Supplement on paired tests:

There are two other simple tests available for paired data:

The sign test.

- this works under almost any circumstances, but the power is not terribly great.

Here’s how it works:

- get the sign of the difference in your paired samples.
- count up the number of positive and number of negative signs
- the higher number is your test statistic, $B^*$ (it’s “$B$” because the distribution we use to carry out our test is the binomial)
- compare our result with the $B$ listed in table 7 (or simply calculate out the binomial probability!)


- researcher looked at the number of times each subspecies of Junco was dominant in a 45 minute period (birds were evenly matched as to size):

- develop our hypothesis:

  - $H_0$: there is no difference in dominance between the two subspecies.
  - $H_1$: there is a difference in dominance.

(What $H_0$ implies is that $p = .5$, see below)

# of times dominant:

<table>
<thead>
<tr>
<th>northern</th>
<th>southern</th>
<th>sign</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>9</td>
<td>-</td>
</tr>
<tr>
<td>0</td>
<td>6</td>
<td>-</td>
</tr>
<tr>
<td>0</td>
<td>22</td>
<td>-</td>
</tr>
<tr>
<td>2</td>
<td>16</td>
<td>-</td>
</tr>
<tr>
<td>0</td>
<td>17</td>
<td>-</td>
</tr>
<tr>
<td>2</td>
<td>33</td>
<td>-</td>
</tr>
<tr>
<td>1</td>
<td>24</td>
<td>-</td>
</tr>
<tr>
<td>0</td>
<td>40</td>
<td>-</td>
</tr>
</tbody>
</table>

So, we have 8 (-)’s, and 0 (+)’s, and our $B^* = 8$
Going into table 7, we have:

\[ B^* \geq B_{\text{table}} = 8, \text{ so we reject and conclude there is that the southern species is more dominant (note that we did not use a one-sided alternative, but if the difference is significant, we are certainly allowed to look at the direction of the difference).} \]

- see also example 9.11, p. 365 [9.12, p. 364] [8.11, p. 315].

Some theory for the sign test:

If the null hypothesis of no difference is true, what’s the probability of being dominant or submissive in the above example?

1/2, or .5, so if \( H_0 \) is true, then \( p = .5 \)

How many trials do we have above? 8 trials.

What outcome did we observe? The southern subspecies was dominant eight times.

So what is the probability of the southern species being dominant 8 times in 8 trials if \( p = .5 \)?

Binomial! (what is the probability of 8 heads in 8 tosses - exactly the same!)

And we know how to do this:

\[
\binom{8}{8} .5^8 .5^0 = .003906
\]

We still need to double this. Why?

- because we used a two sided hypothesis, so we would have rejected for either a lot of (-)’s or a lot of (+)’s.
- so we need to add the probability of 8 (-)’s to that of 8 (+)’s.
- so 2 x .003906 = .0078, which is still less than \( \alpha = .01 \), so we reject.

**Important:** Suppose we had gotten 7 (-)’s?

Then our \( B^* = 7 \), and we would have “failed to reject”.
What is our probability now?

\[
\binom{8}{8}.5^8.5^0 + \binom{8}{7}.5^7.5^1 = .003906 + .03125 = 0.0352
\]

We need to add the probability of our outcome plus the probability of a “worse” outcome (in this case, if we’d gotten \(B^* = 7\), the worse outcome would have been \(B^* = 8\)).

Now of course, we need to double this: \(2 \times .0352 = .0703\) (a slight rounding error here).

And since .0703 is greater than .01, we’d fail to reject (we’d fail to reject even for \(\alpha = .05\), but would be able to reject for \(\alpha = .1\); this does agree with table 7).

Comment: we had to double all our calculated probabilities above because we were interested in both lots of (+)’s and lots of (-)’s. But suppose we were only interested in one or the other. For example, we have an idea about our birds above and know that the southern subspecies is more aggressive. Would it make sense that we wouldn’t be interested in (+)’s?

For example, suppose we did this:

- \(H_0\): there is no difference in dominance between the two subspecies
- \(H_1\): The southern subspecies is dominant

What changes?

- \(\alpha = .01\)

- calculate \(B^*\) as before (so \(B^* = 8\))

- compare to \(B\) in table 7 or calculate probability.

- probability is now .0039 (no doubling). We’re not interested in those outcomes where the northern subspecies was dominant (so we don’t need to calculate the probability that the northern subspecies was more dominant 8 times).

- and we reject again and conclude that the southern subspecies is dominant.
- why the big fuss? Well, suppose that like before, we’d gotten \( B^* = 7 \). If we had picked \( \alpha = .05 \) what happens?

- we get to reject!! \( (p = .03906 + .03125 = .0352, \text{ but we don’t have to double this, so } p\text{-value} < \alpha \text{ and we can reject}) \).

- before, with this same example, we would not have been able to reject.

- this is our first example of a one sided test. Note that one-sided tests have more power! This, of course, leads directly into the next set of notes.

The Wilcoxon signed rank test (essentially, a Mann-Whitney U test for paired data):

(Note: this is not covered in the 2\(^{nd} \) edition. If you have the 2\(^{nd} \) edition, you may want to look at someone else’s text (see section \([9.5] / [8.5] \)).

The advantage of this test (say, over the sign test), is that it has a lot more power.

Of course, the point behind the test is that it can be used with non normal data (if your differences are not normal).

Here's an outline of the test (obviously your data need to be paired):

1) calculate the differences between your samples.

2) get the absolute value of each of your differences.

3) Now rank these values (the smallest absolute difference gets “1”, the second smallest absolute difference gets “2”, etc.).

4) Now get rid of the absolute values, and give your ranks the appropriate sign (i.e., (+) or (-)).

5) Add up all the (+) ranks and all the (-) ranks (Important: use the absolute values of your ranks for calculating the sums).

6) \( W^* \) = the larger of the two sums you got in step (5)

7) Get \( W_{\text{table}} \) from table 8 in your text and make the usual comparison:

\[
\text{If } W^* \geq W_{\text{table}}, \text{ then reject } H_0, \text{ otherwise fail to reject.}
\]

Let's try this with example 9.17, p. 372 \([8.5.1, p. 322]\), in your text.
Some comments:

The test can be used to test for equivalence in means (since it assumes symmetry).

As such, our $H_0$ would be: $\mu_1 = \mu_2$

and of course, our $H_1$ would reflect $H_0$, as usual

Ignore 0's in your differences. They simply don't count.

If some of your differences are identical (so you get identical ranks),
average the ranks for these differences. Note that you don't want too many
ties or the test will not work that well (loose power).

The test does assume the distribution of the differences is symmetric (i.e.,
don't use it on highly skewed data).

Without going into the details of the theory, consider:

If our null hypothesis is true, in other words, if $\mu_1 = \mu_2$, then one
would expect negative and positive ranks to occur at about the
same rate.

Or, another way of saying it, the sum of our negative ranks
should be about the same as the sum of our positive ranks.

If our null hypothesis is not true, then one or the other of these
should be a lot larger.

The table in the back of your book lists the values for
various sample sizes at which one or the other of these
sums is so much larger that it's doubtful the observed
differences are due to chance.

We use simple probability rules to calculate the values in
the table. But it's not necessary for us to go into the details.

(If you're really curious, suppose $n = 8$, and all our
differences are negative. Then the sum of the
positive ranks is 0 (and the sum of all the negative
ranks would be $36 = W^*$).

Since the null hypothesis implies positive
differences are as common as negative differences,
the probability of this happening is $(1/2)^8 = .0039,$
and we would reject (you can confirm this looking
at table 8). The rest of the probabilities would be calculated in a similar way)

Finally, there are several variations on the Wilcoxon signed rank test that can deal with larger sample sizes (e.g., that are not covered by the table in your book). These are actually based on the normal approximation

Let's do one more example, exercise 9.31, p. 376:

We're interested in evaluating the effectiveness of a weight loss drug (mCPP). A group of 9 men was first measured while on the drug, then while on a placebo. We want to know if the weight change while on the drug was more effective.

Being a bit lazy, let's just do: $H_0: \mu_1 = \mu_2$, and $H_1: \mu_1 \neq \mu_2$

The book says to use $\alpha = 0.05$

The book does most of the work for us, and gives us the differences:

| difference | |difference| | rank | signed rank |
|------------|-----|-----------|-------|----------|
| 1.1        | 1.1 | 6         | 6     |
| -1.6       | 1.6 | 7         | -7    |
| -2.1       | 2.1 | 8         | -8    |
| -0.3       | 0.3 | 1         | -1    |
| -0.6       | 0.6 | 3         | -3    |
| -2.2       | 2.2 | 9         | -9    |
| 0.9        | 0.9 | 5         | 5     |
| 0.7        | 0.7 | 4         | 4     |
| -0.4       | 0.4 | 2         | -2    |

And we add up the negative ranks to get 30, and positive ranks to get 15.

So, $W^* = 30$, and with $\alpha = 0.05$, $W_{table} = 40$, so we fail to reject and conclude that we do not have enough evidence to show a difference between the drug and a placebo.