Statistical methodology for sensory discrimination tests and its implementation in sensR

Rune Haubo Bojesen Christensen

May 16, 2011

Abstract

The statistical methodology of sensory discrimination analysis is described. This forms the basis of the implementation in the sensR package for R. Implementation choices will be motivated when appropriate and examples of analysis of sensory discrimination experiments will be given throughout using the sensR package. This document currently covers parameterizations, hypothesis tests, confidence intervals, and power and sample size calculations for the four common discrimination protocols: 2-AFC, 3-AFC, triangle and duo-trio; analysis of replicated experiments with the four common discrimination protocols using the beta-binomial and chance-corrected beta-binomial models.

Contents

1 Introduction

The aim of this document is 1) to describe the statistical methodology for sensory discrimination testing and analysis, and 2) to describe how such analyses can be performed in R using package sensR (Christensen and Brockhoff, 2010) co-developed by the author of this document.

This document is divided into sections that cover topics with similar statistical methodology. Implementation choices in the sensR package will be described in connection with the statistical methodology whenever appropriate. Small examples illustrating the use of functions in the sensR package will appear throughout.

This is not a hands-on practical tutorial to analysis of sensory discrimination experiments with the sensR package, neither is it a user friendly introduction to discrimination and similarity testing in sensory discrimination protocols. The former document does not really exist[1] (yet), and for the latter document, we refer the reader to (Næs et al., 2010, chapter 7). We will assume throughout that the reader has basic statistical training and is familiar with sensory discrimination testing to the level of (Næs et al., 2010, chapter 7).

2 Classification of sensory discrimination protocols

The most common and simplest discrimination protocols comprise the 2-AFC, 3-AFC, triangle, duo-trio, A-not A and same-different protocols. The first four protocols are designed such that the response follows a binomial distribution in the simplest experimental setting. On the other hand responses from A-not A and same-different protocols are distributed according to a compound or product binomial distribution in the simplest experimental setting. An extension of the A-not A method known as the A-not A with sureness is a classical SDT method which leads to multinomially distributed responses. Similarly the same-different method extends to the degree-of-difference protocol also resulting in multinomially distributed responses. An experiment using one of the first four simple protocols can be summarized with the proportion of correct responses or similarly the probability of discrimination or d-prime. The Thurstonian models for the remaining protocols involve one or more additional parameters each with their particular cognitive interpretation.

The 2-AFC and 3-AFC protocols are so-called directional protocols since they require that the nature of the difference (e.g. sweetness) is provided as part of the assessor instructions. On the other hand the triangle and duo-trio protocols are not directional since these protocols are used to test un-specified differences. From a Thurstonian point of view, the sensory dimension or the perceptual dimension is fixed in the 2-AFC and 3-AFC methods. The cognitive decision strategy is consequently assumed different in these two classes of protocols. When the perceptual dimension is fixed, the assessors may use the more effective skimming strategy, while assessors are forced to use the inferior comparison of distances strategy when using the un-directional protocols.

The A-not A and same-different protocols are methods with so-called response bias. Response bias refers to the concept that one type of response is preferred over another despite the sensory distance remains unchanged. For instance some assessors may prefer the "A" response over the "not A" response.

[1] this is on the to-do list of the author of this document, so there is hope it will appear in the future.

The four simple protocols are without response bias since no response can be consistently preferred over another without affecting the discriminative effect. The decision criterion is said to be fixed or stabilized.

3 Four common sensory discrimination protocols: 2-AFC, 3-AFC, triangle and duo-trio

The four common sensory discrimination protocols are often used in practical applications in the food industry as well as in other areas. They are also of considerable interest in the scientific literature about sensory discrimination.

The protocols have one important thing in common from a statistical perspective: their statistical models can all be described as variants of the binomial distribution. That is, the answer from any one of these protocols is either correct or incorrect and the sampling distribution of answers is therefore a binomial distribution.

For the duo-trio and 2-AFC protocols the *guessing probability*, p_g is 1/2. This means that if there is no discriminative difference between the products, then the probability of a correct answers, p_c is one half. Similarly for the triangle and 3-AFC protocols the guessing probability is 1/3. The four common discrimination protocols are said to be free of *response bias* in contrast to the A-not A and same-different protocols. Response bias will be described in section 4.

If we assume for a moment that the population of assessors (be that judges in an expert panel or consumers) is comprised of ignorants who are always guessing and discriminators who always discriminate correctly and provide the appropriate answer (though this will not always be the *correct* answer). One way to express the sensory distance of the objects (or discriminative ability of the assessors — we will treat these viewpoints synonymously throughout) is the *proportion of discriminators*, p_d in the population of interest. It is almost always an unreasonable assumption that some assessors are either always discriminating or always guessing (Ennis, 1993), but we may still talk about the *probability of discrimination*. This probability may refer to particular individuals or to a population; in this section we will adopt a population perspective and in section 5 we will include an individual perspective.

The relation between the probability of a correct answer and the probability of discrimination is

$$p_c = p_g + p_d(1 - p_g), \tag{1}$$

where the guessing probability, p_g is 1/2 for the duo-trio and 2-AFC protocols and 1/3 for the triangle and 3-AFC protocols. The reverse relation is

$$p_d = (p_c - p_g)/(1 - p_g). \tag{2}$$

Another way to summarize the sensory distance is through a measure known as d' (pronounced "d-prime") from signal detection theory (SDT, Green and Swets, 1966; Macmillan and Creelman, 2005), or equivalently *the Thurstonian delta*, δ (Thurstone, 1927a,b,c). These two concepts are identical and will be used synonymously throughout, and they are actually based on the same underlying psychophysical model for the cognitive process. Whereas p_c is a measure and parameter completely free of reference to any particular discrimination protocol, p_d depends on the discrimination protocol through the guessing probability, but

d' depends on the discrimination protocol through the so-called *psychometric function*, for the discrimination protocol. The psychometric function maps from d' to the probability of a correct answer:

$$p_c = f_{ps}(d'). \tag{3}$$

For the m-AFC method, where m denotes the number of "forced choices", the psychometric function is given by

$$f_{m\text{-AFC}}(d') = \int_{-\infty}^{\infty} \phi(z - d')\Phi(z)^{m-1} \, \mathrm{d}z, \tag{4}$$

where ϕ is the standard normal probability density function and Φ is the standard normal cumulative distribution function. The psychometric functions for the four common discrimination protocols are given by

$$f_{3\text{-AFC}}(d') = \int_{-\infty}^{\infty} \phi(z - d')\Phi(z)^2 \, \mathrm{d}z \tag{5}$$

$$f_{2\text{-AFC}}(d') = \int_{-\infty}^{\infty} \phi(z - d')\Phi(z) \, \mathrm{d}z = \Phi(d'/\sqrt{2}) \tag{6}$$

$$f_{\text{tri}}(d') = 2\int_{0}^{\infty} \left\{ \Phi\left[-z\sqrt{3} + d'\sqrt{2/3}\right] + \Phi\left[-z\sqrt{3} - d'\sqrt{2/3}\right] \right\} \phi(z) \, \mathrm{d}z \tag{7}$$

$$f_{\text{duo-trio}}(d') = 1 - \Phi(d'/\sqrt{2}) - \Phi(d'/\sqrt{6}) + 2\Phi(d'/\sqrt{2})\Phi(d'/\sqrt{6}). \tag{8}$$

Further information can be found in Ennis (1993) and Brockhoff and Christensen (2010).

The relations between the three scales at which a sensory difference is described are illustrated in Fig. 1. In the relation between p_d and d' the alternative forced choice protocols behave similarly, while the duo-trio and triangle protocols behave similarly. The gradient of the psychometric functions (cf. eq. (17)) goes to zero when d' goes to zero for the duo-trio and triangle protocols.

The result of a simple discrimination protocol is a number of correct answers, $X = x$ out of n trials. Under the assumption of independent observations, the sampling distribution of X is the binomial:

$$X \sim \text{Binom}(p_c, n), \tag{9}$$

so

$$P(X = x) = \binom{n}{x} p_c^x (1 - p_c)^{n-x}. \tag{10}$$

There is a subtle but important distinction between the *proportion* of a correct answer and the *probability* of a correct answer. The proportion of correct answers is x/n which can be any number between 0 and 1. The probability of a correct answer, which we denote by p_c, is a parameter and represents a true underlying value. As such p_c cannot be lower than the guessing probability for the discrimination protocol that was used and cannot exceed 1. The usual estimator of a binomial probability is just the sample proportion, x/n, but this is not the case here, and it is exactly this feature that makes discrimination testing interesting statistically.

The maximum likelihood (ML) estimator[2] of p_c is given by:

$$\hat{p}_c = \begin{cases} x/n & \text{if} \quad x/n \geq p_g \\ p_g & \text{if} \quad x/n < p_g \end{cases} \tag{11}$$

[2]Following standard statistical practice we use the hat-notation to denote an estimator or an estimate

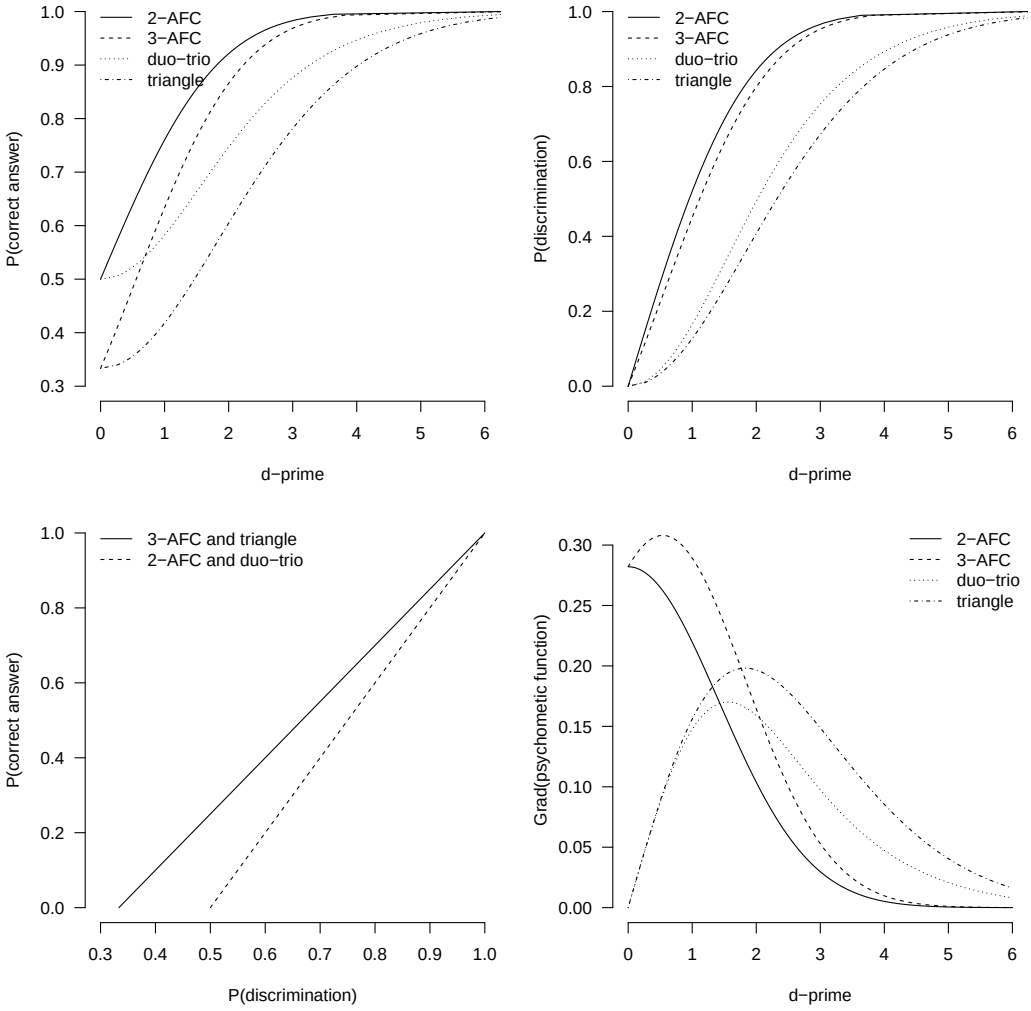

Figure 1: The connection between d', p_c and p_d for the four common sensory discrimination protocols. The so-called psychometric functions; P_c as a function of d', are shown in the upper left figure.

The ML estimator of p_d is given by application of eq. (2), and the ML estimator of d', by inversion of eq. (3), given by

$$\hat{d'} = f_{ps}^{-1}(\hat{p}_c), \tag{12}$$

where $f_{ps}^{-1}(\cdot)$ (which should not be confused with $f_{ps}(\cdot)^{-1} = 1/f_{ps}(\cdot)$) is the inverse psychometric function.

The allowed ranges (parameter space) for these three parameters are given by

$$d' \in [0, \infty[, \quad p_d \in [0, 1], \quad p_c \in [p_g, 1]. \tag{13}$$

Negative d' values are sometimes mentioned in the literature, but negative d' values are not possible in the discrimination protocols that we consider here. They are possible in preference tests and theoretically possible in Thurstonian models based on other assumptions, see section XXX for more background information on this topic.

3.0.1 Implementation in sensR

In package sensR there is a function `rescale` that maps between the three scales; p_c, p_d and d'. A value on one of these scales is given as argument and values on all three scales are given in the results. The results respect the allowed ranges of the parameters in eq. (13), so if the supplied p_c is less than p_g, then $p_c = p_g$ is returned with p_d and d' at the appropriate levels:

```
> rescale(pc = 0.25, method = "triangle")

Estimates for the triangle protocol:
         pc pd d.prime
1 0.3333333  0       0
```

Function `rescale` use a number of auxiliary functions for its computations; these are also directly available to the package user:

- `pc2pd`: maps from the p_c-scale to the p_d-scale.
- `pd2pc`: maps from the p_d-scale to the p_c-scale.
- `psyfun`: implements the psychmetric functions $p_c = f_{ps}(d')$ for the four common discrimination protocols, cf. eq. (3).
- `psyinv`: implements the inverse psychometric functions, $d' = f_{ps}^{-1}(p_c)$ for the four common discrimination protocols, cf. eq. (12).
- `psyderiv`: implements the derivative of the psychometric functions, $f'_{ps}(d')$ for the four common discrimination protocols.

3.1 Inference in simple discrimination protocols

To obtain inference in simple discrimination protocols, we need measures such as standard errors, confidence intervals (CIs) and p-values from significance tests.

3.1.1 Standard errors

The standard error of p_c is given by:

$$\mathrm{se}(p_c) = \sqrt{p_c(1-p_c)/n}. \tag{14}$$

The standard error of p_d and d' can be found by application of the Delta method (see for example Pawitan, 2001):

$$\mathrm{se}\{f(x)\} = \frac{\partial f(x)}{\partial x}\mathrm{se}(x) \tag{15}$$

The standard error of p_d is therefore

$$\mathrm{se}(p_d) = \frac{1}{1-p_g}\mathrm{se}(p_c) \tag{16}$$

since $\partial p_d/\partial p_c = 1/(1-p_g)$, cf. eq. (2). The standard error of d' can similarly be found as

$$\mathrm{se}(d') = \frac{\partial f_{ps}^{-1}(p_c)}{\partial p_c}\mathrm{se}(p_c) = \frac{1}{f'_{ps}(d')}\mathrm{se}(p_c) \tag{17}$$

where $f'_{ps}(d')$ is the derivative of the psychometric function with respect to d'; expressions are given by Brockhoff and Christensen (2010).

Standard errors are only defined and only meaningful as measures of uncertainty when the parameter estimate is at the interior of the parameter space, i.e. when the parameter estimate is not at the boundary of its allowed range, cf. eq. (13).

Even when the parameter estimate is close, in some sense, to the boundary of its parameter space, the standard error is not a meaningful measure of uncertainty, because the uncertainty is in fact asymmetric. This means that symmetric confidence intervals based on the standard error will also be misleading and other techniques should be applied.

3.1.2 The likelihood function

The (log-)likelihood function can be used to obtain likelihood ratio or likelihood root statistics for hypothesis tests, and it can be used to construct confidence intervals with good properties.

The log-likelihood function for a model based on the binomial distribution is given by

$$\ell(p_c; x, n) = C + x\log p_c + (n-x)\log(1-p_c), \tag{18}$$

where $C = \log\binom{n}{x}$ is a constant with respect to p_c. The log-likelihood function for p_d or d' is given by combining eq. (18) with (2) or (12).

In general, standard errors can be found as the square root of the diagonal elements of the variance-covariance matrix of the parameters. The variance-covariance matrix can be found as the inverse of the negative Hessian matrix (the matrix of second order derivatives) of the log-likelihood function evaluated at the ML estimates. Here there is only one parameter (either one of p_c, p_d or d'), so the matrices are merely scalars.

It can be shown that the same standard errors as those derived in eq. (14), (16) and (17) can be derived by differentiating (18) twice and using the chain rule to obtain the standard errors of p_d and d'.

3.1.3 Confidence intervals

There are several general approaches to get CIs for parameters. One general way that applies (with varying success) to almost all parameters with a standard error is the traditional Wald interval:

$$CI: \ \hat{\mu} \pm z_{1-\alpha/2}\text{se}(\hat{\mu}), \tag{19}$$

where $z_{1-\alpha/2} = \Phi^{-1}(1-\alpha/2)$ is the upper $\alpha/2$ quantile of the standard normal distribution. This CI is based on the Wald statistic[3]:

$$w(\mu_0) = (\hat{\mu} - \mu_0)/\text{se}(\hat{\mu}). \tag{20}$$

The CI may also be expressed more generally for a statistic $t(\mu_0)$ that follows a standard normal distribution under the null hypothesis as:

$$CI: \ \{\mu; |t(\mu)| < z_{1-\alpha/2}\}. \tag{21}$$

Using w as t in (21) gives the interval (19).

Another general approach is to use the likelihood root statistic (inverted likelihood ratio test) which applies to all likelihood based models and almost always impressively successful. The likelihood root statistic is given by:

$$r(\mu_0) = \text{sign}(\hat{\mu} - \mu_0)\sqrt{2\left\{\ell(\hat{\mu}; x) - \ell(\mu_0; x)\right\}} \tag{22}$$

Both the Wald and likelihood root statistics asymptotically follow standard normal distributions under the null hypothesis. Even though their asymptotic behavior is in fact identical, their finite sample properties may be quite different and often favor the likelihood root statistic since it removes nonlinear parameterization effects.

A disadvantage of Wald intervals is that they are not invariant to nonlinear transformations of the parameter. This means that a Wald CI for p_c and a Wald CI for d' provides different kinds of evidence about the parameters and could, for instance, lead to inclusion of p_g in the CI on the p_c scale, but exclusion of $d' = 0$ on the d' scale. More generally the Wald CI for p_c cannot be found by transforming the Wald CI limits for d' through the psychometric function. The CIs based on the likelihood root statistic is on the other hand invariant to nonlinear transformations of the parameter. This means the likelihood CI for d' can be found by either computing the likelihood CI for d' directly or by transforming the limits of the likelihood CI for p_c through the inverse psychometric function — they give the same answer. The evidence provided by the likelihood CI is therefore invariant to the choice of scale.

Another approach to generate CIs consistent across parameter scales would be to compute an appropriate CI for, say, p_c and then transform the CI limits through the appropriate functions to obtain CIs for p_d and d'. For likelihood CIs this does not make any difference, of course. If an appropriate CI can be computed on any one scale, this would provide appropriate CIs on the other scales as well. There exists a wide range of CIs for the binomial probability parameter (refs), for instance the score interval and the so-called exact interval in addition to the Wald and likelihood intervals.

The 'exact' binomial interval is given by inversion of the 'exact' binomial test and known as the Clopper-Pearson interval (Clopper and Pearson, 1934). The lower and upper limits are

[3] actually the original definition used $\text{se}(\mu_0)$ in the denominator.

defined as the values of p_c that solve:

$$LL: \ P(X \geq x) = \alpha/2, \quad UL: \ P(X \leq x) = \alpha/2, \tag{23}$$

where $X \sim \text{binom}(p_c, n)$. Rather than solving these equations numerically, the limits can be found directly as quantiles of the beta distribution, $\text{Beta}(a, b)$: the lower limit is the $\alpha/2$ quantile of $\text{Beta}(x, n-x+1)$ and the upper limit is the $1-\alpha/2$ quantile of $\text{Beta}(x+1, n-x)$.

Another commonly applied statistic is based on the normal approximation of the binomial distribution. Asymptotically $(X - np_c)/\sqrt{np_c(1-p_c)}$ behaves like a standard normal random variable, so we may use

$$w^*(p_{c0}) = \frac{x - np_{c0}}{\sqrt{np_{c0}(1-p_{c0})}}, \tag{24}$$

as test statistic. This statistic is in fact identical to the Wald statistic (20) if $\text{se}(\mu_0)$ is used in the denominator instead of $\text{se}(\hat{\mu})$.

The statistic w^* is related to the Pearson χ^2 statistic

$$X^2(p_{c0}) = \frac{(x - np_{c0})^2}{np_{c0}} + \frac{(n - x - n(1-p_{c0}))^2}{n(1-p_{c0})} \tag{25}$$

since $w*$ is the signed square root of X^2. Similarly the likelihood root statistic, $r(p_{c0})$ is related to the likelihood ratio statistic

$$G^2(p_{c0}) = x \log \frac{x}{np_{c0}} + (n-x) \log \frac{n-x}{n(1-p_{c0})} \tag{26}$$

since $r(p_{c0})$ is the signed square root of $G^2(p_{c0})$.

3.1.4 Sensory difference tests

A sensory difference test is a test of

$$H_0: \begin{array}{l} p_c \leq p_{c0} \\ p_d \leq p_{d0} \\ d' \leq d'_0 \end{array} \quad \text{versus} \quad H_A: \begin{array}{l} p_c > p_{c0} \\ p_d > p_{d0} \\ d' > d'_0 \end{array}, \tag{27}$$

where the traditional tests of no-difference is given by choosing $p_{c0} = p_g$, $p_{d0} = 0$ and $d'_0 = 0$ making the null hypothesis an equality rather than an inequality.

The p-value of a difference test is the probability of observing a number of successes that is as large or larger than that observed given the null hypothesis that the probability of a correct answer is p_{c0}. The p-value based on the 'exact' binomial test is therefore:

$$p\text{-value} = P(X \geq x) = 1 - \sum_{i=0}^{x-1} \binom{n}{i} p_{c0}^i (1-p_{c0})^{n-i}, \tag{28}$$

where $X \sim \text{binom}(p_{c0}, n)$

The p-value for a difference based on a statistic, $t(\mu_0)$ that follows a standard normal distribution under the null hypothesis is given by:

$$p\text{-value} = P\{Z \geq t(\mu_0)\} = 1 - \Phi\{t(\mu_0)\}, \tag{29}$$

where Z is a standard normal random variable and Φ is the standard normal cumulative distribution function.

3.1.5 Sensory similarity tests

A sensory similarity test is a test of

$$H_0: \begin{array}{l} p_c \geq p_{c0} \\ p_d \geq p_{d0} \\ d' \geq d'_0 \end{array} \quad \text{versus} \quad H_A: \begin{array}{l} p_c < p_{c0} \\ p_d < p_{d0} \\ d' < d'_0 \end{array}, \tag{30}$$

where subject matter considerations and possibly power computations will guide the choice of p_{c0}, p_{d0} or d'_0. Observe that d'_0 has to be positive for the test to make sense.

The p-value of a similarity test is the probability of observing a number of successes that is as large or less than that observed given the null hypothesis that the probability of a correct answer is p_{c0}. The p-value based on the 'exact' binomial test is therefore:

$$p\text{-value} = P(X \leq x) = \sum_{i=0}^{x} \binom{n}{i} p_{c0}^i (1 - p_{c0})^{n-i}, \tag{31}$$

where $X \sim \text{binom}(p_{c0}, n)$

The p-value for a difference based on a statistic, $t(\mu_0)$ that follows a standard normal distribution under the null hypothesis is given by:

$$p\text{-value} = P\{Z \leq t(\mu_0)\} = \Phi\{t(\mu_0)\}, \tag{32}$$

3.1.6 Confidence intervals and hypothesis tests

Confidence intervals are often described by their relation to hypothesis tests such that a two-sided hypothesis test should be accompanied by a two-sided confidence interval and one-sided hypothesis tests should be accompanied by one-sided confidence intervals. This will make the $1 - \alpha$ level confidence interval the region in which an observation would not lead to rejection of the null hypothesis. A confidence interval should, however, provide more than a rejection region; it should provide an interval in which we can have confidence that the true parameter lies. This corresponds to the interval which provides most support for the parameter. As such confidence intervals should be two-sided even if the appropriate test may be one-sided (Boyles, 2008). We will use two-sided confidence intervals throughout and use these in conjunction with p-values from one-sided difference and similarity tests. This is also implemented in sensR.

Confidence intervals may, however, be one-sided in a slightly different respect since it may happen, for instance, that the lower confidence limit is at the guessing probability, p_g. If the observed proportion of correct answers is less than p_g, the lower confidence limit will also be higher than the observed proportion.

Confidence intervals may be degenerate in the sense that both limits can be zero; this is obviously not very informative. This may happen if, for instance, the observed proportion is below p_g and α is large enough. For small enough α, the upper confidence limit for d' will, however, exceed zero.

Confidence intervals can be used for difference and similarity testing as argued by MacRae (1995) and Carr (1995) when it is enough to know if the alternative hypothesis is rejected or not. Comparing the formulas for the 'exact' Clopper-Pearson confidence limits (23) with

the formulas for p-values in difference and similarity tests also based on the exact test, it is clear that there is a close connection.

If p_{c0} under H_0 is below the lower confidence limit in a $1-\alpha$ level interval, then the p-value of a difference test will be below $\alpha/2$, i.e. the test will be significant at the $\alpha/2$-level. Thus, if p_{c0} is below the lower confidence limit in a 90% interval, then the difference test is significant at the 5% level. Similarly, if p_{c0} is above the upper confidence limit in a 90% interval, then the similarity test is significant at the 5% level.

In difference testing the binomial test is not too liberal even if there is variability in p_d under the alternative hypothesis, because there can be no variability under the null hypothesis that $p_d = 0$. In similarity testing, however, $p_d > 0$ under H_0 and the standard binomial test could possibly be liberal. Also not that p_d under H_A will be less than p_d under H_0, and if there is variation in p_d in the distribution, this variation could be larger under H_0 than under H_A. Also, the power and sample size computations in the following assume that zero variability in p_d. Possibly the power will be lower and sample sizes higher if there really is variation in p_d in the population.

The similarity tests discussed so far are targeted toward equivalence in the population on average. There is no consideration of equivalence on the level of individual discrimination.

A general problem with discrimination testing outlined so far is the assumption that all assessors have the same probability of discrimination. This is hardly ever a priory plausible. The so-called guessing model (refs) assumes that there are two kinds of assessors; non-discriminators that always guess and true discriminators that always perceive the difference and discriminate correctly. This assumption is also hardly ever a priory plausible. More plausible is perhaps that the probability of discrimination has some distribution across the population of assessors as is assumed in the chance-corrected beta-binomial distribution.

3.1.7 Implementation in sensR

The function `rescale` that was described in section 3.0.1 has an additional optional argument `std.err` which allows one to get the standard error of, say, p_d and d' if the standard error of p_c is supplied. This is done through application of eq. (16) and (17) and by using the user visible function `psyderiv`, which implements the derivative of the psychometric functions, $f'_{ps}(d')$ for the four common discrimination protocols:

```
> rescale(pd = 0.2, std.err = 0.12, method = "triangle")

Estimates for the triangle protocol:
         pc  pd  d.prime
1 0.4666667 0.2 1.287139

Standard errors:
    pc   pd   d.prime
1 0.08 0.12 0.4424581
```

The `discrim` function is the primary function for inference in the duo-trio, triangle, 2-AFC and 3-AFC protocols. Given the number of correct answers, x and the number of trials, n, `discrim` will provide estimates, standard errors and confidence intervals on the scale of p_c, p_d and d'. It will also report the p-value from a difference or similarity test of the users choice. p-values will be one-sided while confidence limits will be two-sided, cf. section 3.1.6. Confidence intervals are computed on the scale of p_c and then transformed to the p_d and d'

scales as discussed in section 3.1.3. The user can choose between several statistics including the 'exact' binomial, likelihood, Wald and score statistics. The score option leads to the so-called Wilson or score interval, while the p-value is based on the w^* statistic, cf. eq. (24).

Estimates and confidence intervals reported by `discrim` respect the allowed range of the parameters, cf. eq. (13) and standard errors are not reported if the parameter estimates are on the boundary of their parameter space (allowed range).

Strictly speaking the Wald statistic (20) is not defined when $x/n \leq p_g$, since the standard error of $\hat{p}_c$ is not defined. However, it makes sense to use $\sqrt{\frac{x}{n}\left(1-\frac{x}{n}\right)\frac{1}{n}}$ as standard error in this case. This is adopted in `discrim`.

Similarity testing does not make sense if $p_{c0} = 0$ under the null hypothesis, cf. eq. (30), so a positive p_{d0} has to be chosen for similarity testing.

Example: Suppose we have performed a 3-AFC discrimination test and observed 10 correct answers in 15 trials. We want estimates of the p_c, p_d and d', their standard error and 95% confidence intervals. We are also interested in the difference test of no difference and decide to use the likelihood root statistic for confidence intervals and tests. Using the `discrim` function in R we obtain:

```
> discrim(10, 15, method = "threeAFC", statistic = "likelihood")

Estimates for the threeAFC discrimination protocol with 10 correct
answers in 15 trials. p-value and 95 percent confidence intervals
are based on the likelihood root statistic.

          Estimate Std. Error     Lower     Upper
pc       0.6666667  0.1217161 0.4154537 0.8652194
pd       0.5000000  0.1825742 0.1231806 0.7978291
d-prime 1.1159025  0.4359153 0.2802776 1.9966789

Result of difference test:
likelihood root statistic = 2.632769 p-value = 0.0042346
Alternative hypothesis: d-prime is greater than 0
```

If instead we had observed 4 correct answers in 15 trials and were interested in the similarity test with $p_{d0} = 1/5$ under the null hypothesis, we get using the 'exact' binomial criterion for confidence intervals and tests:

```
> discrim(4, 15, method = "threeAFC", test = "similarity",
     pd0 = 0.2, statistic = "exact")

Estimates for the threeAFC discrimination protocol with 4 correct
answers in 15 trials. p-value and 95 percent confidence intervals
are based on the 'exact' binomial test.

          Estimate Std. Error     Lower     Upper
pc       0.3333333         NA 0.3333333 0.5510032
pd       0.0000000         NA 0.0000000 0.3265049
d-prime 0.0000000         NA 0.0000000 0.7226962

Result of similarity test:
'exact' binomial test: p-value = 0.096376
```

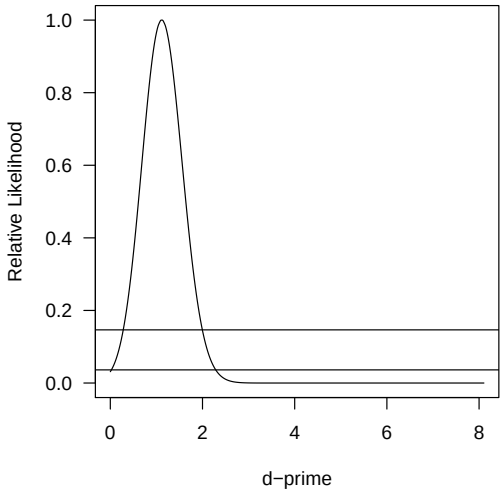

Figure 2: Relative likelihood function for a 3-AFC experiment with 10 correct answers in 15 trials. The maximum likelihood estimate is at $d' = 1.12$ and the two horizontal lines determine the 95% and 99% likelihood based confidence intervals.

```
Alternative hypothesis: d-prime is less than 0.4482204
```

A few auxiliary methods for `discrim` objects are available. `confint` returns the confidence intervals computed in the `discrim` object, `profile` extracts the (profile) likelihood function and `plot.profile` plots the likelihood function.

Example To illustrate the auxiliary methods consider the 3-AFC example above where 10 correct answer were observed in 15 trials.

```
> fm1 <- discrim(10, 15, method = "threeAFC", statistic = "exact")
> confint(fm1)

            Lower     Upper
pc      0.3838037 0.8817589
pd      0.0757056 0.8226383
d-prime 0.1744201 2.1015693
attr(,"method")
[1] "threeAFC"
attr(,"conf.level")
[1] 0.95
attr(,"statistic")
[1] "exact"

> plot(profile(fm1))
```

The resulting graph is shown in Fig. 2. Observe that the likelihood (profile) function may be extracted from a `discrim` object that is not fitted with `statistic = "likelihood"`. Further information about the use and interpretation of (profile) likelihood curves in sensory experiments is given in (Brockhoff and Christensen, 2010; Christensen and Brockhoff, 2009).

3.2 Sample size and power calculations for simple discrimination protocols

The power of a test is the probability of getting a significant result for a particular test given data, significance level and a particular difference. In other words, it is the probability of observing a difference that is actually there. Power and sample size calculations require that a model under the null hypothesis and a model under the alternative hypothesis are decided upon. The null model is often implied by the null hypothesis and is used to calculate the critical value. The alternative model has to lie under the alternative hypothesis and involves a subject matter choice. Power is then calculated for that particular choice of alternative model.

In the following we will consider calculation of power and sample size based directly on the binomial distribution. Later we will consider calculations based on a normal approximation and based on simulations.

3.2.1 The critical value

Formally the critical value, x_c of a one-sided binomial test where the alternative hypothesis is *difference*, or equivalently *greater*, is the smallest integer number that satisfies

$$P(X \geq x_c) \leq \alpha \quad \text{where} \quad X \sim \text{binom}(p_{c0}, n) \tag{33}$$

and p_{c0} is the probability of a correct answer under the null hypothesis. Similarly the critical value, x_c of a one-sided binomial test where the alternative hypothesis is *similarity*, or equivalently *less*, is the largest integer number that satisfies

$$P(X \leq x_c) \leq \alpha \quad \text{where} \quad X \sim \text{binom}(p_{c0}, n) \tag{34}$$

If the sample size is small for the desired α, there may not be a possible critical value that satisfies (33) or (34). In a difference test it may not be enough to observe $x = n$ correct answers, i.e. all correct answers for the test to be significant at the required α. Similarly, it may not be enough to observe no correct answers ($x = 0$) for the similarity test to be significant at the required α.

A simple way to compute x_c is to use a small while loop (shown here for a difference test):

$i = 0$
while $P(X \geq i) > \alpha$ **do**
  $i = i + 1$
end while
return $i + 1$

However, if x_c is a large number, many iterations of the loop would be required, so instead in the `findcr` function in package sensR eq. (33) and (34) are solved numerically for x_c. One complication with this method is that $P(X \geq x_c)$ is discontinuous in x_c and that requires special attention.

Example: Consider the situation that $X = 15$ correct answers are observed out of $n = 20$ trials in a duo-trio test. The exact binomial p-value of a no-difference test is $P(X \geq 15) =$

$1 - P(X \leq 15 - 1) = 0.021$, where $X \sim \text{binom}(0.5, 20)$ so this is significant. If on the other hand we had observed $X = 14$, then the p-value would have been $P(X \geq 14) = 0.058$, which is not significant. We say that $x_c = 15$ is the *critical value* for this particular test on the $\alpha = 5\%$ significance level because $x_c = 15$ is the smallest number of correct answers that renders the test significant.

In R we can find the p-values with

```
> 1 - pbinom(q = 15 - 1, size = 20, prob = 0.5)

[1] 0.02069473

> 1 - pbinom(q = 14 - 1, size = 20, prob = 0.5)

[1] 0.05765915
```

The while loop looks like

```
> i <- 0
> while (1 - pbinom(q = i, size = 20, prob = 0.5) > 0.05) {
     i <- i + 1
 }
> i + 1

[1] 15
```

while we could also use the `findcr` function in package sensR:

```
> findcr(sample.size = 20, alpha = 0.05, p0 = 0.5)

[1] 15
```

3.2.2 The power of difference tests

The power of a difference test is

$$\text{power} = P(X \geq x_c) \quad \text{where} \quad X \sim \text{binom}(p_{cA}, n), \tag{35}$$

where p_{cA} is the probability of a correct answer under the alternative hypothesis and x_c is the critical value of the test, which depends on the probability of a correct answer under the null hypothesis and the significance level, α.

Power increases with the difference between p_{c0} and p_{cA}, the sample size and α. Power can be computed directly once the critical value, p_{cA} and n are known, so the only computational challenge is in the computation of the critical value.

Example: The power of the test considered in the previous example is the probability of getting this p-value or one that is smaller. This depends on the actual sensory difference of the objects/the proportion of discriminators. If half the population are discriminators or equivalently if each assessor has a 50% of correctly discriminating a set of samples, then $p_c = 1/2 + 1/2p_d = 3/4$. The power is the probability of observing 15 or more correct answers:

$$\text{power} = P(X \geq 15) = 1 - P(X \leq 15 - 1) = 0.617 \quad \text{where} \quad X \sim \text{binom}(3/4, 20) \tag{36}$$

This can be obtained in R with

```
> 1 - pbinom(q = 15 - 1, size = 20, prob = 3/4)

[1] 0.6171727
```

or directly using the `discrimPwr` function from sensR:

```
> discrimPwr(pdA = 0.5, sample.size = 20, alpha = 0.05, pGuess = 1/2)

[1] 0.6171727
```

Observe that `discrimPwr` requires that the effect size under the alternative hypothesis is given in terms of p_d rather than p_c or d'. If the effect size under the alternative hypothesis is formulated in terms of d', then `rescale` can be used to convert from d'_A to p_{dA}, but it would be easier to use `d.primePwr`, which accepts d'_A directly and internally calls `discrimPwr`.

If the significance test of interest is not that of no-difference, but that of a small difference versus a relevant difference, the computation of the critical value is slightly different. The power calculation remain essentially the same.

If the limit between irrelevant and relevant differences is at $p_d = 0.1$, so $p_c = 1/2 + 1/2 \cdot 0.1 = 0.55$, then $P(X \geq 16|p_{c0} = 0.55, n = 20) = 1 - P(X \leq 16 - 1) = 0.019$ while $P(X \geq 15|p_{c0} = 0.55, n = 20) = 1 - P(X \leq 15 - 1) = 0.055$. The critical value is therefore 16 and the power of the test is

$$\text{power} = P(X \geq 16) = 0.415 \quad \text{where} \quad X \sim \text{binom}(p_{cA} = 3/4, n = 20) \tag{37}$$

In R we could get the power of this test with

```
> discrimPwr(pdA = 0.5, pd0 = 0.1, sample.size = 20, alpha = 0.05,
     pGuess = 1/2)

[1] 0.4148415
```

Note the `pd0` argument which should match the value of p_d under the null hypothesis.

3.2.3 The power of similarity tests

The power of a similarity test is

$$\text{power} = P(X \leq x_c) \quad \text{where} \quad X \sim \text{binom}(p_{cA}, n), \tag{38}$$

and p_{cA} is the probability of a correct answer under the alternative hypothesis and x_c is the critical value of the test, which depends on the probability of a correct answer under the null hypothesis and the significance level, α.

Example: Assume that we want to calculate the power of a similarity test using the duo-trio protocol with $n = 100$, and that we want to show that the probability of discrimination is less than 1/3, while we believe that there is actually no difference between the objects, so the true probability of discrimination is zero. The null hypothesis is therefore $H_0 : \ p_c \geq 1/2 + 1/2 \cdot 1/3 = 2/3$ and the alternative hypothesis is $H_A : \ p_c < 2/3$. The critical value of this test is $x_c = 58$ since $p = P(X \leq 58|p_c = 2/3, n = 100) = 0.042 \leq 0.05$ while $P(X \leq 59) = 0.064 > 0.05$. The power of this test is therefore

$$\text{power} = P(X \leq 58|p_c = 0.5, n = 100) = 0.956 \tag{39}$$

We would compute this power in R with

```
> discrimPwr(pdA = 0, pd0 = 1/3, sample.size = 100, alpha = 0.05,
     pGuess = 1/2, test = "similarity")

[1] 0.955687
```

If in fact there is a small difference between the objects, so that there is a positive probability of discrimination, say $p_d = 1/5$, then the power is (the critical value remains the same):

$$\text{power} = P(X \leq 58 | p_c = 0.5(1 + 1/5), n = 100) = 0.377 \tag{40}$$

We would compute this power in R with

```
> discrimPwr(pdA = 1/5, pd0 = 1/3, sample.size = 100, alpha = 0.05,
     pGuess = 1/2, test = "similarity")

[1] 0.3774673
```

Observe how the power of the similarity test is quite good if there is absolutely no observable difference between the objects, while if there is in fact a small probability that a difference can be observed, the power is horrible and the sample size far from sufficient.

3.2.4 Power calculation based on simulations

In more complicated models it is not possible to determine an explicit expression for the power of a test and calculation of power based on simulations can be an attractive approach. Sometimes it may also just be easier to let the computer do the job by running simulations rather than to get bugged down in derivations of explicit expressions for power even though they may in fact be possible to derive.

Recall that power is the probability of getting a significant result when there is in fact a difference, thus in the long run it is the proportion of significant results to the total number of tests:

$$\text{power} = \frac{\text{no. } p\text{-values} < \alpha}{\text{no. tests}} \tag{41}$$

We can let the computer generate random data from the model under the alternative hypothesis and then perform the significance test. We can even do that many many times and record the p-values allowing us to calculate the power via eq. (41). In the following we will do exactly that for a binomial test for which we know the right answer.

Consider the no-difference example above in section 3.2.2 where $n = 20$ and the power of a no-difference test was 0.617 when $p_d = 1/2$, so $p_c = 3/4$. We will estimate the power via simulation by generating 10,000 (pseudo) random draws, X_i, $i = 1, \ldots, 10,000$ from $X_i \sim \text{binom}(p_c = 3/4, n = 20)$. For each of these draws we calculate the p-value as $p_i = P(X \geq x_i | p_c = 1/2, n = 20)$. Among these p-values 6184 were below 0.05, so the power estimated by simulation is 0.6184. Observe that this is close to, but not exactly the power that we obtained analytically (0.617). If we did the power calculation over again, we would most likely get a slightly different power estimate although probably also close to 0.617 because we would obtain a slightly different set of random draws. This illustrates that although power calculation via simulation is simple, the result varies a little from one run to another.

Fortunately we can estimate the uncertainty in the estimated power from standard binomial principles. The standard error of the estimated power is $\text{se}(\hat{\text{power}}) = \sqrt{\text{power}(1 - \text{power})/n_{\text{sim}}} =$

$\sqrt{0.6814(1-0.6814)/10,000} = 0.0049$ and an approximate Wald 95% CI for the estimated power is $[0.609; 0.628]$, which covers the true value (0.617) as one would expect.

3.2.5 Power calculation based on the normal approximation

An often used approximation for power and sample size calculations is the normal approximation; the idea is to use a statistic that asymptotically follows a standard normal distribution. For a binomial parameter power and sample size calculation may be based on the Wald statistic (20) as for example described by Lachin (1981) and advocated by Bi (2006) in a sensometric context. We are not aware of any numerical assessments of the accuracy of the normal approximation for power and sample size calculations, but we may expect that for small n or p (under the null or alternative) close to one, the approximation may be rather inaccurate. Since power and sample size determinations are readily available for the exact binomial test, we see no reason to use approximate statistics with doubtful properties for these purposes.

Consider the following hypotheses for a binomial parameter:

$$H_0 : \ p = p_0 \quad H_A : \ p > p_0, \tag{42}$$

then under the null hypothesis approximately

$$\frac{\hat{p} - p_0}{\sigma_0} \sim N(0,1) \tag{43}$$

and under the alternative hypothesis approximately

$$\frac{\hat{p} - p_A}{\sigma_A} \sim N(0,1), \tag{44}$$

where p_A is the probability under the alternative hypothesis, $\sigma_0 = \sqrt{p_0(1-p_0)/n}$, $\sigma_A = \sqrt{p_A(1-p_A)/n}$ and $\hat{p} = X/n$ is the estimator of a binomial parameter. The critical point above which the null hypothesis is rejected is then

$$\frac{\hat{p} - p_0}{\sigma_0} > \Phi^{-1}(1-\alpha) = z_{1-\alpha} \tag{45}$$

i.e. when

$$\hat{p} > z_{1-\alpha}\sigma_0 + p_0. \tag{46}$$

Under H_A the null hypothesis is rejected if

$$\frac{\hat{p} - p_A}{\sigma_A} > \frac{z_{1-\alpha}\sigma_0 + p_0 - p_A}{\sigma_A} \tag{47}$$

and the power is

$$\text{power} = P\left(Z > \frac{z_{1-\alpha}\sigma_0 + p_0 - p_A}{\sigma_A}\right) = 1 - \Phi\left(\frac{z_{1-\alpha}\sigma_0 + p_0 - p_A}{\sigma_A}\right) \tag{48}$$

Equivalent considerations for the equivalence hypotheses lead to

$$\text{power} = P\left(Z < \frac{z_{\alpha}\sigma_0 + p_0 - p_A}{\sigma_A}\right) = \Phi\left(\frac{z_{\alpha}\sigma_0 + p_0 - p_A}{\sigma_A}\right) \tag{49}$$

Isolating n in eq. (48) leads to the following expression for the sample size of difference tests:

$$\text{sample size} = \left(\frac{z_\beta \sqrt{p_A(1-p_A)} - z_{1-\alpha}\sqrt{p_0(1-p_0)}}{p_0 - p_A} \right)^2, \tag{50}$$

where $z_\beta = \Phi^{-1}(1 - \text{power})$. Equivalently for similarity tests:

$$\text{sample size} = \left(\frac{z_{1-\beta} \sqrt{p_A(1-p_A)} - z_{\alpha}\sqrt{p_0(1-p_0)}}{p_0 - p_A} \right)^2, \tag{51}$$

where $z_{1-\beta} = \Phi^{-1}(\text{power})$. The sample sizes given by (50) and (51) should be rounded up to the nearest integer.

3.2.6 Sample size determination

In principle sample size determination is simple; find the sample size such that the power is sufficiently high for a particular test at some significance level given some true difference. Computationally, however, it can be a challenge.

Formally, the required sample size, n^* for a sensory difference test is the smallest integer number, n^* that satisfies

$$P(X \geq x_c) \geq \text{target-power} \quad \text{where} \quad X \sim \text{binom}(p_c, n^*), \tag{52}$$

and $P(X \geq x_c)$ is the *actual power* of the test. Power for a difference test only increases with increasing sample size if the true difference, p_d is larger than the null difference, p_{d0}, so it is a requirement that the value of p_d specified as the true difference is actually covered by the alternative hypothesis.

Similarly, the required sample size, n^* for a similarity test is the smallest integer number, n^* that satisfies

$$P(X \leq x_c) \geq \text{target-power} \quad \text{where} \quad X \sim \text{binom}(p_c, n^*), \tag{53}$$

and $P(X \leq x_c)$ is the *actual power* of the test. Power only increases with increasing sample size if the true difference, p_d is less than the null difference, p_{d0}, so as for difference tests, the value specified as the true difference has to be covered by the alternative hypothesis.

The sample size depends on the particulars of the null and alternative hypotheses as well as the significance level of the test, i.e. α and the desired minimum power; the *target-power*.

So much for the formal definitions: practical sample size determination is in fact not as simple as the definitions may lead one to believe. Consider a situation in which we want to know which sample size to choose in a difference test using the triangle protocol where the null hypothesis is no difference, target power is 0.80, and we believe the actual difference is $d' = 0.9$ under the alternative hypothesis. Standard sample size calculations under the definition (52) tells us that 297 tests are enough; this leads to an actual power of 0.802. However, had we decided to use, say, 300 tests—for convenience and just to be on the safe side, the power of the test is only 0.774; much less than the power with 297 tests and below our target power. This is truly worrying; how many samples do we need to be sure that all larger sample sizes also lead to a power above 0.80? It is natural to expect power to increase

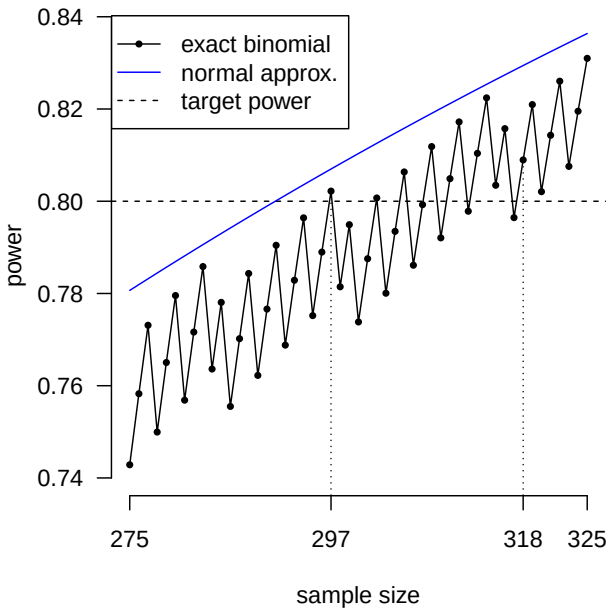

Figure 3: The relation between sample size and power for a difference test with the triangle protocol. The null hypothesis is that of no difference and $d' = 0.9$ is assumed under the alternative model. A power of 0.80 is desired.

with every increase in sample size (a monotonic increase in power with sample size), but this is not the case as is illustrated in Fig. 3.

Power generally increases with the sample size, but it does so in a zig-zag way due to the discreteness of the binomial distribution. As is seen in Fig. 3, the smallest sample size for which power is higher than 0.80 is 297 (actual power = 0.802). The next sample size that gives a power above 0.80 is 302, but the actual power is now less than 0.801. We would need 305 samples (actual power = 0.806) to obtain a power that is higher than the power with 297, and no less than 318 samples (actual power = 0.802) if no larger sample size should lead to a power less than 0.80.

Even though an increase in sample size may lead to a decrease in power, it will instead lead to a decrease in the actual α-level. This occurs because the critical value of the test is at times piece-wise constant as a function of sample size, cf. Fig. 4.

The sample size for the exact binomial test may be computed with much the same while loop that could also be used to find the critical value (cf. section 3.2.1):

$i = 1$
while actual power(i) < target power **do**
  $i = i + 1$
end while
return i

where actual power depends on the hypothesis, cf. (52) and (53). The problem with this approach is that if the required sample size is large, it may take some time to get there;

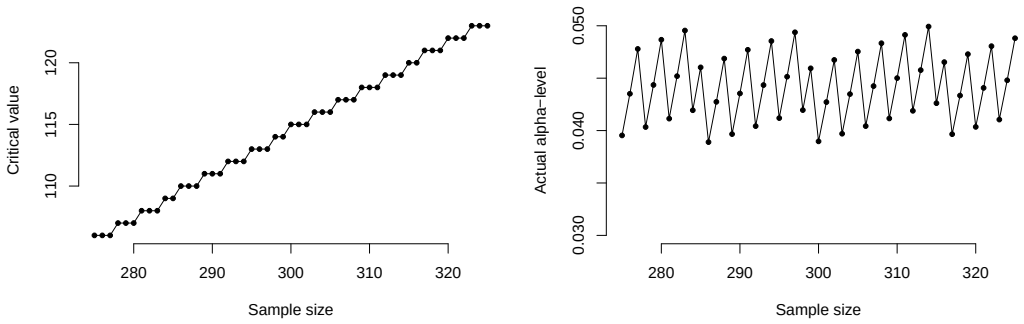

Figure 4: Left: critical value of the test. Right: actual α-level of the test.

recall that at every evaluation of the actual power, the critical value has to be determined. Due to the non-monotonicity of the relationship between power and sample size (cf. Fig. 3), it is not possible to simply solve for the required sample size numerically.

An improvement over the simple while loop is suggested by the normal approximation to the required sample size shown in Fig. 3 in blue. This approximation seems to estimate the sample size too low, and to do so consistently. For the example considered here, the normal approximation estimates that 291 samples is enough to obtain a power of 0.80 (actual power $= 0.8001$). The while loop could simply be started at $i = 291$ rather than at $i = 1$. A problem with this approach is that the normal approximation is not always strictly liberal. In the function `discrimSS` in package sensR a compromise is used, where the while loop is started at one if the sample size estimated by the normal approximation is less than 50. Otherwise the while loop is started at 90% of the normal approximation estimate and sometimes even lower if necessary. If the normal approximation estimate is larger than 10,000, the function will inform of that and not attempt to estimate the sample size due to the expected large computation time. In addition to the sample size for the 'exact' binomial test, it is also possible to ask for the sample size based on the normal approximation.

Example: Consider the example above illustrated in Fig. 3; we wanted to know the sample size for a difference test where the null hypothesis is that of no difference using the triangle protocol. We want a power of 0.80, take $\alpha = 0.05$ and we assume the actual difference is $d' = 0.9$ under the alternative hypothesis. Using package sensR we may get the sample size for the exact binomial test with

```
> (pd <- coef(rescale(d.prime = 0.9, method = "triangle"))$pd)

[1] 0.1044969

> discrimSS(pdA = pd, pd0 = 0, target.power = 0.8, alpha = 0.05,
     pGuess = 1/3, test = "difference", statistic = "exact")

[1] 297
```

We could also obtain the normal approximation with

```
> discrimSS(pdA = pd, pd0 = 0, target.power = 0.8, alpha = 0.05,
     pGuess = 1/3, test = "difference", statistic = "normal")

[1] 291
```

3.3 Related literature

4 A-not A and same-different protocols

5 Replicated simple discrimination protocols

5.1 The beta-binomial model

introduce data and model here

The beta-binomial model is derived from the marginal distribution of the binomial response, X_i, when the probabilities, π_i are allowed to have a beta-distribution on the interval $(0;1)$.

The likelihood function for the beta-binomial model is

$$\ell(\alpha,\beta;x,n) = -N\log Beta(\alpha,\beta) + \sum_{i=1}^{N}\log Beta(\alpha+x_i,\beta-x_i+n_i) \tag{54}$$

where $i = 1,\ldots,N$ is the number of independent binomial observations each with x_i successes out of n_i trials, $Beta$ is the beta function with parameters α and β. The beta-binomial model can be parameterized in terms of the mean binomial parameter, μ and the degree of over-dispersion, γ both living in the interval $(0;1)$. The relation to the (α,β) parameters is:

$$\mu = \alpha/(\alpha+\beta) \quad \gamma = 1/(\alpha+\beta+1) \tag{55}$$

γ can also be interpreted as a correlation...

In a sensory experiment μ can be interpreted as the probability of a correct answer, p_c averaged over subjects. The corresponding values of p_d and d' can be found by eq. (2) and (12).

The standard error of the μ and γ can be found from the variance-covariance matrix of the parameters, which in turn can be found from the Hessian of the likelihood function. The standard errors of the population p_d and population d' can also be found from eq. (16) and (17).

5.2 The chance-corrected beta-binomial model

In the chance-corrected beta-binomial model the probability of a correct answer is only allowed in the interval $(p_g;1)$. This means that the probability of a correct answer for any one individual is not allowed to be less than the guessing probability. This is sensible since it is impossible for the underlying probability of a correct answer for any individual to be less than the guessing probability for any of the simple sensory discrimination protocols.

The likelihood function for the chance-corrected beta-binomial model is

$$\begin{aligned}\ell(\alpha,\beta;x,n) = & -N\log Beta(\alpha,\beta) + \sum_{j=1}^{N}\log \\ & \left\{\sum_{i=1}^{x_j}\binom{x_j}{i}(1-p_g)^{n_j-x_j+i}p_g^{x_j-i}Beta(\alpha+i,n_j-x_j+\beta)\right\}\end{aligned} \tag{56}$$

The parameters, μ and γ still live in the interval $(0;1)$, but μ now has to be interpreted on the p_d scale rather than the p_c scale. Transformation to the other scales and computation of standard errors follow the approach in section 3.

5.3 A mixture of discriminators and non-discriminators

It does not have to be assumed that assessors are either discriminators or non-discriminators. These assessor types can be regarded as the extreme endpoints in-between which the population of assessors distribute.

5.4 Difference testing in replicated experiments

5.4.1 Model based likelihood ratio tests

A likelihood ratio test of over-dispersion can be calculated by comparing the likelihood of the beta-binomial model with the likelihood of the standard binomial model. Suppose 20 panelists each conduct 10 triangle tests, with x_i number of correct responses out of $n_i = 10$ trials for all i, where $i = 1, \ldots, N$, $N = 20$.

The likelihood ratio test statistic for the test of over-dispersion is

$$LR_{over-disp} = 2\{\ell_{beta-bin}(\hat{\mu}, \hat{\gamma}; x; n) - \ell_{binom}(\hat{\mu}; x, n)\}, \tag{57}$$

where $\ell_{beta-bin}(\hat{\mu}, \hat{\gamma}; x; n)$ is the log-likelihood of the (chance-corrected) beta-binomial model given by by (54) or (56) evaluated at the ML estimates and $\ell_{binom}(\hat{\mu}; x, n)$ is the value of

$$\ell_{binom}(\mu; x, n) = \sum_{i=1}^{N} \left\{ \log \binom{n_i}{x_i} x_i \log \mu + (n_i - x_i) \log(1 - \mu) \right\} \tag{58}$$

evaluated at the ML estimate $\hat{\mu} = \sum_i x_i / \sum_i n_i$. Here the alternative hypothesis is that $X_i \sim \text{binom}(\hat{\mu}, n_i)$; that observations from different individuals are independent and binomially distributed with the same probability of a correct answer; $p_c = \mu$.

The likelihood ratio statistic asymptotically follows a χ^2_1-distribution, so the p-value is given by:

$$p\text{-value} = P(\chi^2_1 \geq LR_{over-disp}) \tag{59}$$

We assume a 1 degree of freedom reference distribution, but this may not be appropriate since this is in fact a test of $\gamma = 0$ versus $\gamma > 0$ and hence a test on the boundary of the parameter space for γ. The appropriate df may be closer to one half, but this has not been empirically justified. One df corresponds to the two-sided test and the test is in fact one-sided, which motivates halving the degrees of freedom. Choosing df = 1 is on the safe side since this leads to p-value which may be a little too large.

The test of "any difference", is a test of $H_0: \mu_i = p_g$ for all i, i.e. for all individuals, versus the general alternative that for some individuals at least the probability of a correct answer is different from the guessing probability, p_g and *possibly* varies among individuals. The likelihood under the null hypothesis, $\ell_{binom}(p_g; x, n)$ is given by eq. (58) where $\mu = p_g$, and the likelihood ratio statistic is given by

$$LR_{any-diff} = 2\{\ell_{beta-bin}(\hat{\mu}, \hat{\gamma}; x; n) - \ell_{binom}(p_g; x, n)\}. \tag{60}$$

We assume a χ^2_2 reference distribution for this statistic, so the p-value is

$$p\text{-value} = P(\chi^2_2 \geq LR_{any-diff}) \tag{61}$$

while this is still partly a test at the boundary of the parameter space, so the stated p-value may be a little too large.

Yet another test of "any difference" is possible. First realize that if there really is no sensory difference between the objects/products, then the probability of a correct answer will for all assessors be p_g and there is no room for over-dispersion. We may therefore test $H_0 : \mu = p_g$ versus $H_1 : \mu > p_g$ where μ is the average probability of a correct answer and $\hat{\mu} = \sum_i x_i / \sum_i n_i$. The likelihood ratio statistic for this test is

$$LR_{mean-diff} = 2\{\ell_{binom}(\hat{\mu}; x, n) - \ell_{binom}(p_g; x, n)\} \tag{62}$$

which asymptotically follows a χ^2_1-distribution, so the p-value is given by:

$$p\text{-value} = P(\chi^2_1 \geq LR_{mean-diff}). \tag{63}$$

These three models are nested, so the LR-statistics are additive in the following way

$$LR_{any-diff} = LR_{mean-diff} + LR_{over-disp} \tag{64}$$

and the degrees of freedom add up similarly. This can be used to provide some insight in to the power of these tests under various scenarios. If there is only little or perhaps no over-dispersion at all, then the mean difference test will provide the most powerful test. If most of the structure in the data is due to over-dispersion, then the test directly targeted on over-dispersion will be the most powerful. It is, however, not possible to observe over-dispersion without a difference in mean, so the joint test of mean and dispersion; the test of any difference, may be almost as powerful. We expect that when the beta-binomial model is appropriate, i.e. when there is appreciably over-dispersion, the test of any difference will be Superior to the other two tests or at least at least as good as them.

References

Bi, J. (2006). *Sensory Discrimination Tests and Measurements—Statistical Principles, Procedures and Tables*. Blackwell Publishing.

Boyles, R. A. (2008). The role of likelihood in interval estimation. *The American Statistician 62*(1), pp. 22–26.

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

Carr, B. T. (1995). Confidence intervals in the analysis of sensory discrimination tests—the integration of similarity and difference testing. In *Proceedings of the 4th AgroStat, Dijon, 7.-8. December*, pp. 23–31.

Christensen, R. H. B. and P. B. Brockhoff (2009). Estimation and Inference in the Same Different Test. *Food Quality and Preference 20*, pp. 514–524.

Christensen, R. H. B. and P. B. Brockhoff (2010). sensR: An R-package for Thurstonian modelling of discrete sensory data. R package version 1.2.0 http://www.cran.r-project.org/package=sensR/.

Clopper, C. J. and E. S. Pearson (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika 26*, pp. 404–413.

Ennis, D. M. (1993). The power of sensory discrimination methods. *Journal of Sensory Studies 8*(353-370).

Green, D. M. and J. A. Swets (1966). *Signal Detection Theory and Psychophysics*. John Wiley & Sons.

Lachin, J. M. (1981). Introduction to sample size determination and power analysis for clinical trials. *Controlled Clinical Trials 2*, pp. 93–113.

Macmillan, N. A. and C. D. Creelman (2005). *Detection Theory, A User's Guide* (Second ed.). Lawrence Elbaum Associates, Publishers.

MacRae, A. W. (1995). Confidence intervals for the triangle test can give assurance that products are similar. *Food Quality and Preference 6*(61-67).

Næs, T., P. B. Brockhoff, and O. Tomic (2010). *Statistics for sensory and consumer science*. John Wiley & sons Ltd.

Pawitan, Y. (2001). *In All Likelihood—Statistical Modelling and Inference Using Likelihood*. Oxford University Press.

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

Thurstone, L. L. (1927b). Psychophysical analysis. *American journal of Psychology 38*, pp. 368–389.

Thurstone, L. L. (1927c). Three psychological laws. *Psychological Review 34*, pp. 424–432.