What to do if your assumptions fail

I. Review - assumptions of the t-test:

1) equal variances - usually you don’t even bother to check this; unless you have a good reason to think the variances are equal just use the t-test for unpooled variances.

2) data (population) are normally distributed.

3) data are collected randomly, both (a) between groups and (b) within groups.

II. Dealing with assumption 2.

1) Remember, if your n is reasonably large, don’t bother doing anything. Your $\bar{y}$ will have a distribution that’s more and more normal.
   
a) large? Remember between 20 and 30.

2) But suppose you only have small sample sizes, and your data is not normal. Then what???

3) The Mann-Whitney $U$-test
   
a) The only assumption for this test is # 3 above. You still need to make sure you collected the data randomly.

   b) What are you testing? That the two distributions are the same!
      
      - if you are willing to assume that the distributions are only different in location, but not shape (e.g., two identical binomials, two identical uniforms, etc.), then you can actually use this to test means (see below).

   c) Your hypotheses:

      $H_0$: The populations distributions of $Y_1$ and $Y_2$ are the same (note capitalized $Y$’s).

      $H_1$: The population distributions of $Y_1$ and $Y_2$ are not the same.

      OR

      $H_1$: the distribution of $Y_2$ tends to be greater than that of $Y_1$.

      $H_1$: the opposite of $H_2$.

      (as usual, let's not worry about these for now)

   d) However, we'll make the assumption of equal distributions except for location and use the following hypotheses:
\[ H_0: \mu_1 = \mu_2 \]
\[ H_1: \mu_1 \neq \mu_2 \]

d) Now figure out your \( \alpha \), just like before.

e) Calculate your test statistic. As you might guess, this is called “\( U \)”, not “\( t \):

   i) rank your data, from smallest value to largest.

   ii) for each data point, write down the number of values that are smaller IN THE OTHER SAMPLE next to it (use \( \frac{1}{2} \) if a value is perfectly tied).

   iii) Add up these numbers for both samples

   iv) Choose the larger of these two sums, that’s our \( U^* \).

f) check your work - the two sums should add up to the product of the sample sizes (see the example below).

g) compare your \( U^* \) with the tabulated value of \( U \) (table 6 in your text (or the table printout)):

   i) \( n = \) larger sample size,
   \( n' = \) smaller sample size

   ii) \( \alpha \) as before.

   iii) then, if \( U^* \geq \) tabulated \( U \), reject \( H_0 \).

   otherwise, “fail to reject” \( H_0 \).

h) let’s do an example. Exercise 7.62a, p. 301

   i) researchers measured dopamine levels in the brains of six control rats and in six rats exposed to toluene. The concentrations were:

<table>
<thead>
<tr>
<th>Toluene</th>
<th>Control</th>
</tr>
</thead>
<tbody>
<tr>
<td>3420</td>
<td>1820</td>
</tr>
<tr>
<td>2314</td>
<td>1843</td>
</tr>
<tr>
<td>1911</td>
<td>1397</td>
</tr>
<tr>
<td>2464</td>
<td>1803</td>
</tr>
<tr>
<td>2781</td>
<td>2539</td>
</tr>
<tr>
<td>2803</td>
<td>1990</td>
</tr>
</tbody>
</table>

   ii) q-q plots show the data are seriously not normal (not shown), so we can’t do a t-test (you should try this on your own to verify this).
iii) So we need to use the MWU test. Let's do:

\[ H_0: \text{mean dopamine for toluene is the same as the mean dopamine for the control} \]

\[ H_1: \text{mean dopamine for toluene is NOT the same as the mean dopamine for the control} \]

let’s pick \( \alpha = 0.05 \)

iv) Now we rank the data:

<table>
<thead>
<tr>
<th>( Y_1 )</th>
<th>( Y_2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1397</td>
<td>0</td>
</tr>
<tr>
<td>1803</td>
<td>0</td>
</tr>
<tr>
<td>1820</td>
<td>0</td>
</tr>
<tr>
<td>1843</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>1911</td>
</tr>
<tr>
<td>1990</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>2314</td>
</tr>
<tr>
<td>5</td>
<td>2464</td>
</tr>
<tr>
<td></td>
<td>2539</td>
</tr>
<tr>
<td>6</td>
<td>2781</td>
</tr>
<tr>
<td>6</td>
<td>2803</td>
</tr>
<tr>
<td>6</td>
<td>3420</td>
</tr>
</tbody>
</table>

(this is arranged just a little differently than in your book, but I think it’s a little easier to understand).

v) Now add to get \( K_1 \) and \( K_2 \):

\[ K_1 = 32 \quad K_2 = 4 \]

(check: \( 32 + 4 = 36 = 6 \times 6 \))

vi) Pick the larger of \( K_1 \) and \( K_2 \) and set equal to \( U^* \):

\[ U^* = 32 \]

vii) Now look this up in the tables:

\( n = 6, n' = 6, \alpha = 0.05 \)

So we get: \( U_{\text{table}} = 31 \), and since:

\[ U^* > U_{\text{table}}, \] we reject \( H_0 \) and conclude that we have a difference here.
viii) Note: Some people (and several computer packages) will call this the Wilcoxon rank sum test. It's mathematically equivalent.

i) So, why does it work (or, “let’s do a little theory”)?

You’re book has a pretty good outline of this (see p. 298 and 299 [p. 293 and 294] [286 and 287] [295 and 296]):

Suppose $n_1 = 5$ and $n_2 = 4$, then $K_1 + K_2$ must be 20 ($5 \times 4 = 20$).

So, suppose there is no overlap at all (the data is least compatible with $H_0$).

- this implies that either $K_1$ or $K_2$ = 20 (illustrate - see fig. 7.30a [7.30a] {7.10.3a}, and therefore $U = 20$.

If the data overlap perfectly, then $K_1 = K_2 = 10$, and $U = 10$. (fig. 7.30c [7.30c] {7.10.3c}).

This implies that large values of $U$ are incompatible with $H_0$, and this is the basis of the test.

So where does $U$ come from?? Better, how is $U$ distributed?

- remember, $t$ has a $t$-distribution, $z$ a normal, etc.

- Note that $U$ is discrete. We can calculate the probabilities associated with any given arrangement of the two samples. As your text points out, for example (see appendix 7.2 [7.2] {7.2} [missing in 5th edition] if you’re curious about the exact method of calculating this):

$$\Pr\{K_1 = 0 \text{ and } K_2 = 20\} = .008$$

So we just go through and figure out the probabilities for all possible combinations of $K_1$ and $K_2$, for as many different values of $n_1$ and $n_2$ that we can (which is why many tables stop at $n_1 = n_2 = 20$).

Another note. $\Pr\{K_1 = 0 \text{ and } K_2 = 20\} = \Pr\{K_1 = 20 \text{ and } K_2 = 0\} = .008$

therefore, the probability of either of these events happening is:

$.008 + .008 = .016$

We were using a two tailed test, so for a two tailed test we just use twice the probability of a one tailed test (you can see this real easily in the table).

(look also at figs. 7.30, 7.31, and p. 299 [7.30, 7.31, and p. 294] [7.10.3].
Finally, $U$, in this example, can go from 10 to 20. So how do we actually put the distribution together? Well, we just plot the probabilities (calculated as in the example above, for each value of $U$ from 10 to 20 (as in fig. 7.31 [7.31] {7.10.4} {7.10.4})).

But the tables, of course, only give the probabilities for $\alpha = .05, .01, etc.$

Almost, but not quite finished. The probability of $U = 20$ = .016. But what if we choose $\alpha = .01$??

- we loose! $U$ is discrete, and can’t take on values greater than 20 (if $n = 5$ and $n' = 4$).

- this also points out that the table gives you the value you need to reject at a certain level.

- For example, the probability that $U = 18 = .064$ (see text). Your table lists the associated probability as .10, not .05. $U=18$ isn’t good enough for $\alpha = .05$.

4) The Mann-Whitney test versus the t-test.

i) You now have two tests that do the same kind of thing - test to see if two things are equal.

- which is better? Again, this depends on the power of the test.

- if the data are normal, the t-test IS better.

- if the data are not normal, and $n$ is not very big, the Mann-Whitney $U$-test is better. Sometimes, particularly if you have very long tails, it is much better.

- so how do you decide? Follow the guidelines given above. If $n_1$ and $n_2$ are both less than about 20 to 30, and the data are not normal, use this test. Otherwise, you can probably use the t-test. Of course the less normal your data are, the bigger the $n$'s should be.

- if you can’t check your assumptions, or forget in a couple of years what to do to make sure the t-test works, use this one.

It’s almost always valid, unless the data are not random. Also, even if the data are normal, it still has pretty good power, though, of course, not as good as the t-test.
5) An assumption of the Mann-Whitney U-test that's not obvious.

Just like the t-test, the Mann-Whitney U-test is designed for continuous data. It may not do well with discrete data. In particular, it doesn't like ties. Still, if there are not too many ties, it'll do okay. Here's how to deal with ties:

- when you have data that are tied, you use ½ for EACH of the data points in the other sample that are tied with the data point you're looking at. Probably the easiest way to describe this with an example or two:

- Example 1:

<table>
<thead>
<tr>
<th>Y₁</th>
<th>Y₂</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>2.5</td>
</tr>
<tr>
<td>5</td>
<td>2.5</td>
</tr>
<tr>
<td>5</td>
<td>2.5</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>6</td>
<td>8</td>
</tr>
</tbody>
</table>

| Sum | 22 = K₁ | 14 = K₂ |

- Note that the 5 in sample 1 (Y₁) gets a value of 3. this is because it is tied with four 5's in sample 2 (4 x 1/2), and then there’s the number 3 that’s less than 5, so 1 + (4 x 1/2) = 3. Incidentally, the check still works (n₁ x n₂ = 6 x 6 = 36, and K₁ + K₂ = 36)

II. Dealing with independence between columns assumption

Remember, the data are random between the samples (being in one sample doesn’t affect the probability of being in the other sample)

So can we violate this assumption? As usual, it depends...

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 $\alpha$, 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 $t^*$:

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

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

and $SE_d = \frac{s_d}{\sqrt{n_d}}$

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.


Setup: Oocytes (egg cells) from four frogs where divided into two batches (two batches from each frog, here’s the pairing).

One group of oocytes was exposed to progesterone (a hormone), the other not.

The levels of cAMP (a substance that reacts to hormones) were then measured.

First we set up our hypotheses:

$H_0: \mu_1 = \mu_2$

$H_1: \mu_1 \neq \mu_2$

Decide on $\alpha$ (the problem says to use .10)

then we calculate $t^*$ as follows:

$d = 0.68$

$s_d = 0.40$, which $\Rightarrow s_d = \frac{0.40}{\sqrt{4}} = 0.20$

$so:$

$t^* = \frac{0.68}{0.20} = 3.4$

and if we look up $t$ with $d.f. = 3$ (we use just $n-1$ (all $n$’s here are the same)) and $\alpha = .10$, we get 2.353 (remember that the table gives one-tailed probabilities, so we need to use $\alpha = .05$ to put .10 in both tails).

because $|t^*| >$ tabulated $t$, we reject $H_0$.

we can conclude that progesterone does affect production of cAMP

(our units were pmol/oocyte)

(note: the 5th edition uses $\bar{d} = 0.675$).
g) 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 examples above.

- 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.

4) Concluding remarks on the paired $t$-test.

a) 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 (example 9.5 on page 352 [9.5, p. 351] [8.2.4, p. 303] [8.2.1, p. 308] for which a regular $t$-test “fails to reject” while a paired t-test rejects).

- again, 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.

b) a final note on assumptions of the paired t-test:

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

c) there is an equivalent Mann-Whitney type test if the data are not normal, but we don’t have time to cover it. See [9.5] in the 3rd edition or [8.5] [8.5] in the 4th or 5th editions if you’re interested (sorry, this isn’t in the 2nd edition). It's called the Wilcoxon rank sum test.

- however, the sign test that we discussed earlier does allow for non-normal paired data, though it's not very powerful.