

Chapter 257

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

Introduction

This procedure provides sample size and power calculations for one- or two-sided two-sample Mann-Whitney U or Wilcoxon rank-sum test on continuous data without ties as presented by Noether (1987). 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 design corresponding to this test is referred to as a *parallel-groups* design. It is used to compare two independent groups.

Sample Size Calculation for Ordinal Data

If, instead of continuous data, the data are ordered categorical (such as Likert scale data), Zhao *et al.* (2008) presents an alternative formulation for computing power and sample size. This procedure is available in **PASS** in the procedure entitled *Tests for Two Ordered Categorical Variables (Non-Proportional Odds)*.

Test Assumptions

When running a Wilcoxon-Mann-Whitney test, the basic assumptions are that there are two independent populations with continuous responses that does not have ties. If we can additionally assume that the two populations are identical except for a difference in location, this test can be used as a test of equal means or medians.

Mann-Whitney U Test

Suppose that samples of sizes N_1 and N_2 are available from two independent random variables Y_1 and Y_2 . Further suppose that the statistical hypotheses that are to be compared are that the two variables have the same distribution versus that Y_1 tends to be larger than Y_2 . The Mann-Whitney U statistic is found by counting the number of times an observation randomly selected from the first population is greater than an observation randomly drawn from the second population.

This is often written as

$$U = \#(Y_{1i} > Y_{2j}), \quad i = 1, \dots, N_1; j = 1, \dots, N_2.$$

Nowadays, U includes an adjustment for ties so that its calculation becomes

$$U' = \#(Y_{1i} > Y_{2j}) + \frac{\#(Y_{1i} = Y_{2j})}{2}, \quad i = 1, \dots, N_1; j = 1, \dots, N_2.$$

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

If U , or U' , is divided by the number of possible comparisons between the two samples, N_1N_2 , an estimate of the *competing probability*, $P_1 = Pr(Y_1 > Y_2)$, is found. The null hypothesis is that $P_1 = 0.5$ versus the alternative that $P_1 > 0.5$.

Planning Estimates of P_1

We see that in order to compute an estimate of the required sample size, we need planning estimates of P_1 . This can be obtained from previous studies, from pilot studies, or from a form of simulation.

The simulation could proceed as follows. Generate random samples of Y_1 and Y_2 that adhere to the sample size structure defined by the N_1 , and N_2 . Y_1 and Y_2 are generated so that the variances are the same and the means are different by the minimum amount that the experiment is to detect. From this simulated sample, estimate P_1 using

$$\hat{P}_1 = \frac{1}{N_1N_2} \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} \delta(Y_{1i} - Y_{2j})$$

where

$$\delta(x) = \begin{cases} 1 & \text{if } x > 0 \\ 0.5 & \text{if } x = 0 \\ 0 & \text{if } x < 0 \end{cases}$$

Repeat this simulation several times until a reasonable set of planning estimates of P_1 is obtained. Use the mean (or median) of the simulated \hat{P}_1 as the planning estimate for the sample size calculation.

Wilcoxon Rank-Sum Test

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 Wilcoxon 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

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

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

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 $H_1: \delta \neq 0$,

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

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 α .

Power Calculation for Mann-Whitney U or Wilcoxon Rank-Sum Tests

The sample size formula for the Mann-Whitney U or Wilcoxon rank-sum two-sided tests is provided by Zhao *et al.* (2008) and Noether (1987) for the case when there are no ties in the data as

$$N = \frac{\left(z_{1-\frac{\alpha}{2}} + z_{\beta}\right)^2}{12r(1-r)\left(P_1 - \frac{1}{2}\right)^2}$$

where $r = N_1/(N_1 + N_2)$.

Example 1 – Finding the Sample Size

Researchers wish to compare two types of local anesthesia using a Mann-Whitney U test 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 competing probability, P_1 , is in the range 0.58 to 0.64. How many participants are needed to achieve 90% power at a significance level of 0.05?

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**
 Power..... **0.90**
 Alpha..... **0.05**
 Group Allocation **Equal (N1 = N2)**
 P1 (Probability $Y_1 \geq Y_2$)..... **0.58, 0.6, 0.62, 0.64**

Output

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

Numeric Reports

Numeric Results

Solve For: Sample Size
 Alternative Hypothesis: $H_1: |P_1 - 0.5| \neq 0$

Power		Sample Size			P($Y_1 \geq Y_2$)	
Target	Actual	N1	N2	N	P1	Alpha
0.9	0.90038	274	274	548	0.58	0.05
0.9	0.90141	176	176	352	0.60	0.05
0.9	0.90090	122	122	244	0.62	0.05
0.9	0.90205	90	90	180	0.64	0.05

Target Power The desired power. May not be achieved because of integer N1 and N2.
 Actual Power The achieved power. Because N1 and N2 are integers, 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. $N = N_1 + N_2$.
 P1 The probability that $P(Y_1 \geq Y_2)$. Y_i ($i = 1, 2$) is a random response from group i . This is the effect size.
 Alpha The probability of rejecting a true null hypothesis.

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

Summary Statements

Group sample sizes of 274 and 274 achieve 90% power to reject the null hypothesis of equal group response distributions when the effect size as measured by P1 is 0.58 and the significance level (alpha) is 0.05 using a two-sided Wilcoxon-Mann-Whitney test. P1 is the probability that an observation from group 1 exceeds an observation from group 2.

Dropout-Inflated Sample Size

Dropout Rate	Sample Size			Dropout-Inflated Enrollment Sample Size			Expected Number of Dropouts		
	N1	N2	N	N1'	N2'	N'	D1	D2	D
20%	274	274	548	343	343	686	69	69	138
20%	176	176	352	220	220	440	44	44	88
20%	122	122	244	153	153	306	31	31	62
20%	90	90	180	113	113	226	23	23	46

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 Lohknygina, 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, 343 subjects should be enrolled in Group 1, and 343 in Group 2, to obtain final group sample sizes of 274 and 274, respectively.

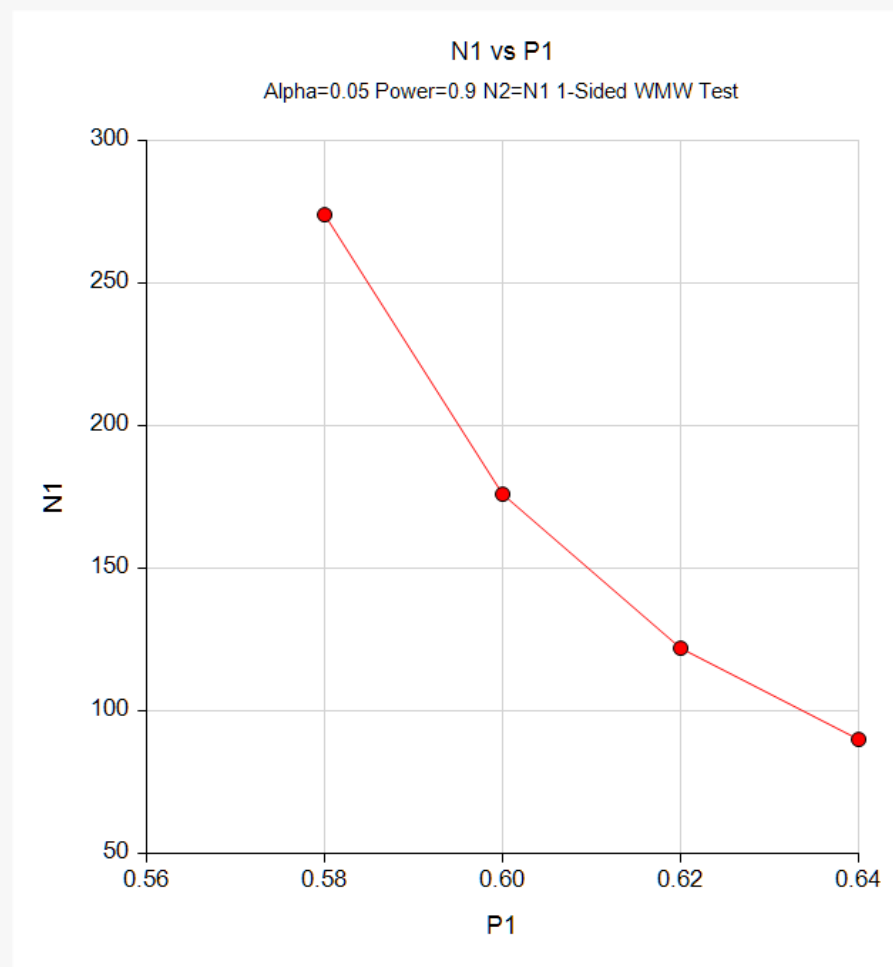
References

- Lehmann, E.L. 1975. 'Nonparametrics - Statistical Methods Based on Ranks.' Holden-Day, Inc. Oakland, California.
- Machin, D., Campbell, M., Tan, S.B., and Tan, S.H. 2018. Sample Size Tables for Clinical, Laboratory and Epidemiology Studies, 4th Edition. John Wiley & Sons. Hoboken, NJ.
- Noether, G.E. 1987. 'Sample size determination for some common nonparametric tests.' Journal of the American Statistical Association, Volume 82(1), 645-647.
- Zhao, Y.D., Rahardja, D., Qu, Y. 2008. 'Sample size calculation for the Wilcoxon-Mann-Whitney test adjusting for ties.' Statistics in Medicine, 27, pages 462-468.

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 – Validation using Machin et al. (2018)

Machin *et al.* (2018) page 75 presents an example in which P_1 is 0.726, alpha (two-sided) is 0.05, power is 0.90, and the allocation ratio, R , is 2/3. The required sample size is given as 72.

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 **Sample Size**
 Alternative Hypothesis **Two-Sided**
 Power..... **0.90**
 Alpha..... **0.05**
 Group Allocation **Enter R = N2/N1, solve for N1 and N2**
 R **0.666667**
 P1 (Probability $Y1 \geq Y2$)..... **0.726**

Output

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

Numeric Results

Solve For: [Sample Size](#)
 Alternative Hypothesis: $H1: |P1 - 0.5| \neq 0$

Power		Sample Size			R (N2 / N1)		P(Y1 ≥ Y2) P1	Alpha
Target	Actual	N1	N2	N	Target	Actual		
0.9	0.90288	43	29	72	0.667	0.674	0.726	0.05

PASS has also calculated the total sample size to be 72. Thus, the procedure is validated.