Chapter 430

Mann-Whitney U or Wilcoxon Rank-Sum Tests (Simulation)

Introduction

This procedure provides sample size and power calculations for one- or two-sided two-sample Mann-Whitney U or Wilcoxon Rank-Sum Tests based on simulation. This test is the nonparametric alternative to the traditional two-sample t-test. Other names for this test are the Mann-Whitney-Wilcoxon test or the Wilcoxon-Mann-Whitney test. The user can also conduct non-inferiority, superiority by a margin, and non-zero null tests.

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.

Other PASS Procedures for Comparing Two Means

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 procedure based on analytic results 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 μ_1 and μ_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

$$H_1: \delta \neq 0$$

$$H_1: \delta > 0$$

$$H_1: \delta < 0$$

depending upon the desire of the researcher or the protocol instructions. A suitable Type I error probability (α) 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).

$$z = \frac{W_1 - \frac{N_1(N_1 + N_2 + 1)}{2} + C}{s_W}$$

where

$$W_1 = \sum_{k=1}^{N_1} Rank(X_{1k})$$

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

$$s_W = \sqrt{\frac{N_1 N_2 (N_1 + N_2 + 1)}{12} - \frac{N_1 N_2 \sum_{i=1}(t_i^3 - t_i)}{12(N_1 + N_2)(N_1 + N_2 - 1)}}$$

where t_1 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 $H_1: \delta \neq 0$,

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

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 α.

Power Calculation

Simulation

Simulation allows us to estimate the power and significance level that is actually achieved by a test procedure in situations that are not mathematically tractable.

The steps to a simulation study are

1. Specify how the test is carried out. This includes indicating how the test statistic is calculated and how the significance level is specified.
2. Generate random samples from the distributions specified by the alternative hypothesis. Calculate the test statistics from the simulated data and determine if the null hypothesis is accepted or rejected. Tabulate the number of rejections and use this to calculate the test's power.
3. Generate random samples from the distributions specified by the null hypothesis. Calculate each test statistic from the simulated data and determine if the null hypothesis is accepted or rejected. Tabulate the number of rejections and use this to calculate the test's significance level.
4. Repeat steps 2 and 3 several thousand times, tabulating the number of times the simulated data leads to a rejection of the null hypothesis. The power is the proportion of simulated samples in step 2 that lead to rejection. The significance level is the proportion of simulated samples in step 3 that lead to rejection.

Generating Random Distributions

Two methods are available in **PASS** to simulate random samples. The first method generates the random variates directly, one value at a time. The second method generates a large pool (over 10,000) of random values and then draw the random numbers from this pool. This second method can cut the running time of the simulation by 70%.

As mentioned above, the second method begins by generating a large pool of random numbers from the specified distributions. Each of these pools is evaluated to determine if its mean is within a small relative tolerance (0.0001) of the target mean. If the actual mean is not within the tolerance of the target mean, individual members of the population are replaced with new random numbers if the new random number moves the mean towards its target. Only a few hundred such swaps are required to bring the actual mean to within tolerance of the target mean. This population is then sampled with replacement using the uniform distribution. We have found that this method works well as long as the size of the pool is the maximum of twice the number of simulated samples desired and 10,000.

Standard Deviations

Care must be used when either the null or alternative distribution is not normal. In these cases, the standard deviation is usually not specified directly. For example, you might use a gamma distribution with a shape parameter of 1.5 and a mean of 4 as the null distribution and a gamma distribution with the same shape parameter and a mean of 5 as the alternative distribution. This allows you to compare the two means. However, although the shape parameters are constant, the standard deviations, which are based on both the shape parameter and the mean, are not.

A Note on Specifying the Difference in Means/Medians

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.

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 distribution shape that most closely resembles the two populations is Tukey's Lambda distribution with Skewness value 0.12 and Elongation 0.07 (Tukey's Lambda distribution adjusts the Normal distribution by a specified Skewness factor and Kurtosis factor – you can examine the distribution with the Data Simulator tool, in the Tools menu).

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 (H1: μ1 - μ2 ≠ δ0)**
Simulations	**2000**
Random Seed	**3915264** (for Reproducibility)
Power	**0.90**
Alpha	**0.01 0.05**
Group Allocation	**Equal (N1 = N2)**
Input Type	**Simple (Differences)**
Distribution to Simulate	**TukeyGH**
G (Skewness)	**0.12**
H (Kurtosis)	**0.07**
δ0 (Null Difference)	**0**
δ1 (Actual Difference)	**3**
μ2 (Group 2 Mean)	**0**
Group Standard Deviations	**Equal**
σ (Standard Deviation)	**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
Hypotheses: H0: δ = δ0 vs. H1: δ ≠ δ0
Test Statistic: Mann-Whitney U or Wilcoxon Rank-Sum Test
Simulated Distribution: TukeyGH

	Sample Size			Mean			Difference		Standard Deviation	Alpha		
Power	N1	N2	N	Null μ1.0	Actual μ1.1	Reference μ2	Null δ0	Actual δ1	σ	Target	Actual	G | H
0.9520 (0.0094) [0.9426 0.9614]	8	8	16	0	3	0	0	3	1	0.01	0.002 (0.002) [0 0.004]	0.12 | 0.07
0.9185 (0.012) [0.9065 0.9305]	16	16	32	0	3	0	0	3	2	0.01	0.007 (0.004) [0.004 0.011]	0.12 | 0.07
0.9140 (0.0123) [0.9017 0.9263]	31	31	62	0	3	0	0	3	3	0.01	0.011 (0.005) [0.006 0.016]	0.12 | 0.07
0.9150 (0.0122) [0.9028 0.9272]	52	52	104	0	3	0	0	3	4	0.01	0.008 (0.004) [0.004 0.013]	0.12 | 0.07
0.9010 (0.0131) [0.8879 0.9141]	78	78	156	0	3	0	0	3	5	0.01	0.009 (0.004) [0.005 0.014]	0.12 | 0.07
0.9275 (0.0114) [0.9161 0.9389]	5	5	10	0	3	0	0	3	1	0.05	0.027 (0.007) [0.019 0.034]	0.12 | 0.07
0.9110 (0.0125) [0.8985 0.9235]	11	11	22	0	3	0	0	3	2	0.05	0.046 (0.009) [0.036 0.055]	0.12 | 0.07
0.9350 (0.0108) [0.9242 0.9458]	22	22	44	0	3	0	0	3	3	0.05	0.052 (0.01) [0.042 0.061]	0.12 | 0.07
0.9065 (0.0128) [0.8937 0.9193]	36	36	72	0	3	0	0	3	4	0.05	0.051 (0.01) [0.041 0.061]	0.12 | 0.07
0.9005 (0.0131) [0.8874 0.9136]	54	54	108	0	3	0	0	3	5	0.05	0.050 (0.01) [0.04 0.059]	0.12 | 0.07

Pool Size: 10000. Simulations: 2000. Run Time: 23.63 seconds.
User-Entered Random Seed: 3915264

Power	The probability of rejecting a false null hypothesis when the alternative hypothesis is true. The second row provides the precision and a 95% confidence interval for Power, (Power Precision) [95% LCL and UCL], based on the size of the simulation.
N1 and N2	The number of items sampled from each population.
N	The total sample size. N = N1 + N2.
μ1.0	The mean of group 1 under the null hypothesis.
μ1.1	The actual mean of group 1 at which power and sample size are calculated.
μ2	The mean of group 2 under both the null and alternative hypotheses.
δ0	The mean difference under the null hypothesis. δ0 = μ1.0 - μ2.
δ1	The actual mean difference at which power and sample size are calculated. δ1 = μ1.1 - μ2.
σ	The assumed population standard deviation for both groups 1 and 2.
Target Alpha	The probability of rejecting a true null hypothesis. It is set by the user.
Actual Alpha	The alpha level that was actually achieved by the experiment. The second row provides the precision and a 95% confidence interval for Alpha, (Alpha Precision) [95% LCL and UCL], based on the size of the simulation.
Additional Columns	The other parameters required to specify the distributions.

Summary Statements

A parallel two-group design will be used to test whether the Group 1 population distribution is different from the Group 2 population distribution (H0: δ = 0 versus H1: δ ≠ 0, δ = population distribution difference). 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 (α) of 0.01. A standard deviation of 1 is assumed for both Group 1 and Group 2. To detect a difference in means of μ1 - μ2 = 3 - 0 = 3 with 90% power, the number of needed subjects will be 8 in Group 1 and 8 in Group 2. These results are based on 2000 simulations (Monte Carlo samples) from the TukeyGH distribution with G = 0.12, H = 0.07.

Dropout-Inflated Sample Size

	Sample Size			Dropout-Inflated Enrollment Sample Size			Expected Number of Dropouts		
Dropout Rate	**N1**	**N2**	**N**	**N1'**	**N2'**	**N'**	**D1**	**D2**	**D**
20%	8	8	16	10	10	20	2	2	4
20%	16	16	32	20	20	40	4	4	8
20%	31	31	62	39	39	78	8	8	16
20%	52	52	104	65	65	130	13	13	26
20%	78	78	156	98	98	196	20	20	40
20%	5	5	10	7	7	14	2	2	4
20%	11	11	22	14	14	28	3	3	6
20%	22	22	44	28	28	56	6	6	12
20%	36	36	72	45	45	90	9	9	18
20%	54	54	108	68	68	136	14	14	28

Dropout Rate	The percentage of subjects (or items) that are expected to be lost at random during the course of the study and for whom no response data will be collected (i.e., will be treated as "missing"). Abbreviated as DR.
N1, N2, and N	The evaluable sample sizes at which power is computed. If N1 and N2 subjects are evaluated out of the N1' and N2' subjects that are enrolled in the study, the design will achieve the stated power.
N1', N2', and N'	The number of subjects that should be enrolled in the study in order to obtain N1, N2, and N evaluable subjects, based on the assumed dropout rate. After solving for N1 and N2, N1' and N2' are calculated by inflating N1 and N2 using the formulas N1' = N1 / (1 - DR) and N2' = N2 / (1 - DR), with N1' and N2' always rounded up. (See Julious, S.A. (2010) pages 52-53, or Chow, S.C., Shao, J., Wang, H., and Lokhnygina, Y. (2018) pages 32-33.)
D1, D2, and D	The expected number of dropouts. D1 = N1' - N1, D2 = N2' - N2, and D = D1 + D2.

Dropout Summary Statements

Anticipating a 20% dropout rate, 10 subjects should be enrolled in Group 1, and 10 in Group 2, to obtain final group sample sizes of 8 and 8, respectively.

References

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.

Devroye, Luc. 1986. Non-Uniform Random Variate Generation. Springer-Verlag. New York.

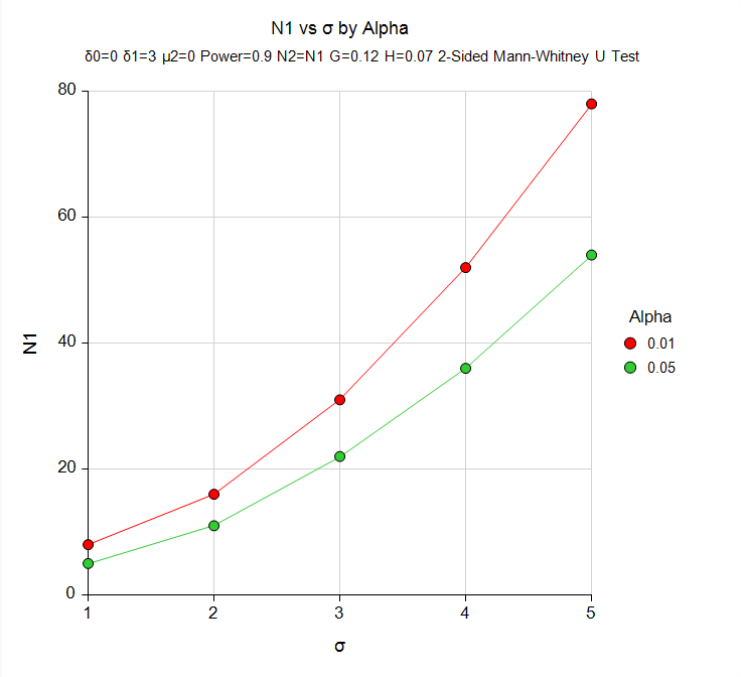
Matsumoto, M. and Nishimura,T. 1998. 'Mersenne twister: A 623-dimensionally equidistributed uniform pseudorandom number generator.' ACM Trans. On Modeling and Computer Simulations.

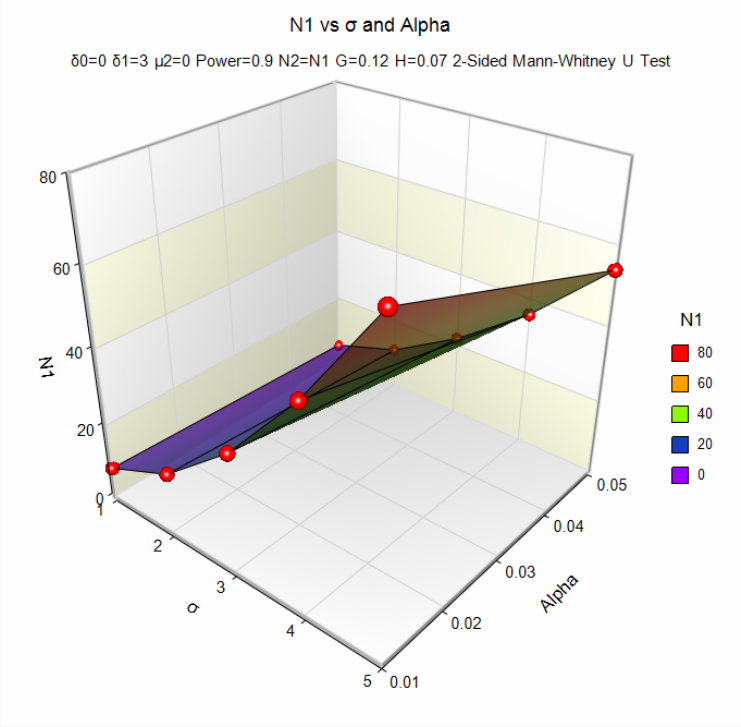
Zar, Jerrold H. 1984. Biostatistical Analysis (Second Edition). Prentice-Hall. Englewood Cliffs, New Jersey.

These reports show the values of each of the parameters, one scenario per row. Since these results are based on simulation, they will vary from one calculation to the next.

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 for Normal Data

Suppose a new corn fertilizer is to be compared to a current fertilizer. 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-Whitney-Wilcoxon 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 2** 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	**Power**
Alternative Hypothesis	**One-Sided (H1: μ1 - μ2 > δ0)**
Simulations	**100000**
Random Seed	**2344877** (for Reproducibility)
Alpha	**0.05**
Group Allocation	**Equal (N1 = N2)**
Sample Size Per Group	**45**
Input Type	**Simple (Differences)**
Distribution to Simulate	**Normal**
δ0 (Null Difference)	**0**
δ1 (Actual Difference)	**10**
μ2 (Group 2 Mean)	**0**
Group Standard Deviations	**Equal**
σ (Standard Deviation)	**25**

Output

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

Numeric Results

Solve For: Power
Hypotheses: H0: δ ≤ δ0 vs. H1: δ > δ0
Test Statistic: Mann-Whitney U or Wilcoxon Rank-Sum Test
Simulated Distribution: Normal

	Sample Size			Mean			Difference		Standard	Alpha	
				Null	Actual	Reference	Null	Actual	Deviation		
Power	N1	N2	N	μ1.0	μ1.1	μ2	δ0	δ1	σ	Target	Actual
0.5814	45	45	90	0	10	0	0	10	25	0.05	0.051
(0.0031) [0.5783 0.5844]											(0.001) [0.05 0.052]

Pool Size: 200000. Simulations: 100000. Run Time: 20.20 seconds.
User-Entered Random Seed: 2344877

The power of the Mann-Whitney test in this scenario is 0.5814 (0.5783 0.5844). This power is only slightly less than the power of the t-test (0.594) for the corresponding scenario.

Example 3 – Validation using Zhao, Rahardja, and Qu (2008)

Zhou, Rahardja, and Qu (2008) page 467 present a table of sample sizes and powers based on their formulas and simulation results. In the example, the data considered follow a multinomial distribution, which heavily increases the likelihood of ties. For the first set of scenarios, the power is 80%, the significance level is 0.05, and Group 2 is assumed to have 53% of the total number of individuals in the sample. The ninth line has the two different multinomial distributions defined as M(0.66, 0.15, 0.19) for Group 1 and M(0.55, 0.15, 0.30) for Group 2. The total sample size is estimated at 502.

For reproducibility, we'll use a random seed of 6283155.

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 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 (H1: μ1 - μ2 ≠ δ0)**
Simulations	**50000**
Random Seed	**6283155 (for Reproducibility)**
Power	**0.80**
Alpha	**0.05**
Group Allocation	**Enter R = N2/N1, solve for N1 and N2**
R	**1.12766**
Input Type	**General**
Group 1 Distribution|H0	**Multinomial(0.66 0.15 0.19)**
Group 1 Distribution|H1	**Multinomial(0.66 0.15 0.19)**
Group 2 Distribution|H0	**Multinomial(0.66 0.15 0.19)**
Group 2 Distribution|H1	**Multinomial(0.55 0.15 0.30)**
M0 (Mean|H0) Parameter Value(s)	**0**
M1 (Mean|H1) Parameter Value(s)	**1**

Output

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

Numeric Results

Solve For:	Sample Size
Hypotheses:	H0: δ = δ0 vs. H1: δ ≠ δ0
Test Statistic:	Mann-Whitney U or Wilcoxon Rank-Sum Test
Group 1 Distributions:	H0: Multinomial(0.66 0.15 0.19) H1: Multinomial(0.66 0.15 0.19)
Group 2 Distributions:	H0: Multinomial(0.66 0.15 0.19) H1: Multinomial(0.55 0.15 0.30)

	Sample Size			Difference		Alpha	
				Null	Actual		
Power	N1	N2	N	δ0	δ1	Target	Actual
0.8019	236	266	502	0	-0.2	0.05	0.05
(0.0035) [0.7984 0.8054]							(0.002) [0.048 0.052]

Pool Size: 100000. Simulations: 50000. Run Time: 5.14 minutes.
User-Entered Random Seed: 6283155

The estimated sample size is 502, which matches the expected result exactly.