor})
, \text{data} = \text{dat},
, \text{family} = \text{binomial(probit)},
, \text{contrasts} = \text{list("Product"=contr.sum),}
, \text{control=glmerControl(optimizer="bobyqa")}
\]

The likelihood ratio test for the assessor-by-product interaction is obtained by:

\[
(\text{LRT} \leftarrow 2*(\text{logLik}(\text{fm2}) - \text{logLik}(\text{fm}))
\]

and the p-value is obtained by:

\[
(\text{pVal} \leftarrow 1 - \text{pchisq}(\text{LRT}, df = 1))
\]
3.9 Concluding remarks

It is possible to choose other expressions for $\eta_{ij}$ in (3.19). One possibility is to include session as an explanatory variable. Allowing the probability of a test product being chosen to possibly depend on the session means that a model with only main effects reads:

$$g(p_{ijk}) = \mu + \alpha_i + \tau_k + b_j$$

where $p_{ijk} = P(Y_{ijk} = 1)$.

Another explanatory variable that could be interesting to consider is time. It could be of interest to investigate if there is an effect of time in situations where the testing is spread out using more than one day. This way of modelling opens up for many possibilities for investigating which variables that possibly affect the probability of choosing a specific product.
Can a Thurstonian mixed model be trusted?

In Linander et al. (2018a) and in Chapter 3 Thurstonian models for the binary paired comparison are considered. In Linander et al. (2018a) it is briefly considered how the hypothesis test of the product effect is affected by ignoring replications from the assessors or including them. In this chapter, the importance of handling the replications correctly will be investigated. This will be considered by investigating the test for products.

Another interesting, as well as important, matter to investigate, is how good a Thurstonian mixed model is at detecting a true assessor-by-product interaction. In Linander et al. (2018a) the hypothesis test for the assessor-by-product interaction results in non-significant interactions for all the sensory attributes. Thus, it is relevant to investigate whether the model will be able to detect true assessor-by-product interactions.

Generally, it is important that a sensory discrimination study has high power such that the results can be trusted (Ennis, 1993; Bi and Ennis, 1999).
4.1 Handling replications

Replications often occur in sensory discrimination testing. It is important to handle the replications carefully to be able to obtain proper results regarding the products.

When considering studies with one product, many suggestions exist of how to handle the replications. Two methods that are widely used are the so-called beta-binomial and corrected beta-binomial models (Næs et al., 2010; Brockhoff, 2003; Ennis and Bi, 1998).

4.1.1 The hypothesis test for products - an example

A comparison of the values of the likelihood ratio test statistics, for the discrimination study introduced in Linander et al. (2018a), for the models with and without assessor is shown in Figure 4.1 (which is the same figure as in Linander et al. (2018a)).

For the majority of the attributes the likelihood ratio test statistics are extremely large, thus it has no practical impact which of the two models are considered. However, the values for the test using the model including assessors are larger than when ignoring the replications. This is important for Greasy evaluated initially after application. Using a significance level of 0.01, the conclusion regarding a product effect depends on which model is used. Furthermore, the difference between the likelihood ratio statistic and the critical value, using the 0.05, level is small.

4.1.2 The test of product main effect - a simulation study

In this section it will be investigated how the test of the main effect of products is affected by the choice of model in the case of a true assessor main effect.

This simulation study is investigating the importance of modelling the assessors when data consist of replications for the assessors.

The model that is considered is the model with the main effects of products and
Figure 4.1: Comparing the likelihood ratio test statistics for hypothesis test of product main effect. The horizontal lines are the critical values for the Chi-squared distribution with 7 degrees of freedom; the 0.05 critical value (dotted line) and the 0.01 critical value (dashed line).
30 Can a Thurstonian mixed model be trusted?

assessors:

\[ g(p_{ij}) = \tilde{\alpha}_i + b_j \Leftrightarrow p_{ij} = f_{\text{paired}}^{-1}(\tilde{\alpha}_i + b_j) \]

\[ = P(Y_{ijk} = 1) \tag{4.1} \]

where \( \tilde{\alpha}_i = \mu + \alpha_i \) and \( b_j \sim N(0, \sigma_b^2) \) being independent for all \( j \). The model in (4.1) is used as the so-called simulating model.

The simulation of data consists of two steps:

**Step 1:** Simulating the probabilities \( p_{ij} \)s from a set of parameter values using model (4.1)

**Step 2:** Simulating the binomially distributed data \( Y_{ijk} \):

\[ Y_{ijk} \sim \text{Binomial}(p_{ij}, 1) \tag{4.2} \]

using the simulated probabilities \( p_{ij} \)s from **Step 1**, where observations are independent over \( k \). Meaning that \( P(Y_{ij1} = 1) = P(Y_{ij2} = 1) = p_{ij} \).

Two estimating models will be used to model the simulated data. One ignoring the replications for the assessors:

\[ g(p_{ij}) = \tilde{\alpha}_i \tag{4.3} \]

and the model given by [4.1] where the replications are modelled by including assessor in the model.

The hypothesis test of interest is the hypothesis test for product. Thus the hypotheses are given as:

\[ H_0 : \tilde{\alpha}_1 = \tilde{\alpha}_2 = \cdots = \tilde{\alpha}_I \quad H_1 : \tilde{\alpha}_i \neq \tilde{\alpha}_{i'} \quad \text{for some } i \neq i' \tag{4.4} \]

One aim of this simulation study is to compare the size of the likelihood ratio test statistics using the two different estimating models. This will illustrate the differences between the test statistics. However, this does not illustrate the consequence of the differences between the test statistics. Thus, to investigate how the conclusions are affected the power of detecting a product difference for the two estimating models will be considered.

The power is the probability of rejecting the null hypothesis when the alternative hypothesis is true:

\[ \text{power} = P(\text{correctly rejecting } H_0) = 1 - \beta \tag{4.5} \]

where \( \beta \) is the probability of accepting the null hypothesis when the alternative hypothesis is true.
Table 4.1: Parameter values of the $\tilde{\alpha}_i$s used for the simulation study.

<table>
<thead>
<tr>
<th>$\tilde{\alpha}_1$</th>
<th>$\tilde{\alpha}_2$</th>
<th>$\tilde{\alpha}_3$</th>
<th>$\tilde{\alpha}_4$</th>
<th>$\tilde{\alpha}_5$</th>
<th>$\tilde{\alpha}_6$</th>
<th>$\tilde{\alpha}_7$</th>
<th>$\tilde{\alpha}_8$</th>
</tr>
</thead>
<tbody>
<tr>
<td>small</td>
<td>−0.5</td>
<td>0.5</td>
<td>0.2</td>
<td>−0.5</td>
<td>0.1</td>
<td>0.2</td>
<td>−0.3</td>
</tr>
<tr>
<td>medium</td>
<td>−1.0</td>
<td>−0.6</td>
<td>0.4</td>
<td>−0.7</td>
<td>0.5</td>
<td>−0.3</td>
<td>−0.1</td>
</tr>
<tr>
<td>large</td>
<td>−1.1</td>
<td>1.5</td>
<td>1.2</td>
<td>−1.5</td>
<td>−1.2</td>
<td>−0.9</td>
<td>−1.3</td>
</tr>
</tbody>
</table>

Table 4.2: Parameter values of $\sigma_b$ used for the simulation study.

| $\sigma_b$ | 0.5 | 1   | 2   |

The $Y_{ijk}$s will be simulated for eight products ($I = 8$), 25 assessors ($J = 25$) and two sessions ($K = 2$). Three sets of values for $\tilde{\alpha}_1, \ldots, \tilde{\alpha}_8$ will be considered. The three sets corresponds to different sizes of the d-prime values. The idea is that one set is for small d-prime values, another set for medium d-prime values and a third for large d-prime values. Looking at the values of the product specific d-prime values for the analysis of the discrimination study, the three groups are defined such that a difference is:

- small when $\alpha_{\text{max}} - \alpha_{\text{min}} \leq 1$
- medium when $\alpha_{\text{max}} - \alpha_{\text{min}} \leq 1.5$
- large when $\alpha_{\text{max}} - \alpha_{\text{min}} \leq 3$

Three different values of the variance in the normal distribution for the assessors are considered. The parameter values used in the simulation study are listed in Table 4.1 and 4.2. The values for the parameters for the $\tilde{\alpha}_i$s are on the d-prime scale.

4.1.3 Results

The results of the simulation study are summarized in two ways; a histogram of the differences of the likelihood ratio test statistics and the power for the two models.

The difference of the likelihood ratio test statistics for the $l$th simulation reads:

$$D_{\text{LRT}} = X_{\text{LRT, assessor}} - X_{\text{LRT, ignoring assessor}}$$ (4.6)
Can a Thurstonian mixed model be trusted?

Figure 4.2: The differences of the likelihood ratio test statistics for the small sized d-prime values.

where \( X_{LRT, \text{assessor}} \) is the likelihood ratio test statistic obtained using model (4.1) and \( X_{LRT, \text{ignoring assessor}} \) is the likelihood ratio test statistic obtained using model (4.3).

The simulated power is found as the number of times the null hypothesis is rejected out of the total number of tests (Bi (2011)):

\[
\text{simulated power} = \frac{\# H_0 \text{ is rejected}}{\# \text{ p-values}} \quad (4.7)
\]

The power will be found using a level of \( \alpha \) of 0.05.

The values of \( D_{LRT} \) for the small sized d-prime values is shown in Figure 4.2.

I would expect that the larger the variance the bigger the differences between the test statistics become. This tendency is seen in Figure 4.2. Furthermore, as
expected the values for $\sigma_b = 0.5$ are smaller than for the other two values of $\sigma_b$, with some very close to 0.

The values of $D_{LRT}$ for the medium sized d-prime values is shown in Figure 4.3.

As before, the differences increase with increasing standard deviation. Furthermore, the differences are a bit larger than in Figure 4.2.

The values of $D_{LRT}$ for the large sized d-prime values is shown in Figure 4.4.

As before, the differences increase with increasing standard deviation. Furthermore, the differences are larger than in Figure 4.2 and 4.3.
**Figure 4.4:** The differences of the likelihood ratio test statistics for the large sized differences.
4.1 Handling replications

**Table 4.3:** Estimated power for the hypothesis test for product for the 0.05 level.

<table>
<thead>
<tr>
<th>Model</th>
<th>$\sigma_b = 0.5$</th>
<th>$\sigma_b = 1$</th>
<th>$\sigma_b = 2$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>small</td>
<td>medium</td>
<td>large</td>
</tr>
<tr>
<td>with</td>
<td>78</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>without</td>
<td>73</td>
<td>99</td>
<td>100</td>
</tr>
</tbody>
</table>

For all of the different d-prime values the differences between the test statistics for the two models increase with increasing assessor variability. This is expected, since the more the assessors vary the more improper the model ignoring the assessors become. When the assessor variability is small it seems more reasonable to ignore the assessor variability. However, it is not possible to know from the differences whether the conclusions are affected, since this also depends on the actual size of the test statistics and not only their difference.

The tendency that is seen in this simulation study also appears to be recognizable in the analysis of the discrimination study. From Figure 4 in Linander et al. (2018a) it is seen that Silky after five minutes is the sensory attribute with the largest effect of the assessors. It is also the sensory attribute with the largest difference between the test statistics in Figure 4.1. Furthermore, the sensory attributes with non-significant assessor effects are Absorption and Thickness which are among the attributes with the smallest differences for the test statistics.

The estimated powers for the simulation study, using the 0.05 level, are shown in Table 4.3.

Considering $\sigma_b = 0.5$ the two models perform equally well with a maximum difference of 5. For $\sigma_b = 1$ the two models perform equally well for medium and large sized d-prime values. For small sized d-prime values the two methods differ with the model including assessor having the highest power of 74. Considering $\sigma_b = 2$ the models have the same power for large sized d-prime values. However, for small and medium sized d-prime values the model including assessor is far superior than the model ignoring the assessors.

Generally, when assessor is included in the model the value of the likelihood ratio test statistic becomes larger than when assessor is omitted from the model. Larger values of the likelihood ratio test statistic means that the null hypothesis
of no product differences will more often be rejected. Thus, the test when assessor is included is more sensitive than when assessor is omitted.

4.2 Will the model detect significant assessor-by-product interactions - an investigation

In Linander et al. (2018a) the test for the assessor-by-product interaction resulted in non-significant interactions for all the attributes. A natural question that arises is whether the model will be able to detect a significant assessor-by-product interaction when it exists. It is important that the model will be able to do this, since the interpretations in Linander et al. (2018a) are based on the fact that the assessor-by-product interaction is non-significant. Thus, it is of high importance to be able to trust that the model without the assessor-by-product interaction is a model that describes the data well.

In this section an investigation is made considering a small simulation study to investigate the model’s ability to detect true assessor-by-product interactions.

4.2.1 The investigation

The model that is considered in the simulation study reads:

\[
g(p_{ij}) = \tilde{\alpha}_i + b_j + d_{ij} \iff p_{ij} = f_{\text{pair}}(\tilde{\alpha}_i + b_j + d_{ij})
\]

(4.8)

where

\[
Y_{ijk} \sim \text{Binomial}(p_{ij}, 1), \quad p_{ij} = P(Y_{ijk} = 1)
\]

with \(b_j \sim N(0, \sigma^2_b)\) and \(d_{ij} \sim N(0, \sigma^2_d)\) being independent for all \(i\) and \(j\).

As in Section 4.1.1 the simulation of data consists of two steps:

**Step 1:** Simulating the probabilities \(p_{ij}\)s from a set of parameter values using model (4.8).

**Step 2:** Simulating the binomially distributed data \(Y_{ijk}\):

\[
Y_{ijk} \sim \text{Binomial}(p_{ij}, 1)
\]

(4.9)
4.2 Will the model detect significant assessor-by-product interactions - an investigation

Table 4.4: Parameter values of the $\tilde{\alpha}_i$s used for the simulation study.

<table>
<thead>
<tr>
<th>$\tilde{\alpha}_1$</th>
<th>$\tilde{\alpha}_2$</th>
<th>$\tilde{\alpha}_3$</th>
<th>$\tilde{\alpha}_4$</th>
<th>$\tilde{\alpha}_5$</th>
<th>$\tilde{\alpha}_6$</th>
<th>$\tilde{\alpha}_7$</th>
<th>$\tilde{\alpha}_8$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$-1.57$</td>
<td>$-2.39$</td>
<td>$-0.89$</td>
<td>$0.83$</td>
<td>$-1.47$</td>
<td>$-0.37$</td>
<td>$-1.77$</td>
<td>$-3.03$</td>
</tr>
</tbody>
</table>

Table 4.5: Parameter values of $\sigma_b$ and $\sigma_d$ used for the simulation study.

<table>
<thead>
<tr>
<th>$\sigma_b^2$</th>
<th>$\sigma_d^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Situation 1</td>
<td>1.61</td>
</tr>
<tr>
<td>Situation 2</td>
<td>1.61</td>
</tr>
<tr>
<td></td>
<td>3.22</td>
</tr>
</tbody>
</table>

using the simulated probabilities $p_{ij}$s from Step 1, where observations are independent over $k$. Meaning that $P(Y_{ij1} = 1) = P(Y_{ij2} = 1) = p_{ij}$.

The aim of the simulation study is to investigate how well the model will detect true assessor-by-product effects of various size. This will be investigated by considering the power of the model for two different scenarios.

For the parameters to be realistic, the values for the parameters, except for $\sigma_d$, for the simulation study are the estimates obtained from the analysis with the largest assessor differences. From Section 4.1.3 Silky after five minutes is known as the sensory attribute with the largest effect of the assessors. The d-prime values of the test products are obtained from Linander et al. (2018b) and are listed in Table 4.4.

The variance parameters are chosen such that $\sigma_b$ equals the estimate from the analysis of Silky after five minutes. Furthermore, $\sigma_d^2$ is considered for two scenarios:

$$\sigma_d^2 = \sigma_b^2 \quad \text{and} \quad \sigma_d^2 = 2\sigma_b^2 \quad (4.10)$$

The values of the variances are listed in Table 4.5.

The simulation study is conducted for two replications for 25 assessors and 8 products doing 1000 simulations. The power for the two situations are 0.42 for situation 1 and 0.94 for situation 2. These preliminary findings indicate that the model would be able to detect a true assessor-by-product difference under certain circumstances. Further, investigations are needed to gain knowledge...
about the stability of the model with respect to detecting true assessor-by-product interactions. For the situation where the variances for the main effect of assessors and assessor-by-product interaction are equal the power was low. However, when the variance of the assessor-by-product is twice the variance of the assessors the power was very high.

4.3 Concluding remarks

The simulation studies considered in this chapter are merely considering two aspects with respect to Thurstonian mixed models. It would also be interesting to consider a simulation study to investigate the power of detecting assessor differences. Furthermore, other values of the parameters could be interesting to consider.

More work needs to be done to investigate how the models are performing when unbalanced data are considered.
This chapter is concerned with aspects of principal component analysis (PCA), which are only briefly considered in Linander et al. (2018b).

5.1 Choice of assessor specific d-prime values

When considering Thurstonian mixed models, as in Linander et al. (2018a), two different assessor specific d-prime values are considered; the $b_j$s and the $\tilde{b}_j$s. In Linander et al. (2018b) the PCA using the $b_j$ values is considered. In this section it is investigated what information is gained by considering both types of assessor specific d-prime values.

5.1.1 Centering

When considering the assessor specific d-prime values, it is possible to do several versions of a PCA. In this section the influence of centering will be investigated.
Let $b_{1j}, \ldots, b_{Mj}$ be the predictions for the $M$ assessors for the $j$th attribute, where $b_{mj}$ is the prediction for the $m$th assessor for the $j$th attribute.

Due to the assumption of the normal distribution the $b_{mj}$s are predicted such that $E(b_{mj}) \approx 0$. Therefore, the assessor specific d-prime values $b_{mj}$s are almost centered. Thus, when using the $b_{mj}$ values, the centered and non-centered PCA will give similar results.

Let $\tilde{b}_{mj} = \mu_j + b_{mj}$ be the estimated difference between products and the control for the $m$th assessor for the $j$th attribute.

Now:

$$E(\tilde{b}_{mj}) = E(\mu_j + b_{mj})$$
$$= \mu_j + E(b_{mj})$$
$$\approx \mu_j$$

Therefore:

$$\tilde{b}_{mj} - E(\tilde{b}_{mj}) \approx \mu_j + b_{mj} - \mu_j = b_{mj}$$

Thus, when centering the $\tilde{b}_{mj}$ values the result is similar to the results when considering the $b_{mj}$ values.

The centered versions of both $b_{mj}$ and $\tilde{b}_{mj}$ lead to similar results as considering the non-centered PCA of $b_{mj}$. Therefore, two different versions of PCA will be considered in the next sections. These will be the non-centered PCA of the $\tilde{b}_{mj}$ values as well as the centered PCA of the $b_{mj}$ values.

### 5.1.2 PCA using assessor specific d-prime values

In this section the assessor specific d-prime values are obtained by modelling data from an existing discrimination study provided by Unilever. This study is the same as described and analyzed in Linander et al. (2018a).

The biplot using the $b_{mj}$ values is shown in Figure 5.1 (which is the same as Figure 4 in Linander et al. (2018a)).

The assessors 8 and 7 are the furthest away from 0 in the same direction as the arrow with respect to Silky after five minutes. Whereas the assessors 1,
5.1 Choice of assessor specific d-prime values

Figure 5.1: The biplot for the centered assessor specific d-prime values $b_j$. 
6, 9 and 19 are the furthest in the opposite direction with respect to Silky after five minutes. Regarding Greasy (initially as well as after five minutes) the assessors 4, 19, 23 and 24 are the furthest away from 0 in the same direction as the arrow. Whereas, the assessors 2, 5, 9, 14 and 26 are the furthest in the opposite direction with respect to Greasy.

The biplot using the $\tilde{b}_{mj}$ values is shown in Figure 5.2

Overall the same grouping of assessors is evident from Figure 5.2 as for 5.1. To mention some of the groups; the assessors 7 and 8 are close together, assessors 12 and 15 are close. Moreover, assessors 11, 18 and 22 are close together. Furthermore positions with respect to the attributes are also the same in the two figures. Thus, both the $b_{mj}$ values as well as the $\tilde{b}_{mj}$ values can be used to investigate which assessors are scoring similarly across attributes.
5.2 Significant versus non-significant effects - is scaling needed?

When considering multiple attributes, a situation that can occur is that some attributes have significant effects, whereas other attributes have non-significant effects. The argument in Linander et al. (2018a) for not scaling the d-prime values before doing the PCA is that d-prime values are measured using the same scale. However, when an attribute has a non-significant effect the d-prime values for that attribute will be close to 0. In this section, it is investigated how the principal component analysis is affected by including attributes with non-significant effects. More specifically, the assessor specific d-prime values from the discrimination study used as an example in Linander et al. (2018a) will be considered.

In this section, it will be the $\tilde{b}_{mj}$ values that are considered, since the arrows in Figure 5.1 are much shorter than the arrows in Figure 5.2.

Two approaches will be considered. One approach is omitting the attributes with non-significant assessor effects. The other approach is to scale the d-prime values.

5.2.1 Omitting non-significant effects

In the discrimination study two attributes, Absorption and Thickness, have non-significant assessor effects. The biplot for the PCA using the $\tilde{b}_{mj}$ values, where Absorption and Thickness have been omitted, is shown in Figure 5.3. The positions of the arrows as well as the assessors do not change when omitting Thickness and Absorption. Thus, the interpretations stay the same. Thus, it appears that the attributes with significant assessor main effects have the most impact in the PCA.

5.2.2 Scaling

The biplot for the PCA using the scaled $\tilde{b}_{mj}$ values is shown in Figure 5.4.

When scaling the assessor specific d-prime values $\tilde{b}_j$ the attributes Absorption and Thickness become visible in the biplot. As expected, the length of the
Figure 5.3: The biplot for the non-centered assessor specific d-prime values $\tilde{b}_j$ with Absorption and Thickness omitted.
Figure 5.4: The biplot for the non-centered assessor specific d-prime values $\tilde{b}_j$, where the values have been scaled.
arrows are more or less the same. The plot in Figure 5.3 is flipped over the y-axis, compared to Figure 5.2. Moreover, the area in which the assessors are placed is shrunken when scaling. An assessor worth mentioning is assessor 7. Without scaling assessor 7 and 8 are similar. However, when scaling assessor 7 no longer is near assessor 8. Considering all of the assessors positions with respect to Sticky (initially and after five minutes), Thickness and Greasy (initially and after five minutes) the results are similar. It appears that Absorption is explaining some of the second principal component, that were explained previously by Silky after five minutes. Thus, it appears that the shift of placement for assessor 7 is primarily due to a high proportion of times test products were chosen for Absorption (see Table 5.1).
Table 5.1: Proportions of a test product being chosen for the assessors listed in percentages.

<table>
<thead>
<tr>
<th></th>
<th>Greasier</th>
<th>Greasier.5min</th>
<th>Silkier</th>
<th>Silkier.5min</th>
<th>Stickier</th>
<th>Stickier.5min</th>
<th>Absorb</th>
<th>Thicker</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>62</td>
<td>75</td>
<td>17</td>
</tr>
<tr>
<td>6</td>
<td>38</td>
<td>12</td>
<td>12</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>62</td>
<td>50</td>
<td>22</td>
</tr>
<tr>
<td>16</td>
<td>0</td>
<td>0</td>
<td>12</td>
<td>12</td>
<td>62</td>
<td>0</td>
<td>38</td>
<td>50</td>
<td>22</td>
</tr>
<tr>
<td>2</td>
<td>8</td>
<td>0</td>
<td>25</td>
<td>8</td>
<td>50</td>
<td>17</td>
<td>50</td>
<td>33</td>
<td>24</td>
</tr>
<tr>
<td>11</td>
<td>19</td>
<td>12</td>
<td>25</td>
<td>12</td>
<td>19</td>
<td>19</td>
<td>75</td>
<td>31</td>
<td>27</td>
</tr>
<tr>
<td>18</td>
<td>12</td>
<td>6</td>
<td>19</td>
<td>12</td>
<td>44</td>
<td>44</td>
<td>38</td>
<td>56</td>
<td>29</td>
</tr>
<tr>
<td>1</td>
<td>12</td>
<td>19</td>
<td>19</td>
<td>6</td>
<td>12</td>
<td>25</td>
<td>81</td>
<td>69</td>
<td>30</td>
</tr>
<tr>
<td>22</td>
<td>8</td>
<td>25</td>
<td>17</td>
<td>8</td>
<td>75</td>
<td>33</td>
<td>33</td>
<td>50</td>
<td>31</td>
</tr>
<tr>
<td>12</td>
<td>25</td>
<td>12</td>
<td>25</td>
<td>19</td>
<td>38</td>
<td>19</td>
<td>56</td>
<td>62</td>
<td>32</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0</td>
<td>25</td>
<td>38</td>
<td>44</td>
<td>31</td>
<td>56</td>
<td>69</td>
<td>33</td>
</tr>
<tr>
<td>15</td>
<td>19</td>
<td>19</td>
<td>25</td>
<td>19</td>
<td>56</td>
<td>19</td>
<td>38</td>
<td>69</td>
<td>33</td>
</tr>
<tr>
<td>21</td>
<td>38</td>
<td>19</td>
<td>25</td>
<td>19</td>
<td>19</td>
<td>25</td>
<td>56</td>
<td>62</td>
<td>33</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>19</td>
<td>50</td>
<td>38</td>
<td>44</td>
<td>19</td>
<td>50</td>
<td>56</td>
<td>34</td>
</tr>
<tr>
<td>14</td>
<td>0</td>
<td>0</td>
<td>62</td>
<td>50</td>
<td>38</td>
<td>25</td>
<td>38</td>
<td>62</td>
<td>34</td>
</tr>
<tr>
<td>19</td>
<td>50</td>
<td>50</td>
<td>12</td>
<td>0</td>
<td>38</td>
<td>50</td>
<td>50</td>
<td>50</td>
<td>38</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>6</td>
<td>50</td>
<td>75</td>
<td>44</td>
<td>38</td>
<td>38</td>
<td>62</td>
<td>39</td>
</tr>
<tr>
<td>13</td>
<td>31</td>
<td>31</td>
<td>44</td>
<td>38</td>
<td>44</td>
<td>19</td>
<td>31</td>
<td>75</td>
<td>39</td>
</tr>
<tr>
<td>4</td>
<td>38</td>
<td>75</td>
<td>25</td>
<td>0</td>
<td>12</td>
<td>38</td>
<td>88</td>
<td>50</td>
<td>41</td>
</tr>
<tr>
<td>7</td>
<td>6</td>
<td>6</td>
<td>62</td>
<td>69</td>
<td>25</td>
<td>12</td>
<td>62</td>
<td>81</td>
<td>41</td>
</tr>
<tr>
<td>20</td>
<td>0</td>
<td>50</td>
<td>38</td>
<td>25</td>
<td>50</td>
<td>62</td>
<td>75</td>
<td>38</td>
<td>42</td>
</tr>
<tr>
<td>17</td>
<td>38</td>
<td>31</td>
<td>50</td>
<td>62</td>
<td>31</td>
<td>6</td>
<td>50</td>
<td>75</td>
<td>43</td>
</tr>
<tr>
<td>25</td>
<td>19</td>
<td>31</td>
<td>44</td>
<td>44</td>
<td>50</td>
<td>56</td>
<td>56</td>
<td>50</td>
<td>44</td>
</tr>
<tr>
<td>10</td>
<td>17</td>
<td>42</td>
<td>25</td>
<td>42</td>
<td>58</td>
<td>67</td>
<td>50</td>
<td>75</td>
<td>47</td>
</tr>
<tr>
<td>24</td>
<td>44</td>
<td>38</td>
<td>50</td>
<td>31</td>
<td>69</td>
<td>38</td>
<td>50</td>
<td>56</td>
<td>47</td>
</tr>
<tr>
<td>23</td>
<td>56</td>
<td>62</td>
<td>50</td>
<td>62</td>
<td>31</td>
<td>31</td>
<td>56</td>
<td>56</td>
<td>51</td>
</tr>
</tbody>
</table>
5.2.3 Comparing scaling with omitting

All in all, when attributes are non-significant, it does not seem to matter whether these are included or omitted in the PCA. However, it appears to matter whether the d-prime values are scaled or not. Since a non-significant effect implies that there are no differences among the assessors, it is most likely the most interesting to consider the non-scaled d-prime values. However, specific knowledge about the non-significant attributes might make it interesting to also consider the scaled d-prime values.

5.3 Outliers and their influence

This section investigates the importance of a replacement value for an outlier in the setting of analyzing d-prime values.

For the discrimination study, used as an example, test product H has a d-prime value of minus infinity. It is investigated how the PCA is affected by different imputed values used instead of minus infinity as well as other ways of handling the extreme value.

5.3.1 Based on the raw proportions

In this section, it is investigated how the PCA is affected by different imputed values for an outlier using the d-prime values found by transforming the proportions by the inverse of the psychometric function.

A d-prime value of $-\infty$ only occurs when the test products were chosen zero times. A way to select the value to replace $-\infty$ is by letting the test products being chosen once. Using the proportion of $1/n$ instead of $0/n$, is conceptually the same, since it has no practical implication whether the test products were chosen zero times or one time. From both situations it is clear that the control is perceived to have the strongest sensory intensity. The advantage of using $1/n$ rather than $0/n$ is it ensures that a finite d-prime value is obtained. Using this approach a d-prime value of $-2.83$ is replacing $-\infty$. The result of using this approach is seen in Figure 5.5 (this is the same as Figure 1 in Linander et al. (2018b)).
Figure 5.5: The biplot for the raw d-prime values using the imputed value where the proportion used is $1/n$ rather than $0/n$. 

5.3 Outliers and their influence
Another approach is to choose a completely arbitrary value, in the sense that no conceptual similarity exists between $-\infty$ and the chosen value. It is investigated how the value of an arbitrary d-prime value affects the PCA. The biplot obtained using a d-prime value of $-10$ is shown in Figure 5.6. Compared to Figure 5.5, the biplot in Figure 5.6 is tilted towards the right affecting the values for the second principal component for all the test products. The values for the first principal component are more or less the same, except for test product H which has a value close to $-10$, the imputed value for Silky for test product H.

Thus, the conclusions are the same regardless which value is used to replace $-\infty$. The value for test product H changes with regards to Silky, but the conclusion is the same; that test product H is very different from the rest of the test products. The relations between the other test products are not affected by which value is used. Thus, it appears that the PCA, in this case, is not affected by the choice of the imputed value.

**Figure 5.6:** The biplot for the raw d-prime values using an imputed value of $-10$. 

- PC1 (78.8% explained var.)
- PC2 (12.2% explained var.)
5.3 Outliers and their influence

5.3.2 Based on the d-prime values obtained from a Thurstonian Mixed Model

It is investigated how the principal component analysis is affected when considering d-prime values that are obtained from a Thurstonian mixed model. When considering the product specific d-prime values the centered, as well as the non-centered PCA are considered. Thus, it will be investigated how the principal component analysis is affected in both situations.

5.3.2.1 Centering the d-prime values

Different values will be used to replace $-\infty$ to investigate the impact the chosen value has. Following the reasoning in Linander et al. (2018b) when considering the d-prime values obtained from the Thurstonian mixed model a value of $-3.47$ will be used as an imputed value. Furthermore, as in Section 5.3.1 a value of $-10$ will be considered. The biplot obtained from the PCA with $-3.47$ is shown in Figure 5.7 (this is the same as Figure 2 in Linander et al. (2018b)).

The biplot from doing the PCA with $-10$ is seen in Figure 5.8. Compared to Figure 5.7 the biplot is flipped over the y-axis but the conclusions remain the same. Test product H is further away from 0 with respect to the first principal component. However, this is expected since the first principal component primarily is explained by Silky.

Thus, for the centered product specific d-prime values, it appears as if the PCA is only affected in an expected way when imputing a value for test product H for Silky.
Figure 5.7: The biplot for the centered d-prime values using an imputed value of \(-3.47\).
Figure 5.8: The biplot for the centered d-prime values using an imputed value of $-10$. **5.3 Outliers and their influence**
5.3.2.2 Non-centered d-prime values

As in Section 5.3.2.1 the values $-3.47$ and $-10$ will be considered. The biplot obtained from the PCA with $-3.47$ is shown in Figure 5.9 (this is the same as Figure 3 in Linander et al. (2018b)).

The biplot obtained from the PCA using $-10$ is seen in Figure 5.10. At first sight, it might look as if the conclusions change. However, the biplot in Figure 5.10 is merely tilted towards the right compared to Figure 5.9. Thus, the test products have the same relative positions with respect to the arrows. Therefore, the conclusions remain the same.

Thus, for the non-centered product specific d-prime values, it appears as if the PCA is only affected in an expected way when imputing a value for test product $H$ for Silky.
5.3 Outliers and their influence

Figure 5.9: The biplot for the non-centered d-prime values using an imputed value of −3.47.
Figure 5.10: The biplot for the non-centered d-prime values using an imputed value of $-10$. 
5.3.2.3 Excluding test product H

A different approach than imputing a value, is to exclude test product H from the analysis. The biplot from doing the PCA using the non-centered d-prime values without test product H is seen in Figure 5.11. The biplot in Figure 5.11 is tilted a bit towards the right compared to the biplot in Figure 5.9. However, the conclusions remain the same.

The biplot from doing the PCA using the centered d-prime values without test product H is seen in Figure 5.12. The biplot in Figure 5.12 is different than the biplot in Figure 5.7. The arrows are spread out in all directions as well as the relative positions of the test products change. Thus, the conclusions from the biplot in Figure 5.12 are somewhat different than the conclusions from the biplot in Figure 5.7.

Thus, it appear that the results for the centered product specific d-prime values are affected by the exclusion of test product H.
Figure 5.11: The biplot for the non-centered d-prime values where test product H has been omitted.
Figure 5.12: The biplot for the centered d-prime values where test product H has been omitted.
5.3.2.4 Excluding the attribute Silky

A different approach than imputing a value or omitting test product H, is to exclude Silky from the analysis. The biplot from doing the PCA using the non-centered d-prime values without Silky is seen in Figure 5.13. The biplot in Figure 5.13 is tilted a bit towards the right compared to the biplot in Figure 5.9. Obviously, Silky is not in the biplot, and the arrows for the remaining attributes are a bit further apart. However, the conclusions remain the same.

The biplot from doing the PCA using the centered d-prime values without Silky is seen in Figure 5.14. The biplot in Figure 5.14 is flipped over the x-axis compared to the biplot in Figure 5.7. Obviously, Silky is not in the biplot. For the remaining attributes the positioning of the arrows is the same. Furthermore, the positions of the test products are the same, thus the conclusions remain the same.
5.3 Outliers and their influence

Figure 5.13: The biplot for the non-centered d-prime values where the attribute Silky has been omitted.
Figure 5.14: The biplot for the centered d-prime values where the attribute Silky has been omitted.
5.4 Interpretation of the assessor specific d-prime values

5.3.3 Summary of the findings

All in all, the different approaches lead to similar results, except for the exclusion of test product H for the centered d-prime values obtained from the Thurstonian mixed model.

5.4 Interpretation of the assessor specific d-prime values

The origin, in Figure 5.1, can be thought of as the consensus (μ̂j for all j), corresponding to the bmj values being equal to 0 for all j. Thus, assessors close to 0 are answering as the consensus. Hence, it appears that the assessors 12, 15 and 20 with respect to Greasy (initially and after five minutes) as well as Sticky (five minutes) are answering as the consensus. With respect to Silky after five minutes, it appears that the assessors 7 and 8 are having the largest proportions of times test products were chosen. This is in fact in alignment with the data listed in Table 5.1. However, the assessors furthest to the right in Figure 5.1 are not the assessors with the lowest proportions. Assessors 4, 6, 9 and 19 have proportions of 0, whereas 1 has a proportion of 6%. But assessor 1 is further to the right than assessor 4. Thus, more work is needed to better understand how to interpret the assessor specific d-prime values in Figure 5.1.

Regarding Figure 5.2 it is difficult how to interpret the values. Considering the total proportions for the assessors, it is seen that the assessors with the highest proportions are placed towards the right. Furthermore, the assessors with the smallest proportions are placed towards the left. The order is not completely strict, but there seems to be a pattern. A reason for this discrepancy might be related to the design of the data; not all assessors evaluated all of the test products. Further research is needed to get a better understanding of the interpretation of the assessor specific d-prime values in Figure 5.2.

5.5 Concluding remarks

Some remarks are in order regarding Figures 3 and 4 in Linander et al. (2018b). Our conclusions in Linander et al. (2018b) are based on our understanding from the preliminary investigation illustrated in Linander et al. (2018b). In this sec-
tion I will address some aspects that are not yet fully investigated and in need of more research.

Figure 3 is a biplot of a PCA where the values not are centered, and there is a need for more investigations to fully understand the possible interpretations from such figures. It appears that PC1 is capturing that Greasy evaluated after zero and five minutes are the only attributes with all negative d-prime values. This information can also be obtained by looking at the d-prime values in Table 2 in Linander et al. (2018b). Thus, biplots of other principal components might give important and informative information that cannot be detected in the figure considering principal components 1 and 2.

Figure 4 is a biplot of the assessor specific d-prime values. If the data had been balanced, such that all assessors evaluate all of the test products it probably would be possible to interpret the biplot with respect to the proportions of the test products. However, all assessors did not evaluate all products. Thus, it might be reflected in Figure 4 that some assessors have evaluated products that are silkier than other products and thereby more prone to have high values with respect to silkiness. For such data where the assessors did not evaluate all of the test products it could be investigated whether it would be possible to find patterns in the biplot regarding the assessor specific d-prime values where the information about products have been removed.
Chapter 6

Asking an additional question in the binary paired comparison

The test protocol considered previously in this thesis is the binary paired comparison. In this chapter the binary paired comparison with an additional question is considered. It is of interest to investigate if more information is obtained asking the assessors an additional question. Compared to the information that is obtained from the binary paired comparison. The data used in this chapter is from an existing discrimination study, provided by Unilever.

6.1 Data structure

In the binary paired comparison, an assessor is choosing the sample with the strongest intensity of the attribute in question. In this chapter, an extension of the binary paired comparison is considered. Each assessor is getting a two-step task. First an assessor is asked to choose the sample, of the two samples, with the strongest intensity of the attribute in question. Subsequently, an assessor has to quantify the difference between the two samples. An assessor is to rate
the magnitude of the difference between the two samples using a 5-point scale. The categories range from 'not different' to 'extremely different'. A response is recorded as 0 when the chosen category is 'not different'. For the remaining four categories the response is recorded as a positive value, when an assessor is choosing a test product and as a negative value when an assessor is choosing the control. An overview of the data values is seen in Table 6.1.

Table 6.1: An overview of how the response is defined.

<table>
<thead>
<tr>
<th>Data value</th>
<th>Chosen Product</th>
<th>Chosen category</th>
</tr>
</thead>
<tbody>
<tr>
<td>-4</td>
<td>Control</td>
<td>Extremely different</td>
</tr>
<tr>
<td>-3</td>
<td>Control</td>
<td>Very different</td>
</tr>
<tr>
<td>-2</td>
<td>Control</td>
<td>Moderately different</td>
</tr>
<tr>
<td>-1</td>
<td>Control</td>
<td>Slightly different</td>
</tr>
<tr>
<td>0</td>
<td>Control/test product</td>
<td>Not different</td>
</tr>
<tr>
<td>1</td>
<td>Test product</td>
<td>Slightly different</td>
</tr>
<tr>
<td>2</td>
<td>Test product</td>
<td>Moderately different</td>
</tr>
<tr>
<td>3</td>
<td>Test product</td>
<td>Very different</td>
</tr>
<tr>
<td>4</td>
<td>Test product</td>
<td>Extremely different</td>
</tr>
</tbody>
</table>

These data will be analyzed by two approaches. One is to consider the values as quantitative and the other is to consider the values as ordered values.

6.2 Considering data as ordered values

When considering the data as ordered values, the models that are used to model such data are known as the cumulative linear models (CLMs) and cumulative linear mixed models (CLMMs) (see e.g. Agresti (2013) and Agresti (2015)). CLMs and CLMMs are used widely in many applications, including sensory science, Christensen et al. (2012) and Christensen and Brockhoff (2013). The difference between a CLM and a CLMM is whether random effects are included in the model or not. When random effects are included the model is a mixed model and thus a CLMM. The model that will be considered in this chapter is a CLMM, since as for the analysis of the binary paired comparison assessors are included as a random effect.

Let $Y_{ijk}$ be the response variable that can fall in the nine categories $-4, -3, \ldots, 3, 4$.
and let
\[ \pi_{hij} = P(Y_{ijk} = h), \quad h = -4, -3, \ldots, 3, 4 \] (6.1)
be the probability that the \( ijk \)th observation falls in the \( h \)th category, which is independent of sessions.

The CLMM with a fixed effect of test products and random effects of assessor as well as the assessor-by-product interaction, reads:
\[ g(P(Y_{ijk} \leq h)) = \theta_h - \alpha_i - b_j - d_{ij} \] (6.2)
where \( i = 1, \ldots, 8 \) represents test products, \( j = 1, \ldots, n_i \) represents the assessors that evaluated the \( i \)th test product, \( k = 1, 2 \) represents sessions and \( h = -4, -3, \ldots, 3, 4 \) represents the categories. Furthermore, \( b_j \sim \mathcal{N}(0, \sigma_b^2) \) are the random effect of assessor being independent for all \( j \), and \( d_{ij} \sim \mathcal{N}(0, \sigma_d^2) \) are the random assessor-by-product interaction being independent for all \( i \) and \( j \).

It is of interest to investigate whether the model can be simplified or not. Firstly, the hypothesis test of a significant assessor-by-product interaction is considered. The main effects of test products and assessors are nested within the assessor-by-product interaction and the hypothesis test of these depend on whether the assessor-by-product interaction is significant or not. The hypothesis test of a random effect is regarding the variance parameter. More specifically, the hypothesis test is considering whether the variance parameter equals zero or is greater than zero:
\[ H_0 : \sigma_d^2 = 0 \quad H_1 : \sigma_d^2 > 0 \] (6.3)
where the alternative hypothesis is one-sided, since a variance is non-negative.

The likelihood ratio test statistic is Chi-squared distributed with 1 degree of freedom. For further details regarding this test see Christensen and Brockhoff (2013).

The test of the likelihood ratio test for the assessor-by-product interaction is seen in Figure 6.1.

The attributes Thickness and Greasy (initially and after five minutes) have significant assessor-by-product interactions. Thus, the differences between the assessors depend on the test products for these attributes.

When doing the test of the main effects of assessor and test products, the attributes with significant assessor-by-product interactions will be omitted. The
Figure 6.1: Likelihood Ratio Test statistics for the test of the assessor-by-product interaction using the ordinal approach. The vertical lines are critical values for the corresponding Chi-squared distribution; the 0.05 critical value (full line), the 0.01 critical value (dashed line) and the 0.001 critical value (dotted line). The symbol shows the size of the corresponding p-value; a p-value that is less than 0.001 (square), a p-value between 0.001 and 0.01 (triangle), a p-value between 0.01 and 0.05 (plus) or a p-value larger than 0.05 (dot).
reasoning for doing this, is that it is not clear how a test of a main effect should be defined when it is nested in a significant interaction.

The likelihood ratio test statistics for the assessor main effect as well as the product main effect are, for the attributes with non-significant assessor-by-product interactions, seen in Figure 6.2.

Absorption is the only attribute with a non-significant effect of the assessors. The remaining attributes have significant assessor main effects as well as product main effects.

6.3 Considering data as quantitative

The quantitative approach is considering the data without the ordering that is used for the ordinal approach.

The model reads:

$$ Y_{ijk} = \mu + \alpha_i + b_j + d_{ij} + \epsilon_{ijk} $$ (6.4)

where $\epsilon_{ijk}$ is the only term that has not yet been introduced. The remaining terms are defined as previously. $\epsilon_{ijk} \sim N(0, \sigma^2)$ are the residuals which are independent for all $i, j$ and $k$.

The likelihood ratio test for the assessor-by-product interaction, using model (6.4), is seen in Figure 6.3.

The attributes Greasy (initially and after five minutes) and Thickness have significant assessor-by-product interactions. The remaining five attributes have non-significant assessor-by-product interactions. This result is similar to the result obtained considering the values as ordered values, which was shown in Figure 6.1.

6.4 Comparison of likelihood ratio test statistics

The values of the likelihood ratio test statistics are compared for the binary paired comparison, the ordinal approach as well as the quantitative approach.
Figure 6.2: Likelihood Ratio Test statistics for the test of the assessor main effect and the product main effect for the ordinal approach. The vertical lines are critical values for the corresponding Chi-squared distribution; the 0.05 critical value (full line), the 0.01 critical value (dashed line) and the 0.001 critical value (dotted line). The symbol shows the size of the corresponding p-value; a p-value that is less than 0.001 (square), a p-value between 0.001 and 0.01 (triangle), a p-value between 0.01 and 0.05 (plus) or a p-value larger than 0.05 (dot).
Figure 6.3: Likelihood Ratio Test statistics for the test of the assessor-by-product interaction for the quantitative approach. The symbol shows the size of the corresponding p-value; a p-value that is less than 0.001 (square), a p-value between 0.001 and 0.01 (triangle), a p-value between 0.01 and 0.05 (plus) or a p-value larger than 0.05 (dot).
Figure 6.4: Likelihood Ratio Test statistics for the test of the assessor-by-product interaction as well as the main effects of assessor and products.

The values of the likelihood ratio test statistics for the different approaches are seen in Figure 6.4. The values for the test of the assessor-by-product interaction do not vary much. Furthermore, all of these values are almost equal to 0. For the test of the main effects, the values obtained using the binary paired comparison are the smallest values. The values obtained from the ordinal and quantitative are similar for most of the tests.

From this initial investigation, it appears that the binary paired comparison is less sensitive than the ordinal as well as the quantitative approach. Further research is needed to fully understand the impact of the different approaches.
6.5 Concluding remarks

Future work is needed to get a better understanding of which model is the most appropriate to use. Considering this additional question, in my opinion it is not obvious whether it is more appropriate to consider the values as ordered or not. I believe that an absolute value of four is more than that of three since this means that the difference is bigger. However, a question that could be important is whether it is possible to determine that a value in favor of the control is lower than a value favoring the test product. Maybe the answer is to think carefully about the coding of the observations.

When considering the values as ordered values an advantage is that it might be possible to develop a Thurstonian model, similar to the model for the 2-AC [Christensen et al. (2012)], providing d-prime interpretations of the estimates.
Asking an additional question in the binary paired comparison
Chapter 7

Comparison of d-prime values

This chapter is concerned with comparing multiple d-prime values. It has previously been proposed by Bi et al. (1997) how to compare multiple d-prime values. The approach in Bi et al. (1997) is to use a Wald-type test. A disadvantage of this method is that when data include observations that lead to parameter estimates on the boundary of the parameter space the test statistic is not well-defined. In Section 7.3 it is described how these situations affect the Wald-type test statistic.

In Linander et al. (2018) we suggest another test for the comparison of multiple d-prime values. This test is based on likelihood theory and does not have the deficiency of being unable to cope with parameter estimates on the boundary of the parameter space. Furthermore, the power of the likelihood test is higher than the power of the Wald-type test. Investigations of the power is considered in Section 7.4.

7.1 The d-prime values

The d-prime values considered are coming from independent discrimination studies using one of the so-called simple-binomial discrimination test pro-
Comparison of d-prime values

tocols; the 2-AFC, 3-AFC, duo-trio, triangle and the tetrad. By definition, these test protocol lead to binomially distributed data. Let $X_i$ be the number of correct answers out of a total of $N_i$ answers for the $i$th sensory discrimination study. Now:

$$X_i \sim \text{Binomial}(p_i, N_i)$$

(7.1)

where $p_i$ is the probability of a correct answer in the $i$th sensory discrimination test.

7.2 Comparing multiple d-prime values

There are many situations in which it is desirable to compare multiple d-prime values. An example is to compare the performance of sensory panels in different laboratories [Sauvageot et al. 2012]. When comparing multiple d-prime values it is investigated if all the d-prime values are identical. Let $d'_1, \ldots, d'_k$ be d-prime values obtained from $k$ independent simple-binomial test protocols. The hypotheses regarding the test of multiple d-prime values being equal, which is referred to as the any-differences hypothesis, are given as:

$$H_0 : d'_1 = d'_2 = \cdots = d'_k$$

$$H_1 : d'_i \neq d'_j \text{ for at least one pair } (i, j) \text{ where } i \neq j$$

7.2.1 Using a Wald-type test

Bi et al. [1997] proposed to use a Wald-type test statistic when considering the any-differences hypothesis. The test of comparing multiple d-prime values is given by this Wald-like test statistic:

$$X^2_{Wald} = \frac{(d'_1 - d'_e)^2}{se(d'_1)^2} + \frac{(d'_2 - d'_e)^2}{se(d'_2)^2} + \cdots + \frac{(d'_k - d'_e)^2}{se(d'_k)^2}$$

$$= \sum_{i=1}^{k} \left( \frac{d'_i - d'_e}{se(d'_i)} \right)^2$$

(7.2)
7.2 Comparing multiple d-prime values

where \( d'_{e} \) is the common d-prime value for the d-prime values under the null hypothesis which is estimated by:

\[
\begin{align*}
    d'_{e} &= \frac{d'_{1}^{2} + d'_{2}^{2} + \cdots + d'_{k}^{2}}{se(d'_{1})^{2} + se(d'_{2})^{2} + \cdots + se(d'_{k})^{2}} \\
    &= \frac{\sum_{i=1}^{k} d'_{i}^{2}}{\sum_{i=1}^{k} se(d'_{i})^{2}}
\end{align*}
\]

where \( se(d'_{i}) \) is the standard error of the \( i \)th d-prime value.
The test statistic follows a Chi-square distribution with \( k - 1 \) degrees of freedom (Bi et al. (1997)):

\[
    X_{Wald}^{2} \sim \chi_{k-1}^{2}
\]

The standard error of \( d' \) can be found from the standard error of \( p_{c} \) by using that (see e.g. Pawitan (2001)):

\[
    se(g(\hat{\theta})) = se(\hat{\theta}) \left| \frac{\partial g}{\partial \hat{\theta}} \right|
\]

Therefore

\[
    se(\delta) = se(p_{c}) \left| \frac{\partial f^{-1}_{psy}(p_{c})}{\partial p_{c}} \right|
    = se(p_{c}) \frac{1}{f_{psy}'(\delta)}
\]

where \( f_{psy}'(\delta) \) is the partial derivative of \( f_{psy}(\delta) \) with respect to \( \delta \):

\[
    f_{psy}'(\delta) = \frac{\partial f_{psy}(\delta)}{\partial \delta}
\]

Thus

\[
    se(d') = se(\hat{p}_{c}) \frac{1}{f_{psy}'(d')}
\]  

\[ (7.3) \]

7.2.2 Using a likelihood test

In Linander et al. (2018) we suggest a test of the any-differences hypothesis constructed by the use of likelihood theory. This section gives a brief description
of the test for a more detailed definition see Linander et al. (2018).

Let $d'_1, \ldots, d'_k$ be d-prime values obtained from $k$ independent simple-binomial tests. Let $x_1, \ldots, x_k$ and $n_1, \ldots, n_k$ be the number of correct and total answers respectively observed in the $k$ discrimination tests. Furthermore, $m_i$ denotes the test that the $i$th d-prime value is obtained from. Due to the independence of the discrimination tests the likelihood function is the product of the individual likelihood functions. Therefore the log-likelihood function under the alternative hypothesis is given as the sum of the individual log-likelihood functions:

$$
\ell_1(d'; x, n, m) = \sum_{i=1}^{k} \log L(d'_i; x_i, n_i, m_i) \quad (7.4)
$$

where $p_i = f_{psy}(d'_i)$.

Under the null hypothesis the log-likelihood function reads:

$$
\ell_0(d'; x, n, m) = \sum_{i=1}^{k} \log L(d'_i; x_i, n_i, m_i) \quad (7.6)
$$

where $p_c = f_{psy}(d'_c)$ is the transformed probability corresponding to the common d-prime value under the null. The common d-prime value $d'_c$ is the estimated value which is found by maximum likelihood.

The likelihood ratio test statistic reads:

$$
-2 \log Q = 2 (\ell_1(d'; x, n, m) - \ell_0(d'_c; x, n, m)) \quad (7.8)
$$

### 7.3 Boundary situations

The Wald-type test given by (7.2) is not well-defined in situations where data consist of boundary situations. Boundary situations are when the parameter $p_c$ is on the boundary of its parameter space. According to (7.2) the two boundary situations are:

$$
\hat{p}_c = p_g \iff x/n < p_g \iff x < np_g \\
\hat{p}_c = x = n \quad (7.9, 7.10)
$$
7.3 Boundary situations

The standard error of $\hat{p}_c$ is given as:

$$se(\hat{p}_c) = \sqrt{\hat{p}_c(1 - \hat{p}_c)/n} \quad (7.11)$$

Thus for $x = n$ the standard error of $\hat{p}_c$ equals 0. Therefore, $se(d')$ equals 0 and the test statistic is not well-defined.

When $x/n < p_g$ challenges arise for some of the test protocols. To realize this, the partial derivatives of the psychometric functions with respect to $\delta$ are considered. First note these two relations regarding the density and distribution functions for the standard normal distribution:

$$\frac{\partial \varphi(\theta)}{\partial \theta} = -\theta \varphi(\theta) \quad (7.12)$$

and

$$\frac{\partial \Phi(\theta)}{\partial \theta} = \varphi(\theta) \quad (7.13)$$

Furthermore, the Leibniz rule can be applied:

$$\frac{\partial}{\partial x} \int_a^b f(x,t)dt = \int_a^b \frac{\partial f(x,t)}{\partial x} dt \quad (7.14)$$

when $f(x,t)$ and the partial derivative of $f(x,t)$ with respect to $x$ are continuous. Considering the psychometric functions defined in Section 2.2 $f(x,t)$ will correspond to products of $\varphi(x)$ and $\Phi(x)$ which are continuous functions and thus $f$ and its derivative are continuous.

The parameter space for $\delta$ is given as non-negative values. Thus, the partial derivatives are defined for $\delta \geq 0$.

For the 2-AFC test protocol the derivative becomes:

$$f'_{2 \text{AFC}}(\delta) = \varphi \left( \frac{\delta}{\sqrt{2}} \right) \frac{1}{\sqrt{2}} \quad (7.15)$$

where for $\delta = 0$:

$$f'_{2 \text{AFC}}(0) = \varphi(0) \frac{1}{\sqrt{2}} > 0 \quad (7.16)$$

For the 3-AFC test protocol the derivative becomes:

$$f'_{3 \text{AFC}}(\delta) = \int_{-\infty}^{\infty} (z - \delta) \varphi(z - \delta) \Phi^2(z)dz \quad (7.17)$$
Comparison of d-prime values

where for $\delta = 0$:

$$f'_{\text{AFC}}(0) = \int_{-\infty}^{\infty} z\varphi(z)\Phi^2(z)dz > 0 \quad (7.18)$$

For the duo-trio test protocol the derivative becomes:

$$f'_{\text{d-t}}(\delta) = -\varphi\left(\frac{\delta}{\sqrt{2}}\right) \frac{1}{\sqrt{2}} - \varphi\left(\frac{\delta}{\sqrt{6}}\right) \frac{1}{\sqrt{6}}$$

$$\quad + 2 \left(\varphi\left(\frac{\delta}{\sqrt{2}}\right) \frac{1}{\sqrt{2}} \Phi\left(\frac{\delta}{\sqrt{6}}\right) + \varphi\left(\frac{\delta}{\sqrt{6}}\right) \frac{1}{\sqrt{6}} \Phi\left(\frac{\delta}{\sqrt{2}}\right)\right) \quad (7.19)$$

where for $\delta = 0$:

$$f'_{\text{d-t}}(0) = -\varphi(0) \frac{1}{\sqrt{2}} - \varphi(0) \frac{1}{\sqrt{6}} + 2 \left(\varphi(0) \frac{1}{\sqrt{2}} \Phi(0) + \varphi(0) \frac{1}{\sqrt{6}} \Phi(0)\right) = 0 \quad (7.21)$$

For the triangle test protocol the derivative becomes:

$$f'_{\text{tri}}(\delta) = 2\sqrt{2/3} \int_{0}^{\infty} \left(\varphi(-z\sqrt{3} + \delta\sqrt{2/3}) - \varphi(-z\sqrt{3} - \delta\sqrt{2/3})\right) \varphi(z)dz \quad (7.22)$$

where for $\delta = 0$:

$$f'_{\text{tri}}(0) = 2\sqrt{2/3} \int_{0}^{\infty} \left(\varphi(-z\sqrt{3}) - \varphi(-z\sqrt{3})\right) \varphi(z)dz = 0 \quad (7.23)$$

For the tetrad test protocol the derivative becomes:

$$f'_{\text{tetrad}}(\delta) = -2 \int_{-\infty}^{\infty} \varphi(z) \left(2\Phi(z)(z-\delta)\varphi(z-\delta)(-1) - 2\Phi(z-\delta)(z-\delta)\varphi(z-\delta)(-1)\right)dz$$

$$\quad = -4 \int_{-\infty}^{\infty} (z-\delta)\varphi(z)\varphi(z-\delta) (\Phi(z-\delta) - \Phi(z)) dz \quad (7.24)$$

where for $\delta = 0$:

$$f'_{\text{tetrad}}(0) = -4 \int_{-\infty}^{\infty} z\varphi(z)\varphi(z) (\Phi(z) - \Phi(z)) dz = 0 \quad (7.25)$$

For three of the discrimination test protocols the derivative evaluated in zero equals 0 and for these test protocols the standard error of $\delta$ is not defined. Thus, when $x/n \leq p_g$ the Wald-type test statistic is undefined.

Boundary situations occur rather often, why this is a non-negligible deficiency of the method proposed by [Bi et al. (1997)].

The likelihood test is well-defined even when data consist of boundary situations.
7.4 Power

This section investigates the power of the likelihood test suggested in Linander et al. (2018). Furthermore, the power is compared to the Wald-type test suggested by Bi et al. (1997).

7.4.1 A simulation study

In this section a simulation study, investigating the power, is carried out. In the simulation study four different simple-binomial tests are considered; triangle, duo-trio, 2-AFC and 3-AFC. Data are simulated for $N = 10$, $N = 30$ and $N = 60$ for 12 different combinations of d-prime values. These combinations are listed in Table 7.1.

<table>
<thead>
<tr>
<th>Setup 1</th>
<th>triangle</th>
<th>d-prime values</th>
<th>duo-trio</th>
<th>2-AFC</th>
<th>3-AFC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td></td>
</tr>
<tr>
<td>Setup 2</td>
<td>2.0</td>
<td></td>
<td>2.0</td>
<td>2.0</td>
<td></td>
</tr>
<tr>
<td>Setup 3</td>
<td>3.0</td>
<td></td>
<td>3.0</td>
<td>3.0</td>
<td></td>
</tr>
<tr>
<td>Setup 4</td>
<td>1.1</td>
<td>1.7</td>
<td>2.3</td>
<td>2.9</td>
<td></td>
</tr>
<tr>
<td>Setup 5</td>
<td>1.1</td>
<td>1.8</td>
<td>2.5</td>
<td>3.2</td>
<td></td>
</tr>
<tr>
<td>Setup 6</td>
<td>1.1</td>
<td>1.9</td>
<td>2.7</td>
<td>3.5</td>
<td></td>
</tr>
<tr>
<td>Setup 7</td>
<td>1.4</td>
<td>2.0</td>
<td>2.6</td>
<td>3.2</td>
<td></td>
</tr>
<tr>
<td>Setup 8</td>
<td>1.3</td>
<td>2.0</td>
<td>2.7</td>
<td>3.4</td>
<td></td>
</tr>
<tr>
<td>Setup 9</td>
<td>1.2</td>
<td>2.0</td>
<td>2.8</td>
<td>3.6</td>
<td></td>
</tr>
<tr>
<td>Setup 10</td>
<td>1.8</td>
<td>2.4</td>
<td>3.0</td>
<td>3.6</td>
<td></td>
</tr>
<tr>
<td>Setup 11</td>
<td>1.6</td>
<td>2.3</td>
<td>3.0</td>
<td>3.7</td>
<td></td>
</tr>
<tr>
<td>Setup 12</td>
<td>1.4</td>
<td>2.2</td>
<td>3.0</td>
<td>3.8</td>
<td></td>
</tr>
</tbody>
</table>

The power of the Wald-type test as well the power of the likelihood test are calculated based on 10,000 simulations and listed in Table 7.2. Generally, the power of the likelihood test is much better than the power for the Wald-type test.
Table 7.2: Estimated power for the any-differences hypothesis test.

<table>
<thead>
<tr>
<th></th>
<th>$N = 10$</th>
<th></th>
<th>$N = 30$</th>
<th></th>
<th>$N = 60$</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>likelihood</td>
<td>wald</td>
<td>likelihood</td>
<td>wald</td>
<td>likelihood</td>
<td>wald</td>
</tr>
<tr>
<td>Setup 1</td>
<td>5 1</td>
<td></td>
<td>4 3</td>
<td></td>
<td>4 3</td>
<td></td>
</tr>
<tr>
<td>Setup 2</td>
<td>6 0</td>
<td></td>
<td>6 7</td>
<td></td>
<td>6 15</td>
<td></td>
</tr>
<tr>
<td>Setup 3</td>
<td>3 0</td>
<td></td>
<td>4 1</td>
<td></td>
<td>6 7</td>
<td></td>
</tr>
<tr>
<td>Setup 4</td>
<td>23 0</td>
<td></td>
<td>70 24</td>
<td></td>
<td>94 77</td>
<td></td>
</tr>
<tr>
<td>Setup 5</td>
<td>26 0</td>
<td></td>
<td>80 18</td>
<td></td>
<td>98 61</td>
<td></td>
</tr>
<tr>
<td>Setup 6</td>
<td>30 0</td>
<td></td>
<td>87 10</td>
<td></td>
<td>99 40</td>
<td></td>
</tr>
<tr>
<td>Setup 7</td>
<td>18 0</td>
<td></td>
<td>67 14</td>
<td></td>
<td>92 61</td>
<td></td>
</tr>
<tr>
<td>Setup 8</td>
<td>22 0</td>
<td></td>
<td>79 11</td>
<td></td>
<td>97 47</td>
<td></td>
</tr>
<tr>
<td>Setup 9</td>
<td>27 0</td>
<td></td>
<td>86 8</td>
<td></td>
<td>99 34</td>
<td></td>
</tr>
<tr>
<td>Setup 10</td>
<td>10 0</td>
<td></td>
<td>51 3</td>
<td></td>
<td>86 27</td>
<td></td>
</tr>
<tr>
<td>Setup 11</td>
<td>15 0</td>
<td></td>
<td>67 3</td>
<td></td>
<td>95 25</td>
<td></td>
</tr>
<tr>
<td>Setup 12</td>
<td>21 0</td>
<td></td>
<td>80 4</td>
<td></td>
<td>98 21</td>
<td></td>
</tr>
</tbody>
</table>

7.5 Comparison of multiple d-prime values using sensR

In this section, it is illustrated how tests of the any-differences hypothesis is conducted using the sensR package.

The test of the any-differences is done by using the dprime_compare function where the option statistic defines whether the likelihood test or the Wald-type test is conducted. With the estim option it is selected whether to estimate the common d-prime under the null hypothesis by a weighted average or the maximum likelihood. Thus, a test of the any-differences hypothesis for the likelihood test using the maximum likelihood estimator can be made by:

```r
dprime_compare(correct, total, protocol, statistic = "likelihood", estim = "ML")
```
Bibliography


Appendix A

Individual differences in replicated multi-product experiments with Thurstonian models for binary paired comparison data

Individual differences in replicated multi-product experiments with Thurstonian mixed models for binary paired comparison data

Christine Borgen Linander, Rune Haubo Bojesen Christensen, Graham Cleaver, Per Bruun Brockhoff

Abstract

Often sensory discrimination tests are performed with replications for the assessors. In this paper, we suggest a new way of analyzing data from a discrimination study. The model suggested in this paper is a Thurstonian mixed model, in which the variation from the assessors is modelled as a random effect in a generalized linear mixed model. The setting is a multi-product discrimination study with a binary paired comparison. This model makes it possible to embed the analyses of products into one analysis rather than having to do an analysis for each product separately. In addition, it is possible to embed the model into the Thurstonian framework obtaining d-prime interpretations of the estimates. Furthermore, it is possible to extract information about the assessors, even across the products. More specifically, assessor specific d-prime estimates are obtained providing a way to monitor the panel. These estimates are interesting because they make it possible to investigate if some assessors are assessing differently.

Keywords: Thurstonian modelling, binary paired comparison, assessor information, multi-product setting, Generalized Linear Mixed Model

1. Introduction

It is a recurrent scenario that discrimination tests are conducted with replications for the assessors (Ennis (2012)). Thus, it is important to handle the possible differences between the assessors correctly. Suggestions in the literature are e.g. the so-called beta-binomial models as well as corrected beta-binomial models. In this paper, we suggest a new way of modelling the potential assessor differences.

It has been described in the literature how Thurstonian modelling is the preferred approach to quantify the difference between products (e.g. Ennis (1993), Ennis & Jesionka (2011), Næs et al. (2010)). In this paper, we follow this recommendation, thus we will consider the analyses on the d-prime scale. Hence, we will be considering Thurstonian models.

This work is part of an overall objective of aligning Thurstonian d’ analysis with the modern world of statistical modelling. Brockhoff & Christensen (2010) show how a Thurstonian model for sensory discrimination tests can be seen as a Generalized Linear Model (GLM).

The way we suggest to handle the possible assessor differences is by adding assessor as a random effect to a GLM. This results in a Generalized Linear Mixed Model (GLMM), which is a way to analyze categorical data like binomial data. Categorical data analysis is a common well-known framework, which is used in many applications. The book by Agresti (2013) gives a thorough description of categorical data analysis.

The setting that is considered is a multi-product setting giving the possibility to investigate for possible assessor-by-product interactions. In discrimination testing test protocols exist where there is no correct answer. The test protocol that is considered in this paper is the binary paired comparison. This allows for the d-prime values to be positive as well as negative. In Section 2.2 we will go into details about the Thurstonian model for this setup.

We believe that adding this level of details to the models give us valuable insights about data that would have been undetected otherwise. Not only do we get the d-prime interpretation of our parameters, in addition, we
gain information about the assessors. Moreover, it is possible to embed the analysis of the products into one analysis instead of having to do an analysis for each product separately. Furthermore, the replications of the assessors are handled correctly when testing for a significant effect of products.

We consider figures of the assessor specific d-prime values, giving an opportunity to get insights about the assessors, which is only possible due to the level of details in the model. From these figures, it will be possible to gain knowledge about how the panel performs. Additionally, these figures make it possible to realize whether some assessors are assessing differently than the rest of the panel. Furthermore, since no correct answer exists it will be possible to detect if the panel is in agreement about which sample had the strongest intensity.

In the remainder of this section a discrimination study is described. In Section 2 we define the methodology we suggest. We will throughout Section 2 illustrate the methodology by using the study described in section 1.1. At the end of this paper we have a discussion in Section 3.

1.1. The discrimination study

In this section an existing discrimination study is explained. We use this as an ongoing example throughout this paper to illustrate the methodology we introduce in Section 2.

The overall aim of this study was to find a new product that has some of the same characteristics as an existing product.

In this study, the assessors were comparing different test products to a control product. A sample of a test product as well as a sample of the control product were applied to an assessor’s own skin. The assessor had to choose the sample with the strongest intensity of the attribute in question.

The organization of evaluations of the test products is illustrated in Figure 1. In one day, assessors evaluated two test products, where each assessor evaluated each test product twice in two different sessions.

In total eight test products (denoted by A, B, C, D, E, F, G and H) were compared to the control product.

The assessors that participated in the assessments of the test products were the same from day to day. Some assessors were not able to participate in the assessments for some days. If an assessor assessed the test products within a day, the assessor participated in both sessions carried out that day. For an assessor to be included in the analysis, the assessor had to participate in at least 50% of the assessments. 25 assessors (denoted by 1, . . . , 25) made enough assessments to be included in the analysis (two assessors did not make enough assessments).

The assessors assessed multiple attributes and their characteristics are listed in Table 1. The assessors evaluated five different attributes all of which were evaluated immediately after application of the samples. In addition, three of these attributes were re-evaluated after five minutes. Thus, in total eight attributes were assessed by the assessors.

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Evaluated after 0 min</th>
<th>Evaluated after 5 min</th>
<th>Tactile</th>
<th>Visual</th>
</tr>
</thead>
<tbody>
<tr>
<td>Thickness</td>
<td>✓</td>
<td>✓</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Absorption</td>
<td>✓</td>
<td>✓</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Greasy</td>
<td>✓</td>
<td>✓</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Sticky</td>
<td>✓</td>
<td>✓</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Silky</td>
<td>✓</td>
<td>✓</td>
<td>✔</td>
<td>✔</td>
</tr>
</tbody>
</table>

Table 1: Overview of the attributes.

Figure 1: Organization of days, sessions, assessors as well as test products.

2. Methodology

In this section we explain the methodology as well as applying this methodology to the data described in Section 1.1. When analyzing such data a model is fitted
for each attribute at a time, thus results are obtained for each attribute separately.

2.1. Explorative investigation of data

A way to gain information about the data obtained from a sensory discrimination study, is to examine proportions. In this section an explorative investigation of the data from Section 1.1 is given.

One aim of analyzing the data is to gain knowledge about which (if any) of the test products that have the characteristics that are desired for this type of product. To gain information about which test product that has the most interesting sensory characteristics we can look at proportions. The proportions, the number of times the test product was chosen, are aggregated over assessors as well as sessions. These proportions (in percentages) are given in Table 2.

An important sensory characteristic is that the test product should be at least as silky as the control product. The most silky test product is having the highest value of the proportions. Therefore, test product D is the most promising test product with respect to silkiness, initially and after five minutes. When a test product was chosen more often than the control, the proportion is larger than 50%. Thus, for a test product to be silkier than the control the proportion must exceed 50%. The percentages for D exceed 50%, thus test product D is silkier than the control.

2.2. d-prime values for test products

When considering the proportions from the previous section an overview of data is given. However, it can be rather difficult to comprehend how similar (or different) the products are. Thus, the proportions are transformed into d-prime values for a better comparison of the products. We will in this section find the d-prime values for the test products to express the sensory difference between the test products and the control for the eight attributes.

These d-prime values are found by transforming the proportion of times the test product was chosen for each attribute via the inverse of the so-called psychometric function.

To develop the Thurstonian model for our setting let $C$ and $T$ denote the distribution of the sensory intensity for the control product and a test product respectively. We assume that $C$ and $T$ are independent and that:

$$C \sim N(\mu_c, \sigma^2) \quad \text{and} \quad T \sim N(\mu_t, \sigma^2).$$

The underlying Thurstonian relative sensory difference $\delta$ is the difference in means scaled by the common standard deviation:

$$\delta = \frac{\mu_t - \mu_c}{\sigma}.$$  

An advantage of using $\delta$: the measure for sensory differences is that $\delta$ does not depend on the discrimination test protocol, see e.g. Ennis (1993)

The psychometric function $f_{psy}$ can for this setting be defined as the probability that the test product is chosen which is the probability of the test product having a larger sensory intensity than the control:

$$f_{psy}(\delta) = P(T > C) = \Phi\left(\frac{\delta}{\sqrt{2}}\right) = p \quad (1)$$

where $\Phi$ is the cumulative distribution function for the standard normal distribution and $p$ is the probability that the test product is chosen over the control product. The reader is referred to Appendix A for the details of the derivation of the psychometric function in (1).

$d'$, the estimate of $\delta$, is the estimated sensory difference between the test product and the control product. d-prime values can be computed using the inverse of the psychometric function:

$$f_{psy}^{-1}(p) = \Phi^{-1}(p) \sqrt{2} \equiv d' \quad (2)$$

A d-prime value for each comparison of a test product to the control is obtained.

The psychometric function given in (1) is illustrated in Figure 2. When $p = 0.5$, corresponding to a d-prime value of 0, the assessors chose the test product half the time. Thus, there is no perceivable difference between the test product and the control product. When $p > 0.5$ the d-prime value is positive and the psychometric function is the same as for the 2-AFC protocol. Additionally, for all d-prime values the setting corresponds to the paired comparison protocol, which in some situations also is the paired preference (Christensen et al. (2012)). A positive or negative d-prime value corresponds to the test product having the strongest or weakest intensity of the attribute in question.

The d-prime values for the test products, for the eight attributes, are shown in Table 3. As expected from the values of the proportions, D is the only test product with a positive d-prime for Silky both evaluated initially and after five minutes.

2.3. Generalized Linear Models

The d-prime values from Section 2.2 are calculated from the data without other assumptions than those regarding the underlying distributions for the sensory intensities. Another way to gain information about the
Table 2: The number of times a test product was chosen for the eight attributes in percentages. The number of evaluations for the test products range from 40 to 46.

<table>
<thead>
<tr>
<th>Test Product</th>
<th>Sticky 0 min</th>
<th>Sticky 5 min</th>
<th>Greasy 0 min</th>
<th>Greasy 5 min</th>
<th>Silky 0 min</th>
<th>Silky 5 min</th>
<th>Thickness 0 min</th>
<th>Thickness 5 min</th>
<th>Absorption 0 min</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>65.2</td>
<td>19.6</td>
<td>26.1</td>
<td>13.0</td>
<td>23.9</td>
<td>21.7</td>
<td>97.8</td>
<td>17.4</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>20.0</td>
<td>52.5</td>
<td>15.0</td>
<td>40.0</td>
<td>20.0</td>
<td>12.5</td>
<td>65.0</td>
<td>85.0</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>5.0</td>
<td>10.0</td>
<td>17.5</td>
<td>25.0</td>
<td>50.0</td>
<td>35.0</td>
<td>47.5</td>
<td>62.5</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>2.3</td>
<td>2.3</td>
<td>34.1</td>
<td>34.1</td>
<td>93.2</td>
<td>70.5</td>
<td>50.0</td>
<td>72.7</td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>12.5</td>
<td>10.0</td>
<td>7.5</td>
<td>15.0</td>
<td>27.5</td>
<td>25.0</td>
<td>50.0</td>
<td>70.0</td>
<td></td>
</tr>
<tr>
<td>F</td>
<td>41.3</td>
<td>13.0</td>
<td>10.9</td>
<td>21.7</td>
<td>34.8</td>
<td>43.5</td>
<td>47.8</td>
<td>37.0</td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>67.5</td>
<td>50.0</td>
<td>17.5</td>
<td>25.0</td>
<td>12.5</td>
<td>20.0</td>
<td>95.0</td>
<td>75.0</td>
<td></td>
</tr>
<tr>
<td>H</td>
<td>90.9</td>
<td>70.5</td>
<td>27.3</td>
<td>13.6</td>
<td>0.0</td>
<td>6.8</td>
<td>22.7</td>
<td>13.6</td>
<td></td>
</tr>
</tbody>
</table>

Table 3: d-prime values found by using the psychometric function on the proportions.

<table>
<thead>
<tr>
<th>Test Product</th>
<th>Sticky 0 min</th>
<th>Sticky 5 min</th>
<th>Greasy 0 min</th>
<th>Greasy 5 min</th>
<th>Silky 0 min</th>
<th>Silky 5 min</th>
<th>Thickness 0 min</th>
<th>Thickness 5 min</th>
<th>Absorption 0 min</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.55</td>
<td>-1.21</td>
<td>-0.91</td>
<td>-1.59</td>
<td>-1.00</td>
<td>-1.10</td>
<td>2.86</td>
<td>-1.33</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>-1.19</td>
<td>0.09</td>
<td>-1.47</td>
<td>-0.36</td>
<td>-1.19</td>
<td>-1.63</td>
<td>0.54</td>
<td>1.47</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>-2.33</td>
<td>-1.81</td>
<td>-1.32</td>
<td>-0.95</td>
<td>0.00</td>
<td>-0.54</td>
<td>-0.09</td>
<td>0.45</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>-2.83</td>
<td>-2.83</td>
<td>-0.58</td>
<td>-0.58</td>
<td>2.11</td>
<td>0.76</td>
<td>0.00</td>
<td>0.86</td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>-1.63</td>
<td>-1.81</td>
<td>-2.04</td>
<td>-1.47</td>
<td>-0.85</td>
<td>-0.95</td>
<td>0.00</td>
<td>0.74</td>
<td></td>
</tr>
<tr>
<td>F</td>
<td>-0.31</td>
<td>-1.59</td>
<td>-1.74</td>
<td>-1.10</td>
<td>-0.55</td>
<td>-0.23</td>
<td>-0.08</td>
<td>-0.47</td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>0.64</td>
<td>0.00</td>
<td>-1.32</td>
<td>-0.95</td>
<td>-1.63</td>
<td>-1.19</td>
<td>2.33</td>
<td>0.95</td>
<td></td>
</tr>
<tr>
<td>H</td>
<td>1.89</td>
<td>0.76</td>
<td>-0.86</td>
<td>-1.55</td>
<td>-Inf</td>
<td>-2.11</td>
<td>-1.06</td>
<td>-1.55</td>
<td></td>
</tr>
</tbody>
</table>

data is by imposing a model to the probabilities of a test product being chosen. The observations from the binary paired comparison test protocol are binomially distributed:

\[ Y_{ijk} \sim \text{binomial}(p_{ij}, 1) \]

where \( i = 1, \ldots, l \) represents the test products, \( j = 1, \ldots, n_i \) represents the assessors for the \( i \)th test product and \( k = 1, \ldots, r \) \((r = 2 \text{ and } l = 8 \text{ for the discrimination study used in this paper})\) represents the sessions carried out on the same day. In addition, we assume that \( p_{ij} \) is the probability of the \( j \)th assessor choosing the \( i \)th test product, is independent of the sessions:

\[ p_{ij} = P(Y_{ijk} = 1) \]

where \( f_{psy} \) is the psychometric function given in (1). Thus, the psychometric function is describing how the parameters \( \mu \) and \( \alpha_i \) are relating to the probability \( p_{ij} \). According to Brockhoff & Christensen (2010) this way of writing a Thurstonian model is a Generalized Linear Model and we refer the reader to Brockhoff & Christensen (2010) for further details on this matter. The parameter \( \mu \) is the average difference between test products and the control product. \( \alpha_i \) is the difference for the \( i \)th test product to the average product-difference \( \mu \). Or put differently, \( \alpha_i \) is the magnitude of how much the \( i \)th test product is different from the average product-difference. Thus, the relation between the underlying sensory difference \( \delta_i \) for the \( i \)th test product to the control product and the model parameters is:

\[ \delta_i = \mu + \alpha_i \]  

(4)

The d-prime value \( d' \), the estimate of \( \delta_i \) given in (4), is the estimated sensory difference between the \( i \)th test product and the control product. These estimates can be found using standard statistical software fitting Generalized Linear Models with the probit link. The d-prime values obtained from using model (3) are listed in Table
3. These values are also the values obtained by transforming the proportions in Section 2.2. Thus, analyzing data with a GLM gives the same d-prime values as transforming the proportions. An advantage of using the GLM approach is that the statistical software provides additional information to the d-prime estimates e.g. standard errors and p-values. Furthermore, realizing that a GLM is another way to write the transformation of the proportions, makes it possible to consider other ways of defining the linear model structure.

2.4. Generalized Linear Mixed Model as a Thurstonian Mixed Model

It was, in the previous section, established that the d-prime values are obtainable using a generalized linear model. In this section, the linear model structure is extended to include a random effect. For other applications, an extension of a GLM to include a random effect is known as a Generalized Linear Mixed Model (GLMM). In this section the linear model structure is extended by adding the effect of the assessors as a random component. Thus, this section is considering a Thurstonian Mixed Model with a fixed effect of test products as well as a random effect of the assessors. The linear model structure for this model reads:

\[ p_{ij} = f_{ij}(\mu + \alpha_i + b_j) \]  (5)

where \(i, j, \mu\) and \(\alpha_i\) are defined as described in Section 2.3. Furthermore, \(b_j \sim N(0, \sigma^2_b)\) is the random effect of the \(j\)th assessor which are independent for all \(j\). \(b_j\) is the difference for the \(j\)th assessor to the average product-difference \(\mu\) on the \(d'\)-scale. Thus, the sensory difference, on the d-prime scale, between the test products and the control product for the \(j\)th assessor is \(\hat{b}_j = \mu + b_j\).

The relation between the product d-prime value \(\delta_i\) and the model parameters is not affected by the random effect of the assessors. This is because the value of \(\delta_i\) is for an average assessor, thus \(b_j\) equals 0, hence the relation is the same as in equation (4). The size of \(d'_i\), the estimate of \(\delta_i\), depends on how the linear model structure is defined. The values of \(d'_i\) using the model structures defined in (3) and (5) for Silikey after 5 minutes are shown in Figure 3.

2.5. Extending the Thurstonian Mixed Model

The model from Section 2.4 considers the main effect of products and assessors. In this section, the Thurstonian Mixed Model given in (5) is extended, such that the interaction of the products and assessors is included in the linear model structure of the probabilities. The

Figure 2: The psychometric function for the study. For \(p > 0.5\)(solid) and \(p < 0.5\)(dashed).

Figure 3: The d-prime values for test products for Silikey after 5 minutes for model (3) (circles) and model (5) (triangles). Generally, the estimates are further away from zero when the effect of the assessors is taken into account.
assessor-by-product interaction is a random effect because assessor is included as a random effect. Thus, this section is considering a Thurstonian Mixed Model with a fixed effect of products as well as random effects of the assessors and the assessor-by-product interaction. The linear model structure for this model reads:

\[ p_{ij} = f_{psy}(\mu + \alpha_i + b_j + d_{ij}) = f_{psy}(\eta_{ij}) \]  

where \( d_{ij} \sim N(0, \sigma^2_d) \) is the random effect of the interaction of the \( i \)th test product and the \( j \)th assessor, which are independent for all \( i \) and \( j \). \( d_{ij} \) is the difference for the \( j \)th assessor for the \( i \)th test product to the average product-difference \( \mu \) on the d-prime scale.

The relation between the product d-prime value, \( \delta_i \), and the model parameters is not affected by the random effect of the assessors nor the assessor-by-product interaction. This is because the value of \( \delta_i \) is for an average assessor, thus \( \beta_i \) and \( d_{ij} \) are 0 and the relation remains that \( \delta_i = \mu + \alpha_i \).

The model defined by (6) relates to other well-known models in the sensory field. The structure of \( \eta_{ij} \) in (6) resembles the usual 2-way mixed structure for sensory profile data. The usual 2-way analysis of sensory profile data can be done in Panelcheck. If we were to consider a setting with only one test product and multiple observations for each assessor, this corresponds to the usual replicated difference test, which can be modelled by e.g. beta-binomial models.

2.6. Simplification of a Thurstonian Mixed Model

It is of interest to investigate the possibility to describe the data with a simpler model. It will become easier to interpret the results in situations with a simpler model e.g. models with a non-significant assessor-by-product interaction. Thus, it is important to consider the tests of the variables that are included in the linear predictor. This section describes how to investigate whether the linear model structure in (6) can be simplified. The first test that is considered is the test of the assessor-by-product interaction. Both assessor and product effects are nested within the assessor-by-product interaction, thus it is important to consider the test of the interaction before testing for assessor and product effects. The interpretation of the assessor-by-product interaction is that the differences between the assessors depend on the products. Therefore, when testing for a significant assessor-by-product interaction it is investigated whether the assessor differences vary with the products. Since the assessor-by-product interaction is a random effect the hypotheses are statements about the variance parameter. For the test of a significant assessor-by-product interaction, the null hypothesis is that the variance equals zero, while the alternative hypothesis is given as the variance being larger than zero:

\[ H_0 : \sigma^2_d = 0 \quad H_1 : \sigma^2_d > 0 \]

The alternative hypothesis is one-sided since the variance is non-negative; see Christensen & Brockhoff (2013) for details. The distribution of the test statistic is the Chi-squared distribution with 1 degree of freedom. The likelihood ratio test statistics for the test of a significant assessor-by-product interaction are shown in Figure 4. The eight attributes have non-significant assessor-by-product interaction. Thus, there is no evidence that the differences between assessors depend on the test products. The model that is used for testing the main effects of assessors and test products is the model without the assessor-by-product interaction. This model is given in (5). When the assessor-by-product interaction is significant, the understanding of the model becomes more difficult. It is a scope of future research how to define...
and interpret the test of the main effects of products as well as assessors in the case of a significant assessor-by-product interaction.

The hypothesis test of a significant effect of test products investigates whether the difference between the control and the test products is the same for all the test products. The likelihood ratio test statistic becomes 
\[-2 \log(Q) = 2 \ell_{H_2} - 2 \ell_{H_0} \sim \chi^2(l - 1)\] (Pawitan (2001)), where \( \ell_{H_0} \) and \( \ell_{H_2} \) are the log likelihood functions under the null and alternative hypothesis respectively. Furthermore for the data used as an ongoing example in this paper \( l - 1 = 7 \). The model under the alternative hypothesis is given by (5) allowing for the test products to have different sensory characteristics for that attribute. Furthermore, the model under the null hypothesis is stating that the test products are perceived to be similar compared to the control:

\[ p_{ij} = f_{p_{ij}}(\mu + b_j) \]

The likelihood ratio test statistics for the test of a significant product main effect are shown in Figure 4. For all attributes, the product main effect is significant, meaning that the test products are perceived differently compared to the control for all the attributes. Currently assessor replication is often ignored in the analysis of these types of studies, e.g. due to limitations of available software. In such analyses the model reads:

\[ p_{ij} = f_{p_{ij}}(\mu + \alpha_i) \] (8)

where \( \mu \) and \( \alpha_i \) are defined as previously described. The likelihood ratio test of the product main effect is equivalent to the test for the model including assessor. Thus, the model under the null hypothesis reads:

\[ p_{ij} = f_{p_{ij}}(\mu) \]

The values of the likelihood ratio test statistic, as well as the values for the test with assessor included in the model, are shown in Figure 5. The value of the likelihood ratio test statistic is generally higher for the test when assessor is included in the model. For some attributes, the difference is small, whereas the difference for other attributes is rather large. The size of the likelihood ratio statistics is just as important as the difference between them, regarding the impact of which model is used. Silky (0 minutes) and Greasy (0 minutes) approximately have the same size of the difference (approximately 8 and 9 respectively). For Silky (0 minutes) the difference is unimportant because both values are large. However, the difference for Greasy (0 minutes) is important because both values are small. For the 0.01 level the conclusion, for Greasy (0 minutes), depends on which model is used; when ignoring the assessor replicates (model (8)) the null hypothesis is not rejected, whereas inclusion of assessors (model (5)) results in a rejection of the null hypothesis. It is a scope of future research to investigate how much the test of product main effect is affected by ignoring the assessor replicates.

The hypothesis test of a significant assessor main effect is considering whether the assessors perceive the test products differently. Thus, the null hypothesis is assuming that the assessors perceive the products similarly, whereas the alternative hypothesis allows for differences between the assessors. The hypothesis test of a significant assessor main effect is equivalent to the hypothesis test of a significant assessor-by-product interaction, with \( \sigma^2_b \) being replaced by \( \sigma^2_m \) in (7). The likelihood ratio test statistics for the test of a significant assessor main effect are shown in Figure 4. The attributes Thickness and Absorption have non-significant assessor main effects. Hence, there is not enough evidence to claim a significant effect of the assessors for these two attributes. Thus, the assessors perceive the test products similarly for Thickness and Absorption. For the remaining six attributes, the assessor main effect is
strongly significant. Therefore, the assessors perceive the test products compared to the control differently for these attributes.

2.7. Product specific d-prime values

It is of interest to find the product specific d-prime values because this will make it possible to compare the sensory characteristics of the different products. The product specific d-primes are estimated from the model without the assessor-by-product interaction. Thus, the Thurstonian mixed model in (5) is used when finding the product specific d-prime values. Therefore, the estimate, on the d-prime scale, for the \( i \)th product reads:

\[
d'_i = \hat{\mu} + \hat{\alpha}_i
\]

where \( \hat{\mu} \) and \( \hat{\alpha}_i \) are the estimates of \( \mu \) and \( \alpha_i \). The estimates of \( \mu \) and \( \alpha_i \) are obtainable from the output in the statistical software.

When the assessor-by-product interaction is significant, the interpretation of the product specific d-prime values become more difficult. In the situation with a significant assessor-by-product interaction one must be cautious when interpreting the product specific d-prime values, because these estimates do not contain all information about the products. It is a scope of future research to investigate the interpretation of the product specific d-prime values when the assessor-by-product interaction is significant.

Confidence intervals for the d-prime values can be found using the Wald-based approach. The 95% Wald-based confidence interval for \( d_i \) reads:

\[
d'_i \pm z_{97.5} \text{se}(d'_i) \tag{9}
\]

where \( z_{97.5} \) is the 97.5% quantile for the standard normal distribution. Furthermore, \( \text{se}(d'_i) \) is the standard error of \( d'_i \). The standard errors are obtained from the output in the statistical software used when analyzing data with a generalized linear mixed model.

The product specific d-prime estimates as well as the 95% confidence intervals for Sticky, Greasy (0 minutes) and Silky (5 minutes) are shown in Figure 6. Test products A, G and H are more sticky than the control product, whereas the remaining test products are less sticky. The test products furthest to the left (C, D and E) are the most promising test products with respect to stickiness, since the desired characteristic is to be less sticky than the control. All the test products are perceived to be less greasy than the control product, since the d-prime values for Greasy are negative. All test products are good candidates with respect to greasiness, since a desired characteristic for the new product is not to be greasier than the control product. The only test product that is perceived to be more silky after 5 minutes than the control product, is test product D. The d-prime values for test products C and F are close to 0, which indicates that these are among the most silky test products after 5 minutes. All in all when considering the results for the attributes Sticky, Greasy (0 minutes) and Silky (5 minutes) the most promising test product is test product D.

2.8. Assessor specific d-prime values

It is of interest to find the assessor specific d-prime values because these values enable a comparison of the assessors. As for the product specific d-prime values the interpretation of the assessor specific d-prime values is more difficult when the assessor-by-product interaction is significant. Thus, d-prime values for the assessors will be calculated using model (5). The average sensory difference between the test products and the control product for the \( j \)th assessor is on the d-prime scale:

\[
b_j = \mu + b_j \tag{10}
\]

The estimate of \( b_j \) in (10) is obtained from the output in the statistical software used when analyzing data with a generalized linear mixed model.
For a balanced design the assessor with the smallest value has been choosing the control most often of all the assessors, whereas the assessor with the highest value has been choosing a test product most often. Assessors with a value of 0 have been choosing the control and a test product half of the times each. The assessors with larger values than the consensus (µ) have on average chosen a test product more often than the control. The assessors with smaller values than the consensus have on average chosen the control more often than the average.

The d-prime estimates for the assessors, \( \bar{b}_j \), is for Silky (0 minutes) shown in Figure 7. The assessor specific d-prime estimates, \( \bar{b}_j \), are negative for Silky evaluated after 0 minutes. Thus, the assessors are in agreement that the control, on average, is silkier than the test products.

The assessors furthest to the left, assessors 9, 16 and 19, are assessing similarly. The d-prime values for these assessors are close to -2.5, which is rather far away from 0. This implies that these assessors have chosen the control much more than the test products. In addition, these assessors are the assessors with the smallest proportions of times the test products were chosen. There is a group of assessors, from 1 to 21 looking at the y-axis, whose estimates are close to -2. These assessors have larger proportions, of times the test products were chosen, than the group furthest to the left. The assessors from these two groups, the assessors from 9 to 21 looking at the y-axis, have d-prime estimates less than the consensus, the estimate of µ.

The assessors 4 and 20 are somewhat different from the other assessors in the sense that they do not appear to be in a group of assessors.

There is a group of assessors, from 13 to 14 looking at the y-axis, that has d-prime values close to -1. These assessors have larger d-prime estimates than the consensus.

Assessor 7 is the assessor with the d-prime value closest to 0. Thus, assessor 7 is the assessor with the largest proportion, of times the test products were chosen, of the assessors.

The d-prime values for the assessors, \( \bar{b}_j \), is for Sticky (0 minutes) shown in Figure 8. The assessor specific d-prime values for Sticky (0 minutes) are negative as well as positive, with the majority being negative. Therefore, some assessors have chosen the test products more often than the control, however the majority of the assessors have chosen the control more often than the test products.

Assessors 1, 6 and 9 are the assessors with the smallest proportion of times the test products were chosen. Assessor 24 is the assessor with the largest proportion of
times the test products were chosen.

Assessor 9 is among the assessors furthest to the left for both attributes. This means that assessor 9 tend to choose the control more often than the test products.

3. Summary and Discussion

We have in this paper suggested a way to analyze data from a binary paired comparison. The analysis that is suggested is to handle the replications of assessors by including them in the model, thus obtaining a Thurstonian mixed model.

When considering Thurstonian mixed models an important gain is that the hypothesis test of a significant product effect handles the replications correctly. In addition, d-prime values of products as well as assessors are obtained from a Thurstonian mixed model. The assessor specific d-prime values enable a way to monitor a panel.

In the situation with a non-significant assessor-by-product interaction, hypothesis tests and d-prime values are well-defined and interpretable. When the assessor-by-product interaction is significant, further research is needed to define and interpret hypothesis tests as well as the d-prime values for the main effects of products and assessors.

Throughout the paper, an analysis has been made for each attribute separately. Future work could be to investigate the possibility to account for correlations between the attributes.

The difference data considered in this paper is from a binary paired comparison. An interesting continuation of the work presented in this paper is to consider other types of difference data. E.g. where products are compared to each other rather than a control like in Gabrielsen (2000) and Gabrielsen (2001).

Acknowledgments

The research that lead to this paper is funded by the Technical University of Denmark and Unilever U.K. Central Resources Limited. Unilever also provided the data that were used as an example of the analyses in this paper. Furthermore, the first author would like to thank Rebecca Evans for many nice and rewarding discussions.

Appendix A. Thurstonian framework

In this section we include the details about the calculation of the psychometric function defined in Section 2.2:

\[ P(C < T) = P(C - T < 0) \]

\[ = P \left( \frac{C - T}{\sigma \sqrt{2}} < 0 \right) \]

\[ = P \left( \frac{C - T}{\sigma \sqrt{2}} + \frac{\delta}{\sqrt{2}} < \frac{\delta}{\sqrt{2}} \right) \]

\[ = P \left( \frac{C - T + \mu_c - \mu_c}{\sigma \sqrt{2}} < \frac{\delta}{\sqrt{2}} \right) \]

\[ = P \left( \frac{C - T}{\sigma \sqrt{2}} < \frac{\delta}{\sqrt{2}} \right) \]

\[ = \Phi \left( \frac{\delta}{\sqrt{2}} \right) \]

where the second equality is valid since \( \sigma \sqrt{2} > 0 \). In addition the last equality is true since \( Z \sim N(0,1) \) and \( \Phi \) is the cumulative distribution function for the standard normal distribution.

References

Individual differences in replicated multi-product experiments with Thurstonian models for binary paired comparison data
Appendix B

Principal Component Analysis of d-prime values

Abstract

When considering sensory discrimination studies, often multiple d-prime values are obtained across sensory attributes. In this paper, we introduce a way to gain information about the d-prime values across sensory attributes by analyzing these with principal component analysis. We consider d-prime values obtained solely by using the inverse of the psychometric function for the test protocol used in the sensory discrimination study. Additionally, we consider d-prime values obtained by analyzing data with a Thurstonian mixed model. From this analysis product specific as well as assessor specific d-prime values are obtained. Thus, by analyzing these by principal component analysis, information regarding products and assessors is obtained across attributes.

Keywords: d-prime values, discrimination testing, assessor information, multi-product setting, Principal Component Analysis

1. Introduction

In discrimination studies, several attributes can be considered. One approach is to do an analysis for one attribute at a time (e.g. Linander et al. (2018)). For other types of sensory data, e.g. sensory profiling data, many attributes can be present. Such data are often analyzed by Principal Component Analysis (PCA) (Næs & Risvik (1996), Næs et al. (2010), Lawless & Heymann (1998)). PCA is a well-known multivariate analysis that is used in many applications e.g. in Chemometrics (Varmuza & Filzmoser (2009)).

When considering discrimination studies, an advantage of using δ; the measure for sensory differences, is that δ does not depend on the discrimination test protocol, see e.g. Ennis (1993). Thus, it is common to obtain d-prime values, the estimates of δ, when analyzing discrimination studies. In this paper, we will be considering principal component analysis of d-prime values obtained from sensory discrimination tests. We will consider two approaches; one analyzing the raw d-prime values, transforming data by the inverse of the psychometric function, and one where the d-prime values are obtained from analyzing data by a model. The d-prime values will be for different individuals (e.g. products or assessors) for different variables (typically sensory attributes). In the analysis of sensory data Luciano & Næs (2009) considered PCA using the estimates obtained from a regular two-way ANOVA.

Linander et al. (2018) analyzed data from a discrimination study of binary paired comparison data, by using a Generalized Linear Mixed Model embedded into a Thurstonian framework. This model can be used to analyze multiple sensory attributes one at a time. From each analysis, assessor specific as well as product specific d-prime values are obtained. These d-prime values will be used in a principal component analysis. Thus, information about products as well as assessors are obtained across attributes.

In the remainder of this section a discrimination study is described. This is used, as an ongoing example, throughout the paper to illustrate the methods we explain in Section 3 and Section 4. We conclude the paper with a discussion in Section 5.

1.1. Discrimination study

The discrimination study that is used as an ongoing example throughout this paper, is the same discrimination study that is used in Linander et al. (2018). In this section, we briefly explain the structure of the discrimination study and refer the reader to Linander et al.
The study includes eight test products, each compared to the same control product, for up to 25 assessors. Not all assessors evaluated all the test products. The test products are products which are applied to the skin. The assessors evaluated eight attributes, five of these were evaluated immediately after application to the skin. In addition, three of these attributes were re-evaluated 5 minutes after application. An assessor had to choose the sample with the strongest intensity of the attribute in question. Each assessor evaluated each test product twice by making one comparison in two consecutive sessions.

2. Principal Component Analysis

In this section, we give an introduction to principal component analysis. For a more detailed description of PCA, many books and papers exist; e.g. Næs et al. (2010), Varmuza & Filzmoser (2009), Bro & Smilde (2014).

The data used in PCA are collected in a data matrix. Typically the objects are in the rows and the variables in the columns. Let \( y_{ij} \) be the observation for the \( i \)th object on the \( j \)th variable. Then the data matrix \( Y \) is given as:

\[
Y = \begin{bmatrix}
  y_{11} & \cdots & y_{1J} \\
  y_{21} & \cdots & y_{2J} \\
  \vdots & \ddots & \vdots \\
  y_{I1} & \cdots & y_{IJ}
\end{bmatrix}
\]

(1)

where \( i = 1, \ldots, I \) is the number of objects, and \( j = 1, \ldots, J \) is the number of variables.

The overall purpose of a PCA is to explain the structure in the data by fewer dimensions than in the original data. The new dimensions are the so-called Principal Components (PC). The first principal component is defined such that it is explaining most of the variation. The second principal component is explaining the second most of the variation under the restriction that it is orthogonal to the first principal component and so forth. For each principal component a set of scores (for the objects) and a set of loadings (for the variables) are obtained. In many applications, data are centered as well as scaled before doing the PCA. The scaling is typically important since variables can be measured using different scales. However, when considering variables within sensory panel studies, such differences in scales rarely occur. Thus, it is often the case, that PCA is applied without scaling the variables when considering experiments in sensory panel studies (Borgogone et al., 2001; Næs et al., 2010; Lawless & Heymann, 1998).

In most applications, including sensory science, centering is most often done. However, important information could be ignored, when d-prime values are centered. Thus, there exist situations where using the non-centered d-prime values is informative.

3. PCA using d-prime values

In this section, principal component analysis is considered using d-prime values obtained from a sensory discrimination study. Basically, any set of d-prime values can be used; from discrimination studies with a correct answer as well as studies using paired comparisons. Here binary paired comparisons are thought of as unbounded 2-AFC tests where no correct answer exists. The implication of no correct answer is that the d-prime values can be negative.

Let \( x_{ij} \) be the number of correct answers, out of a total of \( n_{ij} \) answers, for the \( i \)th object (e.g. product) and the \( j \)th variable (e.g sensory attribute). Let \( f_{psy} \) be the psychometric function for the test protocol used in the discrimination study. Data are transformed into d-prime values using the inverse of the psychometric function:

\[
f_{psy}^{-1}(p_{ij}) = d_{ij}
\]

(2)

where

\[
p_{ij} = x_{ij}/n_{ij}
\]

(3)

for the binary paired comparison. For discrimination studies with a correct answer a slight modification of (3) is needed. Therefore, for studies with a correct answer:

\[
p_{ij} = \begin{cases} 
  x_{ij}/n_{ij} & \text{if } x_{ij}/n_{ij} \geq p_{g} \\
  p_{g} & \text{if } x_{ij}/n_{ij} < p_{g}
\end{cases}
\]

(4)

where \( p_{g} \) is the guessing probability for that study.

The data-matrix, with the d-prime values as the entries, is organized such that the individuals are in the rows and the variables are in the columns:

\[
Y = \begin{bmatrix}
  d'_{11} & \cdots & d'_{1J} \\
  d'_{21} & \cdots & d'_{2J} \\
  \vdots & \ddots & \vdots \\
  d'_{I1} & \cdots & d'_{IJ}
\end{bmatrix}
\]

(5)

When doing the principal component analysis using \( Y \) in (5) scores as well as loadings are obtained such that (Næs et al., 2010):

\[
Y = TP^T
\]

(6)
where $P$ is the matrix with the loadings and $T$ is the matrix with the scores.

For the d-prime values, no scaling is used since all of the values are on the same scale, namely the d-prime scale. An absolute d-prime value of 3 is large. Thus, d-prime values will mostly lie in the interval from $-3$ to 3 when allowing for negative d-prime values as in the binary paired comparison. For studies with a correct answer the interval containing most of the d-prime values is from 0 to 3. Therefore, with the same reasoning as in sensory science in general, no scaling is done using d-prime values in a principal component analysis.

3.1. Example

In this section, the data from the discrimination study explained in Section 1.1 are used. The data are the number of times test products were chosen. The proportions of times the test products were chosen, are transformed into d-prime values using the inverse of the psychometric function:

$$f_{\text{paired}}^{-1}(p) = \Phi^{-1}(p) \sqrt{2} = d'$$  \hspace{1cm} (7)

We refer the reader to Linander et al. (2018) for details of the derivation of (7). There are no restrictions regarding the d-prime values. Positive as well as negative d-prime values can occur, since the test protocol is the binary paired comparison. A negative d-prime value indicates that the control product has the strongest intensity of the attribute in question.

The d-prime values, transformed by (7), are shown in Table 1. It is worth mentioning, that test product $H$ for Silky has a d-prime value of $-\infty$. It is not possible to do a PCA with a value of $-\infty$. Therefore, it must be decided how to handle the infinite value for test product $H$ for Silky.

A d-prime value of $-\infty$ only occurs when the test products were chosen zero times. A way to select the value to replace $-\infty$, is by letting the test products being chosen once. Using the proportion of $1/n$ instead of $0/n$, is conceptually the same, since it has no practical implication whether the test products were chosen zero times or one time. From both situations it is clear that the control is perceived to have the strongest sensory intensity.

The advantage of using $1/n$ rather than $0/n$ is it ensures that a finite d-prime value is obtained. $d'_{H,\text{Silky}}$, the new d-prime value for test product $H$ for Silky, is given as:

$$d'_{H,\text{Silky}} = f_{\text{paired}}^{-1}(1/n_{H,\text{Silky}}) = -2.83$$  \hspace{1cm} (8)

where $n_{H,\text{Silky}}$ is the number of evaluations for test product $H$ for Silky. It is a scope of future research to investigate how different approaches to handle a value of minus infinity, affect the PCA.

It can be difficult to realize if any of the test products have the same sensory characteristics by looking at Table 1. The aim of the PCA is to be able to identify the test products, if any, that have the same sensory characteristics.

Using the d-prime values from Table 1, with $-\infty$ replaced by the d-prime value in (8), the biplot shown in Figure 1 is obtained doing a PCA using the centered values.

The first principal component, PC1, is explained by Sticky (initially and after five minutes) as well as Silky (initially and after five minutes). The test products H and D are placed in opposite directions with respect to PC1; H is stickier and less silky than D. Test products A and G are very similar with respect to PC1. Thus, A and G have similar sensory properties regarding stickiness and silkiness. The second principal component, PC2, is primarily explained by Thickness. The test products A and G are the thickest products. H is the least thick product. The test products B, F, E, C and D are similar with respect to Thickness. In conclusion, test products A and G are similar with respect to the sensory attributes Thickness, Silky (initially and after 5 minutes) and Sticky (initially and after 5 minutes). Test product H is very different from all of the other test products. Test product D is by far the most silky test product.
4. PCA using d-prime values obtained from a Thurstonian Mixed Model

In Section 3, the PCA was applied using a set of d-prime values found by using the inverse of the psychometric function for that discrimination test protocol. In this section, we will model the probabilities of a test product being chosen when considering the binary paired comparison test protocol. The d-prime values obtained from such a model are then analyzed by PCA. It is not necessary to fully understand the model to be able to comprehend the results of the PCA of the d-prime values. Thus, for readers without interest in how the model is defined, Section 4.1 can be omitted.

4.1. Thurstonian Mixed Model

This section defines the model that is used to find the d-prime values. The model was suggested in Linander et al. (2018). In this section, we will include the information necessary to understand the model. The model described in this section is modelling the data for one sensory attribute at a time. Thus, to get the results for all the sensory attributes from the sensory discrimination study, the analysis is repeated for each attribute.

The data are obtained from a discrimination study using the binary paired comparison test protocol. Each observation is binomially distributed:

\[ Y_{lk} \sim \text{binomial}(p_{lm}, 1) \]

where \( l = 1, \ldots, L \) represents the test products, \( m = 1, \ldots, m \) represents the assessors for the \( l \)th test product and \( k = 1, \ldots, K \) represents the sessions carried out on the same day (\( K = 2 \) and \( L = 8 \) for the discrimination study used as an example in this paper). We assume that \( p_{lm} \), the probability of the \( m \)th assessor choosing the \( l \)th test product, is independent of the sessions:

\[ p_{lm} = P(Y_{lk} = 1) \]

It is possible to impose a linear structure of \( p_{lm} \) which explains the variables that are affecting these probabilities. We consider a model where the probabilities are explained by products as well as assessors:

\[ p_{lm} = f_{\text{paired}}(\mu + \alpha_l + b_m) \]  

(9)

where \( f_{\text{paired}} \) is the psychometric function with the inverse given in (7). Additionally, \( \mu \) is the overall average difference between the test products and the control product and \( \alpha_l \) is the difference for the \( l \)th test product to the average product-difference \( \mu \). Thus, the sensory difference for the \( l \)th test product to the control product is

\[ \delta_l = \mu + \alpha_l \]  

(10)

Furthermore, \( b_m \) is the random effect of the \( m \)th assessor where \( b_m \sim N(0, \sigma^2_{m}) \) which are independent for all \( m \). \( b_m \) is the difference for the \( m \)th assessor to the average product-difference \( \mu \) on the \( d' \)-scale. Thus, the sensory difference, on the \( d' \) scale, between the test products and the control product for the \( m \)th assessor is \( b_m = \mu + b_m \).

For this setup, the interest lies in what information it is possible to extract regarding product specific as well as assessor specific d-prime values across the attributes. Thus, the attributes will be considered as the variables (columns in (5)) and the test products as well as the assessors will be considered as the observations (rows in (5)).

<table>
<thead>
<tr>
<th>Test Product</th>
<th>Sticky 0 min</th>
<th>Sticky 5 min</th>
<th>Greasy 0 min</th>
<th>Greasy 5 min</th>
<th>Silky 0 min</th>
<th>Silky 5 min</th>
<th>Thickness 0 min</th>
<th>Thickness 5 min</th>
<th>Absorption 0 min</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.55</td>
<td>-1.21</td>
<td>-0.91</td>
<td>-1.59</td>
<td>-1.00</td>
<td>-1.10</td>
<td>2.86</td>
<td>-1.33</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>-1.19</td>
<td>0.09</td>
<td>-1.47</td>
<td>-0.36</td>
<td>-1.19</td>
<td>-1.63</td>
<td>0.54</td>
<td>1.47</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>-2.33</td>
<td>-1.81</td>
<td>-1.32</td>
<td>-0.95</td>
<td>0.00</td>
<td>-0.54</td>
<td>-0.09</td>
<td>0.45</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>-2.83</td>
<td>-2.83</td>
<td>-0.58</td>
<td>-0.58</td>
<td>2.11</td>
<td>0.76</td>
<td>0.00</td>
<td>0.86</td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>-1.63</td>
<td>-1.81</td>
<td>-2.04</td>
<td>-1.47</td>
<td>-0.85</td>
<td>-0.95</td>
<td>0.00</td>
<td>0.74</td>
<td></td>
</tr>
<tr>
<td>F</td>
<td>-0.31</td>
<td>-1.59</td>
<td>-1.74</td>
<td>-1.10</td>
<td>-0.55</td>
<td>-0.23</td>
<td>-0.08</td>
<td>-0.47</td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>0.64</td>
<td>0.00</td>
<td>-1.32</td>
<td>-0.95</td>
<td>-1.63</td>
<td>-1.19</td>
<td>2.33</td>
<td>0.95</td>
<td></td>
</tr>
<tr>
<td>H</td>
<td>1.89</td>
<td>0.76</td>
<td>-0.86</td>
<td>-1.55</td>
<td>-Inf</td>
<td>-2.11</td>
<td>-1.06</td>
<td>-1.55</td>
<td></td>
</tr>
</tbody>
</table>
4.2. PCA using product specific d-prime values

The product specific d-prime values; the estimates of $\delta_l$, are for a specific sensory attribute given as

$$d'_{ij} = \hat{\mu} + \hat{\alpha}_{lj}$$

(11)

where $\hat{\mu}$ and $\hat{\alpha}_{lj}$ are the estimates of $\mu$ and $\alpha_l$ respectively. To be able to distinguish the estimates obtained for the different attributes, an additional sub-script will be used:

$$d'_{lj} = \hat{\mu}_j + \hat{\alpha}_{lj}$$

(12)

where $j = 1, \ldots, J$ represents the sensory attribute and $\hat{\mu}_j$ and $\hat{\alpha}_{lj}$ are the estimates obtained from the analysis of the $j$th attribute. Thus, $d'_{lj}$ is the sensory difference for the $l$th test product to the control product for the $j$th attribute.

An important aspect of PCA is whether to center the data before doing PCA or not. For the product specific d-prime values both situations will be considered, since each of these contributes with valuable information regarding the test products. As we will show in Section 4.2.1, when centering the product specific d-prime values, the information regarding the control product is removed. However, when the d-prime values are used, without centering, the information about the control product is maintained in the PCA.

4.2.1. Centering

The model in (9) is over-parameterized, thus it is assumed that for each $j$:

$$\sum_{l=1}^{L} \hat{\alpha}_{lj} = 0$$

(13)

When centering the product specific d-prime values, the mean value of the $d'_{lj}$s for each $j$ is subtracted. $\bar{d}'_{lj}$: the mean value over $i$, for a given $j$, reads:

$$\bar{d}'_{lj} = \frac{1}{L} \sum_{i=1}^{L} d'_{ij}$$

$$= \frac{1}{L} \sum_{l=1}^{L} (\hat{\mu}_j + \hat{\alpha}_{lj})$$

$$= \frac{1}{L} \hat{\mu}_j + \frac{1}{L} \sum_{l=1}^{L} \hat{\alpha}_{lj}$$

$$= \bar{\mu}_j$$

(14)

where the last equality follows from (13). Therefore, the centered d-prime values are given as:

$$d'_{lj} - \bar{d}'_{lj} = d'_{lj} - \bar{\mu}_j$$

$$= (\hat{\mu}_j + \hat{\alpha}_{lj}) - \bar{\mu}_j = \hat{\alpha}_{lj}$$

Thus, when doing the PCA using the centered d-prime values, it is the $\alpha_{ljs}$ that are used. Hence, when interpreting the results of the PCA, the information regarding the control product has been removed. Recall that $\alpha_{lj}$ is merely expressing the difference from the $l$th test product to the average product-difference $\bar{\mu}_j$ for the $j$th attribute. Thus, when considering the PCA on the centered d-prime values it is possible to compare the products to each other but not to the control.

4.2.2. Example

The product specific d-prime values, obtained from model (9), are listed in Table 2. The values in Table 2 are further away from zero than the values in Table 1. As in Section 3.1 the d-prime value for test product $H$ for $S\text{ilk}$ky equals $-\infty$. This value was in Section 3.1 replaced by $-2.83$ corresponding to test product $H$ being chosen once. In Table 1 it is seen that test product $D$ for Stickey (evaluated initially and after 5 minutes) has the same value as the imputed value. With the reasoning that the estimated value for test product $H$ for $S\text{ilk}$ky (0 minutes) would have been similar to that of test product $D$ for Stickey, we use the value (of these two) furthest away from 0 given in Table 2 by $-3.47$. The analysis of the d-prime values from the discrimination study with the centered PCA results in the biplot shown in Figure 2.

The absolute values are a bit larger in Figure 2 compared to the values in Figure 1. However, the conclusions are the same as those for Figure 1.

4.2.3. Non-centered

When the d-prime values are not centered, the information about the control is maintained. The d-prime values that are used are the $\hat{\mu}_j + \hat{\alpha}_{lj}s$. The difference $\hat{\mu}_j + \hat{\alpha}_{lj}$ is the estimated sensory difference between the $l$th test product and the control for the $j$th attribute. Therefore, all the d-prime values used for the PCA are differences between the test products and the control. Thus, the origin in the biplot corresponds to the control product.

4.2.4. Example

The biplot for the PCA using the non-centered product specific d-prime values is shown in Figure 3.

The arrows for Thickness and Absorption are short. That, in combination with many small d-prime values, makes it extremely difficult to conclude anything regarding Thickness and Absorption from Figure 3 (the sign of the values will easily change, since other attributes will affect the values of the scores more heavily). Considering Greasy (initially and after 5 minutes), all of the test products are less greasy than the control,
Table 2: Product specific d-prime values found by using the model given in (9).

<table>
<thead>
<tr>
<th>Test Product</th>
<th>Sticky 0 min</th>
<th>Sticky 5 min</th>
<th>Greasy 0 min</th>
<th>Greasy 5 min</th>
<th>Silky 0 min</th>
<th>Silky 5 min</th>
<th>Thickness 0 min</th>
<th>Thickness 5 min</th>
<th>Absorption 0 min</th>
<th>Absorption 5 min</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.64</td>
<td>-1.36</td>
<td>-1.19</td>
<td>-1.76</td>
<td>-1.15</td>
<td>-1.57</td>
<td>2.94</td>
<td>-1.35</td>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>-1.44</td>
<td>0.08</td>
<td>-1.98</td>
<td>-0.51</td>
<td>-1.47</td>
<td>-2.39</td>
<td>0.58</td>
<td>1.50</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>-2.60</td>
<td>-2.15</td>
<td>-1.79</td>
<td>-1.22</td>
<td>-0.03</td>
<td>-0.89</td>
<td>-0.13</td>
<td>0.46</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>-3.42</td>
<td>-3.47</td>
<td>-0.67</td>
<td>-0.65</td>
<td>2.28</td>
<td>0.83</td>
<td>-0.01</td>
<td>0.88</td>
<td></td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>-1.93</td>
<td>-2.22</td>
<td>-2.74</td>
<td>-1.80</td>
<td>-1.02</td>
<td>-1.47</td>
<td>-0.04</td>
<td>0.76</td>
<td></td>
<td></td>
</tr>
<tr>
<td>F</td>
<td>-0.40</td>
<td>-1.88</td>
<td>-2.21</td>
<td>-1.19</td>
<td>-0.64</td>
<td>-0.37</td>
<td>-0.09</td>
<td>-0.47</td>
<td></td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>0.70</td>
<td>-0.01</td>
<td>-1.81</td>
<td>-1.21</td>
<td>-1.89</td>
<td>-1.77</td>
<td>2.43</td>
<td>0.99</td>
<td></td>
<td></td>
</tr>
<tr>
<td>H</td>
<td>2.06</td>
<td>0.88</td>
<td>-1.04</td>
<td>-1.82</td>
<td>-Inf</td>
<td>-3.03</td>
<td>-1.09</td>
<td>-1.58</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 2: The biplot for the centered product specific d-prime values.

Figure 3: The biplot for the non-centered product specific d-prime values.
since they are placed opposite of the direction of the arrows for Greasy. With respect to silkiness, the majority of the test products are less silky than the control. With our current understanding of the interpretation of Figure 3 it appears that test product C is just about as silky as the control (it is on the same "level" as the origin with respect to the arrow for Silky 0 minutes) and test product D is the only test product, which is silkier than the control (it is above the "level" for the origin with respect to the arrow for Silky 0 minutes). Furthermore, with respect to stickiness (initially), our current understanding is that the test products G and H are stickier than the control (they are above the "level" for the origin with respect to the arrow for Sticky 0 minutes); A is more or less as sticky as the control (it is on the same "level" as the origin with respect to the arrow for Sticky 0 minutes). It appears that the remaining test products are less sticky than the control (they are below the "level" for the origin with respect to the arrow for Sticky 0 minutes).

It appears that after five minutes, the only test product that is stickier than the control is test product H (it is above the "level" for the origin with respect to the arrow for Sticky 5 minutes).

From Figure 3 we believe it is possible to identify the test products that are the most interesting test products, based on the sensory properties that are the most important ones. For a test product to be interesting it must be as silky as the control as well as being less greasy and sticky. Thus, with our current understanding, the most interesting test products are C and D, with D being the more promising of the two, since it is silkier than C.

4.3. PCA using assessor specific d-prime values

To gain knowledge about which assessors are scoring the test products similarly, the \( b_{mj} \)s are considered for each attribute. To be able to distinguish the estimates obtained for the different attributes, an additional sub-script will be used. Thus, \( b_{mj} \) is the difference for the \( m \)th assessor to the average product-difference \( \mu_j \), on the \( d' \)-scale, for the \( j \)th attribute. The centered and non-centered PCA will give similar results since \( E(b_{mj}) \approx 0 \), thus the assessor specific d-prime values \( b_{mj} \)s are almost centered. Considering the \( b_{mj} \)s it is possible to investigate how the assessors are performing compared to the average.

4.3.1. Example

The biplot for the assessor specific d-prime values is shown in Figure 4. The first principal component is explained by Silky, mostly evaluated after five minutes, but also initially. The assessors 7 and 8 are scoring high with respect to silkiness. Assessor 19 is the assessor scoring the lowest regarding Silky. The second principal component is mostly explained by Greasy (initially and after five minutes). Assessor 23 is the assessor scoring highest with respect to Greasy. There are assessors who are performing similarly. The assessors 12 and 15 are close in the biplot. Assessors 22, 18 and 11 are also close together. Furthermore, with our current understanding it appears that the assessors in the lower left quadrant have a tendency to score higher than the average assessor across the attributes since all of the arrows points toward left and/or down.

In our opinion Figure 4 can be used to look for scoring patterns which might be missed otherwise. However, this is not the same as being able to interpret the quality of an assessor. We believe that to be able to interpret the quality of the assessors prior knowledge about the 'correct' product differences must be available.

5. Summary and Discussion

We have in this paper been considering principal component analysis using d-prime values. We have been considering two types of d-prime values. More specifically, we have been considering d-prime values that are obtained from transforming a proportion of times a product was chosen as well as d-prime values.
obtained from a Thurstonian mixed model. When em-
bedding the analysis into a Thurstonian mixed model, it
is possible to get information about the products as well
as the assessors, on the d-prime scale. These d-prime
values can then be used for the PCA to gain knowledge
about the products or assessors across the attributes.
One d-prime value that was used in the PCA was not
finite, and therefore an imputed value was used instead
of −∞. This was handled by using the proportion of \(1/n\)
rather than \(0/n\) since these proportions conceptually are
identical. It is a scope of future research to investigate
how different approaches of imputing a value for an out-
lier affect the results of the PCA using d-prime values.

Acknowledgments

The research that lead to this paper is funded by the
Technical University of Denmark and Unilever U.K.
Central Resources Limited. Unilever also provided the
data that were used as an example of the analyses in this
paper.

References

Borgognone, M. G., Bussi, J., & Hough, G. (2001). Principal com-
ponent analysis in sensory analysis: covariance or correlation ma-
trix’. Food Quality and Preference, 12, 323–326.
Bro, R., & Smilde, A. K. (2014). Principal component analysis. Ana-
lytical Methods, 6, 2812–2831.
Journal of Sensory Studies, 8, 353–370.
Springer.
Linander, C. B., Christensen, R. H. B., Cleaver, G., & Brockhoff, P. B.
(2018). Individual differences in replicated multi-product experi-
ments with thurstonian mixed models for binary paired comparison
data. Submitted to Food Quality and Preference.
ing principal component analysis and analysis of variance. Food
Quality and Preference, 20, 167 – 175.
and Consumer Science. Wiley.
science. Elsevier.
Appendix C

Analysis of multiple d-primes obtained from various discrimination protocols

Analysis of multiple d-primes obtained from various discrimination protocols

Christine Borgen Linander ∗, Rune Haubo Bojesen Christensen, and Per Bruun Brockhoff

1DTU Compute, Section of Statistics and Data Analysis, Technical University of Denmark, Richard Petersens Plads, Building 324, DK-2800 Kongens Lyngby, Denmark
2Christensen Statistics, Bringetoften 7, DK-3500 Værløse, Denmark

Abstract

Sensory discrimination tests can be conducted to investigate the performance of sensory panels or to compare different laboratories. In such situations multiple d-prime values can be obtained. In this paper we propose a new test statistic for the comparison of multiple d-prime values. The test statistic is for independent sensory discrimination tests, which lead to binomially distributed responses. The test statistic we suggest is an improved way of analyzing multiple d-prime values compared to a previous suggested test statistic.

1 Introduction

It has become more and more common to do experiments with many discrimination test protocols. Usually the protocols used are identical, but situations with several different discrimination protocols occur as well.

In this paper we introduce a new way of comparing and analyzing multiple d-primes from various discrimination test protocols. A method for doing this analysis has previously been suggested by Bi et al. (1997). The method we propose is based on likelihood theory and it will be denoted the likelihood method. There is a need for this analysis since the likelihood method has advantages compared to the method put forward by Bi et al. (1997). The likelihood method can handle boundary situations more specifically situations where the number

∗corresponding author, email: chjo@dtu.dk
of correct answers equals the total number of answers or when the proportion of
correct answers is at or below the guessing probability. In addition the likelihood
method has higher power.

We are considering the situation with several experiments obtained from various
discrimination test protocols. More specifically we consider the discrimination
protocols that lead to simple-binomial data. The simple-binomial protocols in-
clude the Duo-Trio, Triangle, Tetrad, 2-AFC and 3-AFC protocols. We call
these protocols simple-binomial since the number of correct answers from ex-
periments involving these protocols follow a simple binomial distribution. This
is in contrast to protocols such as A-not A and same-different for which the un-
derlying statistical model is also binomial, but product-binomial or compound-
binomial rather than simple-binomial.

In discrimination experiments involving one of the simple binomial protocols,
the number of correct answers, \( X \) follows a binomial distribution:

\[
X \sim \text{binom}(p_c, n),
\]

where \( p_c \) is the probability of a correct answer and \( n \) is the sample size in the
experiment.

In section 2.1 we cover the boundary situations and in section ?? we consider
the power.

### 2 Test of any differences among the d-prime values

We are considering \( d'_1, \ldots, d'_k \) which are obtained from \( k \) simple binomial proto-
cols. \( x_i \) and \( n_i \) being respectively the number of correct answers and the total
number of answers from the \( i^{th} \) protocol, \( i = 1, \ldots, k \). The \( k \) discrimination
tests are assumed to be independent.

We want to test the hypothesis that all \( d' \)s are equal versus the alternative that
at least two are different:

\[
H_0 : \ d'_1 = d'_2 = \ldots = d'_k \quad \text{versus} \quad H_A : \ d'_i \neq d'_{i'},
\]

for at least one pair of \( (i, i'), i \neq i' \). We denote the hypothesis in (2) by the
any-differences hypothesis.

We suggest to use the likelihood ratio statistic for the test of the any-differences
hypothesis given in (2). Some likelihood theory will be included in this paper,
but for a more thorough introduction the reader is referred to e.g. Pawitan

Since the separate discrimination tests are assumed to be independent, the joint
likelihood function is the product of the individual likelihood functions:

\[ L_A(\mathbf{d}'; \mathbf{x}, \mathbf{n}, \mathbf{m}) = \prod_{i=1}^{k} L_{A_i}(d_i'; x_i, n_i, m_i), \quad (3) \]

where \( \mathbf{d}' = (d'_1, \ldots, d'_k)^T \), \( \mathbf{x} = (x_1, \ldots, x_k)^T \), \( \mathbf{n} = (n_1, \ldots, n_k)^T \) and \( \mathbf{m} = (m_1, \ldots, m_k)^T \). \( m_i \) is the test protocol used for the \( i \)th test. We use the subscript \( A \) because (3) is the likelihood function under the alternative hypothesis/model where not all \( d' \) values are equal.

Since we are considering the simple binomial protocols, the individual likelihood function is the binomial density:

\[ L_{A_i}(d'_i; x_i, n_i, m_i) = \binom{n_i}{x_i} p_i^{x_i} (1 - p_i)^{n_i - x_i}, \quad (4) \]

where \( p_i = f_{psy}(d'_i) \) and \( f_{psy} \) is the psychometric function corresponding to the \( m_i \)'th protocol. Expressions of \( f_{psy} \) for Duo-Trio, Triangle, 2-AFC and 3-AFC are given in (Ennis, 1993; Brockhoff and Christensen, 2010), while \( f_{Tetrad} \) is given in (Ennis et al., 1998).

The joint log-likelihood function is the sum of the individual log-likelihood functions:

\[ \ell_A(\mathbf{d}'; \mathbf{x}, \mathbf{n}, \mathbf{m}) = \sum_{i=1}^{k} \log L_{A_i}(d'_i; x_i, n_i, m_i) \quad (5) \]

\[ = \sum_{i=1}^{k} \log \left( \binom{n_i}{x_i} p_i^{x_i} (1 - p_i)^{n_i - x_i} \right). \]

Under the null hypothesis all \( d' \) values are equal and the likelihood function is defined as in (3) but with the slight change that \( p_i \) is replaced by \( p_e \):

\[ L_0(\mathbf{d}'_e; \mathbf{x}, \mathbf{n}, \mathbf{m}) = \prod_{i=1}^{k} L_{0_i}(d'_e; x_i, n_i, m_i), \quad (6) \]

where \( p_e = f_{psy}(d'_e) \) with \( d'_e \) being the expected value of the common d-prime obtained from the discrimination tests.

Similarly the log-likelihood function under \( H_0 \) reads:

\[ \ell_0(\mathbf{d}'_e; \mathbf{x}, \mathbf{n}, \mathbf{m}) = \sum_{i=1}^{k} \log L_{0_i}(d'_e; x_i, n_i, m_i) \quad (7) \]

\[ = \sum_{i=1}^{k} \log \left( \binom{n_i}{x_i} p_e^{x_i} (1 - p_e)^{n_i - x_i} \right). \]
The likelihood ratio test statistic reads:
\[-2 \log Q = -2 \log \left( \frac{L_0(d'_e; x, n, m)}{L_A(d'; x, n, m)} \right) = -2 \log \left( L_0(d'_e; x, n, m) \right) - 2 \log \left( L_A(d'; x, n, m) \right) = 2 \log \left( L_A(d'; x, n, m) \right) - 2 \log \left( L_0(d'_e; x, n, m) \right) = 2 \ell_A(d'; x, n, m) - 2 \ell_0(d'_e; x, n, m). \] (8)

Under the null hypothesis that all $d'_i$ are equal, the expected value $d'_e$ can be found in different ways. Bi et al. (1997) proposed a weighted average in which the $d'_i$ estimates are weighted by their uncertainty (squared standard error). We suggest using the maximum likelihood (ML) estimate of the common $d'$ under the null hypothesis.

The ML estimate of $d'_e$ is given by the maximum of the joint (log) likelihood function under the null hypothesis. Since the separate discrimination tests are assumed to be independent, the joint log-likelihood function is the sum of the individual log-likelihood functions:
\[ \ell_0(d'_e; x, n, m) = \sum_{i=1}^{k} \ell_0(d'_i; x_i, n_i, m_i), \] (9)
where $m_i$ indicates the method or protocol in the $i$th discrimination test. If $m_i$ is one of the protocols for which $X_i$ follows a binomial distribution, then the $i$th log-likelihood function is given by:
\[ \ell_0(d'_i; x_i, n_i, m_i) = x_i \log(p_i) + (n_i - x_i) \log(1 - p_i) + \log \left( \frac{n}{x} \right), \] (10)
where $p_i = f_{psy(i)}(d'_e)$.

The ML estimator of $d'_e$ then reads:
\[ \hat{d}'_e = \arg \max_{d'_e} \ell_0(d'_e; x, n, m). \] (11)

### 2.1 Boundary situations

When data consists of the so-called boundary situations the method suggested by Bi et al. (1997) is not well-defined. Since such situations occur rather frequently this is a non-negligible deficiency of the Wald-type test. The likelihood test we suggest in this paper is well-defined in the boundary situations.

### 3 Post hoc analyses

In post hoc analyses we make inference for functions of the parameters such as all paired differences among the $d'$ estimates. The complete procedure involves
several steps including: producing estimates of the parameter functions of interest; computing the variance-covariance matrix of these parameter functions, computing p-values possibly adjusting for multiple testing, and summarizing differences among pairs in a compact letter display.

3.1 Difference from specified value

If we want to test that all $d'$ values are different from a specified value, $d'_0$, the null and alternative hypotheses have the form:

$$H_0 : d'_i = d'_0 \quad \text{versus} \quad H_A : d'_i \neq d'_0$$

such that we are testing $n$ hypotheses, $d'_0$ is the value of $d'$ under the null hypothesis and the hypotheses may be directional instead of two-sided as illustrated here.

A Wald test statistic reads $t_i(d'_0) = (d'_i - d'_0)/\text{se}(d'_i)$

The likelihood root statistic for this situation is

$$r_i(d'_0) = \text{sign}(d'_i - d'_0)\sqrt{2\{\ell_A(d'_i) - \ell_0(d'_0)\}}$$

where $d'_i$ is the value of $d'$ under the alternative hypothesis, $\ell_A(d'_i)$ is the log-likelihood under the alternative and $\ell_0(d'_0)$ is the log-likelihood under the null.

Score and ‘exact’ tests are also possible for these hypotheses, but not considered further.

3.2 Difference from common $d$

If we want to test for each $d'_i$ whether it can be considered different from the common $d'$, we could frame the hypotheses as in (12). This would consider $d'_0$ a fixed number, while in fact it depends on the data. Taking this dependency into account we consider the following hypotheses for $(i,i') \in 1,\ldots,n$:

$$H_0 : d'_i = d'_e \quad \text{for all } i \quad \text{versus} \quad H_A : d'_i = d'_{e(i')} \quad \text{for all } i \text{ except } i'$$

such that the null model is parameterized by $d'_e$ while the alternative model is parameterized by $(d'_{e(i')},d'_{i'})$ for the $i'$th test. Here, $d'_{e(i')}$ is the common $d'$ considering all $i$ except $i'$.

The likelihood root statistic now has the form

$$r_i = \text{sign}(d'_i - d'_e)\sqrt{2\{\ell_A(d'_{e(i')},d'_{i'}) - \ell_0(d'_e)\}}$$

where

$$\ell_A(d'_{e(i')},d'_{i'}) = \ell(d'_{i'}; x_{i'}, n_{i'}) + \sum_{i|i \neq i'} \ell(d'_{e(i')}, x_{i}, n_{i})$$

5
Note that $d'_{e(i')}$ has to be estimated $n$ times — one time for each $i' \in 1, \ldots, n$.

Observe that both null and alternative models involve optimization to estimate $d'_e$ and $d'_{e(i')}$ for $i' = 1, \ldots, n$. This means that Wald and Score tests, which usually have the advantage of being computationally simpler than likelihood ratio tests, are in this case computationally more complicated. 'Exact' tests are not directly available here.

### 3.3 Pairwise differences

Tests of pairwise differences occur in two natural settings: 1) If we want to compare $d'$ in one group with $d'$ in all the others one at a time, and 2) if we want to look at all pairwise differences. In the first setting we are considering one group a standard or baseline and the structure is known as Dunnett’s contrasts.

The second setting is known as Tukey’s contrasts for all pairwise differences.

Let $d'_{i,i'} = d'_i - d'_{i'}$ for $i, i' = 1, \ldots, n$, then the general hypothesis for the test of a pairwise difference reads:

\[
\begin{align*}
H_0 : \quad d'_{i,i'} &= 0 \\
H_A : \quad d'_{i,i'} &\neq 0
\end{align*}
\]

for some pair $(i, i')$. Here $d'_{i,i'}$ is called a parameter function since it is a function of the original $(d')$ parameters.

Let $\hat{\theta}$ be a vector of $d'$ estimates and $\hat{\vartheta} = K\hat{\theta}$ be the parameter functions of interest, where $K$ is a constant matrix of suitable dimensions. When $\vartheta$ are Tukey’s all-pairwise differences, $K$ has the following structure:

```r
library(multcomp)
named.vec <- setNames(rep(1, 4), paste("group", 1:4, sep=""))
(K <- contrMat(named.vec, type="Tukey"))
Multiple Comparisons of Means: Tukey Contrasts

<table>
<thead>
<tr>
<th></th>
<th>group1</th>
<th>group2</th>
<th>group3</th>
<th>group4</th>
</tr>
</thead>
<tbody>
<tr>
<td>group2</td>
<td>-1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>group3</td>
<td>-1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>group4</td>
<td>-1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>group3</td>
<td>0</td>
<td>-1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>group4</td>
<td>0</td>
<td>-1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
```

where we made use of the `multcomp` R package (?). Other parameter functions of interest could be Dunnett’s many-to-one comparisons which requires another structure in $K$.

Assuming approximate or asymptotic normality of $\hat{\theta}$ means that $\hat{\vartheta}$ is also asymptotically normally distributed. Let the variance-covariance matrix of $\hat{\theta}$ be $\Sigma_{\theta}$,
then

\[ \hat{\theta} \sim N(\theta, \Sigma_\theta) \] (18)

The variance-covariance matrix of \( \hat{\theta} \) is then

\[ \Sigma_\theta = K \Sigma_\theta K^T \] (19)

and it follows that

\[ \hat{\theta} \sim N(\bar{\theta}, \Sigma_\theta) \] (20)

We can use the Wald test to compute p-values for these parameter functions individually since under the null hypothesis:

\[ Z = (\hat{\theta}_i - \theta_0)/se(\hat{\theta}_i) \sim N(0,1) \] (21)

where the standard errors are given by the square root of the diagonal elements of \( \Sigma_\theta \).

A test of the pair hypothesis in equation (17) can be performed as a Wald test with test statistic

\[ t_i = d'_{i,i}/se(d'_{i,i}) \]

which asymptotically under the null hypothesis follows a standard normal distribution.

The equivalent likelihood root statistic reads

\[ r_i(d'_{i},d'_{i'}) = \text{sign}(d'_{i} - d'_{i'}) \sqrt{2\{\ell_A(d'_{i},d'_{i'}) - \ell_0(d'_{i,i'})\}} \] (22)

where \( \ell_0(d'_{i,i'}) \) is the log-likelihood under the null hypothesis.

### 3.4 Letter display of groups based on pairwise differences

A convenient way of summarizing the results of a doing all pairwise comparisons is the so-called compact letter display. Here a letter is assigned to each of the groups based on all pairwise comparisons of the groups in such a way that

1. Two groups sharing a letter are not significantly different.
2. Two groups not sharing a letter are significantly different.

A particular letter assignment depends on the \( \alpha \)-level chosen for the letter display and on possible multiplicity adjustments of the p-values. An example of a letter display is given in section ??.

In this implementation we use the non-exported function insert_absorb from the multcomp package (?) implementing the insert-absorb algorithm of ?.
4 Conclusion

References


Analysis of multiple d-primes obtained from various discrimination protocols
Appendix D

Analysis of the Data Using the R package sensR

Analysis of the data
Per Bruun Brockhoff and Christine Borgen Linander
15 March 2017

Introduction

This chapter will cover in more detail how to actually analyze sensory discrimination data. This will include both hypothesis testing by p-value computation as by the use of critical values as confidence intervals. And this will include as well discrimination as similarity focused analyses. It
will also include d-prime calculations in several settings together with replicated data analysis by (corrected) beta-binomial models. It will also include a number of perspectivising tutorials.

The open source software \textit{R} will be used throughout, where the \textit{R}-package \textit{sensR} will play a major role, but other packages, e.g. the ordinal package, may also be used. The general form will be tutorial like with specific examples taken from the main protocol chapters of the book. And it will be shown how \textit{R} and the package can be used instead of statistical tables often otherwise reproduced in textbooks.

\textbf{Introduction to and overview of the \textit{sensR} package}

The \textit{sensR} package is an \textit{R} package for the analysis of data from sensory discrimination testing developed by Christensen and Brockhoff (2015). The package facilitates, among other things:

1. Statistical analysis of a broad range of sensory discrimination data
2. Power and sample size computations
3. Thurstonian analyses via d-prime estimation
4. Facilitating a link to generic statistical modelling
   a. Improved confidence intervals via profile likelihood methods.
   b. Allowing for ANOVA and regression modelling in a Thurstonian framework

Table 1 (below) describes which sensory discrimination methods in writing are supported by the \textit{sensR} package and which features the \textit{sensR} package provides for these discrimination methods. Absent check marks indicate that the feature is not implemented.

\begin{table}[h]
\centering
\begin{tabular}{lcccccccccc}
\hline
 & d-prime & Diff & Simil & Power & SampSize & Simul & Like\textit{CI} & Reps & glm & dcompare \\
\hline
Duo-trio, triangle, tetrad & X & X & X & X & X & X & X & X & X & X \\
2-AFC, 3-AFC & X & X & X & X & X & X & X & X & X & X \\
Double triangle, duo-trio & X & X & X & X & X & X & X & X & X & X \\
Double 2-AFC, 3-AFC & X & X & X & X & X & X & X & X & X & X \\
Unspecified 2-out-of-5 & X & X & X & X & X & X & X & X & X & X \\
Unsp. 2-out-of-5 with forgiveness & X & X & X & X & X & X & X & X & X & X \\
\hline
\end{tabular}
\end{table}
Table 1: These different protocols and analyses are currently explicitly supported in the sensR package (together with the ordinal package)

The R environment for statistical computations is developing constantly, and the sensR-package will also constantly extend its applicability and scope. The chapter represents the scope at time of writing and there could very well be new opportunities already implemented when you read this. One generic point is that some simple R-scripts will be shared with you in the chapter showing how to perform (some of) the analyses in a script-based way of running R. These scripts will run directly in your R Console, if copied directly. These scripts are also shared as supplementary material at the book website.

Basic single proportion of correct data

In the first part of this chapter the protocols providing a single proportion of correct answers are treated. Within the book this would include the duo-trio, 2-AFC, 3-AFC, triangle, tetrad, 2-out-of-5 and dual standard protocols. Some of such protocols are "fully supported" by sensR, others are not, see Table 1 above. As the package develops over time the collection of protocols fully supported will increase. In the package, also the so-called "double versions" of the triangle, duo-trio, 2-AFC and 3-AFC are also fully supported. Examples of analysing data from fully supported protocols as non-supported protocols will be given. In the latter case, other R-features together with the sensR package options can still offer nice analysis of the data.

For all the protocols that are "fully supported" by sensR the following things can be easily done and found, and in exactly the same way for each protocol by just choosing the proper method option in the R-functions:

1) Difference test (exact, likelihood or normal approximation based)
2) Similarity test (exact, likelihood or normal approximation based)
3) Estimation and Confidence intervals (exact, likelihood or normal approximation based) for
a) The proportion of correct $p_c$
b) The proportion of momentary discriminators $p_d$
c) The d-prime

4) Power calculations for as well difference as similarity tests:
   a) Based on $p_d$-alternatives
   b) Based on dprime-alternatives

5) Sample size calculations for as well difference as similarity tests:
   a) Based on $p_d$-alternatives
   b) Based on dprime-alternatives

6) Replicated data analysis based on the corrected (and standard) beta-binomial model

7) Simulation of replicated sensory protocol data

8) Offering the psychometric link functions to perform "Thurstonian regression/anova/ancova" analysis in cases with more challenging design structures using the generic generalized linear model features of R, e.g. the glm-function, cf. Brockhoff and Christensen (2010)

9) Easy transformations between the three "levels of interpretation": dprime, $p_c$ and $p_d$ by various transformation functions.

10) Plotting of the Thurstonian distributions

The plan is to exemplify most of these possibilities first. Then the second part of the chapter will cover other protocols, such as A-not A, same-different, with/without sureness scales, degree-of-difference, ABX, ranking data, R-index computation etc.

The analysis of the basic discrimination test data - difference and similarity

Assume that we had $x = 15$ correct out of $n = 20$ tetrad tests. Before using the sensR for the first time it must be installed from the internet (R CRAN) on your local computer. And obviously you would need to first install the R software itself, and it is strongly recommended to also install Rstudio (https://www.rstudio.com/), as a really nice way to run the R programme. When the package has been installed, in Rstudio: simply click Packages and Install and write sensR, you must load the package whenever initiating an R session, where it is to be used. This and the basic analysis of the tetrad case data is carried out as:

```R
library(sensR)
discrim(15, 20, method = "tetrad", conf.level = 0.90)
```

```
## Estimates for the tetrad discrimination protocol with 15 correct answers in 20 trials. One-sided p-value and 90 % two-sided confidence intervals are based on the 'exact' binomial test.
##
##       Estimate Std. Error  Lower  Upper
## pc    0.750  0.096821 0.5444 0.8959
## pd    0.625  0.145241 0.3166 0.8439
```
This is the basic analysis carried out by the discrimination test function `discrim` of the `sensR` package and having read the relevant parts earlier in this book (chapter 2 and chapter 9) the output is almost self explanatory. And at this point a remark to the reader with no prior experience with R and a certain skepticism towards using script based statistical software like this: To produce this result, you simply have to copy-and-paste a single script line into the R console. And if you use Rstudio, you can have the scripts in a separate subwindow, and you can download the script file with everything from this chapter to get started, and submit either single or several script lines easily with the inbuilt run drop down menu or short cut keys of Rstudio, or just basic cut-and-paste. This is the way to use R: Google and find-and-copy what other people did, or use the inbuilt R help-functionality, e.g. as:

```r
?discrim
```

At the bottom of the help pages of all functions in R, there will be example code that can be copied, used and adapted easily.

The basic idea of the discrimination data analysis in `sensR` is that all three ways of interpreting the results: $p_c$, $p_d$ and dprime are given in parallel. One may say that there is really only one statistical analysis (hypothesis test and confidence interval) but the results can be interpreted at the three different "levels", see also Næs et al (2010), Chapter 7. As an illustration of this, one may find the exact binomial 90% confidence interval for the proportion of correct $p_c$ based on generic binomial statistical methods, using the `binom` package, Dorai-Raj (2014), which then has to be installed first as described above.

```r
library(binom)
binom.confint(15, n = 20, conf.level = 0.90, methods = "exact")
```

And then we can transform the estimate and the lower and upper confidence limits using one of the in `sensR` inbuilt transformation utility functions: (again, the name of the function is self explanatory - a `pd2pc` function also exist)

```r
pc2pd(c(0.75, 0.5444176, 0.8959192), Pguess = 1/3)
```

Note how this is exactly the results provided for $p_d$ in the `discrim` function output above. And note that the `c` stands for "concatenate" and is the way to define basic lists of numbers, vectors, in R. And applying a function to a list of numbers makes R apply the function to each element of the vector and create a similar vector of results. And finally, the results (the $p_c$-values and the two CI-
values) can also be directly transformed to the underlying sensory scale by the inbuilt inverse
psychometric function: (the actual psychometric function is also available)

```r
psyinv(c(0.75, 0.5444176, 0.8959192), method = "tetrad")
```

## [1] 1.889770 1.176030 2.604497

And we have reproduced the results on the underlying sensory scale from the `discrim` function
output above. There is in addition the nice function `rescale`, that automatically, transforms to all
three scales. The following three calls would all produce the same results as already seen:

```r
rescale(pc = c(0.75, 0.5444176, 0.8959192), method = "tetrad")
rescale(d.prime = c(1.889770, 1.176030, 2.604497), method = "tetrad")
rescale(pd = c(0.6250000, 0.3166264, 0.8438788), method = "tetrad")
```

All these transformation utility functions automatically handles the cut-off at the guessing
probability in the proper way. Also by using either the default `statistic = exact` option or the
optional `statistic = likelihood` option to get likelihood based confidence intervals instead,
the function can and will find the proper confidence intervals also in extreme cases of observed
$p_c = 1$ or at or below the guessing level. This is something which the classic procedures for using
the variance of dprime to achieve a dprime ±1.96SE 95% confidence interval for the dprime
cannot do. In such cases the SE is simply not computable. Also, generally this so-called Wald-
based principle of finding a confidence interval in this classical way will give different results
depending on which scale that you decide to use: $p_c$, $p_d$, dprime, or any nonlinear function of
these. They cannot all be correct, and it is generally impossible to know which of all these
performs the best, that is, has the most correct coverage probability. But actually, it is well known
that the likelihood based confidence intervals is the optimal choice in this case. And the
likelihood interval is so-called invariant to monotone transformations, or differently put: The
likelihood and basic probability theory supports the simple transformations of the interval
between any of the three (or yet other) scales of interpretation. The choice between the "exact"
and "likelihood" options is a subtlety, which for most practical applications will not be important.
Some actually argue, to many people surprisingly, that the likelihood based intervals are superior
to the exact ones. But both will generally be fine and both are superior to the wald based
intervals that are only valid for subsequently large sample sizes, whereas the other two works
fine also for small samples, and as mentioned also handles extreme observations properly.

One can also easily identify the critical value of the difference test by the little `findcr` function:

```r
findcr(20, alpha = 0.05, p0 = 1/3)
```

## [1] 11

So with 11 or more correct answers the decision would be that the products are different. Or, to
turn towards similarity testing, the same function is also prepared to find the critical value for a
$\alpha = 0.05$ similarity test defined in terms of the $p_d$-value. E.g. if similarity is specified at $p_d \leq 0.50$,
it can be used as:

```r
findcr(20, alpha = 0.05, p0 = 1/3, pd0 = 0.50, test = "similarity")
```
And we see that with $n = 20$, we would decide in favor of $p_d \leq 0.5$-similarity with 9 or less correct answers in the tetrad. Or for that sake also using the triangle or the 3-AFC, as they have the same guessing probability $1/3$. Critical values for similarity tests based on a dprime-specification of similarity, e.g. $dprime = 1.5$, could then be found by combining the transformation functions and the $p_d$-based findcr function:

```r
findcr(20, alpha = 0.05, p0 = 1/3, pd0 = pc2pd(psyfun(1.5, "tetrad"), Pguess=1/3), test = "similarity")
```

showing that one needs to have 8 or less correct answers to have shown $dprime \leq 1.5$ similarity. The stronger requirement comes from the fact that the $p_d$-value corresponding to a $dprime$ of 1.5 in a tetrad test is smaller than 0.50:

```r
pc2pd(psyfun(1, "tetrad"), Pguess=1/3)
```

and hence it becomes more difficult to show similarity. The reason for choosing the "conf.level=0.90" rather than the default of 95% in the initial call to the discrim function above is in fact that with this choice one may with the same call and result output already also carry out the yes/no/critical value-version of any $\alpha = 0.05$ one-tailed similarity tests defined on any of the three scales! This is so because the upper limit of the confidence interval can then be used for this: Any similarity definition can directly be compared with this limit. E.g. the results above would tell us that we have shown a $p_d \leq 0.9$ similarity, or a $dprime \leq 2.7$ similarity but any stricter than such rather silly liberal definitions of similarity would not be shown with 15 correct out of 20 in a tetrad test.

The $p$-value from the similarity test is not seen from the results above, but could be obtained using the same function with a different option, e.g. if we have specified the similarity definition at $dprime \leq 1.3$ and we observed 28 out of 60 correct in a tetrad test:

```r
discrim(28, 60, d.prime= 1.3, method = "tetrad", conf.level = 0.90, test = "similarity")
```

```
## Estimates for the tetrad discrimination protocol with 28 correct answers in 60 trials. One-sided p-value and 90 % two-sided confidence intervals are based on the 'exact' binomial test.
##
##           Estimate Std. Error    Lower    Upper
## pc       0.4667     0.06441   0.35558   0.5803
## pd       0.2000     0.09661   0.03338   0.3705
## d-prime  0.9001     0.24377   0.35091  1.2968
##
```
## Result of similarity test:
'exact' binomial test:  p-value = 0.04845
Alternative hypothesis: d-prime is less than 1.3

So we just barely have shown dprime = 1.3-similarity, as the p-value is smaller than 0.05 in line with the upper limit of the 90% confidence-interval being just below 1.3. Also note that with 28 correct one would also find a significance in the $\alpha = 0.05$ difference test, as the lower limit(s) of the 90% confidence intervals are above the guessing level. This is in no contradiction with the similarity conjecture, but illustrates one of the basic problems of showing similarity by non-significance of a difference test. And note that the 2-tailed 90% confidence intervals are what they are, and are not affected by the choice of which hypothesis test p-value to provide. See also the subsection on the two different approaches to similarity testing in the next section.

The perceptual Thurstonian distributions can be easily plotted as follows:

```
plot(discrim(15, 20, method = "tetrad"), cex.main=0.8)
```

![Distribution of sensory intensity for the tetrad test](image)

Case study on 2-out-of-5

Even though this protocol was recently added as a fully supported protocol, let us illustrate how one could use the sensR package for the analysis of data from a protocol which is not fully supported in sensR. The data from Chapter 9 is analyzed here. Data from any basic test protocol could still be analysed on the proportions scale $p_c$ and $p_d$ even without an implementation of the psychometric function. If the guessing probability matches one of protocols already supported by the discrim function (1/2, 2/5, 1/3, 1/4, 1/9 or 1/10), one could still run all the functions using a
version with the right guessing probability and interpret all output that has nothing to do with d-prime values, as only the dprime-computation would be off the point.

As above, the exact 90%-confidence limits for each of the four data examples can be found by a generic function as:

```r
CIres <- binom.confint(c(3, 7, 4, 6), 20, conf.level = 0.90, methods = "exact")
```

And then transformed to $p_d$-scale using a sensR utility function:

```r
pc2pd(as.matrix(CIres[,4:6]), Pguess = 1/10)
```

```
##         mean      lower     upper
## 1 0.05555556 0.00000000 0.2707376
## 2 0.27777778 0.08590102 0.5089272
## 3 0.11111111 0.00000000 0.3344757
## 4 0.22222222 0.04394861 0.4531316
```

Here it can be seen what level of $p_d$-similarity can be proven and also which differences were shown. The exact p-values for either a similarity test or the difference test could also easily be found by the base `binom.test` function, e.g.:

```r
binom.test(3, 20, p=1/10, alternative = "greater")
```

```
## Exact binomial test
## data:  3 and 20
## number of successes = 3, number of trials = 20, p-value = 0.3231
## alternative hypothesis: true probability of success is greater than 0.1
## 95 percent confidence interval:
##  0.04216941 1.00000000
## sample estimates:
## probability of success
##                   0.15
```

The procedure provides the so-called one-tailed confidence interval, which generally we do not use here. It would not require much, if you know the mathematical expression for a psychometric function to implement that yourself, and then use this together with all the features of sensR via the utility transformation functions, and hence you could in addition to the basic analysis above get to have a fully supported protocol. In the section on analysing data from ABX discrimination tasks, an example of how this could be done is given.

**The planning of the basic discrimination test data - difference and similarity**

The power of the $n = 20$ tetrad $\alpha = 0.05$ difference test with an alternative $p_d$-value if 0.5 can be found as:

```r
discrimPwr(pdA = 0.5, sample.size = 20, alpha = 0.05, pGuess = 1/3)
```
Note how this can provide the power for any basic protocol as long as you just know the guessing probability. As for findcr above one could then also find power based on alternatives expressed on the dprime scale using the transformation utility functions. However, this was already implemented into a similar power function, so the probability of detecting a dprime of 1.5 could be found directly as:

```r
d.primePwr(d.primeA = 1.5, sample.size = 20, alpha = 0.05, method = "tetrad")
```

And to illustrate the general options and show the classical results of the power ranking of the basic protocols, in the following the similar power for the five basic protocols are found in a little loop: (PBB: we might extend the list with the protocols we manage to implement before the chapter goes in print)

```r
d.primePwr(d.primeA = 1.5, sample.size = 20, alpha = 0.05, method = "duotrio")
d.primePwr(d.primeA = 1.5, sample.size = 20, alpha = 0.05, method = "triangle")
d.primePwr(d.primeA = 1.5, sample.size = 20, alpha = 0.05, method = "tetrad")
d.primePwr(d.primeA = 1.5, sample.size = 20, alpha = 0.05, method = "twoAFC")
d.primePwr(d.primeA = 1.5, sample.size = 20, alpha = 0.05, method = "threeAFC")
```

<table>
<thead>
<tr>
<th>Method</th>
<th>Power</th>
</tr>
</thead>
<tbody>
<tr>
<td>duotrio</td>
<td>0.2868</td>
</tr>
<tr>
<td>triangle</td>
<td>0.4348</td>
</tr>
<tr>
<td>tetrad</td>
<td>0.8596</td>
</tr>
<tr>
<td>twoAFC</td>
<td>0.9423</td>
</tr>
<tr>
<td>threeAFC</td>
<td>0.9914</td>
</tr>
</tbody>
</table>

Without an implementation of the double versions in d.primePwr, one may use the \( p_d \)-based power function via the transformation functions:

```r
discrimPwr(pdA = pc2pd(psyfun(1.5, "duotrio", double = TRUE), Pguess = 1/4), sample.size = 20, alpha = 0.05, pGuess = 1/4)
discrimPwr(pdA = pc2pd(psyfun(1.5, "triangle", double = TRUE), Pguess = 1/9), sample.size = 20, alpha = 0.05, pGuess = 1/9)
discrimPwr(pdA = pc2pd(psyfun(1.5, "twoAFC", double = TRUE), Pguess = 1/4), sample.size = 20, alpha = 0.05, pGuess = 1/4)
discrimPwr(pdA = pc2pd(psyfun(1.5, "threeAFC", double = TRUE), Pguess = 1/9), sample.size = 20, alpha = 0.05, pGuess = 1/9)
```

Without an implementation of the double versions in d.primePwr, one may use the \( p_d \)-based power function via the transformation functions:

```r
discrimPwr(pdA = pc2pd(psyfun(1.5, "duotrio", double = TRUE), Pguess = 1/4), sample.size = 20, alpha = 0.05, pGuess = 1/4)
discrimPwr(pdA = pc2pd(psyfun(1.5, "triangle", double = TRUE), Pguess = 1/9), sample.size = 20, alpha = 0.05, pGuess = 1/9)
discrimPwr(pdA = pc2pd(psyfun(1.5, "twoAFC", double = TRUE), Pguess = 1/4), sample.size = 20, alpha = 0.05, pGuess = 1/4)
discrimPwr(pdA = pc2pd(psyfun(1.5, "threeAFC", double = TRUE), Pguess = 1/9), sample.size = 20, alpha = 0.05, pGuess = 1/9)
```

<table>
<thead>
<tr>
<th>Method</th>
<th>Power</th>
</tr>
</thead>
<tbody>
<tr>
<td>double_duotrio</td>
<td>0.5506</td>
</tr>
<tr>
<td>double_triangle</td>
<td>0.4092</td>
</tr>
<tr>
<td>double_twoAFC</td>
<td>0.9982</td>
</tr>
<tr>
<td>double_threeAFC</td>
<td>0.9976</td>
</tr>
</tbody>
</table>

And with a little bit of R-coding not shown here, all the entire power functions can be plotted next to each other:
Similar to all the power computations above, one may also find similarity test power, either based on \( P_d \)-specifications or as exemplified here based on dprime-specifications:

\[
d.\text{primePwr}(d.\text{primeA} = 0, d.\text{prime0} = 1, \text{sample.size} = 20, \alpha = 0.05, \text{method} = \text{"tetrad"}, \text{test} = \text{"similarity"})
\]

\[
## [1] 0.2972139
\]

\[
d.\text{primePwr}(d.\text{primeA} = 0, d.\text{prime0} = 1, \text{sample.size} = 100, \alpha = 0.05, \text{method} = \text{"tetrad"}, \text{test} = \text{"similarity"})
\]

\[
## [1] 0.9341278
\]

showing that the power of detecting a dprime \( \leq 1 \) similarity with \( n = 20 \) and the assumption that there is truly no difference at all (dprimeA=0) is much too low but nicely high with \( n = 100 \). The additional investigation one could make here is to look into the power, in case that the alternative is not the most optimistic scenario: What if the alternative is really at a smaller difference, e.g. d.primeA=0.5, that is, still a similarity setting, as similarity is defined at a value of 1:

\[
d.\text{primePwr}(d.\text{primeA} = 0.5, d.\text{prime0} = 1, \text{sample.size} = 100, \alpha = 0.05, \text{method} = \text{"tetrad"}, \text{test} = \text{"similarity"})
\]

\[
## [1] 0.7150384
\]

Of course, similarity would then only be detected with much lower power. This would also be seen from the two corresponding sample size computations:
Similarity test sample sizes can in a similar way also be found based on \( p_d \)-specifications and difference test sample sizes can be found exactly similar either \( p_d \)-based or dprime-based: (\( \alpha = 0.05 \) and \( \text{test} = \text{"difference"} \) are default settings)

```r
d.primeSS(d.primeA = 1, target.power = 0.9, method = "tetrad")
## [1] 82
```

```r
discrimSS(pdA = 0.2407, target.power = 0.9, pGuess = 1/3)
## [1] 82
```

```r
rescale(d.prime=1, method="tetrad")
```

As all powers, sample sizes and critical values are available via the functions showed, anyone could design and make any kinds of tables or figures based on these functions and based on any choice of effect sizes, powers, sample sizes, alpha levels, test types and interpretation levels. It is part of ongoing developments to extend the sensR package with such table and figure facilities in a user friendly way.

**Two approaches for similarity testing planning and analysis**

Above and in the sensR package in general the approach to similarity testing is to be formal about which hypothesis is being tested and the Type I (\( \alpha \)) and Type II (\( \beta \)) risks related to that. Compared to the more commonly expressed hypotheses used for difference testing, this amounts to a swap between the use of the null and the alternative hypotheses. For similarity testing purposes the null now includes the "difference statement" and the alternative the "similarity statement", and you then have to consider and define explicitly what the latter really is in the context. It is not uncommon, and used in standards also, to handle this "swapping" by instead swapping the roles of \( \alpha \) and \( \beta \) and then making sure that the power of the difference test is high enough, and claim similarity if there is no significant difference found.

Whereas the latter approach, if used carefully, is valid, it does have its limitations and in the point of view of this author rapidly challenges the understanding and communication of results:
1. Generally it appears questionable (for an outsider) to claim a significance by noting an NS result.
2. If you for some reason suddenly increase the $n$, it is indeed invalid in general!
3. Swapping the names of what is really the Type I and Type II risks for what you are doing may be difficult to understand and communicate.

The benefit of the approach of the standards, combined with actually providing additional tables of critical values for the proper similarity tests to be used for any choice of $n$, was of course that already available difference test power and sample size tables could be used. With the flexibility of the implementations of this in sensR this simplicity argument is no longer relevant.

Let us consider the example of Chapter 9: The aim is to perform a level 10% similarity test to be able prove a d-prime similarity of 1 with power 80%. So for the similarity test the $\alpha = 0.10$ and $\beta = 0.20$ and the definition of similarity is given by the effect being less than or equal to 1. In the standards approach as also taken in Chapter 9, the roles of $\alpha$ and $\beta$ are swapped and the required sample size is found as

```r
d.primeSS(d.primeA = 1, target.power = 0.9, alpha = 0.20, method = "tetrad")
d.primeSS(d.primeA = 1, target.power = 0.9, alpha = 0.20, method = "tetrad",
statistic = "stable.exact")
```

yielding 47 and 49 respectively. The default of the d.primeSS function is to find the difference test sample size, and the default is also to find the smallest $n$ with the required power, the first R-call and the number 47. By specifying "stable.exact", the smallest $n$ for which no larger one has a lower power is found, the second R-call and the number 49. The actual beta-risk for the difference test, corresponding to the alpha-level for the similarity test is:

```r
1-d.primePwr(1, sample.size = 49, method = "tetrad", alpha=0.2)
```

```r
## [1] 0.08943058
```

The critical value for the difference test is found to be

```r
findcr(49, alpha = 0.2, p0 = 1/3 )
```

```r
## [1] 20
```

so with 20 or more one would declare difference and with 19 or less one would conclude similarity. The similarity test critical value can also be found directly as, now using the real similarity test alpha-level as the $\alpha$-level in the call: (and using the $p_d$-definition of similarity coming from the tetrad psychometric function and $d_{prime}=1$, which is 0.24)

```r
pc2pd psyfun(1, method = "tetrad", 1/3)
```

```r
## [1] 0.2407126
```

```r
findcr(49, alpha = 0.1, p0 = 1/3, test = "similarity", pd0 = 0.2407126)
```

```r
## [1] 19
```
So this is completely in line with the decision criterion found from the difference test critical value found above. If you analyze the data \( x = 16 \) using the difference test analysis we get the reported difference test p-value of 0.5933. And we know the conclusion of the similarity test, which we similarly knew directly by noting that \( x = 16 \) is less than 20. With the choice of level such that the upper confidence limits can used for the wanted similarity test

\[
\text{discrim(16, 49, conf.level = 0.80, method = "tetrad")}
\]

## Estimates for the tetrad discrimination protocol with 16 correct answers in 49 trials. One-sided p-value and 80 % two-sided confidence intervals are based on the 'exact' binomial test.

##
## Estimate Std. Error  Lower  Upper
## pc       0.3333         NA 0.3333 0.4269
## pd       0.0000         NA 0.0000 0.1403
## d-prime  0.0000         NA 0.0000 0.7407

## Result of difference test:
## 'exact' binomial test:  p-value = 0.5933
## Alternative hypothesis: d-prime is greater than 0

we see the reported difference test p-value. The actual p-value for the difference test is not known from this but could be found to be 0.01333 from:

\[
\text{discrim(16, 49, conf.level = 0.80, method = "tetrad", test="similarity", pd0 = 0.2407126)}
\]

## Estimates for the tetrad discrimination protocol with 16 correct answers in 49 trials. One-sided p-value and 80 % two-sided confidence intervals are based on the 'exact' binomial test.

##
## Estimate Std. Error  Lower  Upper
## pc       0.3333         NA 0.3333 0.4269
## pd       0.0000         NA 0.0000 0.1403
## d-prime  0.0000         NA 0.0000 0.7407

## Result of similarity test:
## 'exact' binomial test:  p-value = 0.01333
## Alternative hypothesis: pd is less than 0.2407

What if we finally did out study with a larger \( n \) than planned, eg. using \( n=100 \)? We couldn't find the critical value from the difference test anymore:

\[
\text{findcr(100, alpha = 0.2, p0 = 1/3 )}
\]

## [1] 38

14
findcr(100, alpha = 0.1, p0 = 1/3, test = "similarity", pd0 = 0.2407126)
## [1] 42

So, properly with 42 or less one would declare similarity, and not with just 37 or less.
But the analysis from sensR can be used no matter what n was chosen and used, e.g. with the outcome x=42:

```r
discrim(42, 100, conf.level = 0.80, method = "tetrad",
       test="similarity", pd0 = 0.2407126)
```

```r
##
## Estimates for the tetrad discrimination protocol with 42 correct answers in 100 trials. One-sided p-value and 80 % two-sided confidence intervals are based on the 'exact' binomial test.
##
##          Estimate Std. Error  Lower  Upper
## pc         0.420    0.04936 0.3535 0.4890
## pd         0.130    0.07403 0.0302 0.2335
## d-prime    0.711    0.21733 0.3335 0.9827
##
## Result of similarity test:
## 'exact' binomial test:  p-value = 0.08415
## Alternative hypothesis: pd is less than 0.2407
```

The power of the n=100 similarity test could also be easily found:

```r
d.primePwr(d.prime0=1, sample.size = 100, method = "tetrad", alpha=0.1,
            test="similarity", d.primeA = 0)
```

```r
## [1] 0.9724325
```

### Analysing replicated difference test

When N assessors/consumers each performed K discrimination tests we face the additional challenge in the data analysis that persons may be heterogeneous in their response. The betabin function in sensR offers as well the standard beta-binomial as the corrected beta-binomial analysis of replicated data. For paired comparison data where the natural alternative is two-tailed the standard beta-binomial would be the natural choice. The natural model for the one-tailed setting corresponding to the standard assumption of equal perceptual variability in test and reference products is the corrected beta-binomial, which is also the default choice by the R-function betabin.

Before the analysis is exemplified, it is illustrated what the effect of a replicated design could be by simulations using the sensR simulation function discrimSim. The results of having 15 assessors doing each 10 tetrad test in a dprime=1 scenario with a latent standard deviation of individual dprimes of 1 could be simulated by the following:

```r
discrimSim(15, 10, d.prime = 1, method = "tetrad", sd.indiv=1)
```
giving the number of correct answers (out of 10) for each of the 15 assessors. In four different heterogeneity scenarios it could look as follows:

In the top left plot the variability seen is what could be called "usual binomial" variability, whereas the other three plots show an increasing level of so-called "over-dispersed" data. A proper analysis will estimate as well the average dprime-level as well as the (extra) variability. The model used for the simulations in discrimSim is in fact not the corrected beta-binomial model, but rather a latent random varying random d-prime model with the restriction that the individual d-prime could never be lower than 0.

To do the analysis of replicated data the data has to be stored in a matrix or a data.frame with two columns; first column containing the number of success and the second the total number of tests for each assessor. The number of rows should correspond to the number of assessors. In the following the example data from Næs et al (2010) is analysed as if it was the result of tetrad tests, corresponding to \( N = 15 \) assessors each having performed \( K = 12 \) tetrad tests:

```r
# Replicated data from the Næs et al book:
x <- c(2, 2, 3, 3, 4, 4, 4, 4, 4, 4, 5, 6, 10, 11)
X2 <- cbind(x, 12)

# Analyzed as tetrad data
summary(betabin(X2, method = "tetrad"), level = 0.9)
```

```r
## Chance-corrected beta-binomial model for the tetrad protocol
```
The corrected beta-binomial model has two unknown parameters the mean $\mu$ and the scale $\gamma$ expressing the over-dispersion and they are estimated by maximum likelihood. The mean is the same as $p_d$ and can then be transformed to as well the $p_c$ as the dprime scale through the relevant psychometric function. Similarly the standard so-called Wald type confidence intervals for all the parameters are found from classical approximate likelihood theory. The hypothesis test for no product difference is a joint test for the mean being at the guessing level $1/3$ and the over-dispersion being non-existant. The test-statistic is in this case 15.492 and a p-value based on the $\chi^2(2)$-distribution is reported to be 0.0004325.

It is illustrative to compare with the "naive" analysis, where all data are simply pooled across persons:

```r
discrim(sum(x), 180, method = "tetrad", conf.level = 0.9)
```

```
## Estimates for the tetrad discrimination protocol with 70 correct answers in 180 trials. One-sided p-value and 90 % two-sided confidence intervals are based on the 'exact' binomial test.
##
##         Estimate Std. Error  Lower  Upper
## pc       0.38889    0.03634 0.3333 0.4525
## pd       0.08333    0.05450 0.0000 0.1787
## d-prime  0.56191    0.19206 0.0000 0.8455
##
## Result of difference test:
## 'exact' binomial test:  p-value = 0.0678
## Alternative hypothesis: d-prime is greater than 0
```

The hypothesis test part of the naive analysis can be argued to be valid as the null distribution is the right one. It becomes clear from the above though, that it would not be the most powerful analysis in cases with high levels of heterogeneity as here. The replicated analysis provides an extremely smaller p-value for detection of a product difference. This appears meaningful for these data where two of the assessor show strong effects with 10 and 11 correct out of 12. Having two out of 15 assessors showing such extreme data for 12 replications would only happen
very rarely if the null hypothesis of no product difference (and hence also no extra variation) were true.

Also compare the widths of the confidence intervals for the d'prime: They become wider in the replicated analysis, as expected, or at least as expected when it comes to the upper limit, due to the extra variability. Generally, the confidence intervals from the naive analysis cannot be used, as they do not incorporate the individual heterogeneity and hence are based on potentially wrong assumptions of independence/homogeneity. In fact the lower confidence limits provided in the replicated analysis is counter intuitive: In spite of a highly significant difference result, the interval includes 0! This illustrates one of the strong limitations of the methodological status quo of replicated difference testing analysis: It is entirely based on the so-called "asymptotic likelihood theory". The hypothesis tests are using \( \chi^2 \)-distributions and the confidence intervals are classical normal based \( \pm z_{1-\alpha/2} \) versions. And no one investigated how good these methods really are for various values of \( N \) and \( K \). For sure, many times the confidence intervals would not even be estimable, and even when they are, they are likely off the nominal coverage levels. Also the actual type I levels of the hypothesis tests are likely off the nominal levels. This could easily be investigated by simulation or enumeration studies. For e.g. \( N = 20 \) and \( K = 2 \) such (yet unpublished) studies show that the actual type 1 level of the \( \chi^2(2) \) (nominal) level \( \alpha = 0.05 \) test is only around 1%. And the proper level \( \alpha = 0.05 \) critical value for the likelihood ratio statistic is around 3.41 rather than the 5.99 from the \( \chi^2(2) \)-distribution. The field needs improved analysis methods for replicated data for these to be used more substantially, and the recommended "best practice" otherwise given in this book, that replicated data should not be used for similarity purposes, is meaningful in light of this.

With proper developed analysis tools, replicated data could be used for as well discrimination as similarity purposes: If we were completely on top of the Type I and II errors and confidence intervals, we would not do erroneous analysis for either purpose. Improved difference test hypothesis testing for replicated data was already discussed by Meyners (2007a) and Meyners (2007b), but the full methodology including the link to Thurstonian interpretation of the results including confidence intervals and as well difference test as similarity test approaches and power and sample size considerations still needs to be completely developed and implemented.

**Link to more general Thurstonian generalized linear modelling and d'prime comparisons**

The unreplicated analyses provided above for any of the supported protocols by the use of the discrim function corresponds in generic statistical terminology to a "one-sample" analysis of proportions data. And the d'prime estimation is a way to "link" the underlying sensory scale to the proportions scale through thepsychometric function. In Brockhoff and Christensen (2010) it is illustrated how this in fact is exactly the same as what in statistics is called "generalized linear models" (glm), which in short is the name for a way to do everything which can be done in normal distribution based linear modelling (two- and multisample comparisons, multifactorial ANOVA, regression, ANCOVA etc) for other types of responses and probability distributions, by expressing the linear mean part of the model for some non-linear link function of the mean of the actual data distribution. For binomial type data like treated here, well known examples of this are
logistic and probit regression type models where the logit \( \log(p/(1 - p)) \) and probit \( \Phi^{-1}(p) \) are the "link functions". In base R the function \( \text{lm} \) (=linear model) is the function that can analyze data based on any one of all the mentioned normal based linear models, and similarly the \( \text{glm} \) function can analyze data from such extended collection of distributions and settings.

The \( \text{glm} \) function in base R offers through the binomial family object a number of different potential non-linear link functions for the analysis of binomial data. Also the \( \text{glm} \) function allows for user specified link functions, if they are implemented in the right way. In sensR the various protocols supported in the \text{discrim} function, that is, currently 12 different ones, cf. Table 1, are implemented in such a way that the inverse psychometric functions can be used as generic link functions in an analysis by the \( \text{glm} \) function. In Brockhoff and Christensen (2010) this was only explicitly mentioned for the four original protocols supported by the \text{discrim} function. Now it holds for all 12 of them.

Let us re-analyze the (artifically) constructed experimental data from Brockhoff and Christensen (2010) as if it was the results of 160 tetrad tests: 80 males and 80 females were randomly allocated into 4 groups of 20 each. Each two set of groups then tested one of four products versus a reference product. The four products came with an increasing concentration level of some kind:

```r
## Generate the data:
data <- expand.grid(conc = 1:4, gender = c("Males", "Females"))
data$correct <- c(9, 11, 13, 14, 13, 14, 16, 18)
data$total <- rep(20, 8)
data$concGrp <- factor(data$conc)

## View data:
data

# conc gender correct total concGrp
# 1 1 Males 9 20 1
# 2 2 Males 11 20 2
# 3 3 Males 13 20 3
# 4 4 Males 14 20 4
# 5 1 Females 13 20 1
# 6 2 Females 14 20 2
# 7 3 Females 16 20 3
# 8 4 Females 18 20 4
```

So one approach could be to carry out 8 separate analyses with \( n = 20 \) in each analysis, e.g. by 8 calls to the \text{discrim} function. These 8 dprimes and their standard errors would also come directly out of the following glm analysis, where the option \texttt{family = tetrad} is used within the generic \texttt{glm} function of R:

```r
## Fit Initial model:
glm0 <- glm(cbind(correct, total - correct) ~ gender:concGrp - 1, data,
            family = tetrad)
summary(glm0)
```
The model syntax notation using ~ gender:concGrp - 1 means that each of the 8 groups is separated and "no intercept" is estimated. This is linked to basic ANOVA parametrizations and simply ensures that the results are summarized directly for each of the 8 groups, and not by some kind of contrasts. So this is in fact an 8 group one-way ANOVA analysis using the inverse tetrad psychometric function as the link function to ensure that the results are provided directly with a tetrad dprime interpretation. The two-way ANOVA decomposition could also be easily carried out following exactly the way this would be done for regular ANOVA type data:

```
"""
"""
"""
```

---

**Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1**

(Dispersion parameter for binomial family taken to be 1)

---

Null deviance: 9.0733e+01 on 8 degrees of freedom
Residual deviance: 1.5543e-15 on 0 degrees of freedom
AIC: 41.85

---

Number of Fisher Scoring iterations: 4

The model syntax notation using ~ gender:concGrp - 1 means that each of the 8 groups is separated and "no intercept" is estimated. This is linked to basic ANOVA parametrizations and simply ensures that the results are summarized directly for each of the 8 groups, and not by some kind of contrasts. So this is in fact an 8 group one-way ANOVA analysis using the inverse tetrad psychometric function as the link function to ensure that the results are provided directly with a tetrad dprime interpretation. The two-way ANOVA decomposition could also be easily carried out following exactly the way this would be done for regular ANOVA type data:
In the glm universe "ANODE" (analysis of deviance) tables are used to summarize various test results similar to the use of ANOVA tables. We see that the interaction is not significant, genders do not react differently to the four products. It turns out that the product differences can be fully described by a regression model as a linear function of the concentration levels but with two different (gender dependent) intercepts. First, this simpler model is fitted and compared to the general 8-parameter model, and found to fit the data adequately: (p-value is 0.9965)

Next we study the results of this analysis:

```r
summary(glm1)
```

## Call:
glm(formula = cbind(correct, total - correct) ~ gender + conc, family = tetrad, data = data)

## Deviance Residuals:
##          1          2          3          4          5          6          7
##  0.01198   0.10879   0.14734  -0.26259   0.10702  -0.30269  -0.11248
##          8
##  0.37055

## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
```
## (Intercept)     0.5070     0.3841   1.320   0.1869
## genderFemales   0.6615     0.2759   2.398   0.0165 *
## conc            0.3234     0.1262   2.563   0.0104 *
## ---
## Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
## (Dispersion parameter for binomial family taken to be 1)
## Null deviance: 13.05053 on 7 degrees of freedom
## Residual deviance:  0.35567 on 5 degrees of freedom
## AIC: 32.206
## Number of Fisher Scoring iterations: 3

And apart from the difference between gender dprimes of 0.66 the dprime generally is expected to increase with 0.32 for each increasing concentration unit (within the range of the data).

This little example serves as an illustration of how the sensR protocol implementations on top of all the detailed one-sample analyses provided also provides the practical connection to generic statistical modelling which is often called for when having carried out larger research or industrial experimental studies. And the connection allows for maintaining the Thurstonian model and the proper interpretation of the results in light of this. In several developments and uses of the ordinal package similar connections between Thurstonian modelling, complex experimental settings and generic statistical modelling including mixed model versions was done, cf. e.g. Christensen at al (2011) and Christensen at al (2012). The latter lead to the implementation of the 2-AC protocol in sensR via the functions twoAC and twoACpwr functions. The use of mixed model versions for 2-AFC data in replicated multi-product experiments is investigated in Linander et al (2017a).

The sensR package in addition offers detailed, novel and improved tools for the one-way ANOVA type comparisons of two or more d-primes from potentially different protocols in the functions dprime_test, dprime_compare and dprime_table. We refer to Linander et al (2017) for details and examples of this.

### Analysis of A-not A tests

The basic A-not A or Yes/No testing paradigm is probably the most well known in Signal Detection Theory (SDT), MacMillan and Creelman (2005). Consider an example with 8 "yes"-responses to yes-samples, 1 "yes"-responses to no-samples, 17 "no"-response to yes-samples and 24 "no"-responses to no-samples. The classical SDT approach of finding the d-prime would then be the subtraction of the two standard normal quantiles corresponding to the hit rate $H = \frac{8}{25}$ and false alarm rate $FA = \frac{1}{25}$:

```r
H <- 8/(8+17)
FA <- 1/(1+24)
zH <- qnorm(H)
zFA <- qnorm(FA)
```
## d-prime:
\[ z_H - z_{FA} \quad \# \text{d'} \]

## [1] 1.282987

Which can also be found like this by the SDT function of sensR:

```r
data <- rbind(c(8, 17),
              c(1, 24))
SDT(data)
```

```r
## z(Hit rate) z(False alarm rate) d-prime
## 1  -0.4676988           -1.750686 1.282987
```

The basic A-not A model and analysis was also in Brockhoff and Christensen (2010) identified as a probit version of a generalized linear model and implemented in sensR in its own function, which offers improved likelihood based confidence intervals and some plotting features:

```r
(m1 <- AnotA(8, 25, 1, 25))
```

```r
## Call: AnotA(x1 = 8, n1 = 25, x2 = 1, n2 = 25)
##
## Results for the A-Not A test:
##
##         Estimate Std. Error     Lower    Upper    P-value
## d-prime 1.282987  0.5243127 0.2553532 2.310621 0.01160771
## likelihood based confidence intervals:
## confint(m1)[2,]
##     2.5 %    97.5 %
## 0.3375385 2.4593495
par(mfrow = c(1, 2))
ROC(m1)
plot(m1, main = "")
```
The p-value provided by the AnotA function is the one-tailed Fisher's Exact Test p-value and the d-prime confidence interval is the Wald based using the standard error also given. The \texttt{confint} call provides the better likelihood based confidence interval for d-prime. The left plot shows the model estimate of the so-called ROC-curve, see more details in the next section, as the solid curve. The dotted curves show the 95\% confidence interval of the ROC curve. The plot to the right is just showing the perceptual distribution on the standard normal scale - the shift in the distributions is the d-prime value 1.28.

**A-not A with sureness**

Sometimes the A-not A protocol is extended by the use of a sureness scale with e.g. three levels of sureness for each of the two possible responses. This then leads to a 6-level ordinal response scale. In Christensen et al. (2011) it is covered in quite some detail how the Thurstonian model in this case directly becomes a so-called cumulative link model with the probit link, and using the \texttt{clm} and \texttt{clmm} functions of the \texttt{ordinal} package, Christensen (2015) can fit these models including all sorts of extended versions of the basic Thurstonian model including mixed model versions. Below a simple example of this is shown as part of the same-diff with sureness analysis.

**Analysis of Same - different tests**

**Case study 1, Chapter 2: The Same Different Test**

The basic $\chi^2$-analysis of the same different data can be done by one of the inbuilt R-functions for this either in the uncorrected version as found in Chapter 2 or the so-called continuity corrected version, where the $\chi^2$-distributional assumption is usually better. (The latter is the default choice by the R function)
# The standard uncorrected pearson test:
chisq.test(matrix(c(21, 21, 9, 33), ncol=2), correct=F)
##
##  Pearson's Chi-squared test
##
## data:  matrix(c(21, 21, 9, 33), ncol = 2)
## X-squared = 7.4667, df = 1, p-value = 0.006285

# The standard corrected pearson test:
chisq.test(matrix(c(21, 21, 9, 33), ncol=2))
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  matrix(c(21, 21, 9, 33), ncol = 2)
## X-squared = 6.2741, df = 1, p-value = 0.01225

The reported p-value is the two-tailed version, so to get the proper one-tailed p-value for the relevant hypothesis test in the same different setting, this should be divided by 2, so p-value = 0.0061 (or 0.0031). The critical value for the one-tailed \( \alpha = 0.05 \) \( \chi^2 \)-test could also be easily found as:

qchisq(0.9, 1)
## [1] 2.705543

The same different test is treated in much detail in Christensen & Brockhoff (2009), where a thorough thurstonian modelling including likelihood based analysis approach is given. All of this was implemented in some easy to use functions within sensR. Some of the benefits of this are that improved likelihood confidence intervals are available, which also works for extreme data cases (e.g. all same or all diff outcomes), and of course the Thurstonian d-prime together with the "tau" decision criterion are estimated including CIs.

The same-diff analysis is simply carried out in R as follows:

sdres <- samediff(21, 21, 9, 33)
summary(sdres)
##
## Call:
## samediff(nsamesame = 21, ndiffsame = 21, nsamediff = 9, ndiffdiff = 33)
##
## Coefficients
##       Estimate Std. Error  Lower  Upper P-value
## tau     0.9539     0.1717 0.6460 1.3157 < 2e-16 ***
## delta   1.9841     0.4157 1.0213 2.7626 0.00288 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## Sensitivity, AUC value with CI:

AUC(sdres$coef[2], sdres$se[2])

# AUC: 0.919684
# 0.95% CI: [0.7958242, 0.9760976]

Note how the uncorrected one-tailed $\chi^2$ based p-value rounded becomes 0.003 just like the reported p-value for the delta (d-prime) in the R-output. The AUC-value is, like the dprime, another measure of difference between products, also called the sensitivity, and is the probability that a random sample from the low-intensity distribution has a lower intensity than a random sample from the high-intensity distribution, and is linked to the dprime in a simple way, $\Phi(d'/\sqrt{2})$:

```r
pnorm(1.9841/sqrt(2))
```

## 0.9196872

The power of same-diff tests is also available, e.g. like:

```r
samediffPwr(n = 100, tau = 1, delta = 2.5, Ns = 10, Nd = 10)
```

## [1] 0.56

### Case study 2, Chapter 2: The Same Different Test with sureness

The sureness-scale used in this case produces two-sample data on the ordered categorical, ordinal, scale. Such data can be analyzed by the "R index method". The R-index express Area Under the Curve of the empirical Receiver Operator Characteristics (ROC) curve, based on the accumulated hit rates and false alarm rates of the two sample frequencies:

```r
# Accumulated data:
FA <- c(0, 1, 2, 6, 8, 16, 36, 36)/36
H <- c(0, 24, 32, 36, 40, 40, 40, 40)/40

# The empirical ROC curve:
plot(FA, H)
lines(FA, H)
```

26
The dotted diagonal line represents the pattern that the ROC curve would take if the two sets of proportions were exactly the same. It turns out that the R-index is related to the non-parametric two-sample Wilcoxon rank sum test, also called Mann-Whitney’s test, such that one can find the R-index and get the corresponding p-value from basic routines providing this test:

```r
## Constructing the two-sample ordinal data from the frequencies
x <- rep(1:6, c(24, 8, 4, 4, 0, 0))
y <- rep(1:6, c(1, 1, 4, 2, 8, 20))

## Using inbuilt Wilcoxon function:
(U <- wilcox.test(y, x, correct=F))
```

The calculated R-index is

\[
\text{R-index} = \text{AUC} = 0.953
\]
## data: y and x
## W = 1372, p-value = 2.767e-12
## alternative hypothesis: true location shift is not equal to 0

## Finding the R-index from that, U/(n1*n2)
(Rindex <- U$statistic/(length(x)*length(y)))

## W
## 0.9527778

To get the relevant one-tailed p-value the reported p-value should be divided by 2.

Two-sample ordinal data like this could also be analysed as "truly ordinal" (a bit more "Thurstonian") using the ordinal-package, Christensen (2015). For this, the data should be on the "long" form and the response should be stored as a factor:

```r
library(ordinal)
# Making the data frame:
mydata <- cbind(c(x, y), c(rep("Same", 40), rep("Diff", 36))
mydata <- as.data.frame(mydata)
mydata$V1 <- ordered(mydata$V1)

# Analysing by ordinal regression model (cumulative link model, clm):
ordinal_res <- clm(V1~V2, data=mydata, link="probit")
summary(ordinal_res)
```

```
## formula: V1 ~ V2
## data:    mydata
##
##  link   threshold nobs logLik AIC    niter max.grad cond.H
##  probit flexible  76   -92.45 196.90 5(0)  9.46e-11 3.5e+01
##
## Coefficients:
##        Estimate Std. Error z value Pr(>|z|)  
## V2Same  -2.4912    0.3309  -7.528  5.16e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
##
## Threshold coefficients:
##        Estimate Std. Error z value
## 1|2   -2.20082   0.31231  -7.047
## 2|3   -1.65069   0.28261  -5.841
## 3|4    1.12916   0.25133  -4.493
## 4|5   -0.69041   0.22775  -3.031
## 5|6   -0.09869   0.20725  -0.476
##
# Finding the AUC from the (apparent) d-prime with CI:
AUC(-ordinal_res$coefficients[6], 0.3856)
```

AUC(-ordinal_res$coefficients[6], 0.3856)
## AUC: 0.9609293
## 0.95% CI: [0.8901205, 0.9891614]

In the clm-output is seen a "d-prime" of 2.49, which would be the d-prime, if the data had been "A-not-A with sureness data". And note how the AUC extracted from this number is very close to the non-parametric R-index found from these data. So for such data the R-index has a relevant Thurstonian interpretation. But this will not be the right d-prime for the same-diff protocol.

But actually the "same-diff-with-sureness" is just another name for unspecified "Degree-of-difference, DOD" -testing, which is covered in Ennis and Christensen (2015), and implemented as a part of the sensR package:

```r
# Analysing as a Degree-of-difference, DOD data:
dodresults <- dod(mydata$V1[mydata$V2=="Same"], mydata$V1[mydata$V2=="Diff"])
print(dodresults)
```

```r
## Results for the Thurstonian model for the Degree-of-Difference method
## Confidence level for 2-sided profile likelihood interval: 95%
##
## Estimates Std. Error Lower  Upper
## d.prime     4.077     0.4388  3.237   4.959
##
## Boundary coefficients:
## 1  2  3  4  5
## Estimate 1.2035 1.8085 2.4646 3.0721 3.9212
## Std. Error 0.1956 0.2458 0.3038 0.3671 0.4339
##
## Data:
## 1 2 3 4 5 6
## same-pairs 24 8 4 4 0 0
## diff-pairs  1 1 4 2 8 20
##
## Results of discrimination test:
## Likelihood Root statistic = 8.111544, p-value = 2.499e-16
## Alternative hypothesis: d-prime is greater than 0
##
## The AUC
AUC(dodresults$d.prime, dodresults$coefficients[2])
```

## AUC: 0.9980275
## 0.95% CI: [0.9885301, 0.999759]

Note how the d-prime is now estimated considerably larger and also the Sensitivity (AUC).

This appears to a more informative analysis of the same-diff-with-sureness data than merely finding the R-index and noting that it is more extreme than a critical value from a table.
The (simulation based) power of the DOD is also available, as well for difference test: (results not shown)

\[
\text{dodPwr}(d.primeA=1, d.prime0=0, ncat=6, sample.size=100, nsim=1000, \\
alpha=.05, \text{method.tau}="LR.max", \text{statistic}="likelihood")
\]

as for similarity testing:

\[
\text{dodPwr}(d.primeA=0, d.prime0=1, ncat=6, sample.size=100, nsim=1000, \\
alpha=.05, \text{method.tau}="LR.max", \text{statistic}="Pearson", \\
\text{alternative}="similarity")
\]

**Difference from Control (DFC) data**

In chapter 11 DFC protocols and data are discussed. The scales exemplified are all ordinal scales of some kind. The analysis of such data could be with or without formal Thurstonian model embedding. Without any attempts of using Thurstonian models such data could either be naively integer scored and then analyzed by normal based linear models using standard linear model functionality of R, or they could in a somewhat less restrictive approach be analyzed by ordinal regression models using e.g. the R-package ordinal. The former is probably more common than the latter. The former has the benefit that it becomes ANOVA analysis well known from e.g. QDA-type data. So people would not be unfamiliar with the need and tools to take "repeated measures" issues into account. That is, including effects, and potentially interaction effects, related to assessors as random effects leading to the relevant mixed model ANOVA for the design used. Such analysis makes some basic assumption about the scales, that it is (at least) on interval scale, i.e. that the naive integer scoring makes good sense. This assumption is not imposed when such data is analyzed by the ordinal package. And in fact the ordinal package offers also quite general mixed model versions of these models to take repeated measures structures into account in a similar way to the normal mixed linear models. And using much the same syntax as linear mixed models in R, see e.g. Christensen & Brockhoff (2013). But admittedly, such models and the corresponding analysis is more complex to comprehend, and likely for that reason not used as much as the linear ones. A particularly nice feature about the implementation of this in the ordinal package is the fact, that the hypothesis of equidistant thresholds between categories, necessary for the scale assumption needed to do linear model analysis is easily tested by a single option choice, example given below.

For a single product versus control protocol the unspecified DFC protocol is the same as the already discussed degree-of-difference, and we saw above how we could analyse such data in a formal Thurstonian framework using the dod-functionality of sensR, which in fact is based on a symmetric threshold ordinal regression model.

**Case study 1, Chapter 11**

Let us re-analyze the two data sets from Chapter 11, first the paired t-test, which then corresponds to a naive scoring approach where dependencies are taken into account properly by the differencing:
Reading the data from a file with same structure as table 1, Chapter 11

```r
DFCdata1 <- read.table("DFCdata1.txt", header = TRUE, sep = ";")
t.test(DFCdata1$Dif)
```

The same could have been achieved by a mixed model for the raw data as follows:

```r
library(lmerTest)

DFCdata1_long <- gather(DFCdata1[,1:3], value = score, key = Product, Test, Control)
DFCdata1_long$Assessor <- factor(DFCdata1_long$Assessor)
DFCdata1_long$Product <- factor(DFCdata1_long$Product)

lmer1 <- lmer(score ~ Product + (1|Assessor), data = DFCdata1_long)
anova(lmer1)
```

```
Analysis of Variance Table of type III with Satterthwaite approximation for degrees of freedom

<table>
<thead>
<tr>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>NumDF</th>
<th>DenDF</th>
<th>F.value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Product</td>
<td>1.4083</td>
<td>1.4083</td>
<td>1 59</td>
<td>3.3115</td>
<td>0.07387 .</td>
</tr>
</tbody>
</table>

---

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Here we used the packages lmerTest and lme4 for the linear mixed model analysis, Kuznetosva et al (2016), Kuznetosva et al (2017) and Bates et al (2015). Now we could make the similar analysis in an ordinal way:

```r
library(ordinal)
DFCdata1_long$scoreOrd <- ordered(DFCdata1_long$score)

clmm1 <- clmm(scoreOrd ~ Product + (1|Assessor), data = DFCdata1_long, link = "probit")

summary(clmm1)
```
We see that the p-value is 0.0948 - not far from the paired t-test p-value. The four category thresholds are not estimated exactly equidistant, but they are also estimated with uncertainties, so let's compare with an analysis with equidistant thresholds:

```r
clmm2 <- clmm(scoreOrd ~ Product + (1|Assessor), data = DFCdata1_long, 
               link = "probit", threshold = "equidistant")
summary(clmm2)
```

## Cumulative Link Mixed Model fitted with the Laplace approximation
##
## formula: scoreOrd ~ Product + (1 | Assessor)
## data:    DFCdata1_long
##
## link threshold nobs logLik  AIC    niter max.grad cond.H
## probit equidistant 120  -134.10 276.21 146(425) 9.27e-07 2.5e+01
##
## Random effects:
## Groups   Name        Variance Std.Dev.
## Assessor (Intercept) 0.4835   0.6954
## Number of groups: Assessor 60
##
## Coefficients:
##             Estimate Std. Error z value  Pr(>|z|)
## ProductTest  0.3545     0.2122   1.671 0.0948 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Threshold coefficients:
##     Estimate Std. Error z value
## 0|1  -1.7383     0.3116  -5.579
## 1|2   0.2422     0.1892   1.280
## 2|3   1.9119     0.2995   6.383
## 3|4   3.1425     0.5244   5.992
The average threshold spacing is estimated at 1.72. Comparing the log-likelihood values for the two models show that $-2\log \mathcal{Q} = 2(134.10 - 133.01) = 2.18$, which is not significant:

```r
1 - pchisq(2.18, 1)
```

## [1] 0.1398145

So the hypothesis of equidistant thresholds cannot be rejected. The p-value for product difference 0.0707 in this model is very close to the simple paired t-test p-value. We could actually also in a meaningful way analyze these data as DOD-data:

```r
Controldata <- DFCdata1_long$scoreOrd[DFCdata1_long$Product=="Control"]
Testdata <- DFCdata1_long$scoreOrd[DFCdata1_long$Product=="Test"]
dod(Controldata, Testdata)
```

## Results for the Thurstonian model for the Degree-of-Difference method
## Confidence level for 2-sided profile likelihood interval: 95%

<table>
<thead>
<tr>
<th>Estimates</th>
<th>Std. Error</th>
<th>Lower</th>
<th>Upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>d.prime</td>
<td>1.271</td>
<td>0.3086</td>
<td>0.5375</td>
</tr>
</tbody>
</table>

## Boundary coefficients:

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.14033</td>
<td>1.1682</td>
<td>2.9138</td>
<td>4.4453</td>
</tr>
</tbody>
</table>

## Data:

<table>
<thead>
<tr>
<th>same-pairs</th>
<th>diff-pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>3 30 27 0 0</td>
<td>5 24 21 9 1</td>
</tr>
</tbody>
</table>

## Results of discrimination test:

Likelihood Root statistic = 2.487306, p-value = 0.006436
Alternative hypothesis: d-prime is greater than 0

The d-prime estimate of 1.27 is likely the best estimate given the data, and the tools we have available. However, the p-value, very different from above, is not to be trusted as the DOD model
and analysis is assuming that the samples are independent, and hence does not take the repeated measures structure of this design into account. On the other hand, the large difference in p-values cannot be explained purely by the repeated measures issue, as the independent sample ordinal based comparison (not shown in detail) has a p-value even larger than the paired analysis. So it is quite noteworthy that a proper dod-analysis of some data (they could have been the results of an independently designed study) shows a clear significant effect of the relevant test, whereas other types of analysis don’t. For same-diff and DFC type protocols, it may be worthwhile to consider the formal Thurstonian model to make sure that the optimal analysis is performed.

**Case study 2, Chapter 11**

Then we analyze the 3-product design of case study 2 in Chapter 11: Reading the data, making the long version of the data and then, as above, doing the linear analysis followed by the two versions of the ordinal based analysis (without and with equidistant threshold assumption).

```r
# Reading the data from a file with same structure as table 5, Chapter 11
DFCdata2 <- read.table("DFCdata2.txt", header = TRUE, sep = ";")
DFCdata2_long <- gather(DFCdata2[,1:4], value = score, key = Product, C, TestA, TestB)
DFCdata2_long$Assessor <- factor(DFCdata2_long$Assessor)
DFCdata2_long$Product <- factor(DFCdata2_long$Product)

# Analysing by mixed model, random assessor effect
lmer2 <- lmer(score ~ Product + (1|Assessor), data = DFCdata2_long)
anova(lmer2)

# Analysis of Variance Table of type III with Satterthwaite approximation for degrees of freedom
# Sum Sq Mean Sq NumDF DenDF F.value Pr(>F)
# Product 36.796 18.398 2 70 12.013 3.271e-05 ***
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
difflsmeans(lmer2)

# Differences of LSMEANS:
# Estimate Standard Error DF t-value Lower CI
# Product C - TestA 1.2 0.2917 70.0 4.09 0.613
# Product C - TestB -0.1 0.2917 70.0 -0.29 -0.665
# Product TestA - TestB -1.3 0.2917 70.0 -4.38 -1.859
# Upper CI p-value
# Product C - TestA 1.776 1e-04 ***
# Product C - TestB 0.498 0.8
# Product TestA - TestB -0.696 <2e-16 ***
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```
## The ordinal analysis

```r
dfCdata2_long$scoreOrd <- factor(dfCdata2_long$score, ordered = TRUE)

clmm1 <- clmm(scoreOrd ~ Product + (1|Assessor), data = dfCdata2_long, link = "probit")
summary(clmm1)
```

## Cumulative Link Mixed Model fitted with the Laplace approximation

```r
## formula: scoreOrd ~ Product + (1 | Assessor)
## data: dfCdata2_long
##
## link threshold nobs logLik  AIC    niter     max.grad cond.H
## probit flexible 108  -163.05 344.11 556(1112) 1.66e-04 8.9e+01
##
## Random effects:
## Groups   Name        Variance Std.Dev.
## Assessor (Intercept) 0.2506   0.5006
## Number of groups: Assessor 36
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## ProductTestA  -0.9513     0.2663  -3.573 0.000353 ***
## ProductTestB   0.1193     0.2545   0.469 0.639070
##
## Threshold coefficients:
##        Estimate Std. Error z value
## -3|-2  -2.6634     0.3955  -6.734
## -2|-1  -1.7656     0.2781  -6.349
## -1|0   -0.9514     0.2310  -4.118
## 0|1    0.2199     0.2122   1.023
## 1|2    2.4882     0.4565   5.451
```

```r
clmm2 <- clmm(scoreOrd ~ Product + (1|Assessor), data = dfCdata2_long, link = "probit",
               threshold = "equidistant")
summary(clmm2)
```

## Cumulative Link Mixed Model fitted with the Laplace approximation

```r
## formula: scoreOrd ~ Product + (1 | Assessor)
## data: dfCdata2_long
##
## link threshold nobs logLik  AIC    niter     max.grad cond.H
## probit equidistant 108  -181.89 373.78 175(360) 1.21e-06 2.2e+02
```
Here we see an example of data where the equidistant threshold assumption is significantly wrong: Comparing the log-likelihood values for the two ordinal models shows that $-2\log Q = 2(181.89 - 163.05) = 37.68$, which is extremely significant:

```r
1-pchisq(37.68, 1)
```

## [1] 8.335481e-10

So actually the message here would be that the simple mixed linear model analysis is not completely valid. Note how the category widths differs from around 0.3, the center one, to 2.2. Also note that generally the product effects estimated in the ordinal based analysis corresponds to relative effect sizes in the similar linear mixed model, in line with Brockhoff et al (2016) (with an opposite sign). For instance, the difference between the evaluation of the blind control and the A test product has a mean of 1.2. The residual standard deviation from this model is:

```r
summary(lmer2)$sigma
```

## [1] 1.23753

so the relative effect size is $1.2/1.23773 = 0.97$ and the estimate of this number extracted from the clmm1 result is 0.95. So the advantage of the ordinal base analysis is that it does not make any assumptions about the scale and the results automatically come on a "d-prime-like" scale. It should be remembered, though, that only if there also is a meaningful Thurstonian model, these values are "real" dprimes. For blocked/correlated DFC data like this, the sensR package does not have the tools for this.

**Ranking data**

Rank data can be modelled and analyzed by all sorts of methods including Thurstonian and Bradley-Terry type models, and it is beyond the scope of this chapter to cover all these. In Rayner...
et al (2005) methods to decompose and visualize sensory rank data in continuation of classical non-parametric rank based data analysis are presented, and R scripts for this is available at the book website. Examples of relevant R-packages for more model based approaches are the pmr package, Lee & Yuo (2015) and the BradleyTerry2 package, Turner & Firth (2012).

Below we analyze the Chapter 12 case study 1 data by one of the classical rank based methods the Friedman's test. First the reader is imported, next a long version created and then the friedman.test function of base R:

```r
## Reading the data from a file with same structure as table 10, Chapter 12
Rankdata1 <- read.table("Rankdata1.txt", header = TRUE, sep = ";";")
Rankdata1_long <- gather(Rankdata1, value = score,
                        key = Product, Sample1, Sample2, Sample3, Sample4)
Rankdata1_long$Assessor <- factor(Rankdata1_long$Assessor)
Rankdata1_long$Product <- factor(Rankdata1_long$Product)

## Analysing by Friedman's test
(fried1 <- friedman.test(score ~ Product | Assessor, data = Rankdata1_long))

## Friedman rank sum test
## data:  score and Product and Assessor
## Friedman chi-squared = 36.9, df = 3, p-value = 4.831e-08

## Doing the post hoc comapriso
library(PMCMR)
with(Rankdata1_long, posthoc.friedman.conover.test(score, Product, Assessor,
p.adjust="holm"))

## Pairwise comparisons using Conover's test for a two-way balanced complete block design
## data:  score , Product and Assessor
##          Sample1 Sample2 Sample3
## Sample2 0.54    -       -
## Sample3 <2e-16  <2e-16  -
## Sample4 <2e-16  <2e-16  0.71
##
## P value adjustment method: holm
```

The overall significance and the grouping of the four products in two separate groups is clearly seen. Here we used the PMCMR package for the post hoc part, Pohlert (2014), where different multiple testing p-value adjustment methods are available.
ABX and dual standard data

Chapters 13 and 14 present the ABX and Dual Standard protocols. None of these are specifically supported by dedicated functions in sensR. The dual standard protocol is another example of what previously was termed a "basic single proportion of correct" protocol. The guessing probability is 1/2 like the duo-trio protocol, so the functions in sensR could be applied, as long as the thurstonian parts, dprime related issues, are not used. Or one could simply, as also exemplified previously use the exact binomial test function and/or generic binomial confidence interval functions. Thurstonian analysis of dual standard data is not provided. For the ABX protocol a thorough analysis is given which shows the strength of a computational framework like R - only very little is required to perform quite advanced analysis.

Dual standard case studies

In Chapter 14 the first case has n=32, which for a difference test then has the following critical value:

\[
\text{findcr}(32, p_0=1/2)
\]

## [1] 22

With the outcome \(x = 17\) we could use the \texttt{discrim} function as described:

\[
\text{discrim}(17, 32, \text{method} = \text{"duotrio"}, \text{conf.level} = 0.9)
\]

##
## Estimates for the duotrio discrimination protocol with 17 correct answers in 32 trials. One-sided p-value and 90 % two-sided confidence intervals are based on the 'exact' binomial test.
##
##         Estimate Std. Error Lower  Upper
## pc       0.5312    0.08822   0.5 0.6846
## pd       0.0625    0.17643   0.0 0.3691
## d-prime  0.5946    0.87291   0.0 1.6241
##
## Result of difference test:
## 'exact' binomial test:  p-value = 0.43
## Alternative hypothesis: d-prime is greater than 0

confirming the non-significant result. Remember to ignore the dprime-result row.

In the second case, with similarity in focus \(n = 95\) was used with \(x = 58\) correct responses. Without a pre-specified similarity definition we cannot perform formal similarity test, but we can do basic analysis similarly:

\[
\text{discrim}(58, 95, \text{method} = \text{"duotrio"}, \text{conf.level} = 0.9)
\]

##
## Estimates for the duotrio discrimination protocol with 58 correct
## answers in 95 trials. One-sided p-value and 90 % two-sided confidence intervals are based on the 'exact' binomial test.

<table>
<thead>
<tr>
<th></th>
<th>Estimate</th>
<th>Std. Error</th>
<th>Lower</th>
<th>Upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>pc</td>
<td>0.6105</td>
<td>0.05003</td>
<td>0.52121</td>
<td>0.6945</td>
</tr>
<tr>
<td>pd</td>
<td>0.2211</td>
<td>0.10006</td>
<td>0.04243</td>
<td>0.3891</td>
</tr>
<tr>
<td>d-prime</td>
<td>1.1817</td>
<td>0.31227</td>
<td>0.48679</td>
<td>1.6829</td>
</tr>
</tbody>
</table>

Result of difference test:
'exact' binomial test:  p-value = 0.0198
Alternative hypothesis: d-prime is greater than 0

On a $\alpha = 0.05$ the products would be claimed different, but if similarity is in focus that test and p-value is really irrelevant. The relevant number is the upper confidence limit for $p_D$ which is 0.389. So we have shown $p_d \leq 0.39$ similarity with a $\alpha = 0.05$ similarity test.

### ABX case studies

Two data examples are used in Chapter 14. The data is of the same structure as same-diff data: A hit rate HA and a false alarm rate FA, e.g. for initial illustrative example used:

```r
(HA <- 40/50)
## [1] 0.8
(FA <- 20/50)
## [1] 0.4
```

```
qnorm(HA) - qnorm(FA)
## [1] 1.094968
```

The analysis is then carried out by the use of Table A.5.3 in MacMillan and Creelman (2005), where the number $z(H) - z(F) = 1.095$ is used to identify that $p(c)_{unb} = 0.708$ and that the dprime is either 1.57 or 1.765 depending on the decision rule (and using linear interpolation).

The mathematics behind the table is that (equations (9.7), (9.11) and (9.12) in Chapter 9 of MacMillan and Creelman (2005))

$$p(c)_{unb} = \Phi((z(H) - z(F))/2)$$

and

$$p(c)_{ABX,10} = \Phi(d/\sqrt{2}) \cdot \Phi(d/2) + \Phi(-d/\sqrt{2}) \cdot \Phi(-d/2)$$

and

$$p(c)_{ABX,diff} = \Phi(d/\sqrt{2}) \cdot \Phi(d/\sqrt{6}) + \Phi(-d/\sqrt{2}) \cdot \Phi(-d/\sqrt{6})$$
These equations are easily implemented in R via the inbuilt standard normal distribution and quantile functions:

\[
\text{pc\_unb} \leftarrow \text{function}(HA, FA) \{ \text{pnorm}((\text{qnorm}(HA) - \text{qnorm}(FA))/2) \}
\]

\[
\text{pc\_ABX\_IO} \leftarrow \text{function}(d) \{ \text{pnorm}(d/\sqrt(2)) \ast \text{pnorm}(d/2) + \text{pnorm}(-d/\sqrt(2)) \ast \text{pnorm}(-d/2) \}
\]

\[
\text{pc\_ABX\_diff} \leftarrow \text{function}(d) \{ \text{pnorm}(d/\sqrt(2)) \ast \text{pnorm}(d/\sqrt(6)) + \text{pnorm}(-d/\sqrt(2)) \ast \text{pnorm}(-d/\sqrt(6)) \}
\]

# Check:
\[
\text{pc\_unb}(4/5, 2/5)
\]

## [1] 0.7079769

\[
\text{pc\_ABX\_IO}(1.57)
\]

## [1] 0.7080253

\[
\text{pc\_ABX\_diff}(1.765)
\]

## [1] 0.7083505

Now Table A.5.3 is actually the inverse of these functions: For a value of \( p(c)_{\text{unb}} \) it gives the d-value that corresponds to this. A little Google search tells you how to define the inverse of a function in R:

\[
\text{inverse} \leftarrow \text{function} \left( f, \text{lower} = -100, \text{upper} = 100 \right) \{ \text{function} \left( y \right) \text{uniroot}((\text{function} \left( x \right) f(x) - y), \text{lower} = \text{lower}, \text{upper} = \text{upper})[1] \}
\]

\[
\text{pc\_ABX\_IO\_inverse} \leftarrow \text{inverse}(\text{function} \left( x \right) \text{pc\_ABX\_IO}(x), 0.01, 100)
\]

\[
\text{pc\_ABX\_diff\_inverse} \leftarrow \text{inverse}(\text{function} \left( x \right) \text{pc\_ABX\_diff}(x), 0.01, 100)
\]

With these, we have now defined the functions that allows us to reconstruct the entire ABX part of the Table A.5.3, so we can use these instead of the table to find the dprime-values for a given data set, either the differencing version:

\[
\text{pc\_ABX\_diff\_inverse}(\text{pc\_unb}(HA, FA))
\]

## $root

## [1] 1.762769

\[
\text{pc\_ABX\_diff\_inverse}(0.708)
\]

## $root

## [1] 1.762907

or the IO version:

\[
\text{pc\_ABX\_IO\_inverse}(\text{pc\_unb}(HA, FA))
\]
As a final step let us take one step further and make a link to generic statistical modelling to make us able to find likelihood based confidence intervals for the estimated dprimes. So we even bypass the more classical Wald based intervals that people might think of as a first attempt. The key observation is that the core number $z(H) - z(F)$ actually exactly is the difference between the two mean parameters of an independent two-samples probit-regression model:

```r
## Doing the two-sample probit regression
y <- c(rep(0, 40), rep(1, 10), rep(0, 20), rep(1, 30))
trt <- c(rep("same", 50), rep("diff", 50))
myprobit <- glm((1-y) ~ trt, family=binomial(link="probit"))
## model summary
summary(myprobit)
```

Note that the `trtsame` estimate is exactly the same and claimed number.

And now we can use inbuilt likelihood confidence interval routines of R to get these for this difference:

```r
confint(myprobit)[2,]
```
And finally we can transform these confidence intervals though the non-linear functions already defined to get the wanted CIs for the actual dprimes, first for the differencing version:

```r
pcABX_diff_inverse(pnorm(0.572/2)) ## Lower CI limit
## $root
## [1] 1.1944
pcABX_diff_inverse(pnorm(1.632/2)) ## Upper CI limit
## $root
## [1] 2.303498
```

and then for the IO version:

```r
pcABX_IO_inverse(pnorm(0.572/2)) ## Lower CI limit
## $root
## [1] 1.072174
pcABX_IO_inverse(pnorm(1.632/2)) ## Upper CI limit
## $root
## [1] 2.029674
```

Apart from giving a really nice analysis of ABX data including and going way beyond what is presented in MacMillan & Creelman (2005), it also serves as the promised example of how to be able to handle single proportion of correct protocols not covered in sensR: Simply implement the psychometric function as exemplified here, and similarly the inverse, and everything would be available.

**References**


43


Analysis of the Data Using the R package sensR