Statistics for Nuclear Medicine
Part 6: Normal Values, Evaluating a New Diagnostic Procedure, Sequential Methods, and Conclusion

Peter C. O'Brien, Marc A. Shampo, and James S. Robertson

Section of Medical Research Statistics, Mayo Clinic and Mayo Foundation, Rochester, Minnesota

J Nucl Med 24: 535–541, 1983

NORMAL VALUES

In interpreting the significance of a laboratory measurement, it is often helpful to know how the value obtained in a given case relates to a set of values from a healthy reference population. What percentage of healthy persons have higher (or lower) values?

For this purpose, we first must find the distribution of the variable in the healthy population.

Estimating normal values. *Distribution in sample.* The basic approach is the same as in several previous undertakings: since it is not possible to make observations of every member of the population, we rely instead on estimates derived from a sample. For illustration, consider the population of 5,594 serum urea values in Part 2. Since those values were unknown to us, we drew a sample of 100 values randomly from the population with which to estimate certain characteristics of the population (such as its mean). The same values are presented again in Table 1 with percentile values added.

As usual, the high percentiles are matched to the large values, and the percentile matched to a value indicates what percentage of all the values rank lower. Thus the largest of the 100 values (173 mg/dl) is the 99th percentile (P_{99}); the next value (103 mg/dl) is the 98th percentile; and so on. If there were 200 values in the sample, rather than 100, the largest observation would estimate the 99.5 percentile; and if 82 mg/dl were then the 10th largest value, that would still be our estimate of the 95th percentile.

As Table 1 stands, however, with 95% of the observations less than 82 mg/dl, 95% also are less than 69, or any number between 82 and 69. As a result, any of these numbers could be used to estimate the 95th percentile. Rather than choosing the largest, it is conventional to choose a value in between (such as 75). Various strategies for making an appropriate choice have been developed and are commonly used. In general, larger samples produce smaller gaps; and sample sizes should be made large enough so that the ambiguity resulting from this problem is negligible.

Provided with a distribution of percentile values in a sizable sample, a physician can determine approximately how his patient's serum urea value relates to those in the reference population. From Table 1, for example, he would know that a value as large as or larger than 50 mg/dl is uncommon, estimated to occur in only 10% of that population.

Sample size. As in any situation where we must rely on sample estimates, we are concerned with their variability. Here we consider percentile estimates from 10 samples of 100 each, drawn from the population of 5,594 (the same samples drawn in Part 2, now represented by selected percentile values in Table 2). Clearly, the values for P_{50} are less variable than the values for the very high percentiles (P_{90}, P_{95}, and P_{99}). Although a sample consisting of 100 values ordinarily is adequate for estimating the center of a population, it is a very small basis for estimating the outer percentiles (such as P_5 or P_{95}).

Refinements. In our example we have deliberately oversimplified the problem of estimating normal percentiles. Normal values of many variables are affected by the age and sex of the subjects. Statistical methods are available for estimating age- and sex-specific percentiles, but they obviously require data from more subjects overall.

Comment. It is a common misconception that, in general, 95% of population values lie within two standard deviations of the population mean. (The proposition is

For reprints contact: Dr. O'Brien, Mayo Clinic and Mayo Foundation, 200 First Street SW, Rochester, MN 55905.

TABLE 1. DISTRIBUTION OF SERUM UREA VALUES IN A SAMPLE (n = 100) DRAWN RANDOMLY FROM A POPULATION (N = 5,594)*†

Value, mg/dl	Frequency and percentile (P)	Value, mg/dl	Frequency and percentile (P)
173	1	36	2
103	1	35	6
95	1	34	2
88	1	33	3
82	1 (P_{95})	32	9 (P_{50})
68	1	31	4
66	1	30	6
52	2	29	6
50	1 (P_{90})	28	2
46	1	27	2 (P_{25})
45	2	26	2
44	1	25	4
42	5	24	6
41	3	23	3 (P_{10})
40	5 (P_{75})	22	2
39	2	20	5 (P_{5})
38	1	19	1
37	3	18	1
		16	1

* Mean of sample is 36.56; standard deviation is 20.27.

† O'Brien PC, Shampo MA: Statistics for clinicians: 4. Estimations from samples. *Mayo Clin Proc* 56:274–276, 1981

TABLE 2. MEAN AND SELECTED PERCENTILES OF SERUM UREA VALUES IN 10 SAMPLES (EACH n = 100)*

Sample	Mean, mg/dl	Values for selected percentiles			
		P_{50}	P_{90}	P_{95}	P_{99}
1†	36.56	32	50	82	173
2	33.92	31	50	57	103
3	34.24	31	50	62	123
4	33.00	31	43	52	86
5	33.47	31	46	60	220
6	36.67	32	48	56	172
7	35.15	30	52	61	123
8	38.93	32	50	69	388
9	32.31	30	48	56	93
10	36.57	32	46	55	174
s‡	2.07	0.8	2.7	8.8	89.3
Population values§	35.33	31	48	60	124

* O'Brien PC, Shampo MA: Statistics for clinicians: 4. Estimations from samples. *Mayo Clin Proc* 56:274–276, 1981.

† From Table 1.

‡ The standard deviation (s) of the 10 values listed directly above.

§ From population of 5,594 values.

true only under special, infrequently occurring conditions, as when the population values have a Gaussian distribution.) This misconception has given rise to the regrettable practice of estimating the 2.5 and 97.5 percentiles simply as the mean ± two standard deviations ($\bar{x} \pm 2$ s). Applied to the first sample of 100 serum urea values in our example (presented in Table 1), $\bar{x} \pm 2$ s yields 36.56 ± 2 • 20.27, giving the impossible result $P_{2.5} = -3.98$ mg/dl. Clearly, the method is unsuitable for general use.

For two nontechnical papers providing an excellent, more detailed discussion regarding the choice of a suitable reference population and the estimation of population percentiles, see Elveback (*1*,*2*).

EVALUATING A NEW DIAGNOSTIC PROCEDURE

When a new medical procedure has been developed, such as emission computed tomography (ECT), it is necessary to evaluate the contribution to patient care that will result from its use. In this situation, the subjective opinion of the physician responsible for patient care will be essential, and perhaps it will determine the ultimate decision as to the procedure's usefulness. It is also desirable, however, to perform studies that will provide objective, quantitative data. Three aspects that should be considered are (1) the reliability of the procedure, (2) its accuracy, and (3) how its results compare with those of conventional methods.

We shall use evaluation of ECT to illustrate how each of these concerns may be addressed. The statistical methods used will differ slightly, according to whether the measurement of interest is dichotomous (such as presence or absence of a tumor) or continuous (such as tumor size). We shall consider the dichotomous type first.

Analyzing dichotomous data. *Reliability.* The reliability of a method (also called its *precision*) is its ability to provide the same answer in repeated observations. (Whether it provides the correct answer is not at issue here but will be considered in the section on accuracy.) Reliability has two aspects: inter-interpreter and intra-interpreter.

For evaluation of inter-interpreter reliability (consistency of observations by different interpreters—in our example, nuclear medicine physicians), a set of ECT images showing a broad range of the abnormalities of interest, and including some showing normality, are presented in random sequence for interpretation by each physician participating in the study. Of course the actual

TABLE 3. ACCURACY OF TEST RESULTS—SCHEMATIC REPRESENTATION OF THE DATA

	Disease status		
Test result	Positive (has disease)	Negative (does not have disease)	Total
Positive	a	b	a + b
Negative	c	d	c + d
Total	a + c	b + d	a + b + c + d

status of each subject must be unknown to the physician at the time he views the image (but this information must be available for subsequent assessment of the accuracy of image interpretation). This type of evaluation requires a large number of images, at least 100 and sometimes more.

It is also desirable to evaluate intra-interpreter reliability (the consistency with which the same interpreter arrives at the same diagnosis when viewing the same image). This may be accomplished by repetition of the study outlined above; however, any possible learning effect should be minimized. A method often used to accomplish this, at least in part, is to use a large number of images, randomly rearrange the order for each repetition, and separate the repetitions by suitably long time intervals. As before, it is essential that the observer make his judgment without knowing the status of the patient.

Accuracy. The accuracy of a procedure is measured by its ability to give the correct answer. Often this is expressed by the error rates: the proportion of false-positive results and the proportion of false-negative results. A test result is said to be a false positive if it is positive but the actual status of the patient is negative. Similarly, a false negative means that the test result is negative but the patient's actual status is positive. The possibilities for classification of the study results are indicated schematically in Table 3. With the notation in Table 3, the proportions of false positives and false negatives are $b/(b + d)$ and $c/(a + c)$, respectively.

A parallel set of terms is also often used in describing the accuracy of a test. Rather than focusing on the error rates, it focuses on the proportion of cases that are classified correctly. Thus, the *sensitivity* of a test is defined as the proportion of patients with the disease who are correctly classified by the test. Similarly, the *specificity* of a test is defined as the proportion of patients without the disease who are correctly classified by the test as being disease-free. In terms of the notation in Table 3,

$$\text{sensitivity} = a/(a + c)$$

$$\text{specificity} = d/(b + d).$$

TABLE 4A. HYPOTHETICAL DATA FOR EVALUATING ACCURACY OF A NEW DIAGNOSTIC PROCEDURE

	Disease status		
Test result	Positive (has disease)	Negative (does not have disease)	Total
Positive (a)	35	5	40
Probably positive (b)	30	10	40
Uncertain (c)	20	20	40
Probably negative (d)	10	30	40
Negative (e)	5	35	40
Total	100	100	200

TABLE 4B. FALSE-POSITIVE AND FALSE-NEGATIVE RATES FOR HYPOTHETICAL DATA IN TABLE 4A

Declared positive	False-positive rate	False-negative rate
None	0.00	1.00
a	0.05	0.65
a + b	0.15	0.35
a + b + c	0.35	0.15
a + b + c + d	0.65	0.05
a + b + c + d + e	1.00	0.00

Another pair of useful numbers that can be derived from the array shown in Table 3 is the positive and negative predictive values, which give the probability that the patient has the disease when the test is positive or does not have the disease when the test is negative. These are $a/(a + b)$ and $d/(c + d)$.

Dichotomizing complex data. Some test results that one wishes to treat as dichotomous are not that simple. Even when the actual status of the patient must be either positive or negative, the best obtainable test readings may be "definitely positive, probably positive, uncertain, probably negative, definitely negative."

If one forces those data into the dichotomous mold, the numbers of false positives and false negatives will vary with the placement of the dividing line. In such a situation it is desirable to determine the numbers of false positives and false negatives for each possible placement of the division, as illustrated in Table 4A and B. The rates in Table 4B may also be displayed graphically, in what is called a *receiver operating characteristic* (ROC) curve (Fig. 1). Such a graph enables the reader to determine at a glance the continuum of possible false-negative and false-positive rates. It also provides a convenient method for comparing two procedures. For example, if the ROC curve for method A lies entirely below the curve for method B, A would be judged superior to

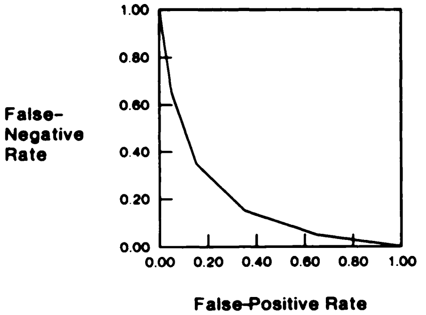

FIG. 1. Receiver operating characteristic (ROC) curve for hypothetical data in Table 4.

B. In applications, one may observe the curves to cross, in which case the ranges where a superiority exists would be of interest.

Comparative studies. The usual objective of a comparative study is to compare the accuracy of an experimental method (ECT imaging in our example) with the accuracy of one or more conventional methods (such as conventional roentgenography). An important first step is to define the patient group to be studied. It is essential in this type of study that eligibility for the study should not depend in any way on the outcome of either the experimental or the conventional method. For this reason, the patient's entry into the study should be determined before he is examined by either method. Once a patient is admitted to the study, examination by each method should be done without knowledge of the results of the competing method. Additional knowledge (certain clinical information, for example) should not be available with either method (unless such information is considered an integral part of that method).

For ascertaining the relative accuracy of the two methods, the true status of the patients must be known. For example, if method A indicates the presence of a tumor when method B does not, resolution of this difference may be obtained from subsequent surgery. In this situation, the willingness to do surgery should be the same when A is positive and B is negative as when A is negative and B is positive. If it is known beforehand that the rate of false positives for each method is near zero, this difficulty does not arise.

In the absence of a definitive diagnosis, the best that can be done is to measure *agreement* between methods A and B without attempting to measure relative accuracy.

When actually comparing two methods, one often finds that whereas one is more sensitive, the other is more specific, making it difficult to ascertain on a quantitative basis that one procedure is superior to the other. (However, a qualitative assessment of the error rates may still be possible.) Where possible, it is best to fix one of the error rates at a desirable level and make comparisons on the basis of the other error rate. As mentioned previously, when test results are not dichotomous, graphing the ROC curves for both methods on the same graph provides an effective basis for making the comparisons.

Analyzing continuous data. The concepts of reliability, accuracy, and comparative studies described above still apply when the measurement of interest is continuous, as is tumor size. However, some of the statistical methods are different.

For example, the reliability (internal consistency) of observations may be expressed by the standard deviation among repeated measurements. It sometimes happens that the error tends to be larger when the quantity under study is large, e.g., errors may tend to be larger in measuring large tumors than very small ones. To counter this, it may be appropriate to express reliability by the coefficient of variation, which is the standard deviation divided by the mean ($s \div \bar{x}$).

In measuring accuracy, we are concerned with how closely a set of measurements cluster about the true value. Sometimes closeness is best measured by the arithmetic difference between the true and observed values. When the difference seems to be proportional to the magnitude of the true value, it may be more appropriate to express the difference as a percentage of the true value. Often a graph, such as that shown in Fig. 2, is helpful in evaluating accuracy. A graph of the cumulative distribution of the error (expressed either as a difference or as percentage error) also may be useful, or perhaps quoting appropriate percentiles from the cumulative distribution will suffice. Sometimes the absolute magnitudes of the error (in which negative signs are disregarded) are most informative.

When results from two methods of measurement are to be compared and the definitive measurement is available (such as tumor size determined at surgery), one can tabulate the errors for each method and compare the two distributions of error. A statistical test of significance also may be performed (based on the values of the observed errors, perhaps using a paired Student *t*-test).

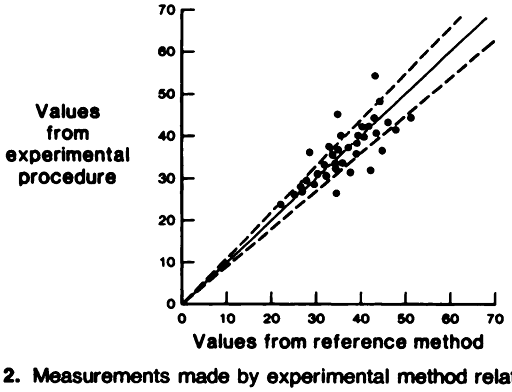

FIG. 2. Measurements made by experimental method related to those from reference method (each point represents the two values in a single case), with lines of identity (—) and ±10% error (- - -). (From O'Brien PC, Shampo MA: Statistics for clinicians: 9. Evaluating a new diagnostic procedure. *Mayo Clin Proc* 56:573–575, 1981.)

However, statistical significance ordinarily is of minor concern in evaluating a new procedure.

Comment. Once the descriptive techniques described above (and perhaps others) have been used, the ultimate question, *Are the reliability, the accuracy, and the improvement over existing methods good enough?* must be answered by the physician.

SEQUENTIAL METHODS

The term *sequential*, in statistics, refers to the approach to study design and data analysis in which the data are reviewed at various points during the course of the study. For an example, we may address the question whether headache is more frequent with drug F or with drug G, which is the same question that was addressed in Part 3, but our method will be different. For the example developed previously, the study was begun by randomly assigning half of a series of patients to receive drug F and the other half drug G; and when the observations had been made, the investigator tested the hypothesis that the frequency of headache with each drug was the same, using an appropriate statistical method.

This time, however, we wish to monitor the data as they are being collected, with a view toward terminating the trial early if either drug appears definitely superior to the other. As in the previous parts of this series, if the data lead us to conclude that there is a difference—whether at an interim review (with consequent termination of the trial) or at completion—we will want to know the corresponding p value. That is, if no real difference existed and our trial were repeated many times, what proportion of those trials would provide such strong evidence of a difference? For reasons to be discussed later (in the Comment), however, the testing methods described previously are not valid for use with a sequential evaluation of the data: modifications are required.

All sequential methods that have been developed use objective predetermined criteria for termination of the trial. To illustrate, let us suppose that the investigator in our hypothetical example decided he would be willing to study a maximum of 120 patients in a clinical trial, 60 to receive drug F and 60 drug G by random assignment. He plans to evaluate the data when each increment of 40 observations becomes available. At each evaluation, he will use the methods described previously to compute a chi-square (χ^2) statistic. If any of these statistics is sufficiently large, the trial will be terminated with the conclusion that one drug is superior to the other.

How large is "sufficiently large"? To ensure that the conclusion of a difference will not be reached erroneously in more than 5% of such studies, specially prepared tables (*not* the tables of the χ^2 distribution referred to in Part 3) must be used. When statistical significance is indicated, the tables also provide the corresponding p value. These tables indicate that the first 40 cases should yield a χ^2 value exceeding 11.8, or the first 80 cases 5.9, or 120 cases 3.94.

Suppose that in the first group of 40 patients, headache is reported by five of the 20 who received drug F and by 12 of the 20 who received drug G. These data yield a χ^2 value of 5.0. Since this is less than 11.8, the observed difference between F and G is not sufficient to warrant stopping the study at this point.

Therefore a second group of 40 patients is enrolled and observed, and the combined results of the two groups are headache in 10 of 40 who received F and in 22 of 40 who received G. These numbers result in a χ^2 value of 7.5; and since 7.5 is greater than 5.9, the evidence at hand is sufficient to warrant termination of the study with small risk that further data would negate the apparent superiority of drug F.

Comment. 1. In this example, it might have been tempting to compare each observed χ^2 value to percentiles of the tabled χ^2 distribution, as in Part 3. With this strategy, one would have obtained a p value of 0.025 at the first test and, since this is less than 0.05, would have concluded that the difference was statistically significant. How often will an experimenter using this strategy reject the null hypothesis incorrectly?

By definition, the probability of obtaining a statistically significant result ($p < 0.05$) at the first review is 0.05. However, the probability of obtaining this result on review of groups 1 and 2 combined (but not group 1 alone) is 0.033; and the probability of obtaining it on review of groups 1, 2, and 3 combined (but not group 1 or groups 1 and 2 combined) is 0.024. Since the null hypothesis will be rejected under any of these three circumstances, the probability of rejection is 0.050 + 0.033 + 0.024, which equals 0.107.

The reader should remember that, if one makes sequential evaluations of data, special methodology should be supplied by a statistician.

2. Procedures have been developed for performing a test of significance as each observation is added to the accumulated evidence, but they are generally impractical and rarely used. Such plans are often referred to as *fully sequential.* On the other hand, the type of sequential design that we have described (where a test is performed as successive groups of observations are added to the accumulation) is referred to as *group sequential.*

CONCLUSION

Initiating a research study. One of the underlying purposes of this series of papers has been to provide the reader with a feel for the situations in which he should consult with a statistician, and an ability to communicate effectively in such consultation.

In undertaking a research effort, the first step is to formulate clearly the question your study is intended to answer.

The next step is to talk with a statistician. The statistician will assist you in the design of the study (to ensure that the answers obtained are valid and that the sample size is sufficient to detect the effects that you are interested in), setting up data collection procedures, planning and implementing data analysis, interpreting and presenting the results, and preparing the manuscript.

How does one go about locating a statistician? Although there are a number of private consulting firms available, probably the simplest and most cost-effective strategy is to seek help from a large university. Most such institutions have a statistics department. A brief phone call to the departmental chairman usually will suffice for referral to a statistician suitable for your needs. (Statisticians tend to specialize in different areas, much as physicians do.)

If you feel uncomfortable as you approach your first meeting with the statistician, recognize this as a normal and quite common reaction. Remember that it is his responsibility to cut through the technical jargon and address your specific needs. You, on the other hand, should be prepared to explain those needs, remembering that he lacks your medical background. Make sure you can state very specifically what questions your study is intended to answer, as this will be the starting point for the collaborative effort. From this point, the process should move quite naturally to considerations of study design, data collection, data analysis, and interpretation of results. Expect to meet with the statistician on a continuing basis as the study moves through these stages. If at any time his recommendations seem to violate "common sense," insist on a satisfactory, comprehensible explanation. The problem may well lie in his mathematical formulation, rather than in your lack of mathematical sophistication.

Summary and final advice. In the preceding series of articles, we have described some of the most elementary concepts and methods in statistics. We started with *descriptive statistics*, discussing methods for describing a data set by use of such descriptors as the mean and standard deviation, median, and range (and interquartile range). Graphic techniques for providing a quick visual impression of the data, such as histograms and scatter diagrams, were presented also.

We then turned our attention to *inferential statistics*, establishing generalizations about a population by use of a sample drawn from it. This process was illustrated by describing some of the more common techniques, such as confidence intervals, *t*-tests, and chi-square (χ^2) tests. In each situation, the basic approach is the same: First, the questions being addressed must be identified and stated precisely. These questions, together with the resources available to the investigator, determine the appropriate study design, which in turn dictates the method used for data analysis. Proper interpretation of the analysis completes the process. It is essential that an investigator who intends to rely on statistical inference work closely with a statistician during the entire process—from questions to study design to data analysis to interpretation.

Two complementary aspects of data analysis were presented: estimation and hypothesis-testing. *Estimation* is attempting (by use of sample data) to ascertain some characteristic of the population, such as the mean serum urea level, or the difference between sets of paired data (such as free thyroxine measurements made before and after heparin infusion, case by case), or the difference between the incidence of side effects associated with two drugs. Because the estimates are based on sample data, which are subject to random variation, we have shown how to assess their precision by deriving standard errors and confidence limits. Since precision improves with increase of sample size, a confidence interval may be viewed as a measure of the adequancy of sample size.

For *hypothesis-testing*, one first transforms the question of interest into a null hypothesis. For example, to determine whether a new treatment modality is more effective than the established modality, one formulates a hypothesis that there is no difference between their effects. To assess the null hypothesis, one collects data and computes a p value. Rejection of the null hypothesis is based on a statement such as, "If the null hypothesis (no difference) is true of the population, then the probability that a sample of this size will show a difference as large as the one that appears in our sample is less than p."

When the data justify rejection of the null hypothesis (that is, when the p value is very small), the results are termed *statistically significant* (not to be confused with *clinically significant*, a judgment to be made by a clinician). When the results of hypothesis-testing do not lead to rejection of the null hypothesis, the interpretation may be less clear. Accepting the null hypothesis may not be justified if the lack of statistical significance may be attributed to small sample size. Again, this question may be addressed by consideration of confidence intervals, if available.

An important principle is that statistics can only establish an association and cannot define the cause and effect. For example, statistics may establish an association between having a yellow-stained index finger and the occurrence of lung cancer. However, it is obvious that although the association is strong, "yellow finger" does not cause cancer. In this case the observed association between yellow finger and lung cancer is merely an artifact resulting from the association between smoking and lung cancer.

Some additional special topics that occur commonly in medical research were discussed: evaluating a new diagnostic procedure, determining normal values, describing survivorship, and using sequential methods. Although we alluded only briefly to some important study-design considerations, it is worthwhile to keep in mind the need for a comparison group, the desirability of random double-blind treatment assignment, and the important distinction between observational and experimental studies.

In all the topics introduced, we only scratched the surface; and of necessity, some topics were omitted entirely. We hope, however, we have provided the reader with an introduction that will encourage a further study of statistics and prepare him for wiser judgment of what he reads.

REFERENCES

1. ELVEBACK LR: How high is high? A proposed alternative to the normal range. *Mayo Clin Proc* 47:93-97, 1972
2. ELVEBACK LR: The population of healthy persons as a source of reference information. *Hum Pathol* 4:9-16, 1973