

DISSERTATION

Nonparametric correlation-based methods
with biomedical applications

zur Erlangung des Grades
Dr. rer. nat.

vorgelegt der Fakultät Statistik
der Technischen Universität Dortmund

von

Claus P. Nowak

Dortmund, Januar 2022

veröffentlicht als Dissertation in der Fakultät Statistik in Dortmund
Erstgutachten erstellt von Markus Pauly
Zweitgutachten erstellt von Kirsten Schorning
mündliche Prüfung abgenommen am 1. Juli 2022

Contents

Summary	2
Extended Summary	3
Univariate distribution functions and the Mann-Whitney parameter	3
Bivariate distribution functions and Kendall's tau	5
Group sequential methodology for the Mann-Whitney parameter	9
References	13
Manuscripts	13
Simultaneous inference for Kendall's tau	14
Group sequential methods for the Mann-Whitney parameter	33
The nonparametric Behrens-Fisher problem in small samples	50

Summary

This dissertation consists of three manuscripts on nonparametric methodology, i.e., *Simultaneous inference for Kendall's tau*, *Group sequential methods for the Mann-Whitney parameter*, and *The nonparametric Behrens-Fisher problem in small samples*. Owing to the cumulative character of this thesis, some parts, in particular those dealing with notation and definitions, had to be repeated throughout the thesis.

The manuscript on Kendall's tau (Nowak & Konietschke, 2021) extends results which I have initially developed in my master's thesis. While the master's thesis only dealt with a linear transformation of Kendall's τ_A and only briefly hinted as to how one could generalise this approach to multivariate factorial designs, this dissertation fully develops a nonparametric estimation theory for multiple rank correlation coefficients in terms of Kendall's τ_B , Somers' D , as well as Kruskal and Goodman's γ , necessitating joint estimation of both the probabilities of ties occurring and the probability of concordance minus discordance. Moreover, I apply the proposed methods to the `iris` flower data set.

As for the second manuscript (Nowak, Mütze, & Konietschke, 2022a), I review and further develop group sequential methodology for the Mann-Whitney parameter. With the aid of data from a clinical trial in patients with relapse-remitting multiple sclerosis, I demonstrate how one could repeatedly estimate the Mann-Whitney parameter during an ongoing trial together with repeated confidence intervals obtained by test inversion. In addition, I give simple approximate power formulas for this group sequential setting.

The last manuscript (Nowak, Pauly, & Brunner, 2022b) further explores how best to approximate the sampling distribution of the Mann-Whitney parameter in terms of the nonparametric Behrens-Fisher problem, an issue that has arisen from the preceding manuscript. In that regard, I explore different variance estimators and a permutation approach that have been proposed in the literature and examine some slightly modified ways as regards a small sample t approximation as well.

In all three manuscripts, I carried out simulations for various settings to assess the adequacy of the proposed methods.

Apart from my supervisor Markus Pauly at the TU Dortmund University, I am particularly grateful to Tobias Mütze, Frank Konietschke, and Edgar Brunner for their comments and suggestions.

Extended summary

As already mentioned in the *summary*, this is a cumulative thesis comprising three manuscripts, namely, *Simultaneous inference for Kendall's tau*, *Group sequential methods for the Mann-Whitney parameter*, and *The nonparametric Behrens-Fisher problem in small samples*.

The main purpose of the extended summary is to bring to the fore the core idea behind deriving the asymptotic sampling distribution of nonparametric effects, that is to say, the probabilities necessary to compute the Mann-Whitney parameter and the different versions of Kendall's tau – no matter what their true population values amount to. In that respect, I will exploit the fact that the proposed approach allows for estimation of the joint asymptotic distribution of various effect estimators and that the resulting test statistics can be inverted to produce confidence limits.

Section deals with univariate distribution and survival functions and provides a definition of the Mann-Whitney parameter and its estimator, followed by a discussion on asymptotics. In Section , I likewise examine different versions of Kendall's tau in terms of bivariate distribution and survival functions to derive their asymptotic sampling distribution. Finally, in Section , I will give a brief introduction to group sequential methodology using a simple example and then focus attention on the Mann-Whitney parameter.

In introducing notation and defining the various effect quantities, I closely follow the pertinent sections in Brunner, Bathke, and Konietzschke (2018), Nowak (2019), Nowak and Konietzschke (2021), Jennison and Turnbull (2000), Nowak et al. (2022a), as well as Nowak et al. (2022b). As for technical details and proofs, I will generally refer to other sources or to those just mentioned.

Univariate distribution functions and the Mann-Whitney parameter

To enhance readability, I will introduce the notation in terms of real-valued random variables only. However, all definitions and results are valid for any random variables mapping to a totally ordered set, covering the more general case of ordered categorical data as well. From a somewhat different point of view, one may just as well assign a real number, often referred to as a score, to each category, while preserving the order, such as “no pain” $\equiv 0$, “moderate pain” $\equiv 1$, and “severe pain” $\equiv 2$ for a three-point pain scale. Framing the problem this way, the consideration of real-valued random variables also suffices for ordered categorical data.

Definition 1 (Univariate cumulative distribution and survival functions). *Let X denote a univariate real-valued random variable defined on the probability space $(\Omega, \mathcal{A}, \mathbb{P})$. Then for any fixed $x \in \mathbb{R}$, we denote by*

$$F^-(x) = \mathbb{P}(X < x), \quad F^+(x) = \mathbb{P}(X \leq x), \quad F(x) = \mathbb{P}(X < x) + 1/2 \cdot \mathbb{P}(X = x)$$

the left-continuous, the right-continuous, and the normalised versions of the cumulative distribution function of X , respectively. In a similar vein, we call

$$S^-(x) = \mathbb{P}(X > x), \quad S^+(x) = \mathbb{P}(X \geq x), \quad S(x) = \mathbb{P}(X > x) + 1/2 \cdot \mathbb{P}(X = x)$$

the right-continuous, the left-continuous, and the normalised versions of the survival function of X .

Naturally, the use of the term *survival function* is questionable if X is not a time-to-event random variable. For purposes of this thesis, however, the survival function only serves as a convenient technical means to define asymptotic variances and their estimators.

To define the empirical distribution and survival functions, it is convenient to first introduce the following count functions, see also Brunner et al. (2018).

Definition 2 (Count functions). *Let x and y denote two real numbers. Then we call*

$$c^-(x, y) = \begin{cases} 0, & x \leq y \\ 1, & x > y \end{cases}, \quad c^+(x, y) = \begin{cases} 0, & x < y \\ 1, & x \geq y \end{cases}, \quad c(x, y) = \begin{cases} 0, & x < y \\ 1/2, & x = y \\ 1, & x > y \end{cases}$$

count functions.

As just mentioned, I will now turn to the empirical distribution and survival functions.

Definition 3 (Univariate empirical distribution and survival functions). *Let $X_1, \dots, X_n \stackrel{iid}{\sim} F(x)$ denote a random sample of n real-valued observations. For any $x \in \mathbb{R}$, we denote by*

$$\widehat{F}^-(x) = \frac{1}{n} \sum_{k=1}^n c^-(x, X_k), \quad \widehat{F}^+(x) = \frac{1}{n} \sum_{k=1}^n c^+(x, X_k), \quad \widehat{F}(x) = \frac{1}{n} \sum_{k=1}^n c(x, X_k)$$

the left-continuous, the right-continuous and the normalised version of the empirical distribution function of X_1, \dots, X_n , respectively. Likewise, we call

$$\widehat{S}^-(x) = \frac{1}{n} \sum_{k=1}^n c^-(X_k, x), \quad \widehat{S}^+(x) = \frac{1}{n} \sum_{k=1}^n c^+(X_k, x), \quad \widehat{S}(x) = \frac{1}{n} \sum_{k=1}^n c(X_k, x)$$

the right-continuous, the left-continuous and the normalised version of the empirical survival function of X_1, \dots, X_n .

Equipped with these definitions, I will now introduce the Mann-Whitney parameter and in due course an estimator together with its asymptotic distribution.

Definition 4 (Mann-Whitney parameter). *Let $X_1 \sim F_1(x)$ and $X_2 \sim F_2(x)$ denote two independent real-valued random variables. Then the probability*

$$p = \mathbb{P}(X_1 < X_2) + 1/2 \cdot \mathbb{P}(X_1 = X_2)$$

is called the Mann-Whitney parameter.

The Mann-Whitney parameter p is also referred to as nonparametric relative effect of X_2 with respect to X_1 or probabilistic index (Brunner et al., 2018; Thas, De Neve, Clement, & Ottoy, 2012). To illustrate its interpretation, suppose that X_1 and X_2 stand for responses from treatment arms 1 and 2 of a parallel two-arm clinical trial and that lower values point to a more favourable outcome. Then the effect p is nothing but the probability that a patient on arm 1 will fare better than one on arm 2, including $1/2$ times the probability of equal outcomes.

To ease estimation, I will now give a representation of the Mann-Whitney parameter p in terms of Lebesgue-Stieltjes integrals.

Corollary 5 (Integral representation of p). *Let $X_1 \sim F_1(x)$ and $X_2 \sim F_2(x)$ denote two independent real-valued random variables. We can then express the Mann-Whitney parameter as*

$$p = \int_{\mathbb{R}} F_1(x) dF_2(x) = \int_{\mathbb{R}} S_2(x) dF_1(x).$$

Proof. See Brunner et al. (2018) or Nowak (2019). □

As a common shorthand notation, one may prefer to drop the function arguments and the integration region, yielding $p = \int F_1 dF_2 = \int S_2 dF_1$. To give an estimator, all that is left to do is replace the theoretical distribution functions with their empirical counterparts.

Definition 6 (Estimation of p). *Let $X_{11}, \dots, X_{1m_1} \stackrel{iid}{\sim} F_1(x)$ and $X_{21}, \dots, X_{2m_2} \stackrel{iid}{\sim} F_2(x)$ denote two independent random samples of real-valued observations. We can then estimate the Mann-Whitney parameter p by*

$$\widehat{p} = \int_{\mathbb{R}} \widehat{F}_1(x) d\widehat{F}_2(x).$$

Again, one could have used $\widehat{p} = \int_{\mathbb{R}} \widehat{S}_2(x) d\widehat{F}_1(x)$ as well. Moreover, as a shorthand we could instead write $\widehat{p} = \int \widehat{F}_1 d\widehat{F}_2 = \int \widehat{S}_2 d\widehat{F}_1$ just as before. As for the limiting distribution, I will rely on the central limit theorem as applied by Brunner and Munzel (2000) and Brunner et al. (2018).

Proposition 7 (Asymptotics of \hat{p}). Let $X_{11}, \dots, X_{1n_1} \stackrel{iid}{\sim} F_1(x)$ and $X_{21}, \dots, X_{2n_2} \stackrel{iid}{\sim} F_2(x)$ denote two independent random samples of real-valued observations. With total sample size $N = n_1 + n_2$, Mann-Whitney parameter p together with its estimator \hat{p} as given in Definitions 4 and 6, variance estimator $\hat{\sigma}_A^2 = N(\int \hat{S}_2^2 d\hat{F}_1 - \hat{p}^2)/(n_1 - 1) + N(\int \hat{F}_1^2 d\hat{F}_2 - \hat{p}^2)/(n_2 - 1)$, and under some mild regularity conditions, we have convergence in distribution to a normal random variate, i.e.,

$$\sqrt{N}(\hat{p} - p)/\hat{\sigma}_A \xrightarrow[N \rightarrow \infty]{\mathcal{D}} \mathcal{N}(0, 1).$$

Proof. As already alluded to, see Brunner and Munzel (2000) and / or Brunner et al. (2018). \square

Since I will later make use of the same approach to derive the joint asymptotic distribution of the probabilities associated with Kendall's tau, I will now informally describe the main idea behind the proof. In essence, Brunner and Munzel (2000) and Brunner et al. (2018) suggest splitting $\sqrt{N}(\hat{p} - p)$ into sums of independent and identically distributed random variables and an asymptotically negligible part. More specifically, consider

$$\begin{aligned} \hat{p} - p &= \int \hat{F}_1 d\hat{F}_2 - p \\ &= \int \hat{F}_1 d\hat{F}_2 - p + \underbrace{\int \hat{F}_1 dF_2 - \int \hat{F}_1 d\hat{F}_2}_{=0} + \underbrace{\int F_1 d\hat{F}_2 - \int F_1 d\hat{F}_2}_{=0} + \underbrace{\int F_1 dF_2 - p}_{=0} \\ &= \int F_1 d\hat{F}_2 + \int \hat{F}_1 dF_2 - 2p + \left(\int \hat{F}_1 d\hat{F}_2 - \int F_1 d\hat{F}_2 - \left(\int \hat{F}_1 dF_2 - \int F_1 dF_2 \right) \right) \\ &= \int F_1 d\hat{F}_2 + \int S_2 d\hat{F}_1 - 2p + \left(\int (\hat{F}_1 - F_1) d\hat{F}_2 - \int (\hat{F}_1 - F_1) dF_2 \right) \\ &= \int S_2 d\hat{F}_1 + \int F_1 d\hat{F}_2 - 2p + \int (\hat{F}_1 - F_1) d(\hat{F}_2 - F_2) \\ &= \frac{1}{n_1} \sum_{i=1}^{n_1} S_2(X_{1i}) + \frac{1}{n_2} \sum_{j=1}^{n_2} F_1(X_{2j}) - 2p + \int (\hat{F}_1 - F_1) d(\hat{F}_2 - F_2). \end{aligned}$$

Multiplication of both sides by the square root of the total sample size produces

$$\begin{aligned} \sqrt{N}(\hat{p} - p) &= \sqrt{N} \underbrace{\left(\frac{1}{n_1} \sum_{i=1}^{n_1} S_2(X_{1i}) + \frac{1}{n_2} \sum_{j=1}^{n_2} F_1(X_{2j}) - 2p \right)}_{=: A_N} + \underbrace{\sqrt{N} \left(\int (\hat{F}_1 - F_1) d(\hat{F}_2 - F_2) \right)}_{=: B_N}. \end{aligned}$$

Observe that the random variables $S_2(X_{1i})$ and $F_1(X_{2j})$ are only random through X_{1i} and X_{2j} respectively, yielding sums of independent and identically distributed random variables. Moreover, one can show that B_N converges to zero in probability as N tends to infinity. Consequently, $\sqrt{N}(\hat{p} - p)$ and A_N share the same limiting distribution and therefore the same asymptotic variance. Thus we can regard

$$\sigma_A^2 = \mathbb{V}(A_N) = \frac{N}{n_1} \mathbb{V}(S_2(X_{11})) + \frac{N}{n_2} \mathbb{V}(F_1(X_{21})) = \frac{N}{n_1} \left(\int S_2^2 dF_1 - p^2 \right) + \frac{N}{n_2} \left(\int F_1^2 dF_2 - p^2 \right)$$

as our variance estimand.

Bivariate distribution functions and Kendall's tau

As intimated at the end of the previous section, I will now consider nonparametric association measures. To begin with, I will give definitions of Kendall's τ_A and τ_B , Somers' D , and Kruskal and Goodman's γ (Kendall & Gibbons, 1990; Kruskal, 1958; Somers, 1962; Goodman & Kruskal, 1954).

Definition 8 (Versions of Kendall's tau). *Let (X_1, Y_1) and (X_2, Y_2) denote two independent and identically distributed two-dimensional real-valued random vectors. With $D_X = \text{sgn}(X_1 - X_2)$ and $D_Y = \text{sgn}(Y_1 - Y_2)$, we call*

$$\tau_A = \mathbb{E}(D_X D_Y), \quad \tau_B = \frac{\mathbb{E}(D_X D_Y)}{\sqrt{\mathbb{E}(D_X^2)} \sqrt{\mathbb{E}(D_Y^2)}}, \quad D_{YX} = \frac{\mathbb{E}(D_X D_Y)}{\mathbb{E}(D_X^2)}, \quad \gamma = \frac{\mathbb{E}(D_X D_Y)}{\mathbb{P}(D_X D_Y \neq 0)}$$

the population versions of Kendall's τ_A and τ_B , Somers' D , and Kruskal and Goodman's γ , respectively.

If ties cannot occur almost surely, then all versions coincide, i.e., $\tau_A = \tau_B = D_{YX} = \gamma$. Similar to Pearson's product moment correlation coefficient $\rho = \text{Cov}(X_1, Y_1) / \sqrt{\mathbb{V}(X_1)\mathbb{V}(Y_1)}$, all four versions of Kendall's tau would then amount to 1 if X_1 and Y_1 are perfectly positively associated, -1 in case of perfect negative, and 0 in case of no association. However, if the occurrence of ties is possible, a variable associated with itself, i.e., $D_Y \equiv D_X$, would no longer imply that $\tau_A = \mathbb{E}(D_X^2) = \mathbb{P}(X_1 \neq X_2)$ equals 1, whereas τ_B and γ scale the association in such a way that we would still have $\tau_B = \gamma = 1$. On the other hand, Somers' definition of D_{YX} is motivated by the slope coefficient in a simple linear regression model.

As was the case with the Mann-Whitney parameter, I will now introduce theoretical bivariate distribution and survival functions as well as their empirical counterparts with a view to expressing all expectations and probabilities in Definition 8 in terms of Lebesgue-Stieltjes integrals to facilitate estimation.

Definition 9 (Bivariate cumulative distribution and survival functions). *Let (X, Y) denote a two-dimensional real-valued random vector defined on the probability space $(\Omega, \mathcal{A}, \mathbb{P})$. For any $(x, y) \in \mathbb{R}^2$, we then denote by*

$$\begin{aligned} F^{--}(x, y) &= \mathbb{P}(X < x, Y < y), \\ F(x, y) &= \mathbb{P}(X < x, Y < y) + 1/2 \cdot \mathbb{P}(X < x, Y = y) \\ &\quad + 1/2 \cdot \mathbb{P}(X = x, Y < y) + 1/4 \cdot \mathbb{P}(X = x, Y = y) \end{aligned}$$

the bivariate cumulative distribution function of (X, Y) left-continuous in both arguments and the normalised version respectively. Likewise, we call

$$\begin{aligned} S^{--}(x, y) &= \mathbb{P}(X > x, Y > y), \\ S(x, y) &= \mathbb{P}(X > x, Y > y) + 1/2 \cdot \mathbb{P}(X > x, Y = y) \\ &\quad + 1/2 \cdot \mathbb{P}(X = x, Y > y) + 1/4 \cdot \mathbb{P}(X = x, Y = y) \end{aligned}$$

the bivariate survival function of (X, Y) right-continuous in both arguments and the normalised version respectively.

Definition 10 (Bivariate empirical distribution and survival functions). *Let $(X_1, Y_1), \dots, (X_N, Y_N)$ iid $F(x, y)$ denote a random sample of N paired real-valued observations. For any $(x, y) \in \mathbb{R}^2$, the bivariate empirical distribution and survival functions corresponding to the ones in Definition 9 are then given by*

$$\begin{aligned} \widehat{F}^{--}(x, y) &= \frac{1}{N} \sum_{k=1}^N c^-(x, X_k) c^-(y, Y_k), & \widehat{F}(x, y) &= \frac{1}{N} \sum_{k=1}^N c(x, X_k) c(y, Y_k), \\ \widehat{S}^{--}(x, y) &= \frac{1}{N} \sum_{k=1}^N c^-(X_k, x) c^-(Y_k, y), & \widehat{S}(x, y) &= \frac{1}{N} \sum_{k=1}^N c(X_k, x) c(Y_k, y). \end{aligned}$$

As mentioned before, I will now go on to express the expectations and probabilities used in the definitions of the different versions of Kendall's tau in terms of Lebesgue-Stieltjes integrals.

Proposition 11 (Integral representation). *Let $(X, Y) \sim F(x, y)$ denote a two-dimensional real-valued random vector and $F_X(x)$ as well as $F_Y(y)$ the corresponding univariate marginal distribution functions. Then, with $D_X = \text{sgn}(X_1 - X_2)$ and $D_Y = \text{sgn}(Y_1 - Y_2)$, it holds*

$$\begin{aligned}\mathbb{E}(D_X D_Y) &= 4 \iint_{\mathbb{R}^2} F(x, y) dF(x, y) - 1, \\ \mathbb{E}(D_X^2) &= 2 \int_{\mathbb{R}} F_X^-(x) dF_X(x), \\ \mathbb{E}(D_Y^2) &= 2 \int_{\mathbb{R}} F_Y^-(y) dF_Y(y), \\ \mathbb{P}(D_X D_Y \neq 0) &= 1 - 4 \iint_{\mathbb{R}^2} (F(x, y) - F^{--}(x, y)) dF(x, y).\end{aligned}$$

Proof. See Nowak and Konietzschke (2021). □

Consequently, using shorthand notation once again, we can treat the estimation problem of the different versions of Kendall's tau as one involving the estimands $\iint F dF$, $\int F_X^- dF_X$, $\int F_Y^- dF_Y$, and $\iint (F - F^{--}) dF$ instead. As in the previous section, I suggest a simple plug-in approach to obtain estimators, which gives rise to the following definition.

Definition 12 (Tau probabilities). *Let $(X_1, Y_1), \dots, (X_N, Y_N) \stackrel{iid}{\sim} F(x, y)$ denote a random sample of N paired real-valued observations and $F_X(x)$ as well as $F_Y(y)$ the corresponding univariate marginal distribution functions. Then we denote by*

$$\begin{aligned}\mathbf{p} &= (p_{\ll}, p_{TX}, p_{TY}, p_T)', \text{ where} \\ p_{\ll} &= \iint_{\mathbb{R}^2} F(x, y) dF(x, y), \\ p_{TX} &= \int_{\mathbb{R}} F_X^-(x) dF_X(x), \\ p_{TY} &= \int_{\mathbb{R}} F_Y^-(y) dF_Y(y), \\ p_T &= \iint_{\mathbb{R}^2} (F(x, y) - F^{--}(x, y)) dF(x, y),\end{aligned}$$

the vector of tau probabilities. Moreover, we can estimate \mathbf{p} by

$$\begin{aligned}\hat{\mathbf{p}} &= (\hat{p}_{\ll}, \hat{p}_{TX}, \hat{p}_{TY}, \hat{p}_T)', \text{ where} \\ \hat{p}_{\ll} &= \iint_{\mathbb{R}^2} \hat{F}(x, y) d\hat{F}(x, y), \\ \hat{p}_{TX} &= \int_{\mathbb{R}} \hat{F}_X^-(x) d\hat{F}_X(x), \\ \hat{p}_{TY} &= \int_{\mathbb{R}} \hat{F}_Y^-(y) d\hat{F}_Y(y), \\ \hat{p}_T &= \iint_{\mathbb{R}^2} (\hat{F}(x, y) - \hat{F}^{--}(x, y)) d\hat{F}(x, y).\end{aligned}$$

For example, to reconstruct Kendall's τ_B , it holds $\tau_B = h(\mathbf{p}) = (2p_{\ll} - 0.5) / \sqrt{p_{TX} p_{TY}}$. Moreover, the estimator $\hat{\tau}_B = h(\hat{\mathbf{p}}) = (2\hat{p}_{\ll} - 0.5) / \sqrt{\hat{p}_{TX} \hat{p}_{TY}}$ coincides with the commonly used empirical version (see also Kendall & Gibbons, 1990; Nowak & Konietzschke, 2021). As for τ_A , D_{YX} and γ , I refer the reader to Nowak and Konietzschke (2021) as well. Similar to the Mann-Whitney parameter, we can rely on the multivariate central limit theorem to derive the asymptotic sampling distribution of these nonparametric association measures.

Proposition 13 (Asymptotics of $\hat{\mathbf{p}}$). *Let $(X_1, Y_1), \dots, (X_N, Y_N) \stackrel{iid}{\sim} F(x, y)$ denote a random sample of N paired real-valued observations and $F_X(x)$ as well as $F_Y(y)$ the corresponding univariate marginal*

distribution functions. With $\hat{\mathbf{p}}$ and \mathbf{p} given as in Definition 12 and under some mild regularity conditions, we have

$$\sqrt{N}(\hat{\mathbf{p}} - \mathbf{p}) \xrightarrow[N \rightarrow \infty]{\mathcal{D}} \mathcal{N}_4(\mathbf{0}, \boldsymbol{\Sigma}),$$

where $\boldsymbol{\Sigma}$ can be consistently estimated by $\hat{\boldsymbol{\Sigma}} = \frac{1}{N-1} \sum_{k=1}^N (\hat{\mathbf{u}}_k - \bar{\hat{\mathbf{u}}})(\hat{\mathbf{u}}_k - \bar{\hat{\mathbf{u}}})'$, with

$$\begin{aligned} \hat{\mathbf{u}}_k &= (\hat{U}_k, \hat{G}_k, \hat{H}_k, \hat{V}_k)', \quad k \in \{1, \dots, N\}, \quad \text{where} \\ \hat{U}_k &= \hat{F}(X_k, Y_k) + \hat{S}(X_k, Y_k), \\ \hat{G}_k &= \hat{F}_X^-(X_k) + \hat{S}_X^-(X_k), \\ \hat{H}_k &= \hat{F}_Y^-(Y_k) + \hat{S}_Y^-(Y_k), \\ \hat{V}_k &= \hat{F}(X_k, Y_k) + \hat{S}(X_k, Y_k) - \hat{F}^{--}(X_k, Y_k) - \hat{S}^{--}(X_k, Y_k). \end{aligned}$$

Proof. See Nowak and Konietzschke (2021). \square

Corollary 14 (Asymptotics of $\hat{\tau}_B$). With $h(x, y, z, \cdot) = (2x-0.5)/\sqrt{yz}$ and under some mild regularity conditions, it holds

$$\sqrt{N}(h(\hat{\mathbf{p}}) - h(\mathbf{p})) = \sqrt{N}(\hat{\tau}_B - \tau_B) \xrightarrow[N \rightarrow \infty]{\mathcal{D}} \mathcal{N}(0, \sigma_h^2),$$

where σ_h^2 can be consistently estimated by $\hat{\sigma}_h^2 = \nabla h(\hat{\mathbf{p}})' \cdot \hat{\boldsymbol{\Sigma}} \cdot \nabla h(\hat{\mathbf{p}})$.

Proof. Straightforward application of the delta method implies the desired result (see e.g. van der Vaart, 1998). \square

One can derive asymptotic results for Somers' D and Goodman and Kruskal's γ in a completely analogous manner. As for Kendall's τ_A , some care needs to be taken (Nowak & Konietzschke, 2021). As I did in the previous section as regards the Mann-Whitney parameter, I will now informally justify the result in Proposition 13 using similar arguments as before. To this end, consider

$$\begin{aligned} \hat{p}_{\ll} - p_{\ll} &= \iint \hat{F} d\hat{F} - p_{\ll} \\ &= \iint \hat{F} d\hat{F} - p_{\ll} + \underbrace{\iint \hat{F} dF - \iint \hat{F} dF}_{=0} + \underbrace{\iint F d\hat{F} - \iint F d\hat{F}}_{=0} + \underbrace{\iint F dF - p_{\ll}}_{=0} \\ &= \iint F d\hat{F} + \iint \hat{F} dF - 2p_{\ll} + \left(\iint \hat{F} d\hat{F} - \iint F d\hat{F} - \left(\iint \hat{F} dF - \iint F dF \right) \right) \\ &= \iint F d\hat{F} + \iint S d\hat{F} - 2p_{\ll} + \left(\iint (\hat{F} - F) d\hat{F} - \iint (\hat{F} - F) dF \right) \\ &= \iint (F + S) d\hat{F} - 2p_{\ll} + \iint (\hat{F} - F) d(\hat{F} - F) \\ &= \frac{1}{N} \sum_{k=1}^N (F(X_k, Y_k) + S(X_k, Y_k)) - 2p_{\ll} + \iint (\hat{F} - F) d(\hat{F} - F). \end{aligned}$$

Multiplication of both sides by the square root of the sample size yields

$$\begin{aligned} \sqrt{N}(\hat{p}_{\ll} - p_{\ll}) &= \sqrt{N} \left(\frac{1}{N} \sum_{k=1}^N \underbrace{(F(X_k, Y_k) + S(X_k, Y_k))}_{=: U_k} - 2p_{\ll} \right) + \underbrace{\sqrt{N} \iint (\hat{F} - F) d(\hat{F} - F)}_{=: B_N}. \end{aligned}$$

Since each U_k is only random through (X_k, Y_k) , the first summand is nothing but a sum of independent and identically distributed random variables, whereas B_N converges to zero in probability as N tends to infinity. Therefore, $\sqrt{N}(\hat{p}_{\ll} - p_{\ll})$ and $\sqrt{N}(\frac{1}{N} \sum_{k=1}^N U_k - 2p_{\ll})$ share the same limiting distribution so that we can treat $\mathbb{V}(U_1)$ as our variance estimand.

By the same token, I will now turn to the probabilities of no ties occurring. More specifically, as to \hat{p}_{TX} and p_{TX} , we have

$$\begin{aligned}
& \hat{p}_{TX} - p_{TX} \\
&= \int \hat{F}_X^- d\hat{F}_X - p_{TX} \\
&= \int \hat{F}_X^- d\hat{F}_X - p_{TX} + \underbrace{\int \hat{F}_X^- dF_X - \int \hat{F}_X^- dF_X}_{=0} + \underbrace{\int F_X^- d\hat{F}_X - \int F_X^- d\hat{F}_X}_{=0} + \underbrace{\int F_X^- dF_X - p_{TX}}_{=0} \\
&= \int F_X^- d\hat{F}_X + \int \hat{F}_X^- dF_X - 2p_{TX} + \left(\int \hat{F}_X^- d\hat{F}_X - \int F_X^- d\hat{F}_X - \left(\int \hat{F}_X^- dF_X - \int F_X^- dF_X \right) \right) \\
&= \int F_X^- d\hat{F}_X + \int S_X^- d\hat{F}_X - 2p_{TX} + \left(\int (\hat{F}_X^- - F_X^-) d\hat{F}_X - \int (\hat{F}_X^- - F_X^-) dF_X \right) \\
&= \int (F_X^- + S_X^-) d\hat{F}_X - 2p_{TX} + \int (\hat{F}_X^- - F_X^-) d(\hat{F}_X - F_X) \\
&= \frac{1}{N} \sum_{k=1}^N (F_X^-(X_k) + S_X^-(X_k)) - 2p_{TX} + \int (\hat{F}_X^- - F_X^-) d(\hat{F}_X - F_X).
\end{aligned}$$

Multiplication of both sides by the square root of the sample size produces

$$\begin{aligned}
& \sqrt{N}(\hat{p}_{TX} - p_{TX}) \\
&= \sqrt{N} \left(\frac{1}{N} \sum_{k=1}^N \underbrace{(F_X^-(X_k) + S_X^-(X_k))}_{=:G_k} - 2p_{TX} \right) + \underbrace{\sqrt{N} \int (\hat{F}_X^- - F_X^-) d(\hat{F}_X - F_X)}_{=:B_N}.
\end{aligned}$$

Again, each G_k is only random through X_k , while $B_N \xrightarrow{\mathbb{P}} 0$ as $N \rightarrow \infty$. Thus, $\sqrt{N}(\hat{p}_{TX} - p_{TX})$ and $\sqrt{N}(\frac{1}{N} \sum_{k=1}^N G_k - 2p_{TX})$ share the same limiting distribution and we can regard $\mathbb{V}(G_1)$ as our variance estimand.

As for $\sqrt{N}(\hat{p}_{TY} - p_{TY})$, we have $\mathbb{V}(H_1)$ with $H_1 = F_Y^-(Y_1) + S_Y^-(Y_1)$. In a similar vein, since $\sqrt{N}(\hat{p}_T - p_T) = \sqrt{N}(\hat{p}_{\ll} - p_{\ll}) - \sqrt{N}(\iint \hat{F}^- d\hat{F} - \iint F^- dF)$, we define the corresponding variance estimand as $\mathbb{V}(V_1)$, where $V_1 = F(X_1, Y_1) + S(X_1, Y_1) - F^--(X_1, Y_1) - S^--(X_1, Y_1)$. More importantly, these asymptotically equivalent statistics $\mathbf{u}_1 = (U_1, G_1, H_1, V_1)'$ readily lend themselves to joint estimation. Indeed, the theoretical asymptotic covariance matrix of the tau probability estimator $\hat{\mathbf{p}}$ is simply $\Sigma = \mathbb{E}(\tilde{\mathbf{u}}_1 \tilde{\mathbf{u}}_1')$, where $\tilde{\mathbf{u}}_1 = \mathbf{u}_1 - \mathbb{E}(\mathbf{u}_1)$. In that regard, one can also easily find the asymptotic covariance matrix of tau probabilities arising from general multivariate distributions as well. For instance, supposing the two subscripts (1) and (2) refer to two $\hat{\mathbf{p}}$ vectors, i.e., $\hat{\mathbf{p}}_{(1)}$ and $\hat{\mathbf{p}}_{(2)}$, assessing the association of a disease progression score and a subjective pain scale in the same patients at time points 1 and 2, respectively. With $\mathbf{u}_{1;(1)}$ and $\mathbf{u}_{1;(2)}$ denoting the corresponding asymptotically equivalent statistics, yielding $\mathbf{w}_1 = (\mathbf{u}'_{1;(1)}, \mathbf{u}'_{1;(2)})'$, we can then express the asymptotic covariance matrix of the estimator $(\hat{\mathbf{p}}'_{(1)}, \hat{\mathbf{p}}'_{(2)})'$ by $\Sigma = \mathbb{E}(\tilde{\mathbf{w}}_1 \tilde{\mathbf{w}}_1')$, where $\tilde{\mathbf{w}}_1 = \mathbf{w}_1 - \mathbb{E}(\mathbf{w}_1)$.

The next section concerns joint estimation as well, but with respect to the Mann-Whitney parameter in the context of accumulating data.

Group sequential methodology for the Mann-Whitney parameter

Group sequential methods address the problem of multiplicity issues arising in clinical trials when performing repeated significance tests as to the same hypothesis on accumulating data. Following the

exposition of Jennison and Turnbull (2000), the key tool is the so-called *canonical joint distribution*. In that regard, consider the following somewhat informal definition.

Definition 15 (Canonical joint distribution). *Suppose a group sequential study with up to K analyses yields the sequence of test statistics $\{Z_1, \dots, Z_K\}$. We say these statistics follow the canonical joint distribution with information levels $\{\mathcal{I}_1, \dots, \mathcal{I}_K\}$ for the parameter θ if*

- (Z_1, \dots, Z_K) is multivariate normal,
- $\mathbb{E}(Z_k) = \theta\sqrt{\mathcal{I}_k}$, $k = \{1, \dots, K\}$, and
- $\text{Cov}(Z_{k_1}, Z_{k_2}) = \sqrt{\mathcal{I}_{k_1}/\mathcal{I}_{k_2}}$, $1 \leq k_1 \leq k_2 \leq K$.

For readers unfamiliar with the concept of information levels, one may generally think of them as the inverse variance of the estimator of the parameter θ . Now, assuming we wished to test $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$, then, the higher the information levels, the closer to the null value the estimate of θ can be while still giving rise to rejection of the null hypothesis. In the context of survival analysis, θ may refer to a log hazard ratio and $\{Z_1, \dots, Z_K\}$ to log-rank tests. If $\{Z_1, \dots, Z_K\}$ stand for two sample t tests, then θ is a theoretical mean difference. To better illustrate the canonical joint distribution, I will now present in more detail the arguably simplest example as given in Jennison and Turnbull (2000).

Example 16 (One sample normal mean). *Let $X_i \stackrel{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$, $i = 1, 2, \dots$ denote the responses of interest. Suppose we wish to test the hypothesis $H_0 : \mu = \mu_0$ and that σ^2 is known. With n_k denoting the cumulative number of observations available at analysis $k \in \{1, \dots, K\}$, we can estimate μ by*

$$\bar{X}^{(k)} = \frac{1}{n_k} \sum_{i=1}^{n_k} X_i \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n_k}\right)$$

and define $\mathcal{I}_k = \{\mathbb{V}(\bar{X}^{(k)})\}^{-1} = n_k/\sigma^2$ the information for $\theta = \mu - \mu_0$ at analysis k . The standardised test statistics are then given by $Z_k = (\bar{X}^{(k)} - \mu_0)\sqrt{\mathcal{I}_k}$. Each Z_k being a linear combination of the independent normal variates X_i , the vector (Z_1, \dots, Z_K) is multivariate normal. Marginally,

$$Z_k \sim \mathcal{N}(\theta\sqrt{\mathcal{I}_k}, 1), \quad k \in \{1, \dots, K\},$$

Finally, for $k_1 \leq k_2$, we have

$$\begin{aligned} \text{Cov}(Z_{k_1}, Z_{k_2}) &= \text{Cov}(\{\bar{X}^{(k_1)} - \mu_0\}\sqrt{\mathcal{I}_{k_1}}, \{\bar{X}^{(k_2)} - \mu_0\}\sqrt{\mathcal{I}_{k_2}}) \\ &= \sqrt{\mathcal{I}_{k_1}}\sqrt{\mathcal{I}_{k_2}} \frac{1}{n_{k_1}} \frac{1}{n_{k_2}} \sum_{i=1}^{n_{k_1}} \sum_{j=1}^{n_{k_2}} \text{Cov}(X_i, X_j) \\ &= \sqrt{\mathcal{I}_{k_1}}\sqrt{\mathcal{I}_{k_2}} \frac{1}{n_{k_1}} \frac{1}{n_{k_2}} n_{k_1} \sigma^2 = \sqrt{\mathcal{I}_{k_1}}\sqrt{\mathcal{I}_{k_2}}(\mathcal{I}_{k_2})^{-1} = \sqrt{\mathcal{I}_{k_1}/\mathcal{I}_{k_2}}. \end{aligned}$$

Thus $\{Z_1, \dots, Z_K\}$ follow the canonical joint distribution with information levels $\{\mathcal{I}_1, \dots, \mathcal{I}_K\}$ for the parameter $\theta = \mu - \mu_0$.

As might be expected, if one replaces σ^2 with a consistent estimator, the resulting standardised statistics would still asymptotically follow the canonical joint distribution. Moreover, one can also drop the normality assumption, which is the subject of the next proposition. However, to avoid running the risk that a simple application of the multivariate central limit theorem gets drowned in unnecessarily complex notation, I will only consider an equally spaced two stage trial. Extending this result to the general case of multi-stage trials with potentially unequal spacing should be straightforward.

Proposition 17 (Asymptotics of the one sample mean in an equally spaced two stage trial). *Let $(X_\ell)_{\ell \geq 1}$ denote a sequence of independent and identically distributed real-valued random variables with $\mu = \mathbb{E}(X_1)$ and $\sigma^2 = \mathbb{V}(X_1) < \infty$. Let further denote by $n \in \mathbb{N}$ the first stage sample size and consider $(Y_\ell)_{\ell \geq 1}$, where $Y_\ell = 1/2 \cdot (X_\ell + X_{n+\ell})$. Defining the first and second stage sample means, i.e., $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$ and $\bar{Y}_n = \frac{1}{n} \sum_{i=1}^n Y_i = \frac{1}{2n} \sum_{i=1}^{2n} X_i$, it then holds*

$$\left(\begin{array}{c} \sqrt{n}(\bar{X}_n - \mu) \\ \sqrt{2n}(\bar{Y}_n - \mu) \end{array} \right) \xrightarrow[n \rightarrow \infty]{\mathcal{D}} \mathcal{N}_2(\mathbf{0}, \mathbf{\Sigma}), \quad \mathbf{\Sigma} = \sigma^2 \cdot \left(\begin{array}{cc} 1 & \sqrt{1/2} \\ \sqrt{1/2} & 1 \end{array} \right).$$

Proof. The random vectors $(X_1, Y_1), \dots, (X_n, Y_n)$ being independent and identically distributed with $\mathbb{V}(X_1) = \sigma^2$ and $\mathbb{V}(Y_1) = \text{Cov}(X_1, Y_1) = 1/2 \cdot \sigma^2$, the result follows directly from the multivariate central limit theorem / Cramér-Wold device (see e.g. van der Vaart, 1998; Cramér & Wold, 1936). \square

Therefore, we can also make use of the canonical joint distribution when considering the Mann-Whitney parameter. By Proposition 7 and the ensuing discussion of A_N at the end of section , we can easily think of $(S_2(X_{1\ell}))_{\ell \geq 1}$ as well as $(F_1(X_{2\ell}))_{\ell \geq 1}$ as the sequence of independent and identically distributed random variables denoted by $(X_\ell)_{\ell \geq 1}$ in Proposition 17. The following proposition intends to give a somewhat more precise account of this result.

Proposition 18 (Group sequential asymptotics regarding the Mann-Whitney parameter). *Let $X_{1i} \stackrel{iid}{\sim} F_1(x)$, $i = 1, 2, \dots$, and $X_{2j} \stackrel{iid}{\sim} F_2(x)$, $j = 1, 2, \dots$, denote two independent random samples of real-valued observations and suppose we wish to test the null hypothesis $H_0 : p = 1/2$. With n_{1k} and n_{2k} as well as $\widehat{F}_1^{(k)}$ and $\widehat{F}_2^{(k)}$ denoting the corresponding cumulative sample sizes and empirical distribution functions at analysis $k \in \{1, \dots, K\}$, we can estimate the Mann-Whitney parameter p by*

$$\widehat{p}^{(k)} = \int \widehat{F}_1^{(k)} d\widehat{F}_2^{(k)} = \frac{1}{n_{1k}} \frac{1}{n_{2k}} \sum_{j=1}^{n_{2k}} \sum_{i=1}^{n_{1k}} c(X_{2j}, X_{1i}).$$

Moreover, with $\mathcal{I}_k = \{(\int S_2^2 dF_1 - p^2)/n_{1k} + (\int F_1^2 dF_2 - p^2)/n_{2k}\}^{-1}$ and $Z_k = (\widehat{p}^{(k)} - 1/2)\sqrt{\mathcal{I}_k}$ denoting the information and the resulting standardised statistic at analysis $k \in \{1, \dots, K\}$, we have that the sequence of statistics $\{Z_1, \dots, Z_K\}$ asymptotically follow the canonical joint distribution with information levels $\{\mathcal{I}_1, \dots, \mathcal{I}_K\}$ for the parameter $\theta = p - 1/2$.

Proof. See Nowak et al. (2022a). \square

The information levels just given are the ones associated with the Brunner-Munzel test (2000) and can be consistently estimated by

$$\widehat{\mathcal{I}}_k = \left(\frac{\int (\widehat{S}_2^{(k)})^2 d\widehat{F}_1^{(k)} - (\widehat{p}^{(k)})^2}{n_{1k} - 1} + \frac{\int (\widehat{F}_1^{(k)})^2 d\widehat{F}_2^{(k)} - (\widehat{p}^{(k)})^2}{n_{2k} - 1} \right)^{-1}.$$

Since the Brunner-Munzel test is known to be too liberal in small samples, one may prefer a version of the test based on a *logit* transformed Mann-Whitney parameter, referred to as *log win odds test* by Nowak et al. (2022a). On the other hand, one may wish to employ the information levels associated with the Wilcoxon-Mann-Whitney test. In that case, however, the resulting standardised statistics asymptotically follow the canonical joint distribution only if both distributions coincide, i.e., if $F_1 = F_2$, see Nowak et al. (2022a). Consequently, we cannot invert the Wilcoxon-Mann-Whitney test to produce repeated confidence intervals.

With the aid of Proposition 18, I will now provide an approximate power formula for this group sequential setting.

Proposition 19 (Power regarding the Mann-Whitney parameter in a group sequential design). *Let $X_{1i} \stackrel{iid}{\sim} F_1(x)$, $i = 1, 2, \dots$, and $X_{2j} \stackrel{iid}{\sim} F_2(x)$, $j = 1, 2, \dots$, denote two independent random samples of real-valued observations, where n_{1k} and n_{2k} are the corresponding cumulative sample sizes available at analysis $k \in \{1, \dots, K\}$, $N_k = n_{1k} + n_{2k}$. Assuming a constant sample size ratio $t = n_{1k}/N_k$ for all stages k , the information levels are then given by*

$$\mathcal{I}_k = \frac{N_k t(1-t)}{(1-t) \int S_2^2 dF_1 + t \int F_1^2 dF_2 - p^2}, \quad k \in \{1, \dots, K\}.$$

To test the hypothesis $H_0 : p \leq 1/2$ against $H_1 : p > 1/2$ at a global one-sided nominal significance level of α , let c_1, \dots, c_K denote the critical values computed from a K -variate normal distribution with mean vector $\mathbf{0}$ and covariance matrix $\mathbf{R} = (r_{ij})_{i,j=1,\dots,K}$, $r_{ij} = \sqrt{N_{\min(k_i, k_j)}/N_{\max(k_i, k_j)}}$, and error spending function of choice. The approximate power is then given by

$$\text{Power} \approx 1 - \Phi_{\mathbf{R}} \left(c_1 - \sqrt{\mathcal{I}_1} \cdot (p - 1/2), \dots, c_K - \sqrt{\mathcal{I}_K} \cdot (p - 1/2) \right),$$

where $\Phi_{\mathbf{R}}$ denotes the cumulative distribution function of a K -variate normal distribution with mean vector $\mathbf{0}$ and covariance matrix \mathbf{R} as just specified.

Proof. See Nowak et al. (2022a). □

Similar to before, Nowak et al. (2022a) give approximate power formulas for the log win odds and Wilcoxon-Mann-Whitney tests as well.

With simulation results indicating that all three tests, i.e., Brunner-Munzel, log win odds, and Wilcoxon-Mann-Whitney, suffer from certain drawbacks, the last manuscript addressing the nonparametric Behrens-Fisher problem in small samples (Nowak et al., 2022b) further explores different ways to approximate the sampling distribution of the Mann-Whitney parameter. In that regard, the manuscript reviews variance estimators proposed by Bamber (1975), DeLong, DeLong, and Clarke-Pearson (1988), Perme and Manevski (2019), Brunner, Happ, and Friedrich (2021), as well as Shirahata (1993). In addition, the manuscript discusses a permutation method developed by Pauly, Asendorf, and Konietzschke (2016). It turns out that the variance estimator of Perme and Manevski (2019) together with a t approximation and a particular choice of degrees freedom performs best in terms of maintaining the nominal significance level. However, I arrived at this conclusion by a somewhat crude trial and error approach based on simulations rather than mathematical arguments or derivations.

References

- Bamber, D. (1975). The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of Mathematical Psychology*, 12(4), 387-415.
- Brunner, E., Bathke, A. C., & Konietschke, F. (2018). *Rank and pseudo-rank procedures for independent observations in factorial designs*. Springer.
- Brunner, E., Happ, M., & Friedrich, S. (2021). *Erwartungstreuer Schätzer für $\text{Var}(\hat{\theta})$ und Konfidenzintervalle für θ* . Unpublished manuscript.
- Brunner, E., & Munzel, U. (2000). The nonparametric Behrens-Fisher problem: Asymptotic theory and a small-sample approximation. *Biometrical Journal*, 42, 17-25.
- Cramér, H., & Wold, H. (1936). Some Theorems on Distribution Functions. *Journal of the London Mathematical Society*, s1-11(4), 290-294.
- DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, 44, 837-845.
- Goodman, L. A., & Kruskal, W. H. (1954). Measures of association for cross classifications. *Journal of the American Statistical Association*, 49(268), 732-764.
- Jennison, C., & Turnbull, B. W. (2000). *Group sequential methods with applications to clinical trials*. Chapman & Hall/CRC.
- Kendall, M. G., & Gibbons, J. D. (1990). *Rank correlation methods* (5th ed.). A Charles Griffin Title.
- Kruskal, W. H. (1958). Ordinal measures of association. *Journal of the American Statistical Association*, 53(284), 814-861.
- Nowak, C. P. (2019). *Test procedures and confidence intervals for Kendall's tau*. Unpublished master's thesis. Humboldt University of Berlin.
- Nowak, C. P., & Konietschke, F. (2021). Simultaneous inference for Kendall's tau. *Journal of Multivariate Analysis*, 185, 104767. doi: 10.1016/j.jmva.2021.104767.
- Nowak, C. P., Mütze, T., & Konietschke, F. (2022a). Group sequential methods for the Mann-Whitney parameter. *Statistical Methods in Medical Research*. doi: 10.1177/09622802221107103.
- Nowak, C. P., Pauly, M., & Brunner, E. (2022b). The nonparametric Behrens-Fisher problem in small samples. arXiv:2208.01231v1 [stat.ME].
- 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.
- Perme, M. P., & Manevski, D. (2019). Confidence intervals for the Mann-Whitney test. *Statistical Methods in Medical Research*, 28(12), 3755-3768.
- Shirahata, S. (1993). Estimate of variance of Wilcoxon-Mann-Whitney statistic. *Journal of the Japanese Society of Computational Statistics*, 6(2), 1-10.
- Somers, R. H. (1962). A new asymmetric measure of association for ordinal variables. *American Sociological Review*, 27(6), 799-811.
- Thas, O., De Neve, J., Clement, L., & Ottoy, J.-P. (2012). Probabilistic index models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74, 623-671.
- van der Vaart, A. W. (1998). *Asymptotic statistics*. Cambridge University Press.

Manuscripts

The following three manuscripts are part of this cumulative dissertation.

- Nowak, C. P., & Konietschke, F. (2021). Simultaneous inference for Kendall's tau. *Journal of Multivariate Analysis*, 185, 104767. doi: 10.1016/j.jmva.2021.104767
- Nowak, C. P., Mütze, T., & Konietschke, F. (2022a). Group sequential methods for the Mann-Whitney parameter. *Statistical Methods in Medical Research*. doi 10.1177/09622802221107103.
- Nowak, C. P., Pauly, M., & Brunner, E. (2022b). The nonparametric Behrens-Fisher problem in small samples. arXiv:2208.01231v1 [stat.ME].

The manuscript on Kendall's tau (Nowak & Konietschke, 2021) is only included in the physical copies of this dissertation.



Group sequential methods for the Mann-Whitney parameter

Statistical Methods in Medical Research
1–17

© The Author(s) 2022



Article reuse guidelines:

sagepub.com/journals-permissions

DOI: 10.1177/09622802221107103

journals.sagepub.com/home/smm



Claus P Nowak^{1,2}, Tobias Mütze³ , and Frank Konietzschke¹

Abstract

Late phase clinical trials are occasionally planned with one or more interim analyses to allow for early termination or adaptation of the study. While extensive theory has been developed for the analysis of ordered categorical data in terms of the Wilcoxon-Mann-Whitney test, there has been comparatively little discussion in the group sequential literature on how to provide repeated confidence intervals and simple power formulas to ease sample size determination. Dealing more broadly with the nonparametric Behrens-Fisher problem, we focus on the comparison of two parallel treatment arms and show that the Wilcoxon-Mann-Whitney test, the Brunner-Munzel test, as well as a test procedure based on the log win odds, a modification of the win ratio, asymptotically follow the canonical joint distribution. In addition to developing power formulas based on these results, simulations confirm the adequacy of the proposed methods for a range of scenarios. Lastly, we apply our methodology to the FREEDOMS clinical trial (ClinicalTrials.gov Identifier: NCT00289978) in patients with relapse-remitting multiple sclerosis.

Keywords

Brunner-Munzel test, error spending, group sequential methods, nonparametric relative effect, Wilcoxon-Mann-Whitney test, win odds

1 Introduction

Since it is not uncommon for phase III clinical trials to run for a number of years, there is much interest in being able to assess safety and efficacy while the trial is still ongoing. Unsurprisingly, regulatory authorities (EMA,¹ FDA²) point out the need to adequately address multiplicity issues and give practical guidance on group sequential methods, which allow for repeated significance testing on accumulating data without inflating the nominal overall type I error rate.

While standard textbooks such as Jennison and Turnbull³, Proschan,⁴ or Wassmer and Brannath⁵ primarily discuss continuous, binary and survival endpoints, the Wilcoxon-Mann-Whitney test^{6–8} has also been extended to group sequential settings.^{9–11} In our view, the estimand most naturally associated with the Wilcoxon-Mann-Whitney test is the probability

$$p = \mathbb{P}(X_1 < X_2) + 1/2 \cdot \mathbb{P}(X_1 = X_2),$$

where $X_1 \sim F_1$ and $X_2 \sim F_2$ denote two independent random variables. The quantity p is called nonparametric relative effect of X_2 with respect to X_1 , probabilistic index or Mann-Whitney parameter.^{12–15} Dividing p by its complement produces

$$p/(1-p),$$

the so-called win odds.¹⁶ Adding half of the probability of equal outcomes to $\mathbb{P}(X_1 < X_2)$ neatly aligns with Putter's generalisation¹⁷ of the Wilcoxon-Mann-Whitney test to the case of ties. By the same token, Brunner et al.¹⁶ regard the win

¹Charité – Universitätsmedizin Berlin, Freie Universität Berlin and Humboldt-Universität zu Berlin, Institute of Biometry and Clinical Epidemiology, Berlin, Germany

²TU Dortmund University, Faculty of Statistics, Dortmund, Germany

³Statistical Methodology, Novartis Pharma AG, Basel, Switzerland

Corresponding author:

Frank Konietzschke, Charité – Universitätsmedizin Berlin, Institute of Biometry and Clinical Epidemiology, Charitéplatz 1, 10117 Berlin, Germany.
Email: frank.konietzschke@charite.de

odds to be a tie corrected version of the win ratio $\mathbb{P}(X_1 < X_2)/\mathbb{P}(X_1 > X_2)$, which has recently attracted attention in the context of time-to-event data,¹⁸ continuous endpoints,¹⁹ and stratification.²⁰ Of course, if tied values cannot occur almost surely, that is, if $\mathbb{P}(X_1 = X_2) = 0$, then p equals $\mathbb{P}(X_1 < X_2)$ and the win odds coincide with the win ratio.

To illustrate the interpretation of the nonparametric relative effect p , let us assume that X_1 and X_2 refer to outcomes from treatment arms 1 and 2, respectively, and that lower values point to a more favourable outcome. Then p is nothing but the probability that patients on arm 1 will fare better than those on arm 2, including $1/2$ times the probability of equal outcomes. Perhaps a little easier to interpret are the win odds. For instance, if $p = 0.75$, then the odds that a patient on arm 1 will fare better than one on arm 2 are $3 : 1$, with the possibility of equal outcomes equally allocated to the ‘fare better’ and ‘fare worse’ scenarios.

However, asymptotic results of the Wilcoxon-Mann-Whitney test as commonly employed are only valid if both distributions coincide, that is, if $F_1 = F_2$. Hence the null hypothesis is usually formulated in terms of the distribution functions as well, that is, $H_0 : F_1 = F_2$ and not the Mann-Whitney parameter p as such. While $F_1 = F_2$ implies $p = 1/2$, the reverse does not hold. For instance, any two symmetric distributions with the same centre of symmetry, such as two normal distributions $\mathcal{N}(0, 1)$ and $\mathcal{N}(0, 4)$, would imply $p = 1/2$. In essence, the nonparametric Behrens-Fisher problem addresses the testing problem $H_0 : p = 1/2$, while making no further assumptions on F_1 and F_2 , which is precisely the scenario that the Brunner-Munzel test¹² was developed to deal with. In that regard, unlike the Wilcoxon-Mann-Whitney test, the limiting distribution of the Brunner-Munzel test is normal with unit variance under both the null and the alternative hypotheses, thus allowing for test inversion and computation of confidence intervals for p , which in turn facilitates the derivation of simple power approximations in the group sequential setting.

A key tool in group sequential theory which we will also rely on here is the so-called *canonical joint distribution*.^{3–5,21} More precisely, a sequence of K test statistics $\{Z_1, \dots, Z_K\}$ with information levels $\{\mathcal{I}_1, \dots, \mathcal{I}_K\}$ for a single parameter θ are said to follow the *canonical joint distribution* if

- (i) $\mathbf{Z} = (Z_1, \dots, Z_K)$ follows a multivariate normal distribution,
- (ii) $\mathbb{E}(Z_k) = \theta\sqrt{\mathcal{I}_k}$, $k = 1, \dots, K$,
- (iii) $\text{Cov}(Z_{k_1}, Z_{k_2}) = \sqrt{\mathcal{I}_{k_1}\mathcal{I}_{k_2}}$, $1 \leq k_1 \leq k_2 \leq K$.

As might be expected, group sequential versions of the nonparametric tests just discussed follow the canonical joint distribution only asymptotically, which is why we will check its applicability for finite sample sizes by way of extensive simulations.

This paper is organised as follows. Section 2 introduces notation and group sequential methods for hypothesis tests based on the nonparametric relative effect p , with derivations concerning the covariance structure of the corresponding group sequential statistics \mathbf{Z} referred to the appendix. Following a discussion on error spending in Section 3, we set out results from simulation studies in Section 4 to assess type I error rates for finite sample sizes. Section 5 deals with the retrospective application of our proposed methodology to a completed clinical trial, whereas Section 6 outlines how to plan a group sequential trial with the aid of simple approximate power formulas. More detailed results and technical considerations regarding the simulations are provided in the Supplemental Material.

2 Nonparametric group sequential models

We start with notation from nonparametric theory necessary to develop group sequential models for the Wilcoxon-Mann-Whitney test, the Brunner-Munzel test and a *logit* transformed version of the latter, which we refer to as the log win odds test. With the asymptotic normality of the test statistics at issue already established for the fixed sample size scenario, a vector \mathbf{Z} of such statistics based on accumulating groups of data is asymptotically multivariate normal by the Crámer-Wold theorem.²² Thus, in order to obtain the asymptotic joint distribution, it remains to properly define the information levels and derive the expectation and covariance matrix of \mathbf{Z} .

2.1 Notation

Let X be a univariate random variable representing real-valued or ordered categorical data, defined on the probability space $(\Omega, \mathcal{A}, \mathbb{P})$. Adopting common notation, we denote by

$$\begin{aligned} F^-(x) &= \mathbb{P}(X < x) \quad \text{the left-continuous,} \\ F^+(x) &= \mathbb{P}(X \leq x) \quad \text{the right-continuous,} \\ F(x) &= \mathbb{P}(X < x) + 1/2 \cdot \mathbb{P}(X = x) \quad \text{the normalised} \end{aligned}$$

version of the cumulative distribution function of X .^{23,24,12}

Now suppose we have a sample of observations $X_1, \dots, X_n \stackrel{iid}{\sim} F$. Then we call

$$\widehat{F}(x) = \frac{1}{n} \sum_{j=1}^n c(x, X_j), \quad c(x, X_j) = \begin{cases} 0 & \text{if } x < X_j \\ 1/2 & \text{if } x = X_j \\ 1 & \text{if } x > X_j \end{cases}$$

the normalised version of the empirical cumulative distribution function. Moreover,

$$R_i = 1/2 + \sum_{j=1}^n c(X_i, X_j), \quad i = 1, \dots, n,$$

denotes the mid-rank of X_i among the observations X_1, \dots, X_n .

For two independent random variables $X_1 \sim F_1$ and $X_2 \sim F_2$, the probability

$$p = \mathbb{P}(X_1 < X_2) + 1/2 \cdot \mathbb{P}(X_1 = X_2) = \int F_1 dF_2$$

is called nonparametric relative effect of X_2 with respect to X_1 (or of F_2 with respect to F_1). We say that

- X_1 tends to smaller values than X_2 if $p > 1/2$,
- X_1 tends to larger values than X_2 if $p < 1/2$,
- X_1 and X_2 are stochastically comparable if $p = 1/2$.

For a more comprehensive discussion on nonparametric theory we refer to Brunner et al.¹³

Throughout the remainder of this paper we will focus on a parallel two-arm clinical trial and consider accumulating responses

$$X_{1i} \stackrel{iid}{\sim} F_1, \quad i = 1, 2, \dots,$$

$$X_{2j} \stackrel{iid}{\sim} F_2, \quad j = 1, 2, \dots,$$

from treatment arms 1 and 2, respectively. Apart from assuming that $0 < p < 1$ and that there exists no x such that $\mathbb{P}(X_{1i} = x) = 1$ or $\mathbb{P}(X_{2j} = x) = 1$, which excludes the degenerate cases of completely separated samples and one-point distributions, F_1 and F_2 are otherwise arbitrary.

With n_{1k} and n_{2k} denoting the cumulative number of observations available at analysis $k = 1, \dots, K$ for the respective treatments, $N_k = n_{1k} + n_{2k}$, we can estimate the nonparametric relative effect p by

$$\widehat{p}^{(k)} = \int \widehat{F}_1^{(k)} d\widehat{F}_2^{(k)} = \frac{1}{n_{1k}} \frac{1}{n_{2k}} \sum_{j=1}^{n_{2k}} \sum_{i=1}^{n_{1k}} c(X_{2j}, X_{1i}) = \frac{1}{N_k} (\bar{R}_{2\bullet}^{(k)} - \bar{R}_{1\bullet}^{(k)}) + 1/2,$$

with $\bar{R}_{g\bullet}^{(k)} = \frac{1}{n_{gk}} \sum_{i=1}^{n_{gk}} R_{gi}^{(k)}$, where $R_{gi}^{(k)}$ is the mid-rank of X_{gi} among all observations

$$X_{11}, \dots, X_{1n_{1k}}, X_{21}, \dots, X_{2n_{2k}}$$

available at analysis k ; $g = 1, 2$; $i = 1, \dots, n_{gk}$.

For asymptotic results, we let both sample sizes tend to infinity such that neither vanishes, that is, $n_{gk}/N_k \rightarrow \gamma_g > 0$ for both $n_{1k} \rightarrow \infty$ and $n_{2k} \rightarrow \infty$, $g = 1, 2$.

2.2 Wilcoxon-Mann-Whitney test allowing for ties

To test the hypothesis $H_0 : F_1 = F_2$ against $H_1 : F_1 \neq F_2$, we employ at each interim analysis k the same test statistic as in the fixed design, namely

$$\widehat{Z}_k = (\widehat{p}^{(k)} - 1/2) \sqrt{\widehat{\mathcal{I}}_k}, \quad k = 1, \dots, K, \quad (1)$$

with estimated information $\widehat{\mathcal{I}}_k = (N_k n_{1k} n_{2k}) / \widehat{\sigma}_{Rk}^2$, where

$$\widehat{\sigma}_{Rk}^2 = \frac{1}{N_k - 1} \sum_{g=1}^2 \sum_{i=1}^{n_{gk}} \left(R_{gi}^{(k)} - \frac{N_k + 1}{2} \right)^2, \quad k = 1, \dots, K.$$

It is well known that each \widehat{Z}_k converges in distribution to a standard normal random variate, provided the null hypothesis is true.¹³

To derive the asymptotic joint distribution of $\widehat{\mathbf{Z}} = (\widehat{Z}_1, \dots, \widehat{Z}_K)$ we need to compute its covariance matrix. Proceeding in accord with Jennison and Turnbull,³ we first replace the estimated information with its population version, resulting in

$$Z_k = (\widehat{p}^{(k)} - 1/2) \sqrt{\mathcal{I}_k} \xrightarrow[H_0]{\mathcal{D}} \mathcal{N}(0, 1), \quad k = 1, \dots, K, \quad (2)$$

$$\mathcal{I}_k = (N_k n_{1k} n_{2k}) / \sigma_{Rk}^2, \quad (3)$$

where we assume the variance

$$\sigma_{Rk}^2 = N_k \left\{ (N_k - 2) \int F^2 dF - \frac{N_k - 3}{4} \right\} - \frac{N_k}{4} \int (F^+ - F^-) dF \quad (4)$$

and therefore the true distribution $F = F_1 = F_2$ to be known.¹³ If F is continuous, the information simplifies to $\mathcal{I}_k = \widehat{\mathcal{I}}_k = (12 n_{1k} n_{2k}) / (N_k + 1)$.

Since $\widehat{\sigma}_{Rk}^2$ are consistent estimators of σ_{Rk}^2 , $k = 1, \dots, K$, the vector of Wilcoxon-Mann-Whitney test statistics $\widehat{\mathbf{Z}}$ has the same limiting distribution as its counterpart $\mathbf{Z} = (Z_1, \dots, Z_K)$ with the true population information. The limiting distribution being multivariate normal, it remains to establish the covariances of the components of \mathbf{Z} .

Proposition 1. Let Z_k and \mathcal{I}_k be defined as in (2) and (3). Then, for $1 \leq k_1 \leq k_2 \leq K$,

$$\text{Cov}(Z_{k_1}, Z_{k_2}) = \sqrt{\mathcal{I}_{k_1} / \mathcal{I}_{k_2}}.$$

2.3 Brunner-Munzel test

To test the null hypothesis $H_0 : p = 1/2$ against $H_1 : p \neq 1/2$, we now compute, analogous to before, for each interim analysis k the Brunner-Munzel test statistic

$$\widehat{Z}_k = (\widehat{p}^{(k)} - 1/2) \sqrt{\widehat{\mathcal{I}}_k}, \quad k = 1, \dots, K, \quad (5)$$

with estimated information $\widehat{\mathcal{I}}_k = (\widehat{\sigma}_{1k}^2 / n_{1k} + \widehat{\sigma}_{2k}^2 / n_{2k})^{-1}$, where

$$\widehat{\sigma}_{1k}^2 = \frac{1}{n_{2k}^2 (n_{1k} - 1)} \sum_{i=1}^{n_{1k}} \left(R_{1i}^{(k)} - R_{1i}^{(1k)} - \bar{R}_{1\bullet}^{(k)} + \frac{n_1 + 1}{2} \right)^2,$$

$$\widehat{\sigma}_{2k}^2 = \frac{1}{n_{1k}^2 (n_{2k} - 1)} \sum_{j=1}^{n_{2k}} \left(R_{2j}^{(k)} - R_{2j}^{(2k)} - \bar{R}_{2\bullet}^{(k)} + \frac{n_2 + 1}{2} \right)^2,$$

and $R_{gi}^{(gk)}$ denotes the mid-rank of X_{gi} among the observations of the g th treatment group $X_{g1}, \dots, X_{gn_{gk}}$ available at analysis k ; $g = 1, 2$; $i = 1, \dots, n_{gk}$.

For the derivation of the asymptotic covariance, we take an approach similar to before. Once again, we substitute the estimated information with the true one

$$\begin{aligned} Z_k &= (\widehat{p}^{(k)} - 1/2) \sqrt{\mathcal{I}_k} \xrightarrow{\mathcal{D}} \mathcal{N}(\theta \sqrt{\mathcal{I}_k}, 1), \quad k = 1, \dots, K, \\ \theta &= p - 1/2, \\ \mathcal{I}_k &= (\sigma_1^2 / n_{1k} + \sigma_2^2 / n_{2k})^{-1}, \end{aligned} \quad (6)$$

where $\sigma_1^2 = \mathbb{V}\{F_2(X_{1i})\}$ and $\sigma_2^2 = \mathbb{V}\{F_1(X_{2j})\}$. However, since the definition of the variance components σ_1^2 and σ_2^2 is actually based on an asymptotically equivalent version of the Z_k s, that is to say,

$$Z_k^U = \left\{ \frac{1}{n_{2k}} \sum_{j=1}^{n_{2k}} F_1(X_{2j}) - \frac{1}{n_{1k}} \sum_{i=1}^{n_{1k}} F_2(X_{1i}) \right\} \sqrt{\mathcal{I}_k} \xrightarrow{\mathcal{D}} \mathcal{N}(\theta \sqrt{\mathcal{I}_k}, 1), \quad (7)$$

we compute the covariance accordingly. This result is given in the following proposition.

Proposition 2. Let Z_k^U and \mathcal{I}_k be defined as in (7) and (6). Then, for $1 \leq k_1 \leq k_2 \leq K$,

$$\text{Cov}(Z_{k_1}^U, Z_{k_2}^U) = \sqrt{\mathcal{I}_{k_1} \mathcal{I}_{k_2}}.$$

Thus, $\widehat{\mathcal{I}}_k$ consistently estimating \mathcal{I}_k , $k = 1, \dots, K$, the sequence of Brunner-Munzel test statistics $\{\widehat{Z}_1, \dots, \widehat{Z}_K\}$ asymptotically follow the canonical joint distribution. In the nonsequential scenario, the test has been shown to be too liberal for small sample sizes when using standard normal quantiles.¹² Analogous to the parametric Behrens-Fisher problem, they propose a Satterthwaite-Smith-Welch t -approximation²⁵⁻²⁷ with degrees of freedom estimated by

$$\widehat{\nu}_k = \frac{\{\widehat{\sigma}_{1k}^2/n_{1k} + \widehat{\sigma}_{2k}^2/n_{2k}\}^2}{\widehat{\sigma}_{1k}^4/\{n_{1k}^2(n_{1k}-1)\} + \widehat{\sigma}_{2k}^4/\{n_{2k}^2(n_{2k}-1)\}}. \quad (8)$$

Another way is to employ a variance stabilising transformation, such as the *logit* function, producing the logarithmised win odds, which we will explore in the next subsection.

2.4 Log win odds test

To address the liberal behaviour of the Brunner-Munzel test, we now consider

$$\begin{aligned} \psi &= \ln \{p/(1-p)\}, \\ \widehat{\psi}^{(k)} &= \ln \{\widehat{p}^{(k)}/(1-\widehat{p}^{(k)})\}, \end{aligned}$$

at stage $k = 1, \dots, K$. Consequently, straightforward application of the delta method yields

$$\widehat{Z}_k = (\widehat{\psi}^{(k)} - 0) \sqrt{\widehat{\mathcal{I}}_k} \xrightarrow{\mathcal{D}} \mathcal{N}(\theta \sqrt{\mathcal{I}_k}, 1), \quad k = 1, \dots, K, \quad (9)$$

$$Z_k = (\widehat{\psi}^{(k)} - 0) \sqrt{\mathcal{I}_k} \xrightarrow{\mathcal{D}} \mathcal{N}(\theta \sqrt{\mathcal{I}_k}, 1), \quad k = 1, \dots, K, \quad (10)$$

with effect $\theta = \psi - 0$ and information levels

$$\begin{aligned} \mathcal{I}_k &= \frac{\{p(1-p)\}^2}{\sigma_1^2/n_{1k} + \sigma_2^2/n_{2k}}, \\ \widehat{\mathcal{I}}_k &= \frac{\{\widehat{p}^{(k)}(1-\widehat{p}^{(k)})\}^2}{\widehat{\sigma}_{1k}^2/n_{1k} + \widehat{\sigma}_{2k}^2/n_{2k}}, \end{aligned}$$

which is nothing but $\{p(1-p)\}^2$ times, or $\{\widehat{p}^{(k)}(1-\widehat{p}^{(k)})\}^2$ times, the information for the corresponding effect $p - 1/2$ from the Brunner-Munzel test as in Section 2.3. Moreover, Proposition 2 together with the information obtained by the delta method directly imply that the log win odds test statistics asymptotically follow the canonical joint distribution.

To recapitulate, in all three cases under the respective assumptions, the standardised test statistics $\{Z_1, \dots, Z_K\}$ with information $\{\mathcal{I}_1, \dots, \mathcal{I}_K\}$ for the parameter θ asymptotically follow the *canonical joint distribution*. The difference between the Wilcoxon-Mann-Whitney and Brunner-Munzel tests arises solely from the way in which we define the information, both distributions F_1 and F_2 needing to coincide for the former but not the latter. The log win odds test is nothing but a Brunner-Munzel test based on the *logit* transformed nonparametric relative effect p .

Before we investigate the adequacy of the proposed methods by means of simulations, we turn our discussion to error spending to explain in more detail the manner in which we wish to reject the null hypothesis.

3 Error spending

Initially, group sequential methods required the number of interim looks to be specified in advance and equally spaced: Pocock²⁸ considered standard normal test statistics and derived local significance levels ('stage levels') which are identical across all stages, while O'Brien and Fleming²⁹ stage levels are extremely low at the first interim and increase with each stage in such a way that the final stage level is quite close to the nominal overall significance level α . To avoid having to specify the time or number of interim looks in advance, Lan and DeMets³⁰ suggested the use of error spending functions, which we will employ in the simulations.

With statistics and information levels $Z_k, \widehat{Z}_k, \mathcal{I}_k, \widehat{\mathcal{I}}_k, k = 1, \dots, K$, given as in the previous section, a right-sided group sequential test for efficacy maintains the nominal significance level α if the stage levels $\alpha_1, \dots, \alpha_K$ are chosen such that

$$\alpha = \mathbb{P}_{H_0}(p_k \leq \alpha_k \text{ for some } k = 1, \dots, K), \quad (11)$$

where we regard the repeated p -values $p_k = 1 - \Phi(\widehat{Z}_k), k = 1, \dots, K$, to be random variables, Φ denoting the cumulative distribution function of the standard normal distribution. The null hypothesis is rejected at stage k if $p_k \leq \alpha_k$ and the trial is consequently stopped. We do not, however, set up futility bounds.

To obtain specific stage levels, we split the global α into K positive parts π_k (' α spent at stage k '), $k = 1, \dots, K$, such that $\sum_{k=1}^K \pi_k = \alpha$ and

$$\mathbb{P}_{H_0}(p_1 > \alpha_1, \dots, p_{k-1} > \alpha_{k-1}, p_k \leq \alpha_k) = \pi_k.$$

To compute the stage levels $\alpha_1, \dots, \alpha_K$, we make use of the underlying limiting *canonical joint distribution* of the statistics $\{\widehat{Z}_1, \dots, \widehat{Z}_K\}$ and estimate the covariance of \widehat{Z}_k and \widehat{Z}_K by $\sqrt{\widehat{\mathcal{I}}_k/\mathcal{I}_{max}}, k = 1, \dots, K - 1$, where \mathcal{I}_{max} is the prespecified information that we believe would be available if the total maximum sample size N_K of the trial were observed under the respective treatment allocation scheme.

The error spending function prescribes precisely how the global α is to be spent across the stages. More formally, an error spending function is defined as a nondecreasing function $f: [0, \infty[\rightarrow [0, \alpha]$ such that $f(0) = 0$ and $f(t) = \alpha$ for all $t \geq 1$. Then the amount of α allocated to stages $k = 1, \dots, K$ is given by

$$\begin{aligned} \pi_1 &= f(\mathcal{I}_1/\mathcal{I}_K), \\ \pi_2 &= f(\mathcal{I}_k/\mathcal{I}_K) - f(\mathcal{I}_{k-1}/\mathcal{I}_K), \quad k = 2, \dots, K. \end{aligned}$$

However, the true information levels are not known in advance. Therefore, we use \mathcal{I}_{max} instead of \mathcal{I}_K and replace the other information levels by their estimates,

$$\begin{aligned} \pi_1 &= f(\widehat{\mathcal{I}}_1/\mathcal{I}_{max}), \\ \pi_2 &= f(\widehat{\mathcal{I}}_k/\mathcal{I}_{max}) - f(\widehat{\mathcal{I}}_{k-1}/\mathcal{I}_{max}), \quad k = 2, \dots, K - 1, \\ \pi_K &= \alpha - f(\widehat{\mathcal{I}}_{K-1}/\mathcal{I}_{max}). \end{aligned}$$

As $\widehat{\mathcal{I}}_K$ might turn out to be lower than \mathcal{I}_{max} , the last equation ensures that the full amount of α still available is spent at the last stage. Moreover, it is important to bear in mind that the information levels $\widehat{\mathcal{I}}_k$ are estimated at stage k and remain unchanged thereafter.

4 Simulations

As the methods developed in Section 2 are of asymptotic nature, we explore their applicability for finite sample sizes in a range of scenarios. To this end, we simulate the group sequential Wilcoxon-Mann-Whitney, Brunner-Munzel, and log win odds tests given as in (1), (5), and (9), respectively. Assuming that lower values correspond to more favourable outcomes, we want to show that treatment 1 is superior to treatment 2, yielding a one-sided efficacy test with $H_0: p \leq 1/2$ against $H_1: p > 1/2$ and a nominal overall significance level of $\alpha = 0.025$. In that regard, it is perhaps more natural to view the Wilcoxon-Mann-Whitney test as a means to test the null hypothesis $H_0: p \leq 1/2$ as well, with $F_1 = F_2$ constituting a model assumption under the null.

To gauge the type I error rate of our proposed methods, we perform 100,000 simulation runs for each scenario, giving rise to a Monte Carlo error of about 0.0003 based on a 95%-precision interval for a global $\alpha = 0.025$. Altogether, we present the results of 120 scenarios for each data generating process, that is all combinations of

- total maximum sample sizes $N_K = \{144, 288, 576, 864, 1008\}$,
- allocation ratios 1:1 or 2:1 (twice as many patients on treatment arm 1),
- two, three, or four stages, and
- two error spending functions.

More specifically, we consider O'Brien and Fleming²⁹ as well as Pocock²⁸ type error spending functions

$$f_{OF}(t) = \min \left\{ 2 - 2\Phi \left(\frac{z_{1-\alpha/2}}{\sqrt{t}} \right), \alpha \right\},$$

$$f_{PO}(t) = \min[\alpha \ln\{1 + (e-1)t\}, \alpha],$$

using the information fractions $\widehat{\mathcal{I}}_k/\mathcal{I}_K$, $k = 1, \dots, K$ to determine the amount of α to be spent since we know the true maximum information \mathcal{I}_K . For the subsequent computation of the stage levels, we make use of the command `getDesignGroupSequential()` from the R package `rpact`.³¹ In addition to using standard normal quantiles for the Wilcoxon-Mann-Whitney, Brunner-Munzel, and log win odds tests, we compute rejection rates based on the Satterthwaite-Smith-Welch t -approximation for the Brunner-Munzel test. As is suggested by Jennison and Turnbull³ and Wassmer and Brannath⁵ to provide satisfactorily accurate results for the two sample t -test, we use the same stage levels for the t -approximation and change the computation of the repeated p -values only, namely $p_k = 1 - F_{\nu_k}^-(\widehat{Z}_k)$, where $F_{\nu_k}^-$ denotes the cumulative distribution function of the t -distribution with $\widehat{\nu}_k$ degrees of freedom as in (8).

It might occur that our methods break down, for instance the variance estimate of the Brunner-Munzel test might be zero in finite samples or the estimated information could actually decrease in a subsequent stage. Since this happened very rarely and has virtually no influence on the results presented in the main paper, we relegate the discussion on exception handling to the supplementary material. Moreover, we only report the overall type I error rate here, that is, the relative frequency of simulation runs, where the null hypothesis could be rejected at some stage. Readers interested in a more detailed presentation of the results such as cumulative rejection rates for each stage are again referred to the supplementary material.

4.1 Normal distribution

First we generated data from normal distributions, namely $X_{gi} \stackrel{iid}{\sim} \mathcal{N}(\mu_g, \sigma_g^2)$, $g = 1, 2$, $i = 1, \dots, n_g$, for three different settings as set out in Figures 1 to 3. In case of equal variances, the Wilcoxon-Mann-Whitney test best maintains the nominal type I error rate for all total maximum sample sizes, whereas the Brunner-Munzel test with or without t -approximation tends to be too liberal and the log win odds test too conservative for smaller samples sizes. In both heteroskedastic settings, that is settings 2 and 3, the Wilcoxon-Mann-Whitney test exceeds the nominal significance level across all sample sizes if the allocation ratio is 1:1. However, if twice as many patients receive treatment 1, then the Wilcoxon-Mann-Whitney test is far too liberal if the data in treatment 1 is less dispersed than in treatment 2 and far too conservative conversely. Again, this behaviour is not affected by sample size.

In line with the simulation results of Brunner and Munzel¹² for the fixed sample size scenario, the rejection rates pattern of the other tests are not affected by heteroskedasticity or different allocation schemes.

4.2 Ordinal data

Now we consider ordinal data divided into five categories $\mathcal{C}_1 < \mathcal{C}_2 < \mathcal{C}_3 < \mathcal{C}_4 < \mathcal{C}_5$, with a smaller index pointing to a more favourable outcome. As in Brunner et al.,¹⁶ the probabilities of each category occurring are derived through a latent Beta distribution: Let $Y_{gi} \stackrel{iid}{\sim} \text{Beta}(\alpha_g, \beta_g)$, $g = 1, 2$, $i = 1, \dots, n_g$, denote a Beta distributed random variable with shape parameters $\alpha_g, \beta_g > 0$, such that the expectation and variance of Y_{gi} are given by

$$\mathbb{E}(Y_{gi}) = \frac{\alpha_g}{\alpha_g + \beta_g}, \quad \mathbb{V}(Y_{gi}) = \frac{\alpha_g \beta_g}{(\alpha_g + \beta_g)^2 (\alpha_g + \beta_g + 1)}.$$

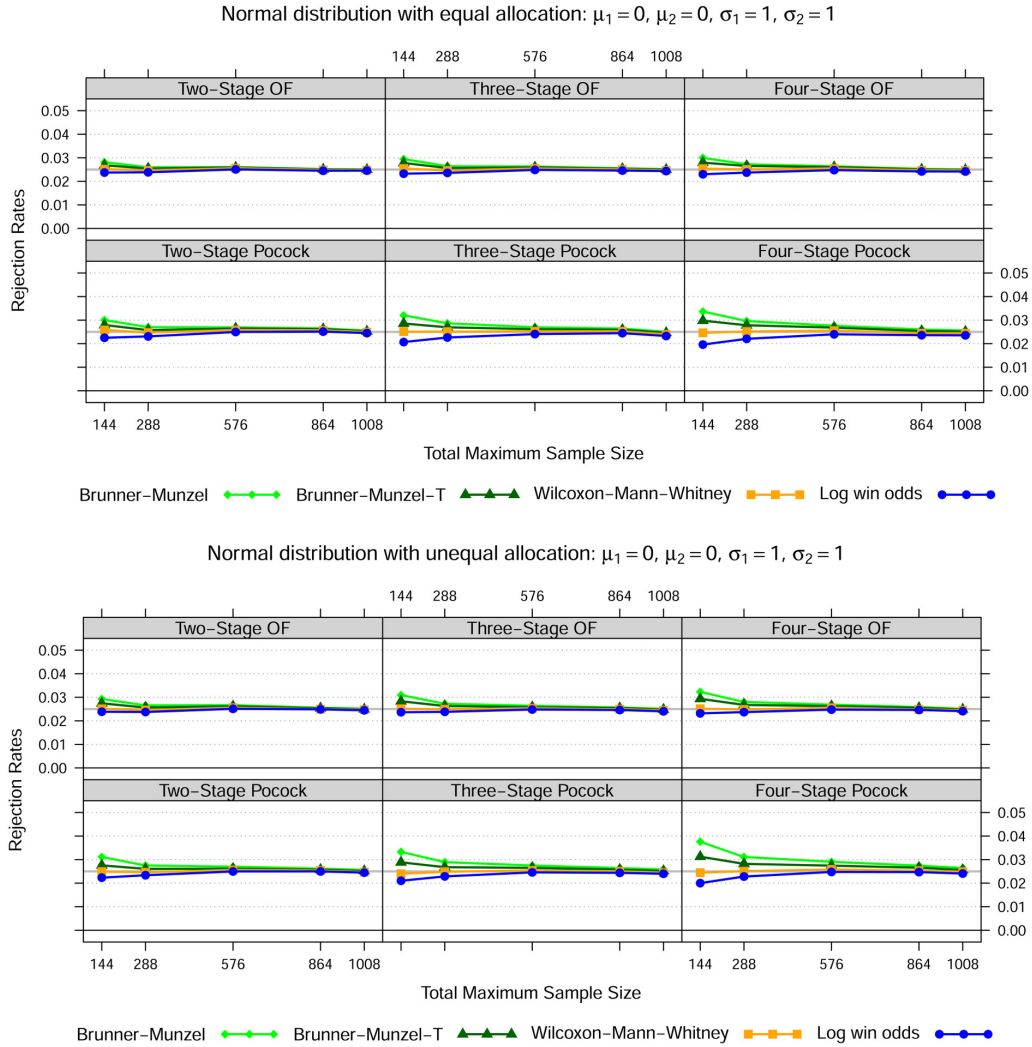


Figure 1. Normal distribution—Setting I

Notes: The lines show the relative frequency of the 100000 simulation runs, where the null hypothesis could be rejected at some stage based on the Brunner-Munzel test (with t-approximation) as in (5), the Wilcoxon-Mann-Whitney test as in (1) and the log win odds test as in (9) for five different total maximum sample sizes, two error spending functions, up to four stages in total as well as two different allocation ratios.

Then, the random variable X_{gi} , $g = 1, 2, i = 1, \dots, n_g$, is defined by

$$X_{gi} = C_k \text{ if } Y_{gi} \in [0.2(k-1), 0.2k] \text{ for } k = 1, \dots, 5.$$

Consequently, the probability mass function of X_{gi} is nothing but

$$\mathbb{P}(X_{gi} = C_k) = \mathbb{P}\{0.2(k-1) \leq Y_{gi} < 0.2k\} \text{ for } k = 1, \dots, 5.$$

We specify three different parameter settings to mimic the homo-/heteroskedasticity pattern for the normal scenarios in Section 4.1. The results exhibit virtually the same behaviour as the normally distributed responses shown previously and are therefore included in the online supplementary material.

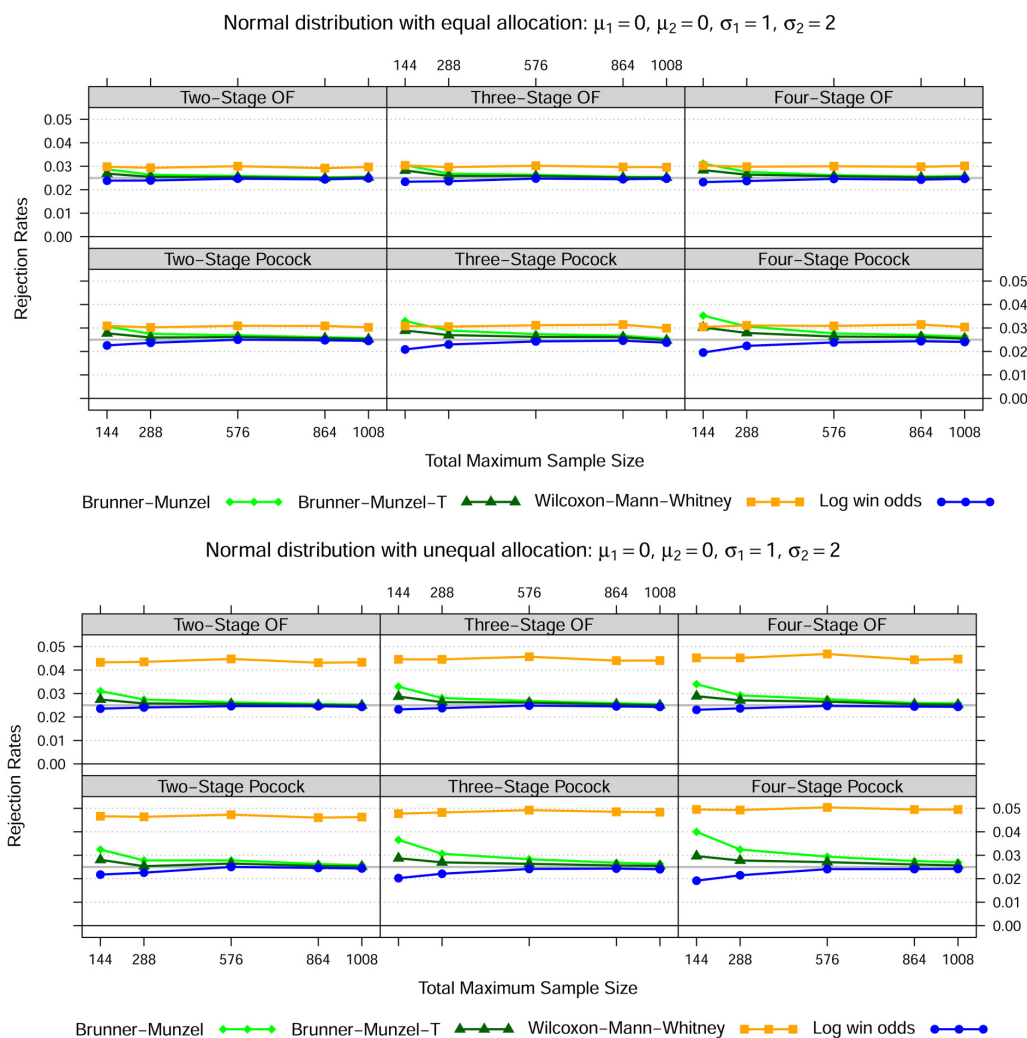


Figure 2. Normal distribution—Setting 2
 Notes: The lines show the relative frequency of the 100000 simulation runs, where the null hypothesis could be rejected at some stage based on the Brunner-Munzel test (with t-approximation) as in (5), the Wilcoxon-Mann-Whitney test as in (1) and the log win odds test as in (9) for five different total maximum sample sizes, two error spending functions, up to four stages in total as well as two different allocation ratios.

5 FREEDOMS clinical trial

The FREEDOMS clinical trial (ClinicalTrials.gov Identifier: NCT00289978) was a placebo-controlled phase III study running from January 2006 to July 2009 to analyse the efficacy and safety of fingolimod in patients with relapsing-remitting multiple sclerosis.³² The primary efficacy endpoint was the annualised relapse rate at 24 months after baseline evaluation. The definition of a relapse was based on the Expanded Disability Status Scale (EDSS),³³ with values ranging from 0 (normal status) to 10 (death due to multiple sclerosis) and a step size of 0.5, although a value of 0.5 is not possible. Thus, a higher score on the EDSS indicates more severe disability.

In this paper, we focus on the EDSS score at 24 months, its change compared to the baseline (post minus prae), and its direction of change, that is, whether the EDSS score at 24 month decreased (−1), stayed the same (0), or increased (+1)

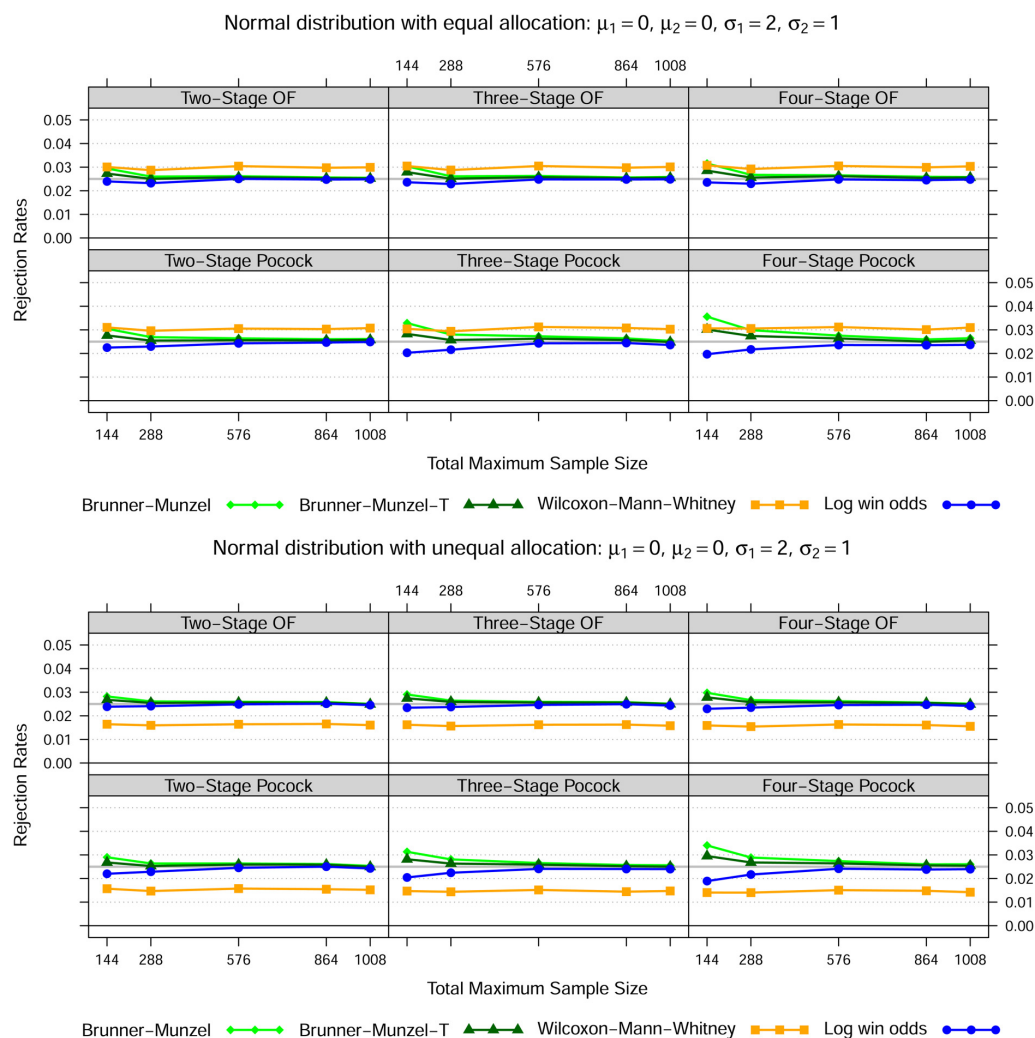


Figure 3. Normal distribution—Setting 3

Notes: The lines show the relative frequency of the 100000 simulation runs, where the null hypothesis could be rejected at some stage based on the Brunner-Munzel test (with t -approximation) as in (5), the Wilcoxon-Mann-Whitney test as in (1) and the log win odds test as in (9) for five different total maximum sample sizes, two error spending functions, up to four stages in total as well as two different allocation ratios.

with respect to the baseline value. To simplify the presentation of the results, we only considered the complete cases data set, that is, patients where the EDSS score was observed both at baseline and 24 months thereafter. Summary descriptive statistics depicted in Table 1 reveal in all three cases that, at the end of the trial, the mean EDSS outcome of patients on the placebo arm is higher and therefore less favourable than for those on the fingolimod 0.5 mg treatment.

While the original design of the FREEDOMS trial did not provide for interim looks, we now retrospectively analyse the data as though there were two equally spaced stages. More specifically, the first 353 patients on either arm who completed the 24 month evaluation form the basis of the first stage analysis, while all 706 patients are taken into account at the second and therefore last stage. As we did in the simulation section, we consider the Wilcoxon-Mann-Whitney test, the Brunner-Munzel test (with t -approximation) as well as the log win odds test and employ O'Brien and Fleming as well

Table 1. Summary descriptive statistics for EDSS data at month 24, month 24 minus baseline (change), and direction of change from the FREEDOMS clinical trial.

EDSS	Treatment	<i>n</i>	Mean	SD	Min	Median	Max
Month 24	Fingolimod 0.5 mg	374	2.269	1.442	0	2	6.5
	Placebo	332	2.545	1.507	0	2	7.0
Change	Fingolimod 0.5 mg	374	0.004	0.878	-3	0	3.5
	Placebo	332	0.131	0.936	-3	0	3.5
Direction	Fingolimod 0.5 mg	374	-0.078	0.734	-1	0	1
	Placebo	332	0.099	0.769	-1	0	1

Table 2. Repeated effect estimates, *p*-values in % based on standard normal and *t* approximation (T), O'Brien and Fleming (α_{OF}) and Pocock type (α_P) error spending stage levels in %.

EDSS	<i>N</i>	Estimate	Wilcoxon-Mann-Whitney			Brunner-Munzel			Log win odds		
			<i>p</i> -value	α_{OF}	α_{PO}	<i>p</i> -value (T)	α_{OF}	α_{PO}	<i>p</i> -value	α_{OF}	α_{PO}
Month 24	353	0.545	7.20	0.16	1.56	7.19 (7.23)	0.15	1.54	7.29	0.16	1.56
	706	0.558	0.34**	2.45	1.38	0.33** (0.33**)	2.45	1.39	0.35**	2.45	1.38
Change	353	0.564	1.60	0.14	1.53	1.60 (1.63)	0.14	1.53	1.69	0.14	1.52
	706	0.560	0.21**	2.45	1.39	0.20** (0.21**)	2.45	1.40	0.22**	2.46	1.40
Direction	353	0.565	1.21*	0.15	1.54	1.20* (1.23*)	0.14	1.53	1.28*	0.14	1.53
	706	0.563	0.09**	2.45	1.39	0.09** (0.09**)	2.45	1.40	0.10**	2.45	1.40

*Rejection with respect to Pocock type stage level only;

**Rejection with respect to both Pocock and O'Brien and Fleming type stage levels.

Table 3. Repeated 95%-confidence intervals based on Pocock type alpha spending function.

EDSS	<i>N</i>	Estimate	Brunner-Munzel		Brunner-Munzel (T)		Log win odds	
Month 24	353	0.545	0.479	0.610	0.479	0.610	0.478	0.609
	706	0.558	0.511	0.606	0.511	0.606	0.511	0.605
Change	353	0.564	0.499	0.628	0.499	0.628	0.499	0.626
	706	0.560	0.514	0.605	0.514	0.605	0.514	0.605
Direction	353	0.565	0.503	0.628	0.503	0.628	0.502	0.626
	706	0.563	0.519	0.608	0.519	0.608	0.518	0.607

as Pocock type error spending functions. Since we do this analysis retrospectively, we can choose $\mathcal{I}_{max} = \widehat{\mathcal{I}}_2$. In all scenarios the estimated information fractions $\widehat{\mathcal{I}}_1/\widehat{\mathcal{I}}_2$ are close to 0.5, essentially coinciding with the sample size fraction 353/706.

Analogous to the simulation section, we aim to reject $H_0 : p \leq 1/2$ at a global significance level of 2.5%. As Tables 2 to 4 demonstrate, we can reject the null hypothesis at some stage in any scenario and conclude that fingolimod treatment is efficacious. Only the direction of change endpoint leads to early rejection, that is, when using Pocock type stage levels. Even if the trial could not have been stopped at the interim, second stage *p*-values in the region of 0.1% would have resulted in rejection eventually. Consistent with the results from the simulations, the *p*-values and confidence intervals from different tests are fairly close.

6 Planning and sample size considerations

In planning a clinical trial, a careful examination of the power of different scenarios under the alternative appears to be advisable at any rate. With the nonparametric relative effect *p* chosen as the efficacy estimand of the primary endpoint, we now extend and slightly modify the approach to sample size planning for the fixed scenario proposed by Happ et al.³⁴ to the group sequential setting.

Table 4. Repeated 95%-confidence intervals based on O'Brien and Fleming type alpha spending function.

EDSS	N	Estimate	Brunner-Munzel		Brunner-Munzel (T)		Log win odds	
Month 24	353	0.545	0.454	0.635	0.453	0.636	0.454	0.633
	706	0.558	0.516	0.601	0.516	0.601	0.516	0.600
Change	353	0.564	0.475	0.652	0.474	0.653	0.474	0.649
	706	0.560	0.519	0.601	0.519	0.601	0.518	0.600
Direction	353	0.565	0.479	0.651	0.478	0.652	0.478	0.649
	706	0.563	0.524	0.603	0.523	0.603	0.523	0.603

Table 5. Power of the Wilcoxon-Mann-Whitney (WMW), Brunner-Munzel (BM), and log win odds (LWO) tests for an equally spaced two stage trial with ordinal data as in Section 4.2, $p = 0.6$, $\alpha_1 = 0.6974797$, $\beta_1 = 1$, $\alpha_2 = 3$, $\beta_2 = 3$.

t	Test	Error spending function	N_1	N_2	Power formula	Simulated power (stage one)
0.5	WMW	Pocock	142	284	0.80382	0.80352 (0.48612)
0.5	BM	Pocock	144	288	0.80231	0.79546 (0.47652)
0.5	LWO	Pocock	152	304	0.80213	0.80372 (0.47272)
0.5	WMW	O'Brien and Fleming	126	252	0.80008	0.79989 (0.16823)
0.5	BM	O'Brien and Fleming	130	260	0.80597	0.79743 (0.19909)
0.5	LWO	O'Brien and Fleming	136	272	0.80232	0.80717 (0.12543)
2/3	WMW	Pocock	153	306	0.80488	0.80571 (0.46197)
2/3	BM	Pocock	132	264	0.80784	0.80016 (0.47790)
2/3	LWO	Pocock	138	276	0.80379	0.80569 (0.47236)
2/3	WMW	O'Brien and Fleming	135	270	0.80472	0.80364 (0.13013)
2/3	BM	O'Brien and Fleming	117	234	0.80417	0.79515 (0.19662)
2/3	LWO	O'Brien and Fleming	123	246	0.80242	0.80582 (0.12398)

As before, we consider the hypothesis pair $H_0: p \leq 1/2$ and $H_1: p > 1/2$ with a nominal overall significance level of $\alpha = 0.025$. To determine the power of a particular alternative, it is convenient to specify the distributions F_1 and F_2 as well as a constant sample size ratio $t = n_{1k}/N_k$ for all stages $k = 1, \dots, K$ such that $F = tF_1 + (1-t)F_2$ is the distribution of the whole data ignoring the group structure, which appears in the variance formula (4) of the Wilcoxon-Mann-Whitney test. If we then choose the sample sizes for the particular stages $k = 1, \dots, K$, we immediately get the true information \mathcal{I}_k^{WMW} , \mathcal{I}_k^{BM} , \mathcal{I}_k^{LWO} as given in (3), (6) and (10), respectively. Approximate power formulas for the group sequential Wilcoxon-Mann-Whitney, Brunner-Munzel and log win odds tests then take the form as provided in the following two propositions.

Proposition 3 Let c_1, \dots, c_K denote the critical values computed from a K -variate normal distribution with mean vector $\mathbf{0}$, covariance matrix $\mathbf{R}^{WMW} = (r_{ij})_{i,j=1,\dots,K}$, $r_{ij} = \sqrt{\mathcal{I}_{\min(k_i,k_j)}^{WMW}/\mathcal{I}_{\max(k_i,k_j)}^{WMW}}$, and error spending function of choice. Then the approximate power of the group sequential Wilcoxon-Mann-Whitney test for $H_1: p > 1/2$ is given by

$$\text{Power}_{WMW} \approx 1 - \Phi_{\mathbf{R}} \left\{ \sqrt{\mathcal{I}_1^{BM}/\mathcal{I}_1^{WMW}} \cdot c_1 - \sqrt{\mathcal{I}_1^{BM}} \cdot (p - 1/2), \dots, \sqrt{\mathcal{I}_K^{BM}/\mathcal{I}_K^{WMW}} \cdot c_K - \sqrt{\mathcal{I}_K^{BM}} \cdot (p - 1/2) \right\},$$

where $\Phi_{\mathbf{R}}$ denotes the cumulative distribution function of a K -variate normal distribution with mean vector $\mathbf{0}$ and covariance matrix $\mathbf{R} = (r_{ij})$, $r_{ij} = \sqrt{N_{\min(k_i,k_j)}/N_{\max(k_i,k_j)}}$.

Proposition 4 Let c_1, \dots, c_K denote the critical values computed from a K -variate normal distribution with mean vector $\mathbf{0}$, covariance matrix $\mathbf{R} = (r_{ij})$, $r_{ij} = \sqrt{N_{\min(k_i,k_j)}/N_{\max(k_i,k_j)}}$, and error spending function of choice. Then the approximate

power of the group sequential Brunner-Munzel and log win odds tests for $H_1: p > 1/2$ is given by

$$\text{Power}_{\text{BM}} \approx 1 - \Phi_{\mathbf{R}} \left\{ c_1 - \sqrt{\mathcal{I}_1^{\text{BM}}} \cdot (p - 1/2), \dots, c_K - \sqrt{\mathcal{I}_K^{\text{BM}}} \cdot (p - 1/2) \right\},$$

$$\text{Power}_{\text{LWO}} \approx 1 - \Phi_{\mathbf{R}} \left(c_1 - \sqrt{\mathcal{I}_1^{\text{LWO}}} \cdot \psi, \dots, c_K - \sqrt{\mathcal{I}_K^{\text{LWO}}} \cdot \psi \right), \quad \psi = \ln \{p/(1-p)\},$$

respectively, where $\Phi_{\mathbf{R}}$ denotes the cumulative distribution function of a K -variate normal distribution with mean vector $\mathbf{0}$ and covariance matrix \mathbf{R} as given above.

The critical values c_1, \dots, c_K as well as $\Phi_{\mathbf{R}}(\cdot)$ can be easily obtained from the commands `getDesignGroupSequential` and `pmvnorm` of the respective R packages `rpact`³¹ and `mvtnorm`.³⁵ To demonstrate the adequacy of the formulas just presented, the results of a small simulation study with 100,000 replications based on the ordinal distribution defined as in Section 4.2 are depicted in Table 5.

7 Discussion

In this paper, we derived group sequential methodology for the Wilcoxon-Mann-Whitney, the Brunner-Munzel, and the log win odds tests, establishing their convergence in distribution to the canonical joint distribution, with simulation studies lending further support to the validity of our approach.

If one is willing both to assume the distributions to be equal under the null and to dispense with confidence intervals, the group sequential Wilcoxon-Mann-Whitney test best maintains the nominal significance level, particularly if sample sizes are small.

In the presence of heteroskedasticity, the Wilcoxon-Mann-Whitney test is either too liberal or too conservative depending on the heteroskedasticity pattern and the sample size allocation ratio. On the other hand, the log win odds test never exceeds the nominal significance level but does have a somewhat conservative tendency in certain scenarios. Nonetheless, the log win odds test allows for test inversion to compute confidence limits for the log win odds, which can readily be converted to the win odds or nonparametric relative effect scales. While the Brunner-Munzel test, with or without t -approximation, can be inverted in the same manner, it tends to be too liberal, especially in case of small sample sizes. In light of the fact that the Brunner-Munzel test gives rise to liberal test decisions for nominal significance levels smaller than 0.05 in the nonsequential setting in small samples, this result is hardly surprising.

In the randomised clinical trial setting, there appears little reason to conclude that distributions under the null are not identical. Still, if the treatment arms produce heteroskedastic outcomes in the alternative, one may well be led to infer from the simulation results that the Wilcoxon-Mann-Whitney test might actually turn out to be less powerful than the log win odds test in certain cases. However, as our case study in Section 5 suggests, the different behaviours of the tests are presumably negligible when sample sizes are reasonably large.

Care should be taken when adopting our methods for multi-arm trials. While Dunnett-type³⁶ many-to-one comparisons should not pose particular difficulties, Tukey-type³⁷ all-pairwise comparisons might lead to Efron's paradox,³⁸⁻⁴⁰ that is, the nonparametric relative effect as defined in this paper may point to nontransitive conclusions. If treatment 1 is more beneficial than treatment 2 and treatment 2 is more beneficial than treatment 3, then it does not necessarily follow that treatment 1 is more beneficial than treatment 3.

Since the variance estimators require the endpoint at issue to induce a rank representation and therefore all pairwise comparisons to be transitive, the methodology presented here does not cover hierarchical composite and possibly censored endpoints in general terms as discussed in Buyse,⁴¹ Cantagallo et al.,⁴² Péron et al.,⁴³ or Buyse and Péron.⁴⁴ However, the idea of linking group sequential theory with generalised U -statistics^{45,46} might prove fruitful in extending our approach in this direction.


Declaration of conflicting interests


The author(s) declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article: Claus P. Nowak and Tobias Mütze are employees of Novartis Pharma AG.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The research is supported by the German Science Foundation awards number DFG KO 4680/4-1.

ORCID iDs

Tobias Mütze  <https://orcid.org/0000-0002-4111-1941>

Frank Konietzschke  <https://orcid.org/0000-0002-5674-2076>

Supplemental materials

The supplemental material as regards the simulations can be found online.

Proofs

Proof of Proposition 1. We begin with the derivation of the covariance for the group sequential Wilcoxon-Mann-Whitney test statistics assuming $F = F_1 = F_2$ and allowing for ties. Setting $\zeta_{ij} = c(X_{2j}, X_{1i})$, we have for $1 \leq k_1 \leq k_2 \leq K$

$$\begin{aligned} \text{Cov}(Z_{k_1}, Z_{k_2}) &= \text{Cov}\left\{\left(\widehat{p}^{(k_1)} - 1/2\right)\sqrt{L_{k_1}}, \left(\widehat{p}^{(k_2)} - 1/2\right)\sqrt{L_{k_2}}\right\} \\ &= \sqrt{L_{k_1}}\sqrt{L_{k_2}} \frac{1}{n_{1k_1}} \frac{1}{n_{2k_1}} \frac{1}{n_{1k_2}} \frac{1}{n_{2k_2}} \sum_{j=1}^{n_{2k_1}} \sum_{i=1}^{n_{1k_1}} \sum_{j'=1}^{n_{2k_2}} \sum_{i'=1}^{n_{1k_2}} \text{Cov}(\zeta_{ij}, \zeta_{i'j'}). \end{aligned}$$

First, we observe that $[\mathbb{E}\{c(X_{2j}, X_{1i})\}]^2 = (\int F_1 dF_2)^2 = (\int F dF)^2 = 1/4$. Now, with $i \neq i'$ and $j \neq j'$, there are four cases to distinguish, that is

$$\begin{aligned} \text{Cov}(\zeta_{ij}, \zeta_{i'j'}) &= 0, \\ \text{Cov}(\zeta_{ij}, \zeta_{ij}) &= \mathbb{P}(X_{1i} < X_{2j}) + 1/4 \cdot \mathbb{P}(X_{1i} = X_{2j}) - 1/4 \\ &= \mathbb{P}(X_{1i} < X_{2j}) + 1/2 \cdot \mathbb{P}(X_{1i} = X_{2j}) - 1/4 \cdot \mathbb{P}(X_{1i} = X_{2j}) - 1/4 \\ &= \int F dF - 1/4 \int (F^+ - F^-) dF - 1/4, \\ &= 1/4 - 1/4 \int (F^+ - F^-) dF, \\ \text{Cov}(\zeta_{ij}, \zeta_{i'j}) &= \mathbb{E}\{c(X_{2j}, X_{1i})c(X_{2j}, X_{1i'})\} - 1/4 \\ &= \int \mathbb{E}\{c(x, X_{1i})c(x, X_{1i'})\} dF_2(x) - 1/4 \\ &= \int \mathbb{E}\{c(x, X_{1i})\}\mathbb{E}\{c(x, X_{1i'})\} dF_2(x) - 1/4 \\ &= \int F_1^2 dF_2 - 1/4 = \int F^2 dF - 1/4, \end{aligned}$$

and by similar arguments, $\text{Cov}(\zeta_{ij}, \zeta_{i'j'}) = \int F^2 dF - 1/4$.

Altogether, there are

- $n_{2k_1}n_{1k_1}$ terms with index combination $i = i'$ and $j = j'$,
- $n_{2k_1}n_{1k_1}(n_{2k_2} - 1)$ terms with $i = i'$ and $j \neq j'$,
- $n_{2k_1}n_{1k_1}(n_{1k_2} - 1)$ terms with $i \neq i'$ and $j = j'$,
- $n_{2k_1}n_{1k_1}(n_{2k_2} - 1)(n_{1k_2} - 1)$ terms with $i \neq i'$ and $j \neq j'$.

Thus, if $F = F_1 = F_2$ but not necessarily continuous, the quadruple sum reduces to

$$\begin{aligned} &\sum_{j=1}^{n_{2k_2}} \sum_{i'=1}^{n_{1k_2}} \sum_{j=1}^{n_{2k_1}} \sum_{i=1}^{n_{1k_1}} \text{Cov}(\zeta_{ij}, \zeta_{i'j'}) \\ &= \left\{1/4 - 1/4 \int (F^+ - F^-) dF\right\} n_{2k_1}n_{1k_1} + \left(\int F^2 dF - 1/4\right) \{n_{2k_1}n_{1k_1}(n_{2k_2} - 1) + n_{2k_1}n_{1k_1}(n_{1k_2} - 1)\} \\ &= n_{2k_1}n_{1k_1} \left\{1/4 - 1/4 \int (F^+ - F^-) dF + \left(\int F^2 dF - 1/4\right)(N_{k_2} - 2)\right\} \\ &= n_{2k_1}n_{1k_1} \left\{(N_{k_2} - 2) \int F^2 dF - \frac{N_{k_2} - 3}{4} - 1/4 \int (F^+ - F^-) dF\right\} \\ &= n_{2k_1}n_{1k_1} \frac{\sigma_{Rk_2}^2}{N_{k_2}}. \end{aligned}$$

Putting everything together, we obtain

$$\begin{aligned}\mathbb{Cov}(Z_{k_1}, Z_{k_2}) &= \sqrt{\mathcal{I}_{k_1}} \sqrt{\mathcal{I}_{k_2}} \frac{1}{n_{1k_1}} \frac{1}{n_{2k_1}} \frac{1}{n_{1k_2}} \frac{1}{n_{2k_2}} n_{2k_1} n_{1k_1} \frac{\sigma_{Rk_2}^2}{N_{k_2}} \\ &= \sqrt{\mathcal{I}_{k_1}} \sqrt{\mathcal{I}_{k_2}} \frac{1}{n_{1k_2}} \frac{1}{n_{2k_2}} \frac{\sigma_{Rk_2}^2}{N_{k_2}} = \sqrt{\mathcal{I}_{k_1}} \sqrt{\mathcal{I}_{k_2}} (\mathcal{I}_{k_2})^{-1} = \sqrt{\mathcal{I}_{k_1} / \mathcal{I}_{k_2}}.\end{aligned}$$

In case of no ties, three of the four cases discussed above further simplify to $\mathbb{Cov}(\zeta_{ij}, \zeta_{ij}) = 1/4$ and $\mathbb{Cov}(\zeta_{ij}, \zeta_{j'j}) = \mathbb{Cov}(\zeta_{ij}, \zeta_{ij'}) = 1/12$, producing the desired result.

Proof of Proposition 2. As for the Brunner-Munzel test, it holds for $k_1 \leq k_2$,

$$\begin{aligned}\mathbb{Cov}(Z_{k_1}^U, Z_{k_2}^U) &= \mathbb{Cov} \left[\left\{ \frac{1}{n_{2k_1}} \sum_{j=1}^{n_{2k_1}} F_1(X_{2j}) - \frac{1}{n_{1k_1}} \sum_{i=1}^{n_{1k_1}} F_2(X_{1i}) \right\} \sqrt{\mathcal{I}_{k_1}}, \left\{ \frac{1}{n_{2k_2}} \sum_{j=1}^{n_{2k_2}} F_1(X_{2j}) - \frac{1}{n_{1k_2}} \sum_{i=1}^{n_{1k_2}} F_2(X_{1i}) \right\} \sqrt{\mathcal{I}_{k_2}} \right] \\ &= \sqrt{\mathcal{I}_{k_1}} \sqrt{\mathcal{I}_{k_2}} \left[\mathbb{Cov} \left\{ \frac{1}{n_{2k_1}} \sum_{j=1}^{n_{2k_1}} F_1(X_{2j}), \frac{1}{n_{2k_2}} \sum_{j=1}^{n_{2k_2}} F_1(X_{2j}) \right\} + \mathbb{Cov} \left\{ \frac{1}{n_{1k_1}} \sum_{i=1}^{n_{1k_1}} F_2(X_{1i}), \frac{1}{n_{1k_2}} \sum_{i=1}^{n_{1k_2}} F_2(X_{1i}) \right\} \right] \\ &= \sqrt{\mathcal{I}_{k_1}} \sqrt{\mathcal{I}_{k_2}} \left(\frac{1}{n_{2k_1}} \frac{1}{n_{2k_2}} n_{2k_1} \sigma_2^2 + \frac{1}{n_{1k_1}} \frac{1}{n_{1k_2}} n_{1k_1} \sigma_1^2 \right) = \sqrt{\mathcal{I}_{k_1} / \mathcal{I}_{k_2}},\end{aligned}$$

which concludes the proof.

Proof of Proposition 3. As for the Wilcoxon-Mann-Whitney test, we first consider the fixed design, that is, $K = 1$, under $H_1 : p > 1/2$. Adopting the notation from Sections 2 and 6 we have

$$\begin{aligned}\text{Power}_{\text{WMW}} &= \mathbb{P} \left\{ \sqrt{\widehat{\mathcal{I}}_1^{\text{WMW}}} \cdot (\widehat{p}^{(1)} - 1/2) \geq c_1 \right\} \\ &\approx \mathbb{P} \left\{ \sqrt{\mathcal{I}_1^{\text{WMW}}} \cdot (\widehat{p}^{(1)} - 1/2) \geq c_1 \right\} \\ &= \mathbb{P} \left\{ \sqrt{\mathcal{I}_1^{\text{BM}}} \cdot (\widehat{p}^{(1)} - 1/2) \geq \sqrt{\mathcal{I}_1^{\text{BM}} / \mathcal{I}_1^{\text{WMW}}} \cdot c_1 \right\} \\ &= \mathbb{P} \left\{ \sqrt{\mathcal{I}_1^{\text{BM}}} \cdot (\widehat{p}^{(1)} - p) \geq \sqrt{\mathcal{I}_1^{\text{BM}} / \mathcal{I}_1^{\text{WMW}}} \cdot c_1 - \sqrt{\mathcal{I}_1^{\text{BM}}} \cdot (p - 1/2) \right\} \\ &\approx 1 - \Phi \left\{ \sqrt{\mathcal{I}_1^{\text{BM}} / \mathcal{I}_1^{\text{WMW}}} \cdot c_1 - \sqrt{\mathcal{I}_1^{\text{BM}}} \cdot (p - 1/2) \right\},\end{aligned}$$

since $\sqrt{\mathcal{I}_1^{\text{BM}}} \cdot (\widehat{p}^{(1)} - p)$ is approximately standard normal under H_1 . Setting $t = n_{1k}/N_k$ for all $k = 1, \dots, K$ immediately gives

$$(\mathcal{I}_k^{\text{BM}})^{-1} = \frac{\sigma_1^2}{n_{1k}} + \frac{\sigma_2^2}{n_{2k}} = \frac{1}{N_k} \cdot \frac{N_k}{n_{1k}} \cdot \frac{N_k}{n_{2k}} \cdot \left(\frac{n_{2k} \sigma_1^2}{N_k} + \frac{n_{1k} \sigma_2^2}{N_k} \right) = N_k^{-1} \cdot \frac{(1-t)\sigma_1^2 + t\sigma_2^2}{t(1-t)},$$

yielding $\sqrt{\mathcal{I}_{k_1}^{\text{BM}} / \mathcal{I}_{k_2}^{\text{BM}}} = \sqrt{N_{k_1} / N_{k_2}}$. The formula for general K follows directly from the canonical joint distribution.

Proof of Proposition 4. The arguments are completely analogous to the ones given for Proposition 3 and are therefore omitted.

References

1. European Medicines Agency. *Reflection Paper on Methodological Issues in Confirmatory Clinical Trials Planned with an Adaptive Design*, 2007. <https://www.ema.europa.eu/en/documents/scientific-guideline/reflection->
2. US Food and Drug Administration. *Adaptive Designs for Clinical Trials of Drugs and Biologics: Guidance for Industry*, 2019. <https://www.fda.gov/media/78495/download> (Accessed November 9, 2020).
3. Jennison C and Turnbull BW. *Group Sequential Methods with Applications to Clinical Trials*. Boca Raton: Chapman & Hall/CRC, 2000.
4. Proschan MA, Lan KKG and Wittes J. *Statistical Monitoring of Clinical Trials: A Unified Approach*. MA, New York: Springer, 2006.
5. Wassmer G and Brannath W. *Group Sequential and Confirmatory Adaptive Designs in Clinical Trials*. Springer International Publishing, 2016.

6. Mann HB and Whitney DR. On a test of whether one of two random variables is stochastically larger than the other. *Ann Math Stat* 1947; **18**: 50–60.
7. Wilcoxon F. Individual comparisons by ranking methods. *Biometric Bull* 1945; **1**: 80–83.
8. Wilcoxon F. Probability tables for individual comparisons by ranking methods. *Biometrics* 1947; **3**: 119–122.
9. Alling DW. Early decision in the Wilcoxon two-sample test. *J Am Stat Assoc* 1963; **58**: 713–720.
10. Phatarfod RM and Sudbury A. A simple sequential Wilcoxon test. *Aust J Stat* 1988; **30**: 93–106.
11. Shuster JJ, Chang MN and Tian L. Design of group sequential clinical trials with ordinal categorical data based on the Mann–Whitney–Wilcoxon test. *Seq Anal* 2004; **23**: 413–426.
12. Brunner E and Munzel U. The nonparametric Behrens-Fisher problem: Asymptotic theory and a small-sample approximation. *Biom J* 2000; **42**: 17–25.
13. Brunner E, Bathke AC and Konietzschke F. *Rank and Pseudo-Rank Procedures for Independent Observations in Factorial Designs*. Springer International Publishing, 2018.
14. Thas O, De Neve J, Clement L et al. Probabilistic index models. *J R Stat Soc B (Statistical Methodology)* 2012; **74**: 623–671.
15. Fay MP, Brittain EH, Shih JH et al. Causal estimands and confidence intervals associated with Wilcoxon-Mann-Whitney tests in randomized experiments. *Stat Med* 2018; **37**: 2923–2937.
16. Brunner E, Vandemeulebroecke M and Mütze T. Win odds: An adaptation of the win ratio to include ties. *Stat Med* 2021; **40**: 3367–3384.
17. Putter J. The treatment of ties in some nonparametric tests. *Ann Math Stat* 1955; **26**: 368–386.
18. Pocock SJ, Ariti CA, Collier TJ et al. The win ratio: A new approach to the analysis of composite endpoints in clinical trials based on clinical priorities. *Eur Heart J* 2011; **33**: 176–182.
19. Wang D and Pocock S. A win ratio approach to comparing continuous non-normal outcomes in clinical trials. *Pharm Stat* 2016; **15**: 238–245.
20. Gasparyan SB, Folkvaljon F, Bengtsson O et al. Adjusted win ratio with stratification: Calculation methods and interpretation. *Stat Methods Med Res* 2020; **0**: 1–32.
21. Scharfstein DO, Tsiatis AA and Robins JM. Semiparametric efficiency and its implication on the design and analysis of group-sequential studies. *J Am Stat Assoc* 1997; **92**: 1342–1350.
22. Cramér H and Wold H. Some theorems on distribution functions. *J Lond Math Soc* 1936; **s1-11**: 290–294.
23. Lévy P. *Calcul des probabilités, volume 9*. Paris: Gauthier-Villars Paris, 1925.
24. Ruymgaart FH (1980) A unified approach to the asymptotic distribution theory of certain midrank statistics. In Raoult JP (eds.) *Statistique non Paramétrique Asymptotique. Lecture Notes in Mathematics, Vol 821*. Springer, Berlin: Heidelberg. <https://doi.org/10.1007/BFb0097422>
25. Satterthwaite FE. An approximate distribution of estimates of variance components. *Biometrics Bull* 1946; **2**: 110–114.
26. Smith HF. The problem of comparing the results of two experiments with unequal errors. *J Council Sci Ind Res* 1936; **9**: 211–212.
27. Welch BL. The significance of the difference between two means when the population variances are unequal. *Biometrika* 1937; **29**: 350–362.
28. Pocock SJ. Group sequential methods in the design and analysis of clinical trials. *Biometrika* 1977; **64**: 191–199.
29. O'Brien PC and Fleming TR. A multiple testing procedure for clinical trials. *Biometrics* 1979; **35**: 549–556.
30. Lan KKG and DeMets DL. Discrete sequential boundaries for clinical trials. *Biometrika* 1983; **70**: 659–663.
31. Wassmer G and Pahlke F. *rpact: Confirmatory Adaptive Clinical Trial Design and Analysis*, 2020. <https://CRAN.R-project.org/package=rpact>. R package version 3.0.1.
32. Kappos L, Radue EW, O'Connor P et al. A placebo-controlled trial of oral fingolimod in relapsing multiple sclerosis. *N Engl J Med* 2010; **362**: 387–401.
33. Kurtzke JF. Rating neurologic impairment in multiple sclerosis: An expanded disability status scale (EDSS). *Neurology* 1983; **33**: 1444–1452.
34. Happ M, Bathke AC and Brunner E. Optimal sample size planning for the Wilcoxon-Mann-Whitney test. *Stat Med* 2019; **38**: 363–375.
35. Genz A, Bretz F, Miwa T et al. *mvtnorm: Multivariate Normal and t Distributions*, 2020. <https://CRAN.R-project.org/package=mvtnorm>. R package version 1.1-1.
36. Dunnett CW. A multiple comparison procedure for comparing several treatments with a control. *J Am Stat Assoc* 1955; **50**: 1096–1121.
37. Tukey J. Comparing individual means in the analysis of variance. *Biometrics* 1949; **5**: 99–114.
38. Gardner M. The paradox of the nontransitive dice and the elusive principle of indifference. *Sci Am: Math Games Column* 1970; **223**: 110–114.
39. Savage RP. The paradox of nontransitive dice. *Am Math Mon* 1994; **101**: 429–436.
40. Thangevelu K and Brunner E. Wilcoxon-Mann-Whitney test for stratified samples and Efron's paradox dice. *J Stat Plan Inference* 2007; **137**: 720–737.
41. Buyse M. Generalized pairwise comparisons of prioritized outcomes in the two-sample problem. *Stat Med* 2010; **29**: 3245–3257.
42. Cantagallo E, De Backer M, Kicinski M et al. A new measure of treatment effect in clinical trials involving competing risks based on generalized pairwise comparisons. *Biom J* 2021; **63**: 272–288.

43. Péron J, Buyse M, Ozenne B et al. An extension of generalized pairwise comparisons for prioritized outcomes in the presence of censoring. *Stat Methods Med Res* 2018; **27**: 1230–1239.
44. Buyse M and Peron J. Generalized pairwise comparisons for prioritized outcomes. In Piantadosi S and Meinert CL (eds.) *Principles and Practice of Clinical Trials*. Cham: Springer, 2020. pp. 1–25.
45. Hoeffding W. A class of statistics with asymptotically normal distributions. *Ann Stat* 1948; **19**: 293–325.
46. Lee AJ. *U-Statistics: Theory and Practice*. New York: Marcel Dekker, 1990.

THE NONPARAMETRIC BEHRENS-FISHER PROBLEM IN SMALL SAMPLES

Claus P. Nowak
Faculty of Statistics
TU Dortmund University
Dortmund, Germany

Markus Pauly
Faculty of Statistics
TU Dortmund University
Dortmund, Germany

Edgar Brunner
Department of Medical Statistics
University Medical Center Göttingen
Göttingen, Germany

ABSTRACT

While there appears to be a general consensus in the literature on the definition of the estimand and estimator associated with the Wilcoxon-Mann-Whitney test, it seems somewhat less clear as to how best to estimate the variance. In addition to the Wilcoxon-Mann-Whitney test, we review different proposals of variance estimators consistent under both the null hypothesis and the alternative. Moreover, in case of small sample sizes, an approximation of the distribution of the test statistic based on the t -distribution, a logit transformation and a permutation approach have been proposed. Focussing as well on different estimators of the degrees of freedom as regards the t -approximation, we carried out simulations for a range of scenarios, with results indicating that the performance of different variance estimators in terms of controlling the type I error rate largely depends on the heteroskedasticity pattern and the sample size allocation ratio, not on the specific type of distributions employed. By and large, a particular t -approximation together with Perme and Manevski's variance estimator best maintains the nominal significance level

Keywords Brunner-Munzel test, Wilcoxon-Mann-Whitney test

1 Introduction

In the biomedical context, nonparametric methods are frequently indicated by ordered categorical data such as pain or clinical severity scores. In order to nonparametrically test the null hypothesis of whether two unpaired samples produce similar outcomes, the Wilcoxon-Mann-Whitney test [Mann and Whitney, 1947, Wilcoxon, 1945, 1947] is arguably the one most commonly used in practice.

Usually, the estimand related to the Wilcoxon-Mann-Whitney test is defined as the probability

$$p = \mathbb{P}(X_1 < X_2) + 1/2 \cdot \mathbb{P}(X_1 = X_2),$$

where $X_1 \sim F_1$ and $X_2 \sim F_2$ denote two independent random variables corresponding to the two samples. The quantity p is referred to as nonparametric relative effect of X_2 with respect to X_1 [Brunner and Munzel, 2000, Brunner et al., 2018], probabilistic index [Thas et al., 2012] or Mann-Whitney parameter [Fay et al., 2018]. In the setting of a parallel two-arm clinical trial, one may regard the random variables X_1 and X_2 as responses from treatment arms 1 and 2 respectively. Assuming that lower values imply a more beneficial outcome, one may interpret p as the probability that a patient on arm 1 will fare better than one on arm 2, including $1/2$ times the probability of equal outcomes.

While the literature seems to agree that the most suitable estimator of p , which we will refer to as \hat{p} , are the corresponding relative frequencies arising from all pairwise comparisons of the sample data, the question of how best to estimate the variance of \hat{p} does not appear to be quite that settled.

The variance estimator employed in the Wilcoxon-Mann-Whitney test is unbiased and consistent, but only under the assumption of equal distributions, i.e., $F_1 = F_2$. Hence the Wilcoxon-Mann-Whitney test can neither be inverted to produce confidence intervals nor does it directly address the nonparametric Behrens-Fisher problem. In that regard, assume both distribution F_1 and F_2 are symmetric with the same centre of symmetry but heteroskedastic such as two normal distributions with the same expectation but different variances, yielding a Mann-Whitney parameter of $p = 1/2$.

To test the null hypothesis $H_0 : p = 1/2$, Shirahata [1993] considers a number of variance estimators of \hat{p} under the assumption that both distributions F_1 and F_2 are continuous while Bamber [1975] proposed an unbiased variance estimator for general F_1 and F_2 , be they continuous, discrete or neither. Moreover, DeLong et al. [1988], Brunner and Munzel [2000] as well as Perme and Manevski [2019] put forward variance estimators consistent for arbitrary F_1 and F_2 as well.

In small samples, Brunner and Munzel [2000] suggest the use of a t -approximation analogous to the Satterthwaite-Smith-Welch approach [Satterthwaite, 1946, Smith, 1936, Welch, 1937] as regards the parametric Behrens-Fisher problem. Using different degrees of freedom and different variance estimators allowing for ties, we carry out simulation studies for a range of scenarios to gauge the performance of the resulting tests in terms of the type I error rate and power. In addition, we consider a permutation test proposed by Pauly et al. [2016].

This manuscript proceeds as follows. In Section 2 we review nonparametric theory and give definitions of the test statistics, whose empirical behaviour we examine as regards type I error rates and power in Section 3 and close with a discussion of the results in Section 4. All proofs and derivations as well as more and more detailed simulation results are given in the appendix.

2 Nonparametric model

We start with notation from nonparametric theory convenient for stating variance formulas and test statistics. Then we go over the variance associated with the Wilcoxon-Mann-Whitney test, as well as variance estimators consistent for arbitrary distributions F_1 and F_2 . For sake of completeness, we will also briefly mention Shirahata's [1993] formulas. With the asymptotic normality of the resulting test statistics already established [Brunner et al., 2018], we will discuss different estimators for the degrees of freedom in a small sample t -approximation as well as the permutation approach developed by Pauly et al. [2016].

2.1 Notation

Let X denote a univariate random variable defined on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$, which stands for real-valued or ordered categorical responses. As is commonly done, we call

$$\begin{aligned} F^-(x) &= \mathbb{P}(X < x) && \text{the left-continuous,} \\ F^+(x) &= \mathbb{P}(X \leq x) && \text{the right-continuous,} \\ F(x) &= \mathbb{P}(X < x) + 1/2 \cdot \mathbb{P}(X = x) && \text{the normalised} \end{aligned}$$

version of the cumulative distribution function of X [Lévy, 1925, Ruymgaart, 1980, Brunner and Munzel, 2000].

For a particular sample of observations $X_1, \dots, X_n \stackrel{iid}{\sim} F$, we further denote by

$$\hat{F}(x) = \frac{1}{n} \sum_{j=1}^n c(x, X_j), \quad c(x, X_j) = \begin{cases} 0 & \text{if } x < X_j \\ 1/2 & \text{if } x = X_j \\ 1 & \text{if } x > X_j \end{cases}$$

the normalised version of the empirical cumulative distribution function. Moreover, we call

$$R_i = 1/2 + \sum_{j=1}^n c(X_i, X_j), \quad i = 1, \dots, n,$$

the mid-rank of X_i among the observations X_1, \dots, X_n .

As for two independent random variables $X_1 \sim F_1$ and $X_2 \sim F_2$, the Mann-Whitney parameter as given in the Introduction has the following integral representation, i.e.,

$$p = \mathbb{P}(X_1 < X_2) + 1/2 \cdot \mathbb{P}(X_1 = X_2) = \int F_1 dF_2.$$

We say that

- X_1 tends to smaller values than X_2 if $p > 1/2$,
- X_1 tends to larger values than X_2 if $p < 1/2$,

- X_1 and X_2 are stochastically comparable if $p = 1/2$.

For a more comprehensive treatment of nonparametric theory we refer to Brunner et al. [2018].

Throughout the remainder of this manuscript we will focus on a parallel two-arm clinical trial with responses

$$\begin{aligned} X_{1i} &\stackrel{iid}{\sim} F_1, \quad i = 1, \dots, n_1, \\ X_{2j} &\stackrel{iid}{\sim} F_2, \quad j = 1, \dots, n_2, \end{aligned}$$

from treatment arms 1 and 2 respectively. With $N = n_1 + n_2$, we can estimate the nonparametric relative effect p by

$$\hat{p} = \int \hat{F}_1 d\hat{F}_2 = \frac{1}{n_1} \frac{1}{n_2} \sum_{j=1}^{n_2} \sum_{i=1}^{n_1} c(X_{2j}, X_{1i}) = \frac{1}{N} (\bar{R}_{2\bullet} - \bar{R}_{1\bullet}) + 1/2,$$

with $\bar{R}_{g\bullet} = \frac{1}{n_g} \sum_{i=1}^{n_g} R_{gi}$, $g = 1, 2$, where R_{gi} is the mid-rank of X_{gi} among all N observations $X_{11}, \dots, X_{1n_1}, X_{21}, \dots, X_{2n_2}$.

To address the nonparametric Behrens-Fisher problem, we consider the null hypothesis $H_0 : p = 1/2$ against $H_1 : p \neq 1/2$. Unsurprisingly, all resulting test statistics are based on the deviation of the Mann-Whitney parameter estimate from $1/2$, i.e.,

$$\hat{p} - 1/2.$$

For asymptotic results, we let both sample sizes tend to infinity such that neither vanishes, i.e., $n_g/N \rightarrow \gamma_g > 0$ for both $n_1 \rightarrow \infty$ and $n_2 \rightarrow \infty$, $g = 1, 2$. Moreover, we assume $0 < p < 1$ and that there exists no x such that $\mathbb{P}(X_{11} = x) = 1$ or $\mathbb{P}(X_{21} = x) = 1$, i.e., excluding the degenerate cases of completely separated samples and one-point distributions.

To obtain asymptotic standard normal test statistics, it remains to define suitable estimators of $\mathbb{V}(\hat{p})$, which is the purpose of the next subsection.

2.2 Variance estimators

Under the assumption of equal distributions, i.e., $F = F_1 = F_2$, the variance estimand $\mathbb{V}(\hat{p})$ takes the following form,

$$\sigma_{WMW}^2 = \frac{\sigma_R^2}{N n_1 n_2}, \quad \text{with } \sigma_R^2 = N \left\{ (N-2) \int F^2 dF - \frac{N-3}{4} \right\} - \frac{N}{4} \int (F^+ - F^-) dF.$$

If $F = F_1 = F_2$ holds, a consistent and unbiased estimator of the variance σ_{WMW}^2 is given by

$$\hat{\sigma}_{WMW}^2 = \frac{\hat{\sigma}_R^2}{N n_1 n_2}, \quad \text{with } \hat{\sigma}_R^2 = \frac{N^3}{N-1} \left(\int \hat{F}^2 d\hat{F} - 1/4 \right) = \frac{1}{N} \sum_{g=1}^2 \sum_{i=1}^{n_g} \left(R_{gi} - \frac{N+1}{2} \right)^2, \quad (1)$$

resulting in $T_{WMW} = (\hat{p} - 1/2) / \hat{\sigma}_{WMW} \xrightarrow{D} \mathcal{N}(0, 1)$, which is nothing but the Wilcoxon-Mann-Whitney test allowing for ties [Brunner et al., 2018]. In the context of this manuscript, we feel it more tenable to regard the Wilcoxon-Mann-Whitney test as a way of testing the null hypothesis formulated in terms of p , i.e., $H_0 : p = 1/2$, whereas $F_1 = F_2$ amounts to an additional assumption on the model under the null.

As for arbitrary distributions F_1 and F_2 , Bamber [1975] as well as [Brunner et al., 2021a] provide a formula of the variance estimand $\mathbb{V}(\hat{p})$, which reads in our notation as

$$\sigma_N^2 = \frac{\tau_0 + (n_2 - 1)\tau_1 + (n_1 - 1)\tau_2 - (n_1 + n_2 - 1)p^2}{n_1 n_2},$$

where $\tau_0 = p - 1/4 \cdot \int (F_1^+ - F_1^-) dF_2$, $\tau_1 = \int (1 - F_2)^2 dF_1$, and $\tau_2 = \int F_1^2 dF_2$.

Bamber [1975] as well as Brunner et al. [2021a] propose an unbiased variance estimator of σ_N^2 as well, namely,

$$\hat{\sigma}_N^2 = \frac{n_2 \hat{\tau}_1 + n_1 \hat{\tau}_2 - \hat{\tau}_0 - (n_1 + n_2 - 1) \hat{p}^2}{(n_1 - 1)(n_2 - 1)}, \quad (2)$$

with $\hat{\tau}_0 = \hat{p} - 1/4 \cdot \int (\hat{F}_1^+ - \hat{F}_1^-) d\hat{F}_2$, $\hat{\tau}_1 = \int (1 - \hat{F}_2)^2 d\hat{F}_1$, and $\hat{\tau}_2 = \int \hat{F}_1^2 d\hat{F}_2$. For a computationally more efficient expression of $\hat{\sigma}_N^2$ in terms of ranks, see Brunner et al. [2021a].

Brunner and Munzel [2000] derived an estimator of σ_N^2 similar in structure to the variance estimator of the two-sample t -test under heteroskedasticity, i.e.,

$$\hat{\sigma}_{BM}^2 = \frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}, \text{ where } \hat{\sigma}_1^2 = \frac{n_1}{n_1 - 1}(\hat{\tau}_1 - \hat{p}^2) \text{ and } \hat{\sigma}_2^2 = \frac{n_2}{n_2 - 1}(\hat{\tau}_2 - \hat{p}^2), \quad (3)$$

with $\hat{\tau}_1$ and $\hat{\tau}_2$ as given in (2). For a computationally more efficient rank representation of $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$, see Brunner and Munzel [2000]. Note that the estimator $\hat{\sigma}_{BM}^2$ is identical to the one given in DeLong et al. [1988].

Perme and Manevski [2019] propose yet another estimator for σ_N^2 , which they refer to as exact, i.e.,

$$\hat{\sigma}_{PM}^2 = \frac{\hat{p}(1 - \hat{p}) + (n_2 - 1)\hat{\sigma}_1^2 + (n_1 - 1)\hat{\sigma}_2^2}{n_1 n_2}, \quad (4)$$

with $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$ as just defined in (3).

In the Introduction, we have vaguely hinted at the consistency of the variance estimators. More precisely, the dominating terms $\hat{\tau}_1$, $\hat{\tau}_2$, \hat{p}^2 are consistent for τ_1 , τ_2 , p^2 and they converge, as weighted with the sample sizes in $\hat{\sigma}_N^2$ (2), $\hat{\sigma}_{BM}^2$ (3), $\hat{\sigma}_{PM}^2$ (4), to zero in probability with the same speed.

2.3 Shirahata's formulas for continuous distributions

Assuming that ties cannot occur almost surely, Shirahata [1993] discusses the following four estimators, i.e., an unbiased one, a bootstrap estimator, an estimator by Fligner and Policello [1981], and a jackknife estimator, which in our notation read as

$$\begin{aligned} \hat{\sigma}_U^2 &= \frac{n_2 \int (1 - \hat{F}_2^-)^2 d\hat{F}_1 + n_1 \int (\hat{F}_1^+)^2 d\hat{F}_2 - \int \hat{F}_1^+ d\hat{F}_2 - (n_1 + n_2 - 1) \int \hat{F}_1^+ d\hat{F}_2^2}{(n_1 - 1)(n_2 - 1)}, \\ \hat{\sigma}_B^2 &= \frac{(n_2 - 1) \int (1 - \hat{F}_2^-)^2 d\hat{F}_1 + (n_1 - 1) \int (\hat{F}_1^+)^2 d\hat{F}_2 + \int \hat{F}_1^+ d\hat{F}_2 - (n_1 + n_2 - 1) \int \hat{F}_1^+ d\hat{F}_2^2}{n_1 n_2}, \\ \hat{\sigma}_{FP}^2 &= \frac{\int (1 - \hat{F}_2^-)^2 d\hat{F}_1}{n_1} + \frac{\int (\hat{F}_1^+)^2 d\hat{F}_2}{n_2} - \frac{\int \hat{F}_1^+ d\hat{F}_2 + (n_1 + n_2 + 1) \int \hat{F}_1^+ d\hat{F}_2^2}{n_1 n_2}, \\ \hat{\sigma}_J^2 &= \frac{\int (1 - \hat{F}_2^-)^2 d\hat{F}_1}{n_1 - 1} + \frac{\int (\hat{F}_1^+)^2 d\hat{F}_2}{n_2 - 1} - \frac{(n_1 + n_2 - 2) \int \hat{F}_1^+ d\hat{F}_2^2}{(n_1 - 1)(n_2 - 1)}. \end{aligned}$$

In case of continuous distributions, however, it follows $\int \hat{F}_1^+ d\hat{F}_2 = \int \hat{F}_1 d\hat{F}_2 = \hat{p} = \hat{\tau}_0$, $\int (1 - \hat{F}_2^-)^2 d\hat{F}_1 = \int (1 - \hat{F}_2)^2 d\hat{F}_1 = \hat{\tau}_1$, $\int (\hat{F}_1^+)^2 d\hat{F}_2 = \int \hat{F}_1^2 d\hat{F}_2 = \hat{\tau}_2$, so that then we can express the variance estimators as follows

$$\begin{aligned} \hat{\sigma}_U^2 &= \frac{n_2 \hat{\tau}_1 + n_1 \hat{\tau}_2 - \hat{\tau}_0 - (n_1 + n_2 - 1) \hat{p}^2}{(n_1 - 1)(n_2 - 1)} = \hat{\sigma}_N^2, \\ \hat{\sigma}_B^2 &= \frac{(n_2 - 1) \hat{\tau}_1 + (n_1 - 1) \hat{\tau}_2 + \hat{\tau}_0 - (n_1 + n_2 - 1) \hat{p}^2}{n_1 n_2}, \\ \hat{\sigma}_{FP}^2 &= \frac{\hat{\tau}_1}{n_1} + \frac{\hat{\tau}_2}{n_2} - \frac{\hat{\tau}_0 + (n_1 + n_2 + 1) \hat{p}^2}{n_1 n_2}, \\ \hat{\sigma}_J^2 &= \frac{\hat{\tau}_1}{n_1 - 1} + \frac{\hat{\tau}_2}{n_2 - 1} - \frac{(n_1 + n_2 - 2) \hat{p}^2}{(n_1 - 1)(n_2 - 1)} = \hat{\sigma}_{BM}^2. \end{aligned}$$

2.4 Degrees of freedom

Analogous to the parametric Behrens-Fisher problem, Brunner and Munzel [2000] propose a Satterthwaite-Smith-Welch t -approximation [Satterthwaite, 1946, Smith, 1936, Welch, 1937] for small samples with degrees of freedom estimated by

$$df = \frac{\{\hat{\sigma}_1^2/n_1 + \hat{\sigma}_2^2/n_2\}^2}{\hat{\sigma}_1^4/\{n_1^2(n_1 - 1)\} + \hat{\sigma}_2^4/\{n_2^2(n_2 - 1)\}}, \quad (5)$$

where $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$ are defined as in (3). In addition, we will consider the degrees of freedom

$$df_1 = \frac{\{\hat{\sigma}_1^2/(n_1 - 1) + \hat{\sigma}_2^2/(n_2 - 1)\}^2}{\hat{\sigma}_1^4/\{(n_1 - 1)^2(n_1 - 2)\} + \hat{\sigma}_2^4/\{(n_2 - 1)^2(n_2 - 2)\}}, \quad (6)$$

$$df_2 = \frac{\{\hat{\sigma}_1^2/(n_1 - 2) + \hat{\sigma}_2^2/(n_2 - 2)\}^2}{\hat{\sigma}_1^4/\{(n_1 - 2)^2(n_1 - 3)\} + \hat{\sigma}_2^4/\{(n_2 - 2)^2(n_2 - 3)\}}, \quad (7)$$

$$df_3 = \frac{2}{1/(n_1 - 1) + 1/(n_2 - 1)}, \quad (8)$$

$$df_4 = \frac{\hat{\sigma}_N^4}{\hat{\sigma}_{1|N}^4/(n_1 - 1) + \hat{\sigma}_{2|N}^4/(n_2 - 1)}, \quad (9)$$

with $\hat{\sigma}_N^2$ as in (2) and $\hat{\sigma}_{1|N}^2 = \frac{n_2\hat{\tau}_1 - 1/2\hat{\tau}_0 - (n_2 - 1/2)\hat{p}^2}{(n_1 - 1)(n_2 - 1)}$, $\hat{\sigma}_{2|N}^2 = \frac{n_1\hat{\tau}_2 - 1/2\hat{\tau}_0 - (n_1 - 1/2)\hat{p}^2}{(n_1 - 1)(n_2 - 1)}$.

The intuition behind using (6) and (7) in small samples is similar to (5), we merely assume that there were one (or two) fewer observations in each of the two groups, with $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$ remaining unchanged as in (5). On the other hand, formulas (8) and (9) are loosely based on a Box-type [Box, 1954, Brunner et al., 2018] approximation as regards the unbiased variance estimator.

Another way to address the liberal behaviour of the tests is to employ a variance stabilising transformation, such as the *logit* function, or a permutation approach as described in Section 2.5 [Brunner et al., 2018, Pauly et al., 2016].

2.5 Test statistics

Collecting the test statistics with regard to the null hypothesis $H_0 : p = 1/2$ allowing for ties, we have

$$T_{WMW} = (\hat{p} - 1/2)/\hat{\sigma}_{WMW}, \quad (10)$$

$$T_N = (\hat{p} - 1/2)/\hat{\sigma}_N, \quad (11)$$

$$T_{BM} = (\hat{p} - 1/2)/\hat{\sigma}_{BM}, \quad (12)$$

$$T_{PM} = (\hat{p} - 1/2)/\hat{\sigma}_{PM}. \quad (13)$$

For the Wilcoxon-Mann-Whitney test (10) we use the standard normal distribution to compute p -values, for the other tests, (11) to (13), a central t -distribution with degrees of freedom given as in (5) to (9). As already alluded to, we will additionally consider the following test statistics based on a *logit* transformation using the delta method, i.e.,

$$T_N^{logit} = \hat{p}(1 - \hat{p}) \cdot \ln\{\hat{p}/(1 - \hat{p})\}/\hat{\sigma}_N, \quad (14)$$

$$T_{BM}^{logit} = \hat{p}(1 - \hat{p}) \cdot \ln\{\hat{p}/(1 - \hat{p})\}/\hat{\sigma}_{BM}, \quad (15)$$

$$T_{PM}^{logit} = \hat{p}(1 - \hat{p}) \cdot \ln\{\hat{p}/(1 - \hat{p})\}/\hat{\sigma}_{PM}. \quad (16)$$

As with the Wilcoxon-Mann-Whitney test, we employ the standard normal distribution to obtain p -values for (14) to (16).

Moreover, we will make use of the studentised permutation approach suggested by Pauly et al. [2016]. To this end, we randomly allocate n_1 out of the entire $N = n_1 + n_2$ observations from the whole sample as originating from the first distribution F_1 , with the remaining n_2 responses regarded as having been drawn from F_2 . Repeating this procedure, say $n_{perm} = 10000$ times, and computing one of the test statistics as in (11) to (16) each time, we obtain a permutation distribution on which to base rejection of the null hypothesis. More formally, we relabel the entire data $(X_{11}, \dots, X_{1n_1}, X_{21}, \dots, X_{2n_2}) =: (X_1, \dots, X_N)$ and define a random variable π uniformly distributed on S_N which is the set of all permutations of $1, \dots, N$. For a particular data set at hand, we then use the permuted pooled sample $(X_{\pi(1)}, \dots, X_{\pi(N)})$ – with the first n_1 and last n_2 components considered as belonging to samples 1 and 2 respectively – to compute (11) to (16), yielding the permuted versions

$$\tilde{T}_N, \tilde{T}_{BM}, \tilde{T}_{PM}, \tilde{T}_N^{logit}, \tilde{T}_{BM}^{logit}, \tilde{T}_{PM}^{logit}. \quad (17)$$

With T_N denoting the test statistic as in (11) based on the original data and $\tilde{T}_N^1, \dots, \tilde{T}_N^{n_{perm}}$ the corresponding test statistics of the n_{perm} random permutations, we calculate the two-sided p -value as follows,

$$2 \min(p_1, p_2), \text{ where } p_1 = \frac{1}{n_{perm}} \sum_{k=1}^{n_{perm}} \mathbf{1}(\tilde{T}_N^k \leq T_N) \text{ and } p_2 = \frac{1}{n_{perm}} \sum_{\ell=1}^{n_{perm}} \mathbf{1}(\tilde{T}_N^\ell \geq T_N).$$

As for the other test statistics, the permutation based p -values are computed in a completely analogous manner.

3 Simulations

As the methods treated in Section 2 are of asymptotic nature, we explore their applicability for finite sample sizes in a range of scenarios. In that regard, we consider the null hypothesis $H_0 : p = 1/2$ against $H_1 : p \neq 1/2$ at a two-sided nominal significance level of $\alpha = 0.05$. We first present simulation results for the asymptotic tests as defined in (10) to (16). As far as the test statistics (11) to (13) are concerned, we only report rejection rates for degrees of freedom df_2 (7) in the main manuscript as they outperformed the other versions. Simulations of permutation tests as in (17) being computationally much more expensive, we restrict our focus to some select scenarios as outlined in Section 3.2.

In extreme cases as alluded to earlier, some variance estimates might actually turn out to be zero or negative as would be the case in two completely separated samples. Since this happened very rarely and has virtually no bearing on the results, we relegate the discussion of exception handling to the appendix.

3.1 Asymptotic tests

First we generate data from normal distributions, namely $X_{gi} \stackrel{iid}{\sim} \mathcal{N}(\mu_g, \sigma_g^2)$, $g = 1, 2$, $i = 1, \dots, n_g$. To gauge the type I error rate of the different tests, we set $\mu_1 = \mu_2 = 0$ and perform 100 000 simulation runs for each scenario, giving rise to a Monte Carlo error of about 0.0006 based on a 95%-precision interval for a nominal significance level of $\alpha = 0.05$. The results depicted in Table 1 indicate that Perme and Manveski's test T_{PM} (13) with df_2 degrees of freedom (7) best maintains the nominal significance level, especially for $\min(n_1, n_2) \geq 15$, although the difference from T_N (11) as well as T_{BM} (12) is not particularly pronounced. However, in the heteroskedastic settings, the Wilcoxon-Mann-Whitney test T_{WMW} (10) is generally either far too liberal or far too conservative depending on sample size allocation. More precisely, if more patients are allocated to the arm producing less dispersed outcomes, then T_{WMW} (10) becomes too liberal, and too conservative otherwise. While *logit* based tests, (14) to (16), virtually never exceed the nominal significance level, they exhibit a somewhat conservative tendency in many cases. In that regard, we only present power graphs for the tests T_{WMW} (10) and T_N (11), T_{BM} (12), T_{PM} (13) with df_2 degrees of freedom (7) as set forth in Figure 1. Unlike before, the power graphs are based on only 10 000 simulation runs per scenario.

Now we choose an ordinal 5-point-distribution with categories $\mathcal{C}_1 < \mathcal{C}_2 < \mathcal{C}_3 < \mathcal{C}_4 < \mathcal{C}_5$. As in Brunner et al. [2021b], the probabilities of each category occurring are derived through a latent Beta distribution, i.e., we consider Beta random variables $Y_{gi} \stackrel{iid}{\sim} \mathcal{B}(\alpha_g, \beta_g)$, $g = 1, 2$, $i = 1, \dots, n_g$, with shape parameters $\alpha_g, \beta_g > 0$, such that the expectation and variance of Y_{gi} are given by

$$\mathbb{E}(Y_{gi}) = \frac{\alpha_g}{\alpha_g + \beta_g}, \quad \mathbb{V}(Y_{gi}) = \frac{\alpha_g \beta_g}{(\alpha_g + \beta_g)^2 (\alpha_g + \beta_g + 1)}.$$

Then we discretise Y_{gi} to the random variable X_{gi} , $g = 1, 2$, $i = 1, \dots, n_g$, as follows

$$X_{gi} = \mathcal{C}_k \text{ if } Y_{gi} \in [0.2(k-1), 0.2k[\text{ for } k = 1, \dots, 5.$$

Consequently, the probability mass function of X_{gi} is nothing but

$$\mathbb{P}(X_{gi} = \mathcal{C}_k) = \mathbb{P}(0.2(k-1) \leq Y_{gi} < 0.2k) \text{ for } k = 1, \dots, 5.$$

Analogous to the normal setting, we consider a homo- and a heteroskedastic scenario as outlined in Table 2. As before, T_{PM} (13) with df_2 degrees of freedom (7) best controls the nominal type I error rate. Moreover, the power graphs in Figure 2 based on 10 000 simulation runs show a pattern similar to the normal scenarios as well.

3.2 Permutation tests

As for the type I error rate of the permutation tests (17) based on the approach proposed by Pauly et al. [2016], we examine some of the scenarios as set out in Tables 1 and 2 using $n_{perm} = 10\,000$ random permutations per simulation run. Bearing in mind that 10 000 simulation runs for each setting give rise to a Monte Carlo error of 0.002 for a two-sided nominal significance level of $\alpha = 0.05$, it still seems fair to us to observe in light of the results depicted in Tables 3 and 4 that Perme and Manevski's original T_{PM} (13) with degrees of freedom df_2 (7) better maintains the nominal significance level on the whole.

More results as regards similar settings as in Pauly et al. [2016], i.e., exponential and binomial distributions, as well as some power scenarios for normal and 5-point-distributions are provided in the appendix.

Table 1: Type I error rates for normal distributions $F_1 = \mathcal{N}(0, \sigma_1^2)$ and $F_2 = \mathcal{N}(0, \sigma_2^2)$ based on 100 000 replications at a two-sided nominal significance level of $\alpha = 0.05$ as regards the test statistics T_{WMW} (10); T_N (11), T_{BM} (12), T_{PM} (13) with degrees of freedom df_2 (7); T_N^{Logit} (14), T_{BM}^{Logit} (15), T_{PM}^{Logit} (16)

n_1	n_2	σ_1	σ_2	T_{WMW}	T_N	T_{BM}	T_{PM}	T_N^{Logit}	T_{BM}^{Logit}	T_{PM}^{Logit}
7	7	1	1	0.05318	0.05527	0.04796	0.04304	0.02886	0.02318	0.01860
10	7	1	1	0.04348	0.05428	0.05003	0.04725	0.03545	0.02981	0.02441
7	10	1	1	0.04290	0.05399	0.04975	0.04708	0.03460	0.02899	0.02386
10	10	1	1	0.05320	0.05696	0.05225	0.04856	0.03583	0.02993	0.02710
15	15	1	1	0.05072	0.05651	0.05290	0.05012	0.04067	0.03691	0.03435
30	15	1	1	0.04906	0.05417	0.05183	0.05001	0.04498	0.04234	0.04050
15	30	1	1	0.04911	0.05431	0.05207	0.05004	0.04504	0.04240	0.04044
30	30	1	1	0.04950	0.05306	0.05138	0.04978	0.04510	0.04316	0.04163
15	45	1	1	0.04891	0.05340	0.05167	0.05040	0.04841	0.04624	0.04461
15	60	1	1	0.04889	0.05192	0.05044	0.04945	0.04957	0.04784	0.04641
15	75	1	1	0.04959	0.05292	0.05144	0.05046	0.05219	0.05059	0.04944
45	15	1	1	0.04945	0.05329	0.05150	0.05002	0.04835	0.04644	0.04489
60	15	1	1	0.04930	0.05262	0.05120	0.04995	0.05003	0.04844	0.04697
75	15	1	1	0.04918	0.05164	0.05055	0.04955	0.05100	0.04978	0.04859
7	7	1	3	0.07223	0.04572	0.04206	0.03917	0.02785	0.02442	0.02135
10	7	1	3	0.08066	0.04509	0.04320	0.04175	0.03273	0.02984	0.02716
7	10	1	3	0.04122	0.05091	0.04734	0.04540	0.03058	0.02694	0.02365
10	10	1	3	0.07141	0.05172	0.04893	0.04705	0.03310	0.02961	0.02835
15	15	1	3	0.06833	0.05235	0.05036	0.04926	0.03822	0.03577	0.03434
30	15	1	3	0.10568	0.05111	0.04995	0.04919	0.04008	0.03887	0.03810
15	30	1	3	0.03163	0.05299	0.05111	0.04973	0.04355	0.04140	0.04013
30	30	1	3	0.07001	0.05308	0.05218	0.05124	0.04572	0.04468	0.04372
15	45	1	3	0.01618	0.05185	0.04994	0.04859	0.04490	0.04314	0.04196
15	60	1	3	0.00965	0.05130	0.04978	0.04853	0.04642	0.04473	0.04332
15	75	1	3	0.00616	0.05263	0.05114	0.04978	0.04871	0.04720	0.04593
45	15	1	3	0.12749	0.05153	0.05080	0.05041	0.04236	0.04145	0.04075
60	15	1	3	0.14104	0.05116	0.05044	0.05000	0.04224	0.04145	0.04093
75	15	1	3	0.14588	0.05068	0.05009	0.04973	0.04233	0.04171	0.04134
7	7	1	5	0.08821	0.03694	0.03496	0.03319	0.02367	0.02168	0.02000
10	7	1	5	0.09850	0.03648	0.03563	0.03504	0.02702	0.02500	0.02380
7	10	1	5	0.04932	0.04573	0.04335	0.04229	0.02883	0.02666	0.02399
10	10	1	5	0.08485	0.04618	0.04453	0.04376	0.03061	0.02890	0.02805
15	15	1	5	0.08132	0.05108	0.04987	0.04928	0.03690	0.03543	0.03458
30	15	1	5	0.12597	0.05022	0.04948	0.04907	0.03803	0.03722	0.03669
15	30	1	5	0.03419	0.05207	0.05079	0.04985	0.04309	0.04149	0.04059
30	30	1	5	0.08136	0.05257	0.05186	0.05128	0.04465	0.04387	0.04345
15	45	1	5	0.01548	0.05039	0.04890	0.04800	0.04333	0.04222	0.04151
15	60	1	5	0.00770	0.05111	0.04984	0.04898	0.04580	0.04473	0.04347
15	75	1	5	0.00431	0.05116	0.05003	0.04899	0.04731	0.04634	0.04530
45	15	1	5	0.15064	0.05092	0.05044	0.05015	0.03894	0.03833	0.03791
60	15	1	5	0.16777	0.05023	0.04993	0.04979	0.03828	0.03769	0.03731
75	15	1	5	0.17407	0.05004	0.04986	0.04961	0.03914	0.03882	0.03852

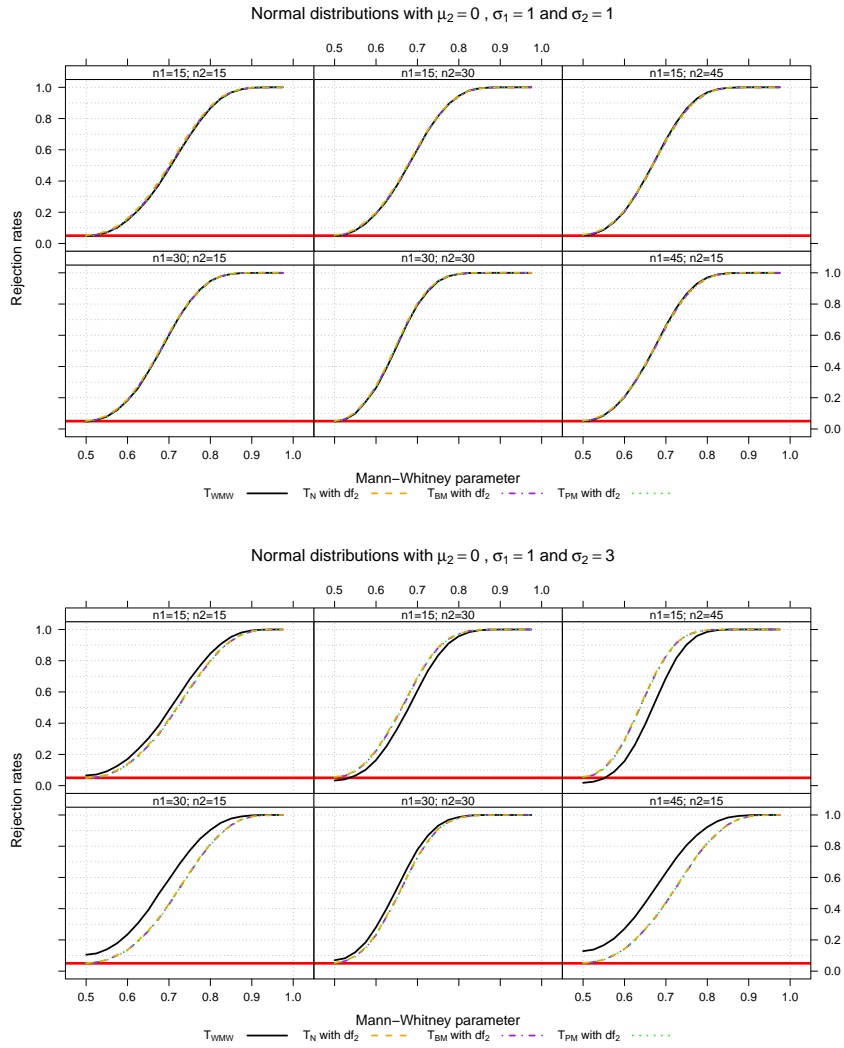


Figure 1: Power graphs for normal distributions based on 10 000 simulation runs

Table 2: Type I error rates for the 5-point distributions with latent $F_1 = \mathcal{B}(\alpha_1, \beta_1)$ and $F_2 = \mathcal{B}(5, 4)$ based on 100 000 replications at a two-sided nominal significance level of $\alpha = 0.05$ as regards the test statistics T_{WMW} (10); T_N (11), T_{BM} (12), T_{PM} (13) with degrees of freedom df_2 (7); T_N^{Logit} (14), T_{BM}^{Logit} (15), T_{PM}^{Logit} (16)

n_1	n_2	α_1	β_1	T_{WMW}	T_N	T_{BM}	T_{PM}	T_N^{Logit}	T_{BM}^{Logit}	T_{PM}^{Logit}
7	7	5	4	0.04611	0.05628	0.05045	0.04129	0.03990	0.03889	0.02723
10	7	5	4	0.04761	0.05391	0.05298	0.04431	0.04309	0.04004	0.03127
7	10	5	4	0.04769	0.05457	0.05350	0.04426	0.04320	0.04016	0.03184
10	10	5	4	0.04832	0.05450	0.05316	0.04798	0.04072	0.03944	0.03263
15	15	5	4	0.04875	0.05440	0.05315	0.04910	0.04330	0.04208	0.03741
30	15	5	4	0.04814	0.05212	0.05140	0.04825	0.04591	0.04513	0.04194
15	30	5	4	0.04902	0.05391	0.05315	0.04990	0.04743	0.04668	0.04344
30	30	5	4	0.04857	0.05175	0.05115	0.04875	0.04577	0.04521	0.04266
15	45	5	4	0.04923	0.05258	0.05201	0.04961	0.04996	0.04937	0.04675
15	60	5	4	0.05026	0.05292	0.05248	0.05041	0.05288	0.05231	0.05004
15	75	5	4	0.04959	0.05263	0.05228	0.05055	0.05426	0.05396	0.05188
45	15	5	4	0.04898	0.05186	0.05144	0.04909	0.04942	0.04892	0.04610
60	15	5	4	0.04868	0.05194	0.05142	0.04940	0.05191	0.05136	0.04899
75	15	5	4	0.04856	0.05090	0.05054	0.04906	0.05263	0.05217	0.05033
7	7	1.2071	1	0.05763	0.05126	0.04755	0.04387	0.03552	0.03366	0.02719
10	7	1.2071	1	0.04264	0.05281	0.05086	0.04506	0.03508	0.03298	0.02710
7	10	1.2071	1	0.07446	0.05213	0.05046	0.04612	0.04219	0.03910	0.03381
10	10	1.2071	1	0.05798	0.05432	0.05293	0.04872	0.03863	0.03638	0.03171
15	15	1.2071	1	0.05579	0.05146	0.05031	0.04752	0.04035	0.03912	0.03606
30	15	1.2071	1	0.03218	0.05345	0.05251	0.05029	0.04528	0.04408	0.04154
15	30	1.2071	1	0.08897	0.05188	0.05131	0.04977	0.04529	0.04446	0.04248
30	30	1.2071	1	0.05827	0.05120	0.05065	0.04917	0.04508	0.04436	0.04264
15	45	1.2071	1	0.10304	0.05027	0.04976	0.04876	0.04582	0.04528	0.04399
15	60	1.2071	1	0.11529	0.05152	0.05116	0.05029	0.04857	0.04813	0.04702
15	75	1.2071	1	0.12079	0.05096	0.05064	0.05008	0.04851	0.04815	0.04739
45	15	1.2071	1	0.02028	0.05296	0.05196	0.04966	0.04718	0.04624	0.04375
60	15	1.2071	1	0.01493	0.05145	0.05075	0.04874	0.04806	0.04725	0.04507
75	15	1.2071	1	0.01190	0.05180	0.05111	0.04917	0.05009	0.04916	0.04721

Table 3: Type I error rates for normal distributions $F_1 = \mathcal{N}(0, \sigma_1^2)$ and $F_2 = \mathcal{N}(0, \sigma_2^2)$ at a two-sided nominal significance level of $\alpha = 0.05$ for the studentised permutation tests as given in (17) based on 10 000 random permutations for each of the 10 000 replications

n_1	n_2	σ_1	σ_2	\tilde{T}_N	\tilde{T}_{BM}	\tilde{T}_{PM}	\tilde{T}_N^{Logit}	\tilde{T}_{BM}^{Logit}	\tilde{T}_{PM}^{Logit}
7	7	1	1	0.0486	0.0492	0.0492	0.0492	0.0482	0.0485
7	10	1	1	0.0473	0.0484	0.0486	0.0471	0.0485	0.0484
10	7	1	1	0.0489	0.0501	0.0502	0.0501	0.0509	0.0498
10	10	1	1	0.0505	0.0507	0.0507	0.0510	0.0505	0.0504
15	15	1	1	0.0507	0.0506	0.0507	0.0500	0.0503	0.0505
15	30	1	1	0.0525	0.0525	0.0523	0.0526	0.0526	0.0526
30	15	1	1	0.0494	0.0494	0.0495	0.0496	0.0496	0.0494
30	30	1	1	0.0526	0.0525	0.0525	0.0523	0.0523	0.0525
15	45	1	1	0.0529	0.0528	0.0529	0.0520	0.0525	0.0525
45	15	1	1	0.0509	0.0510	0.0507	0.0508	0.0510	0.0512
7	7	1	3	0.0477	0.0546	0.0553	0.0381	0.0388	0.0396
7	10	1	3	0.0421	0.0439	0.0466	0.0415	0.0435	0.0445
10	7	1	3	0.0602	0.0638	0.0651	0.0410	0.0414	0.0418
10	10	1	3	0.0563	0.0579	0.0603	0.0494	0.0514	0.0533
15	15	1	3	0.0520	0.0529	0.0538	0.0468	0.0485	0.0500
15	30	1	3	0.0436	0.0445	0.0448	0.0474	0.0480	0.0485
30	15	1	3	0.0548	0.0561	0.0567	0.0436	0.0453	0.0469
30	30	1	3	0.0521	0.0531	0.0535	0.0499	0.0512	0.0516
15	45	1	3	0.0430	0.0431	0.0432	0.0484	0.0487	0.0490
45	15	1	3	0.0542	0.0550	0.0558	0.0403	0.0412	0.0420

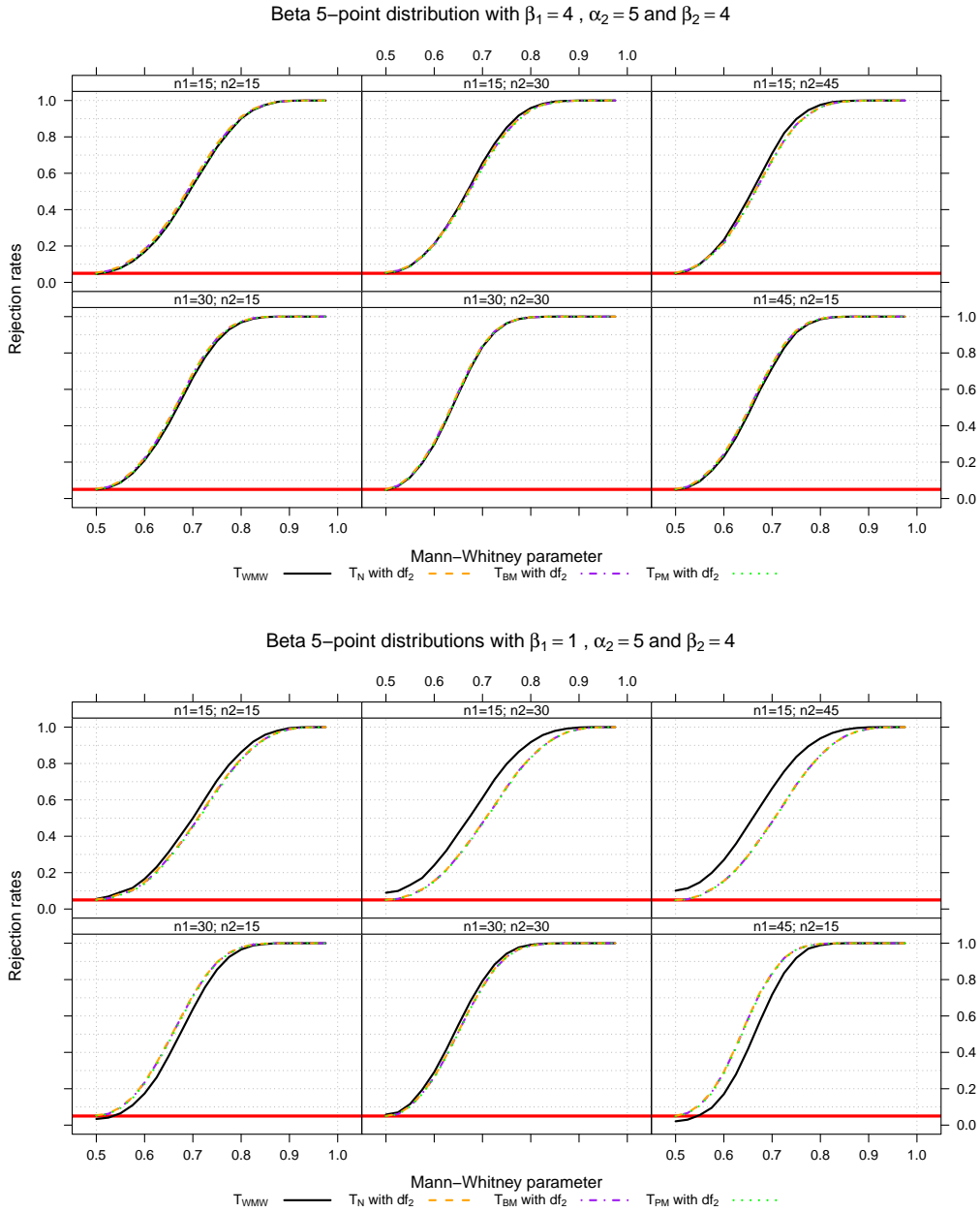


Figure 2: Power graphs for Beta 5-point distributions based on 10 000 simulation runs

Table 4: Type I error rates for the 5-point distributions with latent $F_1 = \mathcal{B}(\alpha_1, \beta_1)$ and $F_2 = \mathcal{B}(5, 4)$ at a two-sided nominal significance level of $\alpha = 0.05$ for the studentised permutation tests as given in (17) based on 10 000 random permutations for each of the 10 000 replications

n_1	n_2	α_1	β_1	\tilde{T}_N	\tilde{T}_{BM}	\tilde{T}_{PM}	\tilde{T}_N^{Logit}	\tilde{T}_{BM}^{Logit}	\tilde{T}_{PM}^{Logit}
7	7	5	4	0.0237	0.0235	0.0235	0.0225	0.0223	0.0224
7	10	5	4	0.0306	0.0307	0.0309	0.0316	0.0309	0.0313
10	7	5	4	0.0275	0.0276	0.0279	0.0271	0.0270	0.0273
10	10	5	4	0.0342	0.0337	0.0338	0.0336	0.0338	0.0341
15	15	5	4	0.0421	0.0421	0.0420	0.0418	0.0419	0.0419
15	30	5	4	0.0464	0.0464	0.0467	0.0469	0.0468	0.0463
30	15	5	4	0.0407	0.0407	0.0406	0.0412	0.0412	0.0410
30	30	5	4	0.0490	0.0489	0.0489	0.0487	0.0488	0.0490
15	45	5	4	0.0481	0.0479	0.0478	0.0483	0.0483	0.0483
45	15	5	4	0.0464	0.0464	0.0465	0.0466	0.0464	0.0462
7	7	1.2071	1	0.0365	0.0366	0.0385	0.0319	0.0321	0.0329
7	10	1.2071	1	0.0474	0.0475	0.0483	0.0413	0.0414	0.0422
10	7	1.2071	1	0.0388	0.0401	0.0405	0.0391	0.0398	0.0421
10	10	1.2071	1	0.0486	0.0487	0.0492	0.0450	0.0450	0.0468
15	15	1.2071	1	0.0531	0.0535	0.0543	0.0502	0.0507	0.0521
15	30	1.2071	1	0.0575	0.0576	0.0586	0.0522	0.0525	0.0541
30	15	1.2071	1	0.0501	0.0501	0.0504	0.0524	0.0528	0.0528
30	30	1.2071	1	0.0566	0.0570	0.0571	0.0554	0.0559	0.0562
15	45	1.2071	1	0.0534	0.0536	0.0544	0.0460	0.0462	0.0468
45	15	1.2071	1	0.0440	0.0442	0.0440	0.0475	0.0477	0.0476

4 Discussion

In this manuscript, we reviewed different variance estimators for the Mann-Whitney parameter and, more generally, different ways of how to approximate its sampling distribution in small samples. To stick to the unbiased variance estimator appears to be somewhat ill-advised. Indeed, in almost all scenarios Perme and Manevski's T_{PM} (13) with degrees of freedom $df_2(7)$ seems preferable in terms of controlling the type I error rate. Of course, Perme and Manevski's variance estimator is not unbiased and the particular choice of degrees of freedom lack a sound theoretical justification, even if they are consistent.

In heteroskedastic settings, the Wilcoxon-Mann-Whitney test T_{WMW} (10) performs poorly, particularly in case of unequal sample sizes, a pattern which also very slightly emerges when using the permutation approach (17) by Pauly et al. [2016].

As far as group sequential models for the Mann-Whitney parameter are concerned, it would be interesting to examine whether the test statistics, in particular T_{PM} (13) with $df_2(7)$, would equally well maintain one-sided nominal significance levels close to zero and up to 0.025. With that caveat in mind, we would further like to point out that T_{PM} (13) with $df_2(7)$ works very well for sample sizes $\min(n_1, n_2) \geq 15$ for a range of different distributions and tends to be somewhat conservative in smaller samples.

References

- D. Bamber. The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of Mathematical Psychology*, 12(4):387–415, 1975.
- G. E. P. Box. Some Theorems on Quadratic Forms Applied in the Study of Analysis of Variance Problems, II. Effects of Inequality of Variance and of Correlation Between Errors in the Two-Way Classification. *The Annals of Mathematical Statistics*, 25(3):484–498, 1954.
- E. Brunner and U. Munzel. The nonparametric Behrens-Fisher problem: Asymptotic theory and a small-sample approximation. *Biometrical Journal*, 42:17–25, 2000.
- E. Brunner, A. C. Bathke, and F. Konietzschke. *Rank and Pseudo-Rank Procedures for Independent Observations in Factorial Designs*. Springer, 2018.
- E. Brunner, M. Happ, and S. Friedrich. *Erwartungstreuer Schätzer für $Var(\hat{\theta})$ und Konfidenzintervalle für θ* . Unpublished manuscript, 2021a.
- E. Brunner, M. Vandemeulebroecke, and T. Mütze. Win odds: An adaptation of the win ratio to include ties. *Statistics in Medicine*, 40(14):3367–3384, 2021b.
- E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, 44:837–845, 1988.
- M. P. Fay, E. H. Brittain, J. H. Shih, D. A. Follmann, and E. E. Gabriel. Causal estimands and confidence intervals associated with Wilcoxon-Mann-Whitney tests in randomized experiments. *Statistics in Medicine*, 37:2923–2937, 2018.
- M. A. Fligner and G. E. Policello. Robust rank procedures for the Behrens-Fisher problem. *Journal of the American Statistical Association*, 76(373):162–168, 1981.
- P. Lévy. *Calcul des probabilités*, volume 9. Gauthier-Villars Paris, 1925.
- H. B. Mann and D. R. Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, 18:50–60, 1947.
- M. Pauly, T. Asendorf, and F. Konietzschke. Permutation-based inference for the AUC: A unified approach for continuous and discontinuous data. *Biometrical Journal*, 58(6):1319–1337, 2016.
- M. P. Perme and D. Manevski. Confidence intervals for the Mann-Whitney test. *Statistical Methods in Medical Research*, 28(12):3755–3768, 2019.
- F. H. Ruymgaart. *A unified approach to the asymptotic distribution theory of certain midrank statistics*. Springer, 1980.
- F. E. Satterthwaite. An approximate distribution of estimates of variance components. *Biometrics Bulletin*, 2:110–114, 1946.
- S. Shirahata. Estimate of variance of Wilcoxon-Mann-Whitney statistic. *Journal of the Japanese Society of Computational Statistics*, 6(2):1–10, 1993.
- H. F. Smith. The problem of comparing the results of two experiments with unequal errors. *Journal of the Council for Scientific and Industrial Research*, 9:211–212, 1936.
- O. Thas, J. De Neve, L. Clement, and J.-P. Ottoy. Probabilistic index models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74:623–671, 2012.
- B. L. Welch. The significance of the difference between two means when the population variances are unequal. *Biometrika*, 29:350–362, 1937.
- F. Wilcoxon. Individual comparisons by ranking methods. *Biometric Bulletin*, 1:80–83, 1945.
- F. Wilcoxon. Probability tables for individual comparisons by ranking methods. *Biometrics*, 3:119–122, 1947.

Appendix

This appendix consists of two main parts. First, we discuss the different variance estimators, translate them into our notation and briefly outline our approach to the Box-type degrees of freedom as regards the unbiased variance estimator. Second, we focus on the simulations, dealing with exception handling and providing more detailed results and results from settings not considered in the main manuscript.

Part I – Variance estimators and degrees of freedom

First we extend our notation of nonparametric theory, then we discuss the variance estimand and its different estimators. Lastly, we briefly explain how we arrived at the Box-type formulas of the degrees of freedom.

General notation

Let X denote a random variable representing ordered categorical or real data, defined on some probability space $(\Omega, \mathcal{A}, \mathbb{P})$. Then we define for each possible value x the following versions of distribution functions

$$\begin{aligned} F^+(x) &= \mathbb{P}(X \leq x), \\ F^-(x) &= \mathbb{P}(X < x), \\ F(x) &= \mathbb{P}(X < x) + 1/2 \cdot \mathbb{P}(X = x). \end{aligned}$$

Now suppose we have a sample of observations $X_1, \dots, X_n \stackrel{iid}{\sim} F(x)$. The empirical distribution functions then take the form

$$\begin{aligned} \hat{F}^+(x) &= \frac{1}{n} \sum_{i=1}^n c^+(x, X_i), & c^+(x, X_i) &= \begin{cases} 0 & \text{if } x < X_i \\ 1 & \text{if } x \geq X_i \end{cases}, \\ \hat{F}^-(x) &= \frac{1}{n} \sum_{i=1}^n c^-(x, X_i), & c^-(x, X_i) &= \begin{cases} 0 & \text{if } x \leq X_i \\ 1 & \text{if } x > X_i \end{cases}, \\ \hat{F}(x) &= \frac{1}{n} \sum_{i=1}^n c(x, X_i), & c(x, X_i) &= \begin{cases} 0 & \text{if } x < X_i \\ 1/2 & \text{if } x = X_i \\ 1 & \text{if } x > X_i \end{cases}. \end{aligned}$$

Now we define the nonparametric relative effect and give some of its properties, which we will use in the derivations later on.

Let $X_{ij} \sim F_i(x)$, $i = 1, 2$, $j = 1, \dots, n_i$ be independent random variables. Then the nonparametric relative effect is given by

$$p = \mathbb{P}(X_{1j} < X_{2j'}) + 1/2 \cdot \mathbb{P}(X_{1j} = X_{2j'}) = \int F_1 dF_2$$

for all $j = 1, \dots, n_1$ and $j' = 1, \dots, n_2$. In particular, if $F = F_1 = F_2$, then $\int F dF = 1/2$.

Let $X_{ij} \sim F_i(x)$, $i = 1, \dots, d$, $j = 1, \dots, n_i$ be independent real-valued random variables. Then for all $i, i' = 1, \dots, d$ as well as $j = 1, \dots, n_i$ and $j' = 1, \dots, n_{i'}$ it holds

$$\begin{aligned} \mathbb{E}(\hat{F}_i(x)) &= \mathbb{E}(c(x, X_{ij})) = F_i(x), \\ \mathbb{E}(\hat{F}_i(X_{i'j'})) &= \mathbb{E}(c(X_{i'j'}, X_{ij})) = \int F_i dF_{i'}. \end{aligned}$$

We will also be using the survival functions

$$\begin{aligned} S^+(x) &= \mathbb{P}(X \geq x), \\ S^-(x) &= \mathbb{P}(X > x), \\ S(x) &= \mathbb{P}(X > x) + 1/2 \cdot \mathbb{P}(X = x), \end{aligned}$$

with their empirical counterparts $\hat{S}^+(x)$, $\hat{S}^-(x)$, and $\hat{S}(x)$ defined accordingly. Note that

$$\begin{aligned} \mathbb{E}(\hat{S}_i(x)) &= \mathbb{E}(c(X_{ij}, x)) = S_i(x), \\ \mathbb{E}(\hat{S}_i(X_{i'j'})) &= \mathbb{E}(c(X_{ij}, X_{i'j'})) = \int S_i dF_{i'}. \end{aligned}$$

Variance estimand

In our notation, we have for the general case of arbitrary F_1 and F_2 ,

$$\sigma_N^2 = \mathbb{V}(\hat{p}) = \frac{\tau_0 + (n_2 - 1)\tau_1 + (n_1 - 1)\tau_2 - (n_1 + n_2 - 1)p^2}{n_1 n_2},$$

where

$$\begin{aligned}\tau_0 &= \int F_1 dF_2 - 1/4 \cdot \underbrace{\int (F_1^+ - F_1^-) dF_2}_{=\mathbb{P}(X_1=X_2)=:\beta}, \\ \tau_1 &= \int S_2^2 dF_1 = \int (1 - F_2)^2 dF_1, \\ \tau_2 &= \int F_1^2 dF_2.\end{aligned}$$

If both distributions coincide, i.e., $F_1 = F_2$, and are continuous, then it holds $\tau_0 = 1/2$, $\tau_1 = \tau_2 = 1/3$, yielding $\sigma_N^2 = \frac{n_1 + n_2 + 1}{12n_1 n_2}$.

In the unpublished preprint of Brunner et al. [2021a], we find formula (1.9), i.e.,

$$\sigma_N^2 = \frac{(n_2 - 1)\sigma_1^2 + (n_1 - 1)\sigma_2^2 + p(1 - p) - 1/4 \cdot \beta}{n_1 n_2},$$

where

$$\begin{aligned}\sigma_1^2 &= \mathbb{V}(F_2(X_{11})) = \int \{F_2 - (1 - p)\}^2 dF_1 = \int (S_2 - p)^2 dF_1 = \int S_2^2 dF_1 - 2p \int S_2 dF_1 + p^2 = \tau_1 - p^2, \\ \sigma_2^2 &= \mathbb{V}(F_1(X_{21})) = \int (F_1 - p)^2 dF_2 = \int F_1^2 dF_2 - 2p \int F_1 dF_2 + p^2 = \tau_2 - p^2.\end{aligned}$$

So it should be evident that both definitions of σ_N^2 are equivalent.

Bamber's [1975] definition of σ_N^2 , which he calls σ_a^2 , is equivalent as well. Assuming that X refers to sample 1 and Y to sample 2, it holds $B_{Y Y X} = 4\tau_1 - 4p + 1$ as well as $B_{X X Y} = 4\tau_2 - 4p + 1$. With $F_X = F_1$ and $F_Y = F_2$ and taking $\mathbb{P}(Y_1, Y_2 < X)$ to mean $\mathbb{P}(Y_1 < X, Y_2 < X)$, we can deduce

$$\begin{aligned}B_{Y Y X} &= \mathbb{P}(Y_1 < X, Y_2 < X) + \mathbb{P}(X < Y_1, X < Y_2) - 2\mathbb{P}(Y_1 < X < Y_2) \\ &= \int \mathbb{P}(Y_1 < x, Y_2 < x) dF_X(x) + \int \mathbb{P}(Y_1 > x, Y_2 > x) dF_X(x) - 2 \int \mathbb{P}(Y_1 < x, Y_2 > x) dF_X(x) \\ &= \int \mathbb{P}(Y_1 < x) \mathbb{P}(Y_2 < x) dF_X(x) + \int \mathbb{P}(Y_1 > x) \mathbb{P}(Y_2 > x) dF_X(x) - 2 \int \mathbb{P}(Y_1 < x) \mathbb{P}(Y_2 > x) dF_X(x) \\ &= \int \{F_Y^-(x)\}^2 dF_X(x) + \int \{S_Y^-(x)\}^2 dF_X(x) - 2 \int \{F_Y^-(x) S_Y^-(x)\} dF_X(x) \\ &= \int \{S_Y^-(x) - F_Y^-(x)\}^2 dF_X(x) \\ &= \int \{S_Y^-(x) + 1/2 \cdot \mathbb{P}(Y = x) - F_Y^-(x) - 1/2 \cdot \mathbb{P}(Y = x)\}^2 dF_X(x) \\ &= \int \{S_Y(x) - F_Y(x)\}^2 dF_X(x) \\ &= \int \{2S_Y(x) - 1\}^2 dF_X(x) \\ &= 4 \int S_Y^2 dF_X - 4 \int S_Y dF_X + 1 \\ &= 4 \int S_2^2 dF_1 - 4 \int S_2 dF_1 + 1 \\ &= 4\tau_1 - 4p + 1.\end{aligned}$$

By the same token, it holds

$$\begin{aligned}
 B_{XXY} &= \mathbb{P}(X_1 < Y, X_2 < Y) + \mathbb{P}(Y < X_1, Y < X_2) - 2\mathbb{P}(X_1 < Y < X_2) \\
 &= \int \mathbb{P}(X_1 < y, X_2 < y) dF_Y(y) + \int \mathbb{P}(X_1 > y, X_2 > y) dF_Y(y) - 2 \int \mathbb{P}(X_1 < y, X_2 > y) dF_Y(y) \\
 &= \int \mathbb{P}(X_1 < y) \mathbb{P}(X_2 < y) dF_Y(y) + \int \mathbb{P}(X_1 > y) \mathbb{P}(X_2 > y) dF_Y(y) - 2 \int \mathbb{P}(X_1 < y) \mathbb{P}(X_2 > y) dF_Y(y) \\
 &= \int \{F_X^-(y)\}^2 dF_Y(y) + \int \{S_X^-(y)\}^2 dF_Y(y) - 2 \int \{F_X^-(y) S_X^-(y)\} dF_Y(y) \\
 &= \int \{F_X^-(y) - S_X^-(y)\}^2 dF_Y(y) \\
 &= \int \{S_X^-(y) + 1/2 \cdot \mathbb{P}(X = y) - F_X^-(y) - 1/2 \cdot \mathbb{P}(X = y)\}^2 dF_Y(y) \\
 &= \int \{F_X(y) - S_X(y)\}^2 dF_Y(y) \\
 &= \int \{2F_X(y) - 1\}^2 dF_Y(y) \\
 &= 4 \int F_X^2 dF_Y - 4 \int F_X dF_Y + 1 \\
 &= 4 \int F_1^2 dF_2 - 4 \int F_1 dF_2 + 1 \\
 &= 4\tau_2 - 4p + 1.
 \end{aligned}$$

Note that in our notation $N_X = n_1$ and $N_Y = n_2$ as well as $A = p$, so that

$$\begin{aligned}
 \sigma_a^2 &= \frac{1}{4N_X N_Y} \{ \mathbb{P}(X \neq Y) + (N_X - 1)B_{XXY} + (N_Y - 1)B_{YYX} - 4(N_X + N_Y - 1)(A - 1/2)^2 \} \\
 &= \frac{1}{4n_1 n_2} \{ 1 - \beta + (n_1 - 1)(4\tau_2 - 4p + 1) + (n_2 - 1)(4\tau_1 - 4p + 1) - 4(n_1 + n_2 - 1)(p - 1/2)^2 \} \\
 &= \frac{(n_1 - 1)\tau_2 + (n_2 - 1)\tau_1 - (n_1 + n_2 - 1)p^2}{n_1 n_2} \\
 &\quad + \frac{1 - \beta - (n_1 + n_2 - 2)(4p - 1) + 4(n_1 + n_2 - 1)(p - 1/4)}{4n_1 n_2} \\
 &= \frac{(n_1 - 1)\tau_2 + (n_2 - 1)\tau_1 - (n_1 + n_2 - 1)p^2}{n_1 n_2} + \frac{4p - \beta}{4n_1 n_2} \\
 &= \frac{\tau_0 + (n_1 - 1)\tau_2 + (n_2 - 1)\tau_1 - (n_1 + n_2 - 1)p^2}{n_1 n_2} = \sigma_N^2.
 \end{aligned}$$

As for Perme and Manevski [2019], they define

$$\mathbb{V}(\hat{\theta}) = \frac{\theta(1-\theta)}{mn} + \frac{n-1}{nm} \mathbb{V}(S_Y(X)) + \frac{m-1}{mn} \mathbb{V}(S_X(Y)),$$

which in our notation should read as

$$\mathbb{V}(\hat{p}) = \frac{p(1-p)}{n_1 n_2} + \frac{n_2 - 1}{n_1 n_2} \mathbb{V}(S_2^-(X_{11})) + \frac{n_2 - 1}{n_1 n_2} \mathbb{V}(S_1^-(X_{21})),$$

However, this formula assumes that F_X and F_Y are both continuous. Perme and Manevski say as much in the supplementary material, i.e., ‘‘For better clarity of all the derivations, we shall assume that both F_X and F_Y are continuous (the extension of formulae to the case of ties is then straightforward)’’. In the main paper they state ‘‘This work focuses on continuous random variables X and Y . In practice, the data may often be documented on a discrete scale and thus ties can occur. Therefore, we shall always extend the definition to the case of ties.’’ Nonetheless, Perme and Manevski seemingly do not explicitly set out what they consider to be the variance estimand in case of ties. In any event, in case of continuity we have

$$\begin{aligned}
 \mathbb{V}(S_2^-(X_{11})) &= \mathbb{V}(S_2(X_{11})) = \mathbb{V}(1 - F_2(X_{11})) = \mathbb{V}(F_2(X_{11})) = \sigma_1^2, \\
 \mathbb{V}(S_1^-(X_{21})) &= \mathbb{V}(S_1(X_{21})) = \mathbb{V}(1 - F_1(X_{21})) = \mathbb{V}(F_1(X_{21})) = \sigma_2^2, \\
 \beta &= 0.
 \end{aligned}$$

Thus Perme and Manevski's formula is equal to our definition of σ_N^2 so long as both distributions F_1 and F_2 are continuous.

Variance estimation

Our plug-in estimators for $\tau_0, \tau_1, \tau_2, p^2$ are given by

$$\begin{aligned}\hat{\tau}_0 &= \hat{p} - 1/4 \cdot \hat{\beta}, \text{ with } \hat{\beta} = \frac{1}{n_1} \frac{1}{n_2} \sum_{j=1}^{n_2} \sum_{i=1}^{n_1} \mathbb{I}(X_{2j} = X_{1i}), \quad \mathbb{I}(X_{2j} = X_{1i}) = \begin{cases} 1 & \text{if } X_{2j} = X_{1i} \\ 0 & \text{if } X_{2j} \neq X_{1i} \end{cases}, \\ \hat{\tau}_1 &= \int \hat{S}_2^2 d\hat{F}_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} \{\hat{S}_2(X_{1i})\}^2 = \frac{1}{n_1} \sum_{i=1}^{n_1} \left\{ \frac{1}{n_2} \sum_{j=1}^{n_2} c(X_{2j}, X_{1i}) \right\}^2, \\ \hat{\tau}_2 &= \int \hat{F}_1^2 d\hat{F}_2 = \frac{1}{n_2} \sum_{j=1}^{n_2} \{\hat{F}_1(X_{2j})\}^2 = \frac{1}{n_2} \sum_{j=1}^{n_2} \left\{ \frac{1}{n_1} \sum_{i=1}^{n_1} c(X_{2j}, X_{1i}) \right\}^2, \\ \hat{p}^2 &= \left(\int \hat{F}_1 d\hat{F}_2 \right)^2 = \left(\frac{1}{n_1} \frac{1}{n_2} \sum_{j=1}^{n_2} \sum_{i=1}^{n_1} c(X_{2j}, X_{1i}) \right)^2.\end{aligned}$$

It can readily be seen that $\mathbb{E}(\hat{\tau}_0) = \tau_0$ since \hat{p} and $\hat{\beta}$ are unbiased. As regards $\hat{\tau}_1$ and $\hat{\tau}_2$, we find

$$\begin{aligned}\mathbb{E}(\hat{\tau}_1) &= \frac{1}{n_1} \sum_{i=1}^{n_1} \left\{ \frac{1}{n_2} \frac{1}{n_2} \sum_{j=1}^{n_2} \sum_{j'=1}^{n_2} \mathbb{E}(c(X_{2j}, X_{1i})c(X_{2j'}, X_{1i})) \right\} \\ &= \frac{1}{n_1} \frac{1}{n_2} \frac{1}{n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \sum_{\substack{j'=1 \\ j' \neq j}}^{n_2} \mathbb{E}(c(X_{2j}, X_{1i})c(X_{2j'}, X_{1i})) + \frac{1}{n_1} \frac{1}{n_2} \frac{1}{n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \mathbb{E}(c(X_{2j}, X_{1i})c(X_{2j}, X_{1i})) \\ &= \frac{n_2 - 1}{n_2} \int S_2^2 dF_1 + \frac{1}{n_2} \mathbb{E}(c(X_{21}, X_{11})^2) = \frac{n_2 - 1}{n_2} \tau_1 + \frac{1}{n_2} \tau_0.\end{aligned}$$

In a similar vein, it follows

$$\begin{aligned}\mathbb{E}(\hat{\tau}_2) &= \frac{1}{n_2} \sum_{j=1}^{n_2} \left\{ \frac{1}{n_1} \frac{1}{n_1} \sum_{i=1}^{n_1} \sum_{i'=1}^{n_1} \mathbb{E}(c(X_{2j}, X_{1i})c(X_{2j}, X_{1i'})) \right\} \\ &= \frac{1}{n_2} \frac{1}{n_1} \frac{1}{n_1} \sum_{j=1}^{n_2} \sum_{i=1}^{n_1} \sum_{\substack{i'=1 \\ i' \neq i}}^{n_1} \mathbb{E}(c(X_{2j}, X_{1i})c(X_{2j}, X_{1i'})) + \frac{1}{n_2} \frac{1}{n_1} \frac{1}{n_1} \sum_{j=1}^{n_2} \sum_{i=1}^{n_1} \mathbb{E}(c(X_{2j}, X_{1i})c(X_{2j}, X_{1i})) \\ &= \frac{n_1 - 1}{n_1} \int F_1^2 dF_2 + \frac{1}{n_1} \mathbb{E}(c(X_{21}, X_{11})^2) = \frac{n_1 - 1}{n_1} \tau_2 + \frac{1}{n_1} \tau_0.\end{aligned}$$

As for \hat{p}^2 , we now look to

$$\begin{aligned}\mathbb{E}(\hat{p}^2) &= \mathbb{E}\left(\frac{1}{n_1} \frac{1}{n_2} \sum_{j=1}^{n_2} \sum_{i=1}^{n_1} c(X_{2j}, X_{1i}) \frac{1}{n_1} \frac{1}{n_2} \sum_{j'=1}^{n_2} \sum_{i'=1}^{n_1} c(X_{2j'}, X_{1i'}) \right) \\ &= \frac{1}{n_1} \frac{1}{n_2} \frac{1}{n_1} \frac{1}{n_2} \sum_{j=1}^{n_2} \sum_{i=1}^{n_1} \sum_{j'=1}^{n_2} \sum_{i'=1}^{n_1} \underbrace{\mathbb{E}(c(X_{2j}, X_{1i}))}_{=: \zeta_{ij}} \underbrace{\mathbb{E}(c(X_{2j'}, X_{1i'}))}_{=: \zeta_{i'j'}}\end{aligned}$$

Looking at the quadruple sum produces

$$\begin{aligned}
 & \sum_{j'=1}^{n_2} \sum_{i'=1}^{n_1} \sum_{j=1}^{n_2} \sum_{i=1}^{n_1} \mathbb{E}(\zeta_{ij} \zeta_{i'j'}) \\
 &= \sum_{j'=1}^{n_2} \sum_{i'=1}^{n_1} \sum_{j=1}^{n_2} \sum_{\substack{i=1 \\ j \neq j', i \neq i'}}^{n_1} \mathbb{E}(\zeta_{ij} \zeta_{i'j'}) + \sum_{j'=1}^{n_2} \sum_{i'=1}^{n_1} \sum_{\substack{j=1 \\ j \neq j'}}^{n_2} \mathbb{E}(\zeta_{i'j} \zeta_{i'j'}) + \sum_{j'=1}^{n_2} \sum_{i'=1}^{n_1} \sum_{\substack{i=1 \\ i \neq i'}}^{n_1} \mathbb{E}(\zeta_{ij'} \zeta_{i'j'}) + \sum_{j'=1}^{n_2} \sum_{i'=1}^{n_1} \mathbb{E}(\zeta_{i'j'} \zeta_{i'j'}), \\
 &= n_1 n_2 (n_2 - 1) (n_1 - 1) \left\{ \int F_1 dF_2 \right\}^2 + n_1 n_2 (n_2 - 1) \int S_2^2 dF_1 + n_1 n_2 (n_1 - 1) \int F_1^2 dF_2 + n_1 n_2 \mathbb{E}(\zeta_{11}^2).
 \end{aligned}$$

Therefore, we have

$$\mathbb{E}(\hat{p}^2) = \frac{(n_2 - 1)(n_1 - 1)}{n_1 n_2} p^2 + \frac{n_2 - 1}{n_1 n_2} \tau_1 + \frac{n_1 - 1}{n_1 n_2} \tau_2 + \frac{1}{n_1 n_2} \tau_0.$$

An unbiased estimator of σ_N^2 should then take the form

$$\hat{\sigma}_N^2 = \frac{n_2 \hat{\tau}_1 + n_1 \hat{\tau}_2 - \hat{\tau}_0 - (n_1 + n_2 - 1) \hat{p}^2}{(n_1 - 1)(n_2 - 1)}.$$

To check the unbiasedness of $\hat{\sigma}_N^2$, consider

$$\begin{aligned}
 & (n_1 - 1)(n_2 - 1) \mathbb{E}(\hat{\sigma}_N^2) \\
 &= n_2 \mathbb{E}(\hat{\tau}_1) + n_1 \mathbb{E}(\hat{\tau}_2) - \mathbb{E}(\hat{\tau}_0) - (n_1 + n_2 - 1) \mathbb{E}(\hat{p}^2) \\
 &= (n_2 - 1) \tau_1 + \tau_0 + (n_1 - 1) \tau_2 + \tau_0 - \tau_0 \\
 &\quad - \frac{n_1 + n_2 - 1}{n_1 n_2} \{ (n_1 - 1)(n_2 - 1) p^2 + (n_2 - 1) \tau_1 + (n_1 - 1) \tau_2 + \tau_0 \} \\
 &= \left(1 - \frac{n_1 + n_2 - 1}{n_1 n_2} \right) \{ \tau_0 + (n_2 - 1) \tau_1 + (n_1 - 1) \tau_2 \} - \frac{(n_1 - 1)(n_2 - 1)}{n_1 n_2} (n_1 + n_2 - 1) p^2.
 \end{aligned}$$

Now since $\left(1 - \frac{n_1 + n_2 - 1}{n_1 n_2} \right) = \frac{(n_1 - 1)(n_2 - 1)}{n_1 n_2}$, it follows that $\mathbb{E}(\hat{\sigma}_N^2) = \sigma_N^2$.

Brunner form of the unbiased variance estimator

Now we want to have a closer look at the estimator in (2.39) derived as in the unpublished preprint of Brunner et al. [2021a],

$$\hat{\sigma}_N^2 = \frac{1}{n_1(n_1 - 1)n_2(n_2 - 1)} \left(\sum_{i=1}^2 \sum_{k=1}^{n_i} \left(R_{ik} - R_{ik}^{(i)} - \left[\bar{R}_{i\bullet} - \frac{n_i + 1}{2} \right] \right)^2 - n_1 n_2 \left[\hat{\theta}(1 - \hat{\theta}) - \frac{1}{4} \hat{\beta} \right] \right),$$

where $\hat{\theta} = \int \hat{F}_2 d\hat{F}_1 = 1 - \hat{p}$ and $\hat{\beta} = \frac{1}{n_1 n_2} \sum_{j=2}^{n_2} \sum_{i=1}^{n_1} \mathbb{I}(X_{2j} = X_{1i})$. First recall the following identities as regards the rank representations,

$$\begin{aligned}
 n_2 \hat{F}_2(X_{1i}) &= R_{1i} - R_{1i}^{(1)}, & 1 - \hat{p} &= \int \hat{F}_2 d\hat{F}_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} \hat{F}_2(X_{1i}) = \frac{1}{n_2} (\bar{R}_{1\bullet} - \frac{n_1 + 1}{2}), \\
 n_1 \hat{F}_1(X_{2j}) &= R_{2j} - R_{2j}^{(2)}, & \hat{p} &= \int \hat{F}_1 d\hat{F}_2 = \frac{1}{n_2} \sum_{j=1}^{n_2} \hat{F}_1(X_{2j}) = \frac{1}{n_1} (\bar{R}_{2\bullet} - \frac{n_2 + 1}{2}),
 \end{aligned}$$

where $R_{1i}^{(1)}$ and $R_{2j}^{(2)}$ denote the so-called internal ranks with respect to the first and second sample. That is to say $R_{1i}^{(1)}$ is the mid-rank of X_{1i} among X_{11}, \dots, X_{1n_1} whereas $R_{2j}^{(2)}$ is the mid-rank of X_{2j} among X_{21}, \dots, X_{2n_2} .

Together with the equality

$$\begin{aligned}
 \int \widehat{F}_2^2 d\widehat{F}_1 - \left(\int \widehat{F}_2 d\widehat{F}_1 \right)^2 &= \int (1 - \widehat{S}_2)^2 d\widehat{F}_1 - (1 - \int \widehat{S}_2 d\widehat{F}_1)^2 \\
 &= 1 + \int \widehat{S}_2^2 d\widehat{F}_1 - 2 \int \widehat{S}_2 d\widehat{F}_1 - \{1 + (\int \widehat{S}_2 d\widehat{F}_1)^2 - 2 \int \widehat{S}_2 d\widehat{F}_1\} \\
 &= \int \widehat{S}_2^2 d\widehat{F}_1 - (\int \widehat{S}_2 d\widehat{F}_1)^2 \\
 &= \int \widehat{S}_2^2 d\widehat{F}_1 - \widehat{p}^2,
 \end{aligned}$$

this produces

$$\begin{aligned}
 &\sum_{i=1}^{n_1} \left(R_{1i} - R_{1i}^{(1)} - \left[\bar{R}_{1\bullet} - \frac{n_1 + 1}{2} \right] \right)^2 \\
 &= \sum_{i=1}^{n_1} \left(n_2 \widehat{F}_2(X_{1i}) - n_2 \int \widehat{F}_2 d\widehat{F}_1 \right)^2 = n_2^2 \sum_{i=1}^{n_1} \left(\widehat{F}_2(X_{1i}) - \int \widehat{F}_2 d\widehat{F}_1 \right)^2 \\
 &= n_2^2 n_1 \frac{1}{n_1} \sum_{i=1}^{n_1} \left(\widehat{F}_2^2(X_{1i}) - 2\widehat{F}_2(X_{1i}) \int \widehat{F}_2 d\widehat{F}_1 + (\int \widehat{F}_2 d\widehat{F}_1)^2 \right) \\
 &= n_2^2 n_1 \left\{ \int \widehat{F}_2^2 d\widehat{F}_1 - (\int \widehat{F}_2 d\widehat{F}_1)^2 \right\} = n_2^2 n_1 \left\{ \int \widehat{S}_2^2 d\widehat{F}_1 - \widehat{p}^2 \right\} = n_2^2 n_1 (\widehat{\tau}_1 - \widehat{p}^2), \\
 &\sum_{j=1}^{n_2} \left(R_{2j} - R_{2j}^{(2)} - \left[\bar{R}_{2\bullet} - \frac{n_2 + 1}{2} \right] \right)^2 \\
 &= \sum_{j=1}^{n_2} \left(n_1 \widehat{F}_1(X_{2j}) - n_1 \int \widehat{F}_1 d\widehat{F}_2 \right)^2 = n_1^2 \sum_{j=1}^{n_2} \left(\widehat{F}_1(X_{2j}) - \int \widehat{F}_1 d\widehat{F}_2 \right)^2 \\
 &= n_1^2 n_2 \frac{1}{n_2} \sum_{j=1}^{n_2} \left(\widehat{F}_1^2(X_{2j}) - 2\widehat{F}_1(X_{2j}) \int \widehat{F}_1 d\widehat{F}_2 + (\int \widehat{F}_1 d\widehat{F}_2)^2 \right) \\
 &= n_1^2 n_2 \left\{ \int \widehat{F}_1^2 d\widehat{F}_2 - (\int \widehat{F}_1 d\widehat{F}_2)^2 \right\} = n_1^2 n_2 \left\{ \int \widehat{F}_1^2 d\widehat{F}_2 - \widehat{p}^2 \right\} = n_1^2 n_2 (\widehat{\tau}_2 - \widehat{p}^2), \\
 &\widehat{\theta}(1 - \widehat{\theta}) - \frac{1}{4} \widehat{\beta} \\
 &= \widehat{p}(1 - \widehat{p}) - \frac{1}{4} \widehat{\beta} = \widehat{p} - \frac{1}{4} \widehat{\beta} - \widehat{p}^2 = \widehat{\tau}_0 - \widehat{p}^2.
 \end{aligned}$$

Now putting everything together we find

$$\begin{aligned}
 \widehat{\sigma}_N^2 &= \frac{n_2^2 n_1 (\widehat{\tau}_1 - \widehat{p}^2) + n_1^2 n_2 (\widehat{\tau}_2 - \widehat{p}^2) - n_1 n_2 (\widehat{\tau}_0 - \widehat{p}^2)}{n_1 (n_1 - 1) n_2 (n_2 - 1)} \\
 &= \frac{n_2 (\widehat{\tau}_1 - \widehat{p}^2) + n_1 (\widehat{\tau}_2 - \widehat{p}^2) - \widehat{\tau}_0 + \widehat{p}^2}{(n_1 - 1)(n_2 - 1)} \\
 &= \frac{n_2 \widehat{\tau}_1 + n_1 \widehat{\tau}_2 - \widehat{\tau}_0 - (n_1 + n_2 - 1) \widehat{p}^2}{(n_1 - 1)(n_2 - 1)}.
 \end{aligned}$$

Bamber form of the unbiased variance estimator

As to Bamber's [1975] notation, he labels observations from sample 1 as X_1, \dots, X_{N_X} and from sample 2 as Y_1, \dots, Y_{N_Y} . He then goes on to define an estimator for $B_{Y Y X}$, that is

$$b_{Y Y X} = p(Y_1, Y_2 < X) + p(X < Y_1, Y_2) - 2p(Y_1 < X < Y_2),$$

where

$$\begin{aligned}
 p(Y_1, Y_2 < X) &= \frac{1}{N_X N_Y (N_Y - 1)} \sum_{i=1}^{N_X} \sum_{j=1}^{N_Y} \sum_{\substack{j'=1 \\ j' \neq j}}^{N_Y} \mathbb{I}(Y_j < X_i) \mathbb{I}(Y_{j'} < X_i) \\
 &= \frac{1}{N_X N_Y (N_Y - 1)} \left\{ \sum_{i=1}^{N_X} \left(\sum_{j=1}^{N_Y} \mathbb{I}(Y_j < X_i) \right)^2 - \sum_{i=1}^{N_X} \sum_{j=1}^{N_Y} \mathbb{I}(Y_j < X_i) \right\} \\
 &= \frac{N_Y}{N_X (N_Y - 1)} \sum_{i=1}^{N_X} \{\widehat{F}_Y^-(X_i)\}^2 - \frac{\sum_{i=1}^{N_X} \sum_{j=1}^{N_Y} \mathbb{I}(Y_j < X_i)}{N_X N_Y (N_Y - 1)}, \\
 p(X < Y_1, Y_2) &= \frac{1}{N_X N_Y (N_Y - 1)} \sum_{i=1}^{N_X} \sum_{j=1}^{N_Y} \sum_{\substack{j'=1 \\ j' \neq j}}^{N_Y} \mathbb{I}(Y_j > X_i) \mathbb{I}(Y_{j'} > X_i) \\
 &= \frac{1}{N_X N_Y (N_Y - 1)} \left\{ \sum_{i=1}^{N_X} \left(\sum_{j=1}^{N_Y} \mathbb{I}(Y_j > X_i) \right)^2 - \sum_{i=1}^{N_X} \sum_{j=1}^{N_Y} \mathbb{I}(Y_j > X_i) \right\} \\
 &= \frac{N_Y}{N_X (N_Y - 1)} \sum_{i=1}^{N_X} \{\widehat{S}_Y^-(X_i)\}^2 - \frac{\sum_{i=1}^{N_X} \sum_{j=1}^{N_Y} \mathbb{I}(Y_j > X_i)}{N_X N_Y (N_Y - 1)}, \\
 p(Y_1 < X < Y_2) &= \frac{1}{N_X N_Y (N_Y - 1)} \sum_{i=1}^{N_X} \sum_{j=1}^{N_Y} \sum_{\substack{j'=1 \\ j' \neq j}}^{N_Y} \mathbb{I}(Y_j < X_i) \mathbb{I}(Y_{j'} > X_i) \\
 &= \frac{1}{N_X N_Y (N_Y - 1)} \left\{ \sum_{i=1}^{N_X} \sum_{j=1}^{N_Y} \sum_{j'=1}^{N_Y} \mathbb{I}(Y_j < X_i) \mathbb{I}(Y_{j'} > X_i) - \sum_{i=1}^{N_X} \sum_{j=1}^{N_Y} \mathbb{I}(Y_j < X_i) \mathbb{I}(Y_j > X_i) \right\} \\
 &= \frac{N_Y}{N_X (N_Y - 1)} \sum_{i=1}^{N_X} \{\widehat{F}_Y^-(X_i) \widehat{S}_Y^-(X_i)\}.
 \end{aligned}$$

Further note that he calls $u_X = \sum_{j=1}^{N_Y} \mathbb{I}(Y_j < X)$ and $v_X = \sum_{j=1}^{N_Y} \mathbb{I}(Y_j > X)$.

Similar to before, it holds for each $i \in \{1, \dots, N_X\}$ that

$$\begin{aligned}
 &\{\widehat{S}_Y^-(X_i)\}^2 + \{\widehat{F}_Y^-(X_i)\}^2 - 2\widehat{F}_Y^-(X_i)\widehat{S}_Y^-(X_i) \\
 &= \{\widehat{S}_Y^-(X_i) - \widehat{F}_Y^-(X_i)\}^2 = \{\widehat{S}_Y^-(X_i) + \frac{1}{2} \sum_{j=1}^{N_X} \mathbb{I}(Y_j = X_i) - \widehat{F}_Y^- - \frac{1}{2} \sum_{j=1}^{N_X} \mathbb{I}(Y_j = X_i)\}^2 \\
 &= \{\widehat{S}_Y^-(X_i) - \widehat{F}_Y^-(X_i)\}^2 = \{2\widehat{S}_Y^-(X_i) - 1\}^2 \\
 &= 4\widehat{S}_Y^2(X_i) - 4\widehat{S}_Y^-(X_i) + 1.
 \end{aligned}$$

Therefore it follows,

$$\begin{aligned}
 b_{Y Y X} &= \frac{N_Y}{N_Y - 1} \left\{ 4 \int \widehat{S}_Y^2 d\widehat{F}_X - 4 \int \widehat{S}_Y d\widehat{F}_X + 1 \right\} - \frac{1 - \widehat{\beta}}{N_Y - 1} \\
 &= \frac{n_2}{n_2 - 1} \left\{ 4 \int \widehat{S}_2^2 d\widehat{F}_1 - 4 \int \widehat{S}_2 d\widehat{F}_1 + 1 \right\} - \frac{1 - \widehat{\beta}}{n_2 - 1} \\
 &= \frac{n_2}{n_2 - 1} \{4\widehat{\tau}_1 - 4\widehat{p} + 1\} - \frac{1 - \widehat{\beta}}{n_2 - 1}.
 \end{aligned}$$

By the same arguments, it should then hold

$$b_{X X Y} = \frac{n_1}{n_1 - 1} \{4\widehat{\tau}_2 - 4\widehat{p} + 1\} - \frac{1 - \widehat{\beta}}{n_1 - 1}.$$

Now putting everything together we again find

$$\begin{aligned}
 s_a^2 &= \frac{p(X \neq Y) + (N_X - 1)b_{XXY} + (N_Y - 1)b_{YYX} - 4(N_X + N_Y - 1)(a - 1/2)^2}{4(N_X - 1)(N_Y - 1)} \\
 &= \frac{1 - \hat{\beta} + n_1(4\hat{\tau}_2 - 4\hat{p} + 1) - 1 + \hat{\beta} + n_2(4\hat{\tau}_1 - 4\hat{p} + 1) - 1 + \hat{\beta} - 4(n_1 + n_2 - 1)(\hat{p} - 1/2)^2}{4(n_1 - 1)(n_2 - 1)} \\
 &= \frac{n_1\hat{\tau}_2 + n_2\hat{\tau}_1 - (n_1 + n_2 - 1)\hat{p}^2}{(n_1 - 1)(n_2 - 1)} \\
 &\quad + \frac{\hat{\beta} - 4\hat{p}(n_1 + n_2) + 4\hat{p}(n_1 + n_2 - 1) + n_1 + n_2 - 1 - (n_1 + n_2 - 1)}{4(n_1 - 1)(n_2 - 1)} \\
 &= \frac{n_1\hat{\tau}_2 + n_2\hat{\tau}_1 - (n_1 + n_2 - 1)\hat{p}^2}{(n_1 - 1)(n_2 - 1)} + \frac{\hat{\beta} - 4\hat{p}}{4(n_1 - 1)(n_2 - 1)} \\
 &= \frac{n_1\hat{\tau}_2 + n_2\hat{\tau}_1 - (n_1 + n_2 - 1)\hat{p}^2 - \hat{\tau}_0}{(n_1 - 1)(n_2 - 1)} = \hat{\sigma}_N^2.
 \end{aligned}$$

So we see that Bamber's definition of the estimator $\hat{\sigma}_N^2$ is equivalent as well.

Perme and Manevski's estimators

As to Perme and Manevski's [2019] notation, they have X_1, \dots, X_m for sample one and Y_1, \dots, Y_n for sample 2. They give two estimators

$$\begin{aligned}
 \hat{V}_{DL}(\hat{\theta}) &= \frac{1}{m} \hat{V}(S_Y(X)) + \frac{1}{n} \hat{V}(S_X(Y)), \\
 \hat{V}_{DLe}(\hat{\theta}) &= \frac{\hat{\theta}(1 - \hat{\theta})}{mn} + \frac{n - 1}{mn} \hat{V}(S_Y(X)) + \frac{m - 1}{mn} \hat{V}(S_X(Y)),
 \end{aligned}$$

where

$$\begin{aligned}
 \hat{V}(S_Y(X)) &= \frac{1}{m - 1} \sum_{i=1}^m (V_{i\bullet} - \hat{\theta})^2, \\
 \hat{V}(S_X(Y)) &= \frac{1}{n - 1} \sum_{j=1}^n (V_{\bullet j} - \hat{\theta})^2,
 \end{aligned}$$

with $V_{i\bullet} = \frac{1}{n} \sum_{j=1}^n V_{ij}$, $V_{\bullet j} = \frac{1}{m} \sum_{i=1}^m V_{ij}$ where $V_{ij} = c(Y_j, X_i)$ according to our notation using the normalised version of the count function.

So these quantities should equal

$$\begin{aligned}
 V_{i\bullet} &= \frac{1}{n} \sum_{j=1}^n V_{ij} = \frac{1}{n} \sum_{j=1}^n c(Y_j, X_i) = \hat{S}_Y(X_i) = \hat{S}_2(X_{1i}), \\
 V_{\bullet j} &= \frac{1}{m} \sum_{i=1}^m V_{ij} = \frac{1}{m} \sum_{i=1}^m c(Y_j, X_i) = \hat{F}_X(Y_j) = \hat{F}_1(X_{2j}).
 \end{aligned}$$

Therefore, it follows

$$\begin{aligned}
 \widehat{\mathbb{V}}(S_Y(X)) &= \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (\widehat{S}_2(X_{1i}) - \widehat{p})^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (\widehat{S}_2^2(X_{1i}) - 2\widehat{p}\widehat{S}_2(X_{1i}) + \widehat{p}^2) \\
 &= \frac{n_1}{n_1 - 1} \left(\int \widehat{S}_2^2 d\widehat{F}_1 - 2\widehat{p} \int \widehat{S}_2 d\widehat{F}_1 + \widehat{p}^2 \right) \\
 &= \frac{n_1}{n_1 - 1} (\widehat{\tau}_1 - \widehat{p}^2) = \widehat{\sigma}_1^2, \\
 \widehat{\mathbb{V}}(S_X(Y)) &= \frac{1}{n_2 - 1} \sum_{j=1}^{n_2} (\widehat{F}_1(X_{2j}) - \widehat{p})^2 = \frac{1}{n_2 - 1} \sum_{j=1}^{n_2} (\widehat{F}_1^2(X_{2j}) - 2\widehat{p}\widehat{F}_1(X_{2j}) + \widehat{p}^2) \\
 &= \frac{n_2}{n_2 - 1} \left(\int \widehat{F}_1^2 d\widehat{F}_2 - 2\widehat{p} \int \widehat{F}_1 d\widehat{F}_2 + \widehat{p}^2 \right) \\
 &= \frac{n_2}{n_2 - 1} (\widehat{\tau}_2 - \widehat{p}^2) = \widehat{\sigma}_2^2.
 \end{aligned}$$

Equipped with these results it follows that

$$\widehat{\mathbb{V}}_{DL}(\widehat{\theta}) = \widehat{\mathbb{V}}_{DL}(\widehat{p}) = \frac{\widehat{\sigma}_1^2}{n_1} + \frac{\widehat{\sigma}_2^2}{n_2} = \frac{\widehat{\tau}_1 - \widehat{p}^2}{n_1 - 1} + \frac{\widehat{\tau}_2 - \widehat{p}^2}{n_2 - 1} = \widehat{\sigma}_{BM}^2,$$

in other words, the variance estimator proposed by DeLong et al. (1988) is identical to Brunner and Munzel's (2000). As for the estimator Perme and Manevski call *exact*, we have

$$\begin{aligned}
 \widehat{\mathbb{V}}_{DLe}(\widehat{\theta}) &= \widehat{\mathbb{V}}_{DLe}(\widehat{p}) = \frac{\widehat{p}(1 - \widehat{p}) + (n_2 - 1)\widehat{\sigma}_1^2 + (n_1 - 1)\widehat{\sigma}_2^2}{n_1 n_2} \\
 &= \frac{\widehat{p}(1 - \widehat{p})}{n_1 n_2} + \frac{(n_2 - 1)(\widehat{\tau}_1 - \widehat{p}^2)}{n_2(n_1 - 1)} + \frac{(n_1 - 1)(\widehat{\tau}_2 - \widehat{p}^2)}{n_1(n_2 - 1)} = \widehat{\sigma}_{PM}^2.
 \end{aligned}$$

Shirahata's formulas

Shirahata [1993] only considers continuous distributions, assuming two independent samples $X_1, \dots, X_m \sim F(x)$ and $Y_1, \dots, Y_n \sim G(x)$. To ease the translation of his formulas into our notation, we exchange the samples, i.e. we set $(X_1, \dots, X_m) = (X_{21}, \dots, X_{2n_2})$ and $(Y_1, \dots, Y_n) = (X_{11}, \dots, X_{1n_1})$. Hence $F(x) = F_2(x)$ and $G(x) = F_1(x)$.

Moreover Shirahata uses the count function $u(x) = 1$ or 0 according as $x \geq 0$ or $x < 0$. Hence we have the following equivalence

$$u(X_i - Y_j) = \begin{cases} 1 & \text{if } X_i - Y_j \geq 0 \\ 0 & \text{if } X_i - Y_j < 0 \end{cases} = \begin{cases} 1 & \text{if } X_i \geq Y_j \\ 0 & \text{if } X_i < Y_j \end{cases} = c^+(X_i, Y_j) = c^+(X_{2i}, X_{1j}).$$

Hence the quantities in Section 2 in Shirahata should read in our notation as

$$\begin{aligned}
 \zeta_{11} &= \theta = \mathbb{E}(u(X_1 - Y_1)) = \mathbb{E}(c^+(X_{21}, X_{11})) = \int F_1^+ dF_2, \\
 \zeta_{21} &= \mathbb{E}(u(X_1 - Y_1)u(X_2 - Y_1)) = \mathbb{E}(c^+(X_{21}, X_{11})c^+(X_{22}, X_{11})) \\
 &= \int \mathbb{E}(c^+(X_{21}, x)c^+(X_{22}, x))dF_1(x) \\
 &= \int (S_2^+)^2 dF_1, \\
 \zeta_{12} &= \mathbb{E}(u(X_1 - Y_1)u(X_1 - Y_2)) = \mathbb{E}(c^+(X_{21}, X_{11})c^+(X_{21}, X_{12})) \\
 &= \int \mathbb{E}(c^+(x, X_{11})c^+(x, X_{12}))dF_2(x) \\
 &= \int (F_1^+)^2 dF_2, \\
 \zeta_{22} &= \theta^2 = \left(\int F_1^+ dF_2 \right)^2.
 \end{aligned}$$

So in case of continuous distributions we have $\zeta_{11} = p = \tau_0$, $\zeta_{21} = \int S_2^2 dF_1 = \tau_1$, $\zeta_{22} = \int F_1^2 dF_2 = \tau_2$, and $\zeta_{22} = p^2$. Hence in case of no ties, it follows that their formula of the theoretical variance coincides with ours,

$$\begin{aligned}\sigma^2 &= \frac{1}{mn} \{ \zeta_{11} + (m-1)\zeta_{21} + (n-1)\zeta_{12} - (m+n-1)\zeta_{22} \} \\ &= \frac{1}{n_1 n_2} \{ \tau_0 + (n_2-1)\tau_1 + (n_1-1)\tau_2 - (n_1+n_2-1)p^2 \} = \sigma_N^2.\end{aligned}$$

As for the estimator, Shirahata considers the quantities B , C^2 and D^2 , which we will now translate into our notation, i.e.,

$$\begin{aligned}B &= \sum_{i=1}^m \sum_{j=1}^n u(X_i - Y_j) = \sum_{i=1}^{n_2} \sum_{j=1}^{n_1} c^+(X_{2i}, X_{1j}) = \sum_{i=1}^{n_2} n_1 \hat{F}_1^+(X_{2i}) = n_1 n_2 \int \hat{F}_1^+ d\hat{F}_2, \\ C^2 &= \sum_{j=1}^n \left(\sum_{i=1}^m u(X_i - Y_j) \right)^2 = \sum_{j=1}^{n_1} \left(\sum_{i=1}^{n_2} c^+(X_{2i}, X_{1j}) \right)^2 = \sum_{j=1}^{n_1} n_2^2 \{ \hat{S}_2^+(X_{1j}) \}^2 \\ &= n_1 n_2^2 \int (\hat{S}_2^+)^2 d\hat{F}_1, \\ D^2 &= \sum_{i=1}^m \left(\sum_{j=1}^n u(X_i - Y_j) \right)^2 = \sum_{i=1}^{n_2} \left(\sum_{j=1}^{n_1} c^+(X_{2i}, X_{1j}) \right)^2 = \sum_{i=1}^{n_2} n_1^2 \{ \hat{F}_1^+(X_{2i}) \}^2 \\ &= n_1^2 n_2 \int (\hat{F}_1^+)^2 d\hat{F}_2.\end{aligned}$$

Shirahata considers a range of variance estimators, the first one being the unbiased one,

$$\begin{aligned}\hat{\sigma}_U^2 &= \frac{1}{m(m-1)n(n-1)} \left(-\frac{m+n-1}{mn} B^2 - B + C^2 + D^2 \right) \\ &= \frac{1}{n_1(n_1-1)n_2(n_2-1)} \left(-\frac{n_1+n_2-1}{n_1 n_2} n_1^2 n_2^2 \left(\int \hat{F}_1^+ d\hat{F}_2 \right)^2 - n_1 n_2 \int \hat{F}_1^+ d\hat{F}_2 \right. \\ &\quad \left. + n_1 n_2^2 \int (\hat{S}_2^+)^2 d\hat{F}_1 + n_1^2 n_2 \int (\hat{F}_1^+)^2 d\hat{F}_2 \right) \\ &= \frac{n_2 \int (\hat{S}_2^+)^2 d\hat{F}_1 + n_1 \int (\hat{F}_1^+)^2 d\hat{F}_2 - \int \hat{F}_1^+ d\hat{F}_2 - (n_1+n_2-1) \left(\int \hat{F}_1^+ d\hat{F}_2 \right)^2}{(n_1-1)(n_2-1)}.\end{aligned}$$

Their bootstrap estimator is given by

$$\begin{aligned}\hat{\sigma}_B^2 &= \frac{1}{m^2 n^2} \left(-\frac{m+n-1}{mn} B^2 + B + \frac{m-1}{m} C^2 + \frac{n-1}{n} D^2 \right) \\ &= \frac{1}{n_1^2 n_2^2} \left(-\frac{n_1+n_2-1}{n_1 n_2} n_1^2 n_2^2 \left(\int \hat{F}_1^+ d\hat{F}_2 \right)^2 + n_1 n_2 \int \hat{F}_1^+ d\hat{F}_2 \right. \\ &\quad \left. + \frac{n_2-1}{n_2} n_1 n_2^2 \int (\hat{S}_2^+)^2 d\hat{F}_1 + \frac{n_1-1}{n_1} n_1^2 n_2 \int (\hat{F}_1^+)^2 d\hat{F}_2 \right) \\ &= \frac{(n_2-1) \int (\hat{S}_2^+)^2 d\hat{F}_1 + (n_1-1) \int (\hat{F}_1^+)^2 d\hat{F}_2 + \int \hat{F}_1^+ d\hat{F}_2 - (n_1+n_2-1) \left(\int \hat{F}_1^+ d\hat{F}_2 \right)^2}{n_1 n_2}.\end{aligned}$$

The simple version of the Fligner and Policello [1981] estimator then takes the form

$$\begin{aligned}\hat{\sigma}_{FP}^2 &= \frac{1}{m^2 n^2} \left(-\frac{m+n+1}{mn} B^2 - B + C^2 + D^2 \right) \\ &= \frac{n_2 \int (\hat{S}_2^+)^2 d\hat{F}_1 + n_1 \int (\hat{F}_1^+)^2 d\hat{F}_2 - \int \hat{F}_1^+ d\hat{F}_2 - (n_1+n_2+1) \left(\int \hat{F}_1^+ d\hat{F}_2 \right)^2}{n_1 n_2} \\ &= \frac{\int (\hat{S}_2^+)^2 d\hat{F}_1}{n_1} + \frac{\int (\hat{F}_1^+)^2 d\hat{F}_2}{n_2} - \frac{\int \hat{F}_1^+ d\hat{F}_2 + (n_1+n_2+1) \left(\int \hat{F}_1^+ d\hat{F}_2 \right)^2}{n_1 n_2}.\end{aligned}$$

Lastly they consider the jackknife estimator

$$\begin{aligned}\hat{\sigma}_J^2 &= \frac{1}{m(m-1)n(n-1)} \left(-\frac{m+n-2}{mn} B^2 - \frac{m-1}{m} C^2 + \frac{n-1}{n} D^2 \right) \\ &= \frac{1}{n_1(n_1-1)n_2(n_2-1)} \left(-\frac{n_1+n_2-1}{n_1n_2} n_1^2 n_2^2 \int \hat{F}_1^+ d\hat{F}_2 \right) \\ &\quad + \frac{n_2-1}{n_2} n_1 n_2^2 \int (\hat{S}_2^+)^2 d\hat{F}_1 + \frac{n_1-1}{n_1} n_1^2 n_2 \int (\hat{F}_1^+)^2 d\hat{F}_2 \\ &= \frac{\int (\hat{S}_2^+)^2 d\hat{F}_1}{n_1-1} + \frac{\int (\hat{F}_1^+)^2 d\hat{F}_2}{n_2-1} - \frac{(n_1+n_2-2) \int \hat{F}_1^+ d\hat{F}_2}{(n_1-1)(n_2-1)}.\end{aligned}$$

Box-type approximation of degrees of freedom

To begin with, we summarise the pertinent results of Box-type [1954] degrees of freedom as developed in Chapter 7.5.1.2 of Brunner et al. [2018].

First, we consider independent normal random variables $X_{11}, \dots, X_{1n_1} \sim \mathcal{N}(\mu_1, \sigma_1^2)$ as well as $X_{21}, \dots, X_{2n_2} \sim \mathcal{N}(\mu_2, \sigma_2^2)$, $N = n_1 + n_2$, $\bar{\mathbf{X}} = (\sum_{i=1}^{n_1} X_{1i}/n_1 \quad \sum_{j=1}^{n_2} X_{2j}/n_2)^\top$. Then we have

$$\begin{aligned}\mathbf{S}_N &= \text{Cov}(\sqrt{N}\bar{\mathbf{X}}) = \bigoplus_{i=1}^2 N\sigma_i^2/n_i = N \begin{pmatrix} \sigma_1^2/n_1 & 0 \\ 0 & \sigma_2^2/n_2 \end{pmatrix}, \\ \hat{\mathbf{S}}_N &= N \begin{pmatrix} \hat{\sigma}_1^2/n_1 & 0 \\ 0 & \hat{\sigma}_2^2/n_2 \end{pmatrix},\end{aligned}$$

where $\hat{\sigma}_g^2 = \frac{1}{n_g-1} \sum_{i=1}^{n_g} (X_{gi} - \bar{X}_{g\bullet})^2$, $g = 1, 2$. Furthermore we will need

$$\mathbf{\Lambda} = \left(\begin{pmatrix} n_1 & 0 \\ 0 & n_2 \end{pmatrix} - \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \right)^{-1} = \begin{pmatrix} 1/(n_1-1) & 0 \\ 0 & 1/(n_2-1) \end{pmatrix}.$$

As for the contrast matrix (centering matrix), we have

$$\mathbf{C} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} - \frac{1}{2} \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} = \frac{1}{2} \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix},$$

resulting in

$$\mathbf{T} = \mathbf{C}^\top (\mathbf{C}\mathbf{C}^\top)^{-1} \mathbf{C} = \mathbf{C}, \quad \mathbf{D}_T = \frac{1}{2} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},$$

since \mathbf{C} is positive semi-definite, symmetric and idempotent.

Then the approximate degrees of freedom for the two sample t test are given by

$$f_0 = [\text{tr}(\mathbf{D}_T \mathbf{S}_N)]^2 / \text{tr}(\mathbf{D}_T^2 \mathbf{S}_N^2 \mathbf{\Lambda}), \quad \hat{f}_0 = \left[\text{tr}(\mathbf{D}_T \hat{\mathbf{S}}_N) \right]^2 / \text{tr}(\mathbf{D}_T^2 \hat{\mathbf{S}}_N^2 \mathbf{\Lambda}),$$

which then simplifies to

$$\begin{aligned}\mathbf{D}_T \hat{\mathbf{S}}_N &= \frac{N}{2} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \hat{\sigma}_1^2/n_1 & 0 \\ 0 & \hat{\sigma}_2^2/n_2 \end{pmatrix} = \frac{N}{2} \begin{pmatrix} \hat{\sigma}_1^2/n_1 & 0 \\ 0 & \hat{\sigma}_2^2/n_2 \end{pmatrix}, \\ \left[\text{tr}(\mathbf{D}_T \hat{\mathbf{S}}_N) \right]^2 &= \frac{N^2}{4} (\hat{\sigma}_1^2/n_1 + \hat{\sigma}_2^2/n_2)^2, \\ \mathbf{D}_T^2 \hat{\mathbf{S}}_N^2 \mathbf{\Lambda} &= \frac{N^2}{4} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \hat{\sigma}_1^4/n_1^2 & 0 \\ 0 & \hat{\sigma}_2^4/n_2^2 \end{pmatrix} \begin{pmatrix} 1/(n_1-1) & 0 \\ 0 & 1/(n_2-1) \end{pmatrix} \\ \text{tr}(\mathbf{D}_T^2 \hat{\mathbf{S}}_N^2 \mathbf{\Lambda}) &= \frac{N^2}{4} \left(\frac{\hat{\sigma}_1^4}{n_1^2(n_1-1)} + \frac{\hat{\sigma}_2^4}{n_2^2(n_2-1)} \right),\end{aligned}$$

yielding $\hat{f}_0 = (\hat{\sigma}_1^2/n_1 + \hat{\sigma}_2^2/n_2)^2 / \left(\frac{\hat{\sigma}_1^4}{n_1^2(n_1-1)} + \frac{\hat{\sigma}_2^4}{n_2^2(n_2-1)} \right)$.

For the degrees of freedom proposed by Brunner and Munzel, the derivation is completely analogous. To this end, consider $X_{11}, \dots, X_{1n_1} \sim F_1$ as well as $X_{21}, \dots, X_{2n_2} \sim F_2$, $N = n_1 + n_2$, $\bar{\mathbf{X}} = (\sum_{i=1}^{n_1} F_2(X_{1i})/n_1 \quad \sum_{j=1}^{n_2} F_1(X_{2j})/n_2)^\top$. Then we have

$$\mathbf{S}_N = \text{Cov}(\sqrt{N}\bar{\mathbf{X}}) = \bigoplus_{i=1}^2 N\sigma_i^2/n_i = N \begin{pmatrix} \sigma_1^2/n_1 & 0 \\ 0 & \sigma_2^2/n_2 \end{pmatrix},$$

$$\widehat{\mathbf{S}}_N = N \begin{pmatrix} \widehat{\sigma}_1^2/n_1 & 0 \\ 0 & \widehat{\sigma}_2^2/n_2 \end{pmatrix},$$

with variances $\sigma_1^2 = \mathbb{V}(F_2(X_{11})) = \tau_1 - p^2$, $\sigma_2^2 = \mathbb{V}(F_1(X_{21})) = \tau_2 - p^2$, and estimators $\widehat{\sigma}_1^2 = \frac{n_1}{n_1-1}(\widehat{\tau}_1 - \widehat{p}^2)$ and $\widehat{\sigma}_2^2 = \frac{n_2}{n_2-1}(\widehat{\tau}_2 - \widehat{p}^2)$, yielding $\widehat{f}_0 = (\widehat{\sigma}_1^2/n_1 + \widehat{\sigma}_2^2/n_2)^2 / \left(\frac{\widehat{\sigma}_1^4}{n_1^2(n_1-1)} + \frac{\widehat{\sigma}_2^4}{n_2^2(n_2-1)} \right)$.

The idea behind the newly adjusted degrees of freedom is to derive them similarly as for the Brunner-Munzel [2000] test, but using the empirical distribution functions to compute the mean vector instead of the theoretical ones, i.e., $\bar{\mathbf{X}} = (\sum_{i=1}^{n_1} \widehat{F}_2(X_{1i})/n_1 \quad \sum_{j=1}^{n_2} \widehat{F}_1(X_{2j})/n_2)^\top$.

To obtain the entries of $\mathbf{S}_N = \text{Cov}(\sqrt{N}\bar{\mathbf{X}})$, we will first consider the following quantities

$$\begin{aligned} \psi_1^2 &= \mathbb{V}(\widehat{F}_2(X_{11})) = \mathbb{V}(1 - \widehat{S}_2(X_{11})) = \mathbb{V}(\widehat{S}_2(X_{11})) = \mathbb{E}((\widehat{S}_2(X_{11}))^2) - (\mathbb{E}(\widehat{S}_2(X_{11})))^2, \\ \psi_{1|1} &= \text{Cov}(\widehat{F}_2(X_{11}), \widehat{F}_2(X_{12})) = \text{Cov}(\widehat{S}_2(X_{11}), \widehat{S}_2(X_{12})), \\ \psi_2^2 &= \mathbb{V}(\widehat{F}_1(X_{21})) = \mathbb{E}((\widehat{F}_1(X_{21}))^2) - (\mathbb{E}(\widehat{F}_1(X_{21})))^2, \\ \psi_{2|2} &= \text{Cov}(\widehat{F}_1(X_{21}), \widehat{F}_1(X_{21})), \\ \psi_{12} &= \text{Cov}(\widehat{F}_2(X_{11}), \widehat{F}_1(X_{21})) = -\text{Cov}(\widehat{S}_2(X_{11}), \widehat{F}_1(X_{21})) \\ &= -\mathbb{E}(\widehat{S}_2(X_{11}), \widehat{F}_1(X_{21})) + \mathbb{E}(\widehat{S}_2(X_{11}))\mathbb{E}(\widehat{F}_1(X_{21})). \end{aligned}$$

So we now consider

$$\begin{aligned} \mathbb{E}(\widehat{S}_2(X_{11})) &= \frac{1}{n_2} \sum_{\ell=1}^{n_2} \mathbb{E}(c(X_{2\ell}, X_{11})) = \int F_1 dF_2 = p, \\ \mathbb{E}(\widehat{F}_1(X_{21})) &= \frac{1}{n_1} \sum_{k=1}^{n_1} \mathbb{E}(c(X_{21}, X_{1k})) = \int F_1 dF_2 = p. \end{aligned}$$

Moreover,

$$\begin{aligned} \mathbb{E}((\widehat{S}_2(X_{11}))^2) &= \frac{1}{n_2^2} \sum_{\ell=1}^{n_2} \sum_{\ell'=1}^{n_2} \mathbb{E}(c(X_{2\ell}, X_{11})c(X_{2\ell'}, X_{11})) \\ &= \frac{1}{n_2^2} \sum_{\ell=1}^{n_2} \sum_{\substack{\ell'=1 \\ \ell' \neq \ell}}^{n_2} \mathbb{E}(c(X_{2\ell}, X_{11})c(X_{2\ell'}, X_{11})) + \frac{1}{n_2^2} \sum_{\ell=1}^{n_2} \mathbb{E}(c(X_{2\ell}, X_{11})c(X_{2\ell}, X_{11})) \\ &= \frac{n_2-1}{n_2} \int S_2^2 dF_1 + \frac{1}{n_2} \mathbb{E}(c(X_{21}, X_{11})^2) = \frac{n_2-1}{n_2} \tau_1 + \frac{1}{n_2} \tau_0, \\ \mathbb{E}(\widehat{S}_2(X_{11})\widehat{S}_2(X_{12})) &= \frac{1}{n_2^2} \sum_{\ell=1}^{n_2} \sum_{\ell'=1}^{n_2} \mathbb{E}(c(X_{2\ell}, X_{11})c(X_{2\ell'}, X_{12})) \\ &= \frac{1}{n_2^2} \sum_{\ell=1}^{n_2} \sum_{\substack{\ell'=1 \\ \ell' \neq \ell}}^{n_2} \mathbb{E}(c(X_{2\ell}, X_{11})c(X_{2\ell'}, X_{12})) + \frac{1}{n_2^2} \sum_{\ell=1}^{n_2} \mathbb{E}(c(X_{2\ell}, X_{11})c(X_{2\ell}, X_{12})) \\ &= \frac{n_2-1}{n_2} p^2 - \frac{1}{n_2} \int F_1^2 dF_2 = \frac{n_2-1}{n_2} p^2 + \frac{1}{n_2} \tau_2, \end{aligned}$$

$$\begin{aligned}
 \mathbb{E}((\widehat{F}_1(X_{21}))^2) &= \frac{1}{n_1^2} \sum_{k=1}^{n_1} \sum_{k'=1}^{n_1} \mathbb{E}(c(X_{21}, X_{1k})c(X_{21}, X_{1k'})) \\
 &= \frac{1}{n_1^2} \sum_{k=1}^{n_1} \sum_{\substack{k'=1 \\ k' \neq k}}^{n_1} \mathbb{E}(c(X_{21}, X_{1k})c(X_{21}, X_{1k'})) + \frac{1}{n_1^2} \sum_{k=1}^{n_1} \mathbb{E}(c(X_{21}, X_{1k})c(X_{21}, X_{1k})) \\
 &= \frac{n_1 - 1}{n_1} \int F_1^2 dF_2 + \frac{1}{n_2} \mathbb{E}(c(X_{21}, X_{11})^2) = \frac{n_1 - 1}{n_1} \tau_2 + \frac{1}{n_1} \tau_0, \\
 \mathbb{E}(\widehat{F}_1(X_{21})\widehat{F}_1(X_{22})) &= \frac{1}{n_1^2} \sum_{k=1}^{n_1} \sum_{k'=1}^{n_1} \mathbb{E}(c(X_{21}, X_{1k})c(X_{22}, X_{1k'})) \\
 &= \frac{1}{n_1^2} \sum_{k=1}^{n_1} \sum_{\substack{k'=1 \\ k' \neq k}}^{n_1} \mathbb{E}(c(X_{21}, X_{1k})c(X_{22}, X_{1k'})) + \frac{1}{n_1^2} \sum_{k=1}^{n_1} \mathbb{E}(c(X_{21}, X_{1k})c(X_{22}, X_{1k})) \\
 &= \frac{n_1 - 1}{n_1} p^2 dF_2 + \frac{1}{n_2} \int S_2^2 dF_1 = \frac{n_1 - 1}{n_1} p_2 + \frac{1}{n_1} \tau_1,
 \end{aligned}$$

as well as

$$\begin{aligned}
 \mathbb{E}(\widehat{S}_2(X_{11}), \widehat{F}_1(X_{21})) &= \frac{1}{n_1 n_2} \sum_{k=1}^{n_1} \sum_{\ell=1}^{n_2} \mathbb{E}(c(X_{2\ell}, X_{11})c(X_{21}, X_{1k})) \\
 &= \frac{1}{n_1 n_2} \sum_{k=2}^{n_1} \sum_{\ell=2}^{n_2} \mathbb{E}(c(X_{2\ell}, X_{11})c(X_{21}, X_{1k})) + \frac{1}{n_1 n_2} \mathbb{E}(c(X_{21}, X_{11})c(X_{21}, X_{11})) \\
 &\quad + \frac{1}{n_1 n_2} \sum_{k=2}^{n_1} \mathbb{E}(c(X_{21}, X_{11})c(X_{21}, X_{1k})) + \frac{1}{n_1 n_2} \sum_{\ell=2}^{n_2} \mathbb{E}(c(X_{2\ell}, X_{11})c(X_{21}, X_{11})) \\
 &= \frac{(n_1 - 1)(n_2 - 1)p^2 + \tau_0 + (n_1 - 1)\tau_2 + (n_2 - 1)\tau_1}{n_1 n_2}.
 \end{aligned}$$

Collecting terms, we have

$$\begin{aligned}
 \psi_1^2 &= \frac{n_2 - 1}{n_2} \tau_1 + \frac{1}{n_2} p - \frac{1}{4n_2} \beta - p^2 = \frac{1}{n_2} \left[(n_2 - 1)(\tau_1 - p^2) + p - p^2 - \frac{1}{4} \beta \right] \\
 &= \frac{1}{n_2} \left[(n_2 - 1)\sigma_1^2 + p(1 - p) - \frac{1}{4} \beta \right], \\
 \psi_2^2 &= \frac{n_1 - 1}{n_1} \tau_2 + \frac{1}{n_1} p - \frac{1}{4n_1} \beta - p^2 = \frac{1}{n_1} \left[(n_1 - 1)(\tau_2 - p^2) + p - p^2 - \frac{1}{4} \beta \right] \\
 &= \frac{1}{n_1} \left[(n_1 - 1)\sigma_2^2 + p(1 - p) - \frac{1}{4} \beta \right], \\
 \psi_{1|1} &= \frac{n_2 - 1}{n_2} p^2 + \frac{1}{n_2} \tau_2 - p^2 = \frac{1}{n_2} (\tau_2 - p^2) = \frac{1}{n_2} \sigma_2^2, \\
 \psi_{2|2} &= \frac{n_1 - 1}{n_1} p^2 + \frac{1}{n_1} \tau_1 - p^2 = \frac{1}{n_1} (\tau_1 - p^2) = \frac{1}{n_1} \sigma_1^2,
 \end{aligned}$$

which are the same terms as in equations (2.32), (2.33), (2.34), and (2.35) of Brunner et al. [2021a] multiplied by either n_2^2 or n_1^2 . Now we turn to

$$\begin{aligned}
 \psi_{12} &= p^2 - \frac{(n_1 - 1)(n_2 - 1)p^2 + \tau_0 + (n_1 - 1)\tau_2 + (n_2 - 1)\tau_1}{n_1 n_2} \\
 &= -\frac{(n_1 - 1)\tau_2 + (n_2 - 1)\tau_1 + \tau_0 - (n_1 + n_2 - 1)p^2}{n_1 n_2} = -\sigma_N^2 = -\mathbb{V}(\widehat{p}),
 \end{aligned}$$

Now, since

$$\begin{aligned}
 \mathbb{V}\left(\frac{1}{n_1} \sum_{i=1}^{n_1} \widehat{F}_2(X_{1i})\right) &= \frac{1}{n_1^2} \sum_{i=1}^{n_1} \mathbb{V}(\widehat{F}_2(X_{1i})) + \frac{1}{n_1^2} \sum_{k=1}^{n_1} \sum_{\substack{k'=1 \\ k' \neq k}}^{n_1} \text{Cov}(\widehat{F}_2(X_{1k}), \widehat{F}_2(X_{1k'})) \\
 &= \frac{1}{n_1^2} n_1 \psi_1^2 + \frac{1}{n_1^2} n_1(n_1 - 1) \psi_{1|1} \\
 &= \frac{1}{n_1 n_2} \left[(n_2 - 1) \sigma_1^2 + p(1 - p) - \frac{1}{4} \beta \right] + \frac{n_1 - 1}{n_1 n_2} \sigma_2^2 = \sigma_N^2, \\
 \mathbb{V}\left(\frac{1}{n_2} \sum_{j=1}^{n_2} \widehat{F}_1(X_{2j})\right) &= \frac{1}{n_2^2} \sum_{j=1}^{n_2} \mathbb{V}(\widehat{F}_1(X_{2j})) + \frac{1}{n_2^2} \sum_{\ell=1}^{n_2} \sum_{\substack{\ell'=1 \\ \ell' \neq \ell}}^{n_2} \text{Cov}(\widehat{F}_1(X_{2\ell}), \widehat{F}_1(X_{2\ell'})) \\
 &= \frac{1}{n_2^2} n_2 \psi_2^2 + \frac{1}{n_2^2} n_2(n_2 - 1) \psi_{2|2} \\
 &= \frac{1}{n_1 n_2} \left[(n_1 - 1) \sigma_2^2 + p(1 - p) - \frac{1}{4} \beta \right] + \frac{n_2 - 1}{n_1 n_2} \sigma_1^2 = \sigma_N^2,
 \end{aligned}$$

and finally

$$\text{Cov}\left(\frac{1}{n_1} \sum_{i=1}^{n_1} \widehat{F}_2(X_{1i}), \frac{1}{n_2} \sum_{j=1}^{n_2} \widehat{F}_1(X_{2j})\right) = \frac{1}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \text{Cov}(\widehat{F}_2(X_{1i}), \widehat{F}_1(X_{2j})) = \frac{n_1 n_2}{n_1 n_2} \psi_{12} = -\sigma_N^2.$$

Therefore, we have

$$\mathbf{S}_N = \text{Cov}(\sqrt{N} \bar{\mathbf{X}}) = N \sigma_N^2 \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}, \quad \widehat{\mathbf{S}}_N = N \widehat{\sigma}_N^2 \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}.$$

If we are to derive the degrees of freedom in an analogous manner as before, we should get

$$\begin{aligned}
 \mathbf{D}_T \widehat{\mathbf{S}}_N &= \frac{N \widehat{\sigma}_N^2}{2} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} = \frac{N \widehat{\sigma}_N^2}{2} \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} \\
 \left[\text{tr}(\mathbf{D}_T \widehat{\mathbf{S}}_N) \right]^2 &= N^2 \widehat{\sigma}_N^4, \\
 \mathbf{D}_T^2 \widehat{\mathbf{S}}_N^2 \boldsymbol{\Lambda} &= \frac{N^2 \widehat{\sigma}_N^4}{4} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 2 & -2 \\ -2 & 2 \end{pmatrix} \begin{pmatrix} 1/(n_1 - 1) & 0 \\ 0 & 1/(n_2 - 1) \end{pmatrix} \\
 &= \frac{N^2 \widehat{\sigma}_N^4}{2} \begin{pmatrix} 1/(n_1 - 1) & -1/(n_2 - 1) \\ -1/(n_1 - 1) & 1/(n_2 - 1) \end{pmatrix} \\
 \text{tr}(\mathbf{D}_T^2 \widehat{\mathbf{S}}_N^2 \boldsymbol{\Lambda}) &= \frac{N^2 \widehat{\sigma}_N^4}{2} \left(\frac{1}{n_1 - 1} + \frac{1}{n_2 - 1} \right),
 \end{aligned}$$

yielding $\widehat{f}_0 = 2 / \left(\frac{1}{n_1 - 1} + \frac{1}{n_2 - 1} \right)$.

Furthermore, we now consider a simple heuristic alternative which may be viewed as a middle ground of the Brunner-Munzel approach and the approach just developed. To this end, recall the Brunner form of the unbiased variance estimator, i.e.,

$$\widehat{\sigma}_N^2 = \frac{1}{n_1(n_1 - 1)n_2(n_2 - 1)} \left(\sum_{i=1}^2 \sum_{k=1}^{n_i} \left(R_{ik} - R_{ik}^{(i)} - \left[\bar{R}_{i\bullet} - \frac{n_i + 1}{2} \right] \right)^2 - n_1 n_2 \left[\widehat{\theta}(1 - \widehat{\theta}) - \frac{1}{4} \widehat{\beta} \right] \right).$$

We now split this unbiased estimator into two ‘‘symmetric’’ parts,

$$\begin{aligned}
 \widehat{\sigma}_{1|N}^2 &= \frac{1}{n_1(n_1 - 1)n_2(n_2 - 1)} \left(\sum_{k=1}^{n_1} \left(R_{1k} - R_{1k}^{(1)} - \left[\bar{R}_{1\bullet} - \frac{n_1 + 1}{2} \right] \right)^2 - \frac{1}{2} n_1 n_2 \left[\widehat{\theta}(1 - \widehat{\theta}) - \frac{1}{4} \widehat{\beta} \right] \right), \\
 \widehat{\sigma}_{2|N}^2 &= \frac{1}{n_1(n_1 - 1)n_2(n_2 - 1)} \left(\sum_{\ell=1}^{n_2} \left(R_{2\ell} - R_{2\ell}^{(2)} - \left[\bar{R}_{2\bullet} - \frac{n_2 + 1}{2} \right] \right)^2 - \frac{1}{2} n_1 n_2 \left[\widehat{\theta}(1 - \widehat{\theta}) - \frac{1}{4} \widehat{\beta} \right] \right),
 \end{aligned}$$

such that $\hat{\sigma}_{1|N}^2 + \hat{\sigma}_{2|N}^2 = \hat{\sigma}_N^2$. Moreover, we artificially set the covariances zero so that now we have

$$\widehat{\mathbf{S}}_N = N \begin{pmatrix} \hat{\sigma}_{1|N}^2 & 0 \\ 0 & \hat{\sigma}_{2|N}^2 \end{pmatrix}.$$

With this artificial covariance matrix, we now compute the degrees of freedom similar to before,

$$\begin{aligned} \mathbf{D}_T \widehat{\mathbf{S}}_N &= \frac{N}{2} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \hat{\sigma}_{1|N}^2 & 0 \\ 0 & \hat{\sigma}_{2|N}^2 \end{pmatrix} = \frac{N}{2} \begin{pmatrix} \hat{\sigma}_{1|N}^2 & 0 \\ 0 & \hat{\sigma}_{2|N}^2 \end{pmatrix}, \\ \left[\text{tr} \left(\mathbf{D}_T \widehat{\mathbf{S}}_N \right) \right]^2 &= \frac{N^2}{4} \left(\hat{\sigma}_{1|N}^2 + \hat{\sigma}_{2|N}^2 \right)^2, \\ \mathbf{D}_T^2 \widehat{\mathbf{S}}_N \boldsymbol{\Lambda} &= \frac{N^2}{4} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \hat{\sigma}_{1|N}^4 & 0 \\ 0 & \hat{\sigma}_{2|N}^4 \end{pmatrix} \begin{pmatrix} 1/(n_1 - 1) & 0 \\ 0 & 1/(n_2 - 1) \end{pmatrix} \\ \text{tr} \left(\mathbf{D}_T^2 \widehat{\mathbf{S}}_N \boldsymbol{\Lambda} \right) &= \frac{N^2}{4} \left(\hat{\sigma}_{1|N}^4 / (n_1 - 1) + \hat{\sigma}_{2|N}^4 / (n_2 - 1) \right), \end{aligned}$$

yielding $\hat{f}_0 = \left(\hat{\sigma}_{1|N}^2 + \hat{\sigma}_{2|N}^2 \right)^2 / \left(\hat{\sigma}_{1|N}^4 / (n_1 - 1) + \hat{\sigma}_{2|N}^4 / (n_2 - 1) \right) = \hat{\sigma}_N^4 / \left(\hat{\sigma}_{1|N}^4 / (n_1 - 1) + \hat{\sigma}_{2|N}^4 / (n_2 - 1) \right)$.

Part II – Simulations

Now, we will briefly discuss how we dealt with cases where the variance estimates turned out to be zero or negative. Thereafter, we will present simulation results in more detail and for more settings than in the main manuscript.

Exception handling

The fact that $\hat{\sigma}_{WMW}^2 \leq 0$ can only occur when all outcomes for patients on both treatment arms coincide, that is to say,

$$x_{11} = \dots = x_{1n_1} = x_{21} = \dots = x_{2n_2},$$

yielding $\int \widehat{F}^2 d\widehat{F} = 1/4$ and consequently $\hat{\sigma}_{WMW}^2 = 0$. With the Mann-Whitney parameter p remaining unchanged, we then pretended that the last observation was different,

$$x_{11} = \dots = x_{1n_1k} = x_{21} = \dots \neq x_{2n_2k},$$

yielding $\hat{\sigma}_{WMW}^2 = 1/4n_1n_2$ and thus $T_{WMW}^2 = 0$ leading to nonrejection of the null hypothesis. We likewise set T_N and T_{BM} to zero as well in such cases, although we replaced the variances in a different manner, i.e., we always used $\max(1/n_1^2n_2^2, \hat{\sigma}_N^2)$ instead of $\hat{\sigma}_N^2$ and $\max(1/n_1^2n_2^2, \hat{\sigma}_{BM}^2)$ instead of $\hat{\sigma}_{BM}^2$. Note that if all values in both samples coincide, we would have $\hat{p}^2 = \hat{\tau}_0 = \hat{\tau}_1 = \hat{\tau}_2 = 1/4$, yielding $\hat{\sigma}_N^2 = \hat{\sigma}_{BM}^2 = 0$. Moreover, $\hat{\sigma}_{PM}^2 = 1/4n_1n_2$.

These lower bounds for the unbiased and Brunner-Munzel variance estimates are motivated by the opposite degenerate case, i.e., completely separated samples without ties such as

$$x_{11} < \dots < x_{1n_1k} < x_{21} < \dots < x_{2n_2k} \quad \text{or} \quad x_{11} > \dots > x_{1n_1k} > x_{21} > \dots > x_{2n_2k},$$

such that either $\hat{p}^2 = \hat{\tau}_0 = \hat{\tau}_1 = \hat{\tau}_2 = 1$ or $\hat{p}^2 = \hat{\tau}_0 = \hat{\tau}_1 = \hat{\tau}_2 = 0$, producing $\hat{\sigma}_N^2 = \hat{\sigma}_{BM}^2 = \sigma_{PM}^2 = 0$. Taking a similar approach as before (see also Brunner et al. 2018 and 2021a), we then pretended the sample was slightly different, i.e.,

$$x_{11} < \dots < x_{1(n_1-1)} < x_{21} < x_{1n_1} < x_{22} < \dots < x_{2n_2k},$$

or

$$x_{11} > \dots > x_{1(n_1-1)} > x_{21} > x_{1n_1} > x_{22} > \dots > x_{2n_2k},$$

yielding a slight change in the effect estimate $\hat{p} = 1 - 1/n_1n_2$ or $\hat{p} = 1/n_1n_2$ respectively. In this changed settings, we would have $\hat{\sigma}_N^2 = \hat{\sigma}_{BM}^2 = 1/n_1^2n_2^2$. As regards the *logit* transformation, we employed the changed Mann-Whitney effect estimates as well so that the resulting test statistics would not be undefined. As for Perme and Manveski's [2019] approach in completely separated samples, we used their new "shift method" to construct confidence intervals for p and rejected the null hypothesis if and only if the number $1/2$ was not an element of this confidence interval.

As far as the degrees of freedom in separated samples are concerned, we assumed $\hat{\sigma}_1^2 = \hat{\sigma}_2^2 > 0$, giving rise to $df = \{N^2(n_1 - 1)(n_2 - 1)\} / \{n_1^2(n_1 - 1) + n_2^2(n_2 - 1)\}$. The other degrees of freedom were chosen accordingly.

However, since we only considered scenarios with $\min(n_1, n_2) \geq 7$, one might just as well have opted to always reject the null hypothesis in case of completely separated samples and never to reject it when all values coincide – no matter which test statistic is at issue. It would have virtually never resulted in a different decision.

More simulation results – Tables and Graphs

Tables 5 to 12 show more detailed simulation results in terms of type I error rates for scenarios already treated in the main manuscript, while Table 13 reports type I error rates for Pauly et al.'s studentised permutation approach [2016] for exponential and binomial distributions not considered before. Figures 3 to 15 depict power curves for a range of distributions, whereas Tables 14 to 16 provide power results for Pauly et al.'s studentised permutation approach as regards a Mann-Whitney effect of $p = 0.7$.

Table 5: Mean variance estimates as regards normal distributions $F_1 = \mathcal{N}(0, \sigma_1^2)$ and $F_2 = \mathcal{N}(0, \sigma_2^2)$ based on 100 000 replications, with $\mathbb{V}(\hat{p})$ and sep denoting the true variance estimand and the relative frequency of the occurrence of completely separated samples respectively

n_1	n_2	σ_1	σ_2	$\mathbb{V}(\hat{p})$	$\hat{\sigma}_N^2$	$\hat{\sigma}_{WMW}^2$	$\hat{\sigma}_{BM}^2$	$\hat{\sigma}_{PM}^2$	sep
7	7	1	1	0.02551020	0.02551208	0.02551020	0.02721286	0.02790682	0.00057
10	7	1	1	0.02142857	0.02142512	0.02142857	0.02261603	0.02321162	0.00009
7	10	1	1	0.02142857	0.02145043	0.02142857	0.02264110	0.02323616	0.00011
10	10	1	1	0.01750000	0.01749250	0.01750000	0.01832616	0.01881817	0.00000
15	15	1	1	0.01148148	0.01147906	0.01148148	0.01184935	0.01211924	0.00000
30	15	1	1	0.00851852	0.00851658	0.00851852	0.00870183	0.00884963	0.00000
15	30	1	1	0.00851852	0.00851616	0.00851852	0.00870137	0.00884913	0.00000
30	30	1	1	0.00564815	0.00564761	0.00564815	0.00574021	0.00582036	0.00000
15	45	1	1	0.00753086	0.00752970	0.00753086	0.00765317	0.00775451	0.00000
15	60	1	1	0.00703704	0.00703405	0.00703704	0.00712667	0.00720374	0.00000
15	75	1	1	0.00674074	0.00674097	0.00674074	0.00681506	0.00687722	0.00000
45	15	1	1	0.00753086	0.00752664	0.00753086	0.00765009	0.00775143	0.00000
60	15	1	1	0.00703704	0.00703678	0.00703704	0.00712939	0.00720646	0.00000
75	15	1	1	0.00674074	0.00674270	0.00674074	0.00681678	0.00687894	0.00000
7	7	1	3	0.02887661	0.02888177	0.02551020	0.03002283	0.03024767	0.00192
10	7	1	3	0.02785149	0.02784806	0.02142857	0.02864640	0.02885283	0.00100
7	10	1	3	0.02089686	0.02089052	0.02142857	0.02169008	0.02199699	0.00017
10	10	1	3	0.01997431	0.01997592	0.01750000	0.02053410	0.02078088	0.00010
15	15	1	3	0.01319212	0.01319536	0.01148148	0.01344344	0.01359980	0.00000
30	15	1	3	0.01253662	0.01253765	0.00851852	0.01266181	0.01274569	0.00000
15	30	1	3	0.00712746	0.00712714	0.00851852	0.00725117	0.00734695	0.00000
30	30	1	3	0.00653401	0.00653393	0.00564815	0.00659594	0.00664655	0.00000
15	45	1	3	0.00510591	0.00510523	0.00753086	0.00518793	0.00525571	0.00000
15	60	1	3	0.00409514	0.00409414	0.00703704	0.00415615	0.00420844	0.00000
15	75	1	3	0.00348867	0.00348763	0.00674074	0.00353725	0.00357981	0.00000
45	15	1	3	0.01231812	0.01231834	0.00753086	0.01240100	0.01245807	0.00000
60	15	1	3	0.01220887	0.01221170	0.00703704	0.01227375	0.01231707	0.00000
75	15	1	3	0.01214332	0.01214928	0.00674074	0.01219889	0.01223372	0.00000
7	7	1	5	0.03104145	0.03103983	0.02551020	0.03182096	0.03174387	0.00386
10	7	1	5	0.03054540	0.03053915	0.02142857	0.03108554	0.03106235	0.00300
7	10	1	5	0.02199142	0.02198072	0.02142857	0.02252815	0.02262856	0.00039
10	10	1	5	0.02156546	0.02156787	0.01750000	0.02194920	0.02203857	0.00026
15	15	1	5	0.01429217	0.01429766	0.01148148	0.01446706	0.01455033	0.00001
30	15	1	5	0.01400327	0.01400415	0.00851852	0.01408901	0.01413427	0.00000
15	30	1	5	0.00735018	0.00735051	0.00851852	0.00743522	0.00749513	0.00000
30	30	1	5	0.00710368	0.00710316	0.00564815	0.00714553	0.00717718	0.00000
15	45	1	5	0.00503618	0.00503644	0.00753086	0.00509293	0.00513693	0.00000
15	60	1	5	0.00387919	0.00387916	0.00703704	0.00392152	0.00395603	0.00000
15	75	1	5	0.00318499	0.00318451	0.00674074	0.00321842	0.00324679	0.00000
45	15	1	5	0.01390698	0.01390648	0.00753086	0.01396295	0.01399380	0.00000
60	15	1	5	0.01385883	0.01386002	0.00703704	0.01390245	0.01392595	0.00000
75	15	1	5	0.01382994	0.01383470	0.00674074	0.01386858	0.01388745	0.00000

Table 6: Type I error rates for normal distributions $F_1 = \mathcal{N}(0, \sigma_1^2)$ and $F_2 = \mathcal{N}(0, \sigma_2^2)$ based on 100 000 replications at a two-sided nominal significance level of $\alpha = 0.05$ as regards the test statistics T_{WMW} and T_N with t -approximation and different degrees of freedom

n_1	n_2	σ_1	σ_2	T_{WMW}	$T_N(df)$	$T_N(df_1)$	$T_N(df_2)$	$T_N(df_3)$	$T_N(df_4)$
7	7	1	1	0.05318	0.06381	0.05821	0.05527	0.05477	0.06041
10	7	1	1	0.04348	0.06358	0.06034	0.05428	0.05297	0.06242
7	10	1	1	0.04290	0.06265	0.05972	0.05399	0.05251	0.06152
10	10	1	1	0.05320	0.06310	0.06058	0.05696	0.05148	0.06199
15	15	1	1	0.05072	0.05809	0.05736	0.05651	0.04973	0.05786
30	15	1	1	0.04906	0.05595	0.05518	0.05417	0.05116	0.05587
15	30	1	1	0.04911	0.05588	0.05505	0.05431	0.05121	0.05585
30	30	1	1	0.04950	0.05336	0.05327	0.05306	0.04917	0.05332
15	45	1	1	0.04891	0.05535	0.05453	0.05340	0.05308	0.05533
15	60	1	1	0.04889	0.05401	0.05292	0.05192	0.05333	0.05401
15	75	1	1	0.04959	0.05562	0.05432	0.05292	0.05596	0.05555
45	15	1	1	0.04945	0.05556	0.05457	0.05329	0.05305	0.05551
60	15	1	1	0.04930	0.05509	0.05388	0.05262	0.05436	0.05502
75	15	1	1	0.04918	0.05434	0.05309	0.05164	0.05512	0.05426
7	7	1	3	0.07223	0.06006	0.04757	0.04572	0.05560	0.04757
10	7	1	3	0.08066	0.05910	0.05070	0.04509	0.06123	0.05169
7	10	1	3	0.04122	0.05436	0.05316	0.05091	0.04560	0.05369
10	10	1	3	0.07141	0.05720	0.05517	0.05172	0.05195	0.05541
15	15	1	3	0.06833	0.05484	0.05383	0.05235	0.05151	0.05380
30	15	1	3	0.10568	0.05408	0.05265	0.05111	0.05651	0.05341
15	30	1	3	0.03163	0.05351	0.05326	0.05299	0.04671	0.05313
30	30	1	3	0.07001	0.05367	0.05338	0.05308	0.05215	0.05338
15	45	1	3	0.01618	0.05240	0.05210	0.05185	0.04451	0.05226
15	60	1	3	0.00965	0.05170	0.05157	0.05130	0.04416	0.05175
15	75	1	3	0.00616	0.05321	0.05297	0.05263	0.04573	0.05328
45	15	1	3	0.12749	0.05448	0.05322	0.05153	0.05882	0.05419
60	15	1	3	0.14104	0.05398	0.05269	0.05116	0.05974	0.05373
75	15	1	3	0.14588	0.05344	0.05207	0.05068	0.05964	0.05307
7	7	1	5	0.08821	0.06395	0.03841	0.03694	0.06155	0.03766
10	7	1	5	0.09850	0.06031	0.04039	0.03648	0.06859	0.04089
7	10	1	5	0.04932	0.04751	0.04705	0.04573	0.04233	0.04723
10	10	1	5	0.08485	0.04977	0.04885	0.04618	0.04762	0.04876
15	15	1	5	0.08132	0.05361	0.05249	0.05108	0.05200	0.05246
30	15	1	5	0.12597	0.05277	0.05160	0.05022	0.05638	0.05224
15	30	1	5	0.03419	0.05273	0.05244	0.05207	0.04689	0.05203
30	30	1	5	0.08136	0.05328	0.05297	0.05257	0.05253	0.05292
15	45	1	5	0.01548	0.05063	0.05052	0.05039	0.04300	0.05038
15	60	1	5	0.00770	0.05121	0.05117	0.05111	0.04305	0.05110
15	75	1	5	0.00431	0.05140	0.05130	0.05116	0.04373	0.05134
45	15	1	5	0.15064	0.05355	0.05230	0.05092	0.05875	0.05322
60	15	1	5	0.16777	0.05333	0.05194	0.05023	0.05951	0.05302
75	15	1	5	0.17407	0.05311	0.05171	0.05004	0.05931	0.05282

Table 7: Type I error rates for normal distributions $F_1 = \mathcal{N}(0, \sigma_1^2)$ and $F_2 = \mathcal{N}(0, \sigma_2^2)$ based on 100 000 replications at a two-sided nominal significance level of $\alpha = 0.05$ as regards the test statistics T_{WMW} and T_{BM} with t -approximation and different degrees of freedom

n_1	n_2	σ_1	σ_2	T_{WMW}	$T_{BM}(df)$	$T_{BM}(df_1)$	$T_{BM}(df_2)$	$T_{BM}(df_3)$	$T_{BM}(df_4)$
7	7	1	1	0.05318	0.05759	0.05527	0.04796	0.04444	0.05527
10	7	1	1	0.04348	0.05732	0.05453	0.05003	0.04652	0.05593
7	10	1	1	0.04290	0.05653	0.05416	0.04975	0.04611	0.05516
10	10	1	1	0.05320	0.05625	0.05454	0.05225	0.04642	0.05547
15	15	1	1	0.05072	0.05454	0.05381	0.05290	0.04624	0.05432
30	15	1	1	0.04906	0.05341	0.05266	0.05183	0.04871	0.05335
15	30	1	1	0.04911	0.05372	0.05303	0.05207	0.04904	0.05361
30	30	1	1	0.04950	0.05170	0.05160	0.05138	0.04745	0.05168
15	45	1	1	0.04891	0.05353	0.05275	0.05167	0.05116	0.05347
15	60	1	1	0.04889	0.05244	0.05159	0.05044	0.05190	0.05241
15	75	1	1	0.04959	0.05420	0.05290	0.05144	0.05464	0.05418
45	15	1	1	0.04945	0.05362	0.05264	0.05150	0.05125	0.05354
60	15	1	1	0.04930	0.05329	0.05239	0.05120	0.05261	0.05331
75	15	1	1	0.04918	0.05292	0.05174	0.05055	0.05384	0.05289
7	7	1	3	0.07223	0.05751	0.04572	0.04206	0.05082	0.04572
10	7	1	3	0.08066	0.05703	0.04741	0.04320	0.05842	0.04918
7	10	1	3	0.04122	0.05191	0.05114	0.04734	0.04171	0.04881
10	10	1	3	0.07141	0.05358	0.05201	0.04893	0.04961	0.05216
15	15	1	3	0.06833	0.05267	0.05154	0.05036	0.04949	0.05156
30	15	1	3	0.10568	0.05276	0.05137	0.04995	0.05525	0.05208
15	30	1	3	0.03163	0.05159	0.05137	0.05111	0.04474	0.05127
30	30	1	3	0.07001	0.05265	0.05244	0.05218	0.05117	0.05245
15	45	1	3	0.01618	0.05026	0.05011	0.04994	0.04294	0.05022
15	60	1	3	0.00965	0.05014	0.05002	0.04978	0.04290	0.05016
15	75	1	3	0.00616	0.05166	0.05146	0.05114	0.04444	0.05177
45	15	1	3	0.12749	0.05373	0.05237	0.05080	0.05804	0.05319
60	15	1	3	0.14104	0.05353	0.05210	0.05044	0.05906	0.05312
75	15	1	3	0.14588	0.05294	0.05153	0.05009	0.05903	0.05264
7	7	1	5	0.08821	0.06286	0.03694	0.03496	0.05923	0.03694
10	7	1	5	0.09850	0.05970	0.03836	0.03563	0.06734	0.03960
7	10	1	5	0.04932	0.04667	0.04620	0.04335	0.04022	0.04383
10	10	1	5	0.08485	0.04802	0.04721	0.04453	0.04680	0.04718
15	15	1	5	0.08132	0.05215	0.05087	0.04987	0.05054	0.05077
30	15	1	5	0.12597	0.05223	0.05097	0.04948	0.05565	0.05149
15	30	1	5	0.03419	0.05133	0.05112	0.05079	0.04559	0.05074
30	30	1	5	0.08136	0.05252	0.05217	0.05186	0.05181	0.05218
15	45	1	5	0.01548	0.04907	0.04899	0.04890	0.04202	0.04884
15	60	1	5	0.00770	0.04998	0.04992	0.04984	0.04198	0.04980
15	75	1	5	0.00431	0.05018	0.05010	0.05003	0.04260	0.05012
45	15	1	5	0.15064	0.05315	0.05184	0.05044	0.05825	0.05266
60	15	1	5	0.16777	0.05297	0.05155	0.04993	0.05919	0.05266
75	15	1	5	0.17407	0.05279	0.05132	0.04986	0.05906	0.05249

Table 8: Type I error rates for normal distributions $F_1 = \mathcal{N}(0, \sigma_1^2)$ and $F_2 = \mathcal{N}(0, \sigma_2^2)$ based on 100 000 replications at a two-sided nominal significance level of $\alpha = 0.05$ as regards the test statistics T_{WMW} and T_{PM} with t -approximation and different degrees of freedom

n_1	n_2	σ_1	σ_2	T_{WMW}	$T_{PM}(df)$	$T_{PM}(df_1)$	$T_{PM}(df_2)$	$T_{PM}(df_3)$	$T_{PM}(df_4)$
7	7	1	1	0.05318	0.05639	0.05139	0.04304	0.04444	0.05411
10	7	1	1	0.04348	0.05525	0.05165	0.04725	0.04510	0.05371
7	10	1	1	0.04290	0.05494	0.05134	0.04708	0.04464	0.05322
10	10	1	1	0.05320	0.05370	0.05181	0.04856	0.04380	0.05302
15	15	1	1	0.05072	0.05183	0.05109	0.05012	0.04405	0.05163
30	15	1	1	0.04906	0.05163	0.05095	0.05001	0.04691	0.05158
15	30	1	1	0.04911	0.05177	0.05110	0.05004	0.04717	0.05169
30	30	1	1	0.04950	0.05013	0.04998	0.04978	0.04604	0.05013
15	45	1	1	0.04891	0.05221	0.05140	0.05040	0.04985	0.05220
15	60	1	1	0.04889	0.05144	0.05046	0.04945	0.05056	0.05138
15	75	1	1	0.04959	0.05293	0.05174	0.05046	0.05350	0.05288
45	15	1	1	0.04945	0.05208	0.05122	0.05002	0.04985	0.05202
60	15	1	1	0.04930	0.05223	0.05121	0.04995	0.05119	0.05217
75	15	1	1	0.04918	0.05185	0.05075	0.04955	0.05288	0.05181
7	7	1	3	0.07223	0.05626	0.04437	0.03917	0.05082	0.04532
10	7	1	3	0.08066	0.05629	0.04612	0.04175	0.05760	0.05443
7	10	1	3	0.04122	0.05375	0.04814	0.04540	0.04084	0.04766
10	10	1	3	0.07141	0.05220	0.05021	0.04705	0.04767	0.05061
15	15	1	3	0.06833	0.05131	0.05036	0.04926	0.04825	0.05034
30	15	1	3	0.10568	0.05191	0.05047	0.04919	0.05437	0.05111
15	30	1	3	0.03163	0.05026	0.05003	0.04973	0.04355	0.04990
30	30	1	3	0.07001	0.05178	0.05157	0.05124	0.05034	0.05156
15	45	1	3	0.01618	0.04904	0.04881	0.04859	0.04187	0.04892
15	60	1	3	0.00965	0.04901	0.04879	0.04853	0.04174	0.04907
15	75	1	3	0.00616	0.05047	0.05023	0.04978	0.04333	0.05054
45	15	1	3	0.12749	0.05307	0.05171	0.05041	0.05746	0.05266
60	15	1	3	0.14104	0.05299	0.05155	0.05000	0.05860	0.05271
75	15	1	3	0.14588	0.05257	0.05124	0.04973	0.05871	0.05227
7	7	1	5	0.08821	0.06181	0.03626	0.03319	0.05923	0.03663
10	7	1	5	0.09850	0.05944	0.03790	0.03504	0.06700	0.05817
7	10	1	5	0.04932	0.05706	0.04410	0.04229	0.03960	0.04338
10	10	1	5	0.08485	0.04748	0.04596	0.04376	0.04546	0.04606
15	15	1	5	0.08132	0.05136	0.05019	0.04928	0.04981	0.05016
30	15	1	5	0.12597	0.05169	0.05037	0.04907	0.05529	0.05113
15	30	1	5	0.03419	0.05050	0.05021	0.04985	0.04479	0.04985
30	30	1	5	0.08136	0.05194	0.05166	0.05128	0.05128	0.05165
15	45	1	5	0.01548	0.04823	0.04810	0.04800	0.04138	0.04784
15	60	1	5	0.00770	0.04914	0.04908	0.04898	0.04093	0.04902
15	75	1	5	0.00431	0.04915	0.04910	0.04899	0.04195	0.04906
45	15	1	5	0.15064	0.05276	0.05164	0.05015	0.05778	0.05235
60	15	1	5	0.16777	0.05273	0.05118	0.04979	0.05890	0.05236
75	15	1	5	0.17407	0.05257	0.05114	0.04961	0.05894	0.05217

Table 9: Mean variance estimates as regards the 5-point distributions with latent $F_1 = \mathcal{B}(\alpha_1, \beta_1)$ and $F_2 = \mathcal{B}(5, 4)$ based on 100 000 replications, with $\mathbb{V}(\hat{p})$ and sep denoting the true variance estimand and the relative frequency of the occurrence of completely separated samples respectively

n_1	n_2	α_1	β_1	$\mathbb{V}(\hat{p})$	$\hat{\sigma}_N^2$	$\hat{\sigma}_{WMW}^2$	$\hat{\sigma}_{BM}^2$	$\hat{\sigma}_{PM}^2$	sep
7	7	5	4	0.02135081	0.02134561	0.02136648	0.02178287	0.02333416	0.00009
10	7	5	4	0.01808278	0.01810689	0.01809247	0.01841288	0.01955837	0.00001
7	10	5	4	0.01808278	0.01808238	0.01809247	0.01838821	0.01953452	0.00001
10	10	5	4	0.01485400	0.01486281	0.01486064	0.01507701	0.01592065	0.00000
15	15	5	4	0.00985519	0.00985734	0.00985846	0.00995248	0.01035631	0.00000
30	15	5	4	0.00736765	0.00737328	0.00736927	0.00742084	0.00762987	0.00000
15	30	5	4	0.00736765	0.00737174	0.00736927	0.00741930	0.00762835	0.00000
30	30	5	4	0.00490385	0.00490477	0.00490462	0.00492854	0.00503659	0.00000
15	45	5	4	0.00653847	0.00654040	0.00653950	0.00657210	0.00671303	0.00000
15	60	5	4	0.00612388	0.00612619	0.00612402	0.00614996	0.00625626	0.00000
15	75	5	4	0.00587513	0.00587719	0.00587532	0.00589620	0.00598153	0.00000
45	15	5	4	0.00653847	0.00653712	0.00653950	0.00656879	0.00670973	0.00000
60	15	5	4	0.00612388	0.00612755	0.00612402	0.00615131	0.00625762	0.00000
75	15	5	4	0.00587513	0.00587810	0.00587532	0.00589710	0.00598242	0.00000
7	7	1.2071	1	0.02458929	0.02473559	0.02303112	0.02530736	0.02629003	0.00037
10	7	1.2071	1	0.01858647	0.01864873	0.01961235	0.01904997	0.01984263	0.00002
7	10	1.2071	1	0.02308031	0.02321632	0.01929195	0.02361703	0.02434381	0.00012
10	10	1.2071	1	0.01712232	0.01719060	0.01596299	0.01747059	0.01805212	0.00002
15	15	1.2071	1	0.01136812	0.01140776	0.01056297	0.01153218	0.01182431	0.00000
30	15	1.2071	1	0.00673455	0.00673987	0.00797797	0.00680209	0.00696066	0.00000
15	30	1.2071	1	0.01026745	0.01029732	0.00774917	0.01035954	0.01051018	0.00000
30	30	1.2071	1	0.00566068	0.00566779	0.00524136	0.00569890	0.00578042	0.00000
15	45	1.2071	1	0.00990409	0.00993534	0.00680155	0.00997680	0.01007819	0.00000
15	60	1.2071	1	0.00972377	0.00974447	0.00632434	0.00977558	0.00985204	0.00000
15	75	1.2071	1	0.00961620	0.00963242	0.00603741	0.00965731	0.00971867	0.00000
45	15	1.2071	1	0.00518592	0.00518722	0.00710263	0.00522868	0.00533715	0.00000
60	15	1.2071	1	0.00441069	0.00441226	0.00666124	0.00444337	0.00452578	0.00000
75	15	1.2071	1	0.00394525	0.00394590	0.00639503	0.00397079	0.00403718	0.00000

Table 10: Type I error rates for the 5-point distributions with latent $F_1 = \mathcal{B}(\alpha_1, \beta_1)$ and $F_2 = \mathcal{B}(5, 4)$ based on 100 000 replications at a two-sided nominal significance level of $\alpha = 0.05$ as regards the test statistics T_{WMW} and T_N with t -approximation and different degrees of freedom

n_1	n_2	α_1	β_1	T_{WMW}	$T_N(df)$	$T_N(df_1)$	$T_N(df_2)$	$T_N(df_3)$	$T_N(df_4)$
7	7	5	4	0.04611	0.06294	0.05821	0.05628	0.04818	0.06200
10	7	5	4	0.04761	0.06245	0.05867	0.05391	0.04932	0.06146
7	10	5	4	0.04769	0.06237	0.05892	0.05457	0.04928	0.06124
10	10	5	4	0.04832	0.06102	0.05821	0.05450	0.04865	0.06063
15	15	5	4	0.04875	0.05580	0.05517	0.05440	0.04825	0.05570
30	15	5	4	0.04814	0.05393	0.05308	0.05212	0.04889	0.05389
15	30	5	4	0.04902	0.05532	0.05466	0.05391	0.05067	0.05526
30	30	5	4	0.04857	0.05226	0.05202	0.05175	0.04802	0.05225
15	45	5	4	0.04923	0.05476	0.05375	0.05258	0.05255	0.05480
15	60	5	4	0.05026	0.05531	0.05411	0.05292	0.05466	0.05537
15	75	5	4	0.04959	0.05517	0.05403	0.05263	0.05590	0.05518
45	15	5	4	0.04898	0.05401	0.05294	0.05186	0.05199	0.05404
60	15	5	4	0.04868	0.05392	0.05296	0.05194	0.05360	0.05393
75	15	5	4	0.04856	0.05345	0.05208	0.05090	0.05408	0.05345
7	7	1.2071	1	0.05763	0.06234	0.05709	0.05126	0.05274	0.05870
10	7	1.2071	1	0.04264	0.05924	0.05621	0.05281	0.04510	0.05696
7	10	1.2071	1	0.07446	0.06349	0.05771	0.05213	0.05829	0.06074
10	10	1.2071	1	0.05798	0.05992	0.05740	0.05432	0.05162	0.05843
15	15	1.2071	1	0.05579	0.05340	0.05245	0.05146	0.04793	0.05281
30	15	1.2071	1	0.03218	0.05412	0.05383	0.05345	0.04716	0.05398
15	30	1.2071	1	0.08897	0.05436	0.05315	0.05188	0.05484	0.05390
30	30	1.2071	1	0.05827	0.05173	0.05149	0.05120	0.04915	0.05156
15	45	1.2071	1	0.10304	0.05312	0.05175	0.05027	0.05639	0.05277
15	60	1.2071	1	0.11529	0.05395	0.05294	0.05152	0.05833	0.05380
15	75	1.2071	1	0.12079	0.05404	0.05257	0.05096	0.05932	0.05384
45	15	1.2071	1	0.02028	0.05380	0.05342	0.05296	0.04658	0.05386
60	15	1.2071	1	0.01493	0.05258	0.05207	0.05145	0.04643	0.05276
75	15	1.2071	1	0.01190	0.05324	0.05252	0.05180	0.04777	0.05340

Table 11: Type I error rates for the 5-point distributions with latent $F_1 = \mathcal{B}(\alpha_1, \beta_1)$ and $F_2 = \mathcal{B}(5, 4)$ based on 100 000 replications at a two-sided nominal significance level of $\alpha = 0.05$ as regards the test statistics T_{WMW} and T_{BM} with t -approximation and different degrees of freedom

n_1	n_2	α_1	β_1	T_{WMW}	$T_{BM}(df)$	$T_{BM}(df_1)$	$T_{BM}(df_2)$	$T_{BM}(df_3)$	$T_{BM}(df_4)$
7	7	5	4	0.04611	0.06189	0.05808	0.05045	0.04776	0.05866
10	7	5	4	0.04761	0.06048	0.05711	0.05298	0.04742	0.05976
7	10	5	4	0.04769	0.06040	0.05723	0.05350	0.04756	0.05953
10	10	5	4	0.04832	0.05881	0.05668	0.05316	0.04747	0.05814
15	15	5	4	0.04875	0.05496	0.05421	0.05315	0.04744	0.05485
30	15	5	4	0.04814	0.05318	0.05232	0.05140	0.04836	0.05308
15	30	5	4	0.04902	0.05477	0.05409	0.05315	0.04977	0.05469
30	30	5	4	0.04857	0.05159	0.05142	0.05115	0.04748	0.05154
15	45	5	4	0.04923	0.05417	0.05315	0.05201	0.05199	0.05416
15	60	5	4	0.05026	0.05486	0.05361	0.05248	0.05422	0.05490
15	75	5	4	0.04959	0.05480	0.05362	0.05228	0.05545	0.05481
45	15	5	4	0.04898	0.05342	0.05239	0.05144	0.05127	0.05339
60	15	5	4	0.04868	0.05351	0.05263	0.05142	0.05318	0.05348
75	15	5	4	0.04856	0.05303	0.05181	0.05054	0.05369	0.05308
7	7	1.2071	1	0.05763	0.05988	0.05598	0.04755	0.05076	0.05678
10	7	1.2071	1	0.04264	0.05560	0.05366	0.05086	0.04357	0.05462
7	10	1.2071	1	0.07446	0.06077	0.05624	0.05046	0.05599	0.05909
10	10	1.2071	1	0.05798	0.05816	0.05533	0.05293	0.04954	0.05635
15	15	1.2071	1	0.05579	0.05221	0.05141	0.05031	0.04682	0.05171
30	15	1.2071	1	0.03218	0.05309	0.05288	0.05251	0.04634	0.05305
15	30	1.2071	1	0.08897	0.05357	0.05250	0.05131	0.05417	0.05325
30	30	1.2071	1	0.05827	0.05115	0.05092	0.05065	0.04840	0.05100
15	45	1.2071	1	0.10304	0.05249	0.05135	0.04976	0.05595	0.05218
15	60	1.2071	1	0.11529	0.05373	0.05260	0.05116	0.05793	0.05350
15	75	1.2071	1	0.12079	0.05372	0.05224	0.05064	0.05903	0.05348
45	15	1.2071	1	0.02028	0.05292	0.05242	0.05196	0.04578	0.05306
60	15	1.2071	1	0.01493	0.05177	0.05129	0.05075	0.04573	0.05192
75	15	1.2071	1	0.01190	0.05247	0.05186	0.05111	0.04696	0.05266

Table 12: Type I error rates for the 5-point distributions with latent $F_1 = \mathcal{B}(\alpha_1, \beta_1)$ and $F_2 = \mathcal{B}(5, 4)$ based on 100 000 replications at a two-sided nominal significance level of $\alpha = 0.05$ as regards the test statistics T_{WMW} and T_{PM} with t -approximation and different degrees of freedom

n_1	n_2	α_1	β_1	T_{WMW}	$T_{PM}(df)$	$T_{PM}(df_1)$	$T_{PM}(df_2)$	$T_{PM}(df_3)$	$T_{PM}(df_4)$
7	7	5	4	0.04611	0.05712	0.05114	0.04129	0.04050	0.05519
10	7	5	4	0.04761	0.05410	0.05124	0.04431	0.04190	0.05298
7	10	5	4	0.04769	0.05465	0.05167	0.04426	0.04192	0.05363
10	10	5	4	0.04832	0.05175	0.05097	0.04798	0.04230	0.05153
15	15	5	4	0.04875	0.05055	0.05006	0.04910	0.04307	0.05048
30	15	5	4	0.04814	0.04998	0.04915	0.04825	0.04543	0.04992
15	30	5	4	0.04902	0.05166	0.05076	0.04990	0.04677	0.05164
30	30	5	4	0.04857	0.04915	0.04894	0.04875	0.04518	0.04914
15	45	5	4	0.04923	0.05158	0.05071	0.04961	0.04925	0.05168
15	60	5	4	0.05026	0.05260	0.05161	0.05041	0.05220	0.05276
15	75	5	4	0.04959	0.05300	0.05186	0.05055	0.05373	0.05302
45	15	5	4	0.04898	0.05093	0.05008	0.04909	0.04871	0.05101
60	15	5	4	0.04868	0.05173	0.05076	0.04940	0.05106	0.05176
75	15	5	4	0.04856	0.05124	0.05019	0.04906	0.05204	0.05126
7	7	1.2071	1	0.05763	0.05605	0.05002	0.04387	0.04589	0.05261
10	7	1.2071	1	0.04264	0.05278	0.04991	0.04506	0.04032	0.05047
7	10	1.2071	1	0.07446	0.05799	0.05296	0.04612	0.05315	0.05534
10	10	1.2071	1	0.05798	0.05378	0.05169	0.04872	0.04558	0.05254
15	15	1.2071	1	0.05579	0.04953	0.04858	0.04752	0.04430	0.04898
30	15	1.2071	1	0.03218	0.05090	0.05065	0.05029	0.04391	0.05086
15	30	1.2071	1	0.08897	0.05202	0.05091	0.04977	0.05278	0.05161
30	30	1.2071	1	0.05827	0.04972	0.04946	0.04917	0.04702	0.04956
15	45	1.2071	1	0.10304	0.05132	0.05009	0.04876	0.05475	0.05104
15	60	1.2071	1	0.11529	0.05284	0.05160	0.05029	0.05679	0.05268
15	75	1.2071	1	0.12079	0.05288	0.05140	0.05008	0.05821	0.05266
45	15	1.2071	1	0.02028	0.05040	0.05004	0.04966	0.04394	0.05048
60	15	1.2071	1	0.01493	0.04991	0.04935	0.04874	0.04385	0.05004
75	15	1.2071	1	0.01190	0.05064	0.05003	0.04917	0.04494	0.05078

Table 13: Type I error rates for exponential and binomial distributions at a two-sided nominal significance level of $\alpha = 0.05$ for the studentised permutation tests based on 10 000 random permutations for each of the 10 000 replications

n_1	n_2	F_1	F_2	\tilde{T}_N	\tilde{T}_{BM}	\tilde{T}_{PM}	\tilde{T}_N^{Logit}	\tilde{T}_{BM}^{Logit}	\tilde{T}_{PM}^{Logit}
7	7	$\mathcal{E}(1)$	$\mathcal{E}(1)$	0.0488	0.0506	0.0502	0.0493	0.0484	0.0479
7	10	$\mathcal{E}(1)$	$\mathcal{E}(1)$	0.0484	0.0495	0.0496	0.0477	0.0485	0.0487
10	7	$\mathcal{E}(1)$	$\mathcal{E}(1)$	0.0460	0.0473	0.0473	0.0465	0.0475	0.0473
10	10	$\mathcal{E}(1)$	$\mathcal{E}(1)$	0.0503	0.0501	0.0504	0.0509	0.0508	0.0509
15	15	$\mathcal{E}(1)$	$\mathcal{E}(1)$	0.0503	0.0501	0.0502	0.0503	0.0504	0.0502
15	30	$\mathcal{E}(1)$	$\mathcal{E}(1)$	0.0478	0.0481	0.0484	0.0484	0.0481	0.0480
30	15	$\mathcal{E}(1)$	$\mathcal{E}(1)$	0.0533	0.0531	0.0531	0.0531	0.0530	0.0531
30	30	$\mathcal{E}(1)$	$\mathcal{E}(1)$	0.0508	0.0508	0.0507	0.0506	0.0506	0.0507
15	45	$\mathcal{E}(1)$	$\mathcal{E}(1)$	0.0498	0.0497	0.0498	0.0503	0.0501	0.0502
45	15	$\mathcal{E}(1)$	$\mathcal{E}(1)$	0.0499	0.0499	0.0499	0.0506	0.0504	0.0501
7	7	$\mathcal{B}(5, 0.6)$	$\mathcal{B}(5, 0.6)$	0.0319	0.0321	0.0325	0.0308	0.0309	0.0311
7	10	$\mathcal{B}(5, 0.6)$	$\mathcal{B}(5, 0.6)$	0.0341	0.0345	0.0350	0.0346	0.0348	0.0350
10	7	$\mathcal{B}(5, 0.6)$	$\mathcal{B}(5, 0.6)$	0.0353	0.0355	0.0356	0.0347	0.0349	0.0353
10	10	$\mathcal{B}(5, 0.6)$	$\mathcal{B}(5, 0.6)$	0.0420	0.0415	0.0415	0.0424	0.0425	0.0422
15	15	$\mathcal{B}(5, 0.6)$	$\mathcal{B}(5, 0.6)$	0.0453	0.0454	0.0456	0.0459	0.0461	0.0458
15	30	$\mathcal{B}(5, 0.6)$	$\mathcal{B}(5, 0.6)$	0.0508	0.0507	0.0510	0.0505	0.0505	0.0507
30	15	$\mathcal{B}(5, 0.6)$	$\mathcal{B}(5, 0.6)$	0.0469	0.0470	0.0470	0.0470	0.0469	0.0469
30	30	$\mathcal{B}(5, 0.6)$	$\mathcal{B}(5, 0.6)$	0.0498	0.0498	0.0497	0.0496	0.0496	0.0496
15	45	$\mathcal{B}(5, 0.6)$	$\mathcal{B}(5, 0.6)$	0.0491	0.0490	0.0493	0.0489	0.0489	0.0490
45	15	$\mathcal{B}(5, 0.6)$	$\mathcal{B}(5, 0.6)$	0.0514	0.0514	0.0515	0.0515	0.0516	0.0516

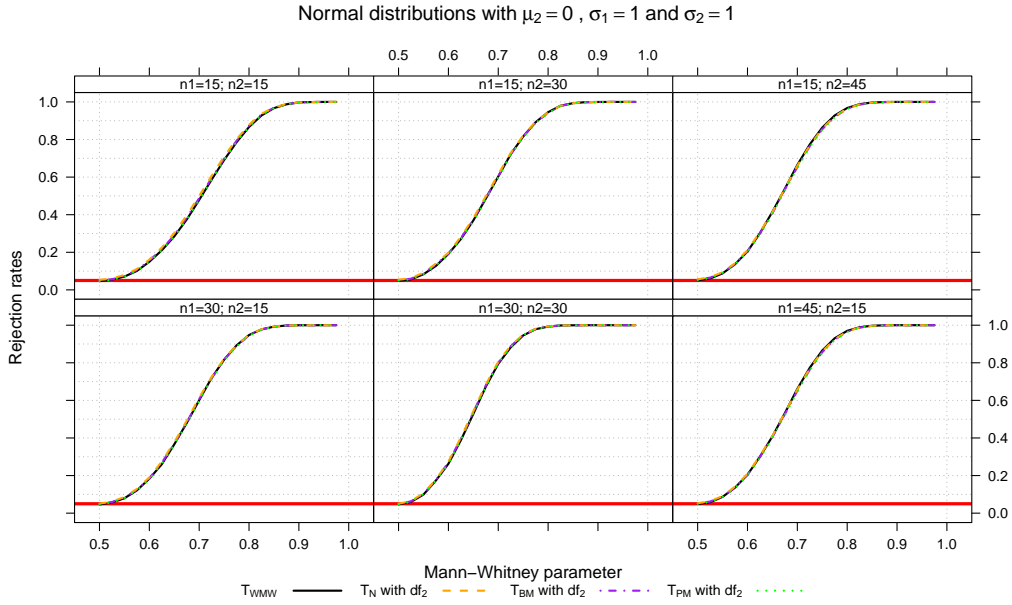


Figure 3: Power graphs for normal distributions based on 10 000 simulation runs

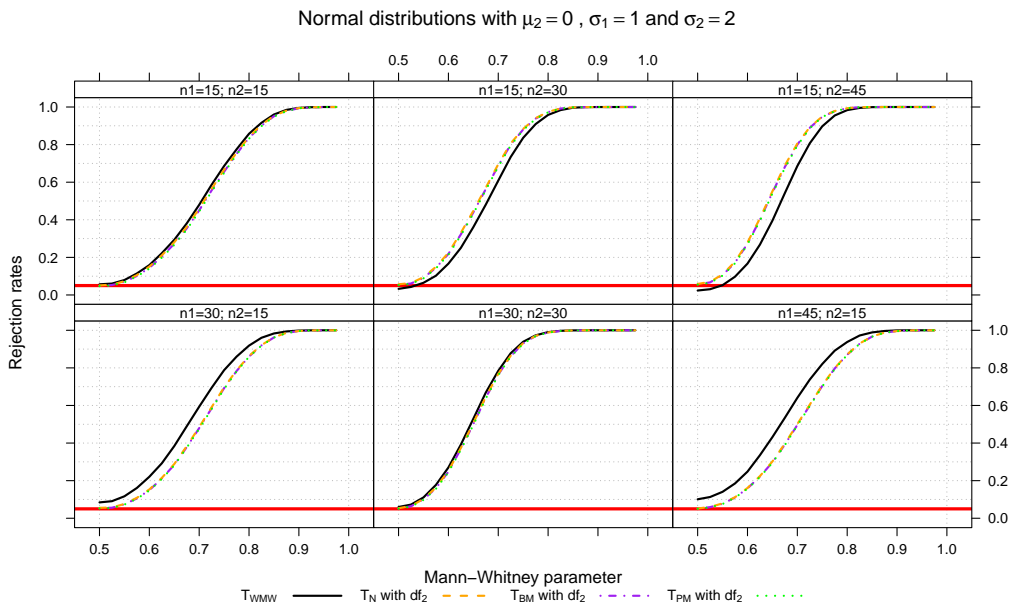


Figure 4: Power graphs for normal distributions based on 10 000 simulation runs

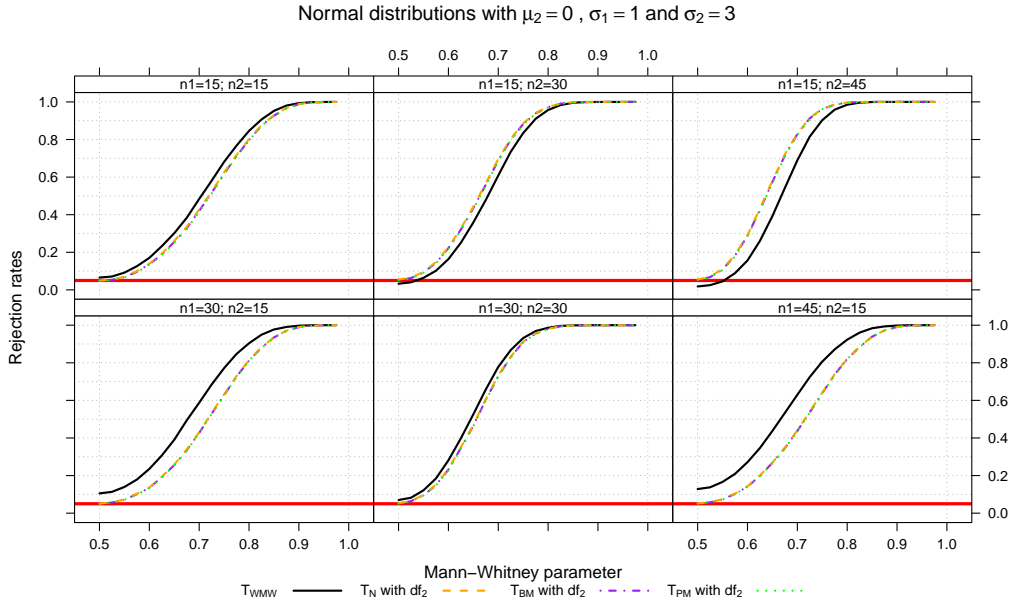


Figure 5: Power graphs for normal distributions based on 10 000 simulation runs

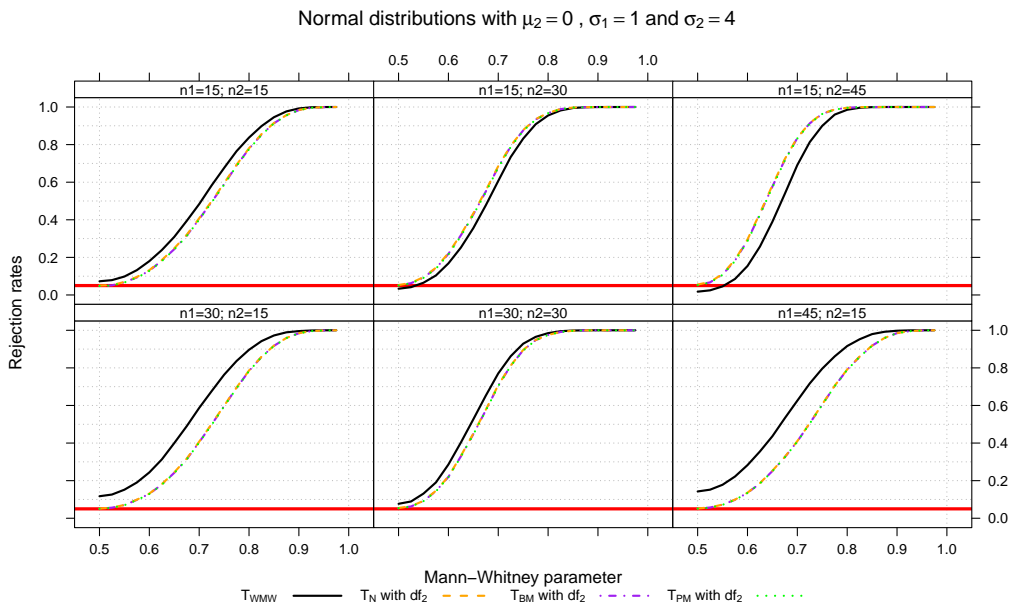


Figure 6: Power graphs for normal distributions based on 10 000 simulation runs

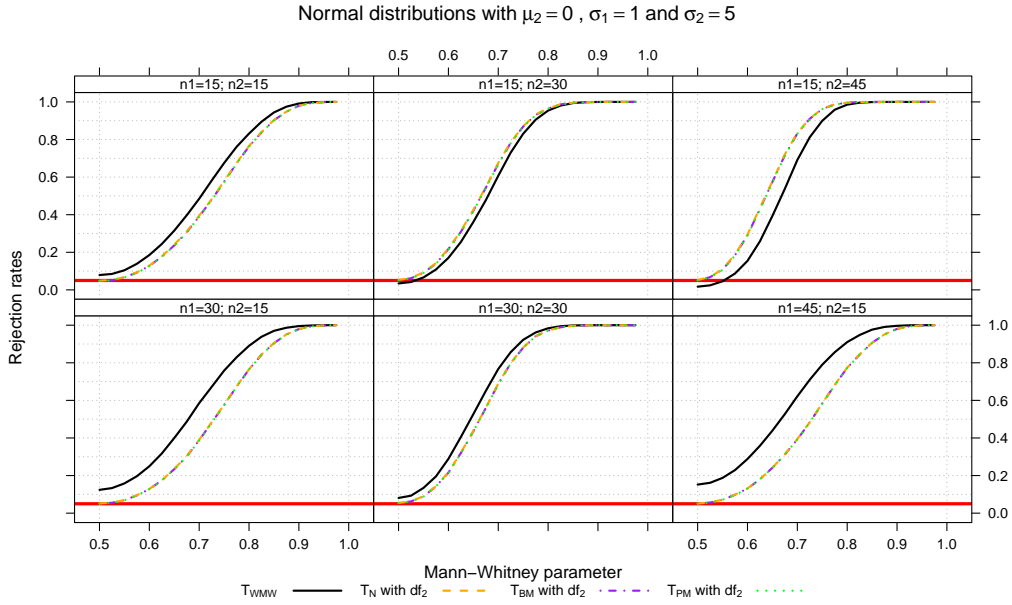


Figure 7: Power graphs for normal distributions based on 10 000 simulation runs

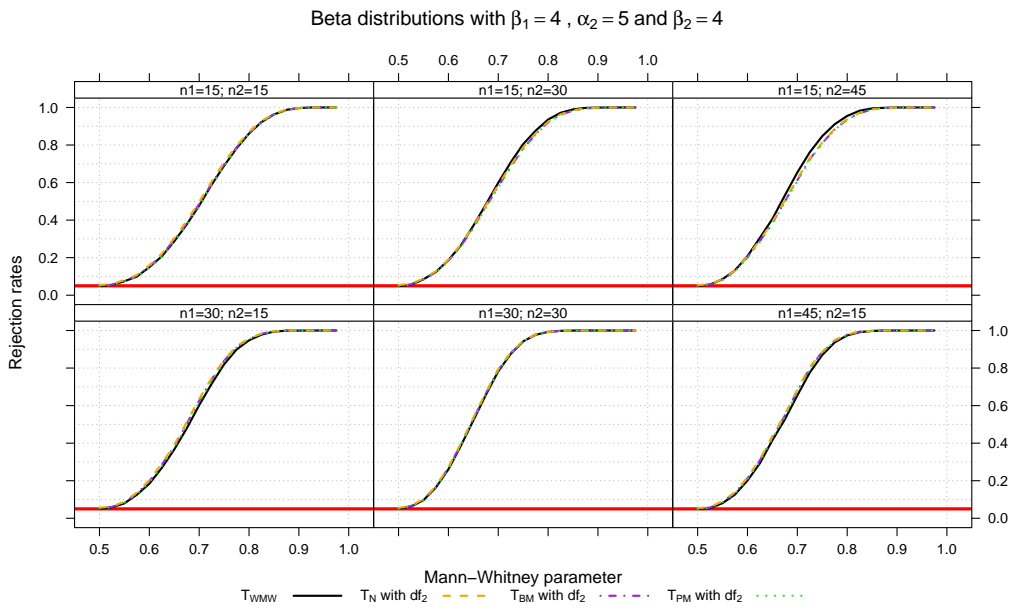


Figure 8: Power graphs for Beta distributions based on 10 000 simulation runs

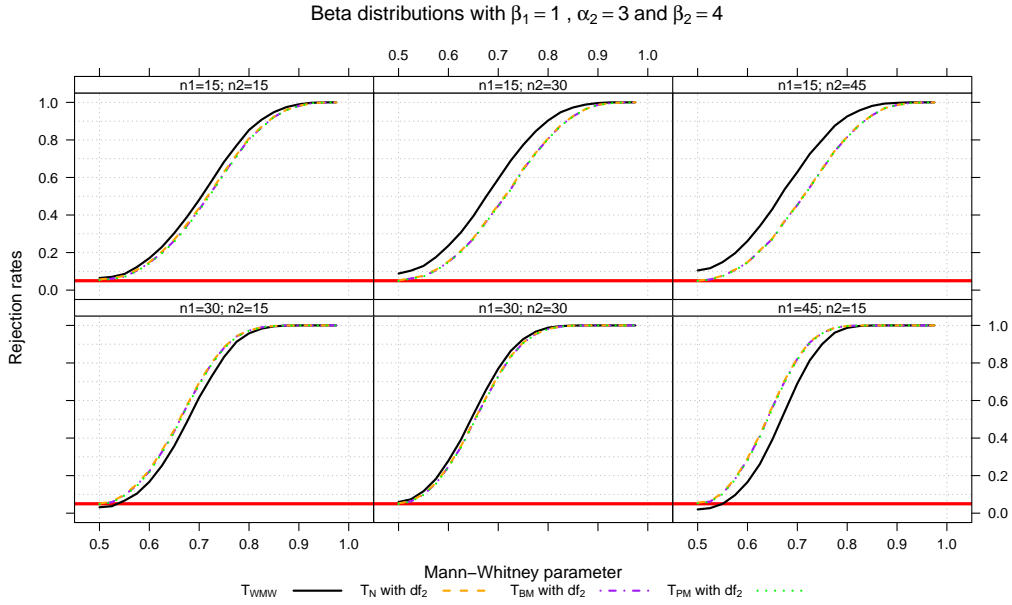


Figure 9: Power graphs for Beta distributions based on 10 000 simulation runs

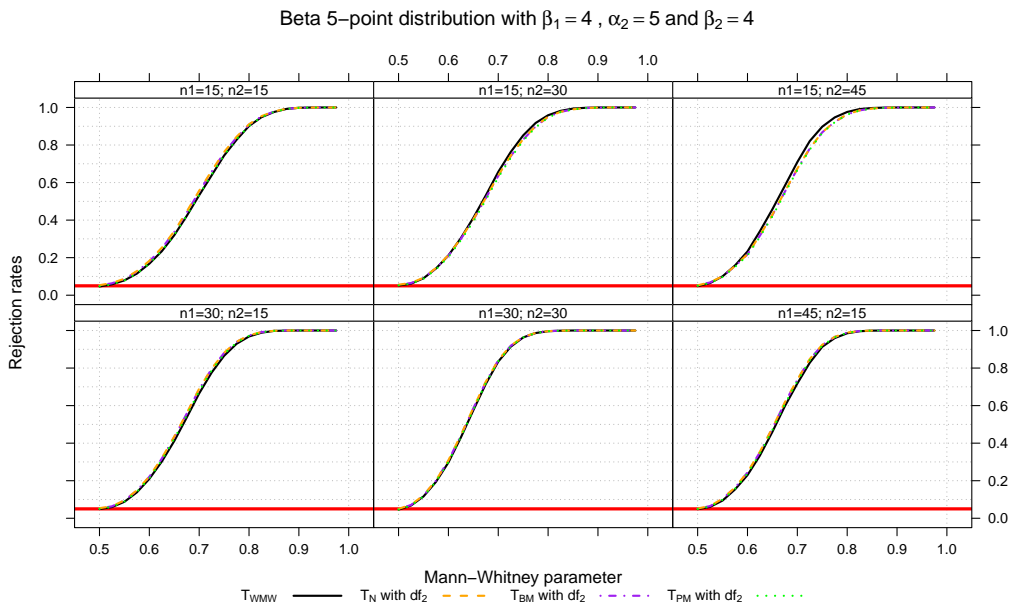


Figure 10: Power graphs for Beta 5-point distributions based on 10 000 simulation runs

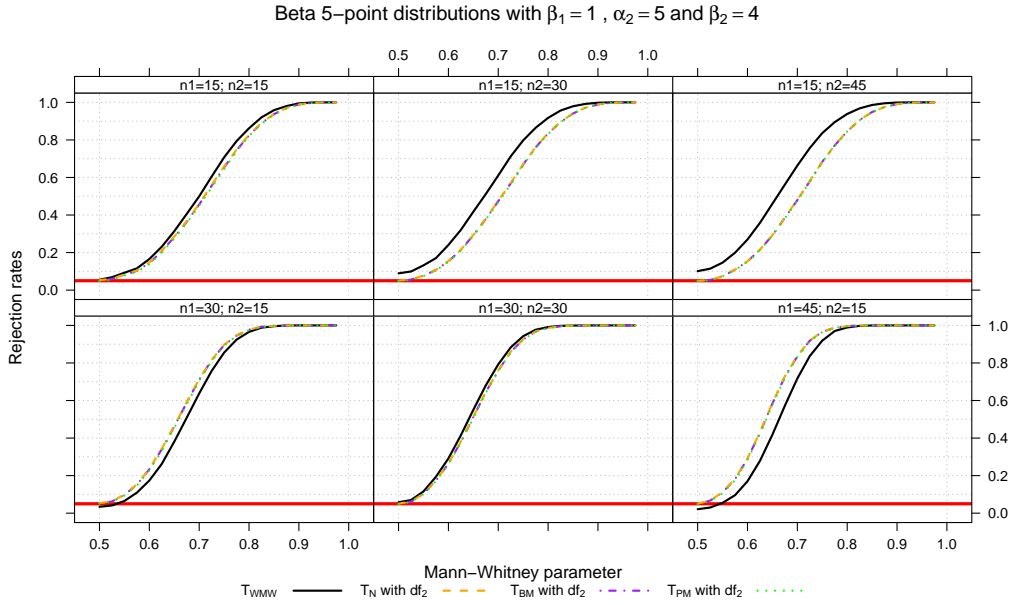


Figure 11: Power graphs for Beta 5-point distributions based on 10 000 simulation runs

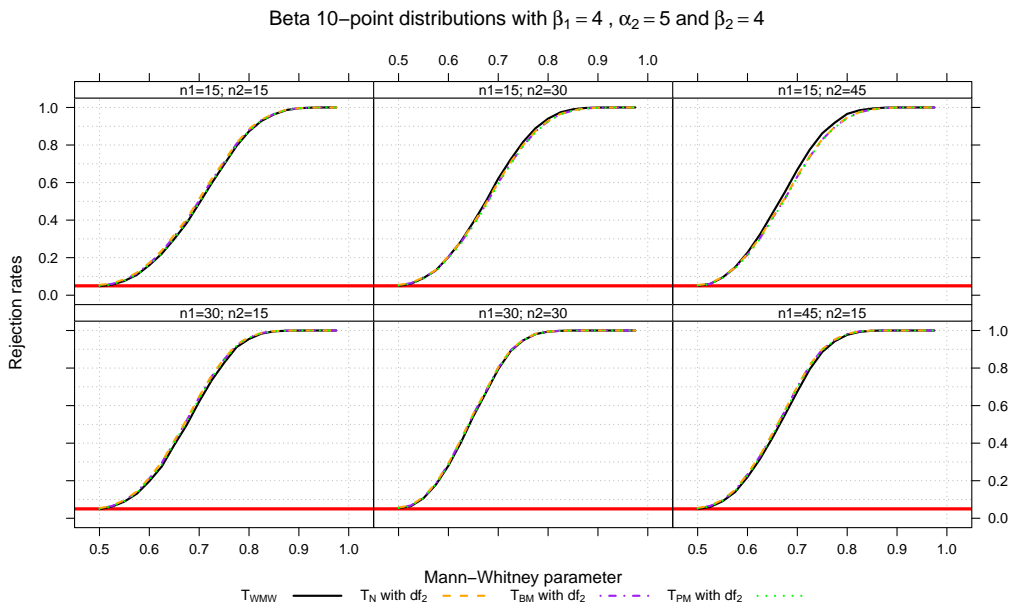


Figure 12: Power graphs for Beta 10-point distributions based on 10 000 simulation runs

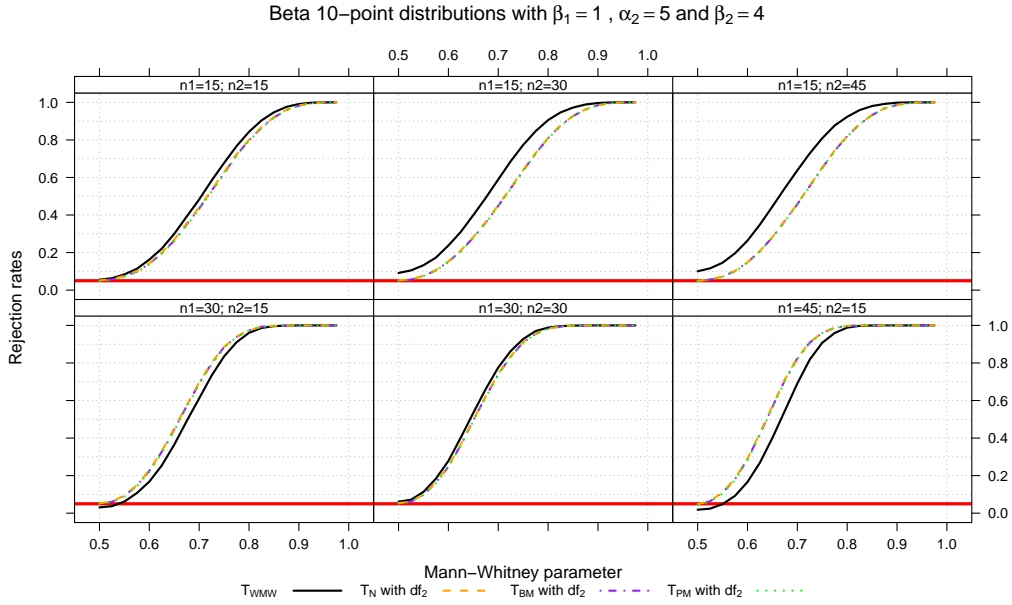


Figure 13: Power graphs for Beta 10-point distributions based on 10 000 simulation runs

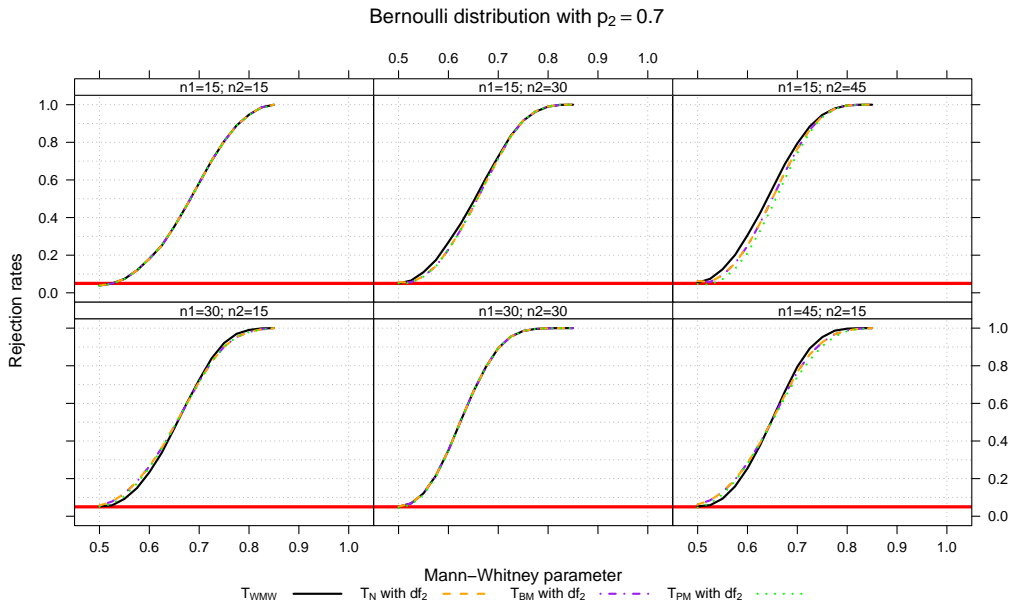


Figure 14: Power graphs for Bernoulli distributions based on 10 000 simulation runs

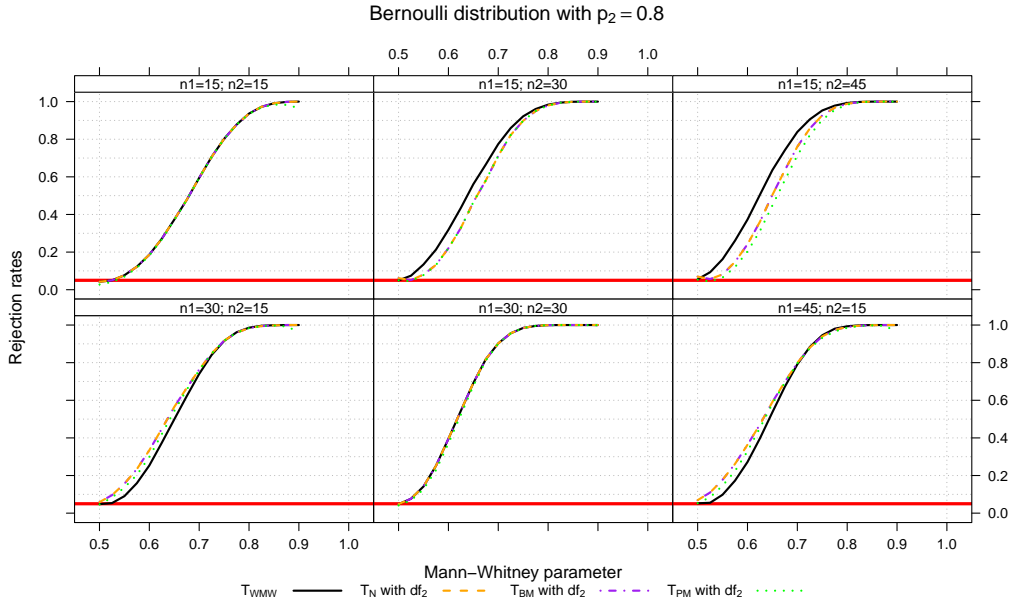


Figure 15: Power graphs for Bernoulli distributions based on 10 000 simulation runs

Table 14: Power results for Mann-Whitney parameter $p = 0.7$ as regards normal distributions $F_1 = \mathcal{N}(\mu_1, \sigma_1^2)$ and $F_2 = \mathcal{N}(\mu_2, \sigma_2^2)$ at a two-sided nominal significance level of $\alpha = 0.05$ for the studentised permutation tests based on 10 000 random permutations for each of the 10 000 replications

n_1	n_2	μ_1	μ_2	σ_1	σ_2	\hat{T}_N	\hat{T}_{BM}	\hat{T}_{PM}	\hat{T}_N^{Logit}	\hat{T}_{BM}^{Logit}	\hat{T}_{PM}^{Logit}
7	7	-0.7416	0	1	1	0.2264	0.2294	0.2293	0.2253	0.2224	0.2248
7	10	-0.7416	0	1	1	0.2719	0.2742	0.2746	0.2676	0.2695	0.2706
10	7	-0.7416	0	1	1	0.2734	0.2760	0.2757	0.2695	0.2715	0.2708
10	10	-0.7416	0	1	1	0.3365	0.3364	0.3358	0.3374	0.3373	0.3367
15	15	-0.7416	0	1	1	0.4883	0.4879	0.4887	0.4882	0.4882	0.4879
15	30	-0.7416	0	1	1	0.6116	0.6116	0.6116	0.6076	0.6079	0.6088
30	15	-0.7416	0	1	1	0.6046	0.6048	0.6045	0.6032	0.6037	0.6034
30	30	-0.7416	0	1	1	0.7839	0.7841	0.7843	0.7848	0.7843	0.7839
15	45	-0.7416	0	1	1	0.6531	0.6529	0.6533	0.6497	0.6502	0.6502
45	15	-0.7416	0	1	1	0.6442	0.6441	0.6445	0.6395	0.6401	0.6398
7	7	-1.6583	0	1	3	0.2110	0.2335	0.2353	0.1712	0.1740	0.1831
7	10	-1.6583	0	1	3	0.2677	0.2748	0.2841	0.2604	0.2733	0.2781
10	7	-1.6583	0	1	3	0.2449	0.2555	0.2598	0.1789	0.1827	0.1843
10	10	-1.6583	0	1	3	0.3052	0.3147	0.3219	0.2793	0.2870	0.2920
15	15	-1.6583	0	1	3	0.4314	0.4358	0.4398	0.4140	0.4199	0.4259
15	30	-1.6583	0	1	3	0.6650	0.6667	0.6686	0.6747	0.6783	0.6803
30	15	-1.6583	0	1	3	0.4582	0.4623	0.4669	0.4184	0.4241	0.4282
30	30	-1.6583	0	1	3	0.7209	0.7237	0.7253	0.7134	0.7161	0.7183
15	45	-1.6583	0	1	3	0.8098	0.8109	0.8111	0.8216	0.8221	0.8224
45	15	-1.6583	0	1	3	0.4560	0.4598	0.4625	0.4079	0.4133	0.4188

Table 15: Power results for Mann-Whitney parameter $p = 0.7$ as regards 5-point Beta distributions with latent $F_1 = \mathcal{B}(\alpha_1, \beta_1)$ and $F_2 = \mathcal{B}(5, 4)$ at a two-sided nominal significance level of $\alpha = 0.05$ for the studentised permutation tests based on 10 000 random permutations for each of the 10 000 replications

n_1	n_2	α_1	β_1	\tilde{T}_N	\tilde{T}_{BM}	\tilde{T}_{PM}	\tilde{T}_N^{Logit}	\tilde{T}_{BM}^{Logit}	\tilde{T}_{PM}^{Logit}
7	7	2.86332	4	0.1948	0.1944	0.1936	0.1915	0.1905	0.1899
7	10	2.86332	4	0.2476	0.2477	0.2480	0.2458	0.2445	0.2455
10	7	2.86332	4	0.2555	0.2559	0.2551	0.2546	0.2553	0.2564
10	10	2.86332	4	0.3263	0.3260	0.3267	0.3251	0.3251	0.3261
15	15	2.86332	4	0.5060	0.5059	0.5054	0.5058	0.5058	0.5059
15	30	2.86332	4	0.6268	0.6270	0.6281	0.6228	0.6230	0.6236
30	15	2.86332	4	0.6586	0.6584	0.6584	0.6598	0.6596	0.6598
30	30	2.86332	4	0.8314	0.8314	0.8313	0.8313	0.8312	0.8314
15	45	2.86332	4	0.6697	0.6697	0.6706	0.6628	0.6631	0.6646
45	15	2.86332	4	0.7183	0.7180	0.7183	0.7192	0.7189	0.7191
7	7	0.57606	1	0.2158	0.2168	0.2238	0.1907	0.1914	0.1956
7	10	0.57606	1	0.2437	0.2464	0.2531	0.1997	0.2004	0.2032
10	7	0.57606	1	0.2710	0.2745	0.2787	0.2722	0.2761	0.2839
10	10	0.57606	1	0.3126	0.3154	0.3200	0.2924	0.2953	0.2998
15	15	0.57606	1	0.4630	0.4647	0.4686	0.4500	0.4516	0.4569
15	30	0.57606	1	0.5012	0.5027	0.5065	0.4704	0.4726	0.4773
30	15	0.57606	1	0.6977	0.6983	0.6991	0.7069	0.7079	0.7090
30	30	0.57606	1	0.7631	0.7634	0.7653	0.7578	0.7588	0.7603
15	45	0.57606	1	0.4971	0.4985	0.5020	0.4576	0.4593	0.4649
45	15	0.57606	1	0.8100	0.8103	0.8097	0.8237	0.8241	0.8239

Table 16: Power results for Mann-Whitney parameter $p = 0.7$ as regards exponential and binomial distributions at a two-sided nominal significance level of $\alpha = 0.05$ for the studentised permutation tests based on 10 000 random permutations for each of the 10 000 replications

n_1	n_2	F_1	F_2	\tilde{T}_N	\tilde{T}_{BM}	\tilde{T}_{PM}	\tilde{T}_N^{Logit}	\tilde{T}_{BM}^{Logit}	\tilde{T}_{PM}^{Logit}
7	7	$\mathcal{E}(2.33333)$	$\mathcal{E}(1)$	0.2302	0.2336	0.2328	0.2274	0.2256	0.2281
7	10	$\mathcal{E}(2.33333)$	$\mathcal{E}(1)$	0.2784	0.2809	0.2810	0.2817	0.2831	0.2839
10	7	$\mathcal{E}(2.33333)$	$\mathcal{E}(1)$	0.2765	0.2797	0.2810	0.2638	0.2658	0.2659
10	10	$\mathcal{E}(2.33333)$	$\mathcal{E}(1)$	0.3417	0.3420	0.3424	0.3412	0.3419	0.3418
15	15	$\mathcal{E}(2.33333)$	$\mathcal{E}(1)$	0.4797	0.4789	0.4796	0.4799	0.4801	0.4796
15	30	$\mathcal{E}(2.33333)$	$\mathcal{E}(1)$	0.6429	0.6425	0.6423	0.6478	0.6471	0.6464
30	15	$\mathcal{E}(2.33333)$	$\mathcal{E}(1)$	0.5646	0.5655	0.5661	0.5562	0.5570	0.5572
30	30	$\mathcal{E}(2.33333)$	$\mathcal{E}(1)$	0.7921	0.7921	0.7917	0.7922	0.7923	0.7922
15	45	$\mathcal{E}(2.33333)$	$\mathcal{E}(1)$	0.7002	0.6993	0.6992	0.7049	0.7040	0.7038
45	15	$\mathcal{E}(2.33333)$	$\mathcal{E}(1)$	0.5947	0.5956	0.5974	0.5808	0.5823	0.5832
7	7	$\mathcal{B}(5, 0.43129)$	$\mathcal{B}(5, 0.6)$	0.2072	0.2090	0.2088	0.2044	0.2041	0.2049
7	10	$\mathcal{B}(5, 0.43129)$	$\mathcal{B}(5, 0.6)$	0.2602	0.2616	0.2620	0.2559	0.2566	0.2582
10	7	$\mathcal{B}(5, 0.43129)$	$\mathcal{B}(5, 0.6)$	0.2679	0.2674	0.2676	0.2653	0.2653	0.2667
10	10	$\mathcal{B}(5, 0.43129)$	$\mathcal{B}(5, 0.6)$	0.3323	0.3324	0.3325	0.3323	0.3321	0.3327
15	15	$\mathcal{B}(5, 0.43129)$	$\mathcal{B}(5, 0.6)$	0.5140	0.5138	0.5131	0.5150	0.5152	0.5148
15	30	$\mathcal{B}(5, 0.43129)$	$\mathcal{B}(5, 0.6)$	0.6426	0.6426	0.6430	0.6393	0.6395	0.6410
30	15	$\mathcal{B}(5, 0.43129)$	$\mathcal{B}(5, 0.6)$	0.6339	0.6341	0.6342	0.6327	0.6322	0.6323
30	30	$\mathcal{B}(5, 0.43129)$	$\mathcal{B}(5, 0.6)$	0.8161	0.8162	0.8160	0.8165	0.8165	0.8162
15	45	$\mathcal{B}(5, 0.43129)$	$\mathcal{B}(5, 0.6)$	0.6803	0.6805	0.6812	0.6738	0.6738	0.6746
45	15	$\mathcal{B}(5, 0.43129)$	$\mathcal{B}(5, 0.6)$	0.6834	0.6835	0.6834	0.6825	0.6824	0.6824