UX Research Method:
Comparing Two Independent Groups

The purpose of this document is to illustrate a method by which two (independent) groups of Likert-scale responses can be compared.

Context

This type of problem commonly arises in User Experience (UX) research, for example, where we need to compare research responses from two exclusively different demographic or ethnographic groups, or where there is a need to contrast research responses to two (exclusively presented) design alternatives.

The Mann-Whitney Test

We can use the Mann-Whitney test to establish whether there is a statistically significant difference between two independent groups. The shape of the underlying distribution (or distributions) from which the samples are drawn is irrelevant and need not be normal. If the shape of the distributions of both groups are the same, then the test can be used to infer whether there is a difference in the median values of both groups. If, however, the shape of the distribution of both groups are different, then the test is used to infer a difference between the distributions.

The Mann-Whitney test was originally developed by H. B. Mann and D. R. Whitney in 1947 as a test of stochastic equality between two groups of sample data. That is, the probability of a random observation from group A being higher than a random observation from group B is the same as that of a random observation from group B being higher than a random observation from group A.

The test is the equivalent to the two-sample t-test for parametric data but can be applied in circumstances where the assumptions necessary to apply the two-sample t-test are not satisfied. That is, the Mann-Whitney test can be applied to non-parametric data, where the data is not necessarily normally distributed and / or the variances are not equal. The shape of the underlying distribution (or distributions) from which the samples are drawn is irrelevant and, in addition, large outliers have no distortionary effect on this test.

Conditions necessary to use the Mann-Whitney Test

There are three key assumptions, or preconditions, to the use the Mann-Whitney test:

1. There must be one dependent variable that is measured at the ordinal level or is continuous and can be converted into ranked data. As such, this makes the test useful for Likert-scale responses.
2. There must be one independent variable that consists of two categorical, independent groups. For example, Group A comprising survey respondents with an annual income below or equal to $50,000 and Group B respondents having an annual income above $50,000. In order to consider three or more groups, an alternative statistical test, such as the Kruskall-Wallis test for unpaired data or Friedman test for paired data, may be appropriate.
3. The observations comprising both sample groups must be independent. That is, there must be no relationship between the observations within each group nor between the groups. Where such a relationship exists, an alternative statistical test, such as the Wilcoxon signed-rank test may be appropriate.

Applying the Mann-Whitney Test

Under the Mann-Whitney test, the null hypothesis (H_0) is that there is no statistical difference between the distributions or medians of the two groups, as appropriate.[1] The alternative hypothesis (H_1) is usually that the distributions are different. In this form, alternative hypothesis is a 'two-tailed' test, in that we're not specifying whether one distribution is stochastically larger than the other, although we could construct the alternative hypothesis in this way to give us a 'one-tailed' test.

In order to test our null hypothesis for significance, we calculate the Mann-Whitney statistic (U) and compare it against a critical value ($U_{(crit,\alpha)}$) for a specified level of significance, α. Typically, α is either 0.01 (90% significance) or 0.05 (95% significance).

If $U \geq U_{(crit,\alpha)}$, the null hypothesis is accepted, otherwise, where $U < U_{crit}$ the alternative hypothesis is accepted.

For smaller group sizes, up to around 30, U_{crit} can be obtained from published statistical tables (for both one-tailed and two-tailed tests).

[1] More accurately, the null hypothesis is that it is equally likely that a randomly selected value from one group will be less than or greater than a randomly selected value from a second group.

Note that for larger group sizes, it is more usual to use the z-test[2] rather than comparing U to $U_{(crit,\alpha)}$. This approach is taken because with larger group sizes, the distribution of U tends towards being normally distributed, and the z-value can be calculated for any group sizes (whereas $U_{(crit,\alpha)}$ values are not readily available for larger group sizes). With the z-test, if the absolute value of the obtained z-value is less than 1.96, then the hull hypothesis (H_0) should be accepted, and if the absolute value of the obtained z-value is greater than 1.96, then it should be rejected.

There are several on-line calculators that can be used to perform the Mann-Whitney test, such as that provided at Social Science Statistics (https://www.socscistatistics.com/tests/mannwhitney/default.aspx).

Example 1: Small sample

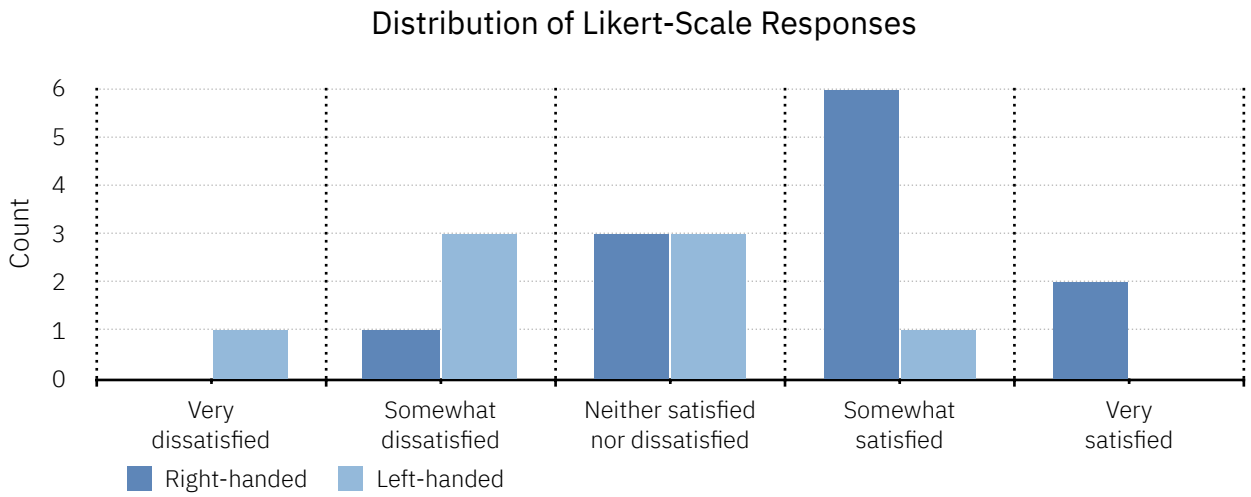
A recent usability test has been undertaken. Twenty people were asked about their level of satisfaction with an-development mobile app and provided their responses as a Likert-score on the scale: (1) Very dissatisfied; (2) Somewhat dissatisfied; (3) Neither satisfied nor dissatisfied; (4) Somewhat satisfied; and, (5) Very satisfied.

Of the 20 people surveyed, 12 identified as right-handed and 8 identified as left-handed. The research question has arisen as to whether there is a difference in between the overall levels of satisfaction between right-handed and left-handed test participants. The table, below, summarises the responses of the 20 participants:

Group	Very dissatisfied (1)	Somewhat dissatisfied (2)	Neither satisfied nor dissatisfied (3)	Somewhat satisfied (4)	Very satisfied (5)	Total responses
Right-handed (R)	0	1	3	6	2	**12**
Left-handed (L)	1	3	3	1	0	**8**

The graph, on the following page, illustrates the distributions of the two groups for this example and, in this case, the test will be to establish whether the distributions are different.

[2] The z-test is based on the z-distribution and is used to determine probabilities and percentiles for normal distributions. The z-distribution is a normal distribution with mean zero and standard deviation of 1.



Here, we can use the Mann-Whitney test to determine whether there is a statistically significant difference between the medians or distribution of the responses from both groups.

For this example, and as shown in the manual workings in Attachment 1, the calculated Mann-Whitney test statistic, U, is 16 and $U_{(crit,\ \alpha=0.05)}$ is 22.

We therefore reject the null hypothesis (since 16 is not greater than or equal to 22) and conclude that there is a statistical basis (95% confidence) to infer that the responses of left and right-handed respondents are meaningfully different.

A single-tailed Mann-Whitney test could be conducted to also conclude that right-handed users are typically happier than left-handed users, giving us a basis to re-examine our mobile app and see whether improvements for left-handed users are practical.

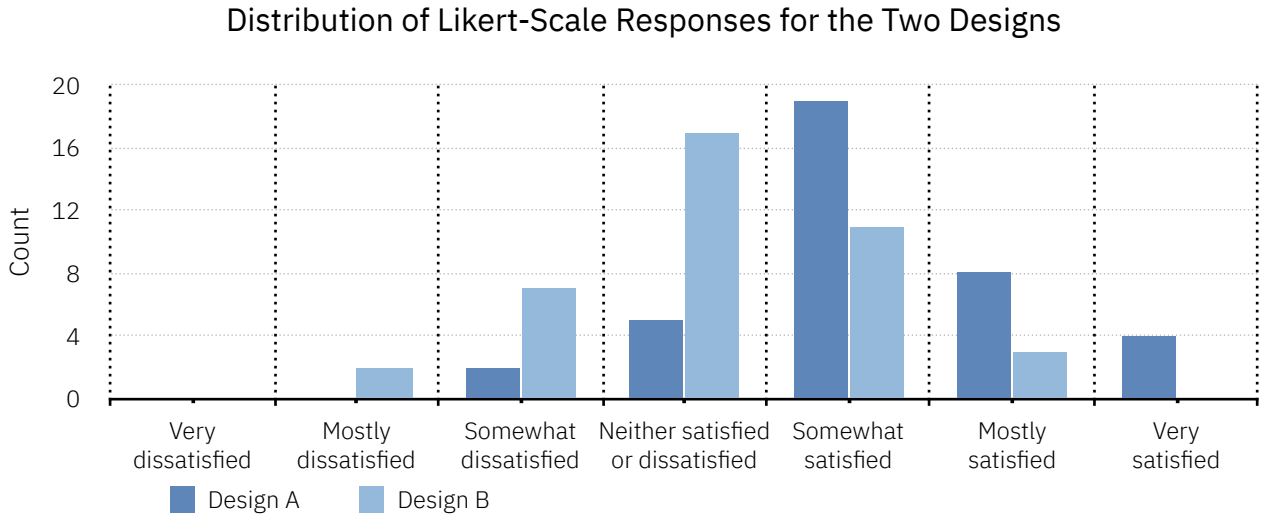
Example 2: Larger sample

In this second example, we have data that comprises responses from a UX research survey of 78 users of a website (with which they were already familiar). These users (or survey participants) were asked a series of questions in relation to a potential redesign. Two interactive designs (“A” and “B”) were presented, with 38 survey responses in relation to design A, and 40 survey responses for design B. No survey participant saw both potential redesigns (and so the groups are independent).

The overall satisfaction of each survey participant with the redesign was assessed using a seven-point Likert-scale response, with the results summarised in the table, on the following page.

Group	Very dissatisfied (1)	Mostly dissatisfied (2)	Somewhat dissatisfied (3)	Neither satisfied or dissatisfied (4)	Somewhat satisfied (5)	Mostly satisfied (6)	Very satisfied (7)	Total responses
Design A	0	0	2	5	19	8	4	**38**
Design B	0	2	7	17	11	3	0	**40**

If we graph this data, then we observe what might be some difference between the responses regarding the two designs (design B seems to perform slightly less well than design A) but its not clear whether this difference is statistically significant.



Again, we can use the Mann-Whitney test to determine whether there is a statistically significant difference between the medians or distribution of the responses from both groups.

With larger sample sizes (say, more than 30 for both groups combined), we may find that statistics tables do not provide $U_{(crit,\alpha)}$ values. Instead we can use the z-test.

Using the Social Science Statistics (https://www.socscistatistics.com/tests/mannwhitney/default.aspx) online calculator we obtain a U value of 355 and a z-value of -4.04365. Since the absolute value of the obtained z-value is greater than 1.96, then the null hypotheses should be rejected, and the medians or distributions of the two samples are statistically different, with a confidence of >95%. Indeed, in this example, the p value is < 0.00001 implying a high level of confidence.

Further reading

- I. E. Allen and C. A. Seaman (2007), “Likert scales and data analyses”, Quality Progress, Vol. 40, No. 7, pp 64.
- H. N. Boone and D. A. Boone (2012), “Analyzing Likert data”, Journal of Extension, Vol. 50, No. 2, pp 1-5.
- J. Carifio, and R. J. Perla (2007), “Ten common misunderstandings, misconceptions, persistent myths and urban legends about Likert scales and Likert response formats and their antidotes”, Journal of Social Sciences, Vol. 3, No. 3, pp 106-116.
- D. L. Clason and T. J. Dormody (1994), “Analyzing data measured by individual Likert-type items”, Journal of Agricultural Education, Vol. 35, pp 4.
- W. J. Conover (1973), “On methods of handling ties in the Wilcoxon signed-rank test”, Journal of the American Statistical Association, Vol. 68, No. 344, pp 985-988.
- J. C. De Winter and D. Dodou (2010), “Five-point Likert items: T test versus Mann-Whitney-Wilcoxon”, Practical Assessment, Research Evaluation, Vol. 15, No. 11, pp 1-12.
- H. B. Mann and D. R. Whitney (1947), “On a test of whether one of two random variables is stochastically larger than the other”, The Annals of Mathematical Statistics, Vol. 18, No. 1 (Mar., 1947), pp. 50-60.
- G. M. Sullivan and A. R. Artino Jr (2013), “Analyzing and interpreting data from Likert-type scales”, Journal of Graduate Medical Education, Vol. 5, No. 4, pp 541-542.
- F. Wilcoxon (1945), “Individual comparisons by ranking methods”, Biometrics Bulletin, Vol. 1, No. 6, pp. 80-83.

Attachment 1: Manual workings for Example 1

Step 1: Order the combined data

The Mann-Whitney test involves ranking the combined data from both right and left-handed groups (a total of 20 responses). To do this, we merge the raw Likert-scores from both our groups and sort into ascending scores. This gives us the data below, with the right-handed and left-handed groups indicated as R and L, respectively:

Score	1	2	2	2	2	3	3	3	3	3	3	4	4	4	4	4	4	4	5	5
Group	L	R	L	L	L	R	R	R	L	L	L	R	R	R	R	R	R	L	R	R

Step 2: Rank the combined data

The next step is to rank the data. Notionally, in this example, our ranking would be from 1 to 20; however, tied ranks must be given equivalent ranking. We can do this by assigning a simple index value to our merged and sorted data, as illustrated in Table 4, below, and then the rank is the average of the index values for the same score. For example, there are four scores of value '2' corresponding to ordinal values 2, 3, 4 and 5. The average of these four ordinal values is 3.5 and so this is the rank allocated to all four scores.

Score	1	2	2	2	2	3	3	3	3	3	3	4	4	4	4	4	4	4	5	5
Group	L	R	L	L	L	R	R	R	L	L	L	R	R	R	R	R	R	L	R	R
Index	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Rank	1	3.5	3.5	3.5	3.5	8.5	8.5	8.5	8.5	8.5	8.5	15	15	15	15	15	15	15	19.5	19.5

Step 3: Sum the ranks for each group

The third step is to add up all of the ranks for each of the two groups; this gives us the 'rank sum' (R) for each group.

$$R_R = \sum_{i=1}^{n_R} r_{R,i} = 3.5+8.5+8.5+8.5+15+15+15+15+15+15+19.5+19.5 = 158$$

and,

$$R_L = \sum_{i=1}^{n_L} r_{L,i} = 1+3.5+3.5+3.5+8.5+8.5+8.5+15 = 52$$

Step 4: Calculate the Mann-Whitney test statistic, U

The next step is to calculate the Mann-Whitney U-value as follows:

$$U = \min(U_R, U_L)$$

where,

$$U_R = n_R \,.\, n_L + \frac{n_R(n_R + 1)}{2} - R_R = 12 \times 8 + (12 \times (12 + 1) / 2) - 158 = 16$$

$$U_L = n_L \,.\, n_R + \frac{n_L(n_L + 1)}{2} - R_L = 8 \times 12 + (8 \times (8 + 1) / 2) - 52 = 80$$

therefore,

$$U = \min(16, 80) = 16$$

Step 5: Calculate the Mann-Whitney test statistic, U

The fifth and final step is to compare the U-value against the critical value for relevant group sizes and given level of statistical significance, α.

The null hypothesis (H_0) is that the distributions (or medians) of the two groups are the same and the alternative hypothesis (H_1) is that they are different. The statistical test is:

if $U \geq U_{crit,\alpha}$ then accept H_0, i.e. the distributions are likely the same, at the given α

or, alternatively

if $U < U_{crit,\alpha}$ then accept H_1, i.e. the distributions are likely different, at the given α

With group sizes of n_R=12 and n_L=8 and a significance level of α=0.05 and a two-tailed hypothesis, the critical value (from tables of critical values) is:

$$U_{crit,\alpha=0.05} = 22$$

Since,

$$U(=16) < (U_{crit,\alpha=0.05} = 22),$$

we accept H_1 and conclude that, at a significance of α=0.05, the distribution of the satisfaction of right-handed respondents is statistically different to the satisfaction of left-handed respondents.