Table 1. Fungitoxicity of some insecticides (% mycelial inhibition)

Plant pathogenic fungi	Concentration (%)																			
	Control	Durmet				Kanodane				Cymbush				Nuvacron				Nuvan			
		0.01	0.05	0.1	0.2	0.01	0.05	0.1	0.2	0.01	0.05	0.1	0.2	0.01	0.05	0.1	0.2	0.01	0.05	0.1	0.2
F. oxysporum	0	43.3	52.62	66.66	81.11	35.55	37.85	44.44	70	22.22	36.66	58.88	80	14.44	38.88	100	100	20	39.24	56.25	100
A. solani	0	66.66	88.88	84.44	100	26.66	50	60.89	75.55	20	64.44	73.33	87.77	0	43.33	60	100	6.66	43.33	65.33	100
C. lunata	0	60	77.77	83.3	90	37.77	53.33	70.55	80.55	27.77	57.32	78.8	90.5	11.11	44.44	91.48	100	11.11	66.66	75.18	100
Helmintho-sporium sp.	0	68.75	83.33	90	100	45	55.55	80.62	88.88	32.5	58.88	83.75	100	18.75	36.61	88.12	100	32.5	55.55	100	100
S. rolfsii	0	70	89.44	100	100	55.55	77.77	90	100	13.33	55.55	81.38	100	0	48.33	100	100	37.77	79.88	100	100

rolfsii, *Curvalaria lunata* and *Alternaria solani*. 90% inhibition was observed at 0.2% with respect to all fungi against different insecticides. As the dose was decreased, there was decrease in inhibition. Except Durmet and Kanodone, the rest of the insecticides did not exhibit any significant inhibition at 0.01%. Durmet was active even at 0.01% on all different fungi.

The fungitoxic properties of pesticides whose primary defined targets are the insects may perhaps be due to their cuticular penetration abilities, which may extend to fungal mycelia. Precise mode of fungitoxic action of such chemicals needs to be elaborated by detailed studies. However, the additional property can perhaps be used to advantage in well-designed pest control strategies in various crops beset with both insect and fungal pests.

1. Leclerg, E. L., *Phytopathology*, 1964, **54**, 1309.
2. Bent, K. J., *Endeavour*, 1969, **28**, 129.
3. Singh Inderjit and Prasad, S. K., *Indian J. Nematol.*, 1973, **3**, 109–133.
4. Khare, M. N., Agarwal, S. C., Kushwaha, L. S. and Tomar, K. S., *Indian Phytopathol.*, 1974, **27**, 364–366.
5. Singh, R. K. and Dwivedi, R. S., *Pesticides*, 1988, **XXII**, 20–23.
6. Nene, Y. L. and Thapliyal, P. N., in *Fungicides in Plant Disease Control*, Oxford and IBH Publishing Co., New Delhi, 1979, p. 507.
7. Vincent, I. M., *Nature*, 1947, **159**, 850.

S. B. BHONDE
S. G. DESHPANDE
R. N. SHARMA

Entomology Section,
National Chemical Laboratory,
Pune 411 008, India

On incorrect use of Student's *t* test in bio-medical research

Student's unpaired *t* test is used to test whether the means of the two groups are statistically different or not. Student's unpaired *t* test is written as *t* test for simplicity in this article. Various types and aspects of *t* test like situations where the test is applicable, assumptions made, calculation procedure, advantages and limitations have been described in standard statistical books[1,2].

While perusing bio-medical articles, the author has come across various types of errors in the application and interpretation of *t* tests. The situations where *t* test is not 'valid' or if it is 'valid' it is with reduced power are defined as errors in this article. The purpose of this article is to illustrate these errors and to indicate correct statistical procedure to be adopted and alternative statistical tests to be used. It is hoped that this would enable research scientists having inadequate statistical knowledge to apply appropriate test correctly and to identify situations where expert statistician's help is essential.

Most of the examples presented in this article are taken from published bio-medical articles with application of *t* tests. No attempt has been made to obtain the raw (basic) data from the authors. The statistical values for appropriate procedures and alternatives are calculated whenever possible, otherwise only references are mentioned.

Definitions used in this study for comparison of two means: Null Hypothesis (NH): A statement concerning the values of a population parameter. Here the means of the two groups are equal.

α (alpha): The significance level of a test: The probability of rejecting the null hypothesis when it is true (or the probability of making a Type I error).

β (beta): The probability of correctly rejecting the null hypothesis when it is true (or the probability of making a Type II error).

Confidence level ($1-\alpha$): The probability that an estimate of a population is within certain specified limits of the true level.

Power of a test ($1-\beta$): The probability of correctly rejecting the null hypothesis when it is false.

Confidence interval of the difference: The probability that an estimate of a difference in two populations is within certain specified limits.

One-tailed (sided) test: In hypothesis testing, when the difference being tested is directionally specified beforehand,

i.e. (when $\bar{X}_1 < \bar{X}_2$, but not $\bar{X}_1 > \bar{X}_2$, is being tested against the null hypothesis $\bar{X}_1 = \bar{X}_2$).

Two-tailed (sided) test: In hypothesis testing, when the difference being tested for significance is not directionally specified beforehand, i.e. when the test takes no account of whether $\bar{X}_1 > \bar{X}_2$ or $\bar{X}_1 < \bar{X}_2$).

Type I and II error may be viewed as 'false positive' and 'false negative' of a diagnostic test.

P value: The *P* value is the probability of committing a Type I error if the actual sample value of the statistic is used as the rejection value. It is the smallest level of significance for which we would reject the null hypothesis for that sample. It is also interpreted as an indicator of the weight of evidence against the null hypothesis.

t test to non-normal data. The *t* test is valid and powerful, if and only if, the assumption of the normality is satisfied. It is always better to check normality of the data of small samples if the data consist of percentage values, counts and enzyme values. For heterogeneous data where the largest value is more than twice the smallest value and for non-negative values where SD (standard deviation) is greater than the mean, *t* test is not powerful.

On considering the data of Table 1, it can be seen that here SD (7.2) is greater than the mean (4.9) suspecting non-normality. However, it is rather difficult to detect non-normality in small samples. A thumb rule to suspect non-normality is to calculate 95% confidence interval (mean ± 2* SD). In Table 1, most of the values of the range are less than zero, thus confirming non-normality. Mild deviations from normality may be permitted for the *t* test analysis for larger samples (greater than 30). It is always advisable to check for causes of the high SD. Only one or two outlier (inconsistent) values may increase SD to a large extent. Under such circumstances, it is always better to ignore these observations in the analysis (after mentioning the reasons of it). If there are no outlier values signifying high SD as the inherent variation, nonparametric tests are more appropriate. These tests do not depend on the assumption of normality. In this study, nonparametric test like 'Mann–Whitney U test' with the median and range values as the summary measures would be more accurate. Another method would be to convert data 'normal' by suitable transformation (logarithmic, square root and inverse, etc.) and then apply *t* test. For the above example, the latter two tests would have given more appropriate result; perhaps even significant difference.

Table 1. Example: In the study of comparisons of GSH hormone levels in acutely ill patients and controls, the investigator applied unpaired *t* test for the following data

Group	Number (*n*)	GSH units mean ± SD	Range
Patients	15	4.9 ± 7.2	(1.3–30.0)
		NS | $t = 1.1$	
Controls	10	2.8 ± 1.7	(1.3–6.6)

Here the investigator would have probably detected statistical significance by nonparametric test.

t test to groups having unequal variances. The *t* test requires assumption of equal variances (homoscedasticity) between two groups to be satisfied. This is checked by *F* test. If the variances are equal by *F* test then the usual *t* test is correct, otherwise modified *t* test or nonparametric test should be applied. This can be explained by the following example:

In Table 2, calculated *F* value came out to be 15.3, which led to the conclusion that variances between the two groups were not equal. Under such circumstances, modified *t* test would be more appropriate. There are many types of modified *t* tests and one of the simplest and powerful would be 'Cochran's modified *t* test[3]'.

The formula:

$$t' = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{[(s_1^2 / n_1) + (s_2^2 / n_2)]}},$$

where s_1 and s_2 are SDs of the 2 groups with n_1 and n_2 observations respectively, and the cutoff value is given

$$t = \frac{[(s_1^2 / n_1) * t_1(\text{table}) + (s_2^2 / n_2) * t_2(\text{table})]}{\sqrt{[s_1^2 / n_1 + s_2^2 / n_2]}}.$$

For the above example table cutoff *t* value ($\alpha = 0.05$) came out to be 2.11. As the calculated *t* value was less than the cutoff *t* value (i.e. 2.07 < 2.11), the difference was not significant. By the usual *t* test, the difference was significant at 0.05 level but with the modified *t* test the difference turned out to be not significant.

Table 2. Example: In the comparison of hypothyroid and normal patients the investigator compared heart rate (part of the study) with *t* test for the following data

Group	Number	Heart rate (units) Mean ± SD
Hypothyroid	16	61.80 ± 2.48
		S, $P < 0.05$ | $t = 2.07$
Normal	20	66.55 ± 9.69

Here the investigator would have got no significant difference (instead of significant difference) by modified *t* test.

(Unpaired) t test to paired (related) data. Unpaired *t* test is not correct for the related data as it requires the assumption of independence between the two groups to be valid.

In Table 3, the values were on the same patients (i) before treatment, and (ii) after treatment. The after-treatment values were related to the before-treatment values, resulting in the violation of the assumption of independence. Hence the (unpaired) *t* test was not valid.

t test applied to more than two groups (without correction). The *t* test is appropriate in comparing means between two groups only. In this category of errors, *t* test is applied many times by comparing all possible pairs of means. This procedure has got the major drawback of increased Type I error. This means that if there are no real differences, the probability of at least one difference being statistically significant by chance at the 5% level can be considerably greater than 5%. As the total number of comparisons increases, the Type I error increases proportionately. This has been discussed in Table 4.

In Table 4, comparisons among six different A, B, C and D groups were made. If for each comparison, Type I error = 0.05 then the effective *P* value was 6*0.05 = 0.30 or 30%, i.e. there was about 30% chance of at least one incorrect conclusion regarding the difference in the groups. The appropriate procedure here would be to test the overall difference among all the groups

Table 3. Example: In the comparison of pepsin output at before- and after-treatment for the group X. Investigator applied *t* test for the following data (part of the data)

		Pepsin output (mean ± SD)		
Group	*n*	Before-treatment	After-treatment	Significance
X	17	1012.0 ± 375.5	863.08 ± 217.5	NS ($t = 1.4$)

Here the investigator would have probably detected significant reduction in pepsin output by paired *t* test.

Table 4. Example: Comparison of blood glucose levels (mean ± SD) in 4 different groups

Group	A	B	C	D
$n = 9$	84.67 ± 5.29	105.78 ± 9.77	93.11 ± 3.62	88.44 ± 8.05

Comparison between	Calculated *t* value	Significance by *t* test	Modified LSD with multiple correction
A–B	5.71	$P < 0.001$	$P < 0.001$
B–C	3.65	$P < 0.01$	$P < 0.01$
C–D	1.59	NS	NS
A–C	3.94	$P < 0.01$	NS
A–D	1.17	NS	NS
B–D	4.11	$P < 0.001$	$P < 0.001$

More sensitive modified LSD detected only 3 out of 4 significant differences obtained by usual *t* test.

Table 5. Example: Data are mean ± SD of the three variables for the two groups of patients with heart disease

Variable	Serum total cholesterol (*X*)	Serum phospholipids (*Y*)	Serum uric acid (*Z*)
Group 1	269.1 ± 60.41	11.92 ± 2.41	5.25 ± 1.00
$n = 17$	| NS	| NS	| NS
Group 2	235.8 ± 36.69	12.55 ± 1.74	4.88 ± 0.91
$n = 15$			

Taken from *The Design and Analysis of Clinical Experiments* (ed. Fleiss, J. L.), John Wiley and Sons, New York, 1986, p. 69. Here multivariate T^2 test detected significant difference between Group 1 and Group 2.

by 'ANOVA (*F*) test'. If and only if *F* test indicated the overall significant difference among the groups, comparison between specific groups could be made by different multiple comparison tests. The simplest and versatile would be 'Modified LSD test'. ANOVA assumes equality of variances (SD^2) among all the groups. For the above example, ANOVA test showed significant differences among the means of all the groups ($F = 11.11$, S, $P < 0.01$). Then to compare means of specific groups, 'Modified LSD test' was applied. Out of the four significant differences obtained by *t* test, only three were found to be significant by the 'Modified LSD test'. Multiple comparison tests protect against calling too many differences significant than does the usual *t* test. This is a rather complex situation where statistician's advice is recommended. In brief, only limited number of comparisons should be made to restrict Type I error.

Application of several t tests to many variables in a single study instead of multivariate test. This is a common error committed by many investigators working in bio-medical field. Most of the studies in bio-medical research consist of comparison between more than two variables of equal importance and interest. For such studies, a multivariate test that compares variables simultaneously would have many advantages over a series of separate *t* tests for each variable. The most important advantage of the former is the possibility of the increased power. If the variables are not much related, the multivariate test has a chance of finding significant differences among the treatments even if none of the (univariate) *t* tests are significant. The example discussed in Table 5 has been well illustrated in *The Design and Analysis of Clinical Experiments* (ed. Fleiss, J. L.), John Wiley and Sons, New York 1986, p. 69.

Thus, in Table 5 *t* test showed no significant differences for *X*, *Y* and *Z* variables between two groups, but 'Hotelling's T^2 test' (multivariate) detected differences ($T^2 = 12.04$, $P < 0.05$) between two groups after considering three variables as a set.

Errors in the computation of t test. In this category of errors, there were few cases where reported *t* values differed from the *t* values computed from the mean and SD. One of the most common errors was to apply the formula $t = (\bar{x}_1 - \bar{x}_2) / \sqrt{(s_1^2 / n_1 + s_2^2 / n_1)}$, instead of $t = (x_1 - x_2)/\sqrt{D}$, where

$$D = \frac{[(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2]}{(n_1 + n_2 - 2)} * [(1/n_1 + 1/n_2)],$$

for markedly different sample sizes.

Number of t tests to repeated measurement studies. Here number of *t* tests are applied at various time points between two groups to study the effect of treatment over a period of time. This procedure has: (i) successive observations in a given study correlated, and (ii) Type I error increased because of many *t* tests in a single study, leading to major defects. The appropriate statistical procedure here would be to calculate suitable summary measures like area under the curve (AUC) and time to maximum effect (T_{max}) etc. for each subject and then to compare mean AUC and mean T_{max} between the two groups by unpaired *t* test[4]. This method is commonly used in clinical pharmacology. Another better alternative would be to reduce number of *t* tests where AUC, T_{max} parameters are of little interest. ANOVA and multivariate test could be other alternatives. By increasing the level of significance from the conventional 0.05 to 0.01 or 0.001, one may arrive at an appropriate conclusion.

Errors in the interpretation of results. This is one of the most important and commonly found errors. This error occurs because of the improper understanding of the statistical hypothesis testing procedure and inadequate communication between investigator and statistician. In this testing procedure, Null Hypothesis (NH) (mean of the first group is equal to the mean of the second group) is tested by calculating mean, SD and t statistic. If the calculated t exceeds the table t value then the NH is rejected, leading to the conclusion 'there is significant difference between the two means'. This conclusion is expressed in probability (P) value. The P value is usually misunderstood as the % chances that the results are wrong or the total of Type I and II errors (expressed in probability). The P value is the probability of obtaining the observed results if the NH were true (assuming the assumptions required for the test to be true). Or it represents the chance that the difference found in the data is the result of random variation when there is no true difference in the population from which the samples have been drawn. For example, a P value of 0.05 indicates one chance in 20 that the trend shown in the data is the result of random variation. The 'power' of the test increases as the number of observations increases. Hence, the investigators obtain statistically significant difference only because of large n and fail to obtain significant difference only because of small n. The statistical significance testing procedure simply cannot prove NH of equality of two (sample) means. Exact P values together with 'confidence interval of difference', should be considered for meaningful conclusions. 'Meta analysis' is a newly-developed statistical technique in which several dissimilar studies with small sample sizes are combined together to draw valid conclusion.

One-tailed t test to get significant result. In this category of error, the investigator applies one-tailed t test to squeeze significance without enough justification. As one-tailed table t value is much less than the two-tailed table t value, the investigator usually gets significant result for one-tailed t test and not for two-tailed t test.

Errors in the design of experiment. These errors are rather complex and difficult to detect. One of the common errors is adopting unifactorial (study of one factor at a time and keeping other factors constant) instead of multifactorial design. Multifactorial designs in bio-medical research are more economical in terms of cost, animals (subjects), etc. and more informative than unifactorial design. The advantages of multifactorial design have been well illustrated by Wallenstein *et al.*[5] hence, these will not be discussed here. In brief, multifactorial designs can measure various separate effects and their interactions with each other which unifactorial designs cannot do. This is a rather difficult procedure and expert biostatistician's help should be taken at the time of planning the experiment. No sensitive or sophisticated statistical method can compensate for a badly planned experiment.

Student's unpaired t test is frequently recommended for comparing the means between the two groups only. The test is valid and powerful if data are 'normally distributed', groups are independent and with equal variances. For dependent data, paired t test is appropriate. If the investigator suspects non-normality of the data, non-parametric tests like 'Mann–Whitney U test' and 'Wilcoxon signed rank test' are more appropriate. In case of large SD, it is always better to detect outlier (inconsistent) values. Presence of only one inconsistent value can increase SD significantly to distort the conclusion. For more than two groups, 'modified t test', 'Analysis of variance' (ANOVA), 'Analysis of covariance' (ANACOVA) and 'multivariate test' are more appropriate tests. The limitation of t test is that it merely tests equality of two means but fails to gauge the magnitude of difference between means of the two groups. 'Confidence interval of the difference' may help the investigator to obtain minimum and maximum difference between the two groups, which is of tremendous use in bio-medical research. Hence t test together with '95% confidence interval of the difference' is usually recommended for greater insights into the data. For borderline significant differences like $P = 0.04$ and $P = 0.06$, results should be viewed cautiously. It should be remembered that a statistical conclusion about significance does not always agree with clinical significance in bio-medical field.

The overall implications of inappropriate use of t test are substandard and misleading research results, wastage of time, money, efforts, etc. With the easy access of statistical (computer) packages, there is a tendency among the research workers to apply t test 'blindly' (without considering various aspects of it). This article is mainly intended for bio-medical research workers with 'inadequate' statistical knowledge but desire to analyse data themselves. No statistical package can prescribe the most appropriate statistical test suitable for analysis of the given data. This does not mean that statistics is of overriding importance in bio-medical research, but it is an area where much improvement is desirable and beneficial for increasing the standard of research. Appropriate statistics should be viewed as an integral part of good bio-medical research.

1. Armitage, P., in *Statistical Methods in Medical Research*, Blackwell Scientific, Oxford, 1971, pp. 116–126.
2. Daniel, W. W., in *Biostatistics: A Foundation for Analysis in the Health Sciences*, John Wiley and Sons, New York, 1987, pp. 207–213.
3. Snedecor, G. W. and Cochran, W. G., in *Statistical Methods*, Oxford and IBH Publishing Co., New Delhi, 1967, p. 115.
4. Matthewes, J. N. S., Altman, D. G., Campbell, M. J. and Royston, P., *Br. Med. J.*, 1990, 300, 232.
5. Wallenstein, S., Zucker, C. L. and Fleiss, J. K., *Circ. Res.*, 1980, 47, 5.

A. S. ARFKAR

Haffkine Institute,
Mumbai 400 012, India