What is Wrong with `wilcox.test()` in R?

Torsten Hothorn
Universität Zürich

Kurt Hornik
Wirtschaftsuniversität Wien

Abstract

We discuss some conceptual issues with the location shift parameterisation underlying parameter and interval estimation in `wilcox.test()`, especially for bounded or ordinal response variables. Looking at other `*.test()` functions in package **stats**, we highlight established and missing links between models and procedures and discuss why diagnostic tools for model assessment and tools for power evaluation and sample size planning have been lacking from **stats** so far.

Reiterating the, unfortunately not widely known, fact that the Wilcoxon-Mann-Whitney Rank Sum statistic can be understood as the empirical score statistic in a proportional odds model, we motivate alternative parameterisations for the Wilcoxon and other linear rank tests where the location shift is not defined on the scale of the response but on some function of the distribution itself. Many known and novel models, parameters, and corresponding test and inference procedures arise, such as log-hazard ratios in proportional hazard models, Lehmann alternatives, or a generalised version of Cohen's d corresponding to the van der Waerden normal scores test.

A unified implementation for empirical maximum likelihood estimation and inference in this framework is provided through the novel `stats::free1way()` function added to the **stats** package in R 4.6.0, along with functionality for model diagnostics and power assessment. The application of these tools is described and illustrated by means of a series of novel analyses of data originating from a randomised trial evaluating neck pain reduction by laser therapy.

Keywords: Area under the ROC curve, Distribution-free oneway layout, Empirical likelihood, Probabilistic index, Proportional odds model, Stratification, Transformation model.

1. Introduction

The history of statistical methodology over the last two centuries can be told as a tale of powerful and ingeniously simple ideas being overshadowed by simplistic yet computable, and thus practically applicable, approximations. The replacement of Laplace' least absolute differences by Gauß' least squares buried median, or more generally, quantile and distribution regression under the normal linear regression tombstone for 150 years (Portnoy and Koenker 1997). Fisher's definition of the likelihood as a quantity being proportional to the *probability* of the data under a given model remains invisible when we introduce a log-likelihood as a log-density to students. Although the core ideas had been developed by Erich Lehmann, John Tukey, George Box, and Sir David Cox during the two decades preceding generalised linear models (GLMs), transformation models featuring data-driven response transformations in addition to interpretable covariate effects on latent scales never gained the same popularity

as GLMs, maybe except for the proportional hazards model and its replacement of the full by the partial likelihood. These three examples are just the tip of the iceberg, and some parts of contemporary methodological research can be understood as an attempt to excavate the origins.

This history still haunts us today, in research, teaching, and everyday practice, and its fallout is hard-wired into software, for example into the **stats** package (R Core Team 2026). Focusing on the suite of classical tests, we hope that diligent students first looking at the **stats** documentation will ask questions like "Why does `wilcox.test()` report p-values and parameter estimates when the latter are missing in `kruskal.test()` and `friedman.test()`?", "Why are odds ratios in `mantelhaen.test()` only available for stratified 2×2 tables?", "Why can't `wilcox.test()` handle ordered or censored outcomes?", "Why are blocks or strata not allowed in `wilcox.test()` or `kruskal.test()`?", or "Why is `power.wilcox.test()` missing?".

We provide a new look at many of the `*.test()` functions in the **stats** package through the lens of semiparametrically efficient population- and permutation-based inference in distribution-free stratified K-sample oneway layouts, whose core ideas are strongly connected to the transformation models mentioned above. The most prominent member of this model family is the Wilcoxon-Mann-Whitney two-sample rank sum test (Wilcoxon 1945; Mann and Whitney 1947, we use the term "Wilcoxon test" hereafter). We will concentrate on this test for the comparison of two parallel groups, developed independently by a number of researchers (Kruskal 1957), and its implementation in `wilcox.test()` as our prime example and will provide explanations for its limitations and introduce a more holistic and model-based understanding aiming to address the conceptual and practical questions in a framework including, but not being limited to, `wilcox.test()`. The new `free1way()` function, added to the **stats** package in R 4.6.0 published 2026-04-24, leverages modern computational tools for a general and reasonably fast implementation of a broad model family.

The Wilcoxon test is often advertised as an nonparametric alternative to the two-sample t-test. In this line of thinking the replacement of confidence intervals for the difference of means by confidence intervals for a location shift seems to make sense. In R, the Wilcoxon test is implemented in `wilcox.test()` with the option to compute confidence intervals for such a parameter using the procedure proposed by Bauer (1972) along with the Hodges-Lehmann estimate of a location shift. The authors of this article implemented these confidence intervals early in 2000, and the method became publically available in `wilcox.test()` as part of R 1.1.0, published 2000-06-15. The impact of the authors' choice to implement confidence intervals for a location shift in `wilcox.test()` is visible in the citation history of the original Bauer (1972) paper (Figure 1). All 23 citations before 2000 were from methodological papers in statistics journals. The method was only applied in empirical research in ecology, biochemistry, genetics, pharmacology, zoology and other fields after the implementation in `wilcox.test()` was published in R 1.1.0.

While the method and its implementation are technically correct, presenting confidence intervals for a location shift in the output of `wilcox.test()` seems suboptimal in hindsight, because the Wilcoxon test cannot be motivated as a test against shift alternatives (Conroy 2012; Divine *et al.* 2018; Fong and Huang 2019; Parker *et al.* 2020, and others). In fact, the Wilcoxon test is the nonparametric or empirical score test for a log-odds ratio in a semiparametric transformation model. Although the connection between the Wilcoxon test and proportional odds models has been discovered decades ago (Whitehead 1993) and is mentioned

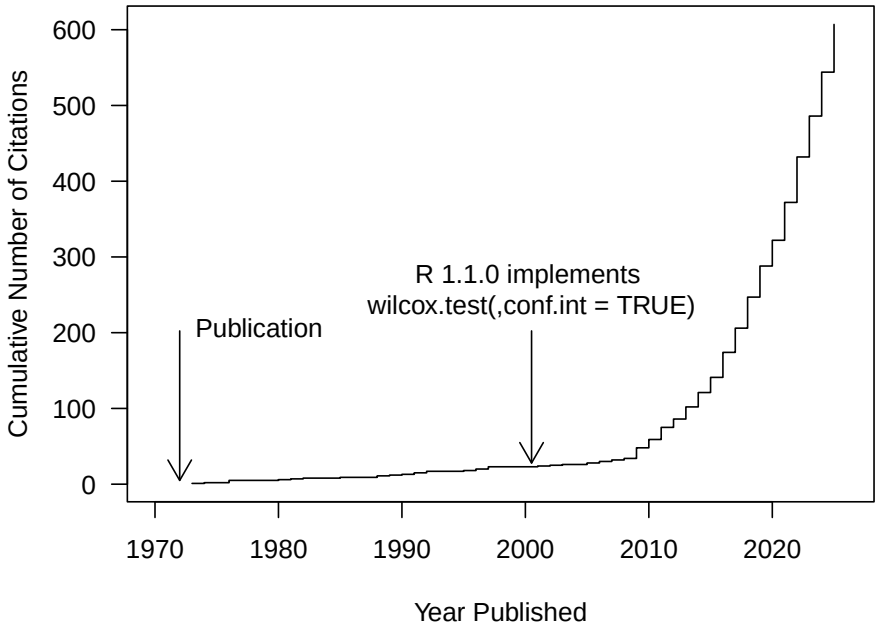


Figure 1: Cumulative number of citations of rank-based confidence intervals for a location shift (Bauer 1972) from 1972 to 2025. Data downloaded from Web of Science (`https://www.webofscience.com`) 2026-06-08.

in prominent textbooks (Harrell Jr 2015), the details for absolutely continuous outcomes have been worked out only more recently (Wang and Tian 2017; Liu *et al.* 2017). In this sense, the rank sum statistic proposed by Wilcoxon (1945) in absence of an underlying probabilistic model later turned out to coincide with the score statistic in the well-defined proportional odds model, first described 35 years later (by McCullagh 1980, but yet without making the connection). In this sense, the Wilcoxon test is not one of the simplistic approximations cited above but the real thing. However, with its implementation being overshadowed by a simplistic location shift model, `wilcox.test()` cannot leverage the full capacity of the proportional odds model for estimation and inference. With a fully specified matching probabilistic model in place, it makes sense to estimate the parameters of this model by optimising some appropriate likelihood function followed by likelihood-based inference procedures for the relevant parameters. The novel `free1way()` function does exactly this, in a model family not only featuring log-odds ratios and corresponding Wilcoxon tests, but also log-hazard ratios (Savage- or logrank tests), log-reverse time hazard ratios (tests against a Lehmann alternative, Lehmann 1953), and a variant of Cohen's d (van der Waerden test), also in stratified designs. In addition to parameter estimates, tests, and confidence intervals in `free1way()`, tools for model diagnostics by means of probability-probability plots and for sample size planning became available in R 4.6.0.

We proceed by introducing randomised clinical trial data we will use as a motivation and

for illustrating concepts and user interfaces throughout this article. Starting with the special case of proportional odds models, and thus Wilcoxon tests, we explain how the model is parameterised and the core concepts of maximum likelihood inference in this special case. Following a reanalysis of the clinical trial, we generalise the framework to other treatment effects covered by `free1way()`.

2. Low-level Laser Therapy Against Neck Pain Study

Chow *et al.* (2006) reported on a double-blind, randomised, placebo-controlled superiority study evaluating the effect of 300 mW, 830 nm laser on chronic neck pain for a total of $N = 90$ subjects at three time points. Pain was measured using a horizontal 10cm Visual Analogue Scale (VAS, mapped to the unit interval) before treatment start ("baseline"), directly after 14 treatments applied over the course of seven weeks ("7 weeks"), and four weeks after end of treatment ("12 weeks"). The VAS trajectories of all subjects are given in Figure 2 and the marginal distributions of VAS at the three time points are presented in Figure 3, separately for the two treatment arms (placebo and active laser treatment). The left panel in this figure highlights quite substantial baseline differences and Table 1 indicates missing values for the VAS measurements directly at the end of treatment.

	time		
laser	Baseline	7 weeks	12 weeks
Placebo	45	41	45
Active	45	43	45

Table 1: Low-level laser therapy against chronic neck pain (Chow *et al.* 2006) data: Sample sizes in two treatment groups over time.

In their "Statistical Methods" section, Chow *et al.* (2006) first reflect on whether or not VAS should be treated as ordinal or as being interval-scaled. Techniques under both paradigms were applied, first comparing change scores (VAS at 12 weeks minus VAS at baseline) using the two-sample t-test and comparing an ordinal assessment of change in pain severity using a proportional odds model. Down the road, the methodological issues arising in this trial motivated the development of ordinal regression models for continuous scales (Manuguerra and Heller 2010) by one of the trial statisticians.

In addition to the analyses presented by Chow *et al.* (2006), one could have asked the question "Are the two marginal distributions of VAS four weeks after end of treatment (third panel in Figure 3) identical?". Because of the boundedness and lack of symmetry, at least in the active laser group, of VAS recordings, many data analysts may opt for a "nonparametric" approach, and thus `wilcox.test()` may seem an attractive choice. Running the following code in R 4.6.0 or later results in

```
R> (w <- wilcox.test(vas ~ laser, data = pain, subset = time == "12 weeks",
+                    exact = TRUE, conf.int = TRUE))

        Wilcoxon rank sum exact test
```

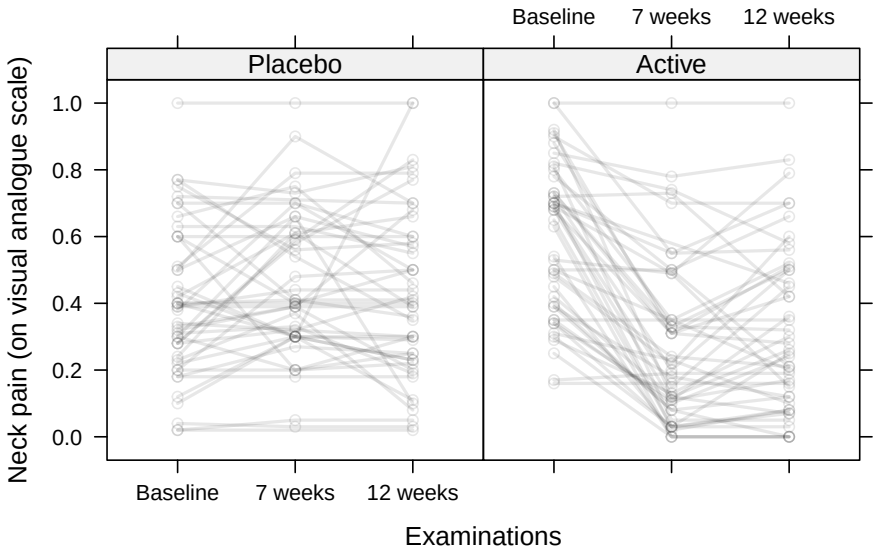


Figure 2: Low-level laser therapy against chronic neck pain (Chow *et al.* 2006) data: Joint distribution of pain measured on Visual Analogue Scale for two treatment groups for three time points.

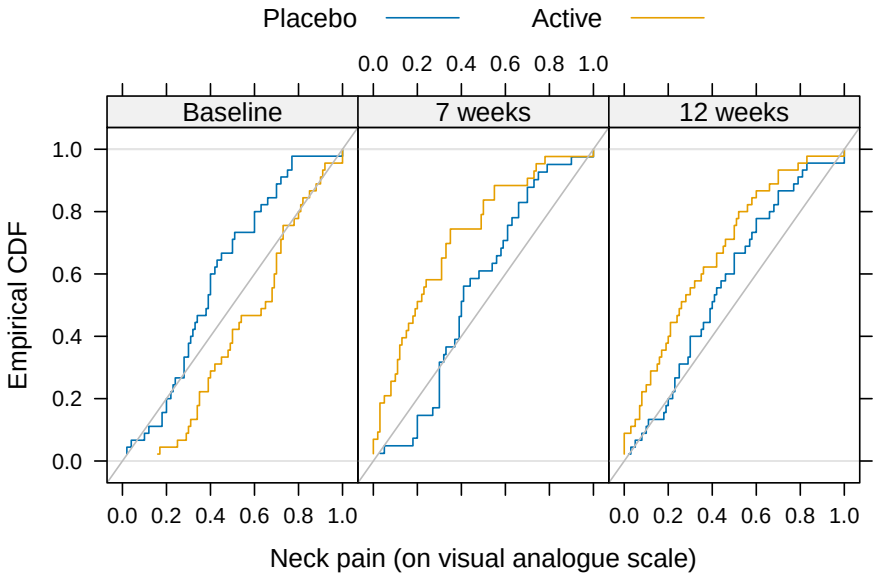


Figure 3: Low-level laser therapy against chronic neck pain (Chow *et al.* 2006) data: Marginal distributions of pain measured on Visual Analogue Scale for two treatment groups for three time points.

```
data:  vas by laser
W = 1260, p-value = 0.046
```

```
alternative hypothesis: true location shift is not equal to 0
96 percent confidence interval:
 0.00 0.22
sample estimates:
difference in location
                  0.11
```

Prior to version R 4.6.0, `wilcox.test()` would have refused to report an exact p-value citing the presence of 40 tied response values. Even without such a warning, the results are puzzling: We picked `wilcox.test()` because of its "nonparametric" nature, so where does the location shift parameter come from? Was VAS treated as an interval-scaled variable? Where do then the ties come from and how do they impact model and estimation? The documentation in `help("wilcox.test")` provides some clues: With $\mathbb{P}(\text{VAS} \leq y \mid \text{Placebo}) = F_{\text{VAS}}(y)$ the marginal distribution of VAS in the placebo group, the results suggest a marginal distribution of VAS in those treated with an active laser $F_{\text{VAS}}(y - \beta)$, with $\hat{\beta} = 0.110$ and a 95.5% confidence interval $(0.000, 0.220)$. The distribution function F_{VAS} is not specified, so the procedure comes with a distribution-free flavour. However, the model assumes that VAS is a continuous variable, because for ordinal variables the expression $F_{\text{VAS}}(y - \beta)$ is ill-defined. Furthermore, the bounds of VAS are not respected, because *e.g.,* by definition $F_{\text{VAS}}(y - 0.22) = 0$ for all $y < 0.22$. One could try to work around this problem by analysing a transformation of VAS, maybe a probit transformation $\Phi^{-1}(\text{VAS})$, but the location shift on this scale is not directly interpretable. Answers regarding the measurement scale of the response, however, are much harder to find, mainly due to a lack of a model motivating the test statistic.

This gap was filled by Wang and Tian (2017), who proved that the Wilcoxon test is a score test in a proportional odds model where the cumulative distribution functions in the two treatment groups are parameterised as

$$\begin{aligned} \mathbb{P}(\text{VAS} \leq y \mid \text{Placebo}) &= \text{expit}(h(y)) \\ \mathbb{P}(\text{VAS} \leq y \mid \text{Active}) &= \text{expit}(h(y) - \delta) \end{aligned}$$

with $\text{expit}(z) = (1 + \exp(-z))^{-1}$ denoting the distribution function of the standard logistic distribution. Consequently, Wang and Tian (2017) suggest to estimate log-odds ratios by means of this model and to provide corresponding confidence intervals. Before we present the theoretical arguments, we follow their advice and analyse the data by means of a Wald test and confidence interval obtained from a proportional odds model. `MASS::polr()` (Ripley and Venables 2002) implements this model, however, only for ordered categorical response variables. Simply treating the observed VAS values as ordered categories leads to

```
R> w12 <- subset(pain, time == "12 weeks")
R> w12$vasE <- ordered(w12$vas)
R> p <- polr(vasE ~ laser, data = w12, Hess = TRUE)
R> summary(p)$coefficients["laserActive",]

     Value Std. Error    t value
-0.746953   0.372347  -2.006068

R> confint(p)
```

```
     2.5 %     97.5 %
-1.4849466 -0.0221237
```

with log-odds ratio $\hat{\delta} = -0.747$ indicating that pain was lower under active laser treatment because the negative shift associates negative values of δ with smaller VAS values in the treated group. This log-odds ratio effect is invariant with respect to monotone transformations of the response. Unfortunately, the Wald p-value is not identical to the p-value reported by `wilcox.test()`, not even when we use an asymptotic p-value instead of an exact one

```
R> wilcox.test(vas ~ laser, data = pain, subset = time == "12 weeks",
+              exact = FALSE)$p.value

[1] 0.0466026
```

In addition, the tests' p-value and the Wald confidence intervals are not congruent

```
R> confint(p, level = 1 - w$p.value)

      2.3 %      97.7 %
-1.49840001 -0.00914961
```

and it can happen that `wilcox.test()` suggests to reject equality of the distribution functions at nominal level α yet a two-sided $(1 - \alpha) \times 100\%$ Wald confidence interval includes zero or vice versa.

3. Theory

We first review the relevant theory of score tests (also known as Rao score and Lagrange Multiplier tests) and corresponding confidence intervals in Section 3.1. The Wilcoxon test is then embedded in this theory as the nonparametric or empirical score test in a proportional odds model in Section 3.2 and we apply this result to the construction of confidence intervals in Section 3.3. We generalise the framework to stratified K-sample oneway layouts with $K > 2$ and describe the main ideas underlying semiparametrically efficient population and permutation inference in this model class in Section 3.5.

3.1. Score Tests and Intervals

Consider a single parameter of interest $\delta \in \mathbb{R}$ and other parameters $\boldsymbol{\vartheta} \in \mathbb{R}^P$ and N independent observations with log-likelihood contributions $\ell_i : \mathbb{R}^{1+P} \to \mathbb{R}$. The log-likelihood is given by $\ell(\delta, \boldsymbol{\vartheta}) = \sum_{i=1}^N \ell_i(\delta, \boldsymbol{\vartheta})$ and its score function by

$$\boldsymbol{S}(\delta, \boldsymbol{\vartheta}) = \frac{\partial \ell(\delta, \boldsymbol{\vartheta})}{\partial (\delta, \boldsymbol{\vartheta}^\top)^\top} = (S_\delta(\delta, \boldsymbol{\vartheta}), \boldsymbol{S}_{\boldsymbol{\vartheta}}(\delta, \boldsymbol{\vartheta})^\top)^\top \in \mathbb{R}^{1+P}.$$

The first element of the gradient is the score with respect to the parameter of interest δ. The observed Fisher information matrix for all parameters $(\delta, \boldsymbol{\vartheta}^\top)^\top$ is

$$\boldsymbol{I}(\delta, \boldsymbol{\vartheta}) = \frac{\partial^2 \ell(\delta, \boldsymbol{\vartheta})}{\partial^2 (\delta, \boldsymbol{\vartheta}^\top)^\top} = \begin{pmatrix} I(\delta, \boldsymbol{\vartheta})_{\delta,\delta} & \boldsymbol{I}(\delta, \boldsymbol{\vartheta})_{\delta,\boldsymbol{\vartheta}} \\ \boldsymbol{I}(\delta, \boldsymbol{\vartheta})_{\delta,\boldsymbol{\vartheta}}^\top & \boldsymbol{I}(\delta, \boldsymbol{\vartheta})_{\boldsymbol{\vartheta},\boldsymbol{\vartheta}} \end{pmatrix} \in \mathbb{R}^{1+P,1+P}$$

with inverse

$$\boldsymbol{A}(\delta, \boldsymbol{\vartheta}) = \boldsymbol{I}(\delta, \boldsymbol{\vartheta})^{-1} = \begin{pmatrix} A(\delta, \boldsymbol{\vartheta})_{\delta,\delta} & \boldsymbol{A}(\delta, \boldsymbol{\vartheta})_{\delta,\boldsymbol{\vartheta}} \\ \boldsymbol{A}(\delta, \boldsymbol{\vartheta})_{\delta,\boldsymbol{\vartheta}}^\top & \boldsymbol{A}(\delta, \boldsymbol{\vartheta})_{\boldsymbol{\vartheta},\boldsymbol{\vartheta}} \end{pmatrix} \in \mathbb{R}^{1+P,1+P}.$$

Both matrices are partitioned such that the first row and column position corresponds to the parameter of interest δ. Under the composite null hypothesis $H_0 : \delta = \delta_0$ the "nuisance" parameters $\boldsymbol{\vartheta}$ are unknown and the restricted maximum likelihood estimator

$$\hat{\boldsymbol{\vartheta}}(\delta_0) = \arg\max_{\boldsymbol{\vartheta} \in \mathbb{R}^P} \ell(\delta_0, \boldsymbol{\vartheta})$$

satisfies $\boldsymbol{S}_{\boldsymbol{\vartheta}}(\delta_0, \hat{\boldsymbol{\vartheta}}(\delta_0)) = \mathbf{0}$. The quadratic score statistic simplifies to

$$\begin{aligned} &\boldsymbol{S}(\delta_0, \hat{\boldsymbol{\vartheta}}(\delta_0))^\top \boldsymbol{I}(\delta_0, \hat{\boldsymbol{\vartheta}}(\delta_0))^{-1} \boldsymbol{S}(\delta_0, \hat{\boldsymbol{\vartheta}}(\delta_0)) \\ &\quad = (S_\delta(\delta_0, \hat{\boldsymbol{\vartheta}}(\delta_0)), \mathbf{0}^\top) \boldsymbol{I}(\delta_0, \hat{\boldsymbol{\vartheta}}(\delta_0))^{-1} (S_\delta(\delta_0, \hat{\boldsymbol{\vartheta}}(\delta_0)), \mathbf{0}^\top)^\top \\ &\quad = (\boldsymbol{I}(\delta_0, \hat{\boldsymbol{\vartheta}}(\delta_0))^{-1})_{11} S_\delta(\delta_0, \hat{\boldsymbol{\vartheta}}(\delta_0))^2 \\ &\quad = A(\delta_0, \hat{\boldsymbol{\vartheta}}(\delta_0))_{\delta,\delta} S_\delta(\delta_0, \hat{\boldsymbol{\vartheta}}(\delta_0))^2 \end{aligned}$$

as explained for example in Rao (2005, Formula 1.18), Lehmann (1998, pages 539–541), Pawitan (2001, example 8.11), or Fahrmeir and Tutz (2001, Formula 4.1.2), and we define the linear score statistic for $H_0 : \delta = \delta_0$ as

$$S(\delta_0) := A(\delta_0, \hat{\boldsymbol{\vartheta}}(\delta_0))_{\delta,\delta}^{1/2} S_\delta(\delta_0, \hat{\boldsymbol{\vartheta}}(\delta_0)).$$

The ancillary statistic $S(\delta_0)$ can be used to construct $(1 - 2\alpha) \times 100\%$ score intervals $\{\delta_0 \in \mathbb{R} \mid Q(\alpha) < S(\delta_0) \leq Q(1-\alpha)\}$ where the quantiles are such that $\mathbb{P}_\delta(S(\delta) \leq Q(\alpha)) \leq \alpha$ and $\mathbb{P}_\delta(S(\delta) > Q(1-\alpha)) \geq 1 - \alpha$ for some $0 < \alpha < 1/2$ and δ denoting the true unknown parameter of the data generating process. Asymptotically and under suitable regularity conditions, $S(\delta)/\sqrt{N} \xrightarrow{d} \mathrm{N}(0, 1)$ and normal quantiles $Q(\alpha) = \sqrt{N}\Phi^{-1}(\alpha)$ can be used to compute score confidence intervals. Alternatively, in the K-sample situation we are concerned with, under the null hypothesis $H_0 : \delta = 0$ stating equivalence of outcome distributions, the term $A(0, \hat{\boldsymbol{\vartheta}}(0))_{\delta,\delta}$ is invariant with respect to permutations, and thus the exact permutation distribution of $S_\delta(0, \hat{\boldsymbol{\vartheta}}(0))$, or a Monte-Carlo approximation thereof, can be used to define $Q(\alpha)$. Thus, exact or approximate conditional tests and confidence intervals based on quantiles of the permutation distribution are available to complement the unconditional asymptotic variants.

Score intervals are invariant with respect to monotone transformations and computationally attractive because only the distribution of $S(0)$ under $\delta_0 = 0$ needs to be computed. However, computation of $S(\delta_0)$ for $\delta_0 \neq 0$ requires updating the maximum likelihood estimators $\hat{\boldsymbol{\vartheta}}(\delta_0)$.

An equivalent formulation of score tests and confidence intervals for δ via the profile log-likelihood

$$\tilde{\ell}(\delta) = \sum_{i=1}^N \ell_i(\delta, \hat{\boldsymbol{\vartheta}}(\delta))$$

is theoretically and computationally attractive. In the semiparametric models outlined in Section 3.5, the maximum likelihood estimator $\hat{\delta}$ is asymptotically normal and efficient (Murphy and van der Vaart 2000) and, because the log-likelihood ℓ and thus the profile log-likelihood

$\tilde{\ell}$ are concave functions of δ, confidence regions are convex sets (Lemma A.1., Murphy and van der Vaart 1997). For scalar parameters δ, these results ensure that we can compute congruent score or likelihood-ratio intervals by inversion of the corresponding tests, that is, $(1-2\alpha)\times 100\%$ confidence intervals will exclude δ_0 iff the hypothesis $H_0 : \delta = \delta_0$ is rejected at level 2α. In Appendix D, we discuss situations where the log-likelihood is very flat and thus the computation of parameter estimates and corresponding confidence intervals fails, despite these theoretical results.

3.2. The Wilcoxon Score Test

We will now discuss a model, featuring nuisance parameters $\boldsymbol{\vartheta}$ in addition to a certain parameter of interest δ, whose likelihood function gives rise to a score statistic $S(\delta_0)$ which, under the null hypothesis $\delta = \delta_0 = 0$, is a linear function of the Wilcoxon Mann-Whitney two-sample rank sum statistic. Based on this model, we will explain how to interpret the parameter δ, how to estimate it using maximum likelihood, and how to obtain confidence intervals. Furthermore, the model allows to identify situations where the test will have high power, and situations where we cannot expect the test to work well, demystifying the notion of its "nonparametric" nature. The model will serve as a blueprint for the development of more general models for $K > 2$ samples in stratified designs in Section 3.5.

For a univariate at least ordered response $Y \in \Xi$ and $X \in \mathbb{R}$ a numeric covariate, with $X \in \{0, 1\}$ being an appropriate contrast in the two-sample situation, we study the proportional odds model for the conditional distribution of Y given $X = x$

$$\mathbb{P}(Y \leq y \mid X = x) = \text{expit}(h(y \mid \boldsymbol{\vartheta}) - \delta x) =: F_Y(y \mid x). \tag{1}$$

The model features a logit-link ($\text{expit} = \text{logit}^{-1}$) and $h(\cdot \mid \boldsymbol{\vartheta}) : \Xi \to \mathbb{R}$, an unknown monotonically increasing transformation function depending on parameters $\boldsymbol{\vartheta} \in \mathbb{R}^P$, and defines the effect of a one-unit increase in x by the log-odds ratio δ such that

$$\frac{\mathbb{P}(Y \leq y \mid X = x + 1)}{\mathbb{P}(Y > y \mid X = x + 1)} = \exp(-\delta) \frac{\mathbb{P}(Y \leq y \mid X = x)}{\mathbb{P}(Y > y \mid X = x)} \quad \forall y \in \Xi.$$

In the two-sample case, the model states that the odds of observing a value of at most y in the second sample ($x = 1$) is $\exp(-\delta)$ times the odds of observing a value of at most y in the first sample ($x = 0$). The proportionality assumption ensures that the factor $\exp(-\delta)$ applies to all possible cut-off values y. Equivalently, one can think of a series of binary logistic regression models with different cut-off points y for dichotomising the response $Y \leq y$. The corresponding intercepts increase as y increases but the log-odds ratio $-\delta$ remains constant. It is well known (Prentice 1978) that the Wilcoxon test is the locally most powerful rank test against shift-alternatives in the logistic model $\mathbb{P}(Y \leq y \mid X = x) = \text{expit}(y - \delta x)$. The shift parameter δ is equivalent to a log-odds ratio in this logistic model. For absolutely continuous responses, model (1) applies a transformation $h(Y \mid \boldsymbol{\vartheta})$ such that $h(Y \mid \boldsymbol{\vartheta}) - \delta x$ follows a standard logistic distribution. The shift parameter δ is no longer interpretable as a difference on the original measurement scale of Y but the interpretation as log-odds ratio remains valid.

In the binary logistic regression point of view, the parameters $\boldsymbol{\vartheta}$ of the monotonically increasing transformation function h are the increasing intercepts. Testing $H_0 : \delta = 0$ in the two-sample situation is equivalent to the null $Y \perp\!\!\!\perp X$ or $F_Y(y \mid 0) = F_Y(y \mid 1)$ for all $y \in \Xi$. Our arguments rely on the fact that the Wilcoxon test is the nonparametric or empirical score

test for δ_0 in the proportional odds model (1) because the score statistic $S(0)$ is permutation-equivalent to the sum of the ranks in the second group ($x_i = 1$). Wang and Tian (2017) demonstrate that the score function $S_\delta(0, \hat{\boldsymbol{\vartheta}}(0))$ is a linear function of the Wilcoxon statistic in the two-sample case, similar arguments are given in the Appendix to Liu *et al.* (2017).

To derive the score statistic $S(0)$ for model (1), consider a sample of N independent observations $(y_i, x_i), i = 1, \ldots, N$ with $C \leq N$ distinct ordered outcomes $\upsilon_1 < \cdots < \upsilon_C$ in $y_1, \ldots, y_N$. The ith observation contributes

$$\ell_i(\delta, \boldsymbol{\vartheta}) = \log(\text{expit}(h(y_i \mid \boldsymbol{\vartheta}) - \delta x_i) - \text{expit}(h(y_i- \mid \boldsymbol{\vartheta}) - \delta x_i)) \tag{2}$$

to the nonparametric or empirical log-likelihood function (Owen 2001) derived from model (1). Here, y_i- is the largest distinct ordered outcome smaller than y_i. The transformation function h is parameterised nonparametrically through parameters $\boldsymbol{\vartheta} \in \mathbb{R}^{C-1}$, more specifically through $\vartheta_0 = -\infty < \vartheta_1 < \cdots < \vartheta_C = \infty$, as

$$h(\upsilon_c \mid \boldsymbol{\vartheta}) = \vartheta_c \text{ and } h(\upsilon_c- \mid \boldsymbol{\vartheta}) = \vartheta_{c-1},$$

i.e., one parameter is assigned to each but the largest of the distinct ordered outcomes.

Under $H_0 : \delta = 0$, we assign upranks $R_1, \ldots, R_N$ to $y_1, \ldots, y_N$, with R_i being the number of observations not larger than y_i. The marginal distribution function is estimated by the nonparametric or empirical maximum likelihood estimator $\hat{\mathbb{P}}(Y \leq y_i) = {}^{R_i}/_N$, *i.e.,* we apply the empirical cumulative distribution function. This estimator is the restricted nonparametric maximum likelihood estimator under $\delta = 0$. We obtain $h(y_i \mid \hat{\boldsymbol{\vartheta}}(0)) = \text{logit}({}^{R_i}/_N)$ and the score statistic is then

$$\begin{aligned} S(0) &= \sqrt{A(0, \hat{\boldsymbol{\vartheta}}(0))_{\delta,\delta}} \sum_{i=1}^{N} \left. \frac{\partial \log(\text{expit}(h(y_i \mid \hat{\boldsymbol{\vartheta}}(0)) - \delta x_i) - \text{expit}(h(y_i- \mid \hat{\boldsymbol{\vartheta}}(0)) - \delta x_i))}{\partial \delta} \right|_{\delta=0} \\ &= \sqrt{A(0, \hat{\boldsymbol{\vartheta}}(0))_{\delta,\delta}} \sum_{i=1}^{N} \frac{\text{expit}'(h(y_i \mid \hat{\boldsymbol{\vartheta}}(0)) - 0x_i) - \text{expit}'(h(y_i- \mid \hat{\boldsymbol{\vartheta}}(0)) - 0x_i)}{\text{expit}(h(y_i \mid \hat{\boldsymbol{\vartheta}}(0)) - 0x_i) - \text{expit}(h(y_i- \mid \hat{\boldsymbol{\vartheta}}(0)) - 0x_i)} x_i. \end{aligned}$$

With $\text{expit}'(z) = \text{expit}(z)(1 - \text{expit}(z))$, $\text{expit}(h(y_i \mid \hat{\boldsymbol{\vartheta}}(0)) - 0x_i) = {}^{R_i}/_N$ and $\text{expit}(h(y_i- \mid \hat{\boldsymbol{\vartheta}}(0)) - 0x_i) = {}^{R_i - t_i}/_N$ (t_i being the number of ties equal to y_i) this simplifies to

$$S(0) = \sqrt{A(0, \hat{\boldsymbol{\vartheta}}(0))_{\delta,\delta}} \sum_{i=1}^{N} \frac{N - 2R_i + t_i}{N} x_i.$$

For the two-sample case ($x_i \in \{0, 1\}$) with $N_1 := \sum_{i=1}^N x_i$ and $N_0 := N - N_1$, we obtain

$$\begin{aligned}
\frac{N}{2}\sqrt{A(0, \hat{\vartheta}(0))_{\delta,\delta}}^{-1} S(0) &= \frac{N}{2} \sum_{i:x_i=1} \left(1 - \frac{2}{N}\left(R_i - \frac{t_i}{2}\right)\right) \\
&= \frac{N}{2}\left(N_1 - \frac{2}{N_0 + N_1} \sum_{i:x_i=1} \left(R_i - \frac{t_i - 1}{2} - \frac{1}{2}\right)\right) \\
&= \frac{N_1}{2} + \frac{N_1(N_0 + N_1)}{2} - \sum_{i:x_i=1} \left(R_i - \frac{t_i - 1}{2}\right) \\
&= \underbrace{\frac{N_0 N_1}{2} + \underbrace{\frac{N_1(N_1 + 1)}{2} - \underbrace{\sum_{i:x_i=1} \left(R_i - \frac{t_i - 1}{2}\right)}_{=W}}_{=-U}}_{=-(U - \mathbb{E}(U))}
\end{aligned}$$

where W is the Wilcoxon statistic, *i.e.,* the sum of the mid-ranks $R_i - (t_i - 1)/2$ in the sample with $x_i = 1$, and $U = W - N_1(N_1 + 1)/2$ is the corresponding Mann-Whitney U-statistic with $\mathbb{E}(U) = N_0 N_1/2$ under H_0. The latter statistic, named `W`, is returned by `wilcox.test()`, with $x_i = 1$ whenever observation i belongs to the first group in terms of the ordering of the levels of the corresponding factor. In contrast, `free1way()` reports the standardised statistic $S(0)$ named `Z` based on $x_i = 1$ when observation i belongs to the second group (corresponding to the default treatment contrasts in most models in R).

This relationship between the score statistic $S(0)$ of a proportional odds model for at least ordered responses and the Wilcoxon statistic W in the two-sample situation is significant from a practical as well as a theoretical point of view. Practically, we can use this result for the construction of score intervals for δ (see Section 3.3). Theoretically, we see that there is no difference in the treatment of binary ($C = 2$), ordinal ($C < N$), and continuous ($C = N$ in the absence and $C < N$ in the presence of ties) responses and, in the latter case, ties do not require special treatment. Shift effects on the original response scale are not present in this model and thus is seems suboptimal for `wilcox.test()` to report corresponding estimates and confidence intervals. However, it is perfectly appropriate to report the maximum likelihood estimator $\hat{\delta}$, interpretable as log-odds ratio, and a corresponding confidence interval in addition to a corresponding p-value; we explain details in Section 3.3. Furthermore, we can apply the arguments presented by van der Vaart (1998, Chapter 15.5) to conclude that the Wilcoxon test is the locally most powerful rank test against proportional odds alternatives for any baseline distribution $\mathbb{P}(Y \leq y \mid X = 0)$ in the very same way as the log-rank (or Savage) test is the locally most powerful rank test against proportional hazards alternatives, see van der Vaart (1998, Example 15.16). In this sense, the Wilcoxon test is distribution-free with maximal, and even optimal, power against proportional odds alternatives. At the same time, it also implies that the Wilcoxon test is inefficient in other situations, such as against hazard ratio alternatives or in cases where the distribution functions cross. Thus, tools for diagnostic assessment of the proportional odds assumption are important.

The unconditional exact or asymptotic score test based on this statistic is commonly applied for computing p-values against the null $H_0 : \delta = 0$ and `wilcox.test()` implements this test.

Asymptotic versions rely on an approximation of the variance of W and exact versions compute the permutation distribution of W, including in the absence of ties where conditioning on the observed mid-ranks leads to conditional or permutation score tests for W, with exact and asymptotic variants now (since R 4.6.0) being available in `wilcox.test()`. The conditional Wilcoxon test is especially attractive for the analysis of ordered categorical responses: The likelihood (2) is the likelihood of a proportional odds model for categorical data and the Wilcoxon statistic based on mid-ranks is proportional to the corresponding score statistic. It seems therefore overly restrictive for `wilcox.test()` to not allow ordered categorical responses.

3.3. Nonparametric or Empirical Score Intervals for Log-odds Ratios

To derive unconditional and conditional $(1 - 2\alpha) \times 100\%$ score confidence intervals for the log-odds ratio δ, we obtain corresponding quantiles $Q(\alpha)$ and $Q(1 - \alpha)$ for $S(0)$. The lower confidence limit is given by the root of $S(\delta_0) - Q(\alpha)$ and the upper limit by the root of $S(\delta_0) - Q(1 - \alpha)$, where

$$S(\delta_0) = \sqrt{A(\delta_0, \hat{\boldsymbol{\vartheta}}(\delta_0))_{\delta,\delta}} \sum_{i=1}^{N} \frac{\text{expit}'(h(y_i \mid \hat{\boldsymbol{\vartheta}}(\delta_0)) - \delta_0 x_i) - \text{expit}'(h(y_i- \mid \hat{\boldsymbol{\vartheta}}(\delta_0)) - \delta_0 x_i)}{\text{expit}(h(y_i \mid \hat{\boldsymbol{\vartheta}}(\delta_0)) - \delta_0 x_i) - \text{expit}(h(y_i- \mid \hat{\boldsymbol{\vartheta}}(\delta_0)) - \delta_0 x_i)} x_i.$$

Evaluating the score function requires computation of the nonparametric maximum likelihood estimator $\hat{\boldsymbol{\vartheta}}(\delta_0)$ and corresponding inverse Fisher information $A(\delta_0, \hat{\boldsymbol{\vartheta}}(\delta_0))_{\delta,\delta}$ under $H_0 : \delta = \delta_0$. With the rows of the design matrices of the proportional odds model being $\bar{\boldsymbol{a}}_i = (x_i, e_{C+1}(y_i)^\top)^\top$ and $\underline{\boldsymbol{a}}_i = (x_i, e_{C+1}(y_i-)^\top)^\top$, where $e_{C+1}(\upsilon_p)$ and $e_{C+1}(\upsilon_p-)$ are unit vectors of length $C + 1$, and parameters $\boldsymbol{\gamma} = (\delta_0, \vartheta_0, \ldots, \vartheta_C)^\top$ (remember $\vartheta_0 = -\infty$ and $\vartheta_C = \infty$, so there are $1 + C - 1 = P + 1$ unknown parameters) such that $\bar{\boldsymbol{a}}_i^\top \boldsymbol{\gamma} = x_i\delta_0 + \upsilon_p$ and $\underline{\boldsymbol{a}}_i^\top \boldsymbol{\gamma} = x_i\delta_0 + \upsilon_{p-1}$ if $y_i = \upsilon_p$, we obtain the Fisher matrix as (cf. Formula (4) in Hothorn *et al.* 2018)

$$\begin{aligned}
\boldsymbol{I}(\delta_0, \boldsymbol{\vartheta}) &= \begin{pmatrix} I(\delta_0, \boldsymbol{\vartheta})_{\delta,\delta} & \boldsymbol{I}(\delta_0, \boldsymbol{\vartheta})_{\delta,\boldsymbol{\vartheta}} \\ \boldsymbol{I}(\delta_0, \boldsymbol{\vartheta})_{\delta,\boldsymbol{\vartheta}}^\top & \boldsymbol{I}(\delta_0, \boldsymbol{\vartheta})_{\boldsymbol{\vartheta},\boldsymbol{\vartheta}} \end{pmatrix} \\
&= \sum_{i=1}^{N} - \left\{ \frac{\text{expit}''(\bar{\boldsymbol{a}}_i^\top \boldsymbol{\gamma}) \bar{\boldsymbol{a}}_i \bar{\boldsymbol{a}}_i^\top - \text{expit}''(\underline{\boldsymbol{a}}_i^\top \boldsymbol{\gamma}) \underline{\boldsymbol{a}}_i \underline{\boldsymbol{a}}_i^\top}{\text{expit}(\bar{\boldsymbol{a}}_i^\top \boldsymbol{\gamma}) - \text{expit}(\underline{\boldsymbol{a}}_i^\top \boldsymbol{\gamma})} \right. \\
&\qquad \left. - \frac{[\text{expit}'(\bar{\boldsymbol{a}}_i^\top \boldsymbol{\gamma}) \bar{\boldsymbol{a}}_i - \text{expit}'(\underline{\boldsymbol{a}}_i^\top \boldsymbol{\gamma}) \underline{\boldsymbol{a}}_i]}{[\text{expit}(\bar{\boldsymbol{a}}_i^\top \boldsymbol{\gamma}) - \text{expit}(\underline{\boldsymbol{a}}_i^\top \boldsymbol{\gamma})]^2} \times [\text{expit}'(\bar{\boldsymbol{a}}_i^\top \boldsymbol{\gamma}) \bar{\boldsymbol{a}}_i^\top - \text{expit}'(\underline{\boldsymbol{a}}_i^\top \boldsymbol{\gamma}) \underline{\boldsymbol{a}}_i^\top] \right\}.
\end{aligned}$$

The Fisher information $A(\delta_0, \hat{\boldsymbol{\vartheta}}(\delta_0))$ for δ_0 for unknown $\boldsymbol{\vartheta}$ is then given by the inverse of the Schur complement $I(\delta_0, \hat{\boldsymbol{\vartheta}}(\delta_0))_{\delta,\delta} - \boldsymbol{I}(\delta_0, \hat{\boldsymbol{\vartheta}}(\delta_0))_{\delta,\boldsymbol{\vartheta}} \boldsymbol{I}(\delta_0, \hat{\boldsymbol{\vartheta}}(\delta_0))_{\boldsymbol{\vartheta},\boldsymbol{\vartheta}}^{-1} \boldsymbol{I}(\delta_0, \hat{\boldsymbol{\vartheta}}(\delta_0))_{\delta,\boldsymbol{\vartheta}}^\top$. It should be noted, however, that for finite sample sizes, $4/N^2 A(0, \hat{\boldsymbol{\vartheta}}(0)_{\delta,\delta}$ is not the inverse variance of $U - \mathbb{E}(U)$ under $H_0 : \delta = 0$. In any case, the analytic expression allows computation of standard errors for δ for both estimated and assumed parameters, the latter option allowing the power curve to be approximated for certain values of δ in the alternative.

3.4. Alternative Two-sample Parameterisation

Sometimes the proportional odds model is parameterised differently in the two-sample situation, most importantly in terms of the probabilistic index, area under the ROC curve (AUC),

or the C-index (a broad overview is provided by Conroy 2012). For $Y(0) \sim Y \mid X = 0$ and $Y(1) \sim Y \mid X = 1$ we have

$$\text{AUC} = \mathbb{P}(Y(0) < Y(1)) + 1/2\mathbb{P}(Y(0) = Y(1)).$$

For $\delta = -\infty$ or ∞, the AUC is zero or one. For $\delta = 0$, the AUC is $1/2$. For all other values, the AUC is a monotone transformation of the log-odds ratio (Fay and Malinovsky 2018)

$$\text{AUC}(\delta) = \frac{\exp(\delta)(\exp(\delta) - 1 - \delta)}{(\exp(\delta) - 1)^2}.$$

Numerous authors proposed confidence intervals for this parameter (DeLong *et al.* 1988; Newcombe 2006a,b; Zhou 2006; Qin and Hotilovac 2008; Pauly *et al.* 2016; Feng *et al.* 2017; Fay and Malinovsky 2018; Perme and Manevski 2019, and many others), either with or without explicit reference to the connection with the Wilcoxon test. Because score intervals are invariant with respect to monotone transformations, we can apply the novel score intervals for log-odds ratios δ developed here directly for the computation of confidence intervals for the AUC. It must be noted, however, that the null hypothesis $H_0 : \text{AUC} = 1/2$ is not equivalent to the null $H_0 : \delta = 0$: The former allows for variance heterogeneity whereas the latter assumes equal distributions (Fay and Proschan 2010).

3.5. Generalisation

The technical arguments presented in this section allow the generalisation of model (1) to the comparison of $K > 2$ samples by introducing additional parameters δ_k, to stratified designs by introducing stratum-specific parameters $\boldsymbol{\vartheta}$, and to other forms of the alternative by replacing the logit-link with some other link function. We introduce the model underlying `free1way()` in this section.

We consider K treatment groups $G \in \{1, \ldots, K\}, K \geq 2$ for an at least ordered outcome $Y \in \Xi$ observed in stratum $S \in \{1, \ldots, B\}$ out of $B \geq 1$ blocks with conditional cumulative distribution function (CDF) $F_Y(y \mid G = k, S = b) = \mathbb{P}(Y \leq y \mid G = k, S = b)$. Detecting and describing differential distributions arising from different treatments is our main objective. We refer to the first treatment $G = 1$ as "control".

With model function $m : [0, 1] \times \mathbb{R} \to [0, 1]$ (introduced by Lehmann 1953, describing the power of rank tests against distributional alternatives), we describe the conditional distribution under treatment k as a function of the conditional distribution under control and a scalar parameter δ_k:

$$F_Y(y \mid G = k, S = b) = m(F_Y(y \mid G = 1, S = b), \delta_k). \tag{3}$$

The model is assumed to hold for all blocks $b = 1, \ldots, B$, treatments $k = 2, \ldots, K$, and outcome values $y \in \Xi$ based on parameters $\delta_2, \ldots, \delta_K \in \mathbb{R}$. For notational convenience, we define $\delta_1 := 0$.

This model formulation gives rise to several specific models, for example, $m_{\text{L}}(p, \delta) = p^{\exp(-\delta)}$ (Lehmann alternatives), $m_{\text{PH}}(p, \delta) = 1 - (1 - p)^{\exp(-\delta)}$ (proportional hazards), $m_{\text{PO}}(p, \delta) = \text{expit}(\text{logit}(p) - \delta)$ (proportional odds), or $m_{\text{Cd}}(p, \delta) = \Phi(\Phi^{-1}(p) - \delta)$ (generalised Cohen's d). Instead of directly working with m, we parameterise the model in terms of some absolute

continuous CDF F with log-concave density $f = F'$ and corresponding derivative f', and quantile or link function F^{-1}. The location model

$$F_Y(y \mid G = k, S = b) = F\left(F^{-1}\left(F_Y(y \mid G = 1, S = b)\right) - \delta_k\right), \quad k = 2, \ldots, K \tag{4}$$

describes different distributions by means of shift parameter on a latent scale defined by F. The negative shift term ensures that positive values of δ_k correspond to the situation of outcomes being stochastically larger in group k compared to control. The shift parameters are invariant with respect to monotone transformations of the response values, that is, transforming the observations of all treatment groups by the same function does not affect the values of δ_k. The choice $F(z) = \exp(-\exp(-z))$ gives rise to m_{L}, $F(z) = 1 - \exp(-\exp(z))$ corresponds to m_{PH}, $F = \text{expit}$ leads to m_{PO}, and $F = \Phi$ results in m_{Cd}. The choice of F is made a priori and determines the interpretation of δ_k.

We are interested in inference for $\delta_2, \ldots, \delta_K$, in terms of confidence intervals and hypothesis tests of the form

$$\begin{aligned}
H_0 &: \delta_k - \mu_k = 0, \text{“two.sided”}, \quad k = 2, \ldots, K, \\
H_0 &: \delta_k - \mu_k \geq 0, \text{“less”}, \quad k = K = 2, \\
H_0 &: \delta_k - \mu_k \leq 0, \text{“greater”}, \quad k = K = 2,
\end{aligned}$$

with the latter two options only for the two-sample case ($K = 2$). The null values $\mu_k, k = 2, \ldots, K$ allow testing hypotheses about specific values of the shift parameters (and can be specified as argument `mu` to `free1way()`).

As before, for an ordered categorical outcome Y from sample space $\Xi = \{\upsilon_1 < \upsilon_2 < \cdots < \upsilon_C\}$, we parameterise the model in terms of intercept ($\vartheta_\cdot$) and shift ($\delta_\cdot$) parameters

$$F_Y(\upsilon_c \mid G = k, S = b) = F(\vartheta_{c,b} - \delta_k), \quad c = 1, \ldots, C,$$

that is, we replace the transformed control outcome $F^{-1}\left(F_Y(\upsilon_c \mid G = 1, S = b)\right) = \vartheta_{c,b}$ with a corresponding intercept parameter. These $C - 1$ intercept parameters are block-specific and monotone increasing $\vartheta_{0,b} = -\infty < \vartheta_{1,b} < \cdots < \vartheta_{C,b} = \infty$ within each block $b = 1, \ldots, B$. We collect all model parameters in a vector

$$\begin{aligned}
\boldsymbol{\theta} = (\theta_1 &:= \delta_2, \\
&\cdots \quad , \\
\theta_{K-1} &:= \delta_K, \\
\theta_K &:= \vartheta_{1,1}, \\
\theta_{K+1} &:= \vartheta_{2,1} > \vartheta_{1,1}, \\
&\ldots, \\
\theta_{K+C-2} &:= \vartheta_{C-1,1} > \vartheta_{C-2,1}, \\
\theta_{K+C-1} &:= \vartheta_{1,2}, \\
&\cdots \quad , \\
\theta_{B(C-1)+K-1} &:= \vartheta_{C-1,B} > \vartheta_{C-2,B}).
\end{aligned}$$

If there is no observation for level c in block b, the corresponding parameter is not identified and removed from $\boldsymbol{\theta}$ (exceptions apply when penalised estimation is performed, see

Appendix C). The parameter space is defined by all parameter vectors $\boldsymbol{\theta}$ satisfying the monotonicity of the intercept parameters. Violations always lead to the log-likelihood function being undefined and thus taking the value $-\infty$. For the ith observation $(y_i = \upsilon_c, g_i = k, s_i = b)$ from block b under treatment k, the log-likelihood contribution is

$$\log(\mathbb{P}(\upsilon_{c-1} < Y \leq \upsilon_c \mid G = k, S = b)) = \log(F(\vartheta_{c,b} - \delta_k) - F(\vartheta_{c-1,b} - \delta_k)) \tag{5}$$

with $\upsilon_0 = -\infty$. For an absolutely continuous outcome $Y \in \mathbb{R}$, we define $\upsilon_c := y_{(c)}$, the cth distinct ordered observation in the sample. The log-likelihood above is then the empirical or nonparametric log-likelihood. If observations were independently right-censored, the contribution of the event $Y > \tilde{y}$ to the log-likelihood is

$$\log(\mathbb{P}(Y > \tilde{y} \mid G = k, S = b)) = \log(1 - F(\vartheta_{c-1,b} - \delta_k))$$

where $\upsilon_{c-1} = \max\{\upsilon \in \Xi \mid \upsilon \leq \tilde{y}\}$, that is, observations right-censored between υ_{c-1} and υ_c correspond to the parameter $\vartheta_{c-1,b}$.

Maximising this form of the log-likelihood leads to semiparametrically efficient estimators (Chapter 15.5, Van der Vaart 1998). In this framework, tests against deviations from the hypothesis H_0 above are locally most powerful rank tests, for example against proportional odds ($F = \text{expit}$) or proportional hazards alternatives ($F(z) = 1 - \exp(-\exp(z))$, Van der Vaart 1998, Example 15.16).

Based on the score function $S(\delta_2, \ldots, \delta_K)$ and the Fisher information $I(\delta_2, \ldots, \delta_K)$, which can be derived in the exact same way as presented for the proportional odds model (Section 3.1), maximum likelihood estimates $\hat{\delta}_2, \ldots, \hat{\delta}_K$ can be supplemented by Wald and likelihood ratio tests and confidence intervals as well as likelihood- and permutation-based Rao score tests and intervals. The power function for the Wald test can be evaluated relatively quickly by means of a small simulation for estimating the inverse Fisher information $I(\delta_2, \ldots, \delta_K)$ for a given alternative, sample size, and control distribution.

4. Implementation

Maximum likelihood estimation and parameter inference for model (3) is implemented in `free1way()` under the following principles. Data is represented in form of a $C \times K \times B$ contingency table, whose element (c, k, b) is the number of observations with configuration $(y = y_c, g = k, s = b)$. In the presence of right-censoring, a fourth dimension is added $(C \times K \times B \times 2)$ whose first $C \times K \times B$ table presents right-censoring and the second table contains numbers of events. The code, roughly 2600 lines, was written in R exclusively, only importing functionality from **Matrix** (Bates *et al.* 2026).

Lessons learned from existing implementations of some aspects of model (3) have been taken into account. Because the likelihood (2) for a nonparametric version of the transformation function h is equivalent to the likelihood of a proportional odds model for ordered categorical outcomes, the nonparametric maximum likelihood estimators $\hat{\delta}$ and $\hat{\boldsymbol{\vartheta}}$ along with $A(\delta_0, \hat{\boldsymbol{\vartheta}}(\delta_0))_{\delta,\delta}$ can be computed from `polr()` or `tram::Polr()` (Hothorn *et al.* 2018; Hothorn 2020), simply by fitting the model to ordered categories obtained by representing a numeric response as an ordinal factor. The same likelihood is implemented for numeric responses in `rms::orm()` (Harrell Jr 2026) which provides an implementation of the model for continuous responses (Liu *et al.* 2017). Only `Polr()` allowed stratification and permutation-based inference so far.

The restricted maximum likelihood estimator $\hat{\boldsymbol{\vartheta}}(\delta_0)$ can also be computed from these functions (using $\delta_0 x_i$ as an offset), only the computation of $A(\delta_0, \hat{\boldsymbol{\vartheta}}(\delta_0))_{\delta,\delta}$ is not commonly implemented. The term $\boldsymbol{I}(\delta_0, \hat{\boldsymbol{\vartheta}}(\delta_0))_{\delta,\boldsymbol{\vartheta}} \boldsymbol{I}(\delta_0, \hat{\boldsymbol{\vartheta}}(\delta_0))^{-1}_{\boldsymbol{\vartheta},\boldsymbol{\vartheta}} \boldsymbol{I}(\delta_0, \hat{\boldsymbol{\vartheta}}(\delta_0))^{\top}_{\delta,\boldsymbol{\vartheta}}$ is relatively simple to compute, because $\boldsymbol{I}(\delta_0, \hat{\boldsymbol{\vartheta}}(\delta_0))_{\boldsymbol{\vartheta},\boldsymbol{\vartheta}}$ is a tri-banded matrix. When the number of distinct observations C is large, the corresponding number of $\boldsymbol{\vartheta}$ parameters is, in the worst case scenario, growing with N. In contrast to all other implementations, `orm()` leveraged the tri-banded structure of the Hessian and employs a Newton-Raphson algorithm where parameter updates for tens of thousands of parameters can be computed very fast thanks to corresponding solvers being available in **Matrix**. The FORTRAN code underlying `orm()`, and the extensive evaluation of unconstrained optimisation in this context by Harrell Jr (2024), inspired much of the tailored optimisation procedures implemented in `free1way()`. Appendices A and B provide more detail regarding these aspects.

5. Pain Reduction using free1way

Performing several potentially interesting analyses of the neck pain randomised trial based on model (3) by means of `free1way()`, we explain how this novel implementation can help to address some methodological difficulties (such as those raised in Chow *et al.* 2006) and how unification of concepts and procedures scattered across different `stats::*.test()` functions previously is achieved.

5.1. Diagnostic Checking

Before we use `free1way()` to fit a proportional odds model corresponding to the Wilcoxon test presented in Section 2, we assess if the data are in line with the proportional odds assumption in two ways. The simplest idea is to reformulate model (1) as

$$\text{logit}\left(\mathbb{P}(Y \leq y \mid X = x)\right) = h(y \mid \boldsymbol{\vartheta}) - \delta x$$

and check if the two empirical distribution functions fitted to VAS separately in the two groups are roughly parallel on the logit-scale. Unlike plotting functionality for the ECDF in packages **stats** or **latticeExtra**, the `plot()` method for `survfit` objects from the **survival** package (Therneau 2026) allows plotting on user-supplied ordinate scales and thus, because the Kaplan-Meier estimator computed by `survival::survfit()` is identical to the ECDF in the absence of right-censoring, we can misuse this functionality for our purposes in Figure 4. The two resulting step functions are reasonably parallel and thus assuming proportional odds is appropriate. We come back to diagnostic testing once we have the fitted model in place.

We proceed by fitting the proportional odds model, that is, we estimate a log-odds ratio $\delta = \delta_2$ in addition to $C - 1 = 49$ increasing threshold parameters $\boldsymbol{\vartheta}$ by maximising the empirical likelihood, requesting the exact permutation distribution to be computed in addition

```
R> f <- free1way(vas ~ laser, data = pain, subset = time == "12 weeks",
+                exact = TRUE)
```

We add a visualisation of these parameters to Figure 4, enriching this model-free plot with the two, now model-based yet distribution-free or "semi-parametric", step-functions $h(y \mid \hat{\boldsymbol{\vartheta}})$ for the placebo group and $h(y \mid \hat{\boldsymbol{\vartheta}}) - \hat{\delta}$ for the active group in Figure 5. The model-free (solid)

```
R> w12 <- cbind(subset(pain, time == "12 weeks"), ev = TRUE)
R> plot(survfit(Surv(vas, ev) ~ laser, data = w12),
+       conf.int = FALSE, mark.time = FALSE,
+       fun = function(p) qlogis(p, lower.tail = FALSE),
+       col = cols, las = 1, main = "12 weeks",
+       xlab = "Neck pain (on visual analogue scale)",
+       ylab = "Distribution function on logit scale")
R> legend("bottomright", legend = levels(w12$laser), col = cols,
+         lty = 1, bty = "n", title = "Empirical")
```

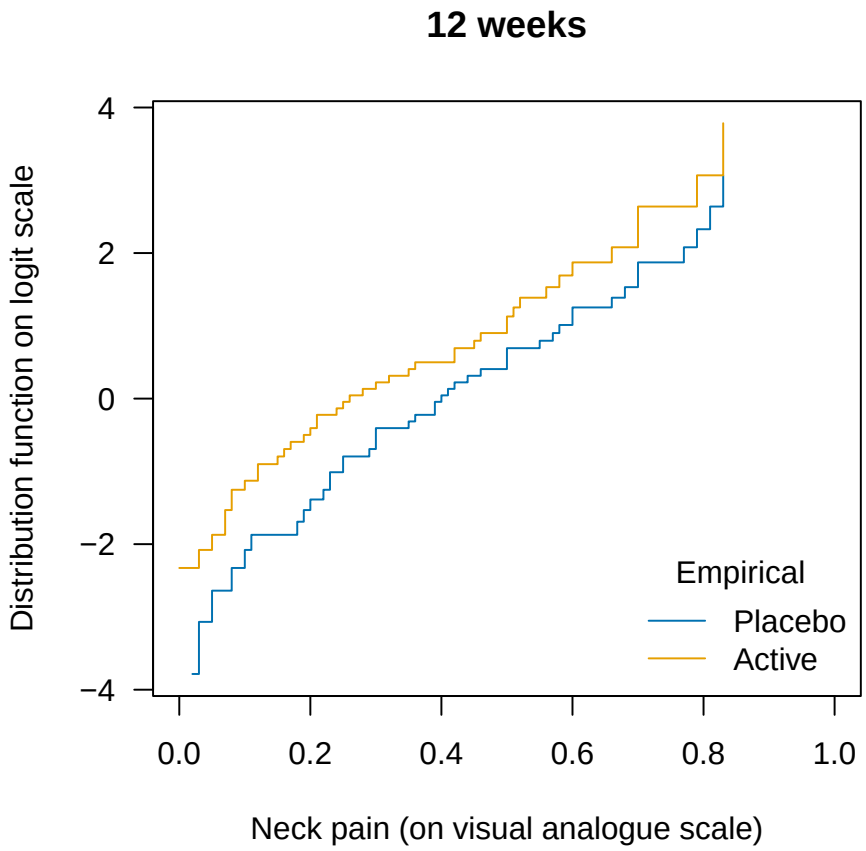


Figure 4: Low-level laser therapy against chronic neck pain (Chow *et al.* 2006) data: Check of proportional odds assumption. The empirical cumulative distribution functions for both treatment groups are plotted on the logit-scale.

and model-based (dashed) lines, the latter featuring a constant vertical shift $\hat{\delta}$, are almost perfectly in line and thus, with any potential worries about model fit out of our way, we can start interrogating the fitted model.

We first assess if the log-likelihood is identical to the log-likelihood of the model estimated by `polr()`

```
R> ### polr
R> logLik(p)

'log Lik.' -338.214 (df=50)

R> ### free1way
R> logLik(f)

[1] -338.214

R> ### number of log-odds ratio + intercept parameters
R> length(coef(f)) + length(f$intercepts[[1L]])

[1] 50
```

where `df` is the number of total parameters (log-odds ratio δ and intercepts $\boldsymbol{\vartheta}$) but not the degrees of freedom, because of the monotonicity constraint on $\boldsymbol{\vartheta}$. We can also check if the parameter estimates and the standard errors are reasonably close

```
R> all.equal(coef(p), coef(f))

[1] "Mean relative difference: 3.96313e-05"

R> all.equal(p$zeta, f$intercepts[[1L]], check.attributes = FALSE)

[1] "Mean relative difference: 1.50774e-05"

R> all.equal(vcov(p)[names(coef(p)),names(coef(p)), drop = FALSE],
+            vcov(f), check.attributes = FALSE)

[1] "Mean relative difference: 2.52501e-06"
```

5.2. Permutation Inference

We compute an exact permutation p-value and corresponding permutation confidence interval for δ. The exact conditional p-value testing $H_0 : \delta = 0$ is identical to the one reported by `wilcox.test()` in Section 2, however, parameter estimate and confidence interval are now reported for log-odds ratios

```
R> (sf <- summary(f, test = "Permutation"))

        2-sample Wilcoxon test against proportional odds alternatives

data:  vas by laser (Placebo, Active)
W = -2.015, p-value = 0.046
alternative hypothesis: true log-odds ratio is not equal to 0
```

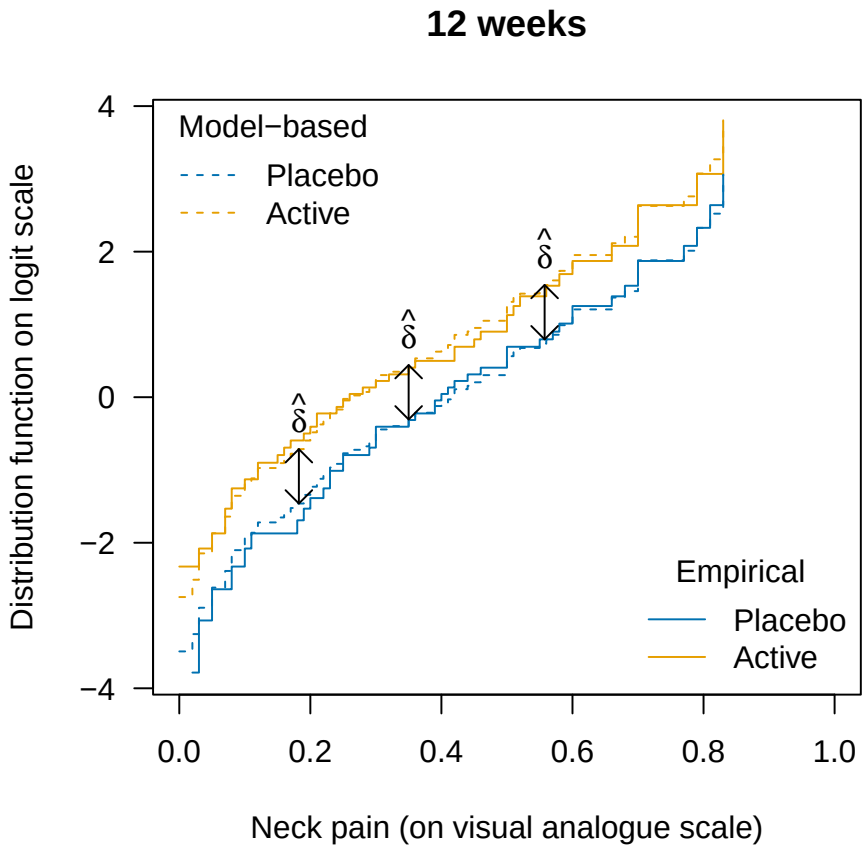


Figure 5: Low-level laser therapy against chronic neck pain (Chow *et al.* 2006) data: Check of proportional odds assumption with empirical and model-based logit-transformed distribution functions. The vertical distance between the model-based step-functions is constant at $|\hat{\delta}| = 0.747$.

```
R> coef(f)

laserActive
  -0.746923

R> confint(f, test = "Permutation")

                2.5 %     97.5 %
laserActive -1.48199 -0.0119882
```

With a negative value of $\hat{\delta} = -0.747$ (which is identical to the estimate $\hat{\delta} = -0.747$ provided by `polr()`), pain measured by VAS was lower in the active group compared to the placebo group. Because of the invariance of the maximum likelihood estimator and the score interval with respect to monotone transformations, we can report the score inference results for odds ratios

```
R> (OR <- exp(coef(f)))

laserActive
   0.473822

R> (ciOR <- exp(confint(f, test = "Permutation")))

               2.5 %   97.5 %
laserActive 0.227186 0.988083
```

concluding that the odds of observing pain on the VAS at most as large as y four weeks after the end of active laser treatment is about half as large as the corresponding odds in the placebo group, with an uncertainty ranging from a relative odds reduction between 23% and 99%.

Closer to the "Wilcoxon parameter" $\mathbb{P}(\text{VAS}_{\text{Placebo}} < \text{VAS}_{\text{Active}}) + 1/2\mathbb{P}(\text{VAS}_{\text{Placebo}} = \text{VAS}_{\text{Active}})$ for a pair of patients sampled independently from the two groups is the model-based probabilistic index (PI), also known as the area under the receiver operating characteristic (ROC) curve (AUC), again being a monotone transformation of δ and thus a permutation score interval being available

```
R> coef(f, what = "AUC")

laserActive
   0.377783

R> confint(f, test = "Permutation", what = "AUC")

               2.5 %   97.5 %
laserActive 0.269764 0.498002
```

Interestingly, not only the model-based AUC but also the model-based ROC curve is a simple function of δ (see Sewak and Hothorn 2023), opening up a second possibility for assessing the proportional odds assumption. We compare the probability-probability plot or empirical ROC curve with the model-based ROC curve in Figure 6 utilising the novel `ppplot()` function added to the **stats** package in R 4.6.0. The empirical and model-based ROC curves are quite close, so no red flags regarding lack of fit of the proportional odds model are raised by this additional diagnostic tool.

Another interesting estimand is the overlap coefficient, the integral over the minimum of the two, in case of VAS Lebesque, densities

```
R> coef(f, what = "OVL")

laserActive
    0.81541

R> confint(f, test = "Permutation", what = "OVL")
```

```
R> x <- subset(pain, subset = time == "12 weeks")
R> with(split(x$vas, x$laser),
+       ppplot(Active, Placebo, conf.level = .95,
+              main = "Visual Analogue Scale at 12 weeks"))
```

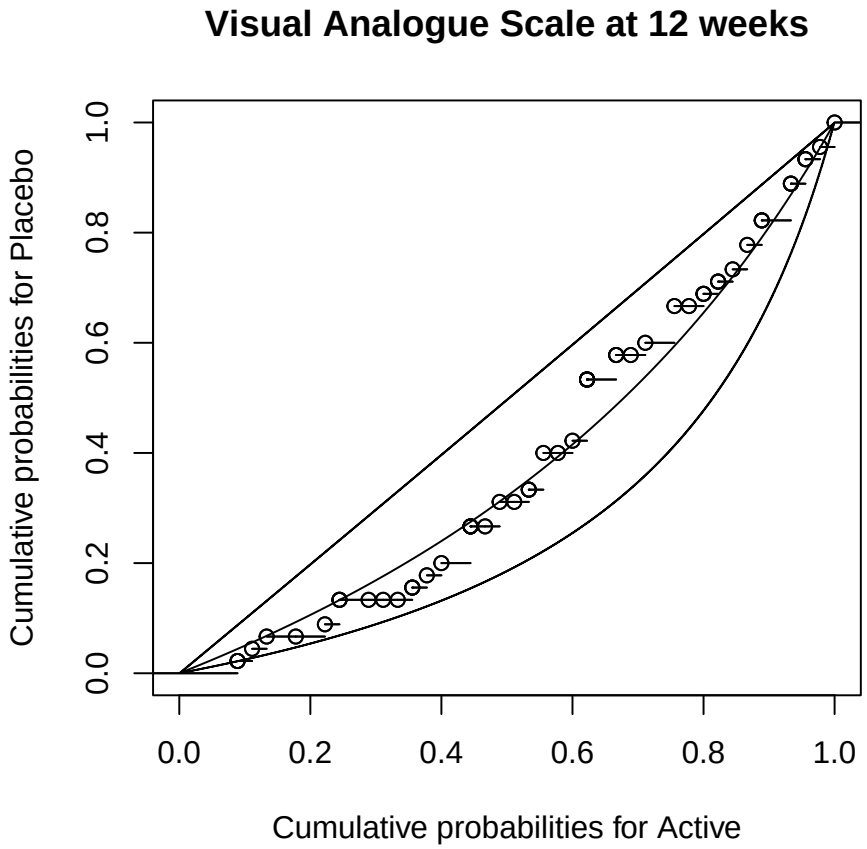


Figure 6: Low-level laser therapy against chronic neck pain (Chow *et al.* 2006) data: Probability-probability plot (dots) and model-based receiver operating characteristic (ROC) curve (lines, with 95% confidence band).

```
                2.5 %    97.5 %
laserActive 0.645574 0.997003
```

This measure is between zero, *i.e.,* a complete separation of VAS in the two groups, and one for identical VAS distributions. The interesting feature here is that we do not have to estimate the two densities directly, under the proportional odd assumption (Sewak and Hothorn 2023).

By changing the confidence level to the p-value, we can check if the p-values and confidence intervals are congruent, *i.e.,* δ_0 is contained in an $(1 - \alpha) \times 100\%$ confidence interval iff the p-value testing $H_0 : \delta = \delta_0$ is smaller than the nominal level α. This works reasonably well for permutation confidence intervals, where the upper bound of the $(1 - p\text{-value}) \times 100\% = 95.398\%$ confidence interval is close to zero

```
R> confint(f, level = 1 - sf$p.value, test = "Permutation")

               2.3 %       97.7 %
laserActive -1.49398 -1.92823e-06
```

Computing tests and confidence intervals using identical inference procedures in `free1way()` ensures that conclusions about effects drawn from tests are in line with effect sizes obtained from confidence intervals.

5.3. Likelihood Inference

For larger sample sizes or under the population model with a priori fix groups, classical likelihood inference procedures can be performed by changing the `test` argument in the `summary()` and `coef()` methods. Rao score tests and intervals are available from

```
R> summary(f, test = "Rao")

        2-sample Wilcoxon test against proportional odds alternatives

data:  vas by laser (Placebo, Active)
Rao chi-squared = 4.062, df = 1, p-value = 0.0439
alternative hypothesis: true log-odds ratio is not equal to 0

R> confint(f, test = "Rao")

               2.5 %     97.5 %
laserActive -1.47361 -0.0203546
```

likelihood-ratio tests and intervals from

```
R> summary(f, test = "LRT")

        2-sample Wilcoxon test against proportional odds alternatives

data:  vas by laser (Placebo, Active)
logLR chi-squared = 4.081, df = 1, p-value = 0.0434
alternative hypothesis: true log-odds ratio is not equal to 0

R> confint(f, test = "LRT")

              2.5 %     97.5 %
laserActive -1.4849 -0.0221214
```

and Wald tests and intervals from

```
R> summary(f, test = "Wald")
```

```
        2-sample Wilcoxon test against proportional odds alternatives

data:  vas by laser (Placebo, Active)
Wald chi-squared = 4.024, df = 1, p-value = 0.0449
alternative hypothesis: true log-odds ratio is not equal to 0

R> confint(f, test = "Wald")

               2.5 %     97.5 %
laserActive -1.47671 -0.0171379
```

Asymptotically, these three procedures are known to provide the same results, but finite sample differences can be quite substantial. Only score and likelihood ratio inference is invariant with respect to monotone transformations, and should be applied when results are reported for functions of δ.

5.4. Power and Sample Size

Other researchers trying to replicate the above results in an independent trial may ask what sample size would be needed to reproduce the same effect size with a certain power. The novel `power.free1way.test()` function implements the power function of the Wald test based on simulation-based estimates of standard errors as a function of effect size (argument `delta`), sample size in each group (argument `n`), and nominal level α (argument `sig.level`). Because these simulations can be performed relatively quickly, the power function can be solved for the sample size if the requested power is specified:

```
R> (pt <- power.free1way.test(delta = coef(f), power = .8, sig.level = .05))

      2-sample Wilcoxon test against proportional odds alternatives

                n = 88
Total sample size = 88 (Control) + 88 (laserActive) = 176
            power = 0.801078
        sig.level = 0.05
   Standard error = 0.266241
   log-odds ratio = -0.746923

NOTE: 'n' is sample size in control group
```

Because this result for the Wald test is not exact, study designers would check if the power of the planned inference procedure (such as an asymptotic permutation test) is in fact close to the power requested by performing a simulation study directly estimating the power. Function `rfree1way()` can be used to sample, by default continuous, response values under a given model, in our case a proportional odds model with given log-odds ratio

```
R> ex <- expression({
+      df <- rfree1way(n = pt$n, delta = pt[["log-odds ratio"]])
```

```
+      summary(free1way(y ~ groups, data = df),
+              test = "Permutation")$p.value
+  })
R> pvals <- replicate(1000, eval(ex))
R> binom.test(x = sum(pvals < pt[["sig.level"]]), n = length(pvals),
+             p = pt[["power"]])$conf.int

R> ex <- expression({
+      df <- rfree1way(n = pt$n, delta = pt[["log-odds ratio"]])
+      summary(free1way(y ~ groups, data = df),
+              test = "Permutation")$p.value
+  })
R> pvals <- replicate(1000, eval(ex))
R> binom.test(x = sum(pvals < pt[["sig.level"]]), n = length(pvals),
+             p = pt[["power"]])$conf.int

[1] 0.788472 0.837664
attr(,"conf.level")
[1] 0.95
```

which suggests that such a replication trial with substantially larger sample size compared to the initial study would achieve the desired power. As a means to increase precision, Chow *et al.* (2006) performed covariate-adjustment for baseline VAS, and we will implement this idea using stratification in the next section.

5.5. Stratification

Before we begin, we follow Section 2.10 in Chow *et al.* (2006) and categorise VAS measurements at all three time points into an ordinal variable at four levels of increasing pain

```
R> pain$vasO <- cut(pain$vas, breaks = c(-Inf, .3, .6, .8, Inf),
+                   ordered_result = TRUE,
+                   labels = c("None to mild", "Moderate",
+                              "Moderately severe", "Severe"))
```

Baseline pain is then added as a stratum variable to all patients

```
R> xB <- subset(pain, time == "Baseline")
R> pain$vasBL <- xB$vasO[match(pain$id, xB$id)]
R> class(pain$vasBL) <- "factor"
```

We repeat the proportional odds model analysis now separately in the four strata defined by ordinal baseline VAS assuming that the log-odds ratio is the same in all four strata. Unfortunately, the exact stratified permutation distribution is not yet implemented, so we use a larger number of random permutation in each stratum to approximate this distribution allowing permutation-based inference for the common odds ratio

```
R> fb <- free1way(vas ~ laser | vasBL, data = pain,
+                 subset = time == "12 weeks", B = 10000)
R> summary(fb, test = "Permutation")

        Stratified 2-sample Wilcoxon test against proportional odds
        alternatives

data:  vas by laser (Placebo, Active)
         stratified by vasBL
Perm Z = -4.474, p-value <2e-16
alternative hypothesis: true log-odds ratio is not equal to 0

R> exp(coef(fb))

laserActive
   0.121984

R> exp(confint(fb, test = "Permutation"))

               2.5 %   97.5 %
laserActive 0.046953 0.319014
attr(,"Attained level")
[1] 0.95
```

This stratified, or conditional, odds ratio estimate is smaller than the marginal odds ratio and less variable, however, the two estimands are not directly comparable because the model is noncollapsible.

Because the proportional odds model and its empirical likelihood describe and handle continuous and ordinal responses in the very same way, we can repeat the analysis with the response on the four pain levels

```
R> f0 <- free1way(vas0 ~ laser | vasBL, data = pain,
+                 subset = time == "12 weeks", B = 10000)
R> summary(f0, test = "Permutation")

        Stratified 2-sample Wilcoxon test against proportional odds
        alternatives

data:  vas0 by laser (Placebo, Active)
         stratified by vasBL
Perm Z = -4.029, p-value <2e-16
alternative hypothesis: true log-odds ratio is not equal to 0

R> exp(coef(f0))

laserActive
  0.0872696
```

```
R> exp(confint(f0, test = "Permutation"))

               2.5 %   97.5 %
laserActive 0.0248541 0.311581
attr(,"Attained level")
[1] 0.9501
```

The results are very close, especially when comparing the two confidence intervals. This might encourage the designers of the replication study to simplify pain assessment by asking study participants if the pain they suffer is "None to mild","Moderate","Moderately severe", or "Severe" instead of going all-in with VAS examinations. We might wonder what the impact on the required sample size for a potential replication study might be when using an ordinal response under stratification for ordinal baseline pain. We first estimate the pain distribution under placebo in the four strata

```
R> (btab <- xtabs(~ vas0 + vasBL, data = pain,
+                 subset = laser == "Placebo" & time == "12 weeks")[,4:1])

                  vasBL
vas0               Severe Moderately severe Moderate None to mild
  None to mild          0                 0        6           12
  Moderate              0                 6        8            3
  Moderately severe     0                 1        4            1
  Severe                1                 1        1            1

R> colSums(btab)

           Severe Moderately severe          Moderate
                1                 8                19
     None to mild
               17

R> (ptab <- prop.table(btab, margin = 2))

                  vasBL
vas0                  Severe Moderately severe  Moderate
  None to mild      0.0000000         0.0000000 0.3157895
  Moderate          0.0000000         0.7500000 0.4210526
  Moderately severe 0.0000000         0.1250000 0.2105263
  Severe            1.0000000         0.1250000 0.0526316
                  vasBL
vas0               None to mild
  None to mild        0.7058824
  Moderate            0.1764706
  Moderately severe   0.0588235
  Severe              0.0588235
```

Because only one patient with severe baseline pain remained at this pain level, we remove the first stratum for the sample size planning. The number of patients in first pain levels at baseline are twice as large compared to the number of baseline patients with moderately severe pain, we use the latter stratum as the first one and double the sample size in the other two strata (argument `strata_ratio`). Furthermore, we define the discrete four point distribution of the pain response for the control group, so that the power of the Wald test for ordinal responses will be approximated in this call to `power.free1way.test()`

```
R> (pbt <- power.free1way.test(prob = ptab[,-1], strata_ratio = c(2, 2),
+                              delta = coef(f0), power = .8,
+                              sig.level = .05))

     Stratified 2-sample Wilcoxon test against proportional odds alternatives

                n = 6
Total sample size = 30 (vas0) + 30 (laserActive) = 60
            power = 0.851626
        sig.level = 0.05
   Standard error = 0.811998
   log-odds ratio = -2.43875

NOTE: 'n' is sample size in control group of first stratum
```

Because the power function is highly discrete in small n, the requested power cannot be achieved exactly. In any case, the number of patients to be enrolled in such a trial would be much smaller than the two times $n = 88$ necessary in a trial without stratification.

5.6. Analysing Binary Responses

Although we saw that converting continuous VAS to an ordinal variable at four levels did not change the results substantially going further analysing VAS dichotomised as VAS $\leq .6$ is probably overdoing it. However, for binary responses in a two-arm trial, the stratified proportional odds model is identical to the model underlying the Cochran-Mantel-Haenszel test:

```
R> mh <- free1way(vasB ~ laser | vasBL, data = pain,
+                 subset = time == "12 weeks")
R> summary(mh, test = "Permutation")$p.value

   Perm Z
0.0141862

R> xt <- xtabs(~ vasB + laser + vasBL,
+              data = pain, subset = time == "12 weeks")
R> mantelhaen.test(xt, correct = FALSE, exact = FALSE)$p.value

[1] 0.0141862
```

Indeed, the asymptotic score permutation test coincides with the Cochran-Mantel-Haenszel test uncorrected for small samples (without continuity correction, as demonstrated by Day and Byar 1979).

The two estimates of the common odds ratio and their confidence intervals are not identical, because the Mantel-Haenszel and conditional maximum likelihood estimator (Mehta and Patel 1985) reported by `mantelhaen.test()` are not identical to the maximum likelihood estimator computed by `free1way()`.

```
R> ### Mantel-Haenszel
R> mantelhaen.test(xt, correct = FALSE, exact = FALSE)$estimate

common odds ratio
         0.119372

R> ### Conditional ML
R> mantelhaen.test(xt, correct = FALSE, exact = TRUE)$estimate

common odds ratio
         0.154894

R> ### ML
R> exp(coef(mh))

laserActive
   0.142811
```

The latter estimator and corresponding inference is only valid when the number of blocks is fixed and the sample size grows within each block. For many strata with low number of observations, with matched-pairs designs being the most extreme example, this is typically not the case as the number of blocks corresponds to the number of observations and thus sample size. In this case, the maximum likelihood estimator computed by `free1way()` is known to suffer finite sample bias and both the Mantel-Haenszel and conditional maximum likelihood estimator correct this bias. The phenomenon also occurs for ordered categorical outcomes (Liu and Agresti 1996). One possible remedy to this problem is discussed in the next section and in Appendix C.

5.7. Repeated Measures Analyses

The analyses presented so far compare two conditional pain distributions observed in the two treatment groups and thus, by definition, assess the treatment effect for groups of patients. It is hard to say something about the pain reduction in single patients. Therefore, in our final analysis, we look at the study from a complete different angle, treating the three pain measurements on the VAS as repeated measurements. In the formulation of model (3), each patient is its own block or stratum. This is dangerous, because now the sample size within blocks is fixed (and equal to one) whereas the number of blocks grows with number of patients. For ordinal outcomes like VAS, Liu and Agresti (1996) described the bias inherent

in the estimation procedure implemented in `free1way()`. We will now discuss how we can still make use of `free1way()` to analyse such repeated measurement data.

We first look at the data from a slightly different perspective, by comparing the individual pain measurements to baseline pain in Figure 7. Almost all patients under active laser treatment reported a pain reduction after seven weeks. In contrast, more than half the number of patients in the placebo group reported more pain in this period of time. After 12 weeks, the pain reduction in the active group is still substantial whereas the variability seemed to have increased in the placebo group.

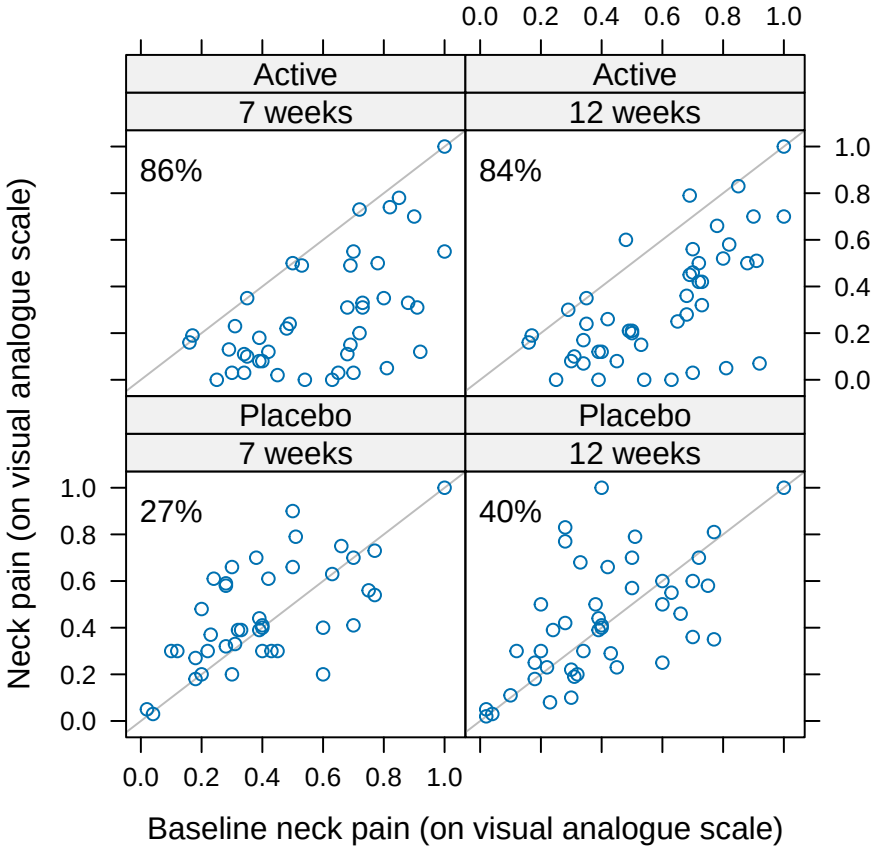


Figure 7: Low-level laser therapy against chronic neck pain (Chow *et al.* 2006) data: Comparison to follow-up pain measurements to baseline pain in two treatment groups. The percentages refer to the fraction of patients reporting an improvement (follow-up VAS $<$ baseline VAS).

We assume a patient-specific yet common log-odds ratio δ_2 comparing pain at the end of treatment with baseline and a common log-odds ratio δ_3 comparing the pain four weeks after end of treatment with baseline for all patients in the placebo group. In contrast to the model employed before, the parameters δ_2 and δ_3 do not compare the treatment-specific but the patient-specific distribution functions of VAS. Treating patients as blocks allows each patient to carry his or her own baseline VAS distribution function, from which we only sample one

observation. This also means that each patient comes with her or his own VAS distribution at seven or 12 weeks. We make the strong assumption that, for each patient, the follow-up VAS distributions are linked to the baseline VAS distribution of this patient in terms of a proportional odds model with, now common to all patients, log-odds ratio δ_2 (seven weeks) and δ_3 (12 weeks). In the same way, such effects also exists in the active laser group (let's call them $\tilde{\delta}_2$ and $\tilde{\delta}_3$), so we perform the analysis twice, starting with the placebo group testing the hypothesis $H_0 : \delta_2 = \delta_3 = 0$:

```
R> fP <- free1way(vas ~ time | id, data = pain,
+                 subset = laser == "Placebo" & !(id %in% miss))
R> unlist(summary(fP, test = "Permutation")[c("statistic", "p.value")])

statistic.Perm chi-squared   p.value.Perm chi-squared
                  4.158273                   0.125038

R> painP <- subset(pain, laser == "Placebo" & !(id %in% miss))
R> painP$id <- painP$id[,drop = TRUE]
R> unlist(friedman.test(vas ~ time | id,
+                       data = painP)[c("statistic", "p.value")])

statistic.Friedman chi-squared                        p.value
                      4.158273                       0.125038
```

This gives exactly the same results as `friedman.test()` when we remove the observations with missing values from the data, because `friedman.test()` only works for full-factorial designs. There is no evidence against the null hypothesis here, so pain in patients in the placebo group does not change over time. Limiting this analysis to two time points (baseline and 12 weeks) and VAS dichotomised at .6, the permutation score test is identical to McNemar's test without continuity correction. The only relevant information here is the number of patients who switched from VAS larger than .6 to VAS at most as large as .6 and vice versa

```
R> summary(free1way(vasB ~ time | id, data = pain,
+                   subset = laser == "Placebo" & time != "7 weeks"),
+          test = "Permutation")$p.value

  Perm Z
0.781511

R> xt <- xtabs(~ vasBL + vasB, data = pain,
+               subset = laser == "Placebo" & time != "7 weeks")
R> xt <- rbind(colSums(xt[1:2,]), colSums(xt[3:4,]))
R> mcnemar.test(xt, correct = FALSE)$p.value

[1] 0.781511
```

For the active laser group, the repeated measurements analysis leads to a clear rejection of the null $H_0 : \tilde{\delta}_2 = \tilde{\delta}_3 = 0$ for the response on the original VAS scale

```
R> fA <- free1way(vas ~ time | id, data = pain,
+                 subset = laser == "Active" & !(id %in% miss))
R> unlist(summary(fA, test = "Permutation")[c("statistic", "p.value")])

statistic.Perm chi-squared   p.value.Perm chi-squared
              4.63553e+01                8.59177e-11

R> painA <- subset(pain, laser == "Active" & !(id %in% miss))
R> painA$id <- painA$id[,drop = TRUE]
R> unlist(friedman.test(vas ~ time | id,
+                       data = painA)[c("statistic", "p.value")])

statistic.Friedman chi-squared                      p.value
                  4.63553e+01                   8.59177e-11
```

but it remains unclear how large the trend to, hopefully (there is no direction of effects in p-values), less pain actually is. In addition, this analysis does not contrast the temporal trend in the active laser group with the trend in the placebo group.

The permutation p-values are valid, because they exclusively rely on parameters estimated under the constraints formulated by the null hypotheses ($\delta_2 = \delta_3 = 0$) and thus the bias for these log-odds ratios described by Liu and Agresti (1996) cannot occur. Of course, fitting the unrestricted model, parameter estimates for the log-odds ratios will be biased and thus also confidence intervals will suffer from this problem. We can try to debias the maximum likelihood estimates, for example by adding a penalty term to the log-likelihood function. One possibility is to use an idea similar to Firth's correction in generalised linear models (Firth 1993; Kosmidis and Firth 2020), essentially penalising the log-likelihood using a penalty term defined by the determinant of the Fisher information matrix (also known as Jeffrey's prior), further details are provided in Appendix C.

We re-fit the two models to all patients treated in the two arms, compare the parameter estimates and compute the contrast of interest for the penalised parameters. We start with the actively treated group

```
R> fA <- free1way(vas ~ time | id, data = pain,
+                 subset = laser == "Active" & !(id %in% miss))
R> fAp <- free1way(vas ~ time | id, data = pain,
+                  subset = laser == "Active" & !(id %in% miss),
+                  MPL_Jeffreys = TRUE)
R> cbind(unpenalised = coef(fA), penalised = coef(fAp))

R> fA <- free1way(vas ~ time | id, data = pain,
+                 subset = laser == "Active" & !(id %in% miss))
R> fAp <- free1way(vas ~ time | id, data = pain,
+                  subset = laser == "Active" & !(id %in% miss),
+                  MPL_Jeffreys = TRUE)
R> cbind(unpenalised = coef(fA), penalised = coef(fAp))
```

```
             unpenalised penalised
time7 weeks     -5.53997  -1.63711
time12 weeks    -4.25218  -1.17997
```

The log-odds ratios comparing the seven and 12 weeks time points with baseline are all negative, thus pain was reduced, however, more so after seven weeks. The penalised log-odds ratios are much smaller and reasonable.

For the placebo group, we obtain (with one warning indicating that the optimiser tried to evaluate the likelihood (5) for a configuration of the parameters violating the constraints)

```
R> fP <- free1way(vas ~ time | id, data = pain,
+                 subset = laser == "Placebo" & !(id %in% miss))
R> fPp <- free1way(vas ~ time | id, data = pain,
+                  subset = laser == "Placebo" & !(id %in% miss),
+                  MPL_Jeffreys = TRUE)
R> cbind(unpenalised = coef(fP), penalised = coef(fPp))
```

```
R> fP <- free1way(vas ~ time | id, data = pain,
+                 subset = laser == "Placebo" & !(id %in% miss))
R> fPp <- free1way(vas ~ time | id, data = pain,
+                  subset = laser == "Placebo" & !(id %in% miss),
+                  MPL_Jeffreys = TRUE)
R> cbind(unpenalised = coef(fP), penalised = coef(fPp))
```

```
             unpenalised  penalised
time7 weeks     1.275001 -0.0137593
time12 weeks    0.653505 -0.3514360
```

The penalised log-odds ratios clearly hint at a constant pain profile.

At the end of the day, we are interested in comparing the two treatment groups with respect to the patient-specific log-odds ratios, that is, we want to know if the pain reduction at the two time points was different between the two groups, so we study the contrasts

```
R> cbind(unpenalised = coef(fA) - coef(fP),
+        penalised = coef(fAp) - coef(fPp))
```

```
R> cbind(unpenalised = coef(fA) - coef(fP),
+        penalised = coef(fAp) - coef(fPp))
```

```
             unpenalised penalised
time7 weeks     -6.81497 -1.623354
time12 weeks    -4.90569 -0.828534
```

The effect of penalisation is still clearly visible, as well as the dampening of the treatment effect over time which can also be observed from Figure 7. It should be noted that the simultaneous comparison of both time point to baseline gives different results compare to the

analysis of the data when we only compare a single time point to baseline, because both shift parameters impact the intercept parameters, see also Appendix D.

6. Odds and Ends

The default link function `link = "logit"` implements the proportional odds model. Alternative link functions give rise to Lehmann alternatives (`link = "loglog"`), proportional hazards models (`link = "cloglog"`), and latent normal models with parameters comparable to Cohen's d (`link = "probit"`). Independent right-censoring of response values can be specified using `survival::Surv()`. It is important to note that `free1way()` always performs full maximum likelihood estimation of all model parameters, so one must not assume to observe results identical to other estimation techniques of the same model (most prominently, to the partial likelihood for proportional hazards models in `survival::survfit()` or `survival::coxph()`). However, likelihood inference results for the unstratified case should be identical to `orm()` (Harrell Jr 2026). This highly tuned FORTRAN implementation served as a blue print for many aspects of `free1way()`. For large samples with a limited number of response categories, the computing times of the two implementations are comparable. For continuous responses and thus many threshold parameters $\boldsymbol{\vartheta}$, `free1way()` is typically slower by a factor of three or four compared to the lightning fast `orm()`.

7. Discussion

The novel `free1way()` function unifies many classical test procedures available for decades in **stats**, among them `wilcox.test()`, `kruskal.test()`, `friedman.test()`, `mantelhaen.test()`, `prop.test()`, and `mcnemar.test()` and adds new members to this family, most importantly distribution-free models against Lehmann alternatives or Cohen's d. The underlying models, corresponding structural assumptions, and parameterisations surface, allowing data analysts to perform diagnostic tests justifying (or not) the application of such models. Most importantly, `free1way()` equips these classical and novel models with a proper likelihood and corresponding parameter estimates as well as a plethora of permutation- and population-based inference procedures. Power assessment and sample size planning is now directly linked to estimation and inference procedures.

References

Bates D, Maechler M, Jagan M (2026). *Matrix: Sparse and Dense Matrix Classes and Methods*. doi:10.32614/CRAN.package.Matrix. R package version 1.7-6, URL https://CRAN.R-project.org/package=Matrix.

Bauer DF (1972). "Constructing Confidence Sets Using Rank Statistics." *Journal of the American Statistical Association*, **67**(339), 687–690. doi:10.1080/01621459.1972.10481279.

Chow RT, Heller GZ, Barnsley L (2006). "The Effect of 300 mW, 830 nm Laser on Chronic Neck Pain: A Double-blind, Randomized, Placebo-controlled Study." *Pain*, **124**(1–2), 201–210. doi:10.1016/j.pain.2006.05.018.

Conroy RM (2012). "What Hypotheses do "Nonparametric" Two-group Tests Actually Test?" *The Stata Journal*, **12**(2), 182–190. doi:10.1177/1536867X1201200202.

Day NE, Byar DP (1979). "Testing Hypotheses in Case-Control Studies-Equivalence of Mantel-Haenszel Statistics and Logit Score Tests." *Biometrics*, **35**(3), 623–630. doi:10.2307/2530253.

DeLong ER, DeLong DM, Clarke-Pearson DL (1988). "Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach." *Biometrics*, **44**(3), 837–845. doi:10.2307/2531595.

Divine GW, Norton HJ, Barón AE, Juarez-Colunga E (2018). "The Wilcoxon-Mann-Whitney Procedure Fails as a Test of Medians." *The American Statistician*, **72**(3), 278–286. doi:10.1080/00031305.2017.1305291.

Fahrmeir L, Tutz G (2001). *Multivariate Statistical Modelling Based on Generalized Linear Models*. Springer-Verlag, New York, USA.

Fay MP, Malinovsky Y (2018). "Confidence Intervals of the Mann-Whitney Parameter That are Compatible With the Wilcoxon-Mann-Whitney Test." *Statistics in Medicine*, **37**(27), 3991–4006. doi:/10.1002/sim.7890.

Fay MP, Proschan MA (2010). "Wilcoxon-Mann-Whitney or t-test? On Assumptions for Hypothesis Tests and Multiple Interpretations of Decision Rules." *Statistical Surveys*, **4**, 1–39. doi:10.1214/09-SS051.

Feng D, Cortese G, Baumgartner R (2017). "A Comparison of Confidence/credible Interval Methods for the Area under the ROC Curve for Continuous Diagnostic Tests with Small Sample Size." *Statistical Methods in Medical Research*, **26**(6), 2603–2621. doi:10.1177/0962280215602040.

Firth D (1993). "Bias Reduction of Maximum Likelihood Estimates." *Biometrika*, **80**(1), 27–38. doi:10.1093/biomet/80.1.27.

Fong Y, Huang Y (2019). "Modified Wilcoxon-Mann-Whitney Test and Power Against Strong Null." *The American Statistician*, **73**(1), 43–49. doi:10.1080/00031305.2017.1328375.

Harrell Jr FE (2015). "Ordinal Logistic Regression." In *Regression Modeling Strategies*, pp. 311–325. Springer. doi:10.1007/978-3-319-19425-7_13.

Harrell Jr FE (2024). "Statistical Computing Approaches to Maximum Likelihood Estimation." *Technical report*, Blog Post. URL https://www.fharrell.com/post/mle/.

Harrell Jr FE (2026). *rms: Regression Modeling Strategies*. doi:10.32614/CRAN.package.rms. R package version 8.1-1, URL https://CRAN.R-project.org/package=rms.

Hothorn T (2020). "Transformation Boosting Machines." *Statistics and Computing*, **30**, 141–152. doi:10.1007/s11222-019-09870-4.

Hothorn T, Hornik K (????). *free1way.docreg: Additional Documentation and Regression Tests for 'stats::free1way()'*. R package version 1.1-0, URL https://codeberg.org/thothorn/coin/.

Hothorn T, Möst L, Bühlmann P (2018). "Most Likely Transformations." *Scandinavian Journal of Statistics*, **45**(1), 110–134. doi:10.1111/sjos.12291.

Kosmidis I, Firth D (2020). "Jeffreys-prior Penalty, Finiteness and Shrinkage in Binomial-response Generalized Linear Models." *Biometrika*, **108**(1), 71–82. doi:10.1093/biomet/asaa052.

Kruskal WH (1957). "Historical Notes on the Wilcoxon Unpaired Two-Sample Test." *Journal of the American Statistical Association*, **52**(279), 356–360. doi:10.1080/01621459.1957.10501395.

Lehmann EL (1953). "The Power of Rank Tests." *The Annals of Mathematical Statistics*, **24**(1), 23–43. doi:10.1214/aoms/1177729080.

Lehmann EL (1998). *Elements of Large-Sample Theory*. 2nd edition. Springer-Verlag, New York, USA.

Liu M, Agresti A (1996). "Mantel-Haenszel-Type Inference for Cumulative Odds Ratios with a Stratified Ordinal Response." *Biometrics*, **52**(4), 1223–1234. doi:10.2307/2532838.

Liu Q, Shepherd BE, Li C, Harrell Jr FE (2017). "Modeling Continuous Response Variables Using Ordinal Regression." *Statistics in Medicine*, **36**(27). doi:10.1002/sim.7433.

Mann HB, Whitney DR (1947). "On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other." *The Annals of Mathematical Statistics*, **18**(1), 50–60.

Manuguerra M, Heller GZ (2010). "Ordinal Regression Models for Continuous Scales." *International Journal of Biostatistics*, **6**(1), 14. doi:10.2202/1557-4679.1230.

McCullagh P (1980). "Regression Models for Ordinal Data." *Journal of the Royal Statistical Society. Series B (Methodological)*, **42**(2), 109–142. doi:10.1111/j.2517-6161.1980.tb01109.x.

Mehta CR, Patel NR (1985). "Exact Logistic Regression: Theory and Examples." *Statistics in Medicine*, **14**(19), 2143–2160. doi:10.1002/sim.4780141908.

Murphy SA, van der Vaart AW (1997). "Semiparametric Likelihood Ratio Inference." *The Annals of Statistics*, **25**(4), 1471–1509. doi:10.1214/aos/1031594729.

Murphy SA, van der Vaart AW (2000). "On Profile Likelihood." *Journal of the American Statistical Association*, **95**(450), 449–465. doi:10.1080/01621459.2000.10474219.

Newcombe RG (2006a). "Confidence Intervals for an Effect Size Measure Based on the Mann-Whitney Statistic. Part 1: General Issues and Tail-area-based Methods." *Statistics in Medicine*, **25**(4), 543–557. doi:10.1002/sim.2323.

Newcombe RG (2006b). "Confidence Intervals for an Effect Size Measure Based on the Mann-Whitney Statistic. Part 2: Asymptotic Methods and Evaluation." *Statistics in Medicine*, **25**(4), 559–573. doi:10.1002/sim.2324.

Owen AB (2001). *Empirical Likelihood*. Chapman and Hall/CRC, New York, U.S.A. doi:10.1201/9781420036152.

Parker S, Jernigan RW, Lansky JM (2020). "When is the Wilcoxon-Mann-Whitney Procedure a Test of Location? Implications for Effect-size Measures." *British Journal of Mathematical and Statistical Psychology*, **73**(1), 170–183. doi:10.1111/bmsp.12162.

Pauly M, Asendorf T, Konietschke F (2016). "Permutation-based Inference for the AUC: A Unified Approach for Continuous and Discontinuous Data." *Biometrical Journal*, **58**(6), 1319–1337. doi:10.1002/bimj.201500105.

Pawitan Y (2001). *In All Likelihood*. Oxford University Press, Oxford, UK.

Perme MP, Manevski D (2019). "Confidence Intervals for the Mann-Whitney Test." *Statistical Methods in Medical Research*, **28**(12), 3755–3768. doi:10.1177/0962280218814556.

Portnoy S, Koenker R (1997). "The Gaussian Hare and the Laplacian Tortoise: Computability of Squared-error versus Absolute-error Estimators." *Statistical Science*, **12**(4), 279–300. doi:10.1214/ss/1030037960.

Prentice RL (1978). "Linear Rank Tests with Right Censored Data." *Biometrika*, **65**(1), 167–179. doi:10.1093/biomet/65.1.167.

Qin G, Hotilovac L (2008). "Comparison of Non-parametric Confidence intervals for the Area under the ROC Curve of a Continuous-scale Diagnostic Test." *Statistical Methods in Medical Research*, **17**(2), 207–221. doi:10.1177/0962280207087173.

R Core Team (2026). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. doi:10.32614/R.manuals. URL https://www.R-project.org/.

Rao CR (2005). "Score Test: Historical Review and Recent Developments." In *Advances in Ranking and Selection, Multiple Comparisons, and Reliability*, pp. 8–20. Birkhäuser, Boston, U.S.A.

Ripley BD, Venables WN (2002). *Modern Applied Statistics with S*. Fourth edition. Springer, New York. ISBN 0-387-95457-0, URL https://www.stats.ox.ac.uk/pub/MASS4/.

Sewak A, Hothorn T (2023). "Estimating Transformations for Evaluating Diagnostic Tests with Covariate Adjustment." *Statistical Methods in Medical Research*, **32**(7), 1403–1419. doi:10.1177/09622802231176030.

Therneau TM (2026). *A Package for Survival Analysis in R*. R package version 3.8-12, URL https://CRAN.R-project.org/package=survival.

van der Vaart AW (1998). *Asymptotic Statistics*. Cambridge University Press, Cambridge, UK.

Van der Vaart AW (1998). *Asymptotic Statistics*. Cambridge University Press, Cambridge, UK. doi:10.1017/CBO9780511802256.

Wang Y, Tian L (2017). "The Equivalence Between Mann-Whitney Wilcoxon Test and Score Test Based on the Proportional Odds Model for Ordinal Responses." In *4th International Conference on Industrial Economics System and Industrial Security Engineering (IEIS)*, pp. 1–5. IEEE, Kyoto, Japan. doi:10.1109/IEIS.2017.8078606.

Whitehead J (1993). "Sample Size Calculations for Ordered Categorical Data." *Statistics in Medicine*, **12**(24), 2257–2271. doi:10.1002/sim.4780122404.

Wilcoxon F (1945). "Individual Comparisons by Ranking Methods." *Biometrics Bulletin*, **1**(6), 80–83.

Zhou GQXH (2006). "Empirical Likelihood Inference for the Area under the ROC Curve." *Biometrics*, **62**(2), 613–622. doi:10.1111/j.1541-0420.2005.00453.x.

A. Implementation and Quality Assurance

The R code implementing `free1way()` is part of the **stats** package. The add-on package **free1way.docreg** (Hothorn and Hornik) serves as a vehicle for distributing

- A literate programming document linking theory and implementation and explaining all technical details. The document is also available as package vignette

  ```
  R> vignette("free1way", package = "free1way.docreg")
  ```

- Additional regression tests.

B. Optimisation

The choice of optimisation routines in `free1way()` is in line with the extensive experiments on this issue reported on by Harrell Jr (2024) and by corresponding implementation in the **rms** add-on package (Harrell Jr 2026). Two aspects are of prime importance: First, it is not necessary to actively perform constraint optimisation, because the log-likelihood contributions (5) is only defined iff the intercept parameters within each block are ordered. Otherwise, the log-likelihood contribution is `-Inf`. Thus, a parameter configuration is inadmissible iff the log-likelihood is not finite and many optimisers can handle this setup.

Second, the Fisher information is available in analytical form and the block corresponding to the intercept parameters is tri-banded, facilitating very efficient storage and parameter updates even when the number of distinct outcomes and thus the number intercept terms is very large.

Two optimisation routines are currently in-built: `nlminb()` and a tailored version of Newton-Raphson with step-halving, the latter suggested by Harrell Jr (2024). The main difference is that `nlminb()` can only handle Fisher information matrices stored in dense form whereas the latter (internal) `.NewtonRaphson()` function leverages the sparsity through functionality offered by the **Matrix** (Bates *et al.* 2026) add-on package. For small sample sizes (and penalised estimation), the default is `nlminb()`, for all other problems `.NewtonRaphson()` is applied. Users can switch between the two by `dooptim = "nlminb"` or `dooptim = ".NewtonRaphson"`.

It is also possible to apply third-party optimisers, the only condition being that a user-interface compatible with the one sported by `nlminb()` is available. We illustrate ways to experiment with other optimisers using the following problematic model

```
R> fPp <- free1way(vas ~ time | id, data = pain,
+                  subset = laser == "Placebo",
+                  MPL_Jeffreys = TRUE, logdet = TRUE)

R> fPp <- free1way(vas ~ time | id, data = pain,
+                  subset = laser == "Placebo",
+                  MPL_Jeffreys = TRUE, logdet = TRUE)
```

issuing the warning `unsuccessful optimisation in free1way: false convergence (8)`. We can implement the penalised negative log-likelihood function (6) for all model parameters $\theta = (\delta_2, \ldots, \delta_K, \boldsymbol{\vartheta})$ and compare the two penalised log-likelihoods

```
R> pll <- function(theta) {
+      ret <- fPp$profile(theta, fix = seq_along(theta))
+      ret$value - .5 * c(ret$logdet)
+  }
R> logLik(fPp)
R> -pll(fPp$par)

R> pll <- function(theta) {
+      ret <- fPp$profile(theta, fix = seq_along(theta))
+      ret$value - .5 * c(ret$logdet)
+  }
R> logLik(fPp)

[1] -107.716

R> -pll(fPp$par)

[1] -107.716
```

Using the estimated parameters as starting values, we can refit the model using Nelder-Mead

```
R> op <- optim(fPp$par, fn = pll)
R> logLik(fPp)
R> -op$value
R> coef(fPp)
R> op$par[1:2]

R> op <- optim(fPp$par, fn = pll)
R> logLik(fPp)

[1] -107.716

R> -op$value

[1] -107.511
```

```
R> coef(fPp)

 time7 weeks time12 weeks
  -0.0279659   -0.4194244

R> op$par[1:2]

[1] -0.0234542 -0.4213366
```

and obtain a slightly larger value of the penalised log-likelihood but the log-odds ratios are quite similar. We can also write a wrapper around `optim()` and specify the name of this function as argument `dooptim`, together with a named list of control arguments (which must feature `trace` and `iter.max`)

```
R> optNM <- function(start, objective, control = list(trace = FALSE), ...) {
+      ctrl <- list(trace = control$trace)
+      ret <- optim(start, fn = objective, control = ctrl)
+      ret$objective <- ret$value
+      ret
+  }
R> fPpNM <- free1way(vas ~ time | id, data = pain, dooptim = "optNM",
+      subset = laser == "Placebo",
+      control = list(optNM = list(trace = FALSE, iter.max = 100)),
+      MPL_Jeffreys = TRUE, logdet = TRUE)
R> logLik(fPp)
R> logLik(fPpNM)

R> optNM <- function(start, objective, control = list(trace = FALSE), ...) {
+      ctrl <- list(trace = control$trace)
+      ret <- optim(start, fn = objective, control = ctrl)
+      ret$objective <- ret$value
+      ret
+  }
R> fPpNM <- free1way(vas ~ time | id, data = pain, dooptim = "optNM",
+      subset = laser == "Placebo",
+      control = list(optNM = list(trace = FALSE, iter.max = 100)),
+      MPL_Jeffreys = TRUE, logdet = TRUE)
R> logLik(fPp)

[1] -107.716

R> logLik(fPpNM)

[1] -124.206
```

This exercise fails with a reportedly unsuccessful optimisation, and it seems Nelder-Mead was not able reproduce its better result when using the default starting values.

C. Penalisation

The sum ℓ of the log-likelihood contributions (5) is penalised by a factor $1/2A$ of the log-determinant of the Fisher information

$$\ell(\delta_2, \ldots, \delta_K, \boldsymbol{\vartheta}) + \frac{1}{2} A \log(\det(\boldsymbol{I}(\delta_2, \ldots, \delta_K, \boldsymbol{\vartheta}))), \tag{6}$$

where the argument `MPL_Jeffreys =` A uses $A =$`FALSE` by default.

For binary outcomes and two time points, the penalised log-odds ratio is identical to the one obtained by Firth's method in a binary logistic regression model and close to the Mantel-Haenszel estimate

```
R> library("brglm2")
R> xt <- xtabs(~ vasB + time + id, data = pain,
+              subset = laser == "Placebo" & time != "7 weeks",
+              drop.unused.levels = TRUE)
R> dxt <- as.data.frame(xt)
R> Firth <- glm(vasB ~ 0 + id + time,
+               data = dxt, weights = dxt$Freq,
+               family = binomial(link = "logit"),
+               method = "brglmFit", type = "MPL_Jeffreys")
R> exp(coef(Firth)["time12 weeks"])
R> exp(coef(fJ <- free1way(xt, MPL_Jeffreys = TRUE)))
R> mantelhaen.test(xt)$estimate
R> ### biased ML, the default
R> exp(coef(free1way(xt, MPL_Jeffreys = FALSE)))
```

```
R> library("brglm2")
R> xt <- xtabs(~ vasB + time + id, data = pain,
+              subset = laser == "Placebo" & time != "7 weeks",
+              drop.unused.levels = TRUE)
R> dxt <- as.data.frame(xt)
R> Firth <- glm(vasB ~ 0 + id + time,
+               data = dxt, weights = dxt$Freq,
+               family = binomial(link = "logit"),
+               method = "brglmFit", type = "MPL_Jeffreys")
R> exp(coef(Firth)["time12 weeks"])
```

```
time12 weeks
     1.12907
```

```
R> exp(coef(fJ <- free1way(xt, MPL_Jeffreys = TRUE)))
```

```
time12 weeks
     1.12907

R> mantelhaen.test(xt)$estimate

common odds ratio
          1.16667

R> ### biased ML, the default
R> exp(coef(free1way(xt, MPL_Jeffreys = FALSE)))

time12 weeks
     1.36111
```

It is important to note that for null hypothesis testing only, patients or blocks with the same outcome value at the two time points do not play a role and are removed (as noted for McNemar's test). In contrast, for estimating the treatment effect, these observations contribute information and therefore have to be retained. In the analysis above, not only the log-odds ratio but also all patient-specific intercepts are identical (up to their signs, the treatment effect comes last for `glm()` and first for `free1way()`)

```
R> max(abs(coef(Firth)[-length(coef(Firth))] + fJ$par[-1L]))

R> max(abs(coef(Firth)[-length(coef(Firth))] + fJ$par[-1L]))

[1] 1.9237e-06
```

Computing penalised estimators is slow because the penalised log-likelihood is currently implemented but the corresponding score function is still missing.

D. Peculiarities

In the framework implemented by `free1way()`, permutation score tests are always available and valid, because the model is only fitted under the null hypothesis and the intercept parameters can always be estimated. Estimating treatment effects and intercept parameters simultaneously may however fail because of a very flat log-likelihood as it is the case when there is no overlap of the outcomes between the two groups.

For example, with observations $1, \ldots, N$ in one group and $N + 1, \ldots, 2N$ in the other group, we can perform an exact Wilcoxon test both using `wilcox.test()`

```
R> N <- 5
R> d <- data.frame(groups = gl(2, N))
R> d$y <- seq_len(2 * N)
R> wilcox.test(y ~ groups, data = d, exact = TRUE)$p.value

[1] 0.00793651
```

and `free1way()`

```
R> f <- free1way(y ~ groups, data = d, exact = TRUE)
R> summary(f, test = "Permutation")$p.value

[1] 0.00793651
```

The call to `free1way()` produces many warnings about optimisation problems, because the log-likelihood is flat. This is also visible from looking at the Hessian of the profile log-likelihood

```
R> f$hessian

            groups2
groups2 2.75291e-10
```

and we can vary the log-odds ratio δ while refitting the intercept parameters by directly accessing the profile log-likelihood

```
R> coef(f)

groups2
45.8589

R> theta <- f$par[-1L]
R> delta <- 20:60
R> logLik(f)

[1] -16.0944

R> range(sapply(delta, function(d)
+      -f$profile(c(d, theta), fix = 1L)$value))

[1] -16.0948 -16.0944
```

It might also come as a surprise that a joint analysis of all treatment groups produces slightly different results than an analysis contrasting subsets thereof. The reason is that the intercept parameters apply to all groups and thus removing or adding groups has an impact on both the intercept and the treatment effect parameters. In this small example, differences in means computed by `lm()` are not affected by the number of groups in the model but log-odds ratios differ slightly

```
R> d <- rfree1way(25, delta = c(.5, .75))
R> ### identical
R> cbind(coef(lm(y ~ groups, data = d))[-1L],
+        c(coef(lm(y ~ groups, data = d, subset = groups != "C"))[-1L],
+          coef(lm(y ~ groups, data = d, subset = groups != "B"))[-1L]))
```

```
            [,1]     [,2]
groupsB 0.201682 0.201682
groupsC 0.160274 0.160274
```

```
R> ### not identical
R> cbind(coef(free1way(y ~ groups, data = d)),
+        c(coef(free1way(y ~ groups, data = d, subset = groups != "C")),
+          coef(free1way(y ~ groups, data = d, subset = groups != "B"))))
```

```
            [,1]    [,2]
groupsB 1.43169 1.31514
groupsC 1.17760 1.10343
```

E. Computational Environment

- R version 4.6.2 RC (2026-10-23 r90783), `x86_64-pc-linux-gnu`
- Locale: `LC_CTYPE=en_GB.UTF-8`, `LC_NUMERIC=C`, `LC_TIME=en_GB.UTF-8`, `LC_COLLATE=C`, `LC_MONETARY=en_GB.UTF-8`, `LC_MESSAGES=en_GB.UTF-8`, `LC_PAPER=en_GB.UTF-8`, `LC_NAME=C`, `LC_ADDRESS=C`, `LC_TELEPHONE=C`, `LC_MEASUREMENT=en_GB.UTF-8`, `LC_IDENTIFICATION=C`
- Time zone: `Europe/Berlin`
- TZcode source: `system (glibc)`
- Running under: `Debian GNU/Linux 13 (trixie)`
- Matrix products: default
- BLAS: `/home/hothorn/R-rc/lib/libRblas.so`
- LAPACK: `/home/hothorn/R-rc/lib/libRlapack.so` ; LAPACK version3.12.1
- Base packages: base, datasets, grDevices, graphics, methods, splines, stats, utils
- Other packages: Deriv 4.3.6, MASS 7.3-66, Rcpp 1.1.2, bibtex 0.5.3, boot 1.3-32, free1way.docreg 1.1-0, lattice 0.23-1, latticeExtra 0.6-31, memisc 0.99.31.8.3, ordinalCont 2.0.2, survival 3.8-12
- Loaded via a namespace (and not attached): Formula 1.2-6, Matrix 1.7-6, RColorBrewer 1.1-3, abind 1.4-8, car 3.1-5, carData 3.0-6, compiler 4.6.2, data.table 1.18.6.1, deldir 2.0-4, grid 4.6.2, interp 1.1-6, jpeg 0.1-11, jsonlite 2.0.0, png 0.1-9, tools 4.6.2, yaml 2.3.12

Affiliation:

Torsten Hothorn
Institut für Epidemiologie, Biostatistik und Prävention
Universität Zürich
Hirschengraben 84, CH-8001 Zürich, Switzerland
Torsten.Hothorn@R-project.org

Kurt Hornik
Institute for Statistics and Mathematics
Department of Finance, Accounting and Statistics
Wirtschaftsuniversität Wien, Welthandelsplatz 1, AT-1020 Wien, Austria
Kurt.Hornik@R-project.org