Statistics in Clinical Chemistry

Spencer M. Free, Jr.

Many people today think of statistics as a form of mathematical manipulation to be performed on numerical data. A smaller group looks upon the technics of experimental statistics as guides for specific research problems or investigations. Too few realize that statistical thinking can be applied to their routine work.

Many readers have seen statistical procedures applied as an afterthought. Somehow, the investigator amasses some related numbers and feels that an intense mathematical treatment will enhance the value of this work. This could well be true if the investigation was planned with the statistical analysis in mind. It may also be true if the investigator who did not anticipate such a final summary interprets the statistical conclusion as only a suggestive one. It is too easy to "make the data fit the facts" when statistical procedures see their first application after the experiment has been completed.

In the broad field of clinical chemistry, experimental statistics is most valuable in the planning phase of the work. Many times, the investigator can recognize a "medically significant" result from an experiment. However, such judgment is valid only when the experiment has been properly planned and the judgment anticipated. Planning here might be defined as promulgating rules for the unbiased collection of adequate data. Furthermore, it would be desirable to substantiate the medical judgment by a statistical "test of significance," which can be provided for in the planning.

Perhaps an aside is justified here. Statisticians somehow have acquired the reputation for wanting too much data. Experience in

Presented by invitation as part of the symposium of the American Association of Clinical Chemists, Tenth Anniversary Meeting, College of Medicine, State University of Iowa, September 4-6, 1958.

Dr. Free is Head, Quality Control Section, Research and Development Division, Smith, Kline and French Laboratories, Philadelphia, Pa.
our own laboratories, however, has shown that usually just the opposite is true. Planning in advance has reduced the amount of data required for a decision.

The above outlines very briefly some of the better recognized applications of statistics. However, statistical thinking also can be very helpful when applied to routine chemical determinations. For example, in some laboratories many determinations of the cholesterol levels in blood are run every day. Everybody acquainted with this work knows that the results differ considerably. But how many people wonder why they differ, and how many seek ways to reduce the variations and/or make the differences more meaningful? To understand the differences is good; to use this understanding to improve the usefulness of the assay results is much better.

The following series of situations is an attempt to describe the difference in assay results as a clinical chemist might view them. Again, the determination of cholesterol in the blood is used as the assay in this discussion.

Situation 1, shown in Fig. 1, is the typical clinical approach. From 2 patients, single samples of blood are drawn on different days or at different times in the same day. Single cholesterol determinations are run on each sample. The clinical chemist now has 2 numbers representing the cholesterol levels in blood, and they are most likely very different. He attributes this difference to the difference between the patients. This is an everyday occurrence multiplied by many samples from many patients.

Situation 2, shown in Fig. 2, can be observed on some occasions. Two samples of blood are taken at different times and/or on different days from a single patient. Blood samples are submitted to the laboratory for analysis so that the determinations are not run on the same day. Single laboratory samples are run on each blood sample. This time the clinical chemist has 2 results which are still different, but the difference is not quite so large as it was in situation 1. Now the difference in the 2 results is attributed to the difference between the 2 samples taken from the same patient. One should pause here and ask how he would feel about 2 results which were the same.

Situation 3, shown in Fig. 3, almost never occurs. One patient gives one sample of blood. However, this sample is subdivided so that part is submitted to the laboratory and assayed on one day; the second part is submitted and assayed on another day. The results
again are different and this time one attributes the difference to the "day to day variation" known to persist (one should be stronger than to use "exist" in this connection) in most laboratories. A few faint voices would suggest that the differences are due only to assay variation.

Situation 4, shown in Fig. 4, almost always happens. From a single patient, 1 sample of blood is drawn, the sample is submitted for assay, and 2 (duplicate) determinations are run on the same day, most often side by side. Again, the results differ, but only by a small amount; all agree that this is a part of the life of a chemist. The difference is entirely assay variation and represents the inability to reproduce exactly the results for a chemical determination. However, the differences here are very small and are scarcely a serious problem.

Upon review of the above examples, most will agree that the differences between the 2 resulting observations would decrease progressively from situation 1 to situation 4. With a little thought, it will be apparent that each of these situations is most affected by 1 of 4 different major sources of variation, but all 4 sources are present in each. Thus it was the addition of several types of variation which actually contributed to the differences.

Figure 5 diagrams an experiment that would most likely not be done with blood samples from humans. However, such a study could easily be carried out with human urine samples or in animals. The experiment would run as follows. Choose several patients and take 2 samples from each patient. Divide each sample into 2 parts and submit the parts for laboratory analyses on separate days. For each analysis run duplicate laboratory determinations on the same day.

A review of the data collected for such an experiment would demonstrate the sources of variation described above. The differences between the many pairs of determinations from the same samples run on the same day will be a good estimate of laboratory variation, the inability to produce identical results on assays run concurrently.

The second line from the bottom of Fig. 5 will show differences larger than those on the bottom. These differences represent the variations due to the day to day effects plus assay variation. Mathematically, the statistician can separate these components.

The third line from the bottom of the figure would show even larger differences. These differences would be due to the difference between
samples from the same patient plus the day to day and assay variation.

The differences on the top line of the figure would be attributable to the sum of the 4 sources of variation. Thus, one can list the major sources of variability which are a part of any one chemical determination as follows:

- Patient to patient variation ($\sigma_p^2$)
- plus
- Sample to sample variation within a patient ($\sigma_s^2$)
- plus
- Day to day variation for a single sample ($\sigma_d^2$)
- plus
- Assay variation ($\sigma_a^2$)

For brevity, the symbol $\sigma_i^2$ will represent these components as the discussion continues. It should be pointed out that these components are additive effects as derived in the discussion of Fig. 5.

In this notation, it is now possible to describe the variation of a single report to be:

$$\sigma_p^2 + \sigma_s^2 + \sigma_d^2 + \sigma_a^2$$

More important, one can describe the expected variation of any average result to be:

$$\frac{\sigma_p^2}{P} + \frac{\sigma_s^2}{S} + \frac{\sigma_d^2}{D} + \frac{\sigma_a^2}{A}$$

Where
- $P =$ the number of patients included
- $S =$ the total number of samples taken
- $D =$ the total number of days on which the work is done
- $A =$ the total number of assays run and averaged

Again, in this notation, a single report would have $P=S=D=A=1$, and the customary average for duplicate assays on the same day would have $P=S=D=1, A=2$. Here should be some hint as to why duplicate assays on the same day seem to accomplish so little more than a check for the right magnitude.

Many laboratories run duplicate analyses ($A=2$) on most samples.
If one could arrange to run the duplicates on different days (D=A=2) the difference between results would be greater. However, the average would be more meaningful, for it would have a smaller variation. Theoretically, this must always be true. In practice one should determine the magnitude of $\sigma_d^2$ to insure that the gain in precision justifies the additional effort. This should be a very useful procedure for many clinical determinations.

To orient better the readers who think of variation in terms of confidence limits, one can say that the approximate 95 per cent confidence limits for a single report would be:

$$\pm 2 \left( \sigma_s^2 + \sigma_d^2 + \sigma_a^2 \right)$$

The approximate 95 per cent confidence limits for any average would be:

$$\pm 2 \left( \frac{\sigma_s^2}{P} + \frac{\sigma_d^2}{S} + \frac{\sigma_a^2}{D} \right)$$

Obviously, in clinical chemistry, one seeks to minimize the total variation. When considering a single determination, minimization generally is a problem of technic and can operate only on the $\sigma_d^2$ and $\sigma_a^2$ components. However, results are more meaningful if all samples are taken at some fixed time, thus minimizing the $\sigma_a^2$ component.

Minimizing the variance of an average is a different problem. Here one can plan a sampling technic that will determine the size of P, S, D, and A. Conceivably, one could specify the desired magnitude and arrange a work schedule to produce such a result.

Several situations can be outlined to consider sampling plans. The comments below are by no means exhaustive but are included as examples to stimulate the reader’s imagination.

**SINGLE PATIENT APPLICATIONS**

1. Any common assay for a diagnosis would probably have S=D=A=1 and P must be equal to 1. Such a result suggests why our so-called normal values have such a spread and perhaps why 2 supposedly similar patients have quite different reports.

2. A very critical determination for diagnosis would best have S greater than 1 and D at least 2 when possible. When this can be accomplished A=S or D will be enough laboratory work. There is little to be gained from duplicate laboratory analyses when you have more than 1 sample or work on more than 1 day.
3. When considering before and after treatment determinations on the same patient, one is interested in the difference "within" a patient and thus \( \sigma_p^2 \) does not enter into the problem. If the difference is a small one, it would be well to have \( D \) and/or \( S \) greater than 1 both before and after treatment. Not much can be done to reduce the variation of a difference when only after treatment samples are available.

Specifically, consider the variation present in 2 assay results when \( S=D=A=1 \) for the before and after treatment tests. This could be why one observes an apparent increase when the treatment in question was expected to produce a decrease or vice versa.

**MULTI-PATIENT APPLICATIONS**

1. In the comparison of the average between 2 groups of patients, the usual clinical procedure is to collect samples as one encounters the patients. If one has 10 patients per group, quite often \( P=10, S=10, D=\text{about 10, and } A=10 \) per group. This is good, for it effects quite a reduction in the variation of the average. However, on occasions one expects a very small difference and the number of patients is limited.

   Here there are 2 alternatives to increase precision: (a) increase \( S \) and/or \( D \), which in turn effects a sufficient increase in \( A \); (b) do all the chemical work on 1 day, and since the result is to be a comparison between 2 groups, one can assume that \( \sigma_d^2=0 \) for the comparison.

2. For comparisons of the same group of patients from time to time when all patients are represented at each time interval, one can assume \( \sigma_r^2=0 \). However, when patients fade in and out of the group as some tend to do, average results are difficult to compare, for \( P \) is continually changing. This should suggest why group averages plotted over time often appear to have weird peaks and valleys in what one might expect to be a smooth trend.

   All of the above comments have been on a somewhat theoretical plane. However, they have very practical application in the chemical laboratory. The crux of good work is proper planning. There are very few advantages to having an excellent laboratory analytic method if the material you are assaying has a limited potential.

   The chemist's task is to reduce the effects of \( \sigma_d^2 \) and \( \sigma_r^2 \). There is good reason to believe that these components represent the total laboratory variation. The sum \((\sigma_d^2 \text{ and } \sigma_r^2)\) describes the laboratory variation of a single assay run on 1 day.
In practice, the overall improvement appears to require understanding and teamwork between the clinician and the chemist. Certainly both stand to gain when the precision of their data is improved. Better precision does not always demand more work. Much more can often be accomplished by proper planning.