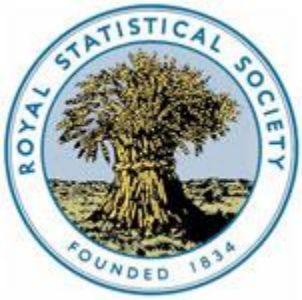

t Is for Trouble (and Textbooks): A Critique of Some Examples of the Paired-Samples t-Test

Author(s): D. A. Preece

Source: *Journal of the Royal Statistical Society. Series D (The Statistician)*, Vol. 31, No. 2 (Jun., 1982), pp. 169-195
Published by: Wiley for the Royal Statistical Society
Stable URL: http://www.jstor.org/stable/2987888
Accessed: 27-07-2017 00:11 UTC

The Statistician, Vol. 31, No. 2 0039-0526/82/00650169 $02.00

t is for Trouble (and Textbooks): A Critique of Some Examples of the Paired-Samples t-test

D. A. PREECE
Statistics Department, Rothamsted Experimental Station, Harpenden, Herts.

Many authors of statistical textbooks have been uncritical of the data in their examples of the paired-samples t-test. Careful criticism requires consideration of the following topics: outliers, non-homogeneity of source of the data, trends, transformations, and degree of precision of data recording. In particular, many examples have data recorded to an insufficient degree of precision.

1 Introduction

The dangers and problems of "service courses" in statistics are well known. The lecturer is under strong pressure from colleagues in, say, applied science departments to teach a wide variety of statistical techniques in a very short time, with insufficient opportunity for detailed and thoughtful study of real data. All too often, the result is that students come out of the course with little more than a set of statistical recipes – including, of course, formulae for the *t*-tests for unpaired and paired samples. These recipes are then used recklessly for statistical cookery, with little attention to the nature and quality of the numerical ingredients.

If concern with quality of data and with the assumptions underlying standard statistical procedures were well to the fore in the generality of current textbooks, the matter would not be so serious. However, many textbooks are startlingly uncritical of the data in their numerical examples, and unthinking statistical data processing is thereby encouraged. Unrealistic and superficial examination questions compound the trouble.

This paper confines itself to examining textbook examples of the paired-samples *t*-test and of corresponding confidence intervals. No systematic attempt has been made to find the worst possible examples: indeed I have merely included examples that have caught my attention because of some special feature of interest. The paper ends with a

re-examination of the well-known "additional hours of sleep" data analysed by Student himself and quoted by R. A. Fisher; these data are not open to the criticisms levelled at other examples, but have problems of their own and admit of a more interesting analysis than just a simple t-test. The paper is directed towards teachers and students alike, as well as towards statistical authors and experimenters.

2 Outliers

Inspection of data for outliers and other possibly anomalous values is often omitted by both research workers and statisticians. And textbooks often do not comment – or invite comment – on outliers in their numerical examples. That this applies to the paired-samples t-test is illustrated by the data of Table 1, which were used by Leonard and Clark (1939, Chapter 6) for such a test. These data are yields, in bushels per acre, of Glabron (x_1) and Velvet (x_2) barley grown side by side in single plots on 12 different

Table 1

Yields, in bushels per acre, of Glabron and Velvet barley grown side by side in single plots on 12 different farms

	Farm											
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*
Glabron (x_1)	49	47	39	37	46	52	51	57	45	45	48	64
Velvet (x_2)	42	47	38	32	41	41	45	56	42	39	47	39
x_1-x_2	7	0	1	5	5	11	6	1	3	6	1	25
x_1+x_2	91	94	77	69	87	93	96	113	87	84	95	103

farms. All 12 differences x_1-x_2 are non-negative, clearly suggesting the superiority of Glabron to Velvet under the conditions prevailing when and where the data were obtained. But one of the differences – that for farm 12 – is more than twice any of the others, and arises from a Glabron yield that is more than half as much again as the corresponding Velvet yield. This suggests that the data from farm 12 should receive careful scrutiny before they are included in an overall analysis. If something is found to be wrong with them, the rest of the data should be checked too; if the cause of the trouble is found, the possibility of the same or similar problems with the other data should be looked into carefully. Such scrutiny cannot, of course, be attempted by a textbook reader who has no access to the field records or to the experimenters, but it is undesirable for the reader to be invited to attempt or accept an uncritical analysis of data whose problems are not discussed.

The Glabron yield for farm 12 is so large (64 bushels per acre) that the question of a recording error (perhaps a transposition error: 64 instead of 46) must be raised. If the possibility of such an error is discounted, after examination of the field records, we should try to check whether, for example, the crops at farm 12 were harvested (or sown) much later (or earlier) than those elsewhere, or whether the amount of nitrogenous fertilizer was anomalous at farm 12. Even in the absence of such differences, there might be many reasons why farm 12 should behave differently from the others. Indeed we must question the use of Varieties × Farms interaction as "error"; if, for example, some of the farms have a different soil type from the others, then part of Varieties × Farms consists of Varieties × Soil types. Even more fundamentally we might ask whether Leonard and Clark were right to indulge in hypothesis testing for this sort of data: the *measurement* of varietal differences in yield for different soil types, different fertilizer levels, etc., might seem more appropriate than *testing the hypothesis* that the varieties do not differ in yield.

If no explanation can be found for seemingly anomalous values, the question remains whether they should be excluded from the analysis. To reject values that disconcert us, while glossing over the ones that we like, is to invite criticism. On the other hand, if we have strong prior reasons for believing that data should have certain properties (e.g. normality, constant variances, additivity), and these properties are violated because of the values under suspicion, we may well feel entitled to reject these values. Many formal procedures for the detection and rejection of outliers are now available, but they are not discussed here; they are of little or no use for very small sets of data.

Another example with 12 pairs of observations was provided by Wetherill (1972, p. 155) as an exercise in obtaining a confidence interval based on the t-distribution. The data, given here in Table 2, came from an experiment in which 12 subjects (people) were asked to steer a pencil along a moving track, in order to provide information on the reflex

Table 2

Average blink-rates per minute of 12 subjects in an experiment on variations in the reflex blink-rate

	Subject											
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*
Straight track (x_1)	24·0	19·5	8·2	8·5	12·1	8·0	8·2	10·1	5·5	10·1[a]	7·2	5·6
Oscillating track (x_2)	15·0	6·6	1·9	1·5	1·1	2·5	0·6	0·5	0·5	3·1	2·1	1·6
$x_1 - x_2$	9·0	12·9	6·3	7·0	11·0	5·5	7·6	9·6	5·0	7·0	5·1	4·0
$x_1 + x_2$	39·0	26·1	10·1	10·0	13·2	10·5	8·8	10·6	6·0	13·2	9·3	7·2

[a] Drew (1951) gave this value as 10·9, but Wetherill (1972) gave 10·1 (p. 155) and used 10·1 in his calculations (p. 328).

blink-rate of the eyes during visual motor tasks. Each subject was tested for two spells each of eight minutes, each spell having alternating periods of straight and oscillating track: the average blink-rate per minute was obtained for each subject for each type of track. For these data, none of the differences $x_1 - x_2$ could be called an outlier, but, on a scatter diagram of x_2 against x_1, the points for subjects 1 and 2 fall well apart from the others; this raises the question whether, in some sense, these two subjects came from a different population (or different populations) from the others and so should perhaps be excluded from the analysis. Also, the fact that the most extreme outlying point is for subject 1, the other being for subject 2, suggests either that the subjects have been ordered according to some unrevealed criterion, or that the experimental conditions took a while to settle down. This merits investigation before the data are analysed. So does the curious distribution of final digits in the 24 x_1 and x_2 values; there seems to be no obvious reason for the final digits of the 24 averages to be restricted to 0, 1, 2, 5, 6 and 9 (Preece, 1981). A further reason for caution can be seen from a scatter diagram of $x_1 - x_2$ against $x_1 + x_2$ where, whether or not subjects 1 and 2 are ignored, there is a suggestion of the difference tending to increase with the total.

Table 3

Further average blink-rates per minute for the 12 subjects of Table 2

	Subject											
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*
When "resting"	28	24	23	18	17	11	10	10	6	5	4	3
Under second set of experimental conditions:												
Straight track	19·0	16·7	2·7	6·6	12·0	7·0	6·0	4·1	3·0	11·3	5·9	3·1
Oscillating track	19·3	9·0	1·1	2·0	1·9	10·2	1·9	1·5	0·5	5·9	4·5	1·2

Table 2 provides a good example of data that make much less sense out of context than when related to other relevant information. Here further information consists of the data in Table 3, which is for the same 12 subjects (Drew, 1951), whose blink-rates were also recorded both during a pre-experimental "resting" period and for each type of moving track under a second set of experimental conditions. The Table 3 data make it clear that the 12 subjects have been arranged in order according to their blink-rates in the "resting" period, and that blink-rates vary, from subject to subject, over a wide range. With Table 3 to hand, there is no longer any question of segregating subjects 1 and 2 from the other 10. It is now natural to ask whether the subjects were of very different ages: Drew reported that they were all undergraduates (both men and women) aged between 19 and 25. With this information established, relevant to the likely homogeneity of the set of experimental subjects, it is much more reasonable than hitherto to proceed to analyse the Table 2 data.

3 Non-homogeneity of Experimental Units

The matter of the homogeneity of the experimental units is raised more vividly by the data used by Tippett (1952, p. 90, Table VII) for a paired-samples *t*-test. These data, reproduced here in Table 4, were obtained from 20 "specimens of meat", or "samples of meat", the percentage fat content of each having been estimated by each of two methods: the standard method of the Association of Official Agricultural Chemists of North

Table 4

Estimates of the percentage of fat in each of 20 "samples of meat"

Estimate from AOAC method (x_1)	*Estimate from Babcock method* (x_2)	*Difference* ($x_2 - x_1$)
22·0	22·3	0·3
22·1	21·8	−0·3
22·1	22·4	0·3
22·2	22·5	0·3
24·6	24·9	0·3
25·3	25·6	0·3
25·3	25·8	0·5
25·6	26·2	0·6
25·6	26·1	0·5
25·9	26·7	0·8
26·0	26·3	0·3
26·2	24·9	−1·3
27·0	26·9	−0·1
27·3	28·4	1·1
27·7	27·1	−0·6
41·5	41·4	−0·1
41·6	41·4	−0·2
45·5	45·5	0·0
48·5	48·2	−0·3
49·1	47·5	−1·6

America (AOAC) and the Modified Babcock method. Inspection of Table 4 shows that the estimates x_1 from the AOAC method are in ascending order of magnitude (which again raises the question whether a time effect is present, perhaps because of changes in the meat as it awaits analysis, perhaps because of a drift in the performance of the apparatus), and that the estimates x_2 from the Modified Babcock method are "close" to the corresponding values of x_1. But by far the most striking feature of the data is the absence of any x_1 and x_2 values between 28·4 and 41·4 per cent, which suggests that the "samples" are of at least two different types; indeed, closer inspection (perhaps aided by a scatter diagram) suggests

possibly three types, as no fat contents are quoted between 22·5 and 24·6 per cent. That this is a matter requiring careful investigation *before* a statistical analysis, is confirmed by the tendency towards positive values of $x_2 - x_1$ for small values of x_1 and x_2, and towards negative values of $x_2 - x_1$ for large x_1 and x_2 values: we must ask whether the algebraic sign of $x_2 - x_1$ tends to depend solely on the amount of fat present, or whether it depends solely on the "type" of meat, irrespective of fat content, or whether the two possibilities are inextricably confounded anyway.

Tippett's text does not help us to resolve these matters, but recourse to the original data (Oesting and Kaufman, 1945), where the x_1 values are *not* in order of magnitude, reveals five types of meat: Wieners, chopped pork, chopped ham, pork sausage links, and bulk pork sausage. As Figure 1 shows, the five "samples" with more than 40 per cent fat were the only pork sausage "samples": two of pork sausage links and three of bulk pork sausage. It is now clear that meat "type" and percentage fat content are confounded, and that the "samples" were not selected so as to permit answers to all the questions that might be asked; the value of the overall *t*-test is questionable. Indeed, as the Modified Babcock method was

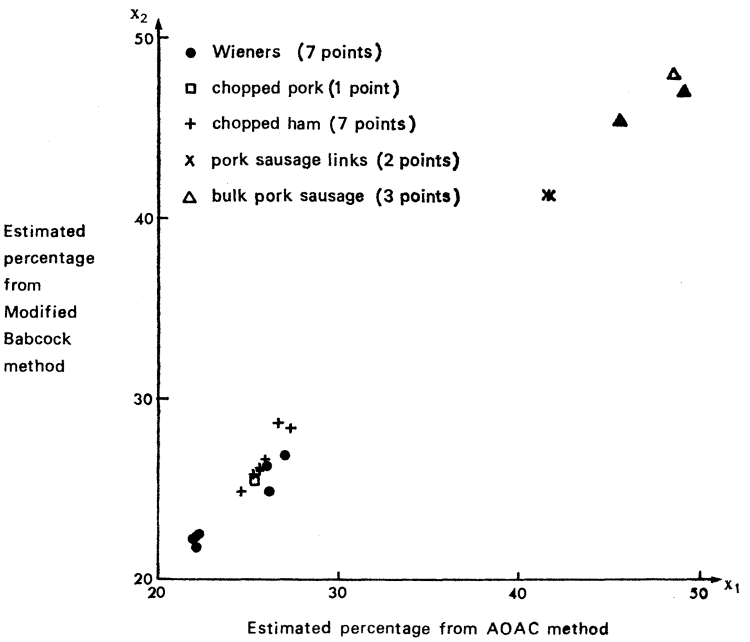

Figure 1

proposed as a rapid control method, whereas the AOAC method might be assumed to be accurate, it would be appropriate to consider the regression of x_2 on x_1 (Youden, 1947, p. 948); but the data are insufficient for fitting different regression lines for each meat "type".

4 Data with a Time Sequence

Table 5 reproduces data used by LeClerg *et al.* (1962, p. 51) to illustrate the paired-samples *t*-test. The time sequence is explicit in the description and tabulation of the data, but was ignored in LeClerg *et al.* calculation

Table 5

Grain yields, in bushels per acre, of Great Northern and Big Four oats grown in adjacent plots from 1912 to 1920

	Year								
	1912	*1913*	*1914*	*1915*	*1916*	*1917*	*1918*	*1919*	*1920*
Great Northern (x_1)	71·0	73·9	48·9	78·9	43·5	47·9	63·0	48·4	48·1
Big Four (x_2)	54·7	60·6	45·1	71·0	40·9	45·4	53·4	41·2	44·8
x_1-x_2	16·3	13·3	3·8	7·9	2·6	2·5	9·6	7·2	3·3

of their *t*-value. Their example is thus very unhappy for its purpose. It is the more unhappy in that, once again, the two largest differences are the first two. Perhaps talk of a "linear trend" in the differences would be extravagant, but the x_1-values clearly suggest a secular decline in the variety Great Northern. Much more evidence (in particular, genetic details) would of course be needed to confirm this impression. Indeed the phrase "grown in adjacent plots" is an inadequate account of how each year's data were obtained; we need to know, at the very least, whether the same pair of plots was used each year and, if not, whether all the results are from the same farm, same soil type, etc.

A more curious example of paired-samples data with a time element is provided by the values in Table 6, taken from Croxton and Cowden (1955, Table 24.2, p. 655) and Croxton *et al.* (1968, Table 24.2, p. 564). These data were quoted as percentages of solids in the shaded and exposed halves of 25 grapefruit from a shading experiment. The most eye-catching feature of these data consists of the final digits of the 50 values, these digits being distributed as in Table 7. Clearly the values have been adjusted from readings obtained to the nearest twentieth of a per cent. Perhaps a tare was subtracted, perhaps calibrations were involved. In any case, the final digits "9" and "4" (and similarly "7" and "2") almost certainly

Table 6

Percentages of solids recorded as being found in the shaded and exposed halves of 25 grapefruit

Fruit	*Shaded* (x_1)	*Exposed* (x_2)	$x_1 - x_2$
1	8·59	8·49	0·10
2	8·59	8·59	0·00
3	8·09	7·84	0·25
4	8·54	7·89	0·65
5	8·09	8·19	−0·10
6	8·49	7·84	0·65
7	7·89	7·89	0·00
8	8·59	7·89	0·70
9	8·54	7·79	0·75
10	7·99	7·84	0·15
11	7·89	7·79	0·10
12	8·09	7·84	0·25
13	7·89	7·89	0·00
14	8·54	8·07	0·47
15	7·84	7·97	−0·13
16	7·49	7·57	−0·08
17	7·89	7·92	−0·03
18	7·79	7·97	−0·18
19	7·84	8·17	−0·33
20	8·89	8·67	0·22
21	8·54	8·07	0·47
22	8·04	7·97	0·07
23	8·59	8·62	−0·03
24	8·19	7·92	0·27
25	8·59	7·97	0·62

appear in readings originally ending in "0" and "5" respectively (Preece, 1981). The interesting feature of this is that the same adjustment seems to have been used throughout the shaded halves and for the first 13 exposed halves, then a different adjustment throughout the remaining 12. We can deduce that the determinations for the shaded fruit were probably made first, probably in the tabulated order, then those for the exposed fruit, in the same order. Thus, the final digits enable us to criticize the running of the experiment. As the aim was to study differences between "shaded" and "exposed", these differences were required to be as precise as possible, which suggests dealing with both halves of a single fruit under as similar conditions as could be obtained. The 25 grapefruit can be thought of as the 25 blocks of a randomized block design with 2 plots per block; then the standard rule "work *by blocks*" means taking the two halves of a single fruit one immediately after the other. As it is, the run of negative

Table 7

Distributions of final digits in the data of Table 6

	Digit										*Total*
	0	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	
Shaded halves	.	.	.	.	7	.	.	.	.	18	25
Exposed halves											
First 13	.	.	.	.	4	.	.	.	.	9	13
Last 12	.	.	3	.	.	.	.	9	.	.	12

differences (fruits 15–19) starting just after the change in adjustment is disconcerting. Having the same adjustment for both halves of a fruit would also have produced all the differences $x_1 - x_2$ to the same degree of precision, all being multiples of 0·05.

5 Data Perhaps Requiring Transformation

Data used by Snedecor and Cochran (1967, pp. 94–5) for the paired-samples t-test are given here in Table 8. These data are slightly altered

Table 8

Numbers of lesions caused by two virus preparations inoculated into the two halves of each of eight tobacco leaves

	Plant[a]							
	5	*6*	*7*	*3*	*1*	*4*	*8*	*2*
Preparation 1 (x_1)	31	20[b]	18	17[b]	9	8	10	7
Preparation 2 (x_2)	18	17	14	11[b]	10	7	5	6
$x_1 - x_2$	13	3	4	6	−1	1	5	1
$x_1 + x_2$	49	37	32	28	19	15	15	11

[a] Youden and Beale's numbering.

[b] The marked values 20, 17 and 11 are Snedecor and Cochran's replacements for Youden and Beale's values 19, 16 and 10 respectively.

(for ease of calculation) from some observations recorded by Youden and Beale (1934, Table II, p. 444), who used the "half-leaf" method to compare how two different virus preparations affected tobacco plants. Eight plants were used, and the data in question are for the second leaf of each; half of each leaf was inoculated with preparation 1, and the other half with preparation 2. The numbers of local lesions that appeared on each half-

leaf were recorded; these were as in Table 8, where the plants have been arranged in descending order of total number of lesions per leaf. (Four of the plants had preparation 1 applied to the left half-leaf, and four to the right; details of this balanced application have however been omitted, as the data came from a series of experiments that afforded no evidence of any systematic difference between left and right.)

Eight pairs of values are of course too few for useful formal tests of the validity of assumptions underlying any analyses that may be attempted. Yet plant 5 has the difference $x_1 - x_2 = 13$, which is more than twice any of the others and is associated with easily the largest value of $x_1 + x_2$; this suggests the desirability of questioning a straightforward analysis of the eight differences. Also, the very fact that the values of x_1 and x_2 are *counts* suggests that they perhaps ought to be transformed before differences are taken. The most appropriate transformation may be either the square root transformation

$$y = \sqrt{x+c}$$

(c being a positive constant or zero), which is often suitable for "small" counts, or the logarithmic transformation

$$y = \log_{10}(x+c)$$

(c again being positive or zero). Another possibility is a power transformation

$$y = (x+c)^{\lambda}$$

with λ not equal to 1, $\frac{1}{2}$ or 0. (The transformations

$$y = \begin{cases} (x+c)^{\lambda}, & \lambda \neq 0 \\ \log(x+c), & \lambda = 0 \end{cases}$$

or

$$y = \begin{cases} [(x+c)^{\lambda} - 1]/\lambda, & \lambda \neq 0 \\ \log(x+c), & \lambda = 0 \end{cases}$$

are those considered by Box and Cox, 1964.) Also possible is

$$y = \log_{10}[\tfrac{1}{2}(x + c + \sqrt{x^2 + 2cx})]$$

(c being a constant usually between 5 and 15) given by Kleczkowski (1955) specifically for lesion numbers.

With positive values for seven of the eight differences $x_1 - x_2$, it seems clear enough that, under the conditions of the experiment, preparation 1 has a greater capacity to produce lesions than preparation 2. But it is instructive to compare the t-values (each with 7 degrees of freedom)

obtained with and without transformation of the counts. Without a transformation, the value is 2·63, which, for a two-sided test of the null hypothesis, is significant at the 5 per cent testing level but not at the 1 per cent level. With the counts transformed by the square root transformation with $c=0$, most of the differences become just over a tenth of what they were previously, the main exception being the new difference of about 0·926 for plant 8; the new t-value is 3·04. With the logarithmic or Kleczkowski transformation with $c=0$, the t-value is 3·11. With the Kleczkowski transformation with $c=5$ (a value used by Kleczkowski himself), the t-value is 3·16. Thus, for a two-sided test for this example, the t-value is significant at the 5 per cent testing level but not 1 per cent, whether the data are transformed or not.

6 Degree of Precision of Recording the Data Values: Illustrative Examples

Another example with eight pairs of observations was given by Wetherill (1972, pp. 153–4). The data, here reproduced in Table 9, are from an experiment to compare two methods of chlorinating sewage: method A (producing values x_1) involved rapid mixing at the start, but method B (x_2) did not. Both methods were used, one after the other, on each of 8 days; the order for running the methods was randomized afresh for each day.

Table 9

Log coliform densities per ml for each of two sewage-chlorination methods on each of eight days

	Day							
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*
Method A (x_1)	2·8	3·1	2·9	3·0	2·4	3·0	3·2	2·6
Method B (x_2)	3·2	3·1	3·4	3·5	2·7	2·9	3·5	2·8
x_2-x_1	0·4	0·0	0·5	0·5	0·3	−0·1	0·3	0·2

The 8 values of x_2-x_1 (of which six are positive and one zero) immediately suggest the superiority of method A and thus the advantage of including the rapid mixing in the sewage treatment process; the t-value is 3·38, which, for a two-sided test of the null hypothesis, is significant at the 5 per cent testing level but not 1 per cent. However, a glance at the 8 values shows that they are effectively from a seven-point scale with points −0·1, 0·0, 0·1, . . ., 0·5, the respective frequencies being 1, 1, 0, 1, 2, 1, 2; the variate, supposedly continuous, has thus become very *dis*continuous. Also, although the values x_2-x_1 each have one decimal place, they are

not to full one-decimal place precision, as they are differences between values themselves given to one decimal place. These considerations raise the question of how our *t*-value may have been affected by use of data given to such a poor degree of precision. When the "rounding interval", "grouping interval" or "group interval" of the data values is small compared with the standard error per value, rounding error can be regarded as just another component of error. But how small is "small", and what are the consequences of very coarse grouping?

Fertig and Heller (1950), one of whose experiments was similar to Wetherill's example, gave x_1 and x_2 values (ranging from 0·15 to 3·38) to two decimal places; their differences are effectively on a 252-point scale (Fertig and Heller's Table 2). However, the densities obtained, before taking logarithms to the base 10, were MPN (most probable number) estimates obtained "to two significant figures"; the log values thus do not have full two-decimal-place precision. (An estimate recorded as 1 200 stands for any density from 1 150 to 1 250. The corresponding range of logs, to two decimal places, is 3·06–3·10, the log of 1 200 being 3·08.) Indeed great *precision* for such data seems inappropriate, because of limitations on the *accuracy* with which the density determinations can be made. So, in investigating how the degree of precision of data-recording can affect a *t*-value, we are not *necessarily* criticizing data having a small degree of precision.

As log density per ml, if precisely measured, is effectively a continuous variate, each x_1 and x_2 value represents a range of possible values, this range being of width 0·1 units, the width of the rounding interval I for the data. Thus the value 2·8 in Table 9 stands for any value from 2·75 to 2·85. Consequentially, each difference $x_2 - x_1$ also represents a range of values, this being of width 0·2 units. In particular, the sets of values

$$0{\cdot}3 \quad -0{\cdot}1 \quad 0{\cdot}6 \quad 0{\cdot}4 \quad 0{\cdot}2 \quad -0{\cdot}2 \quad 0{\cdot}2 \quad 0{\cdot}1$$

and

$$0{\cdot}35 \quad 0{\cdot}10 \quad 0{\cdot}40 \quad 0{\cdot}40 \quad 0{\cdot}35 \quad 0{\cdot}00 \quad 0{\cdot}35 \quad 0{\cdot}30$$

are compatible with the recorded data, and these sets – chosen deliberately to represent extremes – give *t*-values as far apart as 2·05 (not significant at the 5 per cent testing level) and 5·35 (significant at the 0·2 per cent testing level) respectively. The estimated standard error of any one of the x_1 or x_2 values is, from Table 9,

$$([\{0{\cdot}4^2+0{\cdot}0^2+0{\cdot}5^2+0{\cdot}5^2+0{\cdot}3^2+(-0{\cdot}1)^2+0{\cdot}3^2+0{\cdot}2^2\}-(2{\cdot}1^2/8)]/(2\times 7))^{1/2}$$
$$= \pm 0{\cdot}156.$$

We use the term "*range of possible* t-*values*" for ranges such as 2·05–5·35;

details of their determination are too complex to be given in this paper. We discuss these ranges in section 8 below. The ranges for examples in this section and in section 9 are listed in Table 13, along with the numbers m of points on the effective scales for the differences $x_1 - x_2$, the widths I of the rounding intervals for the data values, and the estimated standard errors $\hat{\sigma}$ of the values.

Table 10

Percentages of iron found by each of two methods in each of ten compounds

	Compound									
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
Method A (x_1)	13·3	17·6	4·1	17·2	10·1	3·7	5·1	7·9	8·7	11·6
Method B (x_2)	13·4	17·9	4·1	17·0	10·3	4·0	5·1	8·0	8·8	12·0
$x_2 - x_1$	0·1	0·3	0·0	−0·2	0·2	0·3	0·0	0·1	0·1	0·4

Several other authors use data similar to Wetherill's to illustrate the paired-samples t-test. For example, Chatfield (1978, pp. 147–8) used, untransformed, the data reproduced here in Table 10. (Inspection of the data suggests that transformation is unnecessary.) Again, the differences are effectively on a seven-point scale; the "range of possible t-values" is 0·54 (not significant at the 5 per cent level) to 4·77 (significant at the 1 per cent level).

Likewise Paterson (1939, pp. 19–21) used the data given here in Table 11. It may well have been unrealistic or pointless to try to record the chicks' weights to a greater degree of precision, but the differences emerge as being effectively on a nine-point scale. Paterson reported his t-value to 3 decimal places (2·857), but this clearly cannot be justified; indeed his calculation involved division by a quantity calculated to only 2 significant figures, whereas better arithmetic gives 2·860. The "range of possible

Table 11

Weights, in ounces, of 20 seven-week-old chicks, two from each of ten families, one chick of each pair having been reared in confinement (Series A) and the other on open range (Series B)

Series A (x_1)	9	17	14	13	15	10	11	13	13	15
Series B (x_2)	8	15	11	11	9	12	11	10	9	14
$x_1 - x_2$	1	2	3	2	6	−2	0	3	4	1

t-values" is from 1·43 (again not significant at the 5 per cent level) to 4·87 (again significant at the 1 per cent level).

Paterson concluded that his t-value

> proves that the chicks reared in confinement have increased in weight more rapidly than those allowed free range.

Quite apart from considerations of precision, this is an unsound statement. The crucial phrase "on average" is missing. Also, the phrases "the chicks reared in confinement" and "those allowed free range" refer to the chicks in the experiment, not to a population or populations that might have been supposed to have been sampled, whereas the null hypothesis tested is about *population* weight gain; conclusions about populations are based on probabilities, not certainties, and so should hardly use the word "proves". (And, strictly speaking, Paterson's experiment provides no direct evidence on chicks from one-chick families, if such occur. This is doubtless a minor quibble, but deserves mention here because of the many studies done on twins: in many species, twins are uncommon and cannot be taken for granted as typical of their species.)

One of the skinniest sets of data that a textbook has offered for the paired samples t-test is the clearly artificial set reproduced in Table 12 and given by Balaam (1972, pp. 142, 240). A quick glance at the data is sufficient to

Table 12

Grain yields, in bushels per plot, of two wheat varieties grown on pairs of plots on six different properties in a single region

	Property					
	1	*2*	*3*	*4*	*5*	*6*
Variety A (x_1)	25	16	13	19	23	17
Variety B (x_2)	18	14	9	15	18	12
$x_1 - x_2$	7	2	4	4	5	5

herald a t-value significant at a standard testing level, and indeed the "range of possible t-values" is 4·05–12·04, whereas the tabulated 1 per cent point (two-sided test with 5 degrees of freedom) is 4·03. An even slighter set of obviously artificial data is in the "M & E Handbook" by Harper (1977, pp. 184, 303). For these, Harper obtained a t-value of 3·252 (better arithmetic gives 3·27), with 4 degrees of freedom.

It has been argued that simple examples such as Balaam's and Harper's

are appropriate for teaching, especially to students without calculating machines. Perhaps there is a little truth in this. But more than a very few such artificial examples serve only to inculcate the unthinking routine use of standard statistical recipes. (An unthinking approach is particularly encouraged by having the data in bushels *per plot*, without the plot size being given; students thus have no way of telling whether the yields are good or bad.) Such simple examples frequently appear as questions on examination papers, where they do untold harm by encouraging bad statistical practice – particularly when they are phrased in the well-known "Analyse first; think afterwards" format.

A particularly sorry examination question on the paired-samples t-test was quoted by Harper (1977, p. 345), who attributed it to the Association of Certified Accountants. This question has 6 pairs of values, each of the 12 values being recorded in tenths of hours. The 6 differences, obviously contrived for easy arithmetic, are effectively on a *three*-point scale:

$$+0{\cdot}2 \quad 0{\cdot}0 \quad +0{\cdot}1 \quad +0{\cdot}2 \quad +0{\cdot}1 \quad 0{\cdot}0$$

Limiting t-values of 0·00 and $+\infty$ are obtained by interpreting the differences as, respectively,

$$+0{\cdot}1 \quad -0{\cdot}1 \quad 0{\cdot}0 \quad +0{\cdot}1 \quad 0{\cdot}0 \quad -0{\cdot}1$$

and

$$+0{\cdot}1 \quad +0{\cdot}1 \quad +0{\cdot}1 \quad +0{\cdot}1 \quad +0{\cdot}1 \quad +0{\cdot}1$$

(For consistency, Table 13 gives two-sided percentage points for this example, but the wording of the examination question indicates that a one-sided test was required.) An almost identical examination question, with differences effectively on a four-point scale, was quoted by Owen and Jones (1977, pp. 343–4), who attributed it to the same Association.

7 Degree of Precision of Recording the Data Values: Literature and Rules

In his classic paper on "The probable error of a mean", "Student" (1908, p. 14) discussed the statistical effects of using "wide groups" (i.e. a poor degree of precision of recording) for data. Then Fisher (1922, pp. 361–2) obtained results on the loss of efficiency, due to grouping, when the parameters of a normal distribution with variance σ^2 are estimated from a random sample of observations drawn from the distribution. He showed that the loss in the estimation of the standard deviation is nearly $a^2/6$, where $a\sigma$ is the group interval, and that the loss in the estimation of the mean is nearly $a^2/12$. Thus the loss for the standard deviation is less than 1 per cent, provided the group interval does not exceed one-quarter of the

Table 13

Details of examples quoted in sections 6 and 9 (m *denotes the number of points on the effective scale for the differences* $x_1 - x_2$; I *is the width of the rounding interval for a single observation, and* $\hat{\sigma}$ *is the estimated s.e. of a single observation*)

Author	*D.F.*	*t-value (standard formula)*	*Range of possible t-values*	*Tabulated 5% and 1% points (two-sided test)*	*m*	*I*	$\hat{\sigma}$
Wetherill (1972, pp. 153–4)	7	3·38	2·05–5·35	2·36, 3·50	7	0·1	0·156
Chatfield (1978, pp. 147–8)	9	2·33	0·54–4·77	2·26, 3·25	7	0·1	0·125
Patterson (1939, pp. 19–21)	9	2·86	1·43–4·87	2·26, 3·25	9	1	1·56
Balaam (1972, pp. 142, 240)	5	6·71	4·05–12·04	2·57, 4·03	6	1	1·16
Harper (1977, pp. 184, 303)	4	3·27	2·17–5·05	2·78, 4·60	8	0·1	0·194
Examination question as quoted by Harper (1977, p. 345)	5	2·74	0·00–$+\infty$	2·57, 4·03	3	0·1	0·063
Examination question as quoted by Owen and Jones (1977, pp. 343–4)	5	1·75	−0·35–5·00	2.57, 4·03	4	0·1	0·083
"Student" (1908, pp. 20–1)	9	4·06	3·69–4·47[a]	2·26, 3·25	47	0·1[a]	0·870

[a] On the assumption that the x_1 and x_2 values are to the nearest one-tenth of an hour (see text).

standard deviation. This is the basis of the recommendation (Eisenhart, 1947, p. 192) that

> the width of the rounding interval [should] be less than one-third, or, better, less than one-fourth the standard deviation of random sampling.

Snedecor and Cochran (1967, p. 81) echoed this:

> For accurate work, the advice commonly given is that I [the width of the grouping interval] should not exceed $\sigma/4$. This requires about 24 classes to cover the frequency distribution when the sample is large.

However, Nicholson (1979) wrote of the "often quoted" rule that the width of the rounding interval should not exceed $\sigma/2$. Cochran and Cox (1957, section 3.33) translated the $\sigma/4$ rule across to experimental data in the words

> the rounding interval should not exceed one-quarter of the standard error per observation.

For the examples of section 6, the standard error of a single observation can be estimated only poorly, both because of the poor degree of precision of recording of the data and because of the paucity of degrees of freedom. However, a comparison of the estimated standard errors $\hat{\sigma}$ with the corresponding values of I is salutary: as Table 13 shows, I is greater than $\hat{\sigma}/2$ (let alone $\hat{\sigma}/3$ or $\hat{\sigma}/4$) for *all* the section 6 examples. This seems to confirm the impression of an inadequate degree of precision in data in so many textbook examples. Table 13 also shows that – as one would expect – the excess of I over $\hat{\sigma}/2$ tends to increase as m decreases.

A classic statistical procedure for grouped data is use of "Sheppard's corrections". However, Fisher (1932, Chapter III, Appendix D) advised that

> These adjustments should be used for purposes of estimation, but not for tests of significance.

This advice later (Fisher, 1936) became

> . . . but not usually for tests of significance.

In this form it was reiterated by Eisenhart (1947, p. 203), who pointed out that use of a Sheppard's correction can make the t-value imaginary (in the mathematical sense) as the corrected estimate of the variance can be negative. (Fisher's "estimation" must presumably be interpreted to mean "point estimation", as interval estimates can hardly be based on imaginary t-values.) Further reiteration came from Gjeddebaek (1978), who went on to say that "Sheppard's correction should be avoided in analysis of variance".

Eisenhart (1947, pp. 191–215) considered how rounding affects statistical analyses of (a) large numbers of observations, and (b) small numbers of

observations from a normal population. His conclusions need not be summarized here, except for how he proposed judging the suitability of a particular coarseness of grouping for a small sample: his criterion is based on the probability of the sample variance being zero, i.e. the probability that all the observations are the same when that coarseness of grouping is adopted.

Also relevant to the degree of precision of recording is the work of Fortunato (1980), who devised a Monte Carlo study of the performance of Student's t for samples from discrete distributions.

8 Degree of Precision of Recording the Data Values: Effect on t

When rounding errors are small, and can be regarded as contributing just another component of error to the experimental results, a standard calculated t-value takes full account of this extra component, and can be taken at its face value for hypothesis-testing and calculating confidence limits. However, when rounding errors are large, distributional assumptions lying behind the standard supporting mathematical theory are violated; the "ranges of possible t-values" introduced in section 6 may perhaps then serve to provide useful warnings of shaky inference-making.

The end-points of these ranges are very unlikely to be attained in practice, and indeed any study of the ranges must take account of some sort of distribution of t-values within them. In general, such distributions will not be elegant or symmetrical. To give some idea of the distributional difficulties involved, and of how t-values can be affected by gross rounding of data, we now consider analyses based on the Table 5 yields, both with the given degree of precision and also further rounded, with the time sequence ignored. (These yields give differences that are effectively on a respectably wide 139-point scale.)

We first consider the analyses reported in Table 14, which shows that the data, originally to the nearest one-tenth of a bushel per acre, were rounded variously to the nearest one-fifth of a bushel, the nearest bushel, the nearest two bushels, and so on. (For interval-widths I that are even multiples of 0·1, a data value found on a boundary between two rounding intervals was rounded to the nearest *even* multiple of I. This may have caused an occasional rounded value to differ from what would have been obtained by initial recording with rounding interval I, but the effect of the disturbance can be ignored.) As Table 14 shows, I is initially much less than $\sigma/4$, and the number of points m on the effective scale for the differences is ample. However, as I increases beyond 1·0 the t-value becomes erratic, and when we come to three successive roundings all with $m=3$ (as in the examination question criticized in section 6) we find these yielding t-values as far apart as 6·40 and 2·29 – which provides

Table 14

Summary of analyses of the oats data of Table 5, first as given, then as rounded to the nearest I *bushels per acre, for various values of* I *(the time sequence being ignored)*

I	$\hat{\sigma}$	*t-value*	m
0·1	3·50	4·48	139
0·5	3·57	4·43	29
1·0	3·37	4·68	14
2·0	3·94	3·95	9
2·5	3·58	4·60	6
4·0	3·16	5·37	4
5·0	2·95	6·40	3
10·0	4·71	3·50	3
12·5	6·42	2·29	3
20·0	6·24	1·51	2
25·0	8·84	2·00	2
40·0	12·47	1·51	2
50·0	11·79	1·00	2
100·0	0	—	1
125·0	44·19	2·00	2
200·0	0	—	1

a forceful warning of the dangers of recording data to a very poor degree of precision.

This erratic behaviour calls for explanation, and can be illuminated by noting that the information in Table 14 is dependent on the position of the original data relative to the origin. This can be seen by shifting the data relative to the origin by successive amounts 0·1 and then, for each increment, rounding afresh and recalculating the t-value. With $I=0{\cdot}1$, the t-value is of course unaffected by any number of such shifts. But with $I=0{\cdot}5$, up to 5 different t-values may be obtained; with $I=2{\cdot}5$, up to 25 different values; and so on. Thus, for $I=0{\cdot}5$, the data taken first as they stand, and then increased by 0·1 four times successively, give t-values of

4·43, 4·46, 4·50, 4·49 and 4·51

respectively. For $I=1{\cdot}0$, the t-values similarly obtained are

4·68, 4·47, 4·54, 4·47, 4·32,
4·29, 4·34, 4·30, 4·42, 4·49,
4·49, 4·75, 4·54, 4·54, 4·64
4·47, 4·40, 4·45, 4·30 and 4·42

with mean 4·47; here there are 20 values, not 10, because of the rule for rounding data values falling on interval-boundaries. For values of I from 0·1 to 20·0, Table 15 gives a summary of the t-values generated.

Inspection of Table 15 now helps to explain the behaviour of the t-values of Table 14: those for $I=4·0$ and 5·0 are at the tops of the generated ranges, whereas that for $I=20·0$ is at the bottom of its range. But

Table 15

Details of t*-values generated by shifting the Table 5 data relative to the origin by increments of 0·1*

I	*No. of t-values generated*	*Details of t-values*				*t-value from Table 14*
		Minimum	*Mean*	*Median*	*Maximum*	
0·1	1	–	–	–	–	4·48
0·5	5	4·43	4·48	4·49	4·51	4·43
1·0	20	4·29	4·47	4·47	4·75	4·68
2·0	40	3·83	4·45	4·36	5·24	3·95
2·5	25	4·04	4·41	4·35	4·75	4·60
4·0	80	2·94	4·44	4·54	5·38	5·37
5·0	100	2·44	4·48	3·83	6·40	6·40
10·0	200	1·84	3·69	3·50	6·00	3·50
12·5	125	2·00	3·53	2·83	8·00	2·29
20·0	400	1·51	2·25	2·00	4·00	1·51

proper understanding of the distributions of generated t-values requires much more detail than merely minimum, mean, median and maximum values. Thus, for $I=20·0$ only 5 distinct t-values are obtained; these, with numbers of occurrences in parentheses, are as follows:

1·51 (140), 2·00 (136), 2·53 (28), 3·16 (46), 4·00 (50)

And when the t-values are taken in the order generated, the distinct values come in batches:

1·51 (12),	2·00 (5),	2·53 (3),	3·16 (2),	4·00 (25),
3·16 (3),	2·53 (3),	2·00 (13),	2·53 (4),	3·16 (19),
2·53 (3),	2·00 (2),	1·51 (18),	2·00 (41),	etc.

As I increases, this sort of pattern begins to emerge with $I=2·5$, and is clear from $I=5·0$ onwards; the numbers of distinct t-values for the larger values of I are

$I=5·0$: 15; $I=10·0$: 9; $I=12·5$: 10; $I=20·0$: 5

These results can be summed up by saying that, for coarse rounding, the value obtained for a *t*-statistic depends crucially both on the rounding interval adopted and on the position of the rounding grid relative to the data. Final conclusions can, of course, hardly be drawn from only a single example, or indeed from several, but the results reported here, and others obtained separately, suggest clearly that drastic rounding can be more dangerous than has commonly been supposed. Further work on this topic is called for.

9 Student's Example Involving Two Supposedly Soporific Drugs

It seems appropriate to conclude with an examination of the data given by Student himself (1908, pp. 20–1) and made famous through their use by Fisher (1925, pp. 107–10). The data, here reproduced in Table 16, relate (in Student's words) to

> the different effects of the optical isomers of hyoscyamine hydrobromide in producing sleep. The sleep of 10 patients was measured without hypnotic and after treatment (1) with D. hyoscyamine hydrobromide, (2) with L. hyoscyamine hydrobromide. The average number of hours sleep gained by use of the drug is tabulated . . .

Table 16

Hours of sleep gained by ten patients by use of each of two isomers (viz. Dextro- and Laevo-) of hyoscyamine hydrobromide

	Patient									
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
Dextro- (x_1)	+0·7	−1·6	−0·2	−1·2	−0·1	+3·4	+3·7	+0·8	0·0	+2·0
Laevo- (x_2)	+1·9	+0·8	+1·1	+0·1	−0·1	+4·4	+5·5	+1·6	+4·6	+3·4
x_2-x_1	+1·2	+2·4	+1·3	+1·3	0·0	+1·0	+1·8	+0·8	+4·6	+1·4

Thus there were seemingly three "treatments" in the experiment, namely a control ("without hypnotic") and the two isomers, and the number of hours of sleep was obtained for each patient for each treatment.

If the numbers of hours of sleep for the three treatments are represented by, respectively, y_0, y_1 and y_2, the values in the rows of Table 16 are obtained from

$$x_1 = y_1 - y_0$$
$$x_2 = y_2 - y_0$$
$$x_2 - x_1 = y_2 - y_1$$

Thus, analysis of the values of either x_1 or x_2 is as much a paired-samples job as the analysis of $x_2 - x_1$. Student did indeed do separate analyses of

x_1, x_2 and $x_2 - x_1$, without reference to later worries about the propriety of doing the three tests when there are only two degrees of freedom for comparisons between treatments.

The most striking value in Table 16 is the $x_2 - x_1$ value for patient 9, but, without knowing how many nights of sleep were averaged over, and without knowing y_0 for the different patients, we are in no position to say whether $x_2 - x_1 = +4{\cdot}6$ hours seems to require further explanation.

If the $x_2 - x_1$ values are taken as they stand, they are effectively on a 47-point scale, and give a t-value of 4·06, with 9 degrees of freedom. On the assumption (to be shaken below) that the x_1 and x_2 (or y_1 and y_2) values were to the nearest one-tenth of an hour, the width I of the rounding interval for the values is comfortably less than $\hat{\sigma}/4$ (see Table 13). On this same assumption, the range of possible t-values is from 3·69 to 4·47, which is much narrower than the ranges for the examples of section 6; the end-values are both significant at the 1 per cent testing level (two-sided test). We have strong evidence of the second isomer being a better soporific than the first.

As there were ten patients and three treatments, the data provide $9 \times 2 = 18$ degrees of freedom for patients × treatments interaction, which suggests the possibility of comparing the effects of the isomers by a t-test with 18 degrees of freedom instead of 9. However, Student noted that the standard deviation for $x_2 - x_1$ is "low" by comparison with those for x_1 and x_2. This can be interpreted in analysis of variance terms, as sums of

Table 17

Sums of squares and mean squares for the data of Table 16

S.V.	*D.F.*	*S.S.*	*M.S.*
Between treatments:			
Between isomers	1	12·4820	12·4820
Isomers versus control	1	15·8107	15·8107
Patients × treatments:			
Patients × isomers	9	6·8080	0·7564
Patients × (isomers versus control)	9	19·3593	2·1510

squares for treatments and patients × treatments can be obtained from Table 16 and can be partitioned as in Table 17. The difference between the two interaction mean squares reflects Student's findings. The ratio of the mean squares for isomers and patients × isomers, namely

$$12{\cdot}4820/0{\cdot}7564 = 16{\cdot}50$$

Table 18

The "hours of sleep" data, as given by Cushny and Peebles (1905, p. 509) (n *denotes the number of nights averaged over,* y *denotes average number of hours sleep, and* x *denotes hours gained, by comparison with the control; the data were obtained in the Michigan Asylum for insane at Kalamazoo*)

	Patient									
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
Control										
n	9	9	8	9	9	8	8	7	8	9
y	0·6	3·0	4·7	5·5	6·2	3·2	2·5	2·8	1·1	2·9
Laevo-hyoscyamine hydrobromate										
n	6	6	6	3	3	4	3	6	5	5
y	1·3	1·4	4·5	4·3	6·1	6·6	6·2	3·6	1·1	4·9
x	+0·7	−1·6	−0·2	−1·2	−0·1	+3·4	+3·7	+0·8	0·0	+2·0
Laevo-hyoscine hydrobromate										
n	6	6	6	3	3	3	3	6	6	5
y	2·5	3·8	5·8	5·6	6·1	7·6	8·0	4·4	5·7	6·3
x	+1·9	+0·8	+1·1	+0·1	−0·1	+4·4	+5·5	+1·6	+4·6	+3·4
Racemic hyoscine hydrobromate										
n	6	6	6	3	3	3	3	5	5	6
y	2·1	4·4	4·7	4·8	6·7	8·3	8·2	4·3	5·8	6·4
x	+1·5	+1·4	0·0	−0·7	+0·5	+5·1	+5·7	+1·5	+4·7	+3·5

is of course the square of the previously obtained t-value of 4·06. It can readily be seen that the difference between the isomers is still significant at the 5 per cent testing level, even if the pooled patients × treatments interaction is used as the error term.

Student (1908, p. 20) quoted his data as coming

> from a table by A. R. Cushny and A. R. Peebles in the *Journal of Physiology* for 1904 . . .

However, his treatment names (as well as the year) differ from those of the source paper by Cushny and Peebles (1905). Also, as our Table 18 shows, the original experiment had *four* treatments (a control and three others), and the 40 averages were variously based on 3, 5, 6, 7, 8 or 9 nights' sleep. Thus, a strict analysis of the patients' averages for some or all of the treatments should perhaps incorporate weights equal to the numbers of nights averaged over. (But it would be preferable to have the full data – not just averages – before taking a decision on this, so that a patient's variance between nights of the same treatment could be compared with components of the patients × treatments interaction.) Further complications are introduced by the pre-Fisherian design of the experiment, as described by Cushny and Peebles (1905, p. 509):

> As a general rule a tablet was given on each alternate evening, and the duration of sleep and other features noted and compared with those of the intervening control night on which no hypnotic was given. Hyoscyamine was thus used on three occasions, and then racemic hyoscine, and then laevo-hyoscine. Then a tablet was given each evening for a week or more, the different alkaloids following each other in succession.

Thus, if the treatments, as listed in Table 18, are labelled 0, A, B and C (of which 0, A and B are those considered by Student), the basic order of administration seems to have been something like

0A0A0A0C0C0C0B0B0BACBACBACB

Such a design makes no proper allowance for the elimination of carry-over or compensation effects, and interpretation of the data must therefore be problematical. (Modern designs that can take proper account of carry-over effects are discussed by Cochran and Cox, 1957, section 4.6a) Table 18 does, however, suggest a plausible explanation for the behaviour of patient 9, mentioned above: this patient slept badly under the control or hyoscyamine treatment, but responded well to both forms of hyoscine.

If the design deficiencies (including the unequal replication) are ignored, Table 18 permits calculation of a mean square for patients, for insertion in Table 17, namely 7·4349. Similarly an analysis of variance for *all* the

Table 19

An analysis of variance for the data of Table 18, with the unequal replication ignored

S.V.	*D.F.*	*S.S.*	*M.S.*
Between patients	9	89·5000	9·9444
Between treatments			
Between forms of hyoscine (X)	1	0·0005	0·0005
Hyoscyamine versus hyoscine (Y)	1	16·5375	16·5375
Control versus others (Z)	1	24·3000	24·3000
Patients × treatments			
Patients × X	9	1·6445	0·1827
Patients × Y	9	9·7775	1·0864
Patients × Z	9	24·9200	2·7689
Total	39	166·6800	–

data can be presented as in Table 19, which shows (once again) the clear non-homogeneity of the interaction between patients and treatments: each of the main effect components X, Y and Z should be assessed by comparing it with the corresponding interaction component.

Returning to the paired-samples data as given by Fisher (1925, pp. 107–10), we can now see them as illustrative of those common situations, well-known to practising statisticians, where careful enquiry about the origins of some data reveals concealed complexities, and where the production and justification of an appropriate analysis may be much less straightforward than was originally supposed. In Fisher's text, the complexities are hinted at by the simplicity of the description "Additional hours of sleep gained". "Additional to what?", we may ask, and "Surely a patient's reaction to a drug would not be assessed from a single night's observations?" Thus are we naturally led to consider whether there has been some averaging.

Shorn of its complexities, Fisher's example is a clear enough illustration of the paired samples t-test, and is of special historic interest. With its complexities restored, it is a useful reminder of the realities of applied statistics.

Acknowledgement

This paper was prepared and written whilst the author held a post funded by the Overseas Development Administration.

REFERENCES

BALAAM, L. N. (1972). *Fundamentals of Biometry*. The Science of Biology Series (ed. J. D. Carthy and J. F. Sutcliffe), No. 3. Allen and Unwin, London.

BOX, G. E. P. and COX, D. R. (1964). An analysis of transformations (with discussion). *Journal of the Royal Statistical Society, Series B*, **26**, 211–52.

CHATFIELD, C. (1978). *Statistics for Technology: A Course in Applied Statistics*, 2nd edn. Chapman and Hall, London.

COCHRAN, W. G. and COX, GERTRUDE M. (1957). *Experimental Designs*, 2nd edn. Wiley, New York.

CROXTON, F. E. and COWDEN, D. J. (1955). *Applied General Statistics*, 2nd edn. Prentice-Hall, Englewood Cliffs; Pitman, London.

CROXTON, F. E., COWDEN, D. J. and KLEIN, S. (1968). *Applied General Statistics*, 3rd edn. Pitman, London.

CUSHNY, A. R. and PEEBLES, A. R. (1905). The action of optical isomers. II. Hyoscines. *Journal of Physiology*, **32**, 501–10.

DREW, G. C. (1951). Variations in reflex blink-rate during visual-motor tasks. *Quarterly Journal of Experimental Psychology*, **3**, 73–88.

EISENHART, C. (1947). Effects of rounding or grouping data. Chapter 4 of *Techniques of Statistical Analysis* (ed. C. Eisenhart, M. W. Hastay and W. A. Wallis), pp. 185–223. McGraw-Hill, New York.

FERTIG, J. W. and HELLER, A. N. (1950). The application of statistical techniques to sewage treatment processes. *Biometrics*, **6**, 127–35.

FISHER, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London, Series A*, **222**, 309–68. [Reprinted (1950), with corrections, as Paper 10 in *Contributions to Mathematical Statistics*. Wiley, New York.] [Again reprinted (1971), with corrections, in *Collected Papers of R. A. Fisher* (ed. J. H. Bennett), Vol. I, pp. 275–335. University of Adelaide.]

FISHER, R. A. (1925). *Statistical Methods for Research Workers*. Oliver and Boyd, Edinburgh.

FISHER, R. A. (1932). *Statistical Methods for Research Workers*, 4th edn. Oliver and Boyd, Edinburgh.

FISHER, R. A. (1936). *Statistical Methods for Research Workers*, 6th edn. Oliver and Boyd, Edinburgh.

FORTUNATO, E. (1980). Problems concerning the robustness of Student *t* statistical test. In *Monte Carlo Studies on Robustness* (ed. A. Rizzi), pp. 95–109. CISU (Cooperativo Informazione Stampa Universita a r.l.), Rome.

GJEDDEBAEK, N. F. (1978). Statistical analysis, III: Grouped observations. In *International Encyclopedia of Statistics* (ed. Judith Tanur and W. Kruskal), pp. 1056–60. Macmillan, Free Press, New York; Collier Macmillan, London.

HARPER, W. M. (1977). *Statistics*, 3rd edn. Macdonald and Evans, Plymouth.

KLECZKOWSKI, A. (1955). The statistical analysis of plant virus assays: a transformation to include lesion numbers with small means. *Journal of General Microbiology*, **13**, 91–8.

LECLERG, E. L., LEONARD, W. H. and CLARK, A. G. (1962). *Field Plot Technique*. Burgess, Minneapolis.

LEONARD, W. H. and CLARK, A. G. (1939). *Field Plot Technique*. Burgess, Minneapolis.

NICHOLSON, M. D. (1979). On expressing the mean of rounded data. *Biometrics*, **35**, 873–4.

OESTING, R. B. and KAUFMAN, I. P. (1945). Rapid determination of fat in meat and

meat products. *Industrial and Engineering Chemistry, Analytical Edition*, **17**, 125.

OWEN, F. and JONES, R. (1977). *Statistics*. Polytech Publishers, Stockport.

PATERSON, D. D. (1939). *Statistical Technique in Agricultural Research*. McGraw-Hill, New York.

PREECE, D. A. (1981). Distributions of final digits in data. *The Statistician*, **30**, 31–60.

SNEDECOR, G. W. and COCHRAN, W. G. (1967). *Statistical Methods*, 6th edn. Iowa State University Press, Ames.

"STUDENT" [W. S. GOSSET] (1908). The probable error of a mean. *Biometrika*, **6**, 1–25. [Reprinted (1942) in "*Student's*" *Collected Papers* (ed. E. S. Pearson and J. Wishart), pp. 11–34. Cambridge University Press.]

TIPPETT, L. H. C. (1952). *Technological Applications of Statistics*. Williams and Norgate, London.

WETHERILL, G. B. (1972). *Elementary Statistical Methods*, 2nd ed. Chapman and Hall, London.

YOUDEN, W. J. (1947). Technique for testing the accuracy of analytical data. *Analytical Chemistry*, **19**, 946–50.

YOUDEN, W. J. and BEALE, HELEN P. (1934). A statistical study of the local lesion method for estimating tobacco mosaic virus. *Contributions from Boyce Thompson Institute*, **6**, 437–54.