RANK TESTS USING THE PIM-FRAMEWORK IN R

CONVENIENT CODE FOR NON-EXPERTS

Tim Bal

Student ID: 20040659

Promotor: Prof. dr. Jan De Neve

A dissertation submitted to Ghent University in partial fulfilment of the requirements for the degree of Master of Science in Statistical Data Analysis.

Academic year: 2018 - 2019

GHENT UNIVERSITY

Gent, September 2, 2019

The promotor,	The author,
Prof. dr. Jan De Neve	Tim Bal

FOREWORD

General information

This dissertation is the result of a collaboration between the student, Tim Bal, and his promotor, Prof. dr. Jan de Neve. The starting point of this dissertation is mainly De Neve and Thas (2015) and De Neve (2013). Since the main goal of this thesis is to write convenient R-code, as a possible extension to the package pim (Meys et al., 2017), the theoretical part is mostly a summary of De Neve and Thas (2015) and De Neve (2013) where the most important parts are selected, in order to understand and follow this dissertation. The R-code is completely written by the student, where the syntax and structure is derived from existing R-code, in order to be implemented in the package. The data used in this dissertation is taken from Keuleers et al. (2010) and is freely avaible on http://crr.ugent.be/programs-data/lexicon-projects.

The complete code and transformed/cleaned datasets can be found on the github-page of the student: https://github.com/trbal/pim_rank.

Acknowledgements

The writing of a dissertation is always a long-time effort, and often I compare it to dancing procession of Echternach. Whenever I took three steps forward, it was followed with two steps backward. But in the end, I achieved my goal.

During this process(ion), I have had support without which I could not have achieved my goal. I started the Master of Statistical Data Analysis five years ago, after taking the course of Bayesian Statistics as an elective course when I was still in my Master of Theoretical and Experimental Psychology. It was in that year, my love for statistics was really born, thanks to Yves Rosseel and Dries Benoit. During my last year of my Master of Theoterical and Experimental Psychology, I decided to enroll for the Master of Statistical Data Analysis. It was a hard year, but in the end it was satisfying. When I started to work as a teaching assistant at the Department of Data Analysis, under the supervision of Thierry Marchant, I met the promoter of this dissertation, Jan De Neve.

When I approached Jan to ask if he wanted to be my promoter, he did not hesitate and the first seed for this dissertation was planted. Although I coped with a severe depression last year, the support and empathy I got from both Thierry and Jan gave me the much-needed energy to start this final year of my Master of Statistical Data Analysis.

But as I experienced already before, writing a dissertation is not limited to professional life or the academics. With ups and downs (very deep downs), I have probably annoyed my friends and family. But without their support, I would not have come this far. During the most dark days, my parents always provided a shoulder to cry on. Whenever I needed to talk away frustrations, I knew I could call to some of my closest friends, Rodney and André. During the last year, I also found support with Chris and Evy, a couple I met in the whisky society, so it would only be fair to thank them too for the much appreciated conversations and complementary drams of liquid gold. Finally, I would also like to thank my brother. Although we are not the best friends thinkable and we often have the obligatory brotherly feuds, I have learned he appreciates me much more than I could have hoped and this provides me hope.

I would like to dedicate this dissertation to the people mentioned, but in particular my parents. We have gone through some dark times, some moments I feared my mother would not be able to live the day I would finish my dissertation. Thanks for breaking a lots of eggs, just to let me find my way. I hope I have made you proud!

Tim,
Gent,
September 2, 2019

CONTENTS

ABSTRACT

In previous research De Neve and Thas (2015) have shown that rank tests can be modeled using the PIM-framework. The advantage the PIM-framework has, compared to classic rank tests, is that effect sizes, standard errors, and confidence intervals can be estimated. Next to the traditional Rank Tests, they provide a method extend rank tests to complicated designs. The goal of this dissertation is to make the most used Rank Tests available in one package in R, and more important, user-friendly and flexible, in order to provide effect sizes, standard errors, and confidence intervals. We have written a convenient code, so no complex code need to be constructed by users who have no in-depth knowledge of PIM. Next to this, we also provide extensions on the traditional rank tests, such as handling unbalanced data and missing data as an extension of the Mack-Skillings test, and adding a continuous covariate. We also show that these methods can easily be implemented in research. Keuleers et al. (2010) investigated the difference between categorization and recognition of words versus non-words using a lexical decision task. Using rank tests within the PIM-framework, we replicated the conclusions of Keuleers et al. (2010).

CHAPTER 1

A BRIEF HISTORY OF PIM AND RANK TESTS

1.1 Introduction

The main goal of this dissertation is to create convenient code in R for conducting rank tests using the PIM-framework, with the possibility to extend the rank test with the implementation to control for a covariate. The first advantage of using the PIM-framework is that effect sizes, standard errors, and confidence intervals can be achieved, which is not possible in the classic approach of the rank tests. The second advantage is that a continuous covariate can be taken into account when conducting a rank test.

But before we go into depth of the two advantages, and how they are implemented in R, a short introduction of PIM on the one hand and rank tests on the other hand are needed, and will be given in this chapter.

1.2 Probabilistic Index Model, or PIM in short

Until now, we only used the abbreviation “PIM”. But what is “PIM”? “PIM” stands for “Probabilistic Index Model”. Throughout this dissertation, we will use the abbreviation “PIM”.

PIMs were first introduced by Thas et al. (2012), as an alternative to the linear regression model, the restricted moment model, the cumulative logit model, and the quantile regression model. The probabilistic index is the central summary measure associated with a PIM.

1.2.1 Probabilistic Index

Different names have been used for the Probabilistic Index: $P(Y < Y')$ (Enis and Geisser, 1971; Halperin et al., 1987), probabilsitic index (Acion et al., 2006), exceedance probability (Senn, 1997), stochastic improvement (Lehmann, 1998), common language effect size (McGraw and Wong, 1992), probability of superiority (Grissom, 1994). All have the same goal, namely calculating the probability that a random observation of one group exceeds a random observation of another group.

Suppose we have two groups, where one group receives a treatment (T) and the other group is a control group (C). For both groups, we register an outcome variable (e.g. reaction time = Y), then, we can condition the observed outcome on the group (X), giving $Y|X = T$ for the treatment group and $Y'|X' = C$ for the control group, where (Y, X) and (Y', X') are independent. The probablistic index will become

$$P\left(Y < Y'|X = T, X' = C\right), \tag{1.1}$$

which can be interpreted as the probability that a random reaction time from the treatment group will be faster than a random reaction time from the control group.

1.2.2 The Linear Regression Model as a PIM

The Linear Regression Model is defined as

$$Y = \boldsymbol{Z}^T\boldsymbol{\beta} + \varepsilon, \tag{1.2}$$

with $\boldsymbol{\beta} \in \mathbb{R}^p$, $\mathrm{E}(\varepsilon) = 0$, $\boldsymbol{X} \perp \varepsilon$, and the errors (i.e. ε) have a constant variance and are uncorrelated. The p-dimensional vector $\boldsymbol{Z}$ is a function of the d-dimensional covariate $\boldsymbol{X}$, e.g. if $d = 1$, with $\boldsymbol{X} = X$, then $\boldsymbol{Z}^T = (1, X)$ corresponds to a linear model with intercept, which is also linear in the parameters.

Now, let $(Y, \boldsymbol{X})$ and $\left(Y', \boldsymbol{X}'\right)$ be independent and identically distributed, then a PIM is defined as

$$P\left(Y < Y'|\boldsymbol{X}, \boldsymbol{X}'\right) + \frac{1}{2}P\left(Y = Y'|\boldsymbol{X}, \boldsymbol{X}'\right) = m\left(\boldsymbol{X}, \boldsymbol{X}'; \boldsymbol{\beta}\right), \tag{1.3}$$

with $m(\cdot)$ being a function with range $[0, 1]$. We intrduce the notation $P\left(Y \preccurlyeq Y'|\boldsymbol{X}, \boldsymbol{X}'\right)$ as defined in (1.3). We can rewrite (1.3) using the notation of the Linear Regression Model as defined in (1.2) and the newly introduced notation $P\left(Y \preccurlyeq Y'|\boldsymbol{X}, \boldsymbol{X}'\right)$

$$P\left(Y \preccurlyeq Y'|\boldsymbol{X}, \boldsymbol{X}'\right) = g^{-1}\left(\boldsymbol{Z}^T\boldsymbol{\beta}\right), \tag{1.4}$$

with $\boldsymbol{Z}^T$ a function of $\boldsymbol{X}$ and $\boldsymbol{X}'$ (often $\boldsymbol{Z} = \boldsymbol{X}' - \boldsymbol{X}$ is a meaningful and convenient choice, but not necessarily the only one), $g(\cdot)$ a sufficiently smooth link function that maps $[0, 1]$ onto the range of $\boldsymbol{Z}^T\boldsymbol{\beta}$. In this dissertation we will use the identity link for $g(\cdot)$.

For a more extensive discussion and explanation, we refer to Thas et al. (2012) and De Neve (2013), from which we selected the main concepts in order to understand the contents of this dissertation (proofs, theorems, and complex theoretical mathematics were disregarded).

1.3 Rank tests

Rank Tests are an example of non-parametric (or in some cases semi-parametric) tests. This gives the advantage that minimal assumptions are made concerning the underlying distribution from which the data are sampled. Because no underlying distribution is defined, this gives the advantage that data of almost all underlying distributions can be analyzed, and does not need to fulfill strong assumptions as we make in parametric statistics. Next to this, exact p-values can be obtained, but most of the test statistics of Rank Tests can be approximated by an asymptotic distribution, given that the sample is large enough.

Savage (1953) states that the true beginning of Rank Tests (and non-parametric statistics in general) is the publication of the paper on rank correlation by Hotelling and Pabst (1936). In the 30s and 40s of previous century we saw an immense expansion of non-parametric methods, e.g. Friedman (1937); Smirnov (1939); Wilcoxon (1945). Wilcoxon (1945) creates the famous two-sample rank sum test for equal sample sizes, which is later expanded to the general case by Mann and Whitney (1947). The following decades show an increase in development of non-parametric statistics and Rank Tests.

1.3.1 Quick example

To get a grasp on what Rank Tests are, we give a quick numeric example showing how the Wilcoxon-Mann-Whitney test is being calculated. Section 2.4 will go more in detail, so we will only provide some necessary information to get a basic understanding of Rank Tests.

The goal of the Wilcoxon-Mann-Whitney test is to compare two samples by means of their underlying distributions, where no assumptions are made about the distribu-

Treatment	Placebo
121 (7)	117 (4)
107 (1)	124 (10)
118 (5)	122 (8)
123 (9)	137 (14)
127 (11)	116 (3)
119 (6)	131 (13)
115 (2)	130 (12)

Table 1.1: Fictituous data for the blood pressure example. Systolic blood pressure (mmHg) per group after receiving the medication or placebo. Overall rank between parentheses.

tions. We can state a null hypothesis as $H_0 : F_1 = F_2$, where F_1 is the underlying distribution of the first sample and F_2 of the second sample (Hollander et al., 2013). The corresponding alternative hypothesis is that the distributions are not equal to each other. No further assumptions or restrictions are imposed on the distributions, and no common distribution is assumed. Suppose we have two groups, where one group takes a medication in order to lower blood pressure, and the other group receives a placebo. We can check if the first group shows a lower blood pressure compared to the second by ranking them. The overall rank is given by ranking all observations, regardless of the group the observation belongs to. The observed outcomes can be seen in Table 1.1. *(Note: These are completely fictitious numbers, and are just used for creating an easy example.)*

Under H_0 all $\binom{n_{\text{Placebo}}+n_{\text{Treatment}}}{n_{\text{Placebo}}}$ possible assignments for the ranks of the placebo group are equally likely, each having probability $1/\binom{n_{\text{Placebo}}+n_{\text{Treatment}}}{n_{\text{Placebo}}}$. Thus, each possible assignment for the ranks of the placebo group are 1/3432. Summing the ranks in the placebo group for each possible assignment would give results going from $\sum_{i=1}^{7} i = 28$ to $\sum_{i=8}^{1} 4i = 77$. Since most values are achieved using different combinations, not every value has the same probability of being achieved (the further from the mean, i.e. $(77+28)/2$, the lower the probability). Figure 1.1 shows the frequency distribution of the possible values we can achieve for the placebo group under H_0.

In a first step we sum the ranks of the placebo group:

$$W = 4 + 10 + 8 + 14 + 3 + 1312 = 64,$$

which is the original Wilcoxon two-sample rank sum statistic (Wilcoxon, 1945). We will reject H_0 if $W \geq w_\alpha$, where we choose the constant w_α to make the Type I Error Rate probability equal to α (we have a one-sided test since we want to investigate if the treatment lowers the blood pressure). If we take $w_\alpha = 65$, we obtain $\alpha = 0.064$ (due to the discrete nature of the test-statistic the probability of selecting exactly $\alpha = 0.05$ is almost 0). On the other hand, if we take $w_\alpha = 63$, we obtain $\alpha = 0.104$.

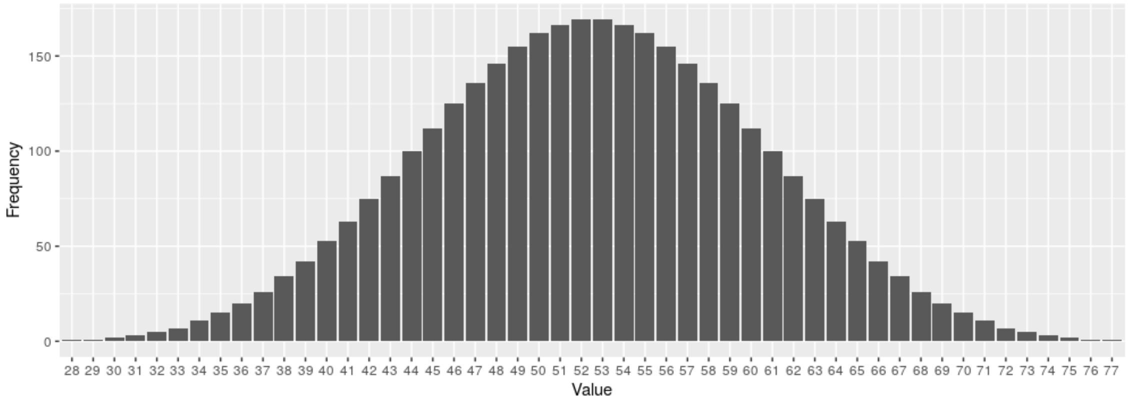

Figure 1.1: Frequency distribution of all possible values for the placebo group under H_0.

We will reject H_0 if we a priori set $\alpha = 0.10$, but not at $\alpha = 0.05$. If we look at the cumulative probabilities of the value 64 and higher, we get the p-value. In our case, this becomes $p = 0.082$ (which is an exact p-value, in contrast to parametric methods where approximations are used).

In Section 2.4, we will handle how the Wilcoxon-Mann-Whitney can be achieved using R.

1.3.2 Conclusion

Rank tests make use of the ranks of the observations, and not of summary-statistics or other statistics used in parametric methods, which allows flexibility concerning underlying distributions in contrast to parametric methods. When looking at our example, we can see some advantages of rank tests. First of all, we have made no assumptions concerning underlying distributions, instead we use the data as it is. Next to this, we have used the test with a small sample size (only 7 per group, making a total of 14). In parametric statistics this would have been a problem, because no normality assumptions can be made. Thirdly, we have obtained an exact p-value, based on the data as it is. But the classic rank tests do not provide effect sizes, standard errors, and confidence intervals. These shortcomings can be tackled using the PIM-framework. But in order to implement this in R, it is expected that the user has a very good knowledge of PIM on the one hand and rank tests on the other hand. For this reason, we provide convenient code in the rest of the dissertation in order for users who want to use a rank test can use it, without the need of digging deep in the theory of PIM, but only the basic knowledge of the interpretation of the provided statistics is needed.

In Chapter 2, we will handle the Wilcoxon-Mann-Whitney test statistic as described before, the Kruskal-Wallis test statistic, the Mack-Skillings test statistic (and by extension the Friedman test statistic), the Jonckheere-Terpstra test statistic. In addition we will provide a more generic method, as an extension of the Mack-Skillings test, where unbalanced data and even missing data can be handled.

1.4 Outline of the dissertation

In Chapter 2, we show how rank tests can be achieved using the PIM-framework, tackling the shortcomings of classic rank tests. This chapter is mainly based on De Neve and Thas (2015), with the main focus (and extension) on creating convenient R-code. In Chapter 3, an extension on Rank Tests is made. More specifically, we explain how a continuous covariate can be introduced in Rank Tests, using the PIM-framework, and of course, a convenient R-code is provided. A case study is provided in Chapter 4 in order to show how different tests described in this dissertation can be used to analyze data. Keuleers et al. (2010) have investigated classification and recognition in the Dutch language, using a lexical decision task comparing how Dutch native speakers react to words compared to non-words. We reproduce their conclusions, showing that the methods are usable in real research. To conclude, we discuss the advantages of rank tests using the PIM-framework and the code provided in Chapter 5. In the same chapter, we also discuss some shortcomings and provide some ideas for further research and development.

CHAPTER 2

RANK TESTS IN THE PIM-FRAMEWORK

2.1 Introduction

In our research concerning rank tests, we noticed that the most common rank tests are scattered around in R in different packages, and the syntax is not always easy or logical. For this reason, we have decided to use the same regression syntax for every function. Every function gives the known statistic from the conventional tests, and next to this, it also provides effect sizes, standard errors and confidence intervals are provided. We try to incorporate as much flexibility as possible within the function.

The theoretical part of this chapter is primarily based on De Neve and Thas (2015), with a few theoretical extensions, and the convenient R-codes as the most important extension on their research.

2.2 Kruskal-Wallis

The Kruskal-Wallis test is a non-parametric test statistic for comparing at least two samples by means of the underlying distribution. The null hypothesis for K samples can be stated as

$$H_0 : F_1 = \ldots = F_K = F_.,\tag{2.1}$$

where F_k denotes the underlying distribution from which sample k is drawn, and $F_. = \sum_{k=1}^{K} \lambda_k F_k$ with $\lambda_k = \lim_{N\to\infty} n_k/N$ where we assume $\lambda_k > 0$, $F_.$ can be interpreted as the marginal distribution. We can represent the relationship between the K different groups by rewriting $F_k = F(t + \tau_k)$, where τ_k is the treatment effect of group k. This let us rewrite (2.1) as

$$H_0 : [\tau_1 = \ldots = \tau_K] \tag{2.2}$$

The alternative hypothesis can then be stated that not all τ_k are equal, or in other words that at least two treament effects are not equal.

The Kruskal-Wallis test statistic can be computed as

$$KW = \frac{12}{N(N+1)} \sum_{k=1}^{K} n_k \left(\bar{R}_k - \frac{N+1}{2} \right)^2 , \tag{2.3}$$

where $\bar{R}_k$ denotes the average rank in the joint ranking of the sample k ($\bar{R}_k = \sum_{i=1}^{n_k} r_{ik}/n_k$, where n_k is the sample size of sample k) (Hollander et al., 2013).

When the smallest n_k tends to infinity, we can approximate the null distribution of the test-statistic KW with an asymptotic χ^2-distribution with $K-1$ degrees of freedom.

2.2.1 PIM alternative

We can model the same test statistic using the marginal PIM (De Neve and Thas, 2015), where marginal means we only condition on one treatment within the Probabilistic Index

$$P\left(Y_i \preccurlyeq Y_j | X_j\right), \tag{2.4}$$

where Y_i (resp. Y_j) denotes a random outcome in group i (resp. j), and X_j denotes the group to which the random outcome j belongs. When we look at this in terms of classical ANOVA, we can set $X_j = k$. Then (2.4) will become

$$P(Y. \preccurlyeq Y_k) = \alpha, \tag{2.5}$$

where Y_k is a random response from group k, with distribution F_k, and $Y.$ a random response, over all group, with marginal distribution $F.$. The interpretation of α is now straightforward, namely the probability that a random observation of group k exceeds a random observation pooled over all groups.

Now, we can rewrite this using the PIM regression model as

$$P\left(Y_i \preccurlyeq Y_j | X_j\right) = \boldsymbol{Z}_{ij}^T \boldsymbol{\alpha}, \tag{2.6}$$

where

$$\boldsymbol{Z}_{ij}^{T} = \left(\mathrm{I}(X_j = 1), \ldots, \mathrm{I}(X_j = K)\right), \tag{2.7}$$

with $\mathrm{I}(\cdot)$ being the indicator function (i.e. $I(\text{true}) = 1$ and $I(\text{false}) = 0$), considering all N^2 pairs of observations (Y_i, Y_j).

Now, let $\hat{\boldsymbol{\alpha}}$ be the estimator of $\boldsymbol{\alpha}$, then the Ordinary Least Squares estimator of α_k can be estimated using

$$\hat{\alpha}_k = \frac{1}{Nn_k} \sum_{i=1}^{N} \sum_{j=1}^{N} \mathrm{I}(X_j = k)\,\mathrm{I}(Y_i \preccurlyeq Y_j), \tag{2.8}$$

with $\mathrm{I}(Y_i \preccurlyeq Y_j)$ equaling 1 if $Y_i < Y_j$, 0.5 if $Y_i = Y_j$, and 0 otherwise.

De Neve and Thas (2015) show that we can rewrite the Kruskal-Wallis test statistic as

$$KW = \left(\hat{\boldsymbol{\alpha}} - \frac{1}{2}\mathbf{1}\right)^{T} \boldsymbol{\Sigma}_0^{-} \left(\hat{\boldsymbol{\alpha}} - \frac{1}{2}\mathbf{1}\right), \tag{2.9}$$

where $\mathbf{1}$ denote the unit vector of length K and $\boldsymbol{\Sigma}_0$ the covariance matrix of $\hat{\boldsymbol{\alpha}}$ under the null hypothesis of equal distributions.

$\boldsymbol{\Sigma}_0$ is the covariance matrix as defined under the null hypothesis, i.e. there is no difference between the groups, and thus every $\alpha_k = 0.5$. Because the pseudo-observations we construct (using the indicator function) possess a cross-correlation structure, i.e. if two pseudo-observations share a common outcome, they will in general not be independent, $\boldsymbol{\Sigma}_0$ is only valid under the null hypothesis (De Neve and Thas, 2015). For this reason De Neve and Thas (2015) also provide a sandwich estimator of the covariance matrix which is also consistent under the alternative hypothesis, i.e. $\hat{\boldsymbol{\Sigma}}$, which takes the possible correlation between the pseudo-observations into account. This sandwich estimator can now be used to construct a Wald-type Kruskal-Wallis test. With this Wald-type Kruskal-Wallis test we can formulate the null hypothesis as

$$H_0 : P(Y \preccurlyeq Y_k) = 0.5 \quad \forall k, \tag{2.10}$$

which is less restrictive than the original null hypothesis. When we use (2.10), we can interpret $\hat{\boldsymbol{\alpha}}$ as effect sizes, and $\hat{\boldsymbol{\Sigma}}$ can now be used for constructing confidence intervals for $\hat{\boldsymbol{\alpha}}$. In order to construct a confidence interval for $\hat{\alpha}_k$, the standard deviation $\hat{\sigma}_k$ (i.e. from the sandwich estimator $\hat{\boldsymbol{\Sigma}}$) is used. Then, the standard procedure is used, namely defining the z-value corresponding with the desired Type I Error Rate = γ[1] (e.g. when we want a Type I Error Rate of 5%, or in other words a 95%-confidence in-

[1] γ is used, instead of the traditional α to note the Type I Error Rate, to avoid misunderstandings, since α is already used as the coefficients of the model.

terval, the needed z-value becomes 1.96). The confidence interval is now calculated using

$$(1-\gamma)\mathrm{CI} = \left[\hat{\alpha}_k \pm z_{\gamma/2}\hat{\sigma}_k\right] \tag{2.11}$$

2.2.2 Convenient R code

We will illustrate this with a dataset as used in Keuleers et al. (2010). In order to investigate word recognition in the Dutch language, in particular lexical decision, a large scale study was performed. 39 participants (7 male, ranging in age from 19 to 46 with a mean age of 23) performed a lexical decision task. This lexical decision task consisted of 57 blocks of 500 trials (except the last block, which consisted of only 178 trials). One trial is defined as the lexical decision (correctly recognizing and classifying) of a stimulus, which can be a word or a pseudo-word. The stimuli comprised 14,089 Dutch words and 14,089 pseudo-words. Of the 14,089 words, 2,807 were one-syllabic and 11,212 were two-syllabic. All stimuli were presented exactly once over al 57 blocks in a completely randomized order. For a more complete overview of the design we refer to the paper of Keuleers et al. (2010) (e.g. how exactly stimuli were presented, how long the blocks took on average, the finanical compensation of the participants, the creation of the stimuli,. . .)

One of their research questions was if there is a difference in reaction time (coded in R as `rt`) between correct recognition and classification of pseudo-words, one-syllabic words and two-syllabic words (coded in R as `nsyl`) in the Dutch language. To investigate this research question, we use the mean reaction time of all participants on a certain stimulus, resulting in 20,178 observations. Because a very large sample will probably result in a large χ^2 value, we take a small sample of 300 stimuli, just to illustrate that the traditional Kruskal-Wallis can be achieved using PIM, and next to this that the PIM-alternative gives more information.

The Kruskal-Wallis test, in the traditional way, can be performed in R using:

```
> kruskal.test(rt~nsyl, data = data)

  Kruskal-Wallis rank sum test

data:  rt by nsyl
Kruskal-Wallis chi-squared = 8.6523, df = 2, p-value = 0.01322
```

We get a χ^2 value of 8.65 with 2 degrees of freedom, resulting in a p-value of 0.0132. So, we have evidence that there is a difference in reaction time between the recognition of non words, one-syllabic words, and two-syllabic words. But we have no idea where the difference can be found. Figure 2.1 gives an idea where this difference can be found, but only a formal test will give conclusive results.

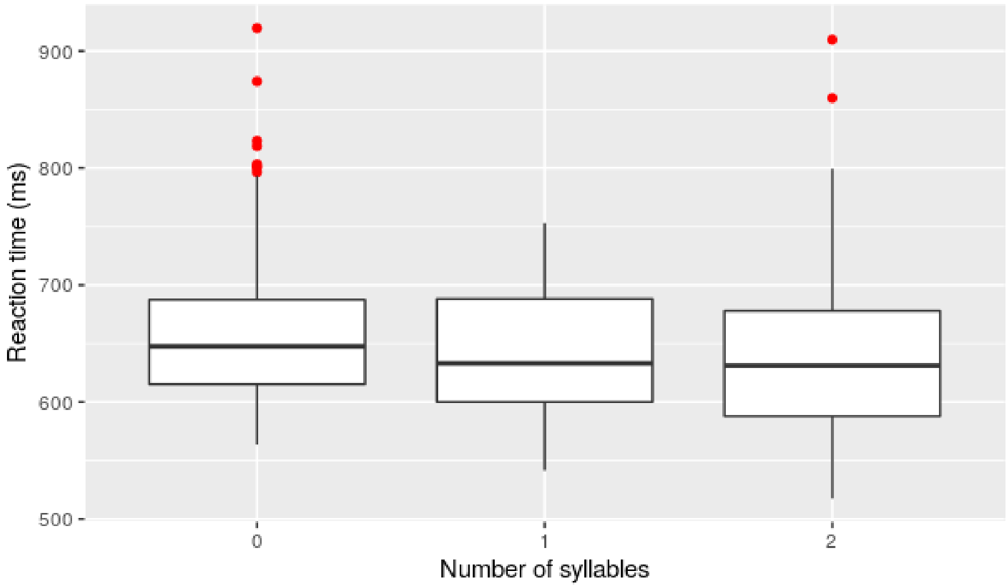

Figure 2.1: Reaction times in ms for pseudo-words (number of syllables = 0) and words of 1 and 2 syllables.

We have written a convenient code, using the syntax of the traditional test, which we illustrate below,

```
> fit <- kw.pim(rt~nsyl, data = data)
> summary(fit)
Summary of following PIM :
 Kruskal-Wallis

Formula:  rt ~ nsyl


      Estimate Std. Error z value Pr(>|z|)
nsyl1  0.54909    0.01709   2.872  0.00407 **
nsyl2  0.48423    0.05221  -0.302  0.76265
nsyl3  0.44581    0.01982  -2.734  0.00625 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

chi-squared = 8.652337 , df = 2 , p-value = 0.0132181

Null hypothesis: b = 0.5
```

We obtain the same value for the χ^2 (and the corresponding p-value), but now, we also have effect sizes. We have coded the first level of the factor as pseudo-words, the second level as one-syllabic words, and the third level as two-syllabic words. We get $\hat{\boldsymbol{\alpha}} = (0.549, 0.484, 0.446)$. Remember that the coefficient is the probability that the value of a random observation of the respective group exceeds a random observation over all groups, as defined in (2.5). Looking at the given p-values, we can say that pseudo-words are recognized slower as a pseudo-word and two-syllabic words are recognized faster as a word. More specific, we have a significant probability of 0.549 that a pseudo-word is recognized as a pseudo-word slower (i.e. have a higher reaction time) than the correct categorization of any stimuli (and a similar reasoning for a faster reaction time for two-syllabic words). This is confirmed when looking at the confidence intervals:

```
> confint(fit)
          2.5 %    97.5 %
nsyl1 0.5158589 0.5823142
nsyl2 0.3856581 0.5828082
nsyl3 0.4066432 0.4849688
```

We can see that 0.5 (the null hypothesis) is not in the 95% confidence intervals for the pseudo-words and the two-syllabic words.

When we only look at lexicality Figure 2.2 shows the difference in terms of the reaction time (in ms). Using the Kruskal-Wallis in the PIM-framework, we confirm the findings of Keuleers, Diependaele and Brysbaert (2010) that words (`lexicality2`) are recognized faster as words than pseudo-words are recognized as pseudo-words (`lexicality1`):

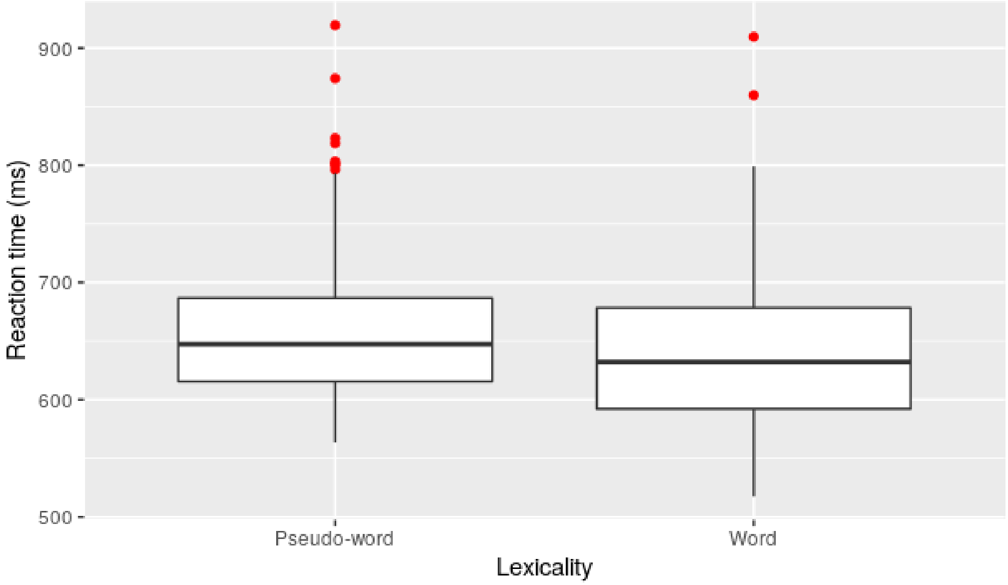

Figure 2.2: Reaction times in ms for pseudo-words and words.

```
> fit <- kw.pim(rt~lexicality, data = data)
> summary(fit)
Summary of following PIM :
 Kruskal-Wallis

Formula:  rt ~ lexicality


            Estimate Std. Error z value Pr(>|z|)
lexicality1  0.54736    0.01743   2.716   0.0066 **
lexicality2  0.45629    0.01609  -2.716   0.0066 **
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

chi-squared = 7.378399 , df = 1 , p-value = 0.006601192

Null hypothesis: b = 0.5
>
> confint(fit)
                2.5 %    97.5 %
lexicality1 0.5135057 0.5812082
lexicality2 0.4249257 0.4876462
```

We also added flexibility to the functions, so the null hypothesis in (2.10) can be changed and does not need to be equal to 0.5. The function `confint` also has the flexibility to differ from the nominal significance level of 0.05 and can be altered, as we are used to in R.

2.3 Mack-Skillings

Mack and Skillings (1980) have proposed a procedure and test statistic for testing the null hypothesis as stated in (2.1) against the corresponding alternative hypothesis for block data, where each block-treatment combination has an equal number of replications, i.e. $n_{ij} = n = N/(KL) \leq 1$, with K the number of treatment groups and L the number of blocks. We can define a block as a homogenous subgroup in which the the subjects are assigned before randomizing the treatment levels (e.g. within-subject design where the subject is a block, comparison of common trends in different languages where we randomize within a language,. . .) The adjusted null hypothesis now becomes

$$H_0 : F_{1l} = \ldots = F_{Kl},$$

with $l = 1, \ldots, L$, where F_{kl} denotes the distribution of group k within block l. In other words, the null hypothesis asserts that the underlying distributions of all K groups are the same within the same block. They propose the test statistic

$$MS = \frac{12}{K(N+L)} \sum_{k=1}^{K} \left(\bar{R}_k - \frac{N+L}{2} \right)^2, \tag{2.12}$$

where $\bar{R}_k = \sum_{l=1}^{L} \sum_{j=1}^{n} R_{klj}/n$, with R_{klj} denoting the rank of the jth replicate in the joint ranking of the response observations of treatment k within block l (in contrast to the Kruskal-Wallis where the joint ranking takes place over all observations, now the joint ranking only takes place within blocks).

We can look at the Mack-Skillings test statistic as an aggregate of several Kruskal-Wallis tests. In order to reduce errors, we only compare within certain blocks, which allows us to compare “apples with apples” and we avoid comparing “apples with oranges”. E.g. when we want to investigate a new medicine, it could be interesting to take into account that every person will not have the same reaction to different dosages of the medicine. Instead of comparing all results over participants, we can compare the effects of dosage within each patient (so-called blocks) and aggregate the findings in 1 test statistic.

When the null hypothesis is true, and when the common number of observations on each treatment, i.e. N/K, tends to infinity, the test statistic (2.12) has an underlying asymptotic χ^2 null distribution with $k-1$ degrees of freedom.

2.3.1 PIM alternative

We can extend the marginal PIM to block designs. Let us consider

$$P(Y_{.l} \preccurlyeq Y_{kl}) = \alpha_k, \tag{2.13}$$

with $k = 1, \ldots, K$ for the treatment groups and $l = 1, \ldots, L$ for the blocks, where $Y_{.l}$ is a random observation, over all groups, within block l, and Y_{kl} is a random observation from group k, within block l. If we integrate $\boldsymbol{Z}_{ij}$, as defined in (2.7), and $\boldsymbol{\alpha}$, using the defintion from (2.13), we can define the regression as

$$P\left(Y_i \preccurlyeq Y_j | B_i, X_j, B_j\right) = \boldsymbol{Z}_{ij}^T \boldsymbol{\alpha}, \tag{2.14}$$

with B_i (resp. B_j) denotes the block from which random observation i (resp. j) is drawn, where the comparisons of the Probabilistic Index are now restricted to comparisons within blocks (i.e. $B_i = B_j$). The interpretation of α_k as defined in (2.13) can now straightforwardly be interpreted as the probability that a random observation of treatment k exceeds a random observation of the marginal distribution within the same block.

Now, let $\hat{\boldsymbol{\alpha}}$ be the estimator of $\boldsymbol{\alpha}$, then the Ordinary Least Squares estimator of α_k can be estimated using

$$\hat{\alpha}_k = d_k \sum_{i=1}^{N} \sum_{j=1}^{N} \mathrm{I}\left(B_i = B_j\right) \mathrm{I}\left(X_j = k\right) \mathrm{I}\big) Y_i \preccurlyeq Y_j\big). \tag{2.15}$$

De Neve and Thas (2015) show that we can rewrite the Mack-Skillings test statistic as

$$MS = \left(\hat{\boldsymbol{\alpha}} - \frac{1}{2}\mathbf{1}\right)^T \boldsymbol{\Sigma}_0^{-} \left(\hat{\boldsymbol{\alpha}} - \frac{1}{2}\mathbf{1}\right), \tag{2.16}$$

where $\boldsymbol{\Sigma}_0$ is the covariance matrix of $\boldsymbol{\alpha}$ under the null hypothesis of equal distributions within blocks.

As explained in Section 2.2.1, $\boldsymbol{\Sigma}_0$ is only valid under the null hypothesis. Following the same reasoning as before, De Neve and Thas (2015) provide a sandwich estimator which can be used for Wald-type Mack-Skillings test statistic, which also results in the estimation of effect sizes and standard errors usable to construct confidence intervals.

2.3.2 Convenient R code

The research done by Keuleers et al. (2010) also compared Dutch proficiency with English (Keuleers et al., 2012) and French (Ferrand et al., 2010) proficiency. Since the English and French data are not restricted to one-syllabic and two-syllabic words, we will only compare the correct classifictation of pseudo-words and words. The Dutch data consists of 14,089 pseudo-words and 14,089 words, the English data consists of 28,515 pseudo-words and 27,134 words, the French data consists of 38,807 pseudo-words and 38,335 words. In the French language (Ferrand et al., 2010) and the English language (Keuleers et al., 2012), similar results have been found, namely that words are correctly classified and recognized faster compared to pseudo-words. But if we would analyze all data, regardless of the language we might get a wrong impression of the effect of lexical decision. Figure 2.3 shows the reaction time per language for pseudo-words and words. We can see that the difference between pseudo-words and words is the largest in the French language, and that the reaction time in the French language is much higher (regardless of lexicality) compared to reaction times in the English language and the Dutch language. For this reason, using blocks is a reasonable choice, so the comparison between pseudo-words and words is only done within the language and not over languages. In this context we consider the language in which the data was collected as the blocks. Because a very large sample will probably result in a large χ^2 value, we take a small sample of 300 stimuli, just to illustrate that the traditional Mack-Skillings test can be achieved using PIM, and next to this, that the PIM-alternative gives more information. Also, the sample of 300 stimuli must consist of 50 stimuli of each group, since the replicates per treatment-block combination must be equal.

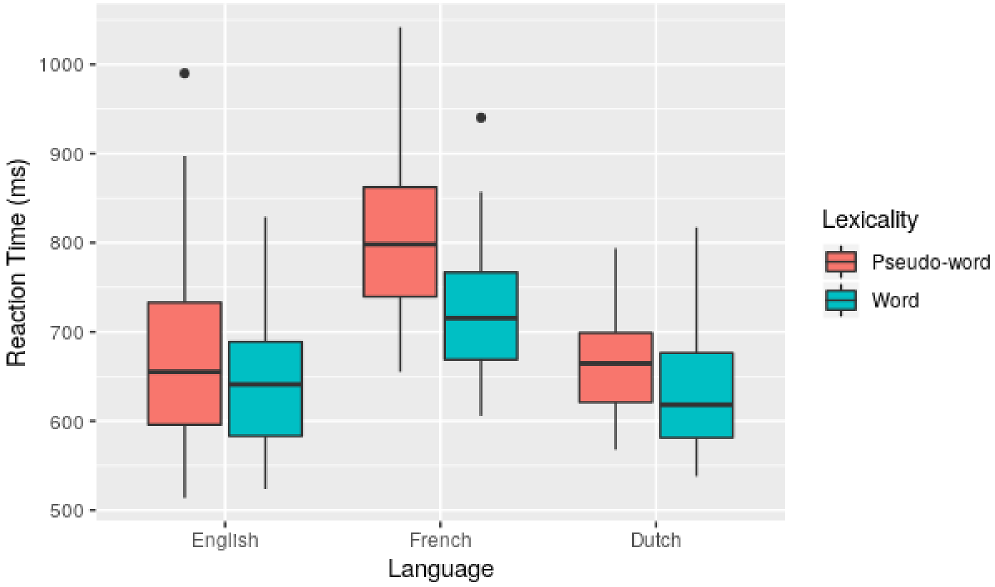

Figure 2.3: Reaction times in ms for pseudo-words and words in Dutch, English and French.

To our knowledge, there is only one active package with the Mack-Skillings test statistic, i.e. `asbio` with the function `MS.test` (Aho, 2019). The function is not easy to use, since the data must be entered in a rather unlogical way. Every block must be one column in a data-frame or matrix. Each row of this matrix must have the same treatment, and this is labeled by a seperate vector (which we create under `trt` in the R-code). Next to this, the number of replicates must be indicated.

The Mack-Skillings test, in the traditional way, can be performed in R using:

```
> library(asbio)
> dataNL <- data$rt[1:100] #selecting Dutch data
> dataEN <- data$rt[101:200] #selecting English data
> dataFR <- data$rt[201:300] #selecting French data
> dataTOT <- cbind(dataNL, dataEN, dataFR)
> trt <- c(rep(1,50),rep(2,50)) #vector to indicate treatment per row
>
> MS.test(dataTOT, trt, reps=50)
  df MS.test.stat P(Chi.sq>MS)
```

```
1  1     23.73348 1.106409e-06
```

We get a χ^2 value of 23.733 with 1 degree of freedom, resulting in a p-value of 1.106×10^{-6}. So, we have evidence that there is a difference in reaction time between the correct classifiction of pseudo-words and words, conditioned on languages. But we have no idea where the difference can be found.

We have written a convenient code, using the syntax of a traditional regression, which we illustrate below.

```
> fit <- mackskillings.pim(rt~lexicality|language, data=data)
> summary(fit)
Summary of following PIM :
 Mack-Skillings

Formula:  rt ~ lexicality | language


            Estimate Std. Error z value Pr(>|z|)
lexicalityN  0.58160    0.01675   4.872 1.11e-06 ***
lexicalityW  0.41840    0.01675  -4.872 1.11e-06 ***
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

chi-squared = 23.73348 , df = 1 , p-value = 1.106409e-06

Null hypothesis: b =  0.5
```

We achieve the same value for the χ^2 (and the corresponding p-value). But now we also have effect sizes. We have coded the first level of the factor as pseudo-words, the second level as words. We get $\hat{\boldsymbol{\alpha}} = (0.582, 0.418)$. We have a significant probability of 0.582 that lexical decision of a pseudo-word has a higher reaction time compared to the reaction time of the lexical decision of any stimuli (and a similar reasoning for the lower reaction time of a word), within the language of interest. Or in other words, a random pseudo-word (resp. word) has a probability of 0.582 (resp. 0.418) of being recognized slower compared to a random stimulus.

This is confirmed when looking at the confidence intervals:

```
> confint(fit)
                2.5 %    97.5 %
lexicalityN 0.5481743 0.6150257
lexicalityW 0.3849743 0.4518257
```

We can see that 0.5 (the null hypothesis) is not in the 95% confidence intervals.

The same flexibility as in the Kruskal-Wallis test is added to the Mack-Skillings test.

2.3.3 Friedman

The Friedman test statistic (Friedman, 1937) is a special case of the Mack Skillings test statistic, more specifically there is only one replication per treatment-block combination, i.e. $n = 1$ (Hollander et al., 2013). De Neve and Thas (2015) show that

$$\left(\hat{\boldsymbol{\alpha}} - \frac{1}{2}\mathbf{1}\right)^T \boldsymbol{\Sigma}_0^- \left(\hat{\boldsymbol{\alpha}} - \frac{1}{2}\mathbf{1}\right) = \frac{12L}{K(K+1)} \sum_{k=1}^{K} \left(\bar{R}_k - \frac{K+1}{2}\right)^2, \tag{2.17}$$

where $\boldsymbol{\Sigma}_0$ is the covariance matrix of $\boldsymbol{Z}_{ij}$ under the null hypothesis of equal distributions within blocks. The right hand side of (2.17) is exactly the Friedman test statistic.

Similar to the Mack-Skillings test, De Neve and Thas (2015) provide a sandwich estimator of the covariance matrix, with the same advantages.

Convenient R code

We used the complete datasets as described in Section 2.3.2. But since we can only have 1 replicate per block-treatment combination we have calculated the mean reaction time of every combination. These results can be seen in Table 2.1. We can see that within a language the mean reaction time of the lexical decision for a pseudo-word is higher compared to the mean reaction time of the lexical decision for a word. But this comparison only holds within a language. The question arises if the differences we observe are significant.

The Friedman test, in the traditional way, can be performed in R using:

```
> friedman.test(rt ~ lexicality|language, data = dataF)

  Friedman rank sum test

data:  rt and lexicality and language
Friedman chi-squared = 3, df = 1, p-value = 0.08326
```

We get a χ^2 value of 3 with 1 degree of freedom, resulting in a p-value of 0.0833. So, we have no evidence that there is a difference in reaction time between the correct classification of pseudo-words and words.

	English	French	Dutch
Word	639.38	739.98	639.08
Pseudo-word	654.50	807.83	662.93

Table 2.1: Mean reaction times (in ms) for pseudo-words and words in the English, French and Dutch language.

We have written a convenient code, using the syntax of the traditional test:

```
> fit <- friedman.pim(rt ~ lexicality|language, data = dataF)
> summary(fit)
Summary of following PIM :
 Friedman

Formula:  rt ~ lexicality | language


            Estimate Std. Error z value Pr(>|z|)
lexicalityN   0.7500     0.1443   1.732   0.0833 .
lexicalityW   0.2500     0.1443  -1.732   0.0833 .
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

chi-squared = 3 , df = 1 , p-value = 0.08326452

Null hypothesis: b =  0.5
```

We achieve the same value for the χ^2 (and the corresponding p-value). But now we also have effect sizes. The interpretation is exactly the same as the interpretation as in Section 2.3.2, namely that we have a probability of 0.75 (resp. 0.25) that the reaction time for the lexical decision of a random pseudo-wprd (resp. word) is higher than the reaction time of a random stimulus, within the language of interest. But, since we have $p > 0.05$, these probabilities do not differ significantly from 0.5 at the nominal level of 0.05.

On the other hand, based on the confidence intervals, we would reject the null hypothesis (since 0.5 is not in the interval), but we have to keep in mind that the confidence intervals are created with a sandwich estimator of the covariance matrix, and can give other significance

```
> confint(fit)
                2.5 %    97.5 %
lexicalityN 0.6085518 0.8914482
lexicalityW 0.1085518 0.3914482
```

For cases like this, there is also an extra flexbility in all functions concerning the estimator. When nothing is given to the argument `method` in the functions `summary`, the given standard errors, z-values, and corresponding p-values are obtained using $\hat{\boldsymbol{\Sigma}}_0$. The user can also give the value `"Wald"` to the argument `method`. In that case the given standard errors, z-values, and corresponding p-values are obtained using the sandwich estimator of the covariance matrix. Since the confidence intervals are only valid with the sandwich estimator, the standard output of function `confint` is based on the sandwich estimator. The user can also give the value `"default"` to use the default estimator of the covariance matrix, i.e. $\hat{\boldsymbol{\Sigma}}_0$ (here, the value `"default"` is used to keep the parallel between all functions). When the user alters the standard method a message is shown, in order to warn the user that non-standard procedures are used.

```
> summary(fit, method="Wald")
Summary of following PIM :
 Friedman

Formula:  rt ~ lexicality | language


            Estimate Std. Error z value Pr(>|z|)
lexicalityN  0.75000    0.07217   3.464 0.000532 ***
lexicalityW  0.25000    0.07217  -3.464 0.000532 ***
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

chi-squared = 48 , df = 1 , p-value = 4.262146e-12

Null hypothesis: b =  0.5

Keep in mind:
 The standard errors, z-values and corresponding p-values are based on the Wald-type covariance matrix. Difference in significance may occur.
>
> confint(fit, method="default")
                  2.5 %    97.5 %
lexicalityN  0.46710357 1.0328964
lexicalityW -0.03289643 0.5328964
Warning message:
In .local(object, parm, level, ...) :
  The confidence intervals are not obtained with the Wald statistic and may be inaccurate.
```

Next to this, the same flexbility as already described is also used for this test.

Skillings-Mack

Skillings and Mack (1981) have proposed a test statistic, based on the Friedman test, which can handle missing data. This can be done in R using the package `Skillings.Mack` (Srisuradetchai, 2015). Just like the Mack-Skillings test, we think the syntax is not very logical, so for this reason, we have made a convenient code with the well-known regression syntax as used in the Friedman-test.

Suppose we use the data as used for the Friedman-test, but we remove the words from the English language. The Skillings-Mack test, in the traditional way, can be performed in R using:

```
> Ski.Mack(data$rt,groups = data$lexicality,blocks = data$language)

Skillings-Mack Statistic =  1.333333 , p-value =  0.248213
Note: the p-value is based on the chi-squared distribution with d.f. =  1
```

When we compare this to the results of our convenient code, we observe the exact same χ^2 value and responding *p*-value. The interpretation remains the same as in the Friedman test.

```
> fit <- skillingsmack.pim(rt ~ lexicality|language, data=dataF2)
> summary(fit)
Summary of following PIM :
 Skillings-Mack

Formula:  rt ~ lexicality | language


             Estimate Std. Error z value Pr(>|z|)
lexicalityN 0.6666667  0.1443376  1.1547  0.24821
lexicalityW 0.3333333  0.1443376 -1.1547  0.24821

chi-squared = 1.333333333 , df = 1 , p-value = 0.248213079

Null hypothesis: b =  0.5
```

2.3.4 Missing data and unbalanced design

As mentioned, the Skillings-Mack test can handle missing data. But, the PIM-framework can also handle missing data and does not need a balanced design. For this reason we have created a more generic function where no restrictions in the design of the data are imposed. When using data suitable for the Mack-Skillings or Friedman, our function will give exactly the same results as those functions, but as soon missing data or an unbalanced design occurs no data is imputed, in contrast to the Skillings-Mack test (Hollander et al., 2013; Skillings and Mack, 1981), and we will achieve other results. The function `blockedrank.pim` takes exactly the same arguments as the Friedman, Mack-Skillings, and Skillings-Mack.

Suppose, we use the data as used for the Mack-Skillings, but instead of selecting 50 replications of each treatment-block combination we use different values for each combination. In our example we use 52 for the Dutch words, 49 for the Dutch pseudo-words, 55 for the English words, 47 for the English pseudo-words, 33 for the French words, and 68 for the French pseudo-words (these values were picked totally random).

```
> fit <- blockedrank.pim(rt~lexicality|language, data=dataMS)
> summary(fit)
Summary of following PIM :
 K-sample rank test with blocks

Formula:  rt ~ lexicality | language


              Estimate Std. Error  z value Pr(>|z|)
lexicalityN 0.56914093 0.01510074  4.57864 4.68e-06 ***
lexicalityW 0.41909123 0.01767090 -4.57864 4.68e-06 ***
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

chi-squared = 4.84070277 , df = 1 , p-value = 0.02779556539

Null hypothesis: b =  0.5
```

The interpretation remains exactly the same as before with the Mack-Skillings test and Friedman test. We have implemented the same flexibility as for all functions before.

2.4 Wilcoxon-Mann-Whitney

The Wilcoxon-Mann-Whitney test is a non-parametric test statistic for comparing two groups by means of the underlying distributions of each sample. Just like in the previous tests, we are now interested in the joint ranking of the observations. We want to know how likely it is that a randomly selected value from one group is greater and/or less than a randomly selected value of the other group.

The null hypothesis can be stated as

$$H_0 : F_1 = F_2 \tag{2.18}$$

where F_1 (resp. F_2) denotes the underlying distribution from which sample 1 (resp. 2) is drawn. In other words, we state in the null hypothesis (2.18) that both samples have the same underlying probability distribution, but the common distribution is not explicitly specified (Hollander et al., 2013). The alternative hypothesis can be stated as $\mathrm{E}(Y_1) - \mathrm{E}(Y_2) \neq 0$ or $P(Y_1 < Y_2) \neq 0.5$ (which can also be specified in a one-sided alternative).

The Wilcoxon-Mann-Whitney test statistic can be computed as

$$W = \sum_{i=1}^{n_2} R_i, \tag{2.19}$$

where n_2 denotes the sample size of sample 2 and R_i denotes the overall joint rank of the ith observation in sample 2 (Hollander et al., 2013).

When the null hypothesis is true and for large samples, we can approximate the distribution of the test-statistic W with a normal distribution with

$$\mathrm{E}_0(W) = \frac{n_2(n_1 + n_2 + 1)}{2} \tag{2.20}$$

$$\mathrm{Var}_0(W) = \frac{n_1 n_2(n_1 + n_2 + 1)}{12}, \tag{2.21}$$

in the absence of ties.

2.4.1 PIM alternative

We can model the standardized Wilcoxon-Mann-Whitney test statistic using the PIM-framework using a PIM that models pairwise comparisons of groups (De Neve and Thas, 2015)

$$P\left(Y_i \preccurlyeq Y_j | X_i, X_j\right), \tag{2.22}$$

with Y_i (resp. Y_j) being a randomly selected value from group i (resp. j), and X_i (resp. X_j) denotes the group to which observation i (resp. j) belongs. When we look at this in terms of classical ANOVA, we can set $X_j = k$. For the 2-sample design (Wilcoxon-

Mann-Whitney) we can rewrite this as following PIM:

$$P(Y_1 \preccurlyeq Y_2) = \alpha_{12}, \tag{2.23}$$

where Y_1 (resp. Y_2) is a random response from group 1 (resp. 2), with distribution F_1 (resp. F_2). The interpretation of α is now straightforward, namely the probability that a random observation of group 2 exceeds a random observation of group 1.

Now, we can rewrite this using the PIM regression model as

$$P\left(Y_i \preccurlyeq Y_j | X_i, X_j\right) = Z_{ij}\alpha, \tag{2.24}$$

where

$$Z_{ij} = \left(\mathrm{I}(X_i = 1)\mathrm{I}\left(X_j = 2\right)\right). \tag{2.25}$$

We restrict the Probablistic Index to all unique treatment combinations.

De Neve and Thas (2015) show that we can rewrite the standardized Wilcoxon-Mann-Whitney test statistic as

$$\frac{\hat{\alpha}_{12} - 0.5}{\sigma_0} = \frac{\sum_{\{i|X_i=1\}} \sum_{\{j|X_j=2\}} \mathrm{I}\left(Y_i \preccurlyeq Y_j\right) - n_1 n_2/2}{\sqrt{[n_1 n_2(n_1 + n_2 + 1)]/12}}, \tag{2.26}$$

where σ_0 denote standard deviation of $\hat{\alpha}$ under the null hypothesis of equal distributions.

We can estimate α_{12} using

$$\hat{\alpha}_{12} = \frac{1}{n_1 n_2} \sum_{i=1}^{N} \sum_{j=1}^{N} \mathrm{I}(X_i = 1)\mathrm{I}\left(X_j = 2\right)\mathrm{I}(X_i = 1)\mathrm{I}\left(Y_i \preccurlyeq Y_j\right). \tag{2.27}$$

Since σ_0 is only valid under the null hypothesis, confidence intervals based on these covariances are not valid. De Neve and Thas (2015) provide a sandwich estimator of the standard error which is also consistent under the alternative hypothesis, i.e. $\hat{\sigma}$. This sandwich estimator can now be used to construct a Wald-type Wilcoxon-Mann-Whitney test. With this Wald-type Wilcoxon-Mann-Whitney test we can formulate the null hypothesis as

$$H_0 : P(Y_1 \preccurlyeq Y_2) = 0.5, \tag{2.28}$$

which is less restrictive than the original null hypothesis. When we use (2.28), we can interpret $\hat{\alpha}$ as an effect size, and $\hat{\sigma}$ can now be used for constructing a confidence interval for $\hat{\alpha}$.

2.4.2 Convenient R code

We will use the dataset from Keuleers et al. (2010) once more. In Section 2.2.2 we have analysed the correct recognition and classification of pseudo-words and words (only looking at lexicality). This was done using the null hypothesis of the Kruskal-Wallis test, which means a common distribution is taken into consideration. When using the Wilcoxon-Mann-Whitney test, we only are intersted in the underlying distributions of both samples and what the probability is that a randomly selected reaction time from the pseudo-words is greater and/or less than a randomly selected reaction time from the words. We will use exactly the same data as the one in Section 2.2.2.

The Wilcoxon-Mann-Whitney test, in the traditional way, can be performed in R using:

```
> wilcox.test(rt~lexicality, data = data)

  Wilcoxon rank sum test with continuity correction

data:  rt by lexicality
W = 13271, p-value = 0.006614
alternative hypothesis: true location shift is not equal to 0
```

Here no large-sample approximation is used, but the p-value is an exact p-value based on permutations. The reported test-statistic is not the W as defined in (2.19). The reported test-statistic can be achieved using

$$U = W - \frac{n_2(n_2+1)}{2} \tag{2.29}$$

$$U' = n_1 n_2 - U, \tag{2.30}$$

where U' is the reported `W` in the `wilcox.test` in R.

So, we have evidence that there is a difference in reaction time between the correct recognition and classification of pseudo words and words. But we have no idea where the difference can be found.

We have written a convenient code, using the syntax of the traditional test:

```
> fit <- wilcox.pim(rt~lexicality, data = data)
> summary(fit)
Summary of following PIM :
 Wilcoxon-Mann-Whitney

Formula:  rt ~ lexicality


             Estimate Std. Error z value Pr(>|z|)
lexicalityNW  0.40923    0.03342  -2.716   0.0066 **
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

z-value = -2.716321 , p-value = 0.006601192

Null hypothesis: b =  0.5
Alternative =  two.sided
```

The *p*-value is almost equal, keeping in mind that the original `wilcox.test` is an exact *p*-value and our *p*-value is a large sample approximation using the standard normal distribution. Remember from (2.23) that the coefficient is interpreted as the probability that a randomly selected value of the second category (`W` in our case) is higher than a randomly selected value of the first category (`N` in our case). In our case, we have a probability of 0.41 that a randomly selected reaction time of the word-group is higher than a randomly selected reaction time of the pseudo-word group. In other words and more clearly, we have evidence that the reaction time for the correct categorization of words are significantly lower than the reaction time for the correct categorization of pseudo-words. To show that both tests are linked, we need to transform the reported `W` from `wilcox.test` to the *z*-value. For this, we need the sample size of both groups

```
> table(data$lexicality)

  N   W
144 156
```

This leads to following conversion

$$\begin{aligned} U &= n_1 n_2 - U^* = 9193 \\ W &= U + \frac{n_2(n_2+1)}{2} = 21,439 \\ W' &= \frac{W - \mathrm{E}_0(W)}{\sqrt{\mathrm{Var}_0(W)}} = -2.716 \end{aligned}$$

where W' is the standardized test statistic, i.e. the z-value, using (2.20) and (2.21).

We also added flexibility to the functions, so the null hypothesis in (2.28) can be changed and does not need to be equal to 0.5. The function `confint` also has the flexibility to differ from the nominal significance level of 0.05 and can be altered, as we are used to in R. As an extra flexibility, the possibility to change the inequality for the alternative hypothesis is added, since the test-statistic can be approximated by the standard normal distribution and can thus be one-sided. In our example, we can a priori assume that words will be correctly recognized and classified faster than pseudo-words, and thus formulate this as the alternative hypothesis. In our R-function this becomes

```
 1 > fit <- wilcox.pim(rt~lexicality, data = data, alternative = "less")
 2 > summary(fit)
 3 Summary of following PIM :
 4  Wilcoxon-Mann-Whitney
 5
 6 Formula:  rt ~ lexicality
 7
 8
 9             Estimate Std. Error z value Pr(>|z|)
10 lexicalityW  0.40923    0.03342  -2.716   0.0066 **
11 ---
12 Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
13
14 z-value = -2.716321 , p-value = 0.003300596
15
16 Null hypothesis: b =  0.5
17 Alternative =  less
```

The overall p-value can be found on line 14, because this is the one-sided p-value (in contrast to the p-value on line 10).

2.5 Jonckheere-Terpstra

Often, we encounter a setting where the treatment groups can be ordered in terms of expected effects. In contrast to the Kruskal-Wallis, the Jonckheere-Terpstra allows for specification of differences between groups in terms of increasing (or decreasing) expected effects (Hollander et al., 2013). The null hypothesis for K treatments can be stated as

$$H_0 : F_1 = \ldots = F_K. \qquad (2.31)$$

When these K treatments can be ordered in an upward trend, De Neve (2013) shows the alternative hypothesis can be formulated as follows

$$H_a : \frac{2}{K(K-1)} \sum_{k=1}^{K-1} \sum_{l=k+1}^{K} P\left(Y_l \preccurlyeq Y_k\right) > 0.5. \tag{2.32}$$

The Jonckheere-Terpstra statistic is given by

$$JS = \frac{\sum_{k=1}^{K-1} \sum_{l=k+1}^{K} \sum_{i|X_i=k} \sum_{j|X_j=l} \left(\mathrm{I}\left(Y_i \preccurlyeq Y_j\right) - \left(\frac{N^2 - \sum_{j=1}^{K} n_j^2}{4} \right) \right)}{\frac{N^2(2N+3) - \sum_{j=1}^{K} n_j^2 (2n_j+3)}{72}}, \tag{2.33}$$

which can be obtained using the PIM regression

$$P\left(Y_i \preccurlyeq Y_j | X_i, X_j\right) = \frac{1}{2} + \alpha Z_{ij}, \tag{2.34}$$

where $Z_{ij} = \mathrm{I}\left(X_i < X_j\right) - \mathrm{I}\left(X_i > X_j\right)$ and all unique treatment combinations are considered. The null hypothesis in this setting can be written as

$$H_0 : P\left(Y_i \preccurlyeq Y_j | X_i < X_j\right) = 0.5, \tag{2.35}$$

with the simple alternative that $P\left(Y_i \preccurlyeq Y_j | X_i < X_j\right) \neq 0.5$ (which can also be formulated as a one sided alternative). The interpretation of α is now the probability that a random outcome of a higher factor level exceeds a random outcome of a lower factor level, reduced with 0.5. By definition it is guaranteed that when the level of both random observations are the same that $P\left(Y_i \preccurlyeq Y_j | X_i = X_j\right) = 0.5$. This gives us the ability to rewrite the null hypothesis as $H_0 : \alpha = 0$ versus the alternative hypothesis $H_a : \alpha \neq 0$ (De Neve and Thas, 2015).

For the large-sample approximation of the Jonckheere-Terpstra test statistic, we use the standard normal distribution. More specifically, as $\min(n_1, \ldots, n_k)$ tends to infinity, we can model the Jonckheere-Terpstra test statistic with an asymptotic standard normal distribution. This allows us to formulate the alternative hypothesis also with a priori specific inequality (Hollander et al., 2013).

2.5.1 Convenient R code

We will use the dataset from Keuleers et al. (2010) once more. In Section 2.2.2 we have analysed the correct recognition and classification of pseudo-words, one-syllabic words and two-syllabic words. Based on research in other languages (Ferrand et al., 2010) we can assume that shorter words will be correctly recognized and classified faster, and words in general are correctly recognized and classified faster compared to pseudo-words. In other words, we can order the alternatives in terms of reaction time as follows One-syllabic words < Two-syllabic words < Pseudo-words.

The Jonckheere-Terpstra test can be found in R in the package `clinfun` (Seshan, 2018) and in the package `kSamples` (Scholz and Zhu, 2019). Both give the same results, so only one will be handled here, specifically the function `jonckheere.test` from the package `clinfun` (due to its flexibility concerning the alternative hypothesis).

```
> jonckheere.test(data$rt, data$nsyl)

  Jonckheere-Terpstra test

data:
JT = 15293, p-value = 0.003241
alternative hypothesis: two.sided
```

We have evidence for our ordered alternative hypothesis. But we still do not know how strong the effect is.

We have written a convenient code, still using the regression syntax

```
> fit <- jt.pim(rt~nsyl, data=data)
> summary(fit)
Summary of following PIM :
 Jonckheere-Terpstra

Formula:  rt ~ nsyl


                            Estimate Std. Error z value Pr(>|z|)
P(rti <= rtj|nsyli < nsylj)  0.58840    0.03003   2.944  0.00324 **
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Jonckheere-Terpstra Statistic =  15293
z-value =  2.943902 p-value =  0.003241024

Null hypothesis: P(Yi <= Yj|Xi < Xj) = 0.5
Alternative =  two.sided
>
> confint(fit)
                                2.5 %    97.5 %
P(rti <= rtj|nsyli < nsylj) 0.5305968 0.6461951
```

We achieve the same p-value. Next to this we report the Jonckheere-Terpstra statistic, although this is not automatically calculated in the PIM-setting. Instead of reporting the coefficient α as defined in (2.34), we report $0.5+\alpha$, since this can be straightforwardly interpreted as the probability that the reaction time of a random one-syllabic word is lower than the reaction time of a random two-syllabic word or a pseudo-word, and the reaction time of a random two-syllabic word is lower than the reaction time of a random pseudo-word (i.e. keeping the ordered alternatives in mind).

We have added the same flexibility as is implemented in the Wilcoxon-Mann-Whitney test (being the flexibility of all functions plus the possibility of a one-sided *p*-value).

CHAPTER 3

RANDOMIZED DESIGNS WITH A COVARIATE

3.1 Introduction

De Neve and Thas (2015) show that, using the PIM-framework, flexibility can be integrated in order to create rank tests for more complicated designs. More specifically, we will handle their proposal for testing for a factor effect while controlling for a continious covariate. If we consider a randomized design with a K-leveled factor X, which is randomized over the units. Given that X is of interest to test the null hypothesis $H_0 : F_1 = \ldots\ F_K$, with F_k the conditional distribution of the outcome Y, conditional on $X = k$. In addition, we measure a continuous covariate Z, associated with Y and independent of X (as a consequence of the complete randomization). Suppose we let X be a binary factor (i.e. $K = 2$), a Wilcoxon-Mann-Whitney test can be performed. De Neve and Thas (2015) construct a rank test where a covariate Z is added as extension of the Wilcoxon-Mann-Whitney:

$$P\left(Y_i \preccurlyeq Y_j | \boldsymbol{X}_i, \boldsymbol{X}_j\right) = \beta + \gamma\left[f(Z_j) - f(Z_i)\right], \tag{3.1}$$

with $\boldsymbol{X} = (X, Z)$, $f(\cdot)$ a known function (i.e. the identity function), using only unique combinations of $\boldsymbol{X}_i$ and $\boldsymbol{X}_j$. They show that

$$\sqrt{N}\frac{\hat{\beta} - P\left(Y_i \preccurlyeq Y_j | X_i = 1, X_j = 2\right)}{\sigma} \xrightarrow{\text{d}} N(0,1), \tag{3.2}$$

as $N \to \infty$, where $\hat{\beta}$ is a consistent estimator for the probability $P\left(Y_i \preccurlyeq Y_j | X_i = 1, X_j = 2\right)$ (which does not depend on Z and corresponds to the population parameter tested for by the Wilcoxon-Mann-Whitney test).

3.2 Extension to 2 or more groups

In this manuscript, we write a code which extends the possibility to more than two groups. Keeping (3.2) in mind, a statistic with more than two groups, will follow a χ^2 distribution with degrees of freedom equal to all possible pairs of the groups, being $df = \frac{k!}{2(k-2)!}$ with k the number of groups (Leemis, 1986). When only two groups are used, the p-value for this test, without any covariates, will be exactly the same as the p-value as achieved with our version of the Wilcoxon-Mann-Whitney test in its two-sided alternative. Evidently, when a covariate is used, the p-value will not be the same. In other words, the two-sided Wilcoxon-Mann-Whitney as described in Section 2.4 can be modelled with this test by just leaving out the covariate, but in addition more power can be added by introducing a covariate.

3.3 Convenient R code

We have written a code with a regression syntax to easily incorporate a covariate. The following syntax is used: `rank_covar.pim(response ~ group | covariate, data)`. The argument for the covariate can be left empty, giving following syntax: `rank_covar.pim(response ~ group, data)`. When the group-variable only contains 2 groups, the result of this function is the same as the Wilcoxon-Mann-Whitney test, with the only difference that a χ^2 statistic is reported and thus only a two-sided test can be performed. But in contrast to the Wilcoxon-Mann-Whitney test, we are not limited to 2 groups.

Keuleers et al. (2010) also checked if there was a learning effect. The test was taken in parts, more specifically in 58 blocks. They found that later blocks showed faster reaction times in general, and that the difference between pseudo-words and words decreased when a learning effect took place.

To see if there is any trend of difference between blocks we can model the data once without the blocks as a covariate and once with the blocks as a covariate. The complete data consists 1,098,942 reaction times[1]. These were found in 57 blocks (1 block was removed from the data because it was a replication of another block) over 39 participants. For pragmatic reasons, we took a random sample of 1000 reaction times, because the larger the dataset, the longer the calculation of the computer takes, and the more memory is needed to allocate temporary matrices (e.g. for the complete dataset this would be approx. 892 gigabytes).

[1] This study will be handled in detail in the next chapter (including structure of data and results). For now, we only wish to illustrate the R-syntax.

We first analyzed the reaction times looking at the lexicality

```
> fit <- rank_covar.pim(rt ~ lexicality | block, data = data.sm)
> summary(fit)
Summary of following PIM :
 Rank test with covariate

Formula:  rt ~ lexicality | block


             Estimate Std. Error z value Pr(>|z|)
lexicalityNW  0.44362    0.01829  -3.082  0.00205 **
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

chi-squared = 9.500161 , df = 1 , p-value = 0.008650999

Null hypothesis: b =  0.5
>
> fit2 <- rank_covar.pim(rt ~ lexicality, data = data.sm)
> summary(fit2)
Summary of following PIM :
 Rank test with covariate

Formula:  rt ~ lexicality


             Estimate Std. Error z value Pr(>|z|)
lexicalityNW  0.44437    0.01829  -3.042  0.00235 **
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

chi-squared = 9.25217 , df = 1 , p-value = 0.002352165

Null hypothesis: b =  0.5
```

The model without the covariate (`fit2` gives a significant effect, with the probability of 0.44 that we observe a lower reaction time for a pseudo-word compared to the reaction time of a word, with a $\chi^2(1) = 9.25.$ When controlling for the covariate, we observe a probability further to 0.5, still significant, with a $\chi^2(1) = 9.50)$. This follows the findings of Keuleers et al. (2010). There was a trend for a learning effect, but this was very small. Keuleers et al. (2010) found no significant interaction.

We can do a similar comparison for the number of syllables

```
> fit3 <- rank_covar.pim(rt ~ nsyl | block, data = data.sm)
> summary(fit3)
Summary of following PIM :
 Rank test with covariate

Formula:  rt ~ nsyl | block


       Estimate Std. Error z value Pr(>|z|)
nsyl12  0.40514    0.03428  -2.767  0.00566 **
nsyl13  0.45007    0.01930  -2.587  0.00969 **
nsyl23  0.55187    0.03512   1.477  0.13974
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

chi-squared = 21.46438 , df = 3 , p-value = 0.000256123

Null hypothesis: b =  0.5
>
> fit4 <- rank_covar.pim(rt ~ nsyl, data = data.sm)
> summary(fit4)
Summary of following PIM :
 Rank test with covariate

Formula:  rt ~ nsyl


       Estimate Std. Error z value Pr(>|z|)
nsyl12  0.40526    0.03428  -2.763  0.00572 **
nsyl13  0.45085    0.01930  -2.547  0.01087 *
nsyl23  0.55208    0.03512   1.483  0.13813
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

chi-squared = 20.16686 , df = 3 , p-value = 0.000156749

Null hypothesis: b =  0.5
```

In both models, we observe significant coefficients. The coding used is 1 for pseudo-words, 2 for 1 syllable and 3 for 2 syllables. When looking at the model without a covariate, he probability that we observe a faster reaction time for pseudo-words in comparison to one-syllabic words (resp. two-syllabic words) is 0.405 (resp. 0.451). One-syllabic words are recognized and classified faster than two-syllabic words, with a probability of 0.552 (but the difference is not significant. The trends are the same when the covariate is used, but no real evidence for a learning effect is seen.

CHAPTER 4

CASE STUDY

4.1 Introduction

Keuleers et al. (2010) created the Dutch Lexicon Project, where participants made lexical decisions for more than 14,000 words and more than 14,000 pseudo-words. 39 participants (all students and employees of Ghent University), from which 7 are male and 32 are female, ranging in age from 19 to 46 with a mean age of 23, performed the lexical decision task. This lexical decision task consisted of 57 blocks of 500 trials (except the last block, which consisted of only 178 trials). One trial is defined as the lexical decision (correctly recognizing and classifying) of a stimulus, which can be a word or a pseudo-word. The stimuli comprised 14,089 Dutch words and 14,089 pseudo-words. Of the 14,089 words, 2,807 were one-syllabic and 11,212 were two-syllabic. All stimuli were presented exactly once over al 57 blocks in a completely randomized order. Pseudo-words are non-words that are pronouncable and have similarities with words, but have no meaning (e.g. keuger, trafend, grering, landast). They investigated the difference between recognition and correct classificiation of psuedo-words and words, where the words could consist of 1 syllable or 2 syllables.

A complete trial consisted of following events:

1. Two vertical fixation lines were presented slightly above and below the center of the monitor
2. 500 milliseconds later the stimulus (i.e. word or pseudo-word) appeared between the two vertical lines
3. The participant had to answer within 2 seconds whether the stimulus is a word or a pseudo-word by pressing one of two buttons (with their dominant hand for a word and with their non-dominant hand for a pseudo-word)
4. The screen was cleared for 500 milliseconds in order for the next trial to start

They hypothesized, based on previous research in French (Ferrand et al., 2010), that words are recognized and classified faster than pseudo-words. In addition, they hypothesized that longer words are recognized and classified slower thant shorter words.

We only report results within the Dutch language, where the analyses are equivalent as in Chapter 3 and Chapter 4. We will skip the Mack-Skillings test, Friedman test, Skillings-Mack test and Blocked Rank test.

4.2 Data preparation

We have used the data of Keuleers et al. (2010, 2012); Ferrand et al. (2010) which can be found on their website http://crr.ugent.be/programs-data/lexicon-projects. We have removed stimuli which had no reaction time, and for the Dutch data we also used the number of syllables. On trial level we removed outliers, based on the median absolute deviation (Leys et al., 2013), looking at the data within blocks and within participants.

Since we are not interested in the reaction times on the level of the participants, but the reaction time for a stimuli within a block, the mean reaction time was calculated if a stimulus appeared in the same block for different participants. This results in $692,143$ mean reaction times, with an average of $12,267$ mean reaction times per block (except for the last block, which has $5,209$ mean reaction times). Within each block, we find an average of $6,405$ mean reaction times for a pseudo-word (except for the last block, which has $2,702$ mean reaction times), an average of 5862 mean reaction times for a word ($2,502$ for the last block). Looking more closely at the words within a block, an average of $1,139$ mean reaction times belong to one-syllabic words (516 in the last block), and an average of $4,674$ mean reaction times belong to two-sylllabic words ($1,970$ in the last block).

Because large datasets would result in high test-values and low *p*-values and give a wrong idea about the results (Lin et al., 2013) and the computational power is often limited, we took a random sample of 1,000 observations of the complete data.

All the transformed/cleaned data can be found on the github-page of this manuscript.

4.3 Results

Figure 4.1 shows that pseudo-words are recognized faster than words, on average, and the reaction time becomes faster when the participant is doing the task in later blocks (i.e. learning effect). Figure 4.2 shows that, when focussing on the distinction within words, one-syllabic words are recognized faster compared to two-syllabic words, but both are still recognized faster compared to pseudo-words.

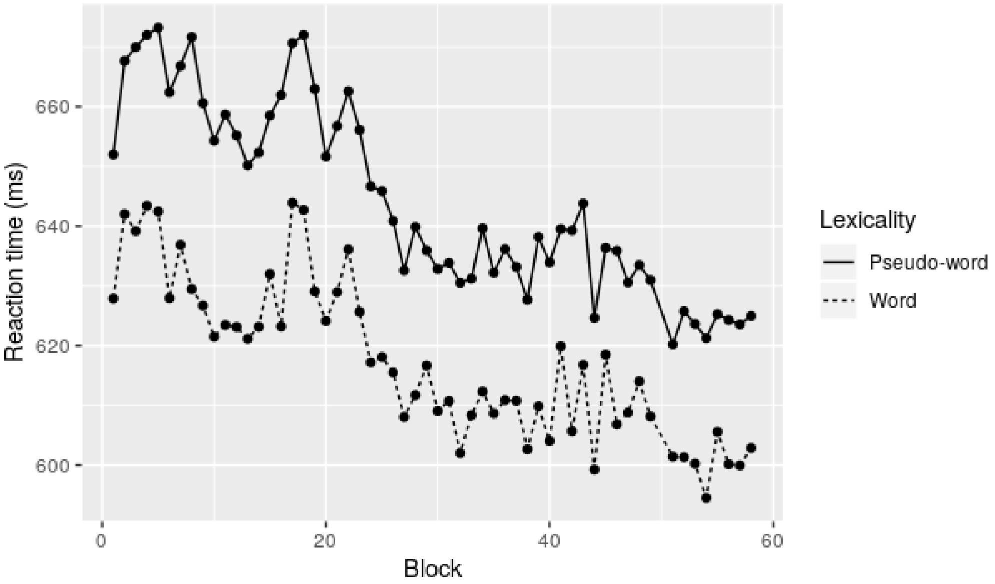

Figure 4.1: Average reaction times in ms per block for the lexicality

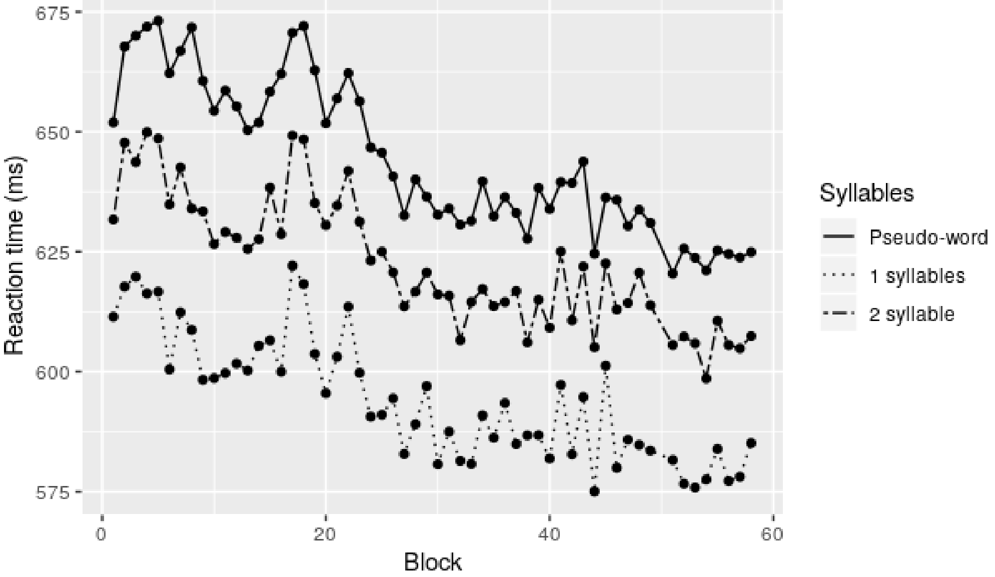

Figure 4.2: Average reaction times in ms per block for the number of syllables

4.3.1 Lexicality

To assess whether there is a difference between the categorization and recognition of pseudo-words versus non-words two different tests can be used, namely the Kruskal-Wallis (Section 2.2) and the Wilcoxon-Mann-Whitney test (Section 2.4). The difference between the two tests is that the Kruskal-Wallis test compares against a common distribution between the two groups, where the Wilcoxon-Mann-Whitney test just compares the two groups against each other.

The Kruskal-Wallis test gives a significance difference between the two groups ($\chi^2(1) = 9.25$, $p = 0.002$), where the probability of correctly classifying a pseudo-word faster than a random stimulus (pseudo-word or word) is 0.47, and the probability of correctly classifying a word faster than a random stimulus is 0.53. The Wilcoxon-Mann-Whitney test gives a the same significance as the Kruskal-Wallis test ($z = -3.04$, $p = 0.002$), but now we can state that the probability of correctly classifying a pseudo-word will be faster than correctly classifying a word is 0.44.

When we control for the possible learning effect, in other words, for the blocks in which the stimuli were presented, we get also significant difference, but now less than when we do not take the covariate into account ($\chi^2(1) = 9.50$, $p = 0.008$). The probability of correctly classyifing a pseudo-word faster than correctly classifying a word is 0.44 (the difference for the probability with the Wilcoxon-Mann-Whitney is the 3rd significant digit, so in this case not that impressive). The difference between the model with the covariate and the model without the covariate is not very large, and maybe the influence of the blocks is negligible. Figure **??** shows the estimated coefficients, together with their 95% confidence interval. Here we can see that the difference between the two models is negligible. Keuleers et al. (2010) found no significant evidence for a learning effect.

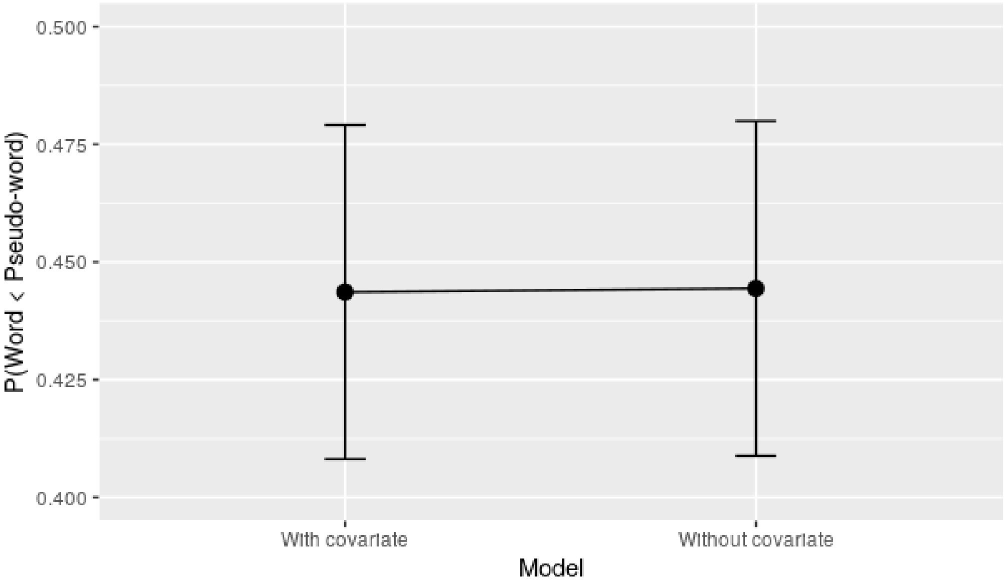

Figure 4.3: Estimated probability that the reaction time of lexical decision of a word is higher than the reaction time of lexical decision of a pseudo-word, including 95% confidence intervals.

4.3.2 Word Length

We now know there is a difference in how fast a word and a pseudo-word are correctly recognized and classified. Figure 4.2 shows a difference between the length of a word (in terms of syllables), but is this difference significant? This can be investigated with the Kruskal-Wallis test, as well with the generic rank test (Section 3.2). Again, the difference between these tests is the assumption concerning a common distribution.

The Kruskal-Wallis test gives a significant difference between the groups ($\chi^2(2) = 11.657$, $p = 0.0029$). The probability that a random pseudo-word is recognized slower than a random stimulus is 0.53 ($z = 3.120$, $p = 0.002$). The probability that a random 1-syllabic word is recognized slower than a random stimulus is 0.43 ($z = -2.304$, $p = 0.021$). We only see a marginal significance for the random 2-syllabic word ($z = -1.899$, $p = 0.058$). The generic rank test, without a covariate, also gives a significant difference between the groups ($\chi_2(3) = 20.167$, $p = 0.0002$, but now we can directly compare the groups. The probability to correctly classify a pseudo-word faster than a one-syllabic word is 0.405 ($z = -2.763$, $p = 0.006$), and to correctly classify it faster than a two-syllabic word is 0.451 ($z = -2.547$, $p = 0.011$). There is no significant difference between the reaction time for correctly classifying a one-syllabic word and a two-syllabic word ($z = 1.483$, $p = 0.138$).

When controlling for the block when the stimulus is presented also gives a significant difference between the groups ($\chi^2(3) = 21.464$, $p = 0.0003$). The pattern of significance is the same, and the probabilities are not very different. The estimates can be seen in Figure **??**, together with their 95% confidence intervals. As with lexicality, we can see that the difference here is also negligible. Again, Keuleers et al. (2010) also did not find any significant evidence for a learning effect.

4.4 Discussion

The results as found by Keuleers et al. (2010) are replicated, namely that words are recognized and classified faster as words than pseudo-words as pseudo-words. We observe a difference between the length of words, but this is only marginally significant. And finally, we have no evidence for a learning effect, in other words no interaction between the block and the lexicality/word length.

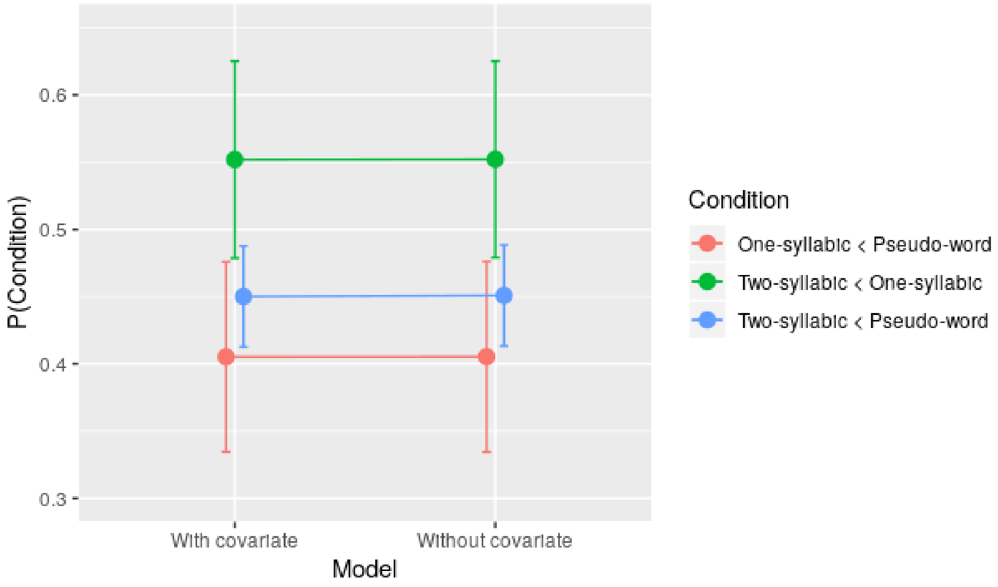

Figure 4.4: Estimated probability for the comparisons between pseudo-words, one-syllabic words, and two-syllabic words, including 95% confidence intervals.

CHAPTER 5

GENERAL DISCUSSION

Although the Probabilistic Index has been around for about 50 years, not necessarily under the same name, it was not until recently this was combined in a complete framework where regression models are incorporated (Thas et al., 2012; De Neve, 2013). Rewriting regression in terms of Probabilistic Index, and more specific using the Probablistic Index Model, has given opportunities to rewrite Rank Tests in order to achieve effect sizes, standard errors, and confidence intervals (De Neve and Thas, 2015). The big drawback was that this was not easily achieved using the PIM-package in R (Meys et al., 2017). In order to perform a Rank Test, using the PIM-framework, in-depth knowledge of the PIM theory on the one hand and the structure of the PIM-package in R on the other hand is needed. To overcome this problem, we have written a convenient R-code, using the standard regression notation. The structure for every formula is completely based on this standard regression notation, and when blocks or covariates are incorporated in a formula, we have used the notation as used in the Friedman-test as provided standard in R (R Core Team, 2018).

Next to the regression notation, we also have encoded the possibility to achieve confidence intervals with a given significance level (this is default 0.05), get only the coefficient(s) with the generic command `coef()`, receive a summary using the generic command `summary()`, and in the summary one can define if the default standard errors need to be shown or the Wald standard errors. When asking for the Wald standard errors, a warning will be shown.

Although we have added lots of flexibility to the tests, some functions still need to be developed. But before these functions are added, a research concerning the validity and reliability should be conducted. When comparing multiple levels, one may be interested in the comparison between two levels, which could be done using a *t*-test. But before this option is added, we still need to correct for multiple testing, but more important, it should be investigated if these tests are valid concerning large-sample approximation of the distribution.

A possible drawback is that effect sizes, when adding a covariate, will not become preciser but might have an inflated standard error (without boosting precision) (Vansteelandt, 2012). Although this could be a problem, the function for adding a covariate is added. Further research is needed to investigate the potential threat, and if this truly poses a problem corrections for the bias can be investigated.

In general, the main goal of this dissertation is to provide convenient code in order to use Rank Tests in the PIM-framework without having to study PIM in-depth. The interpretations of the provided effect sizes are straightforward and confidence intervals can be obtained. Still, caution is needed when using the functions, as with every statistical method. Future research and development can tackle some possible drawbacks and extensions can be designed.

BIBLIOGRAPHY

Acion, L., Peterson, J. J., Temple, S., and Arndt, S. (2006). Probabilistic index: an intuitive non-parametric approach to measuring the size of treatment effects. *Statistics in medicine*, 25(4):591–602.

Aho, K. (2019). *asbio: A Collection of Statistical Tools for Biologists*. R package version 1.5-5.

Browne, R. H. (2010). The t-test p value and its relationship to the effect size and P(X> Y). *The American Statistician*, 64(1):30–33.

De Neve, J. (2013). *Probabilistic Index Models*. Ghent University, Faculty of Sciences, Ghent, Belgium.

De Neve, J. and Thas, O. (2015). A regression framework for rank tests based on the probabilistic index model. *Journal of the American Statistical Association*, 110(511):1276–1283.

Enis, P. and Geisser, S. (1971). Estimation of the probability that Y < X. *Journal of the American Statistical Association*, 66(333):162–168.

Ferrand, L., New, B., Brysbaert, M., Keuleers, E., Bonin, P., Méot, A., Augustinova, M., and Pallier, C. (2010). The French Lexicon Project: Lexical decision data for 38,840 French words and 38,840 pseudowords. *Behavior Research Methods*, 42(2):488–496.

Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the american statistical association*, 32(200):675–701.

Grissom, R. J. (1994). Probability of the superior outcome of one treatment over another. *Journal of Applied Psychology*, 79(2):314.

Halperin, M., Gilbert, P. R., and Lachin, J. M. (1987). Distribution-free confidence intervals for Pr(X1< X2). *Biometrics*, pages 71–80.

Hollander, M., Wolfe, D. A., and Chicken, E. (2013). *Nonparametric statistical methods*, volume 751. John Wiley & Sons.

Hotelling, H. and Pabst, M. R. (1936). Rank correlation and tests of significance involving no assumption of normality. *The Annals of Mathematical Statistics*, 7(1):29–43.

Keuleers, E., Diependaele, K., and Brysbaert, M. (2010). Practice effects in large-scale visual word recognition studies: A lexical decision study on 14,000 Dutch mono-and disyllabic words and nonwords. *Frontiers in psychology*, 1:174.

Keuleers, E., Lacey, P., Rastle, K., and Brysbaert, M. (2012). The British Lexicon Project: Lexical decision data for 28,730 monosyllabic and disyllabic English words. *Behavior research methods*, 44(1):287–304.

Leemis, L. M. (1986). Relationships among common univariate distributions. *The American Statistician*, 40(2):143–146.

Lehmann, E. L. (1998). *Nonparametrics. statistical methods based on ranks*. Prentice Hall, Upper Saddle River, New Jersey, USA.

Leys, C., Ley, C., Klein, O., Bernard, P., and Licata, L. (2013). Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology*, 49(4):764–766.

Lin, M., Lucas Jr, H. C., and Shmueli, G. (2013). Research commentary—too big to fail: large samples and the p-value problem. *Information Systems Research*, 24(4):906–917.

Mack, G. A. and Skillings, J. H. (1980). A Friedman-type rank test for main effects in a two-factor ANOVA. *Journal of the American Statistical Association*, 75(372):947–951.

Mann, H. B. and Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, pages 50–60.

McGraw, K. O. and Wong, S. (1992). A common language effect size statistic. *Psychological bulletin*, 111(2):361.

Meys, J., De Neve, J., Sabbe, N., and Guimaraes de Castro Amorim, G. (2017). *pim: Fit Probabilistic Index Models*. R package version 2.0.1.

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Savage, I. R. (1953). Bibliography of nonparametric statistics and related topics. *Journal of the American Statistical Association*, 48(264):844–906.

Scholz, F. and Zhu, A. (2019). *kSamples: K-Sample Rank Tests and their Combinations*. R package version 1.2-9.

Senn, S. J. (1997). Testing for individual and population equivalence based on the proportion of similar responses. *Statistics in Medicine*, 16:1303–1306.

Seshan, V. E. (2018). *clinfun: Clinical Trial Design and Data Analysis Functions*. R package version 1.0.15.

Skillings, J. H. and Mack, G. A. (1981). On the use of a Friedman-type statistic in balanced and unbalanced block designs. *Technometrics*, 23(2):171–177.

Smirnov, N. V. (1939). On the estimation of the discrepancy between empirical curves of distribution for two independent samples. *Bull. Math. Univ. Moscou*, 2(2):3–14.

Srisuradetchai, P. (2015). *Skillings.Mack: The Skillings-Mack Test Statistic for Block Designs with Missing Observations*. R package version 1.10.

Thas, O., Neve, J. D., Clement, L., and Ottoy, J.-P. (2012). Probabilistic index models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(4):623–671.

Vansteelandt, S. (2012). Discussion of “Probabilistic index models” by O. Thas, J. De Neve, L. Clement and J.P. Ottoy. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(4):623–671.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics*, 1:80–83.