A 't-test' – to pair or not to pair, that is the question

In a newspaper article we read that people shrink during the day. This is caused by gravity and the way our backbones are designed. Suppose now that we want to test this hypothesis (restricting our investigation to concern grown-up men). Also, we have arranged with the necessary procedures to obtain the data in the morning and in the evening. We have also taken care of the measurement and recording and validation questions.

In the classical t-test we then will have a column of data with the lengths of men in the morning and a column of data with lengths of other men in the evening. These columns do not need to be equally long. We also suppose that sigma of process 'A' (length of men in the morning) is equal to sigma of process 'B' (length of men in the afternoon). Therefore we pool our two calculated standard deviations to one estimate of sigma. Using this information, we will test the hypothesis that the true mean of the processes are equal. The result depends on

- the size of the real difference
- the standard deviation of the processes
- the number of values

To pair. What is then a 'paired t-test'? Let us take another look at the example. Perhaps it would be a good idea to record the difference between the morning measurement and the evening measurement for each person. Now we can test the hypothesis that this difference really is zero and any deviation from zero is caused by randomness and not by any real, physical reason. This procedure would be called a paired t-test.

Suppose that the shrinking is exactly 0.7 cm for all men, i.e. no variation. Suppose also that men in the morning are distributed according to N[175; 5] and men in the afternoon are then distributed according to N[175.7; 5]. If we now take a sample from each one and perform a classical t-test, the probability is fairly low in detecting this difference (this probability is usually called the 'power' of the test).

However, if we measured the men in the morning and in the evening we would very quickly realise that the difference is exactly 0.7 cm. But let us try a simulation of the ordinary t-test. Let us create a population of 'men in the morning' and then subtract 0.7 cm to each value to get a population of 'men in the evening'.

```
erase c1-c1000                   # Clears the worksheet.

name c1 'Length of men in the morning' c4 'Sample morning lengths'
name c2 'Length of men in the evening' c5 'Sample evening lengths'

let k1 = 10000                   # Number of rows of data to simulate.
let k2 = 175                     # 'mu' of the normal distribution.
let k3 = 5                       # 'sigma'

random k1 c1;                    # Simulates data, normally
normal k2 k3.                    # distributed.

let c2 = c1 - 0.7                # Subtracts the shrinking.
```

Now we have two sets of data 'Length of men in the morning' and 'Length of men in the evening'. From these sets of data we randomly draw two samples and store these in c4 and c5:

```
sample 50 c1 c4                  # 1st sample in c4.
sample 60 c2 c5                  # 2nd sample in c5.
```

Will an ordinary t-test find this difference? Use the menu [Stat]>[Basic Statistics]>[2-sample t…] or the following command:

```
TwoSample c4 c5                  # Performs the t-test
```

As can be seen on the printout, the t-test cannot pick up the difference that we know exists.

A more realistic simulation. Now we need to realise that there is a variation of the shrinking also. Let us say that 'sigma shrinking' 0.2 cm. We need to simulate this before subtracting from the morning length.

```
erase c1-c1000                    # Clears the worksheet.

name c1 'Length of men in the morning' c4 'Sample morning lengths'
name c2 'Length of men in the evening' c5 'Sample evening lengths'

let k1 = 10000                    # Number of rows of data to simulate.
let k2 = 175                      # 'mu' of the normal distribution.
let k3 = 5                        # 'sigma'.

random k1 c1;                     # Simulates data, normally
normal k2 k3.                     # distributed.

random k1 c8;                     # The shrinking.
normal 0.7 0.2.

let c2 = c1 - c8                  # Subtracts the shrinking.
```

Again we sample a number of data but this time we make sure that we sample both c1 and c2 simultaneously:

```
sample 50 c1 c2 c5 c6             # Sample a number of persons.
```

Will an ordinary t-test find this difference? Use the menu [Stat]>[Basic Statistics]>[Paired t…] or the following command:

```
Paired C5 C6                      # Performs the paired t-test.
```

This time the difference is analysed and spotted very clearly.

Then, when to pair? If we go through necessary formulas concerning the t-test we find the following advice:

if $\sigma_D^2 < 2 \cdot \sigma^2$ (where σ_D is the standard deviation of the difference (here the shrinking) and σ is the standard deviation of the process (here length of men)) then it is important to pair the data.
This means that an investigation using t-test ought to be a designed investigation and if pairing is considered you need to get an approximation of σ_D and σ in order to choose method.

An additional note. If we pair data we lose so-called degrees of freedom. In the non-paired case we use all the data i.e. $(n_1 + n_2)$-values to estimate sigma. In the paired case we have only n values to estimate σ. If we have a small amount of data this has to be taken in consideration.

A small warning. Some years ago we looked in a book that concerned Computer Programming. In a special chapter there was an application of a t-test. The idea was that a group of people used two different computer languages in order to complete a number of tasks and the time needed was recorded. The data was then used to perform an ordinary t-test.
However, the data was not independent of one another as if it took long time for a specific task using language 'A' it also took long time to do the task using language 'B'. Therefore a paired t-test would have been the correct approach. ■