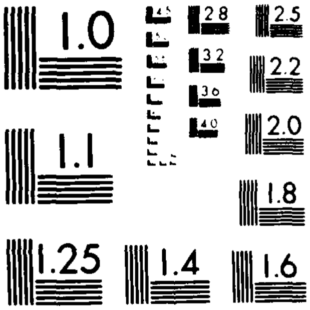


MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS 1963 A

AD-A136 019

...IFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
...NUMBER	2. GOVT ACCESSION NO. AD-A136 019	3. RECIPIENT'S CATALOG NUMBER
...(and Subtitle) ...istical Analysis in Dental Research Papers		5. TYPE OF REPORT & PERIOD COVERED Submission of papaer Jan- Aug 1983
		6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Lewis Lorton		8. CONTRACT OR GRANT NUMBER(s) 11
9. PERFORMING ORGANIZATION NAME AND ADDRESS US Army Institute of Dental Research Walter Reed Army Medical Center Washington,DC		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS US Army Medical Research and Development Command HQDA-IS Ft Detrick, MD 21701		12. REPORT DATE August 8,1983 13. NUMBER OF PAGES 20
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) unclassified 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

Statistics,graphs, T tests ,analysis

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

One volume of a dental journal was reviewed for errors and inconsistencies in statistical analysis and design. The errors were tabulated and correct methods were outlined

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

83 12 16 092

Statistical Analysis in Dental Research Papers

Lewis Lorton, DDS, MSD
Chf, Bioengineering Branch
USA Institute of Dental Research
Letterman Army Institute of Research
Presidio of San Francisco, CA 94129

A-1

Research papers are usually subjected to intense scrutiny for form and content of the text. Strict guidelines apply to format, references, and even footnotes but there seems to be little or no guidance on the statistical analysis of the data. Often-times the statistical analysis or display in the dental research literature obscures the conclusions of the research itself. The purpose of the statistics in a research paper is two-fold: first, to summarize and describe the data and secondly, to test some hypothesis about the data. The statistical errors cited in this review article fall into two main categories-- errors of omission and errors of commission. This paper will give statistical standards applicable to research papers and will explore the misunderstandings which cause the common errors. The simple errors occurring in one volume of the Journal have been tabulated and will be presented to show the ubiquity of errors in even this intensively edited journal.

Previous Reviews

Glantz (1) reviewed 79 original research papers published in 1977. He discovered that, of the 59 papers that incorporated statistical analysis, 46 (or 61%) used the t test incorrectly. The majority of the errors involved "inappropriate use of the t test in a way that often leads the authors to assert that a treatment produced an effect when the data do not support such a conclusion." Schor and Karten (2) found only 28% of analytical papers statistically acceptable in their final form. Gore et al (3) in an epic series of articles mentioned that 52% of the papers published in the prestigious British Medical Journal in the 13 issues published from January to March of 1976 contained at least one statistical error. Badgley (4) reviewed 103 original research papers in the Canadian Medical Journal and The Canadian Journal of Public Health. In 43 of them he found questionable use of statistical inference in drawing conclusions.

Dental researchers probably suffer, in the same proportion as other scientific researchers, from the tendency to make errors of analysis and display. Rare indeed is the journal that provides guidance for its authors in statistical display or analysis. To provide a basis for writing and editing, a short set of standards has been adapted from the Guidelines For Reporting Clinical Trials as presented at the International Conference on Clinical Trials of Agents Used in the Prevention and Treatment of Periodontal Diseases (5).

Standards

summary data must be included

statistical methods that are appropriate for the data should be used

the exact statistical test(s) should be named

statistical displays (tables,graphs) should be completely labelled.

These standards represent probably the minimum that should be required for the presentation of data to the readers of a scientific journal.

Error Categories

The errors in the literature fall in the following categories:

- summary data incomplete
- incompletely or incorrectly labelled figures
- statistical test not named
- inappropriate statistical method
  - most common error- misuse of Student's t test
  - next most common error- standard error used in place of the standard deviation

Errors of Omission

In the first three categories the errors are those of omission, where the author/researcher has failed to provide his audience with the bases used for making the analyses and conclusions.

Because of space limitations it is usually impossible to present an entire set of results, and so it is important to give summary statistics which will inform the reader about the data. These summary statistics should usually include some information about the location of the main body of the results. When data are composed of measurements and the distribution is roughly "normal," the proper measure of the center or the central tendency is usually the average or mean. In data sets where the results are in ranks, the median is the more appropriate measure of central tendency. Some information about the distribution or dispersion should also be given; this can be the range of the data or the standard deviation (these are separate measures and are not synonymous). The number of observations in each sample should also be given in the text and in the legends of figures where other statistical measures are given.

Figures should be completely labelled with the magnitude and identity of the levels. In cases where brackets may be drawn above and below a bar on a bar graph (as in fig. 1), or where a mean is given as ± some figure, the meanings of these brackets or values should be clearly stated. These items should be identified as the standard deviation, the range, the 95% confidence interval or whatever they are.

The exact statistical test should be named. There are many tests which may be applied to each data set and these tests may provide inappropriate answers if they are applied incorrectly. For example, analysis of variance techniques can provide inaccurate and misleading results when used with data that is not normally distributed.

In the instances where summary statistics are not given or the statistical test is not named and significant conclusions are drawn, the interested reader has no opportunity to evaluate the real significance of the results and the accuracy of the conclusions.

These standards are the absolute minimum levels of statistical information which must be included in any paper that is using or implying statistical methods.

Errors of Commission

The effects of the presence of inappropriate statistical methods are not as easily explained as errors of omission. Presentation of some statistical background is necessary.

Mis-use of Student's t tests: The Student's t test is the most commonly used statistical test in the biomedical literature (2).The name of the test is often misspelled as Student t test or student's t test."Student's t test" is an eponym derived from the pen name of the statistician, R.A. Fisher, who wrote under the name of A. Student. In an unmodified state, the use of this test should be limited to the comparison of two means that have been derived from samples of relatively normally distributed, independent, interval data.

Why should the use of multiple t tests be avoided? The reason for this is based on the concept of Type I error and the need to control its size.

When it is printed in a research paper that "these groups were different at a .05 level of significance" the real meaning of the phrase is not obvious. What is really meant by this statement is -if these two groups were actually samples from the same normally distributed population , the difference between their means would only be due to the chance of sampling variation. This observed difference between their means is so large however that it would occur only .05 of the time by chance alone. This is rather a small possibility, and so it is concluded that the two sample groups are not from the same population, and we accept the fact that they are different. However, 5% of the time chance will cause a difference between the means that is so large that we will accept that the sample groups are from different populations , that there are "statistically significant" differences, when they are in fact from the same group. We accept a .05 or 5% chance that we are making an error in saying they are different when they are the same- this is known as Type I error- saying that two groups are different when they are , in reality, the same. By setting the size of the difference between the groups that we will consider significant we control the amount of Type I error we encounter.

For example: if we possessed a large barrel of apples from the same tree, it could be comfortably said that these apples are from the same homogeneous population. Yet they differ in weight due to natural variation; intuitively, we find this acceptable. Suppose we take a sample of 10 apples from the barrel, weigh them, find the average weight per apple and return them to the barrel. Then we proceed to remove another sample, randomly selected from the barrel, and repeat the weighing, recording, returning, and reselection.* We will eventually select weigh and record every possible 10 apple sample from the barrel. If the average weight per apple of these samples were graphed (figure 2) they would fall into a relatively <u>normal</u> distribution- the beloved bell-shaped curve. Note that every point on the distribution is produced by the <u>mean</u> of a sample of ten apples drawn from the parent population. Two vertical lines can be drawn on the curve so that that 95% of the means fall inside the lines.

*if the barrel held 250 apples, there are approximately 2.19 x 10^{17} possible different groups of 10 apples.

Suppose that we were presented with a 10 apple sample (a from fig.2) and were asked to decide statistically whether this group was from the original barrel. We would set 5% as our significance level. We would weigh the sample and determine the mean weight and standard deviation. For apple sample a, the mean would fall within the 95% limits; thus we would accept that this sample comes from the same homogeneous barrel group. If however we were presented with mean b, we would see that the mean of this sample falls outside of that range where 95% of all the means lie. Thus the chance of sample b coming from this barrel is less than 5% and we conclude that sample b <u>does not come</u> from the barrel group even though, in reality it does!! Thus when comparing groups that are equal we conclude 5% of the time that they are different due to the large difference between the means that has occurred by chance alone.

The chance of deciding that differences are significant and caused by some outside influence when ,in fact, the differences are caused by random chance fluctuation is known as Type I error. Since this Type I error can produce disastrous erroneous conclusions in research a great deal of effort in statistical analysis is concerned with keeping the Type I error to its stated or nominal 5%. This percentage is usually the maximum Type I error that most researchers will accept. They will not consider groups as different unless the difference in their means would occur 5% (or less) by chance alone. However if <u>t</u> tests are repeated on several pairs of means from the same data set the chance of declaring a result significant when it is not significant increases with the number of tests performed.

For example, a researcher who has five treatment groups and decides to compare all possible combinations with t tests will need to perform ten comparisons. He may assume that the chance of finding a statistically significant difference that is due only to chance variation in the samples is limited to .05. However, by repeating the test ten times on different pairs of means from the same population, he has increased his chances of making a Type I error (table 1). If there was really no actual difference amongst the five groups and any differences were caused by chance variation, by doing ten tests he would have a 34% chance of finding a seemingly "significant" difference that had been produced by chance variation alone.

Thus when multiple t tests are done the opportunity of declaring a difference to be significant when, it is in fact not significant, increases. For this reason, a researcher should avoid performing large numbers of t tests.

Ordinarily a data set which is divided into several treatment groups can be better analyzed by using Analysis of Variance (ANOVA) techniques. These ANOVA techniques test any number of groups for significant differences and the Type I error can be controlled to eliminate some of the chance of making undiscoverable mistakes.

Standard Error used as a measure of dispersion: The possibility for misleading the reader when the standard error is used as a measure of dispersion instead of the standard deviation is sizeable and the situation should be clarified. The standard error , or the standard error of the mean (SEM) is sometimes confused with the standard deviation. These two similar sounding terms do not describe the same entity and are not interchangeable. Glantz (1) has neatly summarized the sense of the standard deviation.

> When the variable being observed behaves so that any given observation is equally likely to be above or below the mean,and more likely to be near the mean than far from it, it makes sense to quantify the spread of values using the standard deviation. Under these conditions, the standard deviation has the useful property that roughly 68% of the observations will be within 1 standard deviation of the mean and roughly 95% of the observations will be within 2 standard deviations of the mean. This property makes the standard deviation a good way to summarize the varibility in data with a single number.

The standard error is a numerical relative of the standard deviation. It is produced by dividing the standard deviation of a group of observations by the square root of the number of observations.

$$SEM = SD/\sqrt{N} \qquad \text{(in certain situations n-1 is used)}$$

Extensive descriptions of the meaning of the standard error and its role in statistical analysis are given in textbooks. In Glantz (1) there is an excellent short description of its theoretical background. Let it suffice to say here, that the standard error does not measure the distribution of observations within a group; it is used to estimate the theoretical variability between means and is therefore used ,in practical situations, only in testing between means.

The common formula for the Student's t test is given below. The denominator of the equation is the pooled standard error of the two groups.

$$t_s = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

The standard error serves as an intermediate numerical function in the use of t-tests to describe the relationship between pairs of means. The SEM usually is not computed for a single group but the standard deviations of two groups are combined and a "pooled standard error" constructed. It is not, repeat not, an intuitive description of the distribution of the observations from an experimental group because of its relationship through the inverse of the square of the number of the number of observations.

For example:

A researcher measures two groups of apples. The larger group (100 apples) has a mean weight/apple of 200 grams and a standard deviation of 30 grams. The smaller group (50 apples) has a mean weight of 180 grams and a standard deviation of 38 grams.

Mean	number of observations	SD	SEM
200	100	30	3
180	50	38	5.4

The differences between the standard deviations of the two samples, which seemed so benign (38 vs. 30), seem much more imposing when transformed to standard errors (5.4 vs. 3).

If the standard error is not a good intuitive descriptor of the group dispersion and if it usually exists in the form of a pooled standard error of two groups, why then does its use persist in the scientific literature?

First, habit- researchers have been using this term ,SEM , for a long time and see no good reasons to quit; secondly, the standard error is related mathematically to the standard deviation (SD/ n) and so its use is not logically offensive; thirdly, the use of the standard error on graphs or in tables makes the dispersion of the data seem less than if the SD is graphed (fig.2). This is an arithmetic maneuver to decrease the visual effect of the dispersion of the data.

These same three reasons can be reconstituted as arguments against the habitual use of the standard error: first, tradition- the old habits, unfamiliarity with statistics, an unwillingness to change; secondly, the arithmetic relationship (the inverse square root of the observations) is too convoluted for the casual reader to recalculate; and thirdly, the use of the SEM in graphs or in tables may obscure the real nature of the results by understating the variations in the data.

When authors choose to use the standard error of the mean for data display they divide the readership into two groups. The first group understands what an SEM is and knows how to reconstruct the standard deviation from it (SD = SEM x$\sqrt{n}$). This group is merely inconvenienced in reviewing the research. The second group cannot distinguish between a SEM and a SD; this group is misinformed.

Since investigators publish to inform an interested public, they should use the clearest, correct means of doing so.

Survey of the Literature for Statistical Inaccuracies

The articles in one complete volume (vol #48) of the Journal of Prosthetic Dentistry was read by this author with the intent of classifying the papers into one of four categories. Narrative: the philosophy of a particular technique, theory, or method. Clinical or Case Report: techniques or methods, or unusual diagnostic situations under clinical conditions. Research: qualitative research, as in joint tomography, where the results and conclusions are not amenable to statistical analysis. Research/statistical: the results of this paper are expressed in some quantitative way which are, or should be, analyzed by the authors to provide some support for the conclusions.

The papers classified into the research/statistical category were re-read. Instances were tabulated where the papers did not meet the previously outlined standards. No attempt was made to evaluate the experimental design or to determine if the conclusions were derived from proper interpretation of the data analysis.

Results

The tabulation of the incidence of errors is presented in raw form in Table 2. The distribution of the types of papers in each editorial division of the journal is shown in Table 3.

Comment

A review of a single volume of a scientific journal whose readership is composed of both scientific researchers and clinicians revealed some specific failings in the statistical display and analysis of the research papers. These papers were not reviewed with the intent to criticize the experimental design or the conclusions drawn from the data and the analysis. Tabulations were made of instances where omissions of summary data or errors of fact made the presentation unsatisfactory according to rather inexacting standards stated for this review. These standards were derived from guidelines for statistical display and analysis published previously in other journals of biomedical research and communication.

A number of simple errors in statistical display and analysis appeared in these articles. Apparently the same exacting scholarship which is applied to the research efforts is not brought to bear upon the analysis of the data. How may readers exercise critical judgement of research papers if the necessary summary data are not given? Why should a reader have to calculate values or transpose information from the text to understand a figure/graph?

Dedicated researchers do not have the intent of producing inadequate or inaccurate papers; yet there are some inadequacies in either education or application of statistical practice that degrade the level of published dental research papers. A sudden increase in the level of statistical education that will decrease the occurrence of statistical errors is unlikely. Since a scientific journal vouches for the adherence to form of the articles it publishes, should the editors assume responsibility for statistical "consciousness raising" and verify the accuracy of the presentation?

The editors of Circulation Research, after they had reviewed the article by Glantz (1) analyzing the statistical state of articles in their journal, reacted by changing the editing process

> so that published papers use statistical tests in the proper way. It is likely that some of our reviewers cannot or will not provide an appropriate critique of statistical evaluation of data or comment on the need for the statistical evaluation of certain data. For this reason, we are adopting the following policy. Initial review of all manuscripts will be carried out as in the past. When it appears that a manuscript is likely to be accepted, a statistician will review it to determine whether a statistical method should be employed, or whether its use is appropriate. This may prolong the reviewing process, but we

> believe the delay will be small. We believe that such a delay will be justified, as it will insure that the information presented in papers published in the Journal has been analyzed and interpreted appropriately.(7)

The use of statistical analysis in research papers should be as well defined and elegant and as thoughtfully considered as the research itself. The statistical analysis should add to the credibility and clarity of the paper, not detract from it. The evaluation of the statistical portions of a research paper is as difficult and necessary a process in the critical presentation and review as any other portion of the activities of a researcher, author, reviewer, or editor.

References

1.Glantz SA: Biostatistics:How to detect, correct and prevent errors in the medical literature. Circulation 61:1 , 1980.

2.Schor S, Karten I: Statistical evaluation of medical journal manuscripts. JAMA 195:1123, 1966.

3.Gore S, Jones IG, Ritter EC: Misuses of statistical methods: critical assessment of articles in B.M.J. from January to March, 1976. Br.Med.J. 1:85,1977.

4.Badgley RF: An assessment of research methods reported in 103 scientific articles from two Canadian Medical journals. Can Med Assoc J 85:246, 1961.

5.Guidelines for Reporting Clinical Trials. J.Am.Dent.Assoc. 87:557, 1973.

6.Feinstein AR: Clinical Biostatistics.IIV. *Survey of statistical* procedures in general medical journals. Clin Pharmacol Ther 15:97, 1974

7.Rosen MR, Hoffman BF: Statistics,biomedical scientists, and Circulation Research (an editorial). Circ Res 42:739, 1978

Table 1.

Increased chance of Type I error with multiple t- tests

Number of tests	nominal error	real error
1	.05	.05
2	.05	.08
3	.05	.14
4	.05	.17
5	.05	.21
6	.05	.23
8	.05	.28
10	.05	.32

Table II. Statistical Applications (Uses and Misuses) in Volume 48, Journal of Prosthetic Dentistry*

Initial page of article	Statistical Application	Comment/Explanation
23	Incomplete or incorrectly labelled figures Multiple *t*-tests	Used multiple *t*-tests; did not give value of the statistic; used an asterisk (*) to denote significance in Table III, and *non*significance in Table IV
48	. . .	Nice use of ANOVA; well-labelled figures
52	. . .	Nice use of ANOVA; well-labelled figures
135	Summary data incomplete	No indication of variability of data
159	Incomplete or incorrectly labelled figures	Measure of dispersion not labelled
163	Summary data incomplete Statistical test not named	No measure of dispersion; test alluded to but not named; test statistics not given
171	Summary data incomplete Incomplete or incorrectly labelled figures	Measure of dispersion not given in text or labelled on figures; tests alluded to but not named
237	Incomplete or incorrectly labelled figures Statistical tests not named	No measure of dispersion given on fiures; tests not named although they seem to be multiple *t*-tests
282	Summary data incomplete Incomplete or incorrectly labelled figures	Measure of dispersion given as ± but not named
285	Multiple *t*-tests	ANOVA would be more appropriate
289	Multiple *t*-tests	ANOVA would be more appropriate
292	. . .	Uses a multiple Kruskal-Wallis test for ranked data which may cause same error as multiple *t*-test
377	Multiple *t*-tests	ANOVA techniques would be preferred; 16 *t*-tests done
388	. . .	Excellent use of nonparametric tests
401	Multiple *t*-tests	ANOVA would be more appropriate
424	Incomplete or incorrectly labelled figures SEM used when SD indicated	SEM used as indicator of dispersion on figures but not labelled
451	Summary data incomplete	Measure of dispersion given as ± but not labelled
492	Summary data incomplete Multiple *t*-tests	Measure of dispersion given but not identified
555	Summary data incomplete Incomplete or incorrectly labelled figures Statistical test not named	No measure of dispersion given in text; measures indicated on graph are not identified
575	Summary data incomplete	No measure of dispersion given
610	. . .	Sophisticated use of multiple regression
640	Summary data incomplete	Although ANOVA techniques were used, measures of central tendency and dispersion are not given
647	Statistical test not named Multiple *t*-tests	No values for the *t*-statistics (*if* that was the test done)
676	Summary data incomplete	± value given but not identified; good example of ANOVA and post hoc testing
681	Multiple *t*-tests	ANOVA would be preferable
686	. . .	Well-detailed description of statistical analysis
719	Multiple *t*-tests SEM used when SD indicated	Two-way ANOVA should be done

*Total number of articles with statistical analyses = 48. Number of articles without one or more of the five error categories = 27 of the 48. Number of articles with statistical errors = 21. Total errors identified in five categories = 34, i.e. Summary data incomplete = 10; Incomplete or incorrectly labelled figures = 7; Statistical test not named = 5; Multiple *t*-tests = 10, SEM used when SD indicated = 2.

. . . Six articles cited as good examples of statistical applications.

Table 3

Tabulation of Article Type by Editorial Section

	Article Type				
Editorial Section	Narrative	Case Report	Research	Research/statistical	Total
RP	2	5	2	11	20
FP/OP	7	8	5	22	42
MP	3	16		2	21
TMJ	2	3		8	13
R/E	9	1		4	14
DT	4	8		1	13
Tips		10			10
	27	51	7	48	133

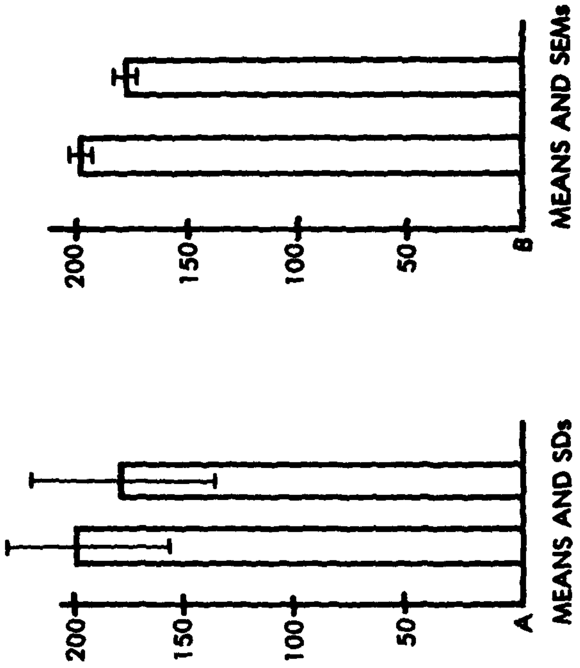
(RP-removable pros; FO/OP- fixed pros/operative; MP- maxillofacial pros; TMJ-temperomandibular joint; R/E- research & education; DT-dental technology;)

Legends

fig.1 When the same two data groups are displayed with the SDs and the SEMs, the bar graphs with the SEMs seem to have a smaller distribution of the observations. This may be misleading to the casual reader who expects to see the SDs, or may not know the difference.

fig.2 A "normal" curve with the 95% limits indicated. In normal distributions, 95% of all the observations fall within 1.96 SDs of the mean. That leaves 5% of the observations that fall outside the "normal" range.

COMPARING THE VISUAL EFFECT OF STANDARD DEVIATION AND STANDARD ERROR FROM THE SAME GROUP



05

"NORMAL" CURVE