The Mann-Whitney U-test -- Analysis of 2-Between-Group Data with a Quantitative Response Variable

Application: Compare the distributions of scores on a quantitative variable obtained from 2 independent groups. Thus, it is applied in the same data situation as a t-test or an ANOVA for independent samples, except that it is used when the data are either importantly non-normally distributed, the measurement scale of the dependent variable is ordinal (not interval or ratio), or from a too-small sample. There are two versions of the Mann-Whitney U test, one for small samples (i.e., when n < 20 for each group) and one for large samples. It is important to remember the null hypothesis for this test, and to differentiate it from the nulls for the t-test and the median test.

H0: The two populations represented by the two conditions (groups, samples) have the same distribution of scores.

To reject H0: is to say that the population distributions are different in some way, center, spread and/or shape. When the forms of the distributions are similar (as is often the case – compare the size and symmetry of the IQR from the two conditions), then rejecting H0: is interpreted to mean that one population tends to have larger scores (or a larger median) than the other.

Special Note: There is another statistical test very similar to the Mann-Whitney U, called the Wilcoxin signed-ranks test. The two tests were designed at about the same time, following the same logic and mathematical principles. In fact, having computed one of these statistical tests, the researcher can transform their results into the other statistic using a fairly simple formula (much as one can transform a t-value into an F-value, and vice versa). The choice to present the Mann-Whitney U (instead of the Wilcoxin), and this version of the computation in particular was made based on simplicity - this is the most straightforward computational scheme I could find. The complexity, especially for the Wilcoxin, comes from the different possible summary values that can be obtained depending whether the larger or the smaller sample has the higher or lower scores. While there are rules for deciding the correct computations, the procedure described below avoids this issue by using a clever variation of the critical value table.
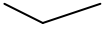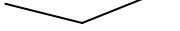
The data: This analysis involves the grouping variable **reptdept** (1 = not separate reptile department, 2 = separate department), and the response variable **reptgood** (rating of reptile quality - 1-10 scale) . Below are the scores for the 12 stores (reptdept, reptgood).

$$1,2 \quad 2,8 \quad 2,9 \quad 2,7 \quad 1,4 \quad 1,7 \quad 2,4 \quad 1,4 \quad 1,5 \quad 2,9 \quad 2,7 \quad 1,2$$

Research Hypothesis: The researcher hypothesized that stores with separate reptile departments would have reptiles of better overall quality than stores that did not have separate reptile departments.

H0: for this analysis: Pet shops that do not have separate reptile departments have the same distribution of reptile quality ratings as shops that do have separate reptile departments.

Step 1 Rearrange the data from lowest to highest score while keeping track of group membership, and assign a rank to each score. If there is a tie, all of the scores that tie receive the average rank of that set of scores.

Group ➔	1	1	1	1	2	1	1	2	2	2	2	2
Reptgood ➔	2	2	4	4	4	5	7	7	7	8	9	9
Rank ➔	1.5	1.5	4	4	4	6	8	8	8	10	11.5	11.5

Ave. of ranks 1 & 2   Ave. of ranks 3,4 & 5   Ave. of ranks 7,8 & 9   Ave. of ranks 11 & 12

Step 2 Compute the sum of the ranks for one of the samples (it really doesn't matter which)

For group 1 (not separate departments) $T_a = 1.5 + 1.5 + 4 + 4 + 6 + 8 = 25$

Step 3 Determine the sample size for the sample for which you computed the total ranks in Step 2

$$n_a = 6$$

Step 4 Determine the sample size for the other sample

$$n_b = 6$$

Step 5 Compute **U** using the following formula

$$U_a = (n_a * n_b) + \frac{n_a * (n_a + 1)}{2} - T_a = (6 * 6) + \frac{6 * (6 + 1)}{2} - 25 =$$

$$= 36 + \frac{42}{2} - 25 = 32$$

Step 6 Determine the value of **U** to compare to the U-critical value.

The value you will compare to the U-critical value will be **the smaller of** : U_a and $U_b = (n_a * n_b) - U_a$

For these data   Ua = 32   and U_b = $(n_a * n_b) - U_a$ = (6*6) − 32 = 4

We would use U_b = 4

For small samples (n < 20 for each group):

Step 7   Determine the critical values of **U** using the table.
The rows and columns represent sample size for each group (it doesn't matter which group is used for rows and which for the columns). Look down the appropriate column and across the appropriate row to find the critical value.

For the example data, you look down the column labeled n_1 = **6** and across the row labeled n_2 = **6**, and find the critical values of **5**.

Step 8 Compare the obtained U and the critical U values to determine whether to retain or reject the null hypothesis.

-- if the obtained U value (from Step 6) is larger than the critical U value, then retain H0:

-- if the obtained U value (from Step 6) is smaller than the critical U value, then reject H0:

For the example data, we would decide to reject the null hypothesis, because U=4
is smaller than the smaller critical value of 5.

For large samples (one or both groups with n > 20):
With samples this large, the value of U approaches a normal distribution, and so the null hypothesis can be tested by a Z-test.

Step 9   Compute the standard deviation of U

$$\text{Std dev} = \sqrt{\frac{(n_a * n_b) * (n_a + n_b + 1)}{12}} \quad \sqrt{\frac{(6 * 6) * (6 + 6 + 1)}{12}} = 6.24$$

Step 10 Compute z

$$Z = \frac{U - ((n_a * n_b)/2)}{\text{Std dev}} = \frac{32 - ((6*6)/2)}{6.24} = 2.24$$

Step 11 Compare the obtained Z value and the critical Z value to determine whether to retain or reject the null hypothesis.

-- if the **absolute value** of the obtained Z is less than 1.96, then retain H0:

-- if the **absolute value** of the obtained Z is greater than 1.96, then reject H0:

```
For the example data, we would decide to reject the null hypothesis, because the
absolute value of the obtained Z (2.24) is greater than the critical value of
1.96.
```

For both small and large samples:

Step 12 IF you reject the null hypothesis, determine whether the data support or do not support the research hypothesis.

-- IF you reject the null hypothesis AND the group that was hypothesized to have the larger scores does, then the research hypothesis is supported

-- IF you retain the null hypothesis OR you reject the null hypothesis BUT the group that was hypothesized to have the larger scores actually has the smaller scores, then the research hypothesis is not supported.

By the way: Usually the researcher hypothesizes that there is a difference between the conditions. Sometimes, however, the research hypothesis is that there is NO difference between the conditions. If so, the research hypothesis and H0: are the same! When this is the case, retaining H0: provides support for the research hypothesis, whereas rejecting H0: provides evidence that research hypothesis is incorrect.

```
      For the example data, we would decide that the research hypothesis is supported,
because the null hypothesis was rejected, and because, as hypothesized, the stores with
the separate reptile department tended to have the larger scores.
```

Step 13 Reporting the results

You will want to compute medians and IQR values to help describe the data before reporting the results of the significance test. As for the other statistical tests, the report includes the "wordy" part and the statistical values you used to make your statistical decision. If you reject H0:, be sure to describe how the groups differed, rather than just reporting that there was "a difference".

```
      Those stores without separate reptile departments displayed reptiles with a
median quality rating of 4 (IQR = 2-5.5), whereas those that did have separate
departments had a median rating of 7.5 (IQR = 6.25-9).  As hypothesized, pet stores
with separate reptile departments tended to have significantly higher ratings than
those which did not, U = 4, p < .05.
```

Critical Values for the Mann-Whitney U-Test

Level of significance: 5% (P = 0.05)

Size of the largest sample (n₂)

Size of the smallest sample (n₁)

n₁\n₂	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
3	0	1	1	2	2	3	3	4	4	5	5	6	6	7	7	8	8	9	9	10	10	11	11	12	13	13
4	1	2	3	4	4	5	6	7	8	9	10	11	11	12	13	14	15	16	17	17	18	19	20	21	22	23
5	2	3	5	6	7	8	9	11	12	13	14	15	17	18	19	20	22	23	24	25	27	28	29	30	32	33
6		5	6	8	10	11	13	14	16	17	19	21	22	24	25	27	29	30	32	33	35	37	38	40	42	43
7			8	10	12	14	16	18	20	22	24	26	28	30	32	34	36	38	40	42	44	46	48	50	52	54
8				13	15	17	19	22	24	26	29	31	34	36	38	41	43	45	48	50	53	55	57	60	62	65
9					17	20	23	26	28	31	34	37	39	42	45	48	50	53	56	59	62	64	67	70	73	76
10						23	26	29	33	36	39	42	45	48	52	55	58	61	64	67	71	74	77	80	83	87
11							30	33	37	40	44	47	51	55	58	62	65	69	73	76	80	83	87	90	94	98
12								37	41	45	49	53	57	61	65	69	73	77	81	85	89	93	97	101	105	109
13									45	50	54	59	63	67	72	76	80	85	89	94	98	102	107	111	116	120
14										55	59	64	67	74	78	83	88	93	98	102	107	112	118	122	127	131
15											64	70	75	80	85	90	96	101	106	111	117	122	125	132	138	143
16												75	81	86	92	98	103	109	115	120	126	132	138	143	149	154
17													87	93	99	105	111	117	123	129	135	141	147	154	160	166
18														99	106	112	119	125	132	138	145	151	158	164	171	177
19															113	119	126	133	140	147	154	161	168	175	182	189
20																127	134	141	149	156	163	171	178	186	193	200
21																	142	150	157	165	173	181	188	196	204	212
22																		158	166	174	182	191	199	207	215	223
23																			175	183	192	200	209	218	226	235
24																				192	201	210	219	228	238	247
25																					211	220	230	239	249	258
26																						230	240	250	260	270
27																							250	261	271	282
28																								272	282	293
29																									294	305
30																										317