Statistics Corner: Wilcoxon–Mann–Whitney Test

Kamal Kishore[1], Vidushi Jaswal[2]

Received on: 03 November 2022; *Accepted on:* 15 November 2022; *Published on:* 31 December 2022

Abstract

Wilcoxon–Mann–Whitney (WMW) test is a nonparametric counterpart of the *t*-test for comparing two unpaired groups. Traditional teaching and many books recommend applying WMW when: (1) continuous outcome variables violate assumptions and (2) data are ordinal. Standard recommendations about the applicability of WMW are not correct. Many health researchers also believe that WMW compares medians between groups; the reporting measure, however, is contextual—it depends on factors such as distribution type, sample size, and heteroscedasticity. A researcher comparing outcomes from two groups found that continuous dependent variables (DVs) do not fulfill the normality and homogeneity of variance assumptions. An initial literature search indicates that nonparametric methods are better for analyzing data. There are, however, a few vital questions concerning analyzing data with WMW:

- Does the test make any assumptions?
- What it compares—median or mean rank?
- What is the null and alternate hypothesis?
- What to report and how to interpret results?

Keywords: Brunner–Munzel test, Data interpretation, Mann–Whitney test, Nonparametric test, Two unpaired groups, Wilcoxon rank sum test.

Journal of Postgraduate Medicine, Education and Research (2022): 10.5005/jp-journals-10028-1613

Introduction

Researchers frequently compared two groups using a *t*-test; it is the first choice when the DV is continuous. The *t*-test compares the equality of means between groups and makes certain assumptions about the data distribution. For more details, the readers can see the previous article comparing two unpaired groups in the series.[1] There are many situations, such as when data are continuous but violate parametric test assumptions or data are categorical—the *t*-test is not applicable. The nonparametric alternative of *t*-tests is WMW. Usually, both *t*-test and WMW give similar results, but what if significance tests show different *p*-values and subsequent conclusions?

The use of nonparametric tests (NPTs) such as WMW is significantly lower than the parametric tests. There can be multiple reasons. First, parametric tests are more powerful, given that they fulfill the assumptions. Second, medical researchers are frequently interested in comparing the mean or median between groups. The *t*-test compares the average score, but the same is less clear about WMW; the interpretation of the outcome from WMW is contextual as compared to the *t*-test. Many software does not give confidence intervals for WMW. Third, NPTs are less emphasized and covered late in the semester. The NPTs are, therefore, less well-understood, applied, interpreted, and reported.

This manuscript will extend the discussion to analyze, report, and interpret study findings from two independent groups started in the previous article.[1] The current article will discuss (1) the problem statement, (2) the NPT, (3) the WMW test, and (4) the interpretation and reporting of study findings. We will begin by framing a research question. All data analysis was conducted using R commander (Rcmdr)—a graphical user interface for free, open-source, and command-driven R software.[2]

Problem Statement

Postgraduate Institute of Medical Education and Research (PGIMER) is a tertiary care institute; the mandate of PGIMER is to undertake intensive research in patient care.[3] Statistical analysis training is an integral part of students' academic learning. The literature has shown increasing statistical anxiety among medical students.[4,5] The researcher selected a reliable and valid SATS-36 questionnaire to capture statistical anxiety. Higher SATS indicate high anxiety—a low score is better. The survey's objective is: is there any significant difference between the statistics anxiety of the two groups? To further clarify, the "research staff and students" are the two independent groups and statistics anxiety is the outcome variable in the study.

Disclaimer

For demonstration, Excel® was used to generate the data for the analysis. However, the SATS-36 questionnaire genuinely assesses statistics anxiety.[5]

Research Question

Is statistics anxiety similar among research staff and students?

Null Hypothesis (H_0)

The population distribution of statistics anxiety is not significantly different between research staff and students. Most researchers,

[1]Department of Biostatistics, Postgraduate Institute of Medical Education and Research, Chandigarh, India

[2]Department of Psychology, Mehr Chand Mahajan DAV College for Women, Chandigarh, India

Corresponding Author: Kamal Kishore, Department of Biostatistics, Postgraduate Institute of Medical Education and Research, Chandigarh, India, Phone: +91 9591349768, e-mail: kkishore.pgi@gmail.com

How to cite this article: Kishore K, Jaswal V. Statistics Corner: Wilcoxon–Mann–Whitney Test. J Postgrad Med Edu Res 2022;56(4):199–201.

Source of support: Nil

Conflict of interest: None

© The Author(s). 2022 Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (https://creativecommons.org/licenses/by-nc/4.0/), which permits unrestricted use, distribution, and non-commercial reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.

however, are interested in comparing the median rather than population distribution. The WMW compares medians between group under specific situations—not in general.

Alternative Hypothesis (H_1)

The population distribution of statistics anxiety differs significantly between research staff and students.

Sample Size

Studies have demonstrated that the *t*-test and Welch's *t*-test perform significantly better than WMW for larger sample sizes.[6] It is, therefore, ideal for applying WMW for a sample size of 30 or less.[7]

Assumptions

In contrast to popular belief, WMW makes certain assumptions. The assumptions, however, are not about a parameter of interest.

- The participants are randomly allocated into groups.
- The DV is measured at least on the ordinal score.
- The DVs of two groups have similar distribution (normal or any other distribution)—a crucial assumption for applicability of WMW.

Nonparametric Test—A Brief Overview

Nonparametric tests do not make assumptions about the parameter of populations. The name, however, is confusing as NPT makes assumptions—it makes comparatively fewer assumptions than parametric tests. The significant advantage of NPT is that it applies to both categorical and continuous data. The NPTs are, however, relatively less powerful than the parametric tests for quantitative data except when the sample size is small, unequal, and violates the assumption of homogeneity; the *t*-test is robust to violation of the normality assumption. The major challenge while applying NPT is to formulate an appropriate hypothesis test. The confusion also stems from the type of outcome variables—nominal (binary vs multiple), ordinal, and continuous. The hypothesis, nature of tests, and interpretation depend on the nature of the outcome variable. Further, interpretation, type of test, and reporting vary as per assumptions. We will demonstrate the importance of context in the WMW test.

WMW Test

Wilcoxon (1945) and Mann and Whitney (1947) independently developed the test; to honor the authors, it is named WMW. Many names for WMW in literature are Mann–Whitney *U* test, Mann–Whitney–Wilcoxon, and Wilcoxon rank sum test. The test is applicable when either outcome variable is ordinal or continuous but does not meet the parametric test assumptions. The conventional understanding is that WMW compares the median between two groups; this, however, is not true in general—it tests whether two independent samples have the same distribution. WMW can give significant results even when the median between the groups is significantly different. The statistically significant result may offer the researcher a false hope that there is a difference (false positive) between intervention groups.

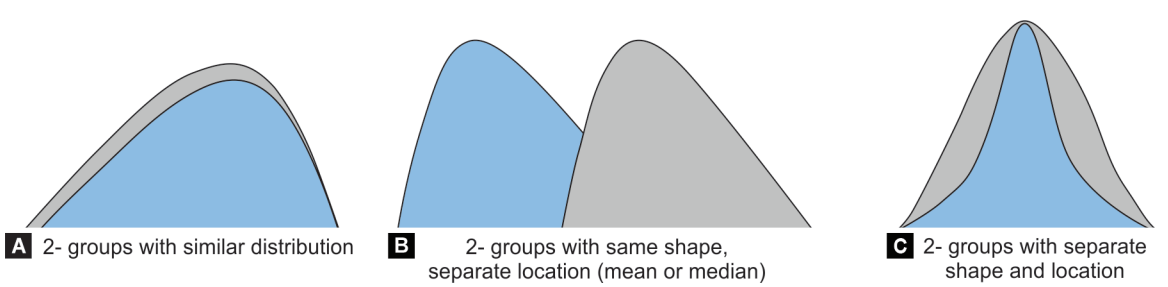
Unlike the *t*-test, interpretations from WMW are contextual or assumption based. When two groups have similar distributions (same shape and variance: pure shift model), such as in Figures 1A and B, WMW compares whether the median is significantly different between the two groups. When each group has a distinct shape, such as in Figure 1C, WMW compares population distribution—report mean rank. It is, therefore, recommended first to check the distribution of DV in both groups for accurate reporting of results from the WMW test. Many simulation studies have discussed the limitations of WMW for unequal variance and tied data—use Brunner–Munzel, an extension of the WMW test.[8]

WMW Test in Rcmdr

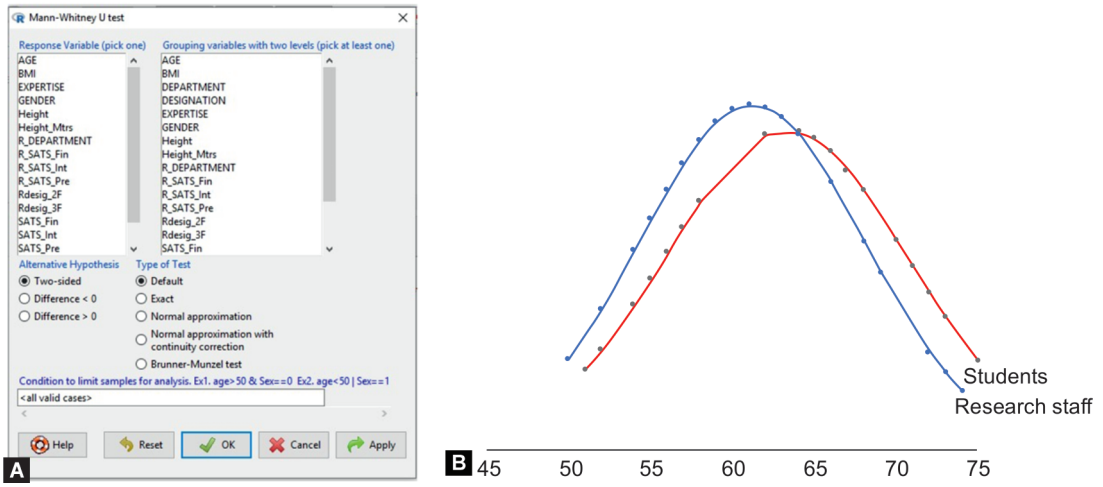
We assume that most researchers use Microsoft Excel® routinely to capture, clean, and code the data. Interested readers can read the previous article entitled "data cleaning and importing data" in the series.[9] There are five subtypes of WMW tests in Rcmdr—the test is available as Mann–Whitney *U* test in R. Readers can access the option using the menu "Statistical analysis<Nonparametric tests<Mann-Whitney *U* test." After two clicks, the variables list, as shown in Figure 2A, will pop up in front of the user. A researcher can select one-response (DV) and grouping variable (independent variable) to apply to the test. The challenge, however, is choosing the appropriate test, as Rcmdr gives five variations of the WMW test. When the sample size is small and ties are present—select the exact test; use WMW for equal variance, and the sample size is 20 or more. Use the Brunner–Munzel test for >20 sample sizes, the presence of ties and unequal variance. By default, Rcmdr reports boxplot; therefore, we plotted the data using Microsoft Excel® to gauge the distribution. Figure 2B displays the data distribution from both groups—the shape is identical. Therefore, we applied "normal approximation with continuity correction."

Reporting and Interpretation

For SATS-36 data, Figure 2B shows that the two groups are identically distributed. Researchers can frame the null hypothesis as the median SATS score is not significantly different between students and research staff—use median for reporting and interpreting results. The median SATS score (65) of students is not significantly different ($p = 0.3$) from the median SATS score (60) of research staff. The scenario where identical distribution assumptions does not meet—frame null hypothesis about population distribution and report mean rank while reporting results.



Figs 1A to C: Display examples of the population distribution of data for reporting from WMW test. (A) Two groups with a similar distribution; (B) Two groups with the same shape and separate location (mean or median); (C) Two groups with separate shape and location



Figs 2A and B: Mann–Whitney *U* test and SATS score distribution. (A) Default window display in Rcmdr to run Mann–Whitney *U* test; (B) SATS score between students and research staff groups are identically distributed* (*Plotted using Microsoft Excel®, Rcmdr gives boxplot)

Conclusion

The conventional wisdom that WMW is a test of the median is wrong and flawed—its interpretation is contextual. The other assumption is that WMW being nonparametric, does not make assumptions—it makes crucial assumptions about distribution and sample size. The researchers must cross-check data distribution and sample size to select the correct statistical analysis, reporting, and interpretation. WMW compares the distribution of population—mean rank is a better indicator than the median. The Brunner–Munzel test performs better than the routine WMW test when data contain tied observations and two heterogeneous groups. The nonparametric statistical tests, however, are not routinely taught in undergraduation and postgraduation classes. The wide availability of laptops, software, and statistical tests should motivate the faculty to teach and emphasize the NPTs for the reliability and validity of study findings.

Acknowledgment

We acknowledge Dr Anuj Kumar Pandey, Senior Research Officer from International Institute of Health Management Research, New Delhi, for his valuable time and input in improving the manuscript's readability.

Orcid

Kamal Kishore https://orcid.org/0000-0001-8936-0843
Vidushi Jaswal https://orcid.org/0000-0002-8914-4283

References

1. Kishore K, Jaswal V. Statistics corner: comparing two unpaired groups. J Postgrad Med Educ Res 2022;56(3):145–148. DOI: 10.5005/jp-journals-10028-1594
2. Fox J. Using the R Commander: A Point-and-click Interface for R. Chapman and Hall/CRC; 2016.
3. PGIMER, Chandigarh [Internet]. [cited 2022 Jun 13]. Available from: https://pgimer.edu.in/PGIMER_PORTAL/PGIMERPORTAL/home.jsp#
4. Zhang Y, Shang L, Wang R, et al. Attitudes toward statistics in medical postgraduates: measuring, evaluating and monitoring. BMC Med Educ 2012;12(1):117. DOI: 10.1186/1472-6920-12-117
5. Hannigan A, Hegarty AC, McGrath D. Attitudes towards statistics of graduate entry medical students: the role of prior learning experiences. BMC Med Educ 2014;14(1):70. DOI: 10.1186/1472-6920-14-70
6. Fagerland MW. t-tests, non-parametric tests, and large studies—a paradox of statistical practice? BMC Med Res Methodol 2012;12(1):78. DOI: 10.1186/1471-2288-12-78
7. Sedgwick P. Parametric v non-parametric statistical tests. BMJ 2012;344(2):e1753. DOI: 10.1136/bmj.e1753
8. Divine GW, Norton HJ, Barón AE, et al. The Wilcoxon–Mann–Whitney procedure fails as a test of medians. Am Stat 2018;72(3):278–286. DOI: 10.1080/00031305.2017.1305291
9. Kishore K, Jaswal V. Statistics corner: data management in Rcmdr. J Postgrad Med Edu Res 2022;56(2):102–105. DOI: 10.5005/jp-journals-10028-1571