bmtest: a Jamovi module for Brunner–Munzel's test: a robust alternative to Wilcoxon–Mann–Whitney's test

Karch, J.D.

psych

MDPI

Article

bmtest: A Jamovi Module for Brunner–Munzel's Test—A Robust Alternative to Wilcoxon–Mann–Whitney's Test

Julian D. Karch

Methodology and Statistics Department, Institute of Psychology, Leiden University, P.O. Box 9555, 2300 RB Leiden, The Netherlands; j.d.karch@fsw.leidenuniv.nl

Abstract: In psychological research, comparisons between two groups are frequently made to demonstrate that one group exhibits higher values. Although Welch's unequal variances *t*-test has become the preferred parametric test for this purpose, surpassing Student's equal variances *t*-test, the Wilcoxon–Mann–Whitney test remains the predominant nonparametric approach despite sharing similar limitations with Student's *t*-test. Specifically, the Wilcoxon–Mann–Whitney test is associated with strong, unrealistic assumptions and lacks robustness when these assumptions are violated. The Brunner–Munzel test overcomes these limitations, featuring fewer assumptions, akin to Welch's *t*-test in the parametric domain, and has thus been recommended over the Wilcoxon–Mann–Whitney test. However, the Brunner–Munzel test is currently unavailable in user-friendly statistical software, such as SPSS, making it inaccessible to many researchers. In this paper, I introduce the bmtest module for jamovi, a freely available user-friendly software. By making the Brunner–Munzel test accessible to a wide range of researchers, the bmtest module has the potential to improve nonparametric statistical analysis in psychology and other disciplines.

Keywords: Brunner–Munzel test; jamovi; Wilcoxon–Mann–Whitney test; nonparametric; robust

Citation: Karch, J.D. bmtest: A Jamovi Module for Brunner–Munzel's Test—A Robust Alternative to Wilcoxon–Mann–Whitney's Test. *Psych* **2023**, *5*, 386–395. https://doi.org/10.3390/psych5020026

Academic Editor: Alexander Robitzsch

Received: 31 March 2023
Revised: 24 April 2023
Accepted: 3 May 2023
Published: 10 May 2023

1. Introduction

Researchers in psychology and other fields often compare two independent groups. Typically, the goal is to demonstrate that a particular group tends to exhibit greater values, for example, to verify that a new therapy method leads to better outcomes. In line with Ref. [1], I call this goal "establishing direction". The standard approaches for establishing direction between two groups are Student's and Welch's *t*-tests, as well as the Wilcoxon–Mann–Whitney test (alternatively known as Mann–Whitney U, Mann–Whitney–Wilcoxon, or Wilcoxon rank-sum test) and their associated confidence intervals.

While the *t*-tests are used most often, nonparametric tests should be preferred over the parametric *t*-tests in some situations. The WMW test is almost always recommended and used for nonparametrically comparing two groups [2]. The main advantage of the WMW test compared to the *t*-tests is that it does not assume interval or normally distributed data. Thus, it can be beneficial when one or both of these assumptions are not met.

Unfortunately, the Wilcoxon–Mann–Whitney test is often wrongly used in psychology. The primary issue is that its corresponding null and alternative hypotheses, as well as associated assumptions, are frequently misinterpreted. The WMW test is typically presented as either a test of equality of population medians [3] or equality of the two populations in all aspects, which is referred to as equality of distributions [4]. If paired with the appropriate assumptions, these hypotheses are correct in the sense that the WMW test is valid (correct Type I error rate) and consistent (power approaches 1 with increasing sample size) [5]. However, the assumptions associated with these hypotheses are not realistic in psychology, and if they are not met, the WMW test can have severely inflated Type I error rates and almost 0 power, even in large samples [2].

The most general perspective on the WMW test, in that all other (correct) perspectives are a special case of it with stricter assumptions, is the Mann–Whitney functional perspective [5]. The null hypothesis is equality of distributions, and the alternative hypothesis is that the relative effect p is unequal to 0.5. Another reason to consider $p \neq 0.5$ as the alternative hypothesis of the WMW test is computational: The WMW test essentially uses a standardized sample estimate of the relative effect p as its test statistic (see Section 2).

To explain the relative effect p, it is temporarily assumed that tied values never occur. Then, the relative effect p is the probability that a random observation from the first group is less than a random observation from the second. If $p = 0.5$, it is equally likely that a random observation from the first group is larger than a random observation from the second group as it is that a random observation from the first group is smaller than one from the second group. In this case, the two groups are referred to as (stochastically) comparable [6]. $p \neq 0.5$ is consequently labeled not comparable.

Even under the Mann–Whitney functional perspective, the WMW test is associated with unrealistic assumptions. As Ref. [5] note, it is hard to imagine a situation in which it is scientifically justified to assume that the distributions are either equal or not comparable ($p \neq 0.5$). The perspective emerges more as a description of situations under which the WMW test is correct. A modification of this perspective, making it more realistic, is to replace the null hypothesis of equal distributions with the null hypothesis of the groups not being comparable: $H_0 : p = 0.5$ [2,5], (Chapter 8.8 of Ref. [7]). In line with Ref. [6], I refer to this as the stochastic comparability perspective (in the statistical literature, this problem is famously known as the nonparametric Behrens–Fisher problem).

Multiple issues can emerge when using the WMW test under the more realistic stochastic comparability perspective. If groups differ, for example, due to unequal variances, under general circumstances, the wrong standard errors are used. This, in turn, leads to inflated Type I error rates, poor power, and unsatisfactory confidence intervals, particularly when sample sizes are different [2], (Chapter 8.8 of Ref. [7]). Thus, the WMW test is associated with shortcomings similar to those of Student's t-test [8]. Indeed, computationally, it is essentially the nonparametric analog to Student's t-test (see Section 2). Much like Welch's t-test in the parametric realm, the Brunner–Munzel (BM) test addresses these shortcomings [2,6,9]. Indeed, the BM test is essentially the nonparametric analog of Welch's t-test (see Section 2). Consequently, in line with the recommendation to use Welch's t-test instead of Student's t-test by default [8], the Brunner–Munzel test has been recommended for use by default [2,9].

However, many researchers may be unable to use the BM test. While the BM test is available in multiple R packages, including lawstat [10], brunnermunzel [11], nparcomp [12], and an SAS macro [6], it is not available in any user-friendly graphical user interface (GUI)-based software, such as SPSS. Thus, researchers who rely exclusively on user-friendly GUI-based software do not have access to it.

To address this, I introduce the bmtest module for jamovi [13]. Like jamovi, the bmtest module is freely available and open source, setting it apart from almost all other user-friendly statistical software programs, including SPSS, SAS, and Stata, which are proprietary and require payment. As a side product, the bmtest jamovi module is also available as the R package bmtest. Its unique advantage over existing R packages is that it offers all commonly used versions of the BM test.

The remainder of this paper is organized as follows. First, I will provide a brief introduction to the BM test. Next, I will offer a step-by-step tutorial on how to use the jamovi module. Finally, I will provide a brief overview of the R package.

2. Brunner–Munzel Test

I will explain the BM test using a fictitious example introduced in a popular statistics book [14] for explaining the WMW test. The example concerns two groups of clubbers: one group was given an ecstasy tablet to take on Saturday night, and another drank alcohol.

Levels of depression were measured using the Beck Depression Inventory on the day after (Sunday) and midweek (Wednesday).

2.1. Relative Effect

I will start by describing the relative effect in more detail. Applied to the example, the relative effect p represents the probability that a randomly selected ecstasy consumer will report less severe depressive feelings than a randomly selected alcohol drinker. To formally define the relative effect, I introduce the (random) variables X_1 and X_2, which represent random observations from the first and second groups, respectively. Excluding ties, the relative effect is calculated as $p = P(X_1 < X_2)$. To allow for ties, the probability of a tie is assigned with equal weight to both possibilities (X_1 smaller, and X_2 smaller) and thus with weight $\frac{1}{2}$ to the relative effect. The resulting relative effect allowing for ties is given by the equation:

$$p = P(X_1 < X_2) + \frac{1}{2}P(X_1 = X_2).$$

If the relative effect is $p = 0.5$, groups 1 and 2 are deemed *(stochastically) comparable*, which is the null hypothesis of the BM test. For two-sided testing, the alternative hypothesis is that $H_A : p \neq 0.5$, indicating that the groups are not comparable. For one-sided testing, the alternative hypothesis can be either $H_A : p > 0.5$, indicating that X_1 *tends to take smaller values*, or $H_A : p < 0.5$, indicating that X_1 *tends to take greater values*.

To further illustrate the relative effect, consider two binary random variables X_1 and X_2 with potential values of 0 and 1 with $P(X_1 = 0) = 0.7$ and $P(X_2 = 0) = 0.1$. Thus, X_1 tends to take smaller values. This is reflected in the relative effect as follows: given that $X_1 = 0$, there is a probability of 90% that X_2 is bigger and a probability of 10% that it is equal. Thus, the relative effect given $X_1 = 0$ is $(0.9 + \frac{1}{2} \times 0.1) = 0.95$. Similarly, given that $X_1 = 1$, there is a probability of 0% that X_2 is bigger and a probability of 90% that it is equal, and thus the relative effect given $X_1 = 1$ is $0 + \frac{1}{2} \times 0.9 = 0.45$. Consequently, considering that $P(X_1 = 0) = 0.7$ and $P(X_1 = 1) = 0.3$, the (unconditional) relative effect is $p = 0.7 \times 0.95 + 0.3 \times 0.45 = 0.8$.

2.2. Test Statistic

The test statistic for the BM test is based on the sample relative effect, which is an unbiased and consistent (approaches the true value with increasing sample size) estimator of the relative effect. First, consider that the observations for the two groups of size n_1, n_2 are $X_{11}, \ldots, X_{1n_1}$, and $X_{21}, \ldots, X_{2n_2}$. The sample relative effect is then defined as follows:

$$\hat{p} = \frac{1}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} S(X_{1i}, X_{2j}).$$

Here, $S(x_1, x_2) = 1$ if $x_1 < x_2$, $S(x_1, x_2)=\frac{1}{2}$ if $x_1 = x_2$, and otherwise $S(x_1, x_2) = 0$. The sample relative effect thus considers all possible pairs of observations (X_{1i}, X_{2j}). For a pair for which X_{1i} is smaller than X_{2j}, one is added to a count. For X_{1i} being equal to X_{2j}, $\frac{1}{2}$ is added, otherwise nothing is added. The sample relative effect is the count divided by the number of pairs ($n_1 n_2$). The sample relative effect thus essentially summarizes the outcome of a competition between all observations from both groups [3]. For every match a group wins, 1 is awarded to it, and for ties, $\frac{1}{2}$. The sample relative effect is the proportion of points won by the second group (X_2).

Computing the sample relative effect in this manner is inefficient and not implemented in software programs, as all $n_1 n_2$ pairs need to be considered. Instead, the relative effect can be computed based on ranked data, which also reveals that the BM test is rank-based. To achieve this, the rank R_{ik} of observation X_{ik} among the pooled sample $X_{11}, \ldots, X_{1n_1}, X_{21}, \ldots, X_{2n_2}$ is calculated. For tied values, the mid-rank is assigned, which

is the average of the ranks among the tied values. Let $\bar{R}_i$ be the average rank of the observations in group i. Then, the sample relative effect can be computed as follows:

$$\hat{p} = \frac{1}{N}(\bar{R}_2 - \bar{R}_1) + \frac{1}{2},$$

where $N = n_1 + n_2$.

The BM test standardizes the sample relative effect $\hat{p}$ such that it approximately has a standard normal distribution under the null hypothesis of $p = 0.5$; that is, it approaches a normal distribution with increasing sample size. The WMW test can be defined equivalently. To achieve this, both tests subtract the expected value under the null (0.5) and divide by an estimated standard error denoted by *se*. This leads to the following equivalent test statistics, which are named U and W; as in the case of the WMW test, they are essentially equivalent to Mann–Whitney's U and Wilcoxon's W.

$$U = \frac{\hat{p} - 0.5}{se(\hat{p})} \text{ or equivalently } W = \frac{\bar{R}_2 - \bar{R}_1}{se(\bar{R}_2 - \bar{R}_1)}$$

The crucial difference between the BM and the WMW tests lies in the employed estimate of the standard error. The estimate employed by the WMW test is only valid under the restrictive assumption of the location shift model [5,9], which, for example, excludes unequal variances between the groups. In contrast, the standard error employed by the BM test is valid without these restrictive assumptions. Indeed, the test statistic of the BM test is approximately standard normally distributed and thus leads to approximately valid tests under very general assumptions [6].

The formulas of standard errors and the resulting test statistics for the WMW and the BM test are as follows.

$$WMW = \frac{\bar{R}_2 - \bar{R}_1}{\sqrt{S_R^2(1/n_1 + 1/n_2)}} \quad BM = \frac{\bar{R}_2 - \bar{R}_1}{N\sqrt{S_{1R}^2/n_1 + S_{2R}^2/n_2}}. \tag{1}$$

Here, S_R^2 is a pooled variance estimator of the ranked data, and S_{iR}^2 is closely related to a variance estimator of the ranked data in group i (see Ref. [9] for the formulas). Equation (1) reveals that the WMW test is essentially the rank-based version of Student's t-test and the BM test, the rank-based version of Welch's t-test (compare Equation (1) to the t-test formulas given, for example, in Ref. [9], Equation (2.2)). This provides further insight into why the BM test provides accurate results under less restrictive assumptions than the WMW test.

The BM statistic can consequently be interpreted equivalently to the t statistic for Welch's t-test. Large positive values indicate that the first group tends to take smaller values, whereas large negative values indicate that the first group tends to take greater values. With increasing sample size, the BM statistic is approximately standard normally distributed. Thus, when testing two-sided with a significance level of $\alpha = 0.05$ and reasonably large sample sizes for both groups, a $|BM| > 1.96$ indicates a significant result.

2.3. From Test Statistic to p-Value

There are different approaches for converting the BM statistic into a p-value, leading to different versions of the BM test: as for Welch's t-test, the normal approximation is not accurate enough in smaller samples. To remedy this, a more accurate approximation based on the t distribution has been proposed [15]. This approach is called the asymptotic approach and is recommended instead of the normal approximation. The asymptotic approach has been shown to perform satisfactorily for sample sizes as small as $n_1, n_2 \geq 10$ [15]. The degrees of freedom of the t distribution are estimated based on the rank data similar to Welch's method. For the precise formula, see (Ref. [7] Section 7.8.6) and for the derivation (Ref. [6] Section 3.5.2).

An alternative approach is the permutation testing method [16]. This method permutes the assignments to the groups and saves the resulting test statistic BM_i for each permutation i. This procedure is repeated for every possible permutation to estimate a permutation sampling distribution under the null hypothesis of no difference between the groups. The p-value is obtained by counting the proportion of permutations for which the observed test statistic BM without permutation is more extreme than the permutation test statistics BM_i. The permutation method is the most accurate approach and shows reasonable performance for group sizes as low as $n_1, n_2 = 7$ [17]. The disadvantage of the permutation approach is that it is computationally unfeasible for moderate sample sizes because the number of possible permutations grows rapidly with sample size [2].

A variant of the permutation approach addressing the computational issue is random permutation testing. Instead of considering all permutations, only a small random subset of permutations is used. This enables permutation testing for all sample sizes at the cost of a decrease in accuracy.

While it is clear that the full permutation approach will provide the most accurate p-value, depending on the circumstances, either the asymptotic or random permutation approach will be more accurate. However, recently, it has been shown that the random permutation approach is more robust in terms of Type I error rates and power compared to the asymptotic approach in many situations as they commonly occur in psychology [9]. Thus, the three approaches to obtain a p-value can be interpreted as different trade-offs between computational complexity and accuracy. The difference between the approaches vanishes with increasing sample size as they all converge to the correct value.

In practice, my advice is as follows. Due to the computational complexity of the full permutation approach, it can only be used in very small samples. At the moment, this is essentially restricted to cases where n_1, n_2 are both not substantially larger than 10 [2]. In such small samples, its increased accuracy is meaningful, and thus it should be used. In all other cases, in principle, I recommend the random permutation approach, following the results and advice given in Ref. [9] with 10,000 random permutations. In case this also takes too much time, which should generally only be the case for quite large samples, one can fall back to the asymptotic approach, which can be expected to be accurate enough in such large samples.

2.4. Confidence Intervals

All three versions of the BM test (even the permutation versions) can be inverted to construct confidence intervals for the relative effect p [18]. The procedure is the same for all versions and also the same as used for other tests, such as the t-tests. Exemplarily, the 95% confidence interval contains all those values a for which the corresponding null hypothesis $H_0 : p = a$ with associated test statistic $\frac{\hat{p}-a}{se(\hat{p})}$ and significance level $\alpha = 0.05$ cannot be rejected. This approach is valid if the true relative effect p is not close to 0 or 1 and the sample size is large enough, which should be the case for most applications in psychology with moderate sample size since a relative effect of 0 or 1 implies that all observations in one group are smaller than all observations of the other group, which is rare in psychology. If these conditions are not met, Ref. [2] provides guidance on alternative approaches that can still provide valid confidence intervals, which, however, are only implemented in other R packages, most prominently, the nparcomp package [12].

3. bmtest Jamovi Module

3.1. Installation

The jamovi program can be downloaded from https://www.jamovi.org/download.html (accessed on 15 March 2023) and is available for all commonly used operating systems. Alternatively, jamovi can also be used via the jamovi cloud. However, this is not recommended, as it currently does not allow for the installation of the bmtest module. Installation varies by operating system but is straightforward; therefore, it is not described in detail. For this tutorial, I used version 2.3.21 for Windows.

The next step is to install the bmtest module. Generally, there are two ways to install jamovi modules: installing from the jamovi store or installing manually from a .jmo file. At the time of writing, the bmtest module can only be installed manually. The latest .jmo file can be obtained from https://github.com/karchjd/bmtest/releases/. For this paper, I used version 0.1.0, which is the latest version at the time of writing. Note that the appropriate jmo file for the operating system in use must be downloaded.

The module can be installed by clicking on "+ Modules", "Jamovi Library", "Sideload", and then the upward-facing arrow (see Figure 1). This opens a file manager window in which the downloaded .jmo file should be opened, which will install the module. If you read this at a later point, the latest version of the bmtest module should also be available in the jamovi store and can be installed directly from the store.

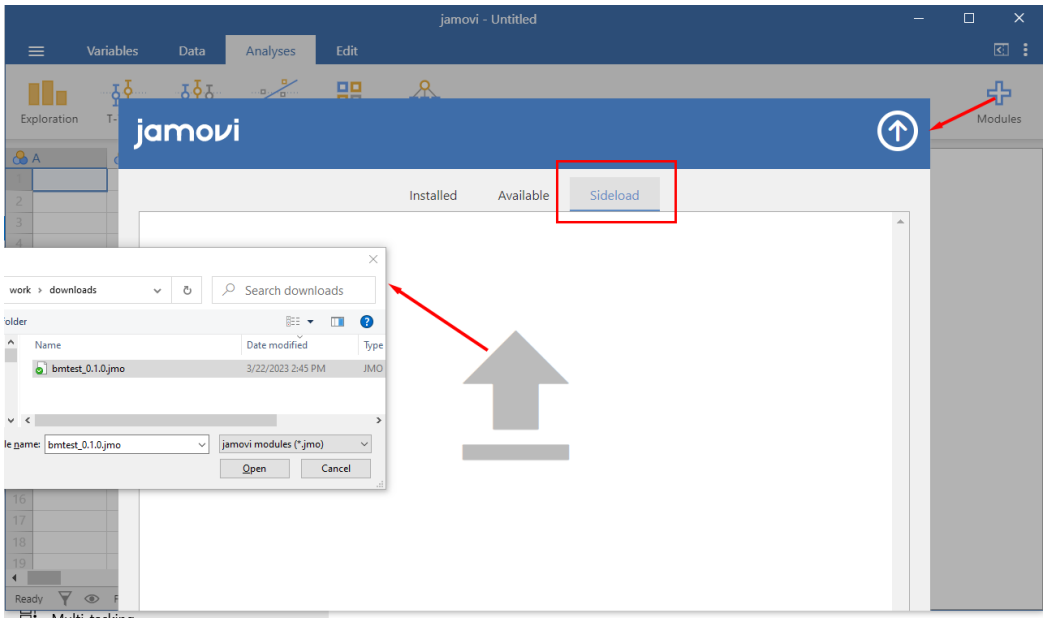

Figure 1. Installing the bmtest Module.

3.2. Usage

The first step is to open a dataset. As mentioned in the previous section, the fictitious example drug dataset contains data from two groups of clubbers: 30 individuals were given an ecstasy tablet to take on Saturday night, and 30 individuals consumed alcohol. Levels of depression were measured using the Beck Depression Inventory (BDI) on the day after (Sunday) and midweek (Wednesday). The drug dataset can be downloaded via this link: https://osf.io/download/nug79/ (accessed on 15 March 2023). The first step is to open the dataset. As this process is straightforward, it will not be explained.

To access the BM test, click on the "BM Test" menu and select "Brunner-Munzel Test". This opens the BM test menu (see Figure 2). The first step is to select the "Dependent Variables" and the "Grouping Variable". The variables "Sunday_BDI" and "Wednesday_BDI" contain the reported frequency of depressive feelings and are thus selected as dependent variables. Note that multiple dependent variables can be specified, and a BM test is performed for each dependent variable separately. The variable "Drug" is the grouping variable and is consequently selected as such. Selecting variables is done in the same manner as always in jamovi and is essentially the same as in SPSS.

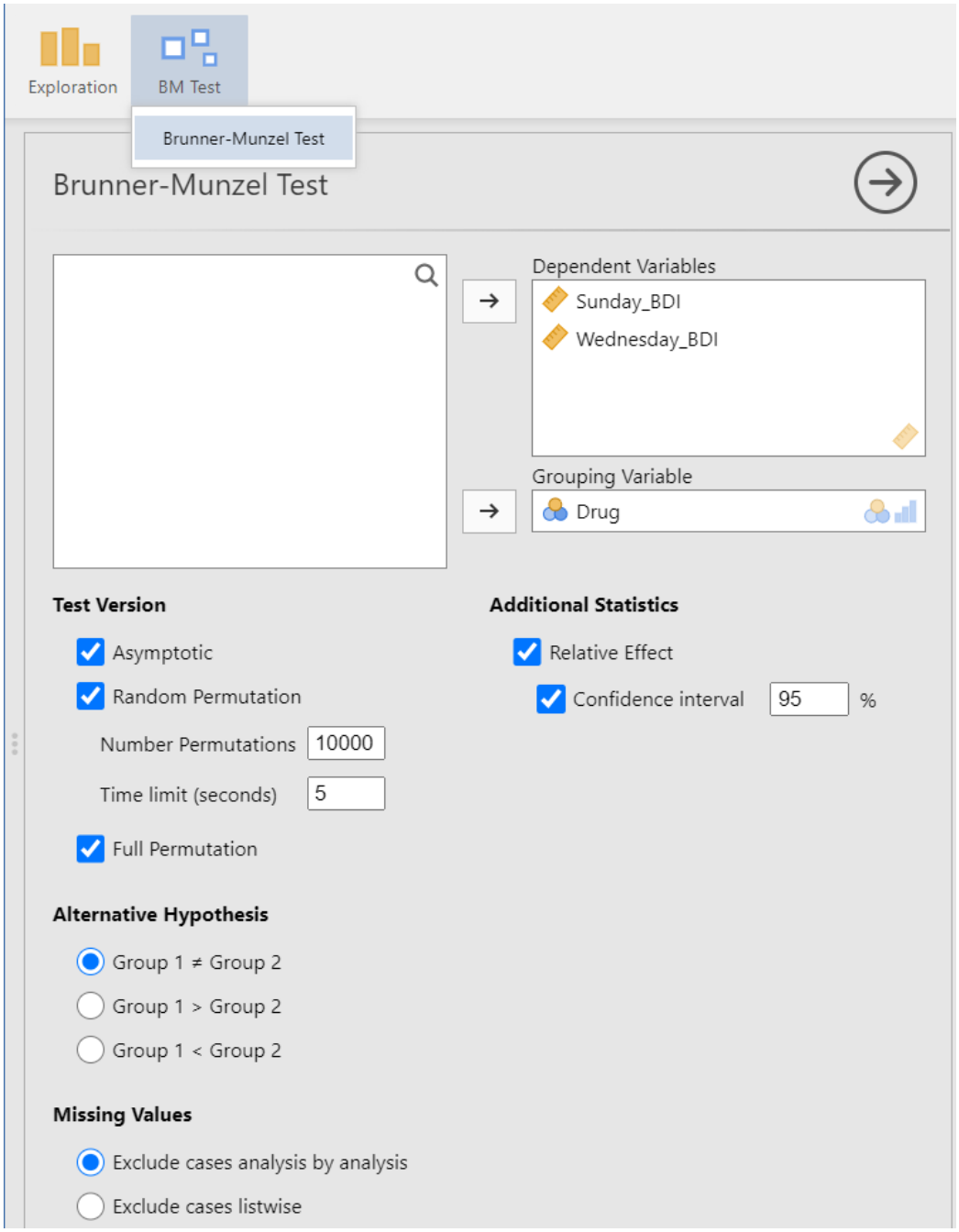

Figure 2. bmtest Main Menu.

After the dependent and grouping variables have been selected, the results are automatically computed and displayed. By default, only the asymptotic version of the test is computed, as it is by far the fastest. The other two versions can be requested by activating the respective checkboxes in the "Test Version" menu section. For the random permutation version, the number of permutations can be chosen and is, by default, set to the recommended value of 10,000. There is also a changeable time limit, as with large datasets, the random permutation test might take too long. The computation will be stopped, and no results will be displayed after this time limit has elapsed. Typically, it takes longer for the program to stop than the specified time limit because the underlying R code has to reach a state where it can interrupt the computation.

The full permutation version is only computed if it is computationally feasible (currently, if the number of permutations does not exceed 40,116,600, which for equal group

sizes amounts to $n_1, n_2 \leq 14$). If not, the respective row in the results table is left empty, and an error message is displayed (see Figure 3).

Brunner-Munzel Test

Brunner-Munzel Test

						95% Confidence Interval	
		Statistic	df	p	P̂(Ecs < Alc) + ½P̂(Ecs = Alc)	Lower	Upper
Sunday_BDI	Asymptotic	-0.111	54.4	0.912	0.492	0.341	0.643
	Random Permutation	-0.111		0.904	0.492	0.337	0.643
	Full Permutation	NaN [a]					
Wednesday_BDI	Asymptotic	-4.068	42.7	< .001	0.238	0.109	0.368
	Random Permutation	-4.068		< .001	0.238	0.111	0.370
	Full Permutation	NaN [a]					

Note. H_a P̂(Ecs < Alc) + ½P̂(Ecs = Alc) ≠ ½

[a] Number of needed permutations too large to be computationally feasible. Use one of the other two options.

Figure 3. Results Table.

In the "Hypothesis" menu section, the user can specify whether a one- or two-sided test should be performed. This only influences the obtained p-value. The confidence intervals reported are always two-sided. This must be crucially taken into account when interpreting confidence intervals. This choice was made to be consistent with most BM test R packages, which also always report two-sided confidence intervals.

The "Missing Value" section allows specifying how missing values should be treated. When selecting "Exclude cases analysis by analysis," missing values are excluded separately for each dependent variable. Otherwise, each row for which any of the involved variables is missing (all dependent variables and the grouping variable) is removed. I recommend choosing "Exclude cases analysis by analysis" as the default option, which is also the default setting.

Activating the checkboxes for the relative effect and the confidence interval adds the sample relative effect $\hat{p}$, as well as a confidence interval for the true relative effect p to the results. The desired confidence level can also be specified. For the full permutation version, confidence intervals are not displayed, as they are currently not supported.

3.3. Interpreting and Reporting Results

The results are displayed in a dynamic table to the right of the menu, as is the default in jamovi (see Figure 3). This table is automatically updated as the settings are changed. The first two unnamed columns contain information about the dependent variables and the test version, with the results for each combination of a dependent variable and test version displayed in one row. The "Statistic" column contains the test statistic BM and is thus the same for all test versions. The "df" column contains the degrees of freedom for the asymptotic version and is thus empty for the permutation versions. The "p" column contains the p-value. The next column contains the estimated relative effect $\hat{p}$ and is thus the same for all versions. The column name informs the reader which group is treated as the first group (X_1). In the example, this is the Ecstasy group. The confidence interval columns contain corresponding confidence intervals.

The results can be interpreted and reported as follows. First, for the day after substance consumption (Sunday_BDI), the null hypothesis of stochastic comparability could not be rejected, as the p-value was higher than the significance level of 5% and the 95% confidence interval contained a relative effect of $p = 0.5$. Considering the results from the, in this case, recommended random permutation version, this can be reported as follows: The

day after the drugs were taken, the data were in line with depression levels in ecstasy users and alcohol users being comparable, $BM = -0.11$, $p = 0.904$. Splitting ties equally, the probability that a random ecstasy user was less depressed than a random alcohol user was $\hat{p} = 0.49$, 95% CI $[0.34, 0.64]$.

The results for Wednesday (Wednesday_BDI) suggest that the ecstasy group tended to greater values; that is, ecstasy consumers tended to feel more depressed than alcohol drinkers. This can be reported as follows: By Wednesday, ecstasy users tended to feel more depressed than alcohol users, $BM = -4.07$, $p < 0.001$. Splitting ties equally, the probability that a random ecstasy user was less depressed than a random alcohol user was $\hat{p} = 0.24$, 95% CI $[0.11, 0.37]$.

4. bmtest R Package

At the time of writing, the package is not yet available via CRAN. The following code installs the package:

```
remotes::install_github("karchjd/bmtest")
```

The package contains only the `bmtest` function. The arguments of the function correspond to the options in the BM test menu just discussed. The analysis conducted in the previous section can be performed as follows:

```
library(bmtest)
drug_data <- haven::read_sav("Drug.sav")
drug_data$Drug <- forcats::as_factor(drug_data$Drug)
bmtest(data = drug_data, vars = c("Sunday_BDI", "Wednesday_BDI"),
       group = "Drug", asym = TRUE, randomPerm = TRUE, fullPerm = TRUE,
       miss = "perAnalysis",  relEff = TRUE, ci = TRUE, ciWidth = 95)
```

This returns the table shown in Figure 3 as output. The help file (`?bmtest`) provides more detailed guidance.

5. Concluding Remarks

This paper presented the jamovi module and R package bmtest, which implements all versions of the Brunner–Munzel (BM) test. The BM test addresses the drawbacks of the Wilcoxon–Mann–Whitney (WMW) test in a manner similar to how Welch's t-test improves upon Student's t-test. The bmtest jamovi package makes the BM test available in user-friendly GUI-based statistical software, thereby making it accessible to a broad audience of researchers.

Funding: This research received no external funding.

Data Availability Statement: The data presented in this study are openly available in https://osf.io/nug79/.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Schlag, K.H. Who gives direction to statistical testing? Best practice meets mathematically correct tests. *SSRN Electron. J.* **2015**. [CrossRef]
2. Karch, J.D. Psychologists should use Brunner-Munzel's instead of Mann-Whitney's U test as the default nonparametric procedure. *Adv. Methods Pract. Psychol. Sci.* **2021**, *4*. [CrossRef]
3. Divine, G.W.; Norton, H.J.; Barón, A.E.; Juarez-Colunga, E. The Wilcoxon-Mann-Whitney procedure fails as a test of medians. *Am. Stat.* **2018**, *72*, 278–286. [CrossRef]
4. Howell, D.C. *Statistical Methods for Psychology*; Cengage Learning: Boston, MA, USA, 2012.
5. Fay, M.P.; Proschan, M.A. Wilcoxon-Mann-Whitney or t-test? On assumptions for hypothesis tests and multiple interpretations of decision rules. *Stat. Surv.* **2010**, *4*, 1–39. [CrossRef] [PubMed]

6. Brunner, E.; Bathke, A.C.; Konietschke, F. *Rank and Pseudo-Rank Procedures for Independent Observations in Factorial Designs: Using R and SAS*; Springer International Publishing: Berlin/Heidelberg, Germany, 2019. [CrossRef]
7. Wilcox, R.R. *Modern Statistics for the Social and Behavioral Sciences*; CRC Press: Boca Raton, FL, USA, 2017.
8. Delacre, M.; Lakens, D.; Leys, C. Why psychologists should by default use Welch's *t*-test Instead of Student's *t*-test. *Int. Rev. Soc. Psychol.* **2017**, *30*, 92. [CrossRef]
9. Noguchi, K.; Konietschke, F.; Marmolejo-Ramos, F.; Pauly, M. Permutation tests are robust and powerful at 0.5% and 5% significance levels. *Behav. Res. Methods* **2021**, *53*, 2712–2724. [CrossRef] [PubMed]
10. Gastwirth, J.L.; Gel, Y.R.; Hui, W.L.W.; Lyubchich, V.; Miao, W.; Noguchi, K. Lawstat: Tools for Biostatistics, Public Policy, and Law. 2022. Available online: https://CRAN.R-project.org/package=lawstat (accessed on 15 March 2023)
11. Ara, T. Brunnermunzel: (Permuted) Brunner-Munzel Test. 2022. Available online: https://CRAN.R-project.org/package=brunnermunzel (accessed on 15 March 2023)
12. Konietschke, F.; Placzek, M.; Schaarschmidt, F.; Hothorn, L.A. nparcomp: An R software package for nonparametric multiple comparisons and simultaneous confidence intervals. *J. Stat. Softw.* **2015**, *64*, 1–17. [CrossRef]
13. The Jamovi Project. *Jamovi [Computer Software]*; Version 2.3; 2022. Available online: https://www.jamovi.org (accessed on 15 March 2023)
14. Field, A. *Discovering Statistics Using IBM SPSS Statistics*, 5th ed.; Sage: London, UK, 2017.
15. Brunner, E.; Munzel, U. The nonparametric Behrens-Fisher problem: Asymptotic theory and a small-sample approximation. *Biom. J.* **2000**, *42*, 17–25. [CrossRef]
16. Good, P. *Permutation, Parametric and Bootstrap Tests of Hypotheses*, 3rd ed.; Springer: New York, NY, USA, 2005.
17. Neubert, K.; Brunner, E. A studentized permutation test for the non-parametric Behrens-Fisher problem. *Comput. Stat. Data Anal.* **2007**, *51*, 5192–5204. [CrossRef]
18. Pauly, M.; Asendorf, T.; Konietschke, F. Permutation-based inference for the AUC: A unified approach for continuous and discontinuous data. *Biom. J.* **2016**, *58*, 1319–1337. [CrossRef]