## Chapter 503

## Mann-Whitney U or Wilcoxon Rank-Sum Tests (Guenther)

## Introduction

This procedure provides sample size and power calculations for one- or two-sided two-sample MannWhitney $U$ or Wilcoxon Rank-Sum Tests. This test is the nonparametric alternative to the traditional twosample t-test. Other names for this test are the Mann-Whitney-Wilcoxon test or the Wilcoxon-Mann-Whitney test.

The design corresponding to this test procedure is sometimes referred to as a parallel-groups design. This design is used in situations such as the comparison of the income level of two regions, the nitrogen content of two lakes, or the effectiveness of two drugs.
There are several statistical tests available for the comparison of the center of two populations. You can examine the sections below to identify whether the assumptions and test statistic you intend to use in your study match those of this procedure, or if one of the other PASS procedures may be more suited to your situation.

This procedure uses the method of Guenther (see Al-Sunduqchi and Guenther (1990)) for power calculations.

## Other PASS Procedures for Comparing Two Means or Medians

Procedures in PASS are primarily built upon the testing methods, test statistic, and test assumptions that will be used when the analysis of the data is performed. You should check to identify that the test procedure described below in the Test Procedure section matches your intended procedure. If your assumptions or testing method are different, you may wish to use one of the other two-sample procedures available in PASS. These procedures are Two-Sample T-Tests Assuming Equal Variance, Two-Sample T-Tests Allowing Unequal Variance, Two-Sample Z-Tests Assuming Equal Variance, and Two-Sample Z-Tests Allowing Unequal Variance. There is also a Mann-Whitney U or Wilcoxon Rank-Sum Tests (Simulation) procedure available. The methods, statistics, and assumptions for those procedures are described in the associated chapters.

If you wish to show that the mean of one population is larger (or smaller) than the mean of another population by a specified amount, you should use one of the clinical superiority procedures for comparing means. Non-inferiority, equivalence, and confidence interval procedures are also available.

## Test Assumptions

When running a Mann-Whitney-Wilcoxon test, the basic assumptions are random sampling from each of the two populations and that the measurement scale is at least ordinal. These assumptions are sufficient for testing whether the two populations are different. If we can additionally assume that the two populations are identical except possible for a difference in location, then this test can be used as a test of equal means or medians.

## Test Procedure

If we assume that the two populations differ only in location, with $\mu_{1}$ and $\mu_{2}$ representing the means of the two populations of interest, and that $\delta=\mu_{1}-\mu_{2}$, the null hypothesis for comparing the two means (or medians) is $H_{0}: \delta=0$. The alternative hypothesis can be any one of

$$
\begin{aligned}
& H_{1}: \delta \neq 0 \\
& H_{1}: \delta>0 \\
& H_{1}: \delta<0
\end{aligned}
$$

depending upon the desire of the researcher or the protocol instructions. A suitable Type I error probability (a) is chosen for the test, the data is collected, and the data from both groups are combined and then ranked.

The Mann-Whitney test statistic is defined as follows in Gibbons (1985).

$$
z=\frac{W_{1}-\frac{N_{1}\left(N_{1}+N_{2}+1\right)+C}{2}}{s_{W}}
$$

where

$$
W_{1}=\sum_{k=1}^{N_{1}} \operatorname{Rank}\left(X_{1 k}\right)
$$

The ranks are determined after combining the two samples. The standard deviation is calculated as

$$
s_{W}=\sqrt{\frac{N_{1} N_{2}\left(N_{1}+N_{2}+1\right)}{12}-\frac{N_{1} N_{2} \sum_{i=1}\left(t_{i}^{3}-t_{i}\right)}{12\left(N_{1}+N_{2}\right)\left(N_{1}+N_{2}-1\right)}}
$$

where $t_{i}$ is the number of observations tied at value one, $t_{2}$ is the number of observations tied at some value two, and so forth.
The correction factor, $C$, is 0.5 if the rest of the numerator of $z$ is negative or -0.5 otherwise. The value of $z$ is then compared to the standard normal distribution.

The null hypothesis is rejected in favor of the alternative if, for $H_{1}: \delta \neq 0$,

$$
z<z_{\alpha / 2} \text { or } z>z_{1-\alpha / 2}
$$

for $H_{1}: \delta>0$,

$$
z>z_{1-\alpha},
$$

or, for $H_{1}: \delta<0$,

$$
z<z_{\alpha} .
$$

Comparing the $z$-statistic to the cut-off $z$-value (as shown here) is equivalent to comparing the $p$-value to $\square$.

## Power Calculation for Mann-Whitney U or Wilcoxon RankSum Tests

The power calculation for the Mann-Whitney U or Wilcoxon Rank-Sum Test is the same as that for the twosample equal-variance $t$-test except that an adjustment is made to the sample size based on an assumed data distribution as described in Al-Sunduqchi and Guenther (1990). For a Mann-Whitney U or Wilcoxon Rank-Sum Test group sample size of $n_{i}$, the adjusted sample size $n_{i}^{\prime}$ used in power calculations is equal to

$$
n_{i}^{\prime}=n_{i} / W,
$$

where $W$ is the Wilcoxon adjustment factor based on the assumed data distribution.
The adjustments are as follows:

| Distribution | $\boldsymbol{w}$ |
| :--- | :--- |
| Double Exponential | $2 / 3$ |
| Logistic | $9 / \pi^{2}$ |
| Normal | $\pi / 3$ |

This section describes the procedure for computing the power from $n_{1}^{\prime}$ and $n_{2}^{\prime}, \alpha$, the assumed $\mu_{1}$ and $\mu_{2}$, and the assumed common standard deviation, $\sigma_{1}=\sigma_{2}=\sigma$. Two good references for these methods are Julious (2010) and Chow, Shao, Wang, and Lokhnygina (2018).

If we call the assumed difference between the means $\delta=\mu_{1}-\mu_{2}$, the steps for calculating the power are as follows:

1. Find $t_{1-\alpha}$ based on the central- $t$ distribution with degrees of freedom,

$$
d f=n_{1}^{\prime}+n_{2}^{\prime}-2
$$

2. Calculate the non-centrality parameter:

$$
\lambda=\frac{\delta}{\sigma \sqrt{\frac{1}{n_{1}^{\prime}}+\frac{1}{n_{2}^{\prime}}}}
$$

3. Calculate the power as the probability that the test statistic $t$ is greater than $t_{1-\alpha}$ under the non-central- $t$ distribution with non-centrality parameter $\lambda$ :

$$
\text { Power }=\operatorname{Pr}_{\text {Non-central-t }}\left(t>t_{1-\alpha} \mid d f=n_{1}^{\prime}+n_{2}^{\prime}-2, \lambda\right)
$$

The algorithms for calculating power for the opposite direction and the two-sided hypotheses are analogous to this method.
When solving for something other than power, PASS uses this same power calculation formulation, but performs a search to determine that parameter.

## A Note on Specifying the Means/Medians or Difference in Means/Medians

When means are specified in this procedure, they are used to determine the assumed difference in means for power or sample size calculations. When the difference in means is specified in this procedure, it is the assumed difference in means for power or sample size calculations. It does not mean that the study will be powered to show that the mean difference is this amount, but rather that the design is powered to reject the null hypothesis of equal means if this were the true difference in means. If your purpose is to show that one mean is greater than another by a specific amount, you should use one of the clinical superiority procedures for comparing means.

## A Note on Specifying the Standard Deviation

The sample size calculation for most statistical procedures is based on the choice of alpha, power, and an assumed difference in the primary parameters of interest - the difference in means in this procedure. An additional parameter that must be specified for means tests is the standard deviation. Here, we will briefly discuss some considerations for the choice of the standard deviation to enter.

If a number of previous studies of a similar nature are available, you can estimate the variance based on a weighted average of the variances, and then take the square root to give the projected standard deviation.

Perhaps more commonly, only a single pilot study is available, or it may be that no previous study is available. For both of these cases, the conservative approach is typically recommended. In PASS, there is a standard deviation estimator tool. This tool can be used to help select an appropriate value or range of values for the standard deviation.

If the standard deviation is not given directly from the previous study, it may be obtained from the standard error, percentiles, or the coefficient of variation. Once a standard deviation estimate is obtained, it may be useful to then use the confidence limits tab to obtain a confidence interval for the standard deviation estimate. With regard to power and sample size, the upper confidence limit will then be a conservative estimate of the standard deviation. Or a range of values from the lower confidence limit to the upper confidence limit may be used to determine the effect of the standard deviation on the power or sample size requirement.
If there is no previous study available, a couple of rough estimation options can be considered. You may use the data tab of the standard deviation estimator to enter some values that represent typical values you expect to encounter. This tool will allow you to see the corresponding population or sample standard deviation. A second rough estimation technique is to base the estimate of the standard deviation on your estimate of the range of the population or the range of a data sample. A conservative divisor for the population range is 4 . For example, if you are confident your population values range from 45 to 105, you would enter 60 for the Population Range, and, say, 4, for ' C '. The resulting standard deviation estimate would be 15 .

If you are unsure about the value you should enter for the standard deviation, we recommend that you additionally examine a range of standard deviation values to see the effect that your choice has on power or sample size.

## Example 1 - Finding the Sample Size

Researchers wish to compare two types of local anesthesia to determine whether there is a difference in time to loss of pain. Subjects will be randomized to treatment, the treatment will be administered, and the time to loss of pain measured. The researchers would like to generate a sample size for the study with $90 \%$ power to reject the null hypothesis of equal loss-of-pain time if the true difference is at least 3 minutes. How many participants are needed to achieve $90 \%$ power at significance levels of 0.01 and 0.05 ?

Past experiments of this type have had standard deviations in the range of 1 to 5 minutes. It is anticipated that the standard deviation of the two groups will be equal.

It is unknown which treatment has lower time to loss of pain, so a two-sided test will be used.
The researchers will be performing a Mann-Whitney-Wilcoxon test instead of the $t$-test because it is anticipated that the distribution of the two populations is not Normal. The researchers assume that the Logistic distribution shape most closely resembles what they expect to observe from the data.

## Setup

If the procedure window is not already open, use the PASS Home window to open it. The parameters for this example are listed below and are stored in the Example 1 settings file. To load these settings to the procedure window, click Open Example Settings File in the Help Center or File menu.

| Design Tab |  |
| :---: | :---: |
| Solve For ...............................................Sample Size |  |
| Alternative Hypothesis | Two-Sided |
| Data Distribution ...... | Logistic |
| Power. | . 0.90 |
| Alpha. | . 0.010 .05 |
| Group Allocation | .Equal ( $\mathrm{N} 1=\mathrm{N} 2$ ) |
| Input Type. | Difference |
| ठ... |  |
| б ........ | . 1 to 5 by 1 |

## Output

Click the Calculate button to perform the calculations and generate the following output.

## Numeric Reports

## Numeric Results

| Solve For: | Sample Size |
| :--- | :--- |
| Test Type: | Two-Sample Mann-Whitney U or Wilcoxon Rank-Sum Test |
| Difference: | $\delta=\mu 1-\mu 2$ |
| Hypotheses: | H0: $\delta=0$ vs. H1: $\delta \neq 0$ |
| Data Distribution: | Logistic |


| Power |  | Sample Size |  |  | Mean Difference ठ | Standard Deviation | Alpha |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| Target | Actual | N1 | N2 | N |  |  |  |
| 0.9 | 0.95643 | 6 | 6 | 12 | 3 | 1 | 0.01 |
| 0.9 | 0.93894 | 4 | 4 | 8 | 3 | 1 | 0.05 |
| 0.9 | 0.90052 | 14 | 14 | 28 | 3 | 2 | 0.01 |
| 0.9 | 0.91690 | 11 | 11 | 22 | 3 | 2 | 0.05 |
| 0.9 | 0.90596 | 30 | 30 | 60 | 3 | 3 | 0.01 |
| 0.9 | 0.91250 | 21 | 21 | 42 | 3 | 3 | 0.05 |
| 0.9 | 0.90260 | 51 | 51 | 102 | 3 | 4 | 0.01 |
| 0.9 | 0.90487 | 36 | 36 | 72 | 3 | 4 | 0.05 |
| 0.9 | 0.90268 | 78 | 78 | 156 | 3 | 5 | 0.01 |
| 0.9 | 0.90312 | 55 | 55 | 110 | 3 | 5 | 0.05 |


| Target Power | The desired power value (or values) entered in the procedure. Power is the probability of rejecting a false null hypothesis. |
| :---: | :---: |
| Actual Power | The power obtained in this scenario. Because N 1 and N 2 are discrete, this value is often (slightly) larger than the target power. |
| N1 and N2 | The number of items sampled from each population. |
| N | The total sample size. $\mathrm{N}=\mathrm{N} 1+\mathrm{N} 2$. |
| $\mu 1$ and $\mu 2$ | The assumed population means. |
| $\delta$ | The difference between population means at which power and sample size calculations are made. $\delta=\mu 1$ $\mu 2$. |
| $\sigma$ | The assumed population standard deviation used for both of the two groups. |
| Alpha | The probability of rejecting a true null hypothesis. |

## Summary Statements

A parallel two-group design will be used to test whether the Group 1 location (distribution center) is different from the Group 2 location ( $\mathrm{H} 0: \mu 1-\mu 2=0$ versus $\mathrm{H} 1: \mu 1-\mu 2 \neq 0$ ). The comparison will be made using a two-sided, two-sample Mann-Whitney U or Wilcoxon Rank-Sum test, with a Type I error rate ( $\alpha$ ) of 0.01 . The common standard deviation for both groups is assumed to be 1, and the underlying data distribution is assumed to be Logistic. To detect a difference in means of 3 with $90 \%$ power, the number of needed subjects will be 6 in Group 1 and 6 in Group 2.

Dropout-Inflated Sample Size

|  |  | Sample Size |
| :--- | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: |

## Dropout Summary Statements

Anticipating a $20 \%$ dropout rate, 8 subjects should be enrolled in Group 1, and 8 in Group 2, to obtain final group sample sizes of 6 and 6 , respectively.

## References

Al-Sunduqchi, Mahdi S. 1990. Determining the Appropriate Sample Size for Inferences Based on the Wilcoxon Statistics. Ph.D. dissertation under the direction of William C. Guenther, Dept. of Statistics, University of Wyoming, Laramie, Wyoming.
Chow, S.C., Shao, J., Wang, H., and Lokhnygina, Y. 2018. Sample Size Calculations in Clinical Research, Third Edition. Taylor \& Francis/CRC. Boca Raton, Florida.
Julious, S. A. 2010. Sample Sizes for Clinical Trials. Chapman \& Hall/CRC. Boca Raton, FL.
Machin, D., Campbell, M., Fayers, P., and Pinol, A. 1997. Sample Size Tables for Clinical Studies, 2nd Edition. Blackwell Science. Malden, MA.
Zar, Jerrold H. 1984. Biostatistical Analysis (Second Edition). Prentice-Hall. Englewood Cliffs, New Jersey.

These reports show the values of each of the parameters, one scenario per row.

## Plots Section

Plots


These plots show the relationship between the standard deviation and sample size for the two alpha levels.

## Example 2 - Comparing the Power to the T-Test with Normal Data

Suppose a new corn fertilizer is to be compared to a current fertilizer. The current fertilizer produces an average of about 74 lbs . per plot. The researchers need only show that there is difference (increase) in yield with the new fertilizer. With 90 plots available, they would like to examine the power of the test if the improvement in yield is at least 10 lbs .

Researchers plan to use a one-sided test with alpha equal to 0.05 . Previous studies indicate the standard deviation for plot yield to be 25 lbs . The distribution of plot yield values is unknown, so the researchers would like to see the loss in power if the distribution turns out to be Normal and the Mann-WhitneyWilcoxon test is used rather than the standard $t$-test. The power for this scenario with the standard t-test is 0.594 .

## Setup

If the procedure window is not already open, use the PASS Home window to open it. The parameters for this example are listed below and are stored in the Example $\mathbf{2}$ settings file. To load these settings to the procedure window, click Open Example Settings File in the Help Center or File menu.


## Output

Click the Calculate button to perform the calculations and generate the following output.

## Numeric Results

| Solve For: | Power |
| :--- | :--- |
| Test Type: | Two-Sample Mann-Whitney U or Wilcoxon Rank-Sum Test |
| Difference: | $\delta=\mu 1-\mu 2$ |
| Hypotheses: | H0: $\delta \leq 0$ vs. H1: $\delta>0$ |
| Data Distribution: | Normal |


| Power | Sample Size |  |  | Mean |  | Mean Difference б | Standard Deviation | Alpha |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | N1 | N2 | N | $\mu 1$ | $\mu 2$ |  |  |  |
| 0.56868 | 45 | 45 | 90 | 84 | 74 | 10 | 25 | 0.05 |

The power of the Mann-Whitney test in this scenario is 0.56868 . This power is only slightly less than the power of the $t$-test ( 0.594 ) for the corresponding scenario.

## Example 3 - Validation using Chow, Shao, Wang, and Lokhnygina (2018)

Chow, Shao, Wang, and Lokhnygina (2018) presents an example on page 53 of a two-sided two-sample $t$-test sample size calculation for equal group sizes in which $\delta=0.05, \sigma=0.1$, alpha $=0.05$, and power $=0.80$. They obtain a sample size of 64 for each group.
The Mann-Whitney U or Wilcoxon Rank-Sum test power calculations are the same as the two-sample $t$-test except for an adjustment factor for the assumed data distribution. If we set the data distribution to Normal, we should get a result of $\mathrm{N} 1=\mathrm{N} 2=64 \times \pi / 3=67.021$, which rounds up to 68 .

## Setup

If the procedure window is not already open, use the PASS Home window to open it. The parameters for this example are listed below and are stored in the Example $\mathbf{3}$ settings file. To load these settings to the procedure window, click Open Example Settings File in the Help Center or File menu.

| Design Tab |  |
| :---: | :---: |
| Solve For .... | ..Sample Size |
| Alternative Hypothesis | ..Two-Sided |
| Power.. | . 0.80 |
| Alpha.. | . 0.05 |
| Group Allocation .... | Equal ( $\mathrm{N} 1=\mathrm{N} 2$ ) |
| Input Type. | Difference |
| б... | . 0.05 |
| O ...... | . 0.1 |

## Output

Click the Calculate button to perform the calculations and generate the following output.

## Numeric Results

| Solve F |  | Sample Size |  |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| Test Ty |  | Two-Sample Mann-Whitney U or Wilcoxon Rank-Sum Test |  |  |  |  |  |
| Differen |  | 1- $\mu$ |  |  |  |  |  |
| Hypoth | s: $\quad \mathrm{H}$ | $\mathrm{H} 0: \delta=0$ vs. $\mathrm{H} 1: \delta \neq 0$ |  |  |  |  |  |
| Data Distribution: Normal |  |  |  |  |  |  |  |
| Power |  | Sample Size |  |  | Mean | Standard |  |
| Target | Actual | N1 | N2 | N | ठ | $\boldsymbol{\sigma}$ | Alpha |
| 0.8 | 0.80146 | 68 | 68 | 136 | 0.1 | 0.1 | 0.05 |

The sample size of 68 in each group matches the expected result.

