Some Malpractices in Medical Statistics

S. JAMES KILPATRICK, JR.

Department of Biometry, Medical College of Virginia, Richmond

Since it is now fashionable for papers in medical journals to contain statistical notations, it also follows that in a certain fraction of these the statistical content is wrongly applied. These malpractices may be classified as numerical, statistical, and methodological. To illustrate some of the most commonly occurring errors the following examples are given.

Numerical Conventions

The number of decimals given in an observation should show the accuracy of the measurement. For example, blood pressure is measured to the nearest mm Hg. However, if one comes across a systolic blood pressure of 123.2, one is entitled to expect that the person used a sensitive manometer which could read to tenths of a mm Hg. Another example, which is especially relevant at Medical College of Virginia, is the question of how many decimals should be shown in students' standard scores.

It is generally assumed that, when repeated observations are made, the person collecting the data is measuring to the same degree of accuracy throughout. The same number of decimal points should therefore be given. Thus, if one were describing the elevation of systolic blood pressure, where this is calculated as "after treatment minus before treatment," one might measure to one decimal, e.g., 0.4 mm Hg. In this case one would expect that a zero reading be given as 0.0. This would mean that no detectable difference was observed to the nearest tenth mm Hg. An anomaly arises when measurements made to the nearest fraction are converted into decimal notations. Thus, if body weight was measured to the nearest quarter pound because the quarter pound was the smallest weight on a scale, one might want to record a weight of 170¼ as 170.25 lb. This figure, however, suggests that weight was measured to the nearest hundredth of a pound, which is not the case. There is no easy solution to this anomaly except to work in the basic units, in this case quarter pounds. In this way decimals are eliminated: 170¼ would then be given as 641 quarter pounds.

Another useful convention regarding the numerical presentation of data is that comparative statistics should be expressed to the same base. In a recent article on cystic fibrosis of the pancreas, an author used on the same page odds, fractions, and percentages as are exemplified in the following hypothetical extracts:

"From an earlier paper we showed that the odds of having a further affected child were 1/13. In this study the proportion of affected children (excluding the propositus) was 8/57." The difficulty of comparing 1 in 13 with 8 in 57 can be alleviated by writing "From an earlier paper we showed that the proportion of having a further affected child was 0.1 or 1/14. In this study the proportion of affected children (excluding the propositus) was 0.14 or 8/57." It is now apparent that the current study revealed a slightly higher frequency of affected children than the older study, but that this was unreliable because of the small number on which the first estimate was based, as is revealed by its expression to only one decimal point.

"The pH of sweat in 11 of the normal sibs was measured, and this was found to be elevated in 5 (45.45%)." This might better be written as "The pH of sweat in all of the normal sibs was measured and this was found to be elevated in about 0.5 (5/11)." The well-known tendency to express everything in terms of a percentage leads to a spurious degree of sensitivity if the base is much less than a hundred. Moreover, percentages are often grossly misleading especially when no denominator is given. Consider the statement "43% of patients in the current study with regional ileitis had blood group O. This was lower than the 56% reported by our earlier study." On the face of it this would suggest that regional ileitis is changing its relationship to blood group O. However, when one realizes that the 43% is based on 3 patients out of a total of 7 having blood group O and the 56% based on 5 out of 9, the difference is immediately seen to be unimportant.

Wrong Denominators

Percentages may also be misleading because they are expressed in terms of the wrong denominator. Mainland (1964) quotes an example taken from the *British Medical Journal* in which 139 members of the Woman's Royal Air Force who showed temporary amenorrhea

used. Perhaps the most serious criticism of the use of the mean is that it covers up what was actually done. A mean may not reveal, for example, that three observations of a given response were made and the figure reported represents the average of the two of these which were closest together. Here the investigator is essentially throwing away one-third of his information. Again, an average says nothing about the underlying distribution which is too often assumed to be normal.

Estimations of Accuracy

A convention has become established of reporting statistics plus or minus a small number. This convention may have been borrowed from engineering or laboratory sciences in which the number following the plus or minus sign is an estimation of the degree of accuracy of the foregoing figure. However, nowadays this small figure usually represents a statistical estimate of variability. The question immediately arises of what estimate this is. Thus, the mean hemoglobin of four aliquots of blood may be given as 10.5 ± 0.2 grams per 100 ml. This figure of 0.2 could either represent a measure of variability (the standard deviation) of the four observations around the mean, or it might refer to an estimation of the variability of this mean and others, based on four estimations around the true value of hemoglobin for this pool of blood (the standard error). One can only discriminate between these two alternatives in the light of other information given in the report or in the context of the use of the 10.5 ± 0.2. There have been occasions in which statistic comparable to 0.2 was calculated as the standard error of the mean, that is, it represented the accuracy of the mean about the true value, but this figure was subsequently used as though it described the variability of the original observations about the sample mean, giving thus, a spurious degree of reproducibility to the technique.

Assumptions of Normality

Parametric statistical techniques are based largely on the assumption of an underlying normal distribution. In a large sample, the assumption of normality can be tested directly. Most medical applications of parametric statistical methods are made, however, to small samples in which the normal assumption cannot properly be tested. Many practitioners of statistics today prefer not to have to rely on an assumption as the cornerstone of their analytical methods. Hence, the increasing tendency to use non-parametric or distribution-free procedures. The results of this new approach are reflected in statistical tables. For example, *Documenta Geigy* (1962) gives exact confidence limits for a sample proportion and exact χ^2 values for 2×2 contingency tables for sample sizes up to 60.

Inappropriate P Values

The use of P in medical journals has become so widespread that it is perhaps useful here to redefine how this is used and what it means. In terms of a comparative trial such as the comparison of two drugs, one may assume generally, for example,

1) that the two treatments have in reality no different effect,

2) that patients or subjects are allocated strictly at random to one or other of these treatments,

3) that the distribution of the response to therapy follows a normal distribution,

4) that the variability of responses is the same in the groups compared.

The statement "$P < 0.05$" then means that the probability is less than 5% of finding a difference as great or greater than that observed due to random sampling variation. Such a low probability is interpreted as spuriously low because one of the four assumptions is not warranted. If the first assumption is wrong, then in fact there is a difference between the effects of treatment.

The statement "$P < 0.05$" clearly then does not *prove* the reality of treatment differences. Other interpretations are always possible unless the other three assumptions are known to hold. Even then, we always have the possibility (even though this is unlikely) that the observed difference was due to random sampling variation and that no real treatment difference exists.

Probability values should not then be calculated or quoted when the four assumptions given above or some other set of assumptions previously specified are unrealistic. An example of the inappropriate use of P (unhappily this is a real example) occurred in a teaching handout to medical students in a British university. In this handout, statistics were given on the differential death rates from leukemia in males and females. This was followed by the statement, "These sex differences are clearly significant ($P < 10^{-10}$)." Such a statement is wrong since the equivalent assumptions to 2), 3), and 4) in this situation are not warranted. Moreover, the statement is superfluous. There is clearly a difference in these population-based sex-specific death rates from leukemia.

Significance of Repeated Tests

Not only are tests of significance inappropriately applied to surveys, but often by the multifactor nature of the data and the lack of specific hypotheses, *batteries* of significance tests are run rather than the appropriate multivariate analogue. Example: Suppose there are 15 different items in a survey. An investigator (especially one who has ready access to a computer and a suitable program) might ask for correlations between every pair of variates. In

used. Perhaps the most serious criticism of the use of the mean is that it covers up what was actually done. A mean may not reveal, for example, that three observations of a given response were made and the figure reported represents the average of the two of these which were closest together. Here the investigator is essentially throwing away one-third of his information. Again, an average says nothing about the underlying distribution which is too often assumed to be normal.

Estimations of Accuracy

A convention has become established of reporting statistics plus or minus a small number. This convention may have been borrowed from engineering or laboratory sciences in which the number following the plus or minus sign is an estimation of the degree of accuracy of the foregoing figure. However, nowadays this small figure usually represents a statistical estimate of variability. The question immediately arises of what estimate this is. Thus, the mean hemoglobin of four aliquots of blood may be given as 10.5 ± 0.2 grams per 100 ml. This figure of 0.2 could either represent a measure of variability (the standard deviation) of the four observations around the mean, or it might refer to an estimation of the variability of this mean and others, based on four estimations around the true value of hemoglobin for this pool of blood (the standard error). One can only discriminate between these two alternatives in the light of other information given in the report or in the context of the use of the 10.5 ± 0.2. There have been occasions in which statistic comparable to 0.2 was calculated as the standard error of the mean, that is, it represented the accuracy of the mean about the true value, but this figure was subsequently used as though it described the variability of the original observations about the sample mean, giving thus, a spurious degree of reproducibility to the technique.

Assumptions of Normality

Parametric statistical techniques are based largely on the assumption of an underlying normal distribution. In a large sample, the assumption of normality can be tested directly. Most medical applications of parametric statistical methods are made, however, to small samples in which the normal assumption cannot properly be tested. Many practitioners of statistics today prefer not to have to rely on an assumption as the cornerstone of their analytical methods. Hence, the increasing tendency to use non-parametric or distribution-free procedures. The results of this new approach are reflected in statistical tables. For example, *Documenta Geigy* (1962) gives exact confidence limits for a sample proportion and exact χ^2 values for 2×2 contingency tables for sample sizes up to 60.

Inappropriate P Values

The use of P in medical journals has become so widespread that it is perhaps useful here to redefine how this is used and what it means. In terms of a comparative trial such as the comparison of two drugs, one may assume generally, for example,

1) that the two treatments have in reality no different effect,

2) that patients or subjects are allocated strictly at random to one or other of these treatments,

3) that the distribution of the response to therapy follows a normal distribution,

4) that the variability of responses is the same in the groups compared.

The statement "$P < 0.05$" then means that the probability is less than 5% of finding a difference as great or greater than that observed due to random sampling variation. Such a low probability is interpreted as spuriously low because one of the four assumptions is not warranted. If the first assumption is wrong, then in fact there is a difference between the effects of treatment.

The statement "$P < 0.05$" clearly then does not *prove* the reality of treatment differences. Other interpretations are always possible unless the other three assumptions are known to hold. Even then, we always have the possibility (even though this is unlikely) that the observed difference was due to random sampling variation and that no real treatment difference exists.

Probability values should not then be calculated or quoted when the four assumptions given above or some other set of assumptions previously specified are unrealistic. An example of the inappropriate use of P (unhappily this is a real example) occurred in a teaching handout to medical students in a British university. In this handout, statistics were given on the differential death rates from leukemia in males and females. This was followed by the statement, "These sex differences are clearly significant ($P < 10^{-10}$)." Such a statement is wrong since the equivalent assumptions to 2), 3), and 4) in this situation are not warranted. Moreover, the statement is superfluous. There is clearly a difference in these population-based sex-specific death rates from leukemia.

Significance of Repeated Tests

Not only are tests of significance inappropriately applied to surveys, but often by the multifactor nature of the data and the lack of specific hypotheses, *batteries* of significance tests are run rather than the appropriate multivariate analogue. Example: Suppose there are 15 different items in a survey. An investigator (especially one who has ready access to a computer and a suitable program) might ask for correlations between every pair of variates. In

all, he has asked for $(15 \times 14)/2 = 105$ values. If the 5% level of significance is applied throughout, approximately five of the 105 values will be sufficiently large to be judged technically significant even though the 15 items are uncorrelated in the population.

Inclusion of Pilot Data in a Subsequent Experiment

This is another malpractice somewhat similar to the above. The experimenter, by inclusion of pilot data, tends to prejudice the result of the experiment in terms of a favorable result. If a full-scale experiment is done, this is often because the results in the pilot have been encouraging. By adding pilot data, the experiment is already halfway towards technical significance. The analysis of a fullscale experiment should not, therefore, incorporate the pilot data except after deep consideration on the effects of such inclusion on the results.

Indices and Ratios

There are 125 indices listed in *Dorland's Illustrated Medical Dictionary*. This shows how fashionable it is to construct an index to report results. Statistically, there are a number of reasons why the use of indices should be avoided where possible. Consider a situation in which there are p different responses and q concomitant factors. Let the responses be $y_1, y_2, \ldots y_p$ where $p \geq 1$, and the concomitant factors be $x_1, x_2, \ldots x_q$ where $q \geq 0$. According to the situation and the number of y's and x's, the research worker tends to use a simple ratio of y/x, or a weighted sum Σwy of the y's, or a combination $\Sigma wy/x$ or $\Sigma wy/\Sigma wx$.

Ratios of y/x of a response y to a concomitant variate x are used extremely frequently, especially in therapeutic experiments, e.g., dose per kg body weight, or in the response to treatment of a part to the whole, e.g., change in weight of an organ/change in body weight. Two statistical criticisms of the use of ratios are that the ratio of two normal variates is not necessarily normal, and that the use of a ratio assumes a linear relationship. Rather than assuming a proportional relationship, it is better to estimate the relationship from the raw data. The original analysis of total acid output (T.A.O.) in rats on different doses of thyroxine (Blair et al., 1965) used T.A.O. mg/100g body weight. This was, however, unnecessary since further analysis showed body weights did not differ significantly among tested groups. If they had differed, the estimated regression of T.A.O. on body weight could have been used.

Weighted sums occur when there are a number of responses to be summarized and are of the form Σwy where w represents the relative weight. Examples occur in diagnostic indices, e.g., in the clinical diagnosis of thyrotoxicosis (Crooks, Murray, and Wayne, 1959), in the diagnosis of rheumatoid arthritis (Mainland, 1964). Another example is the combination of standard scores of medical students. There is a dangerous tendency today to arbitrarily score subjective impressions. This leads to pseudo-quantification. In adding these scores, much information is lost in the process. Moreover, an additive combination is not necessarily the best because of the non-independence of different signs or symptoms. The determination of weights may also be made on an extremely ad hoc basis. Thus, it is often better to use a multiple classification or, if the responses are measured, to use multivariate techniques which, with the advent of fast digital computers, are becoming increasingly practicable.

Indices of the form $\Sigma w(y/x)$ *or* $\Sigma wy/\Sigma wx$, i.e., a weighted sum of ratios or the ratio of two weighted sums. An example of the first occurs in a retrospective study of births (Gruenwald and Minth, 1961). The authors quote a mean ratio of placenta weight (PW) to a birth weight (BW). This might be expressed as $(1/n)\Sigma(PW/BW)$ where n is the number of births. (Unfortunately, the authors mistakenly calculate the mean placenta weight divided by the mean body weight.) The index formed from the ratio of two weighted sums is best exemplified by the Standardized Mortality Ratio (S.M.R.). This compares the mortality in an occupational or other group relative to a standard population. The most common misuse of this index is to form the ratio of two S.M.R.'s which has little meaning or justification since the weighting systems in two S.M.R.'s are different (Kilpatrick, 1963).

The use of indices and index numbers is not then recommended since no single figure can summarize all the relevant information in a comparison and since an index may be misleading because the tacit assumptions underlying its use may be wrong.

Design of the Investigation

All appearances to the contrary, most of the above malpractices are not serious in that they can be remedied by recourse to the original data if this is still available. Much more serious are those errors which affect the basic data recorded. Statisticians prefer to be consulted *before* the study is initiated in order to guard against this type of irremedial error.

The first objective of good design is to provide estimates of important effects which are independent of (not confounded with) other effects or influences. This is achieved by orthogonality in experimental design. In general, in a balanced design, one can estimate the effects of a factor averaged over different levels of other factors. The most frequent criticism made today by N.I.H. reviewers of proposals for medical research projects is that the proposed data will not unambigu-

Table 2

Water	Diet		
	A	B	C
None	3	3	3
Moderate water	3	3	3
Excess water	3	3	3

Table 3

Water	Diet		
	A	B	C
None	9		
Moderate water		9	
Excess water			9

Table 4

Diet	A	B	C
Excess water	9	9	9

Table 5

Water	Diet B
None	9
Moderate	9
Excess	9

ously answer the questions posed (Cochran, 1965).

In a hypothetical experiment to determine the effect of different diets, *A*, *B*, *C*, and the amount of water drunk on weight changes in rats, two different experimental strategies might be as follows in tables 2 and 3, where the numbers indicate the number of animals allocated to different treatment modalities of diet and water. A typical reaction is that the research worker would not do the experiment this way. He might use the single-factor design shown in table 4 to find which diet (say *B*) has the greatest effect on weight when there is no limitation on water and then repeat as follows in table 5. The above procedure implies that he is interested in the combination of diet and water which most increases body weight. If this is so, then the "one factor at a time" approach is inefficient, (more animals, more time) and may even be misleading because of interaction (Diet *C* with moderate water may give best results). Many efficient experimental designs are now available for use in medical research.

Recently, Box (1954) and others have developed designs for industrial multifactorial experiments with the objective of estimating that combination of treatment levels which maximises the response. There is every reason to believe that factorial and response surface designs could be usefully applied in bio-medical research.

References

BLAIR, D. W., M. J. WILLIAMS, A. J. CARR, AND S. J. KILPATRICK. Effect of L-thyroxine on gastric secretion in the pylorus-ligated rat. *Gut* 6: 343–348, 1965.

BOX, G. E. P. The exploration and exploitation of response surfaces: some general considerations and examples. *Biometrics* 10: 16–60, 1954.

COCHRAN, W. G. The planning of observational studies of human population. *J. Roy. Stat. Soc.* Ser. A 128: 234–255, 1965.

CROOKS, J., I. P. C. MURRAY, AND E. J. WAYNE. Statistical methods applied to the clinical diagnosis of thyrotoxicosis. *Quart. J. Med.* 28: 211–234, 1959.

DIEM, KONRAD (ed.). *Documenta Geigy. Scientific Tables.* New York: Geigy Pharmaceuticals, 1962, 778 pp.

GRUENWALD, P., AND H. N. MINH. Evaluations of body and organ weights in the perinatal pathology. II. Weights of body and placenta of surviving and autopsied infants. *Am. J. Obstet. Gynecol.* 82: 312–319, 1961.

KILPATRICK, S. J. Mortality comparisons in socio-economic groups. *Appl. Stat.* 12: 65–86, 1963.

MAINLAND, D. *Elementary Medical Statistics.* 2nd ed. Philadelphia: W. B. Saunders, 1964, 381 pp.