Dealing with the assumption of independence between samples - introducing the paired design.

a) Suppose you deliberately collect one sample and measure something.

Then you collect another sample in such a way that the first item in the second sample is as much like the first item in the first sample as possible. In other words, you're controlling for variation between samples, or "pairing" your data.

An example. You're interested in the effect of some medicine on mice. You put together 10 pairs of mice, and keep each pair in a separate cage. One member of each pair gets the medicine, the other doesn't.

Why not just put 10 mice in one cage, 10 in another?

Suppose the room were the mice are kept doesn't have very good temperature control, or that the lab person doesn't pay attention when feeding and watering the mice.

Your results might show the effect of temperature, feeding or watering NOT the effect of medicine.

By putting two mice in each cage and giving one medicine, the other not, you are eliminating the variation caused by temperature, feeding, watering, etc.

b) This is called a "paired design". You pair things in your two groups to get rid of as much variation as possible, leaving only the change caused by the effect you're interested in (medicine, in the case of our mice).

i) a classic example - take 20 people. Measure their blood pressure.

- give them blood pressure medication.

- measure their blood pressure again.

- any change would have to be caused by the medication (though to be sure, you might do a control as well (another 20 people in the same conditions, whom you then give a placebo)).

c) So we now know a little about the paired *t*-test. How do you actually do it?

i) Same as usual. Figure out your H_0, H_1, then decide on α, etc. etc.

ii) But when we calculate t^*, we now need to do things differently yet again! (We already have three formulas for t^*, now we're getting a fourth!) Here's t^*:

$$t_s = t^* = \frac{\bar{d} - 0}{SE_{\bar{d}}}$$

where d = *the (sample) difference in values between each pair,*

$$\text{and } SE_{\bar{d}} = \frac{s_d}{\sqrt{n_d}}$$

obviously, *n* is the same for both samples (you can't have different sample sizes!).

d) This isn't too bad, and you've actually calculated mean differences and standard deviations a couple of times on various homework problems.

e) Incidentally, note the similarities to the one sample *t*-test. It's not a coincidence.

f) But let's do an example, and then talk a little more about this.

i) example 9.1, page 180

Setup: We're comparing the hindleg with the foreleg of deer. It makes sense to use a paired setup.

A big deer is more likely to have a large hindleg and a large foreleg.

First we set up our hypotheses:

H_0: $\mu_1 = \mu_2$

H_1: $\mu_1 \neq \mu_2$

Decide on α (the problem says to use .05)

10 deer were measured with the following results:

hindleg	foreleg	difference
142	138	4
140	136	4
144	147	-3
144	139	5
142	143	-1
146	141	5
149	143	6
150	145	5
142	136	6
148	146	2

For these data we calculate the following:

$n = 10$ $\qquad$ $\bar{y} = \bar{d} = 3.3$

$s_d^2 = 9.3444$ $\qquad$ $s_{\bar{d}} = 0.97$

$v = d.f. = 9$ $\qquad$ $t^* = \dfrac{3.3}{0.97} = 3.402$

and since $t_{0.05,9} = 2.262$, *we reject* H_0.

(incidentally, p = 0.007703)

We conclude that the fore and hind legs of deer are not the same size.

g) To do it in R, we would simply read in the above data as usual, and then do:

```
t.test(hind,fore,paired = TRUE)
```

As usual, we can modify this with "`greater`" or "`less`" etc.

h) Figuring out if data are paired is sometimes tricky. Usually one knows because the experiment was ***designed*** that way.

A classic example is before and after (using the same person) as in the example above.

Take a closer look at the example above, as well as 9.2 in your text.

Is there a deliberate connection between the first item in sample one and the first item in sample two? If so, the data are probably paired.

If you're just comparing two groups and not trying to match up members of each group, it's almost always just a regular "un"-paired t-test.

i) Concluding remarks on the paired *t*-test.

i) the paired t-test deliberately violates the assumption of independence in between our samples.

A regular *t*-test is not valid anymore, and can give you misleading results

A paired test can have a lot more power than an unpaired t-test.

As usual, the the bottom line is power. You want to use the most powerful test you can. If the data are paired, use a paired *t*-test. If they're not paired, then the paired test is invalid, so use the regular *t*-test.

j) What about the assumptions of the paired t-test?

The d's are random & normally distributed. If sample size is large, or if things are only approximately normal, this test is still okay.

k) There is an equivalent Mann-Whitney type test if the data are not normal. It's next on the agenda.

There are two other simple tests available for paired data:

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

The advantage of this test (say, over the sign test, covered next), 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 (just like in the paired t-test).

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 smaller of the two sums you got in step (5)

Your text calls these sums T_+ and T_- (i.e., $W^* = \min(T_+, T_-)$)

7) Get W_{table} from table B.12, p. 758 in your text and make the usual comparison:

If $W^* \leq W_{table}$, then reject H_0, otherwise fail to reject.

Note that this is backwards from our usual comparison. However it is still true that if $p \leq \alpha$ that you reject (that's always true).

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 (your sample size goes down for each 0).

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 differences).

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 smaller 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.

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 an example from a different text (see also example 9.4 on page 185):

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.

If μ_1 = weight change on the placebo, we'd have:

H_0: $\mu_1 = \mu_2$, and H_1: $\mu_1 < \mu_2$

(or if we want to write it in terms of differences:

H_0: $\mu_d = 0$ and H_1: $\mu_d < 0$ (a negative value means the weight change while the drug was more effective)

The text says to use $\alpha = 0.05$, and then does most of the work for us:

mCpp	Placebo	difference	|difference|	rank	signed rank
0.0	-1.1	1.1	1.1	6	6
-1.1	0.5	-1.6	1.6	7	-7
-1.6	0.5	-2.1	2.1	8	-8
-0.3	0.0	-0.3	0.3	1	-1
-1.1	-0.5	-0.6	0.6	3	-3
-0.9	1.3	-2.2	2.2	9	-9
-0.5	-1.4	0.9	0.9	5	5
0.7	0.0	0.7	0.7	4	4
-1.2	-0.8	-0.4	0.4	2	-2

And we add up the negative ranks to get $T_- = 30$, and positive ranks to get $T_+ = 15$.

So, $W^* = \min(30,15) = 15$, and with $\alpha = 0.05$, $W_{table} = 8$ (one sided!), so we fail to reject and conclude that we do not have enough evidence to show a difference between the drug and a placebo.

If you want to use R (using the deer leg example), just do:

```
wilcox.test(hind,fore,paired = TRUE)
```

The sign test.

The advantage of the sign test is that it can work under almost any circumstances

The disadvantage is poor power.

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 value of B listed in the handout (you could use table B.26b, but this is much easier).

Let's do an example, again from a different text:

A 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 our H_0 implies is that p = 0.5, see below)
>
> # of times dominant:

northern	southern	sign
0	9	-
0	6	-
0	22	-
2	16	-
0	17	-
2	33	-
1	24	-
0	40	-

So, we have 8 (-)'s, and 0 (+)'s, and our B* = 8

Going into the table on the handout, we have:

$B^* \geq B_{table} = 8$, 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 24.10, p. 538, but note that your text calculates the probabilities directly, which is a bit of a pain in this case.

You can also use R for this. In this case, you calculate binomial probabilites using the binom.test function:

```
binom.test(8,8)
```

(which gives us p = 0.007812, so we reject)

This will give you the p-value for our birds above.

If we try this with the deer data (example 24.10), we would have:

```
binom.test(8,10)
```

(8 "successes" in 10 measures)

And we get the same value as in the text (p = 0.1094)

Some theory for the sign test and more on one sided tests:

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 x .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!! ($p = .03906 + .03125 = .0352$, but we don't have to double this, so p-value $< \alpha$ and we can reject).

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

- this is another illustration of why one sided tests are good - they give us more power.