Chapter 251

Tests for Two Ordered Categorical Variables (Non-Proportional Odds, Wilcoxon-Mann-Whitney)

Introduction

This module computes power and sample size for tests of ordered categorical (ordinal) data without making the proportional odds assumption. The Wilcoxon-Mann-Whitney (WMW) test statistic is adopted to test whether the two groups being compared are different or not. Hence, this procedure may be used for sizing studies that use the WMW test on ordinal data.

A common scale the results in ordered categorical data is the Likert scale. Ordinal data often result from surveys in general and quality of life (QoL) surveys in particular in which responses are categories such as *very good, good, moderate, poor*. When there are only two categories, an analysis using two proportions should be used. When there are more than two responses, and those responses can be ordered, the techniques described in this chapter can be used.

COVID-19

Studies of the efficacy and safety of therapeutic agents for the treatment of hospitalized patients with novel coronavirus disease (COVID-19) provides an additional example of a primary endpoint of illness severity that are on an ordinal scale. The World Health Organization (WHO) in their COVID-19 Therapeutic Trial Synopsis document (February 18, 2020) recommends two-arm clinical trials be conducted with a nine-point ordinal scale of illness severity.

Technical Details

The power and sample size formulae presented here are given in Machin *et al*. (2018) and Zhao *et al*. (2008).

Wilcoxon-Mann-Whitney Test Statistic

Suppose ordinal variables Y_1 (control) and Y_2 (experimental) each have the same K possible outcomes $C_1, \ldots, C_K$. Further suppose that these categories can be ordered so that C_k is more desirable than C_j if $k < j$. Hence C_1 is the best outcome and C_K is the worst.

The difference between the distribution of Y_1 and Y_2 is measured by the competing probability

$$\pi = \Pr(Y_1 < Y_2) + \frac{1}{2}\Pr(Y_1 = Y_2)$$

The null hypothesis of no difference between the two distributions is given by $H_0: \pi = \frac{1}{2}$.

Define the counts

$$N_{ik} = \sum_{j=1}^{N_i} I(Y_i = C_k), \qquad i = 1,2 \; and \; k = 1, \ldots, K$$

where $I(x)$ is an indicator variable whose values are 1 or 0 depending on whether x is *true* or *false*. Here, N_1 is the sample size of Y_1 and N_2 is the sample size of Y_2. The total sample size of the study is $N = N_1 + N_2$.

Let $p_{ik} = Pr(Y_i = C_k), \; i = 1,2 \; and \; k = 1, \ldots, K$.

The WMW test is computed as follows.

First, estimate π using

$$\hat{\pi} = \frac{1}{N_1 N_2} \sum_{i_1=1}^{N_1} \sum_{i_2=1}^{N_2} d\left(Y_{1i_1} - Y_{2i_2}\right)$$

where

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

Let $E(\hat{\pi}) = \mu_1$.

Second, estimate the variance of $\hat{\pi}$ under the null hypothesis using

$$\hat{V}_0 = \frac{N+1}{12N_1N_2} - \frac{1}{12N(N-1)N_1N_2} \sum_{k=1}^{K} \left(M_k^3 - M_k\right)$$

where

$$M_k = N_{1k} + N_{2k}, \; k = 1, \ldots, K$$

Finally, construct the z-statistic to test the null hypothesis as

$$z_0 = \frac{\hat{\pi} - 0.5}{\sqrt{\hat{V}_0}}$$

The significance test proceeds using the assumption that z_0 has the standard normal distribution.

Power and Sample Size

Page 59 of Machin *et al.* (2018) provides the sample size for the two-sided test as

$$N = \frac{(1+R)^2}{12R}\left\{\frac{\left(z_{1-\frac{\alpha}{2}} + z_{1-\beta}\right)^2\left[1 - \frac{1}{(1+R)^3}\sum_{k=1}^{K}(RP_{1k} + P_{2k})^3\right]}{\left[\sum_{k=2}^{K}\left(P_{1k}\sum_{h=1}^{k-1}P_{2h}\right) + 0.5\left(\sum_{k=1}^{K}P_{1k}P_{2k}\right) - 0.5\right]}\right\}$$

where $R = N_2/N_1$ and P_{ik} is the assumed population value of p_{ik}.

The power is found by solving this equation for $1-\beta$.

If a one-sided test is needed, replace $\alpha/2$with α.

Example 1 – Finding the Sample Size

Suppose a clinical trial is planned to compare the response to certain treatment. The subjects are divided into two groups: those that will receive the current treatment and those that will receive an experimental treatment. Three months after the administration of the treatment, the subjects rate their response as *very good, good, neutral, poor,* and *very poor*. Historically, the responses have been about 10% *very good*, 20% *good*, 40% *neutral*, 20% *poor*, and 10% *very poor*.

The researchers want to consider several different scenarios for the response distribution of the experiment group, each of which will show a shift toward the positive (very good and good) categories.

These patterns will be loaded in the spreadsheet. The spreadsheet will appear as follows:

C	E1	E2	E3
1	2	3	5
2	4	3	2
4	2	1	1
2	1	1	1
1	1	2	1

Note that the above patterns result in the following response proportions:

C	E1	E2	E3
0.1	0.2	0.3	0.5
0.2	0.4	0.3	0.2
0.4	0.2	0.1	0.1
0.2	0.1	0.1	0.1
0.1	0.1	0.2	0.1

They want to look at the sample size requirements to achieve a power of 0.90. They want to set alpha to 0.05 and analyze the results with a two-sided test. They want the size of the treatment group to be twice the size of the control group.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the procedure window. You may then make the appropriate entries as listed below, or open **Example 1** by going to the **File** menu and choosing **Open Example Template**.

Option	Value
Design Tab	
Solve For	**Sample Size**
Null Hypothesis	**Two-Sided**
Power	**0.8**
Alpha	**0.05**
Group Allocation	**Enter R = N2/N1, solve for N1 and N2**
R	**2**
P1's Input Type	**Enter Columns Containing Sets of P1's**
Columns Containing Sets of P1's	**1**
P2's Input Type	**Enter columns containing sets of P2's**
Columns Containing Sets of P2's	**2-4**

Input Spreadsheet Data

Row	C	E1	E2	E3
1	1	2	3	5
2	2	4	3	2
3	4	2	1	1
4	2	1	1	1
5	1	1	2	1

Annotated Output

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

Numeric Results

Numeric Results

Hypotheses Two-Sided Test
Group 1 Control Group
Group 2 Experimental Group

Power	Sample Size Cntl N1	Sample Size Exp N2	Sample Size Total N	Target N2/N1 R	Number of Categories K	Category Proportions Sets Cntl P1	Category Proportions Sets Trt P2	Competing Probability π	Alpha
0.80472	51	102	153	2	5	C	E1	0.635	0.05
0.80267	85	170	255	2	5	C	E2	0.605	0.05
0.81684	22	44	66	2	5	C	E3	0.710	0.05

Value Lists

Name	Values
C	0.1, 0.2, 0.4, 0.2, 0.1
E1	0.2, 0.4, 0.2, 0.1, 0.1
E2	0.3, 0.3, 0.1, 0.1, 0.2
E3	0.5, 0.2, 0.1, 0.1, 0.1

References

Machin, D., Campbell, M., Tan, S.B., and Tan, S.H. 2018. Sample Size Tables for Clinical Studies, 4th Edition. John Wiley & Sons. Hoboken, NJ.
Zhao, Y.D., Rahardja, D. Qu, Y. 2008. 'Sample size calculation for the Wilcoxon-Mann-Whitney test adjusting for ties.' Statistics in Medicine, 27, 462-468.

Report Definitions

Power is the probability of rejecting a false null hypothesis.
N1 is the number of subjects in the group 1, the control group.
N2 is the number of subjects in the group 2, the experimental group.
N is the total sample size, N1 + N2.
R is the target ratio of N2 to N1, so that N2 = R × N1. It may not be achieved exactly because of rounding
Num Cat's K is the number of categories in the response variable.
Grp 1 Prop's Set P1 is the name of the set containing the response proportions for each of the K categories in group 1, the control group.
Grp 2 Prop's Set P2 is the name of the set containing the response proportions for each of the K categories in group 2, the experimental group.
π, the competing probability, is a measure of the difference between the two group distributions. Here, π = Pr(Y1>Y2) + Pr(Y1=Y2) / 2. Under the null hypothesis π = 0.5. You should avoid P2's that result in π values close to 0.5.
Alpha is the probability of rejecting a true null hypothesis.

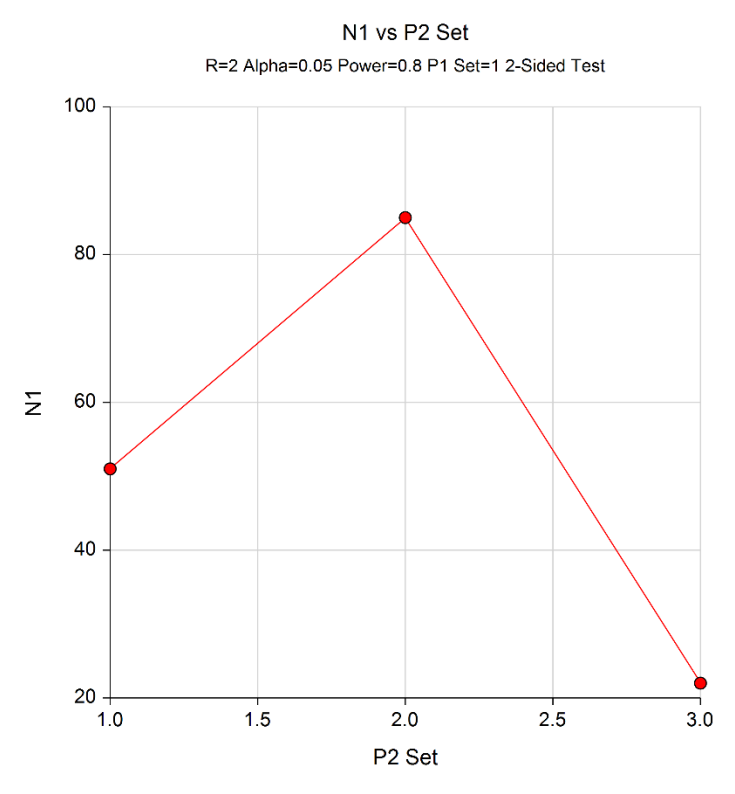
Summary Statements

Samples of 51 subjects in the control group and 102 subjects in experimental group achieve 80% power to detect a difference between the group 1 proportions and the group 2 proportions when the significance level (alpha) is 0.05 using a two-sided Wilcoxon-Mann-Whitney test. The number of response categories is 5. The response proportions in group 1 are 0.1, 0.2, 0.4, 0.2, 0.1. The response proportions in group 2 are 0.2, 0.4, 0.2, 0.1, 0.1.

This report shows the numeric results of this sample size study. The definitions of the items on the report are given in the Report Definitions section.

Chart Section

Chart Section



This plot gives a visual presentation to the results in the Numeric Report.

Example 2 – Validation using Machin *et al.* (2018)

Machin *et al.* (2018) pages 64 - 65 have an example in which they calculate per group sample size to be 3011 when alpha is 0.05, power is 90%, the control group proportions are 0.6632, 0.1458, 0.1910, and the treatment group proportions are 0.6062, 0.2338, 0.1600.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load the procedure window. You may then make the appropriate entries as listed below, or open **Example 2** by going to the **File** menu and choosing **Open Example Template**.

Option	Value
Design Tab	
Solve For	**Sample Size**
Null Hypothesis	**Two-Sided**
Power	**0.8**
Alpha	**0.05**
Group Allocation	**Equal (N1 = N2)**
P1's Input Type	**Enter P11, P12, ..., P1K Pattern**
P11, P12, ..., P1K Pattern	**0.6632 0.1458 0.1910**
P2's Input Type	**Enter P21, P22, ..., P2K Pattern**
P21, P22, ..., P2K Pattern	**0.6062 0.2338 0.1600**

Output

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

Numeric Results

Numeric Results

Hypotheses	Two-Sided Test
Group 1	Control Group
Group 2	Experimental Group

	Sample Size				Category Proportions			
	Cntl	Exp	Total	Number of	Sets		Competing	
Power	N1	N2	N	Categories K	Cntl P1	Trt P2	Probability π	Alpha
0.80009	3011	3011	6022	3	P1	P2	0.482	0.05

Value Lists

Name	Values
P1	0.663, 0.146, 0.191
P2	0.606, 0.234, 0.16

PASS also calculated the required sample size as 3011 per group. Thus, the procedure is validated.

As Machin *et al.* (2018) alluded to in their book, this is an unusually high sample size which they indicate should be investigated further. From the **PASS** report we can see that the high sample size occurs because π happens to be very close to 0.5. If the P2 values were adjusted a little so that π is closer 0.45, the sample size would be reduced to only 417. We used the pattern 5, 2, 2, for P2 to achieve this sample size.