Assumptions are important:
the Paired and Pooled t-tests

KEYWORDS:
Teaching,
Correlation;
Degrees of freedom;
Dependent and independent observations.

J.C.W. Rayner
Department of Applied Statistics
The University of Wollongong
Wollongong, NSW, Australia

Summary
The relationship between the test statistics for the paired and pooled t tests, when applied to the same data set, is obtained. This assists us to understand why each is preferred over the other in appropriate circumstances

◆ INTRODUCTION◆

WHEN introducing students to the paired and pooled *t* tests, I find their most compelling need is to obtain the arithmetically correct value of the test statistics, and hence come to the "correct" conclusions. Far down on the list of priorities is to check the assumptions. So I give them an exercise where making an incorrect assumption can lead to a false conclusion. A typical example of this type is as follows.

◆EXAMPLE◆

This example is adapted from one in Mendenhall et al. (1990, problem 12.17, p. 581). "Muck" is the rich, highly organic type of soil that serves as the primary growth medium for vegetation in the Florida Everglades. Because of the high concentration of organic material, muck can be destroyed by a variety of natural and manmade causes. In May 1972 several plots in the Everglades were staked out, marked, and the depth of muck at each location measured. This was repeated in October 1978. A portion of the data (measured in inches) is given in Table I.

(i) Firstly analyse these (mucky!) data using a paired *t*-test. Give, as nearly as your tables permit, a value. Hence test to see if there is sufficient evidence to indicate that there was a significant loss in the average muck depth between 1972 and 1978. What would you conclude if you desired to implement an $\alpha = 0.01$ level test?

(ii) Now analyse the same data using a pooled *t*-test. Again give a p-value, and implement an $\alpha = 0.01$ level test.

(iii) In each analysis what assumptions should be checked?

The student should find that the p-value is less than 0.5% for the paired test, and between 2.5% and 5% for the pooled test. In discussion we stress that for the pooled test the data are independent random samples, so that every observation is independent of every other observation, whereas in the paired test the paired data may be dependent, frequently being observations on the same individual. It may be argued that if the data are dependent (paired), and if the dependent data were jumbled, then while it might not be valid to use the pooled test, it may be suggestive. In this case, if testing at the 1% level, the pooled test will be not significant, while the paired test will be significant. This point is brought out again in later problems, where for example, a non-parametric test may be not significant at a specified level, but a parametric test will be significant. In that context the question is, is the parametric test appropriate? Here we must ask, is the pairing appropriate?

Table 1: Depth of muck (ins)

Plot#	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1972	34.5	44.0	37.5	27.0	37.0	40.0	47.2	35.2	44.0	40.5	17.0	29.5	31.5	35.0	44.0
1978	31.5	37.9	35.5	23.0	34.5	31.1	46.0	31.0	35.2	37.2	24.7	25.8	29.0	36.8	36.5

◆RELATIONSHIP BETWEEN ◆ PAIRED AND POOLED *t* TESTS

(a) The Pooled Test: Suppose we have two independent random samples, $X_1, X_2, \ldots, X_N$ from a $N(\mu_x,\sigma_x^2)$ distribution, and $Y_1, Y_2 \ldots Y_n$ from a $N(\mu_y,\sigma_y^2)$ distribution. We wish to make inferences about the difference $d = \mu_x - \mu_y$ in the population means. Write $\bar{X}$ and $\bar{Y}$ for the sample means and S_X^2 and S_Y^2 and for the (unbiased) sample variances. If we can assume the unknown variances are equal, $\sigma_x^2 = \sigma_Y^2 = \sigma^2$ say, the common variance can be estimated using

$$S_p^2 = [(m-1) S_X^2 + (n-1) S_Y^2]/(m + n - 2)$$

The resulting test statistic ,

$T_{pooled} = \{(\bar{X} - \bar{Y}) - d\}/\{[S_P\surd[1/m + 1/n]\}$ has the t_{m+n-2} distribution.

This is the test statistic for the **pooled t-test.**

(b) The Paired Test: Suppose now we have observations $(X_1, Y_1), (X_2, Y_2) \ldots (X_n, Y_n)$ occurring as independent pairs, as often arises in before-after situations, such as, is a diet or medical treatment effective? The X's and Ys may **not** be independent; they are frequently observations on the same subject. However the differences $D_1 = X_1 - Y_1$, $D_2 = X_2 - Y_2$ **are** independent. If they can also be assumed to be normally distributed with common mean and variance, so that the D_i are independent $N(d,\sigma^2)$ then inferences about d can be based on the test statistic $T_{paired} = (\bar{D} - d)\surd n/S_D$, where S_D is the sample standard deviation of the differences.

(c) The Relationship: In terms of the X and Y sample means and variances, note that $\bar{D} = \bar{X} - \bar{Y}$ $E(D) = E(X - Y) = \mu_X - \mu_Y$ and that

$$(n-1) S_D^2 = \Sigma_i\{(X_i - Y_i) - (\bar{X} - \bar{Y})\}^2$$
$$=\Sigma_i\{X_i - \bar{X}) - (Y_i - \bar{Y})\}^2$$
$$=\Sigma_i(X_i - \bar{X})^2 - 2\Sigma_i (X_i - \bar{X})(Y_i - \bar{Y}) + \Sigma_i(Y_i - \bar{Y})^2$$
$$= (n-1)\{S_X^2 - 2rS_XS_Y + S_Y^2 \}$$

where r is the sample correlation between the X and Y values.

It follows that

$$T_{Paired} = \frac{(\bar{D} - \delta)\sqrt{n}}{S_D}$$
$$= \frac{\{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)\}\sqrt{n}}{\sqrt{\{S_X^2 - 2rS_XS_Y + S_Y^2\}}}$$

If m = n then $S_P^2 = (S_X^2 + S_Y^2)/2$ and

$T_{pooled} = \{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)\}/\{[S_P\surd[2/n]\}$

Substituting gives

$$T_{Paired} = T_{Pooled}\,\frac{S_P\sqrt{2/n}\sqrt{n}}{\sqrt{2S_P^2 - 2rS_XS_Y}}$$
$$= \frac{T_{Pooled}}{\sqrt{1 - rS_XS_Y / S_P^2}}$$

(d) If the pooled test is valid, then so is the paired. For if every X_1 is paired with a Y, namely Y_1, then for all i, $X_i - Y_i$ is normal, the pairs are independent, $E[X_i - Y_i] = \mu_X - \mu_Y = d$, say, and $var(X_i - Y_i) = s^2$, say. The assumptions noted in (a) are all satisfied.

◆THE EXAMPLE CONTINUED ◆

If the 1972 values are regarded as the X values and the 1978 as the Y, we find $\bar{X} = 36.927$, $\bar{Y} = 33.047$, $S_X^2 = 40.889$, $S_Y^2 = 35.517$. This gives $S_P^2 = 38.203$ and $t_{Pooled} = 1.719$. Further calculation gives the sample covariance as $\Sigma(X_i - \bar{X})(Y_i - \bar{Y})/(n - 1) = 33.989$, and the sample correlation as 0.892.

Substituting in the relationship above gives $t_{Paired} = 5.176$.This is confirmed by direct calculation, for D = 3.880 and $S_D = 2.903$.

◆IMPLICATIONS ◆

If the pooled test really is appropriate, then the sample correlation should be close to zero, and numerically T_{paired} and T_{pooled} will be approximately equal. However T_{paired} will be referred to n - 1 degrees of freedom, while T_{pooled} will be referred to twice this number. Looking at the *t* tables it will be seen that a particular *t* value may be significant at say the 5% level for 2(n - 1) degrees of freedom, but not significant with (n - 1) degrees of freedom. So the pooled test will be more critical of the data than the paired test in that it will have smaller p values.

The pooled test is more likely than the paired test to detect altematives from the null hypothesis. In technical terms, this test has more *power.*

If, as is in the example above and as is often the case, X and Y are positively correlated (high blood pressure before treatment, and relatively high after), then $|T_{paired}|$ will also be numerically greater than $|T_{pooled}.|$ If the correlation is large enough, the large value of $|T_{paired}|$ will overcome the degrees of freedom effect discussed in the previous paragraph. The paired test will then be more critical of the data than the pooled test. This is typically the case when the data occur as dependent pairs; that is, when the paired test is appropriate.

If the paired test is appropriate, but the pairings have been jumbled, the relationship between t_{paired} and t_{pooled} tells us that if the original observations can reasonably be assumed to have been positively correlated, then we can calculate t_{pooled} and the paired test will be at least as critical of the data.

◆A CAUTIONARY NOTE ◆

A colleague has observed that given two independent samples of size n. there are n! ways they can be paired and hence n! values of t_{paired} corresponding to the one t_{pooled} value. So using the n! possible pairings, a plethora of numerical values of t_{pooled} (and hence p-values) can be obtained. But section (d) above suggests that the paired test is valid.

There is no contradiction here. The several p-values are equivalent to taking n! different data sets and calculating test statistics and p-values. One then expects 5% of these to be greater than the 5% critical value and so on.

◆ACCOUNTING FOR THE ◆ DEGREES OF FREEDOM

Given a random sample of size n from a normal distribution with unknown mean and variance, there are n variables giving information about the unknown parameters. One variable, $\overline{X}$, gives information about location and the remaining n - 1 variables give information about dispersion. Given the 2n independent observations in the situation described in section (a), one variable gives information about the X population location, one about the Y population location, and 2(n - 1) variables give information about the common σ^2**.** Given the paired data as in section (b), one variable gives information about the difference in the population locations, one about the average of the population locations, and n - 1 about the dispersion of the differences. The remaining n - 1 variables give information about the dispersion not accounted for by the dispersion of the differences.

If pairing is appropriate, one reason why the paired test is more critical of the data (i.e. why it is a more powerful test) is that it doesn't use the irrelevant n - 1 variables that give information about the dispersion not accounted for by the dispersion of the differences. Using that information is like seeking an object known to be in a northerly direction, but looking east and west also. In the above discussion one may read "degrees of freedom" instead of "variables". But not doing so is perhaps more informative.

Conclusion

The paired and pooled tests are each in their own settings, optimal in a sense. The argument presented here merely gives reasons for this without raising technical issues of optimality.

Reference

Mendenhall, W., Wackerly, D.D. and Scheaffer, R.L. (1990). *Mathematical Statistics with Applications*, Fourth Edition, Boston: PWS-Kent Publishing Company.