Subchapter 11a.
The Mann-Whitney Test

Twenty-one persons seeking treatment for claustrophobia are independently and randomly sorted into two groups, the first of size $n_a = 11$ and the second of size $n_b = 10$. The members of the first group each individually receive Treatment A over a period of 15 weeks, while those of the second group receive Treatment B. The investigators’ directional hypothesis is that Treatment A will prove the more effective.

At the end of the experimental treatment period, the subjects are individually placed in a series of claustrophobia test situations, knowing that their reactions to these situations are being recorded on videotape. Subsequently three clinical experts, uninvolved in the experimental treatment and not knowing which subject received which treatment, independently view the videotapes and rate each subject according to the degree of claustrophobic tendency shown in the test situations. Each judge’s rating takes place along a 10-point scale, with 1="very low" and 10="very high"; and the final measure for each subject is the simple average of the ratings of the three judges for that subject. The following table shows the average ratings for each subject in each of the two groups.

<table>
<thead>
<tr>
<th>Group A</th>
<th>Group B</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.6</td>
<td>5.2</td>
</tr>
<tr>
<td>4.7</td>
<td>5.3</td>
</tr>
<tr>
<td>4.9</td>
<td>5.4</td>
</tr>
<tr>
<td>5.1</td>
<td>6.2</td>
</tr>
<tr>
<td>5.2</td>
<td>6.3</td>
</tr>
<tr>
<td>5.5</td>
<td>6.8</td>
</tr>
<tr>
<td>6.5</td>
<td>7.7</td>
</tr>
<tr>
<td>6.5</td>
<td>8.0</td>
</tr>
<tr>
<td></td>
<td>8.1</td>
</tr>
<tr>
<td>mean</td>
<td>5.6</td>
</tr>
<tr>
<td></td>
<td>6.5</td>
</tr>
</tbody>
</table>

The investigators expected Treatment A to prove the more effective, and sure enough it is group A that appears to show the lower mean level of claustrophobic tendency in the test situations. You might suppose that all they need do now is plug their data into an independent-samples $t$-test to see whether the observed mean difference is significant.

If you ever find yourself with a set of data of this sort, and are seized by the impulse to plug the numbers into a $t$-test, please lie down until the impulse passes. For the fact is that a $t$-test could not legitimately be used in this particular situation, nor in any other where the assumptions of the test are so patently violated. As indicated in the main body of Chapter 11, the $t$-test for independent samples can be meaningfully applied only insofar as

1. the two samples are independently and randomly drawn from the source population(s);
2. the scale of measurement for both samples has the properties of an equal interval scale; and
3. the source population(s) can be reasonably supposed to have a normal distribution.
While our initial sorting of the subjects into groups A and B satisfies the first requirement in the list, the second and third requirements are clearly not met. A rating scale cannot be assumed to possess the properties of an equal interval scale. Moreover, as you can see from the adjacent graph, there is nothing in the data to suggest an underlying normal distribution. The indication, indeed, is that the distribution of the underlying population might have a rather pronounced positive skew. (The graph shows the frequency distribution for the 21 subjects of both groups combined, sorted into the intervals 4.5 to 5.4, 5.5 to 6.4, 6.5 to 7.4, and 7.5 to 8.4.)

The admonition to refrain from using a t-test in situations of this sort is not simply a matter of good manners or good taste. The t-test makes these particular assumptions about the data you are feeding into it. Feed it the type of information it assumes it is getting, and the result it gives you in return will provide a firm basis for drawing rational conclusions. Feed it information that violates any one or several of its assumptions, and it will still patiently crunch the numbers and crank out a result. But that result will be nonsense, and so too will be any conclusion you might draw from it. The "significant" result that would come from applying a t-test to this set of data would in fact be no result at all. It would be merely the semblance of a result.

Assumption 2, an equal-interval scale of measurement; assumption 3, an underlying normal distribution. In cases where data from two independent samples fail to meet either or both of these requirements, one is well advised to forego the t-test and turn instead to its non-parametric alternative, the Mann-Whitney Test. The only assumptions of the Mann-Whitney test are

1. that the two samples are randomly and independently drawn;
2. that the dependent variable (e.g., claustrophobic tendency) is intrinsically continuous, capable in principle, if not in practice, of producing measures carried out to the nth decimal place; and
3. that the measures within the two samples have the properties of at least an ordinal scale of measurement, so that it is meaningful to speak of "greater than," "less than," and "equal to."

We will begin with the sheer mechanics of the Mann-Whitney procedure, and then come back to consider its underlying logic. Actually, there are two separate sequences in the logic of the Mann-Whitney test. The first, described here as Method I, is the more widely applicable. The second, Method II, is needed only when one or another of the samples is fairly tiny.

### Mechanics

It begins by assembling the measures from samples A and B into a single set of size \( N = n_a + n_b \). These measures are then rank-ordered from lowest (rank#1) to highest (rank#N), with tied ranks included where appropriate. For the present set of data, this process would yield the following result.
\[ N = n_a + n_b = 11 + 10 = 21 \]

**Tied Ranks.** Note that two of the raw-data entries each have the value of 5.2. As these two entries fall in the sequence where ranks #5 and #6 would be, in the absence of such a tie, they are each given the average of these two ranks, which is 5.5. Similarly for the two entries whose raw value is 6.5; each is accorded the average of ranks #15 and #16. If there were three raw-data entries tied for ranks 8, 9, and 10, each would receive the average of those ranks, which is 9. And so on.

Once they have been sorted out in this fashion, the rankings are then returned to the sample, A or B, to which they belong and substituted for the raw measures that gave rise to them. Thus, the raw measures that appear in the following table on the left are replaced by their respective ranks, as shown in the table on the right.
Our final bit of mechanics is just a dab of symbolic notation.

\[
\begin{align*}
T_A &= \text{the sum of the } n_a \text{ ranks in group A} \\
T_B &= \text{the sum of the } n_b \text{ ranks in group B} \\
T_{AB} &= \text{the sum of the } N \text{ ranks in groups A and B combined}
\end{align*}
\]

For the present example:

\[
\begin{align*}
T_A &= 96.5 \quad \text{[with } n_a = 11]\] \\
T_B &= 134.5 \quad \text{[with } n_b = 10]\] \\
T_{AB} &= 231 \quad \text{[with } N = 21]\]
\]

---

**Logic & Procedure: Method I**

The effect of replacing raw measures with ranks is two-fold. The first is that it brings us to focus only on the ordinal relationships among the raw measures—"greater than," "less than," and "equal to"—with no illusion or pretense that these raw measures (4.6, 4.7, 4.9, etc.) derive from an equal-interval scale. The second is that it transforms the data array into a kind of closed system, many of whose properties can then be known by dint of sheer logic.

For example: If you had a total of N=4 items ranked in this fashion, the sum of those ranks would inescapably be

\[T_{AB} = 1+2+3+4 = 10\]

This would also be true if some of the ranks were tied; for example:

\[T_{AB} = 1+2.5+2.5+4 = 10\]

Similarly for N=5 ranked items:

\[T_{AB} = 1+2+3+4+5 = 15\]

And for N=21 ranked items, as in the present example:

\[T_{AB} = 1+2+3+4+\cdots+18+19+20 = 231\]

In general, the sum of any set of N ranks will be equal to

\[
T_{AB} = \frac{N(N+1)}{2} \quad \text{for } N=4: \frac{(4\times5)}{2}=10 \\
\text{for } N=5: \frac{(5\times6)}{2}=15 \\
\text{for } N=21: \frac{(21\times22)}{2}=231
\]

and the average of the N ranks will of course be that sum divided by N. In any particular case
you could calculate the value of $T_{AB}$ in the manner shown above and then divide it by N, though a rather more direct way is to calculate it simply as $(N+1)/2$, which is what the mean of the N ranks boils down to, once you perform the requisite algebraic juggling:

$$\text{mean rank}_{AB} = \frac{N(N+1)}{2} \times \frac{1}{N} = \frac{N+1}{2} $$

<table>
<thead>
<tr>
<th>N</th>
<th>5/2</th>
<th>6/2</th>
<th>22/2</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>2.5</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td></td>
<td>11</td>
</tr>
<tr>
<td>21</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Either way, the null hypothesis in our example is that treatments A and B do not differ with respect to their effectiveness. If this were true, then the raw measures within samples A and B would be about the same, on balance, and the rankings that derive from them would be evenly mixed within samples A and B, like cards in a well shuffled deck.

So if the null hypothesis were true, we would expect the separate averages of the A ranks and the B ranks each to approximate this same overall mean value of $(N+1)/2=11$. This entails that the rank-sums of the two groups, $T_A$ and $T_B$, would approximate the values

$$T_A = n_a(N+1)/2 = 11(21+1)/2 = 121$$

and

$$T_B = n_b(N+1)/2 = 10(21+1)/2 = 110$$

Thus we know at the outset that

- the observed value of $T_A = 96.5$ belongs to a sampling distribution whose mean is equal to 121
- and
- the observed value of $T_B = 134.5$ belongs to a sampling distribution whose mean is equal to 110

Now and again in writings on subjects more or less mathematical, you will find an author saying "It can be shown that such-and-such is so." This typically occurs at junctures where the point to be made is not an obvious one; where it would be a distraction to stop and make it obvious; and where you must therefore ask the reader to take it on faith. This is one of those junctures. For any particular situation of the sort now under consideration, it can be shown that the sampling distributions for $T_A$ and $T_B$ both have the same variance and the same standard deviation. Here I will show only the structure for the standard deviation, which in the general case is equal to

$$\sigma_T = \sqrt{n_a n_b (N+1)}$$

Thus, for the present example, the sampling distributions of $T_A$ and $T_B$ each have a standard deviation equal to
\[ \sigma_T = \sqrt{\frac{(11)(10)(21+1)}{12}} = \pm 14.2 \]

Another assertion of the "it can be shown" variety is this. In situations where \( N \) ranks are distributed within two samples of sizes \( n_a \) and \( n_b \), it can be shown that the sampling distributions of \( T_A \) and \( T_B \) tend to approximate the form of a normal distribution, providing that \( n_a \) and \( n_b \) are both equal to or greater than 5. The source of this tendency is similar to what we saw in Chapter 6 when considering the tendency of binomial sampling distributions to approximate the normal.

If the preceding two steps have not been obvious, the next one surely will be. Given the mean and standard deviation of a normally distributed sampling distribution, you can then fold the observed value of either \( T_A \) or \( T_B \) into an appropriate version of a \( z \)-ratio and refer the result to the unit normal distribution. In this particular context the \( z \)-ratio must include a "\( \pm .5 \)" correction for continuity, just as with the binomial \( z \)-ratio, to accommodate the fact that the sampling distributions of \( T \) are intrinsically discrete. (\( T_A \) and \( T_B \) can assume decimal values such as 96.5 and 30.5 only as an artifact of the process of assigning tied ranks. Intrinsically, the ranks—1, 2, 3, 4, etc.—on which they are based are all integers.)

Designating \( T_{obs} \) as the observed value of either \( T_A \) or \( T_B \); \( \mu_T \) as the mean of the corresponding sampling distribution of \( T \); and \( \sigma_T \) as the standard deviation of that sampling distribution,

the general structure of the ratio is

\[
z = \frac{(T_{obs} - \mu_T) \pm .5}{\sigma_T} \quad \text{correction for continuity:} \quad \begin{cases} - .5 \text{ when } T_{obs} > \mu_T \\ + .5 \text{ when } T_{obs} < \mu_T \end{cases}
\]

Shown below are the details of calculation for both of our observed \( T \) values, \( T_A = 96.5 \) and \( T_B = 134.5 \). Keep in mind that the means of the sampling distributions for \( T_A \) and \( T_B \) are 121 and 110, respectively, and that the standard deviation for both distributions is \( \pm 14.2 \).

\[
z_A = \frac{(96.5 - 121) + .5}{14.2} = \frac{-24}{14.2} = -1.69
\]

\[
z_B = \frac{(134.5 - 110) - .5}{14.2} = \frac{+24}{14.2} = +1.69
\]

The remarkable symmetry of \( z_A \) and \( z_B \) is no accident. In all instances, \( z_A \) and \( z_B \) will have the same absolute value and opposite signs, one positive and the other negative. It
therefore makes no difference whether the test of statistical significance is performed with $z_A$ or $z_B$. The only requirement is that we keep a clear idea of what a positive or negative sign for $z$ means in any particular case.

In our example, the investigators begin with the directional hypothesis that Treatment A will prove the more effective. This entails that sample A will tend to have the smaller values among the raw measures (indicating lower levels of claustrophobic tendency), hence the smaller ranks, hence an observed value of $T_A$ smaller than its null-hypothesis value of 121, hence a value of $z_A$ with a negative sign. With the opposite directional hypothesis—that Treatment B will prove the more effective—the expectation would be for $z_B$ to end up with the negative sign. With a non-directional hypothesis there would of course be no specific expectation one way or the other.

Here, at any rate, is the bottom line. If the null hypothesis were true, the mere-chance probability of ending up with a value of $z_A$ falling at or beyond (to the left of) $z = -1.69$ within the normal distribution would be a scant $P = .0455$. For a directional test, our observed value of $z_A = -1.69$ is therefore significant a bit beyond the basic .05 level. (For a non-directional test it would of course be non-significant, since the mere-chance probability in this case would have to be figured as $P = .0455 + .0455 = .091$.)

![Critical Values of ±z](image)

**Critical Values of ±z**

<table>
<thead>
<tr>
<th>Level of Significance for a Directional Test</th>
<th>.05</th>
<th>.025</th>
<th>.01</th>
<th>.005</th>
<th>.0005</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-Directional Test</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$z_{critical}$</td>
<td>1.645</td>
<td>1.960</td>
<td>2.326</td>
<td>2.576</td>
<td>3.291</td>
</tr>
</tbody>
</table>

¶Logic & Procedure: Method II

The limitation of the simple procedure described above is that it is valid only when $n_a$ and $n_b$ are both equal to or greater than 5. This is probably not much of a limitation in actual practice, for one would rarely choose to rely on such tiny samples, except as a last resort. At any rate, if either sample is of a size smaller than 5, it is a different path that must be taken. Up to a point, it is also possible to take this second path even when the sample sizes are both equal to or greater than 5, as in our present example. In this event, the probability assessments arrived at via the two methods are essentially equivalent.
Here as well we begin with some things that can be known by dint of sheer logic. Continuing with our claustrophobia example, recall that $T_A = 96.5$ is the sum of the $n_a = 11$ ranks in sample A and that $T_B = 134.5$ is the sum of the $n_b = 10$ ranks in sample B.

This first point will be fairly obvious. The maximum possible value of $T_A$ in our example would be the sum of the highest $n_a = 11$ of the ranks:

$$11 + 12 + 13 + 14 + 15 + 16 + 17 + 18 + 19 + 20 + 21 = 176$$

Similarly, the maximum possible value of $T_B$ would be the sum of the highest $n_b = 10$ of the ranks:

$$12 + 13 + 14 + 15 + 16 + 17 + 18 + 19 + 20 + 21 = 165$$

For any particular combination of $n_a$ and $n_b$, these maximum possible values can be reached through the formulas

$$T_{A[\text{max}]} = n_a n_b + \frac{n_a (n_a + 1)}{2}$$

which for the present example comes out as

$$T_{A[\text{max}]} = (11)(10) + \frac{11(12)}{2} = 176$$

and

$$T_{B[\text{max}]} = n_a n_b + \frac{n_b (n_b + 1)}{2}$$

which comes out as

$$T_{B[\text{max}]} = (11)(10) + \frac{10(11)}{2} = 165$$

Method II of the Mann-Whitney test turns upon the calculation of a measure known as $U$, which for either sample is equal to the difference between the maximum possible value of $T$ for the sample versus the actually observed value of $T$. Thus for sample A it would be

$$U_A = T_{A[\text{max}]} - T_A$$

$$= n_a n_b + \frac{n_a (n_a + 1)}{2} - T_A$$
and for sample B

\[ U_B = T_{B[\text{max}]} - T_B \]

\[ = n_a n_b + \frac{n_b(n_b+1)}{2} - T_B \]

For the present example we have already obtained the maximum-possible and observed values for \(T_A\) and \(T_B\):

<table>
<thead>
<tr>
<th></th>
<th>maximum possible value</th>
<th>observed value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(T_A)</td>
<td>176</td>
<td>96.5</td>
</tr>
<tr>
<td>(T_B)</td>
<td>165</td>
<td>134.5</td>
</tr>
</tbody>
</table>

so we can easily calculate the value of \(U\) for this example as either

\[ U_A = 176 - 96.5 = 79.5 \]

or

\[ U_B = 165 - 134.5 = 30.5 \]

It does not matter which of these values you use, so long as you are consistent. The reason it does not matter is that \(U_A\) and \(U_B\) are mirror images of each other. For any given values of \(n_a\) and \(n_b\), the sum of \(U_A\) and \(U_B\) will always be equal to the product of \(n_a\) times \(n_b\). Hence the following identities:

\[ U_A + U_B = n_a n_b \]

\[ U_A = n_a n_b - U_B \]

\[ U_B = n_a n_b - U_A \]

The upshot is that once you know one of the values of \(U\), you also implicitly know the other. Note as well that \(U_A\) and \(U_B\) are inversely related. The larger the value of one, the smaller must be the value of the other.

We determined earlier that the values of \(T_A\) and \(T_B\) that would be expected on the null hypothesis would be \(T_A = 121\) (versus the observed \(T_A = 96.5\)) and \(T_B = 110\) (versus the observed \(T_B = 134.5\)). From this you can infer that the values of \(U\) to be expected under the null hypothesis are

\[ U_A = 176 - 121 = 55 \quad \text{[vs observed } U_A = 79.5\text{]} \]

and

\[ U_B = 165 - 110 = 55 \quad \text{[vs observed } U_B = 30.5\text{]} \]

More generally, the null-hypothesis values of \(U\) are given by the identity
\[ U_A = U_B = \frac{(n_a n_b)}{2} \]

which for our example would work out as
\[ U_A = U_B = \frac{(11)(10)}{2} = 55 \]

On the null hypothesis we would therefore expect the values of \( U_A \) and \( U_B \) both to approximate \( \frac{(n_a n_b)}{2}=55 \). So if the null hypothesis is true, if there really is no difference in effectiveness between Treatment A and Treatment B, how likely is it that we could end up, by mere chance, with an observed value of \( U_A \) as large as 79.5; or alternatively, how likely that we could end up with an observed value of \( U_B \) as small as 30.5?

The principle underlying the answer to this question is but one more verse of a by-now familiar song. First you figure out the total number of possible ways in which \( N=21 \) ranks can be combined within two groups of sizes \( n_a = 11 \) and \( n_b = 10 \); and then you figure out the proportion of such combinations that would produce a value of \( U_A \) as large as 79.5, or alternatively a value of \( U_B \) as small as 30.5.

When I tell you that the total number of possible combinations in this example is
\[ \frac{N!}{n_a!n_b!} = \frac{21!}{11!10!} = 352,716 \]
you will see in a flash that here is yet another instance of "easier said than done." Fortunately, it is also another instance where the bedeviling details have already been worked out. In most hard-copy statistics textbooks you will find a multi-page table of critical values of \( U \), applicable to cases where \( n_a \) and \( n_b \) are both equal to or less than 20.

I cannot show you the full scope of this table, because it is under copyright and I would have to pay a fee to reproduce it. Fortunately the underlying principles cannot be held in copyright, so what I have done instead is program the following table to calculate some of the critical values of \( U \) directly. Enter any particular values of \( n_a \) and \( n_b \) into the designated cells, click the "Calculate" button, and the corresponding critical values of \( U \) will appear in the table. The only restriction is that \( n_a \) and \( n_b \) must both be between 5 and 20, inclusive. (For cases where the size of either sample is smaller than 5, you will need to consult a table in the back of some hard-copy textbook.) If the two sample-size values are different, you should designate the larger as \( n_a \) and the smaller as \( n_b \).

Before you start plugging any new numbers, however, please be sure to read the explanatory text that follows the table.

[Interactive portion omitted from this print file.]

As you can see, I have already inserted our current values of \( n_a = 11 \) and \( n_b = 10 \), along with the corresponding critical values of \( U \). As illustrated in the next table, each of the pairs of
numbers designated "lower limit" and "upper limit" mark the boundaries between which falls a certain proportion of all possible mere-chance combinations of $U_A$ and $U_B$ in the case where $n_a=11$ and $n_b=10$.

| lower limit | 31 | 26 | 21 |
| upper limit | 79 | 84 | 89 |

| .90 | .95 | .98 | proportion between lower and upper limits |
| .10 | .05 | .02 | proportion outside middle range; both "tails" combined |
| .05 | .025 | .01 | proportion outside middle range at either end; each "tail" considered separately |

For the present example our focus can be chiefly on the first pair, 31 and 79. If the null hypothesis is true, there is a 90% chance that $U_A$ and $U_B$ will both be larger than 31 and both smaller than 79. The remaining possibilities fall outside that range, 5% below it and 5% above. At the one end, there is a 5% chance that $U_A$ might end up as large as 79 and $U_B$ as small as 31. At the other, there is a 5% chance that $U_B$ might end up as large as 79 and $U_A$ as small as 31. In our example, the investigators begin with the directional hypothesis that Treatment A will prove the more effective. Specifically, what they expect is

- that the raw-measure ratings of claustrophobic tendency will be lower in sample A than in sample B; and
- that sample A will accordingly tend to have the lower ranks; hence $T_B > T_A$ and $U_A > U_B$

And that is exactly what they do find: $U_A = 79.5$, $U_B = 30.5$. As these observed values of $U$ fall just outside the range bounded by 31 and 79, we can say that the result is significant just beyond the .05 level for a directional test. For a non-directional test the mere-chance probability of the obtained result would be only a shade below 10%, hence non-significant. (Note that this is the same conclusion we reached earlier with Method I.)

Similarly, the second set of numbers in the table, 26 and 84, mark the boundaries between which fall 95% of all possible mere-chance outcomes. Observed values of $U$ falling at or beyond these boundary points would therefore be significant at or beyond the .05 level for a non-directional test and at or beyond the .025 level for a directional test. And so on for the third set of numbers in the table, marking the lower and upper boundaries for the directional .01 and the non-directional .02 levels of significance.

The VassarStats web site has a page that will perform all steps of the Mann-Whitney test, including the initial rank-ordering of the raw measures.