

How to Perform a Mann-Whitney U Test to Compare Two Groups

Authored by
stats writer

January 21, 2026

RECOMMENDED CITATION

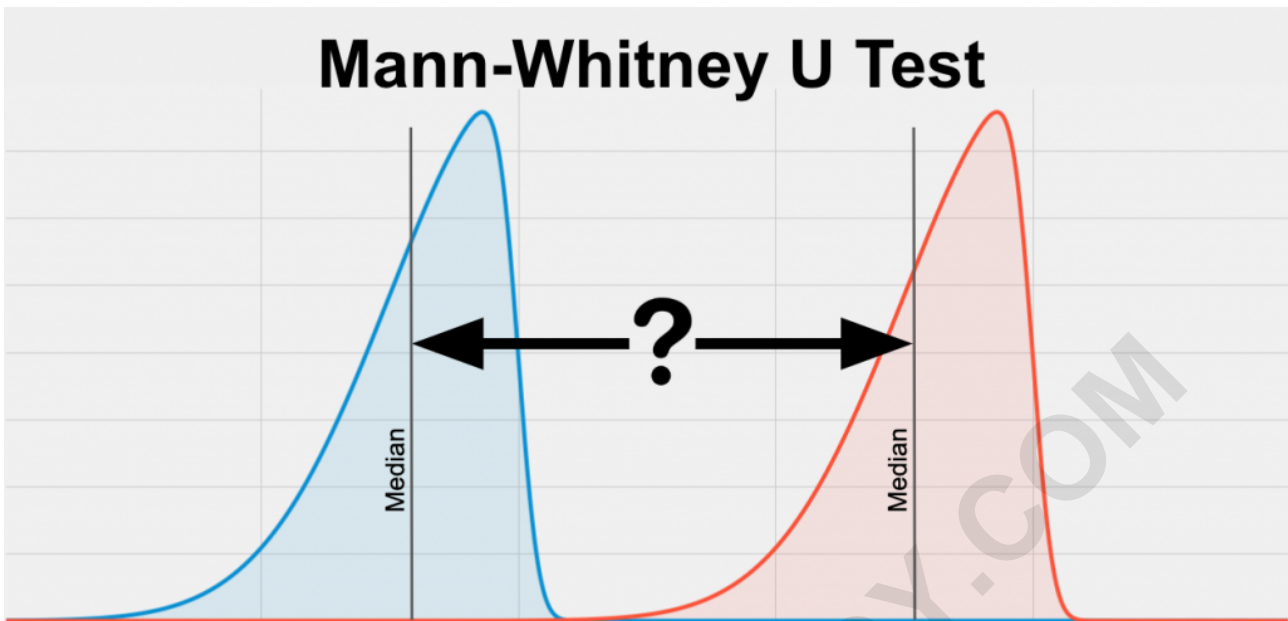
stats writer (2026). *How to Perform a Mann-Whitney U Test to Compare Two Groups*. PSYCHOLOGICAL SCALES. Retrieved from <https://scales.arabpsychology.com/?p=126968>

The Mann-Whitney U Test is a powerful statistical method used to compare two independent groups or samples. It is fundamentally a non-parametric test, meaning it makes no restrictive assumptions regarding the underlying population distribution of the data. This statistical flexibility makes it the preferred choice when data are not normally distributed or when sample sizes are small. The test mechanism involves assigning ranks to the combined data from both groups and subsequently calculating a U statistic, which quantifies the difference between the sum of the ranks of the two samples. By evaluating whether this measured difference is statistically significant, the Mann-Whitney U Test helps researchers determine if the two source populations are genuinely dissimilar in location, informing critical decisions across various research domains.

Defining the Mann-Whitney U Test

The **Mann-Whitney U Test**, also widely known as the Wilcoxon Rank-Sum Test, serves as a crucial inferential statistical procedure designed to ascertain whether two independent groups exhibit a statistically significant difference concerning a variable of interest. This technique is specifically engineered for scenarios where traditional parametric tests, like the t-test, are inappropriate due to distribution characteristics. For the test to be valid, the variable under examination must be a continuous variable, and while the exact values of the variable are used in the ranking process, the interpretation focuses on the medians or general location shifts between the two groups.

A fundamental requirement for employing the Mann-Whitney U Test is that the two comparison groups must be **independent**--meaning the observations in one group are entirely unrelated to the observations in the other. Furthermore, while the test is robust, sufficient data is still necessary for reliable results. It is generally recommended to have more than five observations in each group, although the precise minimum sample size is heavily influenced by the anticipated magnitude of the difference (effect size) between the groups being compared.



The *Mann-Whitney U Test* is also known by several alternative names, including the **Mann-Whitney Wilcoxon Test**, the **Wilcoxon Rank-Sum Test**, or the **Wilcoxon Mann-Whitney Test**.

Key Assumptions for the Mann-Whitney U Test

Every statistical procedure relies on a set of underlying assumptions that must be met for the results generated to be accurate, unbiased, and statistically sound. When these assumptions are violated, the conclusions drawn from the test may be misleading or incorrect. Understanding and verifying the prerequisites for the *Mann-Whitney U Test* ensures that the findings are reliable representations of population differences.

The primary assumptions required for the proper application and interpretation of this statistical test are:

The dependent variable must be **Continuous** (Interval or Ratio level data).

The underlying data distribution is often **Skewed** (Non-Normal).

The samples used must be selected through a **Random Sample** process.

There must be **Enough Data** (Sufficient Sample Size).

The distributions of the two groups should have a **Similar Shape Between Groups** for median comparison.

We will now delve into a detailed explanation of each of these critical requirements to ensure proper implementation of this non-parametric test.

Requirement 1: Continuous Data

The variable that constitutes the outcome measure (the variable of interest upon which the two groups are being compared) must be a continuous variable. This means the variable must be capable of taking on any value within a reasonable range, often measured on an interval or ratio scale, thereby allowing for fine distinctions and ranking between observations.

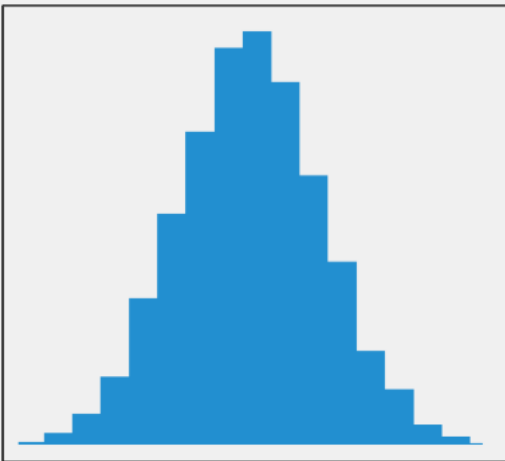
Excellent illustrations of continuous variables include metrics such as age measured in years, weight, height, standardized test scores, comprehensive survey scores (especially those resulting from averaging multiple items), or annual salary figures. These variables possess measurement units that can be infinitely subdivided, although in practice, measurement precision is limited.

*If your variable of interest is expressed as a proportion (e.g., comparing the percentage of males who voted against the percentage of females who voted), a more appropriate alternative is the **Two Proportion Z-Test**.*

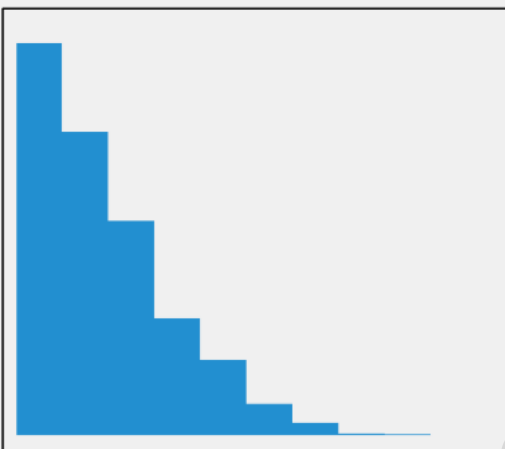
Requirement 2: Dealing with Skewed Distribution

A primary advantage of the Mann-Whitney U Test is its tolerance for data that does not adhere to a Gaussian, or bell-shaped, distribution--a condition known as normal distribution in statistics. Researchers are specifically encouraged to utilize the Mann-Whitney U Test when the variable of interest is **skewed**, meaning the bulk of the data points are concentrated toward one end of the spectrum rather than clustered symmetrically around the central mean.

This non-parametric approach is robust against outliers and violations of normality, making it ideal for real-world data in fields like medicine, economics, and behavioral sciences where variables frequently exhibit asymmetry (e.g., income, reaction times, disease recovery periods).



A normal distribution.
It is bell shaped with most of the data in the middle



A skewed distribution.
It is leaning left or right with most of the data on the edge

Conversely, if your variable exhibits a clear normal distribution, the statistically more powerful test to use would be the **Independent Samples T-Test**.

Requirement 3: Ensuring Random Sample Selection

A critical prerequisite for drawing valid inferences about a larger population is that the data points comprising each group must be derived from a **simple random sample**. This implies that every individual in the population had an equal chance of being selected for the sample, and selection was performed independently for both comparison groups.

For instance, if a study aims to compare the effects of a diet intervention, researchers must randomly assign participants to either the treatment group or the control group. Failure to employ proper randomization introduces the risk of **bias**, where systematic errors in the sampling process lead to inaccurate or misleading statistical results. Random selection ensures that observed differences are attributable to the variable being studied rather than pre-existing differences between the groups.

If genuine randomization is compromised, the generalizability of any conclusions is severely restricted. Furthermore, if your study involves two measurements taken from the same group of subjects (e.g., pre-test/post-test scores), these are considered **paired samples**, necessitating the use of a **Paired Samples T-Test** or the **Wilcoxon Signed-Rank Test**, depending on normality.

Requirement 4: Determining Sufficient Sample Size

While the Mann-Whitney U Test is suitable for relatively small data sets, a sufficient sample size (N) in each group is essential to achieve adequate statistical power. Generally, a minimum of five data points per group is often cited as a loose guideline, though many practitioners argue for larger samples to ensure reliable rank calculations and limit the impact of extreme outliers.

The precise required sample size is fundamentally dependent on the expected **effect size**--the magnitude of the difference hypothesized between the groups. If researchers anticipate a large and obvious difference, a smaller sample size may suffice to detect significance. Conversely, detecting a subtle or small difference necessitates a significantly larger sample size to mitigate the risk of a Type II error (failing to detect a real effect).

Mann-Whitney U Test	
Effect Size	Sample Size Needed*
Small	824 (412 in each group)
Medium	134 (67 in each group)
Large	54 (27 in each group)

*sample size calculation was conducted in G*Power with a power of 0.80, critical value (alpha) of 0.05, and 0.20, 0.50, and 0.80 used as the effect size values for small, medium, and large Cohen's D effect sizes respectively

If your sample size is exceptionally large ($N > 30$), and if you happen to know the population parameters (mean and standard deviation) for a normally distributed variable of interest, you might consider running an **Independent Samples Z-Test** instead.

Requirement 5: Similar Distribution Shape

For the Mann-Whitney U Test to be interpreted as a test of differences in **medians** (or average location), it is vital that the overall distribution shapes of the two groups, when visualized as histograms or density plots, are approximately similar. If the shapes are consistent--even if skewed--a statistically significant result indicates a difference in the central tendency (median) between the two populations.

However, if the distribution shapes differ substantially (e.g., one group is highly skewed right and the other is symmetric), a significant U Test result indicates only that the populations differ in some way, potentially relating to spread (variability) rather than just the median. In scenarios where shapes are dissimilar, researchers must exercise caution in interpreting the result solely as a difference in average value.

Contextual Application of the Mann-Whitney U Test

The Mann-Whitney U Test is specifically tailored for a narrow, yet common, set of research circumstances. Utilizing this test appropriately requires a clear understanding of the nature of the data and the research question being posed. It is the definitive choice when the following five conditions are simultaneously satisfied:

The research goal is to identify if two groups are **significantly different** (a difference question) on the outcome variable.

The outcome variable of interest is **continuous** (measured on an interval or ratio scale).

The analysis involves precisely **two groups** for comparison.

The data collected consists of **independent samples** (unrelated observations).

The variable of interest follows a **skewed distribution** (non-normal).

A brief elaboration on each of these criteria will solidify your understanding of when this robust, non-parametric test should be deployed.

Focus on Differences, Not Relationships

The Mann-Whitney U Test is fundamentally designed to answer a **difference question**: is Group A statistically distinguishable from Group B based on the outcome variable? This distinguishes it from other forms of analysis, such as correlation, which investigates the strength of the linear relationship between two variables, or regression/prediction, which attempts to forecast one variable's value based on another.

Requirement for Continuous Measurement

As previously established, the outcome must be a continuous variable (4/5), capable of assuming a broad range of numerical values. Examples include physiological measures like heart rate or height, behavioral metrics like the duration of a task, or economic data such as quarterly revenue.

It is essential to differentiate continuous data from other data types that cannot use this test, such as ordered or ordinal data (e.g., finishing place in a race, Likert scale responses treated strictly as ordinal), categorical data (e.g., political affiliation, gender), or binary data (e.g., success/failure, presence/absence of a disease).

Limitation to Two Comparison Groups

The structure of the Mann-Whitney U Test limits its application strictly to comparing **two groups** on the variable of interest. It is a bivariate comparison tool, meaning it cannot handle experiments involving three or more independent conditions simultaneously.

*If your experimental design includes three or more independent groups, you must utilize alternative methods. If your variable is normally distributed, the appropriate test is a **One-Way ANOVA**. If the variable remains skewed, the non-parametric equivalent is the **Kruskal-Wallis One-Way ANOVA**. If you are comparing a single group against a known or hypothesized population value, you would use a **Single Sample T-Test** (if normal) or a **Single Sample Wilcoxon Signed-Rank Test** (if skewed).*

Necessity of Independent Samples

The condition of **independent samples** dictates that the participants or observations composing the two comparison groups must be entirely unrelated. For example, comparing the test scores of a randomly selected cohort of college freshmen to those of a randomly selected cohort of high school seniors constitutes independent sampling, as the individuals in one group have no relationship to those in the other.

*In contrast, if the same set of subjects provides data under two different conditions--such as measuring their anxiety levels before an exam and then again after the exam--this generates **paired data**. In such cases, the analysis should switch to a **Paired Samples T-Test** (for normal distribution) or the **Wilcoxon Signed-Rank Test** (for skewed data).*

Prioritizing Skewed Distributions

As a rank-based test, the Mann-Whitney U Test shines when the data strongly deviates from a normal distribution. Data normality, characterized by its symmetric, bell shape, can be formally

assessed using specialized goodness-of-fit tests, such as the **Kolmogorov-Smirnov test** or the **Shapiro-Wilk test**. If these tests confirm non-normality, or if visual inspection clearly shows skewness, the Mann-Whitney U Test becomes the statistically defensible choice over parametric alternatives.

Illustrative Example of the Mann-Whitney U Test

To illustrate the practical application of the Mann-Whitney U Test, consider a clinical trial focused on evaluating a new medication intended to shorten recovery time from a specific illness. The study design involves two independent groups:

Group 1: The Treatment Group, which receives the experimental medical intervention.

Group 2: The Control Group, which receives an inert substance (placebo) or a standard control condition.

Variable of Interest: Time elapsed until full recovery from the disease, measured in days (a continuous variable).

The primary objective is to test the **null hypothesis**--the statistical premise that the experimental treatment has no effect. If the null hypothesis were true, we would expect Group 1 and Group 2 to exhibit approximately the same average recovery time. Conversely, the research hypothesis posits that receiving the experimental medical treatment will significantly reduce the number of days required for patients to fully recover.

Recovery time, particularly for diseases, is often a variable that is highly **skewed**; a few patients may recover very quickly while many others take much longer, thus failing the assumption of normal distribution. This inherent skewness confirms the appropriateness of using the Mann-Whitney U Test to compare the recovery times for both independent groups.

Upon collecting the data, the Mann-Whitney U Test is performed. The analysis yields two critical outputs: the **W-statistic** (related to the U statistic) and the **p-value**. The W-statistic quantifies the magnitude of the difference in ranks between the two samples. More importantly, the p-value represents the probability of observing a difference in recovery times as extreme as the one measured, assuming that the null hypothesis (i.e., the treatment does nothing) is true. If the resultant p-value is less than or equal to the predetermined significance level (typically 0.05), the result is deemed **statistically significant**, allowing the researcher to reject the null hypothesis and conclude that the difference observed is genuinely due to the treatment and not merely random chance.