Statistics for Nuclear Medicine
Part 2: Estimation from Samples and *t*-Tests

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

Mayo Clinic and Mayo Foundation, Rochester, Minnesota

J Nucl Med 24: 165–171, 1983

ESTIMATION FROM SAMPLES

In the preceding paper, we discussed statistical techniques for describing a set of data—descriptive statistics. Here we begin to consider inferential statistics: how to deal with problems wherein it is not practical to obtain and manipulate observations on every member of the population of interest. Our approach is to study a sample from the population. (Indeed, it is a convention of inferential statistics that "population" means a group—not necessarily of persons—that is studied by sampling.) To the extent that the sample group is representative of the population from which it is taken, inferences properly drawn from the sample will apply to the population.

Description of population characteristics. Statisticians often refer to a population characteristic as a variable; for example, height, weight, and bone density would all be considered variables. The distribution of the values of a variable in the population can be represented by a sample histogram constructed with measurements from a sample group. Similarly, the sample mean and standard deviation ($\bar{x}$ and s) may be used to estimate the population mean and standard deviation (μ and σ). Statisticians refer to statistics such as $\bar{x}$ and s as random variables, since they vary randomly in repeated samples from the same population. The corresponding mean and standard deviation of the population—which are constants—are referred to as parameters. In distinguishing population parameters from their sample estimates, Greek symbols are generally used for the former and Roman symbols for the latter.

Received Sept. 30, 1982; revision accepted Sept. 30, 1982.

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

Variability of random samples. The ability of sample statistics to describe population characteristics depends very much on the representativeness of the sample. To get some idea of the variation in random sampling (called random error), consider the data from a Mayo Clinic study of serum urea concentrations in 5,594 subjects.

Suppose that, having been provided with the values, we want to know their mean and standard deviation but do not want to add up 5,594 numbers and do all the necessary further calculations on that large a scale. The 5,594 observations can be considered a population in the statistical sense and a random sample can be selected from it. Such a sample amounting to 100 observations is presented in Table 1, and a mean ($\bar{x}$) of 36.56 and standard deviation(s) of 20.27 have been calculated from it. In fact, when the complete but tedious calculations were performed by a computer, the population mean (μ) was 35.33 and the population standard deviation (σ) was 21.55.

Since the samples drawn from a population vary, so do the estimates derived from them. To illustrate, we have drawn nine additional samples, each of size 100, from the population described above. As Table 2 shows, the means associated with the resulting set of 10 samples varied from 32.31 to 38.93.

Accuracy of sample mean as estimate of population mean. In judging how accurately a sample mean estimates the population mean, one begins with the realization that large samples are more reliable representatives than small ones. The procedures to be described here are suitable for samples containing as few as 60 observations, provided that the population does not have outliers or severe skewness.

With samples of sufficient size, regardless of the underlying distribution in the population, 95% of all

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

Value	Frequency	Value	Frequency
16	1	36	2
18	1	37	3
19	1	38	1
20	5	39	2
22	2	40	5
23	3	41	3
24	6	42	5
25	4	44	1
26	2	45	2
27	2	46	1
28	2	50	1
29	6	52	2
30	6	66	1
31	4	68	1
32	9	82	1
33	3	88	1
34	2	95	1
35	6	103	1
		173	1
			N = 100

* Mean of this sample ($\bar{x}$) is 36.56 and standard deviation (s) is 20.27; for calculations of $\bar{x}$ and s, see paper 1.

† From O'Brien PC, Shampo MA: Statistics for Clinicians: 4. Estimation From Samples. *Mayo Clin Proc* 56:274–276, 1981.

sample means are within two standard errors of the population mean. The standard error of the mean ($SE_{\bar{x}}$) equals the standard deviation of the sample divided by

TABLE 2. MEANS OF 10 100-OBSERVATION SAMPLES FROM POPULATION OF 5,594 SERUM UREA OBSERVATIONS*

Sample No.	Sample mean (mg/dl)
1	36.56
2	33.92
3	34.24
4	33.00
5	35.47
6	36.67
7	35.15
8	38.93
9	32.31
10	36.57

* From O'Brien PC, Shampo MA: Statistics for Clinicians: 4. Estimation from Samples. *Mayo Clin Proc* 56:274–276, 1981.

the square root of the number of observations in the sample:

$$SE_{\bar{x}} = \frac{s}{\sqrt{N}}$$

So in sample 1 (Table 1), in which s = 20.27 and N = 100,

$$SE_{\bar{x}} = \frac{20.27}{\sqrt{100}} = \frac{20.27}{10} = 2.03$$

And since the sample mean lies within two standard errors of the population mean in 95 of 100 instances, one can calculate the 95% confidence interval (CI) having the limits:

$$95\%\ CI = \bar{x} \pm 2 \cdot SE$$

Considered strictly, the "2" in the equation above is an approximation of a quantity that varies with sample size. But with N = 60 it is 2.00, and with extremely large samples it is 1.96; so when sample size is large, 2 usually is satisfactory. Continuing the application to sample 1, whose mean is 36.56:

$$95\%\ CI = 36.56 \pm 2 \cdot 2.03$$
$$= 36.56 \pm 4.06$$

Thus we can be confident, but not absolutely sure, that the population mean lies somewhere between confidence limits 32.50 and 40.62.

The 95% confidence interval provides a valuable indication of how much has been learned about the population mean from the sample. To obtain a narrower confidence interval, a larger sample is necessary. In the example above, if a confidence interval with a width of just 4 units instead of 8.12 (40.62 - 32.50) is desired, the sample will have to be increased to approximately 400 observations.

Influence of small sample size. Thus far we have been using methods suitable for a moderately large sample. When the sample contains fewer than 60 observations, the number 2, by which we multiply the standard error, must be replaced by a larger number (obtained from special tables). This number, which increases as sample size decreases, is designated by the symbol t^*_{N-1}.

Thus, when sample size is less than 60, decreasing the sample size increases the width of the confidence interval in two ways: (a) the standard error of the mean is increased, as illustrated in the previous section, and (b) t^*_{N-1} itself, the multiplier of the standard error, is increased. To illustrate, suppose that the standard deviation of 20.27, derived from the 100-observation sample, had been obtained from a sample of only 10 observations. Then:

$$SE_{\bar{x}} = \frac{20.27}{\sqrt{10}} = \frac{20.27}{3.16} = 6.41$$

But also, the 95% confidence interval must be calculated thus:

$$95\%\ \text{CI} = \bar{x} + t^*_{N-1} \cdot \text{SE}$$

For the present sample (N = 10), $t^*_{N-1} = 2.26$. With this, and with the same mean obtained from sample 1 (36.56),

$$95\%\ \text{CI} = 36.56 \pm 2.26 \cdot 6.41$$
$$= 36.56 \pm 14.49$$

providing 95% confidence limits of 22.07 and 51.05.

So—despite retention of the same sample mean and standard deviation—the change from a basis of 100 observations to only 10 has changed the standard error from 2.03 to 6.41 and the width of the 95% confidence interval from 8.12 (40.62 - 32.50) to 28.98 (51.05 - 22.07).

Comment. 1. Note that the standard deviation is not very helpful in describing the variability of the sample in Table 1. Specifically, the mean minus the usual two standard deviations becomes negative, which no actual serum urea value could be. As mentioned in the first paper of this series, the standard deviation has its greatest usefulness in relating sample means to population means—which is done by converting it to the standard error.

2. It will become more readily apparent in subsequent papers that much of the information required by statisticians in order to make probability statements is available only in special tables. Because the goal of this series is merely to acquaint the reader with basic concepts, the mechanics of working with the tables will not be discussed. It is hoped that the reader will not attempt to analyze his or her data, or even design the experiment, without the assistance of a statistician.

3. In this section we have dealt with the mean of a simple measurement, the serum urea concentration, and of course it might as well have been body weight, days of hospitalization, or any other measurement. But further, the same concept of estimating a population mean from a sample—and for determining the confidence limits of the estimate—can be applied to differences (such as case-by-case differences in bone density before and after treatment) and to proportions (such as proportion of patients benefiting from a drug). The concepts presented here have very wide use in medical statistics.

ONE SAMPLE OF PAIRED OBSERVATIONS (PAIRED *t*-TEST)

The previous section showed how a sample drawn from a large population can be used to provide an estimate of a population statistic (such as an estimate of the mean value of a variable) and also how the accuracy of such estimates can be assessed. In this section, those methods will be applied in a procedure called the "paired *t*-test" to solve a medical problem.

Formulation of the problem. We need to evaluate the effectiveness of heparin in increasing the concentration of free thyroxine. The population of interest consists of all patients who will receive heparin if it is used clinically in the future. The problem may be stated in three questions: (1) Will the drug increase the level of free thyroxine? (2) If so, by how much? (3) Did we have enough data?

With the use of $\mu_{A\text{-}B}$ to represent the mean difference between measurements after (A) and before (B) treatment, if the drug is administered to the entire population as defined, the questions may be stated statistically: (1) Is $\mu_{A\text{-}B} = 0$? (2) If not, how large is $\mu_{A\text{-}B}$? (3) How accurately have we estimated $\mu_{A\text{-}B}$?

Of course, it is not possible to determine $\mu_{A\text{-}B}$ directly by measuring the after-before difference in the total population of future patients. However, the methods described in the preceding section can provide inferences about this parameter.

Collection of data. First, it is necessary to obtain a random sample from the population. Suppose only a very small pilot study consisting of 10 patients (N = 10) is to be done. If it can be assumed that patients present themselves in random order, the sample can be obtained simply by taking the next 10 patients who need treatment. Because it is rarely possible to conduct a truly random collection in medical practice (as is often done in population surveys, for example), the question of the representativeness of the sample is an important aspect of any inferential study but it will not be pursued in the present section.

Suppose the sample is obtained appropriately, the level of free thyroxine is measured, heparin is administered, and the level of free thyroxine is measured again. In Table 3, note that two measurements are made on each patient. It is because these two measurements are made on the same patient, and thus are correlated rather than independent, that the data are regarded as a single sample of paired observations.

In the sample, there is a mean increase ($\bar{\Delta}$) of 0.440 ng/dl, and this serves as an estimate of drug effect in the population. It implies that heparin may increase the level of free thyroxine.

However, that result is based only on sample data, subject to random error (which means that other samples from the same population probably would give different results). So one wonders: If there is no real difference between A and B (the null hypothesis), how often would a difference as large as 0.440 ng/dl occur in repeated samples from the population?

Question 1: Is $\mu_{A\text{-}B} = 0$? The procedure is to make a probability statement of the sort, "*If* a given assumption or hypothesis regarding the population (such as $\mu_{A\text{-}B} = 0$) is true, *then* the probability of obtaining this sample

TABLE 3. MEASUREMENT OF FREE THYROXINE (ng/dl) BEFORE AND AFTER 12-HOUR ADMINISTRATION OF HEPARIN

Patient	After (A)	Before (B)	Δ (A-B)*
1	1.9	2.2	−0.3
2	1.7	0.7	+1.0
3	1.3	1.0	+0.3
4	1.4	2.7	−1.3
5	0.5	0.7	−0.2
6	2.3	1.2	+1.1
7	2.7	0.9	+1.8
8	2.5	1.3	+1.2
9	2.9	1.3	+1.6
10	1.0	1.8	+0.8

$\overline{\Delta}$ = +0.440 ng/dl
s = 1.057 ng/dl
$SE_{\overline{\Delta}}$ = 0.334 ng/dl

* A positive Δ represents an increase in level of free thyroxine.

result is no more than___(a value to be calculated)." And if the probability turns out to be sufficiently small, that will provide a basis for rejecting the null hypothesis. In other words, when the sample result (an observed fact) is nearly impossible in conjunction with the hypothesis, one may reject that hypothesis in favor of an alternative hypothesis that seems more consonant with the data (for example, μ_{A-B} is greater than 0). One must recognize that all probability statements are "If ... then ..." statements, expressing the probability that, under carefully stated circumstances, something will happen or be true.

In the present example (Table 3), the first step toward determining the quantity needed for completing the probability statement is to calculate the size of the mean difference relative to the standard error of the difference. If the pair-by-pair differences include no outliers or evidence of severe skewness, the following formula may be used:

$$t = \overline{\Delta}/SE_{\overline{\Delta}}$$

(Notice that the variation associated with $\overline{\Delta}$, which is $SE_{\overline{\Delta}}$, is based on the variation among the pair-by-pair differences.) Substituting from Table 3,

$$t = 0.440/0.334 = 1.317$$

And, using special tables or computing facilities, we find that, if $\mu_{A-B} = 0$, then the probability of obtaining a value for t greater than 1.317 is 0.110. This probability is often referred to as a P value; so here, $P = 0.110$. It means that the observed difference would occur by random variation (without an underlying real difference) in 11.0% of samples.

The interpretation of this probability must be clear and not careless. What can we say?

1. We cannot reject the hypothesis $\mu_{A-B} = 0$. Since the observed results would occur fairly often even if heparin had no real effect, that may be the case—no real effect.

2. Conversely, we cannot rule out the possibility that a real effect exists, since a real effect may have gone undetected because of the small sample size. We can say only that the evidence in favor of a real increase is not statistically significant.

Question 2: How large is μ_{A-B}? In this situation, it is of interest to ask, "What values of μ_{A-B} are consistent with the observed results of our study?" The methods described in "Estimation From Samples" can provide a 95% confidence interval for μ_{A-B}:

$$95\%\ CI = \overline{\Delta} \pm t^*_{N-1} \cdot SE$$

The value of t^*_{N-1} is obtained from a standard statistical table: for the present example (N = 10) it is 2.26.

Thus,

$$95\%\ CI = 0.440 \pm 2.26 \cdot 0.334$$
$$= 0.440 \pm 0.750$$

So we may be confident that the interval from −0.310 to +1.190 contains the true value of μ_{A-B}. The confidence stems from the fact that intervals constructed by this method contain the true value in 95% of trials with different samples. Obviously, the result obtained in our small sample could have occurred with no real underlying difference or with a sizable positive real difference (level of free thyroxine increased) or even a negative real difference (level of free thyroxine decreased).

Question 3: How accurately have we estimated μ_{A-B}? (Was the sample large enough?) In general, confidence intervals are very useful in assessing the adequacy of sample size. If an effect exists, the harder we look for it (the larger our sample) the more likely we are to find it. A wide confidence interval says that we have not examined a large enough sample, and in that circumstance, failure to produce a small P value should not be regarded as demonstration that no effect exists.

To illustrate this point further, suppose that in the previous example the same mean increase ($\overline{\Delta}$ = 0.440) and standard deviation (s = 1.057) had resulted from a sample of size N = 100. In this case, calculations similar to those described above reveal P <0.001, indicating that (if there is no real difference) random variation would produce the observed effect less than one time in 1,000. (The symbol < is read "less than"; conversely, the symbol > is read "greater than.") Similarly, the 95% confidence interval becomes 0.231 to 0.649—much narrower than with the original small sample and no longer including 0.

A schema for the paired t-test is shown in Figure 1.

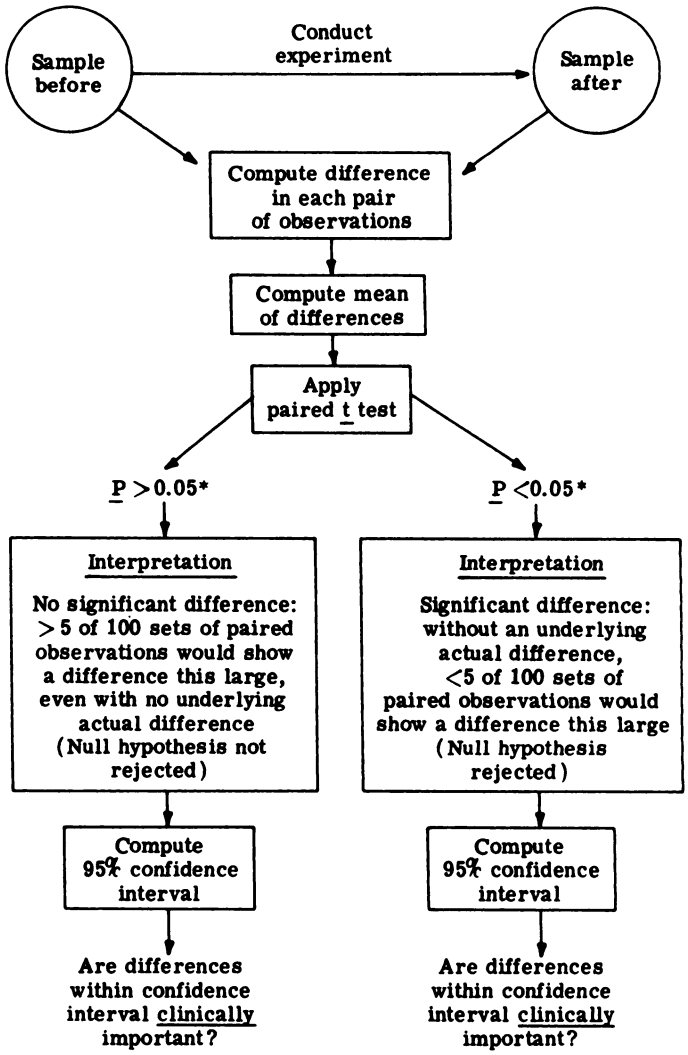

FIG. 1. Schema of analysis for difference in paired data. Asterisks indicate that limits other than 0.05 could be used. (Modified from "Statistics for Family Physicians," in *Family Practice*, O'Brien PC, Shampo MA, Bachman JW. Philadelphia, W. B. Saunders Company [in press])

Comment. 1. One might ask, "How small a *P* value is required to achieve statistical significance?" The answer to this question depends on the circumstances of the particular study, and in general it is best not to think in terms of yes and no—significant or not significant. However, for guidelines one may consider *P* values between 0.10 and 0.05 as suggestive of a difference, though not statistically significant. The term "statistically significant" is usually reserved for situations where P is less than 0.05; and often the evidence of a difference is not considered conclusive unless the P value is less than 0.01.

2. Although the evidence of a heparin effect in the present example, with N = 100, would be described as statistically significant (not likely to occur in the absence of a heparin effect), the more important question—is it clinically significant?—is still unanswered. Whereas the statistician can help in addressing this very important question by providing confidence limits, as in the example, the ultimate decision must come from the clinician.

3. To provide the paired observations for the paired *t*-test, each item in one data set must have an intrinsic correspondence with one—and only one—item in the other set. "Before" and "after" measurements from the same person (as in our example) are a frequent source of paired data. Pairing of data from different persons may be appropriate if the persons have been carefully matched. For instance, in comparing the effects of two drugs, an investigator might exclude genetic variation by using identical twins—giving drug X to one and drug Y to the other. The resulting paired data would be analyzed as in our example. More commonly, there may be two or three factors with major influence on response to treatment, making it desirable to recruit subjects in pairs—the members of each pair being similar to each other with respect to the factors identified as most important. Then, after one member of the pair is treated and the other is not, an observed difference between them should reflect response to treatment.

COMPARING TWO SAMPLES (THE TWO-SAMPLE *t*-TEST)

In the previous example, the study design consisted

simply of obtaining measurements before and after the administration of heparin in a series of consecutive cases. Although this type of study is satisfactory for many research objectives, it is not adequate for a full assessment of the effectiveness of a drug. Specifically, we want to know how the effect observed under our experimental conditions compares with what would happen under other conditions. For example, in many applications the response to treatment may be due—in whole or in part—merely to a psychologic response of the patients who receive the medication. In such situations, one may distinguish the biologic from the psychologic components by including a control group in the study. These patients are similar to the experimental group, but they are to be given only a placebo—a preparation that resembles the experimental drug in all outward respects but has no biologic capability to affect the variable under study. Of course it would be desirable in such situations to ensure that the patient has no knowledge of which medication he is receiving (drug or placebo), in which case the study is said to be "single-blind." If the persons who perform examinations and make decisions in the course of the experiment also are not allowed to know whether drug or placebo is given, the study is said to be "double-blind."

Even when the psychologic effects are not an issue, a meaningful interpretation of the data typically requires comparison with some other regimen. Often, readily available historical data will be helpful. However, in view of potential differences in the patient populations, as well as other factors that may be difficult to quantify, it is usually desirable to incorporate a contemporary comparison group in the study design.

We shall continue our example regarding the effect of heparin on the free thyroxine concentration by supposing that the objective is to compare the results of 12 hr of heparin administration with those of 15 min of heparin administration. Suppose that half the study group was randomly assigned to the 12-hr regimen and half to the 15-min regimen. Suppose the average increase in free thyroxine level was 0.640 ng/dl for the former ($\bar{\Delta}_{12\text{ hr}} = 0.640$), compared with 0.100 ng/dl for the latter ($\bar{\Delta}_{15\text{ min}} = 0.100$), with corresponding standard deviations of $s_{12\text{ hr}} = 1.057$ and $s_{15\text{ min}} = 0.800$, respectively.

Question 1: Is there a difference? Although the apparent effect of 12-hr administration of heparin is greater than that of 15-min administration (0.640 vs. 0.100), we ask the familiar question: What is the probability of obtaining such an apparent difference in the absence of a real difference? Assuming that the data contain no outliers or severe skewness, and noting that the standard deviations are similar, we compare the mean difference between groups with the variability present in both groups:

$$t = \frac{\bar{\Delta}_{12\text{ hr}} - \bar{\Delta}_{15\text{ min}}}{s_{\text{pooled}}\sqrt{\frac{1}{N_{12\text{ hr}}} + \frac{1}{N_{15\text{ min}}}}}$$

in which:

$\bar{\Delta}_{12\text{ hr}} - \bar{\Delta}_{15\text{ min}}$ = difference between mean change in the 12-hr group and mean change in the 15-min group

$N_{12\text{ hr}}$ = number of patients in the 12-hr group

$N_{15\text{ min}}$ = number of patients in the 15-min group

s_{pooled} = a combination of the standard deviations of the two groups (by a method we will not describe)

(Note that the denominator

$$S_{\text{pooled}}\sqrt{\frac{1}{N_{12\text{ hr}}} + \frac{1}{N_{15\text{ min}}}}$$

is analogous to the denominator in the equation for t in the one-sample t-test. It is the standard error of the difference in the numerator. In all of our examples using the t-test, no matter how complicated the equation becomes—how many factors or symbols are included—we are still computing a relative deviate, dividing the numerator by its standard error.)

If we suppose that there were 100 patients in each group, computation of s_{pooled} gives 0.916; and appropriate substitutions yield:

$$t = \frac{0.640 - 0.100}{0.916\sqrt{\frac{1}{100} + \frac{1}{100}}} = 4.17$$

From suitable tables or computing facilities we find that, if there were no difference between the effect of heparin at 12 hr and that at 15 min, a value of t as large as 4.17 would be obtained from 1.8% of repeated experiments ($P = 0.018$). Thus the data are not consistent with the hypothesis of no difference at the $P = 0.018$ level. So we reject the null hypothesis: the observed result is so unlikely to occur without a real underlying difference that there almost certainly is such a difference.

Question 2: How much difference? The next question is: How much drug effect do these data imply? In this example, the 95% confidence interval for the true mean difference (drug effectiveness) is given by:

$$95\%\text{ CI} = (\bar{\Delta}_{12\text{ hr}} - \bar{\Delta}_{15\text{ min}}) \pm t^* \cdot s_{\text{pooled}} \cdot \sqrt{\frac{1}{N_{12\text{ hr}}} + \frac{1}{N_{15\text{ min}}}}$$

which, with appropriate substitutions, becomes:

$$\begin{aligned} 95\%\text{ CI} &= (0.640 - 0.100) \pm 2 \cdot 0.916 \cdot \sqrt{0.02} \\ &= 0.540 \pm 0.260 \\ &= 0.280 \text{ to } 0.800 \end{aligned}$$

Although we cannot be certain that this interval contains the true difference, the method employed does provide

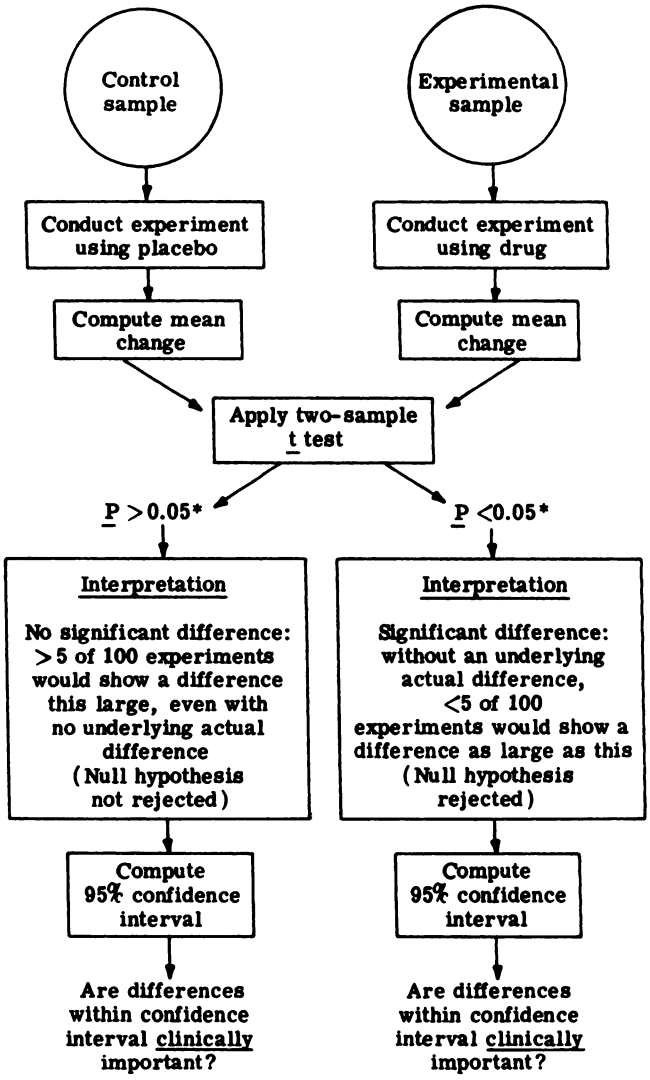

FIG. 2. Schema of analysis for difference between two independent samples. Asterisks indicate that limits other than 0.05 could be used. (Modified from "Statistics for Family Physicians," in *Family Practice*, O'Brien PC, Shampo MA, Bachman JW. Philadelphia, W. B. Saunders Company [in press])

an interval containing the true difference in 95% of applications.

As before, the most important question must now be addressed by the physician: Is the effectiveness of the drug significant clinically?

A schema for the two-sample *t*-test is shown in Figure 2.

Comment. In addition to the design considerations mentioned thus far, numerous others need to be considered in developing a research protocol. Typically, one must be careful to define clearly and in advance of the study what the criteria will be for admission into the study. These should be reported with the study results so that the reader may judge whether the patient population of interest to him is similar. In short, many factors would need to be considered, requiring the close cooperation between clinician and statistician, before the data are collected.