

## Clinical biostatistics

### XXXIV. The other side of 'statistical significance': alpha, beta, delta, and the calculation of sample size

Alvan R. Feinstein, M.D.\* *New Haven, Conn.*

*The Departments of Medicine and Epidemiology of the Yale University School of Medicine*

'Statistical significance' is commonly tested in biologic research when the investigator has found an impressive difference in two groups of animals or people. If the groups are relatively small, the investigator (or a critical reviewer) becomes worried about a statistical problem. Although the observed difference in the means or percentages is large enough to be biologically (or clinically) significant, do the groups contain enough members for the numerical differences to be 'statistically significant'?

For example, if Group A has a mean of 9.8 units and Group B has a mean of 17.3 units, the difference may be biologically impressive because the second mean is almost twice as large as the first. On the other hand, if the two groups each contain only a few members, or if the data are widely dispersed around the mean values, our biologic impression may not be sustained numerically. The statistical assessments may show that the observed difference could quite easily have arisen by chance alone.

The statistical procedures used to test the numerical 'significance' of an observed difference between two groups have been discussed in several previous installments<sup>4, 6</sup> of this series. The calculations used for the procedures depend on the kind of basic data in which the results were expressed. For dimensional data, the results would be cited as means and the usual statistical procedure would be a t test. For nominal or

existential data, the results are expressed as frequency counts that are converted to proportions, percentages, or rates; and the usual statistical procedure would be a chi-square test. (To avoid making unproved assumptions about the distribution of a hypothetical parent population, we can replace the t test by a Pitman permutation test and the chi-square test by a Fisher exact probability test.) If the data are expressed in ranked ordinal values, the usual statistical procedure would be the Wilcoxon rank sum test or the Mann-Whitney U test.

Although each of these tests is chosen according to the type of data under examination, the underlying statistical strategy is identical. It follows the same principle that was used to prove theorems in elementary school geometry. We assume that a particular conjecture is true. We then determine the consequences of that conjecture. If the consequences produce an obvious absurdity or impossibility, we conclude that the original conjecture cannot be true, and we reject it as false.

When this reasoning is used for the statistical strategy that is called "hypothesis testing", the argument proceeds as follows. We have observed a difference, called  $\delta$  (delta), between Groups A and B. To test its 'statistical significance', we assume, as a conjecture, that Groups A and B are actually not different. This conjecture is called the *null hypothesis*. With this assumption, we then determine how often a difference as large as  $\delta$ , or even larger, would arise by chance from data for two groups having the same number of members as A and B. The result of this determination is the *P value* that emerges from the statistical test procedure.

Supported by Public Health Service Grant No. HS00408 from the National Center for Health Services Research and Development.

\*Professor of Medicine and Epidemiology, Yale University School of Medicine, New Haven, Conn. Senior Biostatistician, Cooperative Studies Program Support Center, Veterans Administration Hospital, West Haven, Conn.

At this point in the reasoning, the statistical strategy departs from what was used to prove theorems in grade school geometry. In geometry, there were no problems in deciding whether or not to reject the assumed conjecture, because the geometrical logic regularly brought us to a situation that was impossible, i.e., the P value was zero. In such a circumstance, the original conjecture could not be maintained. It had to be wrong because it could not possibly be right. With statistical inference, however, the results can seldom, if ever, be so conclusive. The P value that emerges from the calculations in the statistical test may be as small as .000001, or even smaller, but it never becomes zero. There is always a possibility, however infinitesimal, that the observed difference arose by chance alone. Accordingly, unlike the situation in geometry, we cannot use a statistical test to prove with total certainty that the original conjecture is wrong. There is always a chance of 1 in 20, or 1 in 50, or whatever the P value is, that the original conjecture (i.e., the null hypothesis) is right.

To draw statistical conclusions, therefore, we must establish a concept that was not necessary for the inferential reasoning of grade school geometry. This concept is called an  $\alpha$  (alpha) level of 'significance'. It is used to demarcate the *rejection zone*. If the P value that emerges from the statistical test is equal to or smaller than  $\alpha$ , we decide that we shall reject the null hypothesis. In doing so, we demarcate  $\alpha$  as the risk of being wrong in this conclusion—but it is a risk we must take in order to have a statistical mechanism for drawing conclusions. In geometrical inference,  $\alpha$  was always zero. In statistical inference,  $\alpha$  is customarily chosen to be .05, i.e., 1 in 20, although some investigators (or editors) may select other boundaries such as .1 or .01.

A previous paper<sup>6</sup> of this series contained a discussion of the arbitrary way in which .05 became designated as the customary level of  $\alpha$ . The designation came, not as a pronouncement of the Deity or from the deliberations of an international committee, but from a habit of R. A. Fisher. Noting that a conclusion had to be drawn after a statistical test was performed, and know-

ing that an  $\alpha$  level was necessary to draw the conclusion, Fisher chose  $\alpha$  to be .05. The rest of the statistical world followed.

#### A. Statistical reasoning and diagnostic analogies

This reasoning is regularly applied in a way that makes the statistical appraisal of 'significance' resemble a clinical diagnostic test.

1.  *$\alpha$  level and 'diagnostic specificity'*. In using .05 or whatever other  $\alpha$  level is selected as boundary for the rejection zone, an investigator specifies the deliberate chance that he wants to allow of being wrong when he decides that the observed difference in his two groups is real. There will exist a probability of magnitude P, however, that the null hypothesis is correct—that the observed difference in the groups has arisen simply by chance, and that the conclusion is wrong.

The  $\alpha$  level is thus analogous to the risk of getting a *false positive* result in a diagnostic test<sup>5</sup>. Suppose we make a diagnosis of lung cancer after finding a positive result in the Pap smear of a patient's sputum. If the patient does in fact have lung cancer, the diagnostic decision is correct—a true positive. If the patient does not have lung cancer, the diagnosis is wrong—a false positive conclusion. In the customary situation of hypothesis testing, we want to make a positive decision, rejecting the null hypothesis and concluding that the observed difference is real. The  $\alpha$  level indicates the statistical risk that this decision may be wrong and that there is actually no difference between the groups. The value  $1-\alpha$  can therefore be likened to the *specificity* of a diagnostic test, which is the likelihood that the test will have a negative result when the disease is absent. The value of  $1-\alpha$  denotes the likelihood of being correct when we do not reject the null hypothesis and thereby conclude that the observed difference is not 'statistically significant'.

The kind of reasoning used in forming a 'null hypothesis' and in establishing levels of  $\alpha$  and  $1-\alpha$  is based on the idea that the 'disease' we are looking for, i.e., a real difference between the groups, is absent. The chance of a false positive diagnosis is  $\alpha$ ; and the chance of a true negative

diagnosis is  $1-\alpha$ . Consequently, if we set  $\alpha$  at .05, we take a 5% chance of being wrong if we reject the null hypothesis (i.e., draw a positive conclusion) and a 95% chance of being right if we concede the null hypothesis (i.e., fail to draw a positive conclusion).

This analogy to a false positive result and to the specificity of a diagnostic test can make the  $\alpha$  and  $1-\alpha$  concepts particularly easy for clinicians to understand, since the idea of a diagnostic test does not appear in the general statistical phrase by which  $\alpha$  level is usually called *Type I error*. Probably the main reason for the statistical nomenclature is the difference in the way investigators and statisticians use the inverted logic that customarily goes into hypothesis testing. To an investigator, the test is usually done for a positive reason—to demonstrate that a biologically impressive difference is also statistically impressive. The investigator thus regards a doubly negative phenomenon (rejection of the null hypothesis) as a *positive* event, analogous to getting a positive result in a diagnostic test. In general statistical usage, however, the acceptance or rejection of the null hypothesis is seldom associated with any negative or positive intellectual virtues. Thus, for statistical definitions, a Type I error consists of rejecting the null hypothesis when it is actually true.

2. *The calculation of  $P_A$* . To apply these principles requires the calculation of a P value for the observed data and the observed difference,  $\delta$ . This particular P value, which is the conventional one usually cited in medical literature, will be designated here as  $P_A$  to distinguish it from other P values that will be discussed later. The procedures used for calculating  $P_A$  are presented in detail in textbooks of statistics and will be summarized as follows:

a. The simplest and most generally applicable statistical strategy rests on the idea of a "critical ratio" or "z-score". For any single value randomly chosen from a Gaussian distribution whose constituent values are  $x_1, x_2, x_3, \dots$ , a critical ratio can be calculated as  $z = (x_1 - \mu)/\sigma$ . In this formula,  $\mu$  is the mean of the parent distribution;  $\sigma$  is its standard deviation; and  $x_1$  is the single value with which we are concerned. If the data under consideration consist of

a series of means of samples, each of which has  $n$  members randomly drawn from the same parent population, these means also have a Gaussian distribution. Their mean is also  $\mu$ , but their standard deviation is  $\sigma/\sqrt{n}$ . This standard deviation of a series of means is called SE, the 'standard error' of the mean. The corresponding critical ratio for any of these means,  $\bar{x}$ , is  $z = (\bar{x} - \mu)/(\sigma/\sqrt{n})$ .

b. If we want to analyze a difference in the means,  $\bar{x}$  and  $\bar{y}$ , of two samples, we seek the distribution of a new variable,  $w = \bar{x} - \bar{y}$ . This variable, which has its own mean, will have a "common" standard deviation that can be calculated in one of two ways. If we assume, by the null hypothesis, that the population variances of  $\bar{x}$  and  $\bar{y}$  are equal, we can create a "pooled variance" for  $w$ . If we do not assume equality of population variances for  $\bar{x}$  and  $\bar{y}$ , then the common variance of  $w$  equals the sum of the variance of  $\bar{x}$  and the variance of  $\bar{y}$ . In particular, if the mean values,  $\bar{x}$  and  $\bar{y}$ , happen to be proportions,  $p_1$  and  $p_2$ , we can calculate the common variance as follows:

1. If we assume that the population values for  $p_1 = p_2$ , the common mean for both samples is  $\bar{p} = (np_1 + np_2)/2n = (p_1 + p_2)/2$ . The common variance of the difference in proportions is  $2\bar{p}(1 - \bar{p})/n$ .

2. If we do not assume that the population values for  $p_1 = p_2$ , the common variance is calculated as  $[p_1(1 - p_1) + p_2(1 - p_2)]/n$ .

The corresponding z values for these two cases would be  $(p_1 - p_2)/\sqrt{2\bar{p}(1 - \bar{p})/n}$  in the first instance, and  $(p_1 - p_2)/\sqrt{[p_1(1 - p_1) + p_2(1 - p_2)]/n}$  in the second. In general, the formula for calculating the z value of a difference in means is

$$z = \frac{\text{difference in means}}{\text{"standard error" of the difference}}$$

c. Regardless of whether a particular critical ratio of z is calculated for a single mean or for a difference in means, the z values have some distinctive, important properties. If we drew a large series of samples, consisting of single means or differences in means, and if the sample sizes were themselves large, and if a z score were calculated for each sampling, the array of z values would approximate a 'standard normal' distribu-

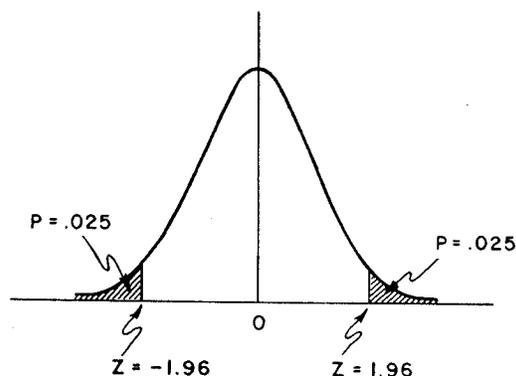


Fig. 1. Standard Gaussian distribution showing values of  $z$  for two-sided  $P = .05$ .

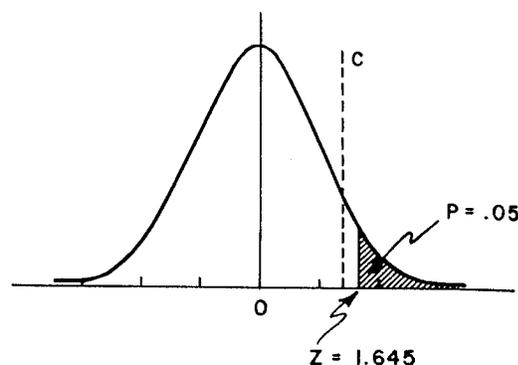


Fig. 2. Standard Gaussian distribution showing value of  $z$  for one-sided  $P = .05$ . Line drawn at  $c$  represents an observed value.

tion, having a mean of 0 and a standard deviation of 1. Furthermore, each positive (or negative) value of  $z$  will be associated with a value of  $P$ , which represents the amount of 'exterior' probability as the area that lies beneath the standard Gaussian curve to the left (or right) of an ordinate erected at  $z$ . Because the entire area below the curve has a probability value of 1, the value of  $P$  for the area beyond  $z$  will be .5 when  $z = 0$  (since half of the curve lies to the right—or left—of the corresponding ordinate). Some other pertinent results, which can be found in the tables of most statistical textbooks, are as follows:

$z$ value	0	1.28	1.645	1.96	2.58
$P$ value	.5	.1	.05	.025	.005

d. For any selected exterior level of probability,  $\alpha$ , there will correspond a value of  $z_\alpha$ , denoting the point on the abscissa at which an area of  $\alpha$  is cut off under the probability curve. Thus, if  $\alpha$  is set at .005,  $z_{.005} = 2.58$ . The choice of the associated  $z$  values for an  $\alpha$  level depends on whether the area of exterior probability is being considered in a one-sided or two-sided direction. For a two-sided test, the exterior probability area is divided symmetrically at the extremes of the curve. Thus, if  $\alpha = .05$  for a two-sided test, we would seek the value of  $z$  for  $\alpha/2$ , which is  $z_{.025} = 1.96$ . This distinction is shown in Fig. 1. For a one-sided test, however, all of the exterior probability is placed on one side of the curve, as shown in Fig. 2. Thus, if  $\alpha = .05$  but if the test is one-sided, we would

seek the value of  $z_{.05}$ , which is 1.645. This shift from a two-sided to a one-sided test therefore reduces the value of  $z$  from 1.96 to 1.645. The line drawn at  $c$  in Fig. 2 shows the location of an observed difference,  $p_2 - p_1$ , and the associated  $z$  value. (The actual  $z$  value is determined as  $z_c = (p_2 - p_1)/SE$ .) Since  $c$  does not lie within the shaded boundary, the observed  $p_2 - p_1$  difference would be regarded statistically as 'not significant' in a one-sided test at the selected level of  $\alpha$ .

e. With an appropriate statistical table showing  $z$  and  $P$  values, we can thus readily move back and forth from an observed difference in two proportions (or means), to a value of  $z$ , to an associated value of  $P$ . When the calculations are performed after the data have been obtained, this  $P$  value is the  $P_A$  mentioned at the beginning of this section.

3. *The calculation of sample size for  $\alpha$  and  $\Delta$  levels.* The Type-I or  $\alpha$ -error approach that has just been described has often been used to determine—before the research is performed—the sample size required to attain 'statistical significance' in a contrast of two groups. For this process, the formula that was used to get the value of  $z$  is algebraically manipulated so that we solve for  $n$  instead.

a. *Information needed for calculations.* Four items of information must be assigned, estimated, known, or assumed in order to do the calculations.

1. We must assign a value of  $\delta$  that

will be regarded as a biologically impressive difference between the means (or rates) found in the control group and in the treated (or 'experimental') group. Because this value of delta is assigned before the research, it will be designated here as  $\Delta$ , to distinguish it from the  $\delta$  value observed after the data are obtained.

2. We must estimate the standard error of the difference in the means or rates of the two groups. For rates or percentages, we will know from previous research the particular value,  $p_1$ , that is to be expected in the control group. The experimental group will then be required to have a value  $p_2$  which is higher or lower than  $p_1$  by the amount,  $\Delta$ . We can then calculate the standard error of the assigned difference,  $p_2 - p_1$ , by using the formula described in the previous section. [For dimensional data, we take  $\Delta$  to be the difference in means, and we assume that the experimental group will have the same variance (or standard deviation) that is expected in the control group.]

3. We assign a level of  $\alpha$  (thereby indicating  $z_\alpha$ ) as the chosen level of 'significance.' [This level of  $\alpha$  actually represents  $P_\alpha$ .]

4. We assume that the two groups will each be of size  $n$ , and that the total sample size will be  $N = 2n$ .

*b. Example of calculations.* To illustrate this process, suppose we will regard a new treatment as 'useful' if it raises the percentage of success by 10% from its customary level of 50% in the control group. For a one-sided result, statistically significant at the level of  $\alpha = .05$ , how large a sample do we need if this difference actually occurs?

We have assigned  $\Delta = .1$  and  $\alpha = .05$ . Since the test is one-sided, we find from the table that  $z_{.05} = 1.645$ . Since we know that  $p_1 = .50$ , we estimate  $p_2 = p_1 + \Delta = .60$ . We quickly determine  $\bar{p}$  as  $(p_1 + p_2)/2 = .55$ . Assuming the null hypothesis, the common variance of the difference in  $p_1$  and  $p_2$  is  $2\bar{p}(1 - \bar{p})/n$ , which is  $(2)(.55)(.45)/n = (.495)/n$ . We now have all the information we need to solve for  $n$  in the formula

$$z = \frac{\Delta}{\sqrt{2\bar{p}(1 - \bar{p})/n}}$$

Squaring both sides and isolating  $n$ , we get

$$n = \frac{z^2 [2\bar{p}(1 - \bar{p})]}{\Delta^2}$$

Substituting appropriately, we find

$$n = \frac{(1.645)^2 (.495)}{(.1)^2} = 133.9 \approx 134.$$

Since  $n$  is the size of one sample, the total group needed for the proposed research would be  $2n$ , or 268 patients. Had we decided to use a two-sided test of probability (just in case the treatment was 10% worse than the controls) the value of  $z$  would have been  $z_{.025} = 1.96$  and  $n$  would have been calculated to be  $(1.96)^2 (.495)/(.1)^2 = 190.2 \approx 191$  patients. The total required sample size would have been 382 patients.

Readers who feel more comfortable with the chi-square test of 'significance' than with the  $z$  procedure might like to see how this same result can be attained using the chi-square formula. As shown elsewhere<sup>3</sup>, the formula for calculating chi-square for frequency data expressed as two proportions, each from a group of size  $n$ , is

$$\chi^2 = \left[ \frac{(n)(n)}{2n} \right] \left[ \frac{\Delta^2}{\bar{p}\bar{q}} \right]$$

where  $\bar{q} = 1 - \bar{p}$ . This formula can be solved for  $n$  to yield  $n = 2\bar{p}(1 - \bar{p})\chi^2/\Delta^2$ . For  $\alpha = .05$  in a two-sided chi-square test at 1 degree of freedom (as befits a test of two proportions), the critical value of  $\chi^2$  is 3.84. Substituting this value of  $\chi^2$  into the formula gives us exactly the same result that was obtained with the  $z$  procedure.

The reason for using the relatively unfamiliar  $z$  procedure, rather than our old friend chi-square, is that the  $z$  procedure can be applied for dimensional data as well as for proportions. More importantly, the  $z$  procedure is especially useful for illustrating the additional concepts that are to appear shortly.

*4. The role of  $\beta$  and 'false negatives'.* All of the strategies and tactics that have just been discussed constitute a long-standing, well established statistical procedure that is still used by many investigators as the 'natural' way to calculate sample size. In 1928, however, Jerzy Neyman and Egon S. Pearson<sup>10</sup> pointed out that the reasoning was incomplete. According to the Neyman-Pearson argument, a statistical test of 'significance', like a medical test of

**Table I.** Analogies of conclusions in diagnostic and statistical reasoning

Diagnostic reasoning		
Result of diagnostic test	Disease is really:	
	Present	Absent
Positive	True positive diagnosis	False positive diagnosis
Negative	False negative diagnosis	True negative diagnosis
Statistical reasoning		
Result of statistical test	Significant difference is really:	
	Present, i.e., null hypothesis not true	Absent, i.e., null hypothesis true
Reject null hypothesis	No error; probability: $1-\beta$	Type I error; Probability: $\alpha$
Concede, i.e., do not reject null hypothesis	Type II error; probability: $\beta$	No error; probability: $1-\alpha$

diagnosis, has another side to it. In diagnosis, when thinking about the situation where the disease is absent, we recognize that a diagnostic test will yield either a *false positive* diagnosis or a *true negative* diagnosis, but what about the situation where the disease is present? We have thus far ignored this other side of diagnostic reasoning. What about the *false negative* or *true positive* diagnoses that will occur if the disease actually exists? In statistical reasoning, the counterpart to a false negative diagnosis is the error we would make if we conceded the null hypothesis and concluded that the observed difference was not statistically significant when, in fact, an important difference really existed.

In Fig. 2, for example, the value of  $c$  that was converted to a  $z$  score for the observed difference,  $p_2 - p_1$ , would have been declared 'not significant' statistically. How can we be sure of this decision? Perhaps the value of  $z_c$  is compatible with some other curve drawn farther to the right, in which the true difference of  $p_2 - p_1$  is larger and more biologically significant than what was observed? If the true difference is biologically important and if we

failed to draw that conclusion, we would have made an erroneous decision. This type of false negative conclusion is what statisticians call a *Type II error*, and a new array of reasoning was created to work out its mathematical relationships.

The first step in the proceedings is to get a mathematical name for the likelihood of this type of error. If  $\alpha$  was the boundary of probability demarcated for a false positive error, we can use the symbol  $\beta$  as the boundary for false negative error. Furthermore, if  $1-\alpha$  corresponds to the *specificity* of a diagnostic test,  $1-\beta$  will correspond to its *sensitivity*—the likelihood of making a positive diagnosis when the disease is present. Consequently, working at selected levels of  $\alpha$  and  $\beta$ , we would have a  $1-\beta$  chance of being right when we reject the null hypothesis (if it is false), and a  $1-\alpha$  chance of being right when we concede it (if it is true). The associated analogies of the "diagnostic" and statistical reasoning are shown in Table I.

The terms *sensitivity* and *specificity* provide an idea of the diagnostic power of a test and help indicate the confidence that can be placed in the test. These same concepts are used (although somewhat differently) for the statistical nomenclature. As shown in Table I, the level of  $1-\beta$ , corresponding to the sensitivity of the test in correctly making a positive diagnosis, is often called the statistical *power* of the test. The level of  $1-\alpha$ , corresponding to the specificity of the test in correctly making a negative diagnosis, does not have a particular statistical word, such as *power*, attached to it. Its level is sometimes cited, however, as the *confidence* with which a null hypothesis is conceded. With this set of vocabulary and concepts, we can characterize a statistical test of significance by citing the  $\alpha$  level and the power (or  $1-\beta$  level) at which the test is employed for a difference,  $\Delta$ . We might thus want to talk about a particular test as having a 95% chance ( $= 1-\beta$ ) of correctly rejecting the null hypothesis at the 5% level ( $= \alpha$ ) if the true difference is  $\Delta$ .

The value of  $\Delta$  is what allows us to specify just what is happening during the tests of statistical hypothesis that are under scrutiny. In ordinary statistical testing, the hypothesis we

really wanted to accept is called the *alternative hypothesis*. It is expressed as  $H_A$  and is written (for a two-sided test) as  $H_A: \text{mean}_1 \neq \text{mean}_2$ . For a one-sided test,  $H_A$  could be written as  $\text{mean}_1 < \text{mean}_2$  or as  $\text{mean}_1 > \text{mean}_2$ . To accept this alternative hypothesis, we engaged in a pattern of reasoning that called for rejection of the null hypothesis, which is written as  $H_0: \text{mean}_1 = \text{mean}_2$ .

When we add a  $\beta$ -type of reasoning to this previous form of  $\alpha$  reasoning, we become interested in testing two separate hypotheses. One of these is the conventional null hypothesis,  $H_0$ , which says  $\text{mean}_1 = \text{mean}_2$ . The other is the alternative hypothesis,  $H_A$ , which is expressed in the form of  $\text{mean}_2 - \text{mean}_1 = \Delta$  or (if two-sided probabilities are used) as  $|\text{mean}_2 - \text{mean}_1| = \Delta$ . The realities, decisions, and corresponding probabilities of error are shown in Table II. The most important point to be noted here is that a Type II (or beta) error can occur when we accept the null hypothesis. Since this is equivalent to the error of falsely rejecting  $H_A$ , we can determine the magnitude of probabilities for a Type II error by considering the consequence of a rejection of  $H_A$ .

An illustration of this situation is shown in Fig. 3. On the left of this figure is the Gaussian distribution of  $z$  values under the assumption that  $H_0$  is true. To the right of the vertical line drawn at  $c$ , we have a zone in which  $H_0$  will be falsely rejected at an alpha level that corresponds to  $c$ . The curve at the right of the figure shows the distribution of  $z$  values under the assumption that  $H_A$  is true. To the left of the vertical line drawn at  $c$ , we have a zone in which the acceptance of  $H_0$  will be associated with a false rejection of  $H_A$ .

### B. Strategies in $\beta$ statistics

We can now contemplate two new kinds of statistical procedures: a new kind of  $P$  value and an additional way of determining sample size.

1. *The calculation of  $\beta$ -error.* In an ordinary *post hoc* test of 'statistical significance', the  $P_A$  value that we determined (from the  $z$  value of the observed results) told us about the possibility of a Type I or  $\alpha$  error. In calculating

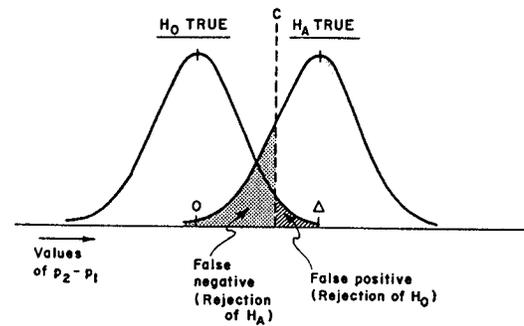


Fig. 3. Distributions for null hypothesis,  $H_0$ , and for alternative hypothesis,  $H_A$ . For further details, see text.

Table II. Hypotheses, conclusions, and types of statistical error

Reality	Decision*	Type of error	Probability
{ $H_0$ true; } { $H_A$ false }	Accept $H_0$	None	$1-\alpha$
	Reject $H_0$	False positive	$\alpha$
{ $H_0$ false; } { $H_A$ true }	Accept $H_0$	False negative	$\beta$
	Reject $H_0$	None	$1-\beta$

\*If  $H_0$  is accepted,  $H_A$  is rejected; and vice versa.

sample size, we chose a specific level for this error and called it  $\alpha$  or  $P_\alpha$ . If we assign a  $\Delta$  value, however, and contemplate just how large the true difference might really be for the two contrasted groups, we can determine the 'other side' of statistical significance. We can calculate a  $P_B$  value for the possibility of a Type II or  $\beta$  error.

The procedure used for this determination can be illustrated in reference to the curves of Fig. 3.

a. The left hand curve in Fig. 3 has its mean at 0, and represents the distribution of values for  $p_2 - p_1$  under the null hypothesis, with  $H_0$  assumed to be true.

b. The right hand curve in Fig. 3 has its mean at  $\Delta$ , and represents the analogous distribution of  $p_2 - p_1$  under the alternative hypothesis,  $H_A$ .

c. The line drawn at  $c$  represents an observed value of  $p_2 - p_1$ . [Alternatively, as

we shall see later, it can represent an assigned value for  $\alpha$  or for  $\beta$ .]

d. A  $z$ -value can be determined for any of these points by appropriate reference to a standard error. Thus, for an observed value of  $p_2 - p_1 = \delta$ , under the null hypothesis,

$$z_c = \frac{\delta}{\sqrt{[2\bar{p}(1 - \bar{p})]/n}}$$

and

$$z_\Delta = \frac{\Delta}{\sqrt{[2\bar{p}(1 - \bar{p})]/n}}$$

The value of  $z_c$ , as shown in the hatched area of Fig. 3, will indicate the probability of a false positive rejection of  $H_0$ .

e. Under the alternative hypothesis,  $H_A$ , the stippled area to the left of  $c$  will indicate the likelihood of a false positive rejection for  $H_A$ . If  $c$  represents the observed value of  $\delta$ , this point is located at  $\Delta - \delta$ , and the associated negative value for  $z_B$  is

$$z_B = \frac{\Delta - \delta}{\sqrt{[p_2(1 - p_2) + p_1(1 - p_1)]/n}}$$

f. For most practical purposes, we can assume that the two standard errors are equal, i.e., that  $2\bar{p}(1 - \bar{p}) = p_2(1 - p_2) + p_1(1 - p_1)$ . For example, if  $p_1 = .50$  and  $p_2 = .70$ , the respective values for these terms are 0.48 and 0.46. Consequently, if we let  $v = 2\bar{p}(1 - \bar{p}) \approx p_2(1 - p_2) + p_1(1 - p_1)$ , the formulas we have been considering become

$$z_c = \frac{\delta\sqrt{n/v}}{\sqrt{v}}$$

$$z_\Delta = \frac{\Delta\sqrt{n/v}}{\sqrt{v}}$$

and

$$z_B = (\Delta - \delta)\sqrt{n/v} = z_\Delta - z_c.$$

g. By referring this  $z_B$  value to the associated one-sided  $P$  value, we obtain  $P_B$ , which is the probability of falsely rejecting the alternative hypothesis.

## 2. Illustration of calculation for $P_B$

a. *Equal sample sizes.* Suppose an investigator, comparing the rates of patient satisfaction with medical care at two hospitals finds that the rate of satisfaction was 70% (16/23) at Hospital A and 84% (19/23) at Hospital B. The investigator concludes that the difference is not statistically 'significant' because chi-

square = 1.08 and  $P$  is too large ( $> .1$ ) to be 'significant'. [Doing the statistical test by the  $z$  procedure, we would choose  $\bar{p} = (16 + 19)/(23 + 23) = 76\%$ . We then get  $z_c = (.83 - .70)(\sqrt{23})/\sqrt{(2)(.76)(.24)} = (.13)(4.8)/\sqrt{.36} = 1.04$ ; and the associated  $P$  value is .15.] After drawing this conclusion, the investigator claims that the care at the two hospitals produces the same degree of satisfaction.

Contrary to his claim, however, we suspect that Hospital B does give better care, that its real level of patient satisfaction is actually 20% higher than in Hospital A, and that 'statistical significance' was absent in this study as an act of chance, possibly because the sample sizes were too small. What we would like to know, therefore, is the likelihood that the investigator may have been wrong in his conclusion. We use the formula  $z_B = (\Delta - \delta)\sqrt{n/v} = (.20 - .13)\sqrt{23/.36} = (.07)(7.99) = 0.56$ . The associated one-sided value for  $P_B$  is .288 or 29%. Thus, there is a good chance (of about 29%) that the investigator falsely accepted the null hypothesis if the true difference in rate of satisfaction at the hospitals is as high as 20%.

b. *Unequal sample sizes.* All of the foregoing calculations were based on the assumption that the observed proportions,  $p_1$  and  $p_2$ , came from groups of equal sizes. If the sample sizes,  $n_1$  and  $n_2$ , are unequal, then  $\bar{p} = (n_1p_1 + n_2p_2)/N$ , where  $N = n_1 + n_2$ . Under the null hypothesis, the standard error of  $p_1 - p_2$

is  $\sqrt{\bar{p}\bar{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$ , where  $\bar{q} = 1 - \bar{p}$ . This

expression becomes  $\sqrt{N\bar{p}\bar{q}/n_1n_2}$ . Under the alternative hypothesis, the standard error

of  $p_1 - p_2$  is  $\sqrt{\frac{p_1q_1}{n_1} + \frac{p_2q_2}{n_2}} = \sqrt{(n_2p_1q_1 + n_1p_2q_2)/(n_1n_2)}$ . For most practical purposes,

we can assume that  $\sqrt{N\bar{p}\bar{q}} = \sqrt{n_2p_1q_1 + n_1p_2q_2}$ . The formula for finding beta error would then be  $z_B = (\Delta - \delta)\sqrt{n_1n_2/N\bar{p}\bar{q}}$ .

## 3. The 'power curve' of a statistical test.

If we want to operate a statistical test at a fixed level of  $\alpha$  for making decisions, we can readily determine the values of  $\beta$  that will be associated for any choice of  $\alpha$ . In this case, the chosen level of  $\alpha$  will determine an assigned value of

$z_\alpha$ , which will be the location of  $z_c$  in Fig. 3. Since we previously developed the formula that  $z_B = z_\Delta - z_c$ , we can substitute the assigned value of  $z_\alpha$  for  $z_c$  and get  $z_\Delta = z_\alpha + z_B$ . Since  $z_\Delta$  has a fixed value that depends on the magnitudes of  $\Delta$ ,  $n_1$ ,  $n_2$ ,  $N$ ,  $\bar{p}$ , and  $\bar{q}$ , the values of  $z_\alpha$  and  $z_B$  must sum to a constant. Consequently, if higher values are assigned to  $z_\alpha$ , the values of  $z_B$  must decrease; and vice versa. This reciprocal relationship is analogous to what we have noted in a previous discussion<sup>5</sup> of sensitivity and specificity for a diagnostic test. If one increases, the other decreases.

This reciprocal aspect of the equation allows various statistical tests to be illustrated with "power" curves, which show the values of  $1 - P_\beta$  that will occur with different choices of  $\alpha$ . For example, consider the situation where the true values of the compared rates are  $p_1 = 27\%$  and  $p_2 = 45\%$ , so that  $\Delta = p_2 - p_1 = .18$  and  $v = 2\bar{p}(1 - \bar{p}) = 0.46$ . If we take a sample size of 80 for each group in a comparison, we will have  $z_\Delta = \Delta\sqrt{n/v} = (.18)\sqrt{80/.46} = 2.37$ . If we decide to reject the null hypothesis at a two-sided  $\alpha$  level of .05, we would have  $z_B = 2.37 - 1.96 = 0.41$ . The associated  $P_B$  value would be .341 and the "power" of the test would be 65.9%. If the null hypothesis were to be rejected at a two-sided  $\alpha$  level of .1,  $z_B = 2.37 - 1.645 = 0.73$ , and the associated  $P_B$  value would be .233, giving the test the higher "power" of 76.7%.

### C. The calculation of a 'doubly significant' sample size

In the foregoing discussion, we worked on the assumption that the research was complete. We had our data; we had determined or assigned the level of  $P_A$  or  $P_\alpha$ ; and we wanted to know what  $P_B$  might be. A different application of these concepts occurs for the modern calculation of a 'doubly significant' sample size, which means that we want a sample large enough to be significant at the levels of both  $\alpha$  and  $\beta$ . In the previous calculations, we began with known data for everything except  $z_B$ , and we solved for  $z_B$ . Now we begin by knowing (or assuming) all the necessary information except  $n$ , and we solve the equation for  $n$ .

1. *Simplified procedure.* The simplest ap-

proach for these calculations is to take the previously developed formula and to substitute the assigned values of  $z_\alpha$  and  $z_\beta$  for their respective counterparts  $z_c$  and  $z_B$ . We then would have

$z_\Delta = \Delta\sqrt{\frac{n}{v}} = z_\alpha + z_\beta$ . If we square both sides and solve for  $n$ , we get  $n = v(z_\alpha + z_\beta)^2/\Delta^2$ . For example, suppose we expect that  $p_2 = .70$  and  $p_1 = .50$  and we want to attain statistical significance for a 2-sided  $\alpha$  level of .05 and a one-sided  $\beta$  level of .05. What size should our sample be? We have assigned  $\Delta = .20$ ,  $z_\alpha = 1.96$  and  $z_\beta = 1.645$ . From previous calculations, we know that  $v$  is either 0.48 or 0.46. Let us call it 0.47. Substituting directly into the cited equation, we get  $n = (0.47)(1.96 + 1.645)^2/(\cdot 20)^2$ , and  $n = 152.7$ . We would thus need 153 patients in each group, for a total sample size of 306 patients.

2. *Stricter procedure.* In a mathematically stricter set of calculations, we make provision for the fact that the value of  $z_\alpha$  is determined using the null hypothesis, whereas the value of  $z_\beta$  depends on the alternative hypothesis. Thus, the point  $c$  in Fig. 3 will define an  $\alpha$  level of

$$z_\alpha = \frac{c}{\sqrt{2\bar{p}(1 - \bar{p})/n}}$$

At this same point  $c$ ,

$$z_\beta = \frac{\Delta - c}{\sqrt{[p_2(1 - p_2) + p_1(1 - p_1)]/n}}$$

If we solve the first of these two equations for  $c$ , and substitute the results into the second, we get

$$z_\beta = \frac{\Delta - (\sqrt{2\bar{p}(1 - \bar{p})/n})(z_\alpha)}{\sqrt{[p_2(1 - p_2) + p_1(1 - p_1)]/n}}$$

This equation becomes

$$\frac{1}{\sqrt{n}} \left\{ z_\beta [\sqrt{p_2(1 - p_2) + p_1(1 - p_1)}] + z_\alpha [\sqrt{2\bar{p}(1 - \bar{p})}] \right\} = \Delta$$

Squaring both sides and solving for  $n$ , we get

$$n = \left( \frac{1}{\Delta^2} \right) \left\{ z_\alpha [2\bar{p}(1 - \bar{p})] + z_\beta [p_2(1 - p_2) + p_1(1 - p_1)] \right\}^2$$

This formula, which is less formidable than it looks, is the one that regularly appears in many statistical discussions<sup>8, 11, 13</sup> of sample size. In

the numerical example just cited, the actual values are

$$\begin{aligned} n &= \frac{1}{(.20)^2} \left\{ 1.96[2(.60)(.40)]^{\frac{1}{2}} + 1.645[(.70)(.30) \right. \\ &\quad \left. + (.50)(.50)]^{\frac{1}{2}} \right\}^2 \\ &= \frac{1}{.04} \left\{ 1.96[.69] + 1.645[.68] \right\}^2 \\ &= \frac{1}{.04} \left\{ 1.35 + 1.12 \right\}^2 = \frac{1}{.04} \left\{ 2.47 \right\}^2 = 153. \end{aligned}$$

The result is identical to what we obtained with the previous set of simplified calculations.

**3. Use of statistical tables.** These computations can be avoided if we make use of appropriate sets of prepared tables. A particularly good collection of tables, showing sample sizes for different values of  $\alpha$ ,  $\beta$ ,  $\Delta$ , and  $p_1$ , is contained in Table A-3 (pages 176-194) of the excellent textbook by Fleiss<sup>7</sup>. For example, if  $\Delta$  is 5%, if  $\alpha$  (two-sided) is .05, and if  $\beta$  (one-sided) is .05, Fleiss' Table A-3 shows that the size of  $n$  will range from 796 if  $p_1 = 5\%$ , to 2669 if  $p_1 = 50\%$ . If  $\Delta$  is 20%, under the same conditions of  $\alpha$  and  $\beta$ , the size of  $n$  will range from 99 if  $p_1 = 5\%$  to 172 if  $p_1 = 50\%$ . The values of  $n$  in Fleiss' tables are higher than those calculated in the preceding illustrations here because Fleiss' computations include a 'correction for continuity', analogous to that of the Yates' 'correction' in chi-square tests. Using a formula derived by Kramer and Greenhouse<sup>9</sup>, the  $n$  values in our calculations can be converted to the  $n'$  values cited by Fleiss. The formula is

$$n' = \frac{n}{4} \left[ 1 + \sqrt{1 + (8/n\Delta)} \right]^2.$$

**4. Differences in sample-size methods.** We can now note the difference between calculating a 'singly' and a 'doubly significant' sample size. In the classical old formula for a 'singly significant' sample

$$n = \frac{z_\alpha^2 v}{\Delta^2}.$$

For the 'doubly significant' sample,

$$n = \frac{(z_\alpha + z_\beta)^2 v}{\Delta^2}.$$

The change to 'double significance' thus increases the sample size by a ratio of  $[(z_\alpha + z_\beta)/z_\alpha]^2$ , which is  $\left[ 1 + \frac{z_\beta}{z_\alpha} \right]^2$ . If we choose equal

values for our levels of  $\alpha$  and  $\beta$ , the right hand ratio of  $z$  values will be 1, and the 'doubly significant' sample will be  $(1 + 1)^2 = 4$  times as large as the first.

To illustrate this point, suppose we want to achieve a one-sided significance level of .05 in a clinical trial where the expected rates of success are 10% in the control group and 20% in the treated group. From these data,  $p_2 = .20$ ,  $p_1 = .10$ ,  $\Delta = .10$ , and  $v$  is estimated as  $(.20)(.80) + (.10)(.90) = 0.25$ . [The alternative estimation for  $v$  would be  $(2)(.15)(.85) = 0.26$ .]

For the 'one-sided' calculations of sample size for  $\alpha$  level significance, we would have

$$n = \frac{z^2 v}{\Delta^2} = \frac{(1.645)^2 (.25)}{(.10)^2} = 67.65$$

and we would need  $2 \times 68 = 136$  patients altogether.

To calculate sample size for both  $\alpha$  and  $\beta$  levels of significance, we would have

$$n = \frac{(z_\alpha + z_\beta)^2 v}{\Delta^2} = \frac{(1.645 + 1.645)^2 (.25)}{(.10)^2} = 270.6.$$

We would therefore need  $2 \times 271 = 542$  patients, an amount that is about four times larger than before.

Two other important features to be noted about these formulas are the crucial roles of  $\Delta$  and  $v$ . Since  $\Delta$  is always a value between 0 and 1, the value of  $\Delta^2$  is always smaller than  $\Delta$ . (For example, if  $\Delta = .3$ ,  $\Delta^2 = .09$ .) Furthermore, the smaller the value of  $\Delta$ , the smaller will be the value of  $\Delta^2$  and the larger will be the corresponding value of  $(1/\Delta^2)$  which is used as a factor in determining  $n$ . Thus, the smaller the difference for which we want to show 'statistical significance', the larger is the sample size that is required. In fact, if we want to prove the null hypothesis exactly, and to show that  $p_1$  and  $p_2$  are absolutely identical, we would need a sample of infinite size because  $\Delta = 0$ .

Since  $v$  appears in the numerator of the factors that are multiplied to calculate  $n$ , the size of  $n$  will decrease as  $v$  decreases. The value of  $v$ , being dependent on  $2\bar{p}(1 - \bar{p})$ , will be at a maximum when  $p = 50\%$  and will take minimum values near the polar extremes of 0% or 100%. Thus, if  $\bar{p}$  is close to 0% or to 100%,  $v$  will be small and  $n$  will be correspondingly

small. On the other hand, when  $\bar{p}$  is very close to a polar extreme, a large value for  $\Delta$  may be extremely difficult or unfeasible to obtain. The advantage of a very high or very low value for  $\bar{p}$  may thus be completely obliterated by the associated disadvantages of a very small value for  $\Delta$ .

#### D. The importance of $\beta$ error

Because scientific research is usually directed at showing that two entities are different, most investigators depend on statistical tests that provide values only for  $P_A$ . The values of  $P_B$  are generally omitted, either because the investigator is unaware of their existence or because he is not concerned about them. The absence of attention to the possibility of  $\beta$  error is equivalent to setting the value of  $z_\beta = 0$ . For this value of  $z_\beta$ , the one-sided  $P_B = .5$ ; and the 'power' of the test is  $1 - P_B$  or 50%. In other words, the investigator takes a 50-50 chance of committing the false negative error of incorrectly rejecting the alternative hypothesis.

Most investigators accept this risk with equanimity, since their main concern in the customary situation of 'significance' testing is with  $\alpha$  error—with a false positive conclusion. There are at least two major scientific circumstances, however, in which the role of  $\beta$  error becomes particularly important.

The first of these circumstances is a clinical trial in which we want to be sure of having a satisfactory chance of detecting a substantial  $\Delta$  when it exists. If we fail to find 'statistical significance' at the  $\alpha$  level, we might like to be reasonably confident about accepting the null hypothesis. This strategy is responsible for sample-size calculations that culminate in such phrases as "a 90% chance of finding a 20% difference at the .05 level". In this phrase, the associated statistical values are  $\Delta = .20$ ,  $\alpha = .05$  and (one-sided)  $\beta = .1$ .

The second (and perhaps more important) role of  $\beta$  error is in the situation where we want to show that two groups are similar rather than different. An example of such a situation is a clinical trial<sup>12, 15</sup> whose conclusion was that the quality of primary care provided by nurse practitioners is essentially equal to what is offered by physicians. Another example, which is an increasingly common situation in clinical pharmacology, occurs for tests of the bioequivalence of

two pharmaceutical preparations. In such circumstances, we assign a value of  $\Delta$  as the maximum permissible difference between the groups. If the observed difference is smaller than  $\Delta$ , we shall conclude that the groups are essentially equivalent. As noted earlier, the value that is chosen for  $\Delta$  and the magnitude of  $p_1$  will be as important as the choices of  $\alpha$  and  $\beta$  in determining sample size.

The high values of  $n$  that can emerge from these calculations will be a major problem in routine studies of bioavailability. In an example cited earlier, for  $\alpha$  and  $\beta$  both equal to .05 and for  $\Delta = 29\%$ , the size of  $n$  could range from 84 to 143, according to the values of  $p_1$ . Since sample sizes of this magnitude will usually be unfeasible, the values of  $\alpha$  and  $\beta$  may have to be made quite liberal. Thus, if  $\alpha$  (two-sided) is increased to 0.2 and if  $\beta$  (one-sided) is increased to 0.15, the size of  $n$  will vary as follows:

$\Delta$	5%	5%	20%	20%
$p_1$	5%	50%	5%	50%
$n$	254	728	38	57

These sample sizes, although smaller than before, are still substantially larger than the 6 to 10 patients whose data have been customarily examined for studies of bioavailability. If the bioavailability research is conducted in a "crossover" manner, in one group of patients rather than two groups, the paired arrangement of data will permit a further reduction in sample size. Nevertheless, if strict statistical standards become demanded for studies of bioequivalence, the problems of obtaining ample numbers of people for the tests may be so formidable that the studies will be impossible to conduct. Just as the old calculations of  $P_A$  for  $\alpha$  error alone made no provision for  $\beta$  error, the new calculations of sample size of  $\beta$  error alone may have to be done without consideration of  $\alpha$  error.

#### E. Caveats and abuses

A knowledge of  $\beta$  reasoning and the 'other side' of 'significance' can lead to prompt detection of a classical abuse in the way that statistical tests are often reported in medical literature. The  $\beta$  reasoning has also been applied to create new problems in the calculation of sample size.

1. *Conclusions when the null hypothesis is conceded.* In a routine statistical test of 'sig-

nificance', what conclusion do we draw if the  $P_A$  value is higher than  $\alpha$ ? The correct answer to this question is that such a high  $P$  value makes us concede, i.e., fail to reject, the null hypothesis. With this concession, we conclude that the observed difference is statistically not significant. The wrong answer to the question is that we accept the null hypothesis and conclude that the difference is insignificant.

The distinctions between *concede* and *accept* and between *not significant* and *insignificant* can be clinically illustrated by recalling the purpose of the sputum Pap smear as a diagnostic test. We order the test in search of a positive diagnosis of lung cancer. If the test is negative, however, we cannot conclude that lung cancer has been ruled out. We would merely concede that we have failed to show its presence. To accept the negative diagnosis that lung cancer is absent, i.e., to rule it out, we would want to check results from additional tests, such as the chest X-ray.

Consequently, in a simple test of 'statistical significance', a high  $P$  value is like the Scottish verdict of *not proved*. When  $P_A$  exceeds  $\alpha$ , we neither reject nor accept the null hypothesis. We concede it, or fail to reject it. Our conclusion must therefore be that the observed difference is *not significant*, rather than *insignificant*. To conclude that it is insignificant, we would have to accept the null hypothesis—a decision that would require additional evidence for the possibility of  $\beta$  error.

The previous example of satisfaction with care at two hospitals provided an illustration of the erroneous conclusions that can occur when  $P_A > \alpha$ . The investigator wanted to claim that the satisfaction was similar at the two hospitals, but we would not accept his claim because it had a 'power' of only 71%. In fact, if our original sample size was quadrupled, and if the proportion of successes remained the same, the resulting numbers would be 64/92 vs. 76/92 and the difference would be statistically significant at  $P < .05$  even though the observed  $\delta$  was only 13%.

Even when the two contrasted results seem quite similar, we still cannot conclude that their difference is *insignificant*. For example, suppose

we have found a success rate of 3/7 (43%) for treatment A and 4/9 (44%) for treatment B. This result seems unimpressive, but it could readily arise by chance if the true success values for treatments A and B were, respectively, 29% and 56%. Thus, if we exchanged one success and one failure in the patients comprising groups A and B, we would get success rates of 2/7 (29%) for A and 5/9 (56%) for B. This difference is impressive although not statistically significant.

One of the main abuses of tests of statistical significance occurs, therefore, when an investigator who gets a high  $P$  value, i.e.,  $P_A > \alpha$ , concludes that the observed difference between two groups is 'insignificant' and that the groups should be regarded as similar. If this type of reasoning were correct, we could always 'prove' that two treatments were identical, merely by using a small sample size for the study. Thus, if we put 3 patients in each group, a result as extreme as 0/3 (0%) successes for treatment A vs. 3/3 (100%) for treatment B could still not achieve 'statistical significance'. (The two-sided  $P$  value is .1.) From this failure to attain 'statistical significance', it would be absurd to conclude that the observed difference is insignificant and that the two treatments are equivalent. Nevertheless, such errors regularly appear in medical literature.

An analogous problem occurs when statistical tests are done to determine whether the act of randomization provided an equitable distribution of the patient groups before treatment began in a clinical trial. When good grounds exist for suspecting baseline inequalities, a high  $P_A$  value cannot alone be accepted as confirmation of their absence. The analysis is incomplete unless attention is also given to  $P_B$ . (A memorable example of such omissions occurred in analyses<sup>1, 17</sup> of the celebrated UGDP study of diabetes. When statistically significant differences were *not* found in certain analyses of baseline distinctions, the data analysts concluded that the baseline differences were insignificant, although no levels of  $\beta$ -error were cited.)

The point to be borne in mind is that an ordinary test of 'statistical significance' can be used only to reject the null hypothesis, not to accept

it. The test either shows or does not show a 'significant' difference. It cannot show an 'insignificant' difference. To draw the latter conclusion, we would need to know the other kind of P value for the possibility of  $\beta$  error.

2. *Problems in calculating sample size.* With the increasing performance of controlled clinical trials, many alternative strategies have been proposed for determining sample size. So many different proposals have been made, in fact, that contriving new ways to gauge sample size seems to have become a favorite indoor sport of statistical theoreticians. The alternative strategies include schemata based on sequential analysis, Bayesian conjectures, and various 'play-the-winner' techniques. Schneiderman<sup>14</sup> has provided a well-written summary of the state of the art in some of the statistical ideologies. For practical purposes, the material presented here has been based on the currently accepted "conventional wisdom".\*

Like many other statistical activities, a preoccupation with the mathematical tactics of determining sample size may often distract both statisticians and investigators from basic challenges that are the really fundamental issues in scientific research. In order to calculate a sample size, we often ignore these issues and assume that they have been taken care of. After the Neyman-Pearsonian, Bayesian, or other strategies have yielded a number in the sample size calculations, the clinical investigator may become awed by the precision of the number ('you will need exactly 984 patients') or flustered by its magnitude ('how the devil can I possibly get so many?'). As this number and negotiations about its reduction become the focus of attention, the clinician and statistician may forget that the basic scientific problems remain unresolved. Among them are the following:

a. *The univariate choice of an endpoint.* To determine  $p_1$ ,  $p_2$ , and  $\Delta$ , we must choose a

single variable whose outcome will be the "endpoint" in the research. This concentration on only one kind of outcome is contrary to every tenet of good clinical investigation, which calls for an appraisal of the multitude of variables that are involved in a patient's responses to treatment. Nevertheless, there currently exist no satisfactory biostatistical procedures for either choosing sample size in a multivariate manner or preparing a clinically effective composite of important multiple variables into a single univariate index.

b. *The focus on 'hard data'.* Since everything in the sample size calculations depends on the endpoint noted in a single variable, statisticians usually want to be sure that this endpoint is an item of 'hard data', such as *death*. Since changes in death rates are usually smaller than the changes that can occur in important 'soft data' variables, the result of the focus on hard data is to create a relatively small value of  $\Delta$ , which may lead to excessively large values for the calculated sample size. A more important consequence of the hard-data focus is that an important soft-data variable, such as vascular complications or quality of life, may become ignored in the early stages of biostatistical planning for the trial and may remain ignored (or poorly managed) thereafter. Because of this inattention to 'soft data', the most important clinical and human aspects of therapy—the associated risks, benefits, costs, joys, and sorrows of treatment—often become grossly neglected in the research<sup>16</sup>.

c. *The current uninformed choice of  $p_1$ .* Because good data are seldom available for 'historical controls', the choice of  $p_1$  (as an estimate of the outcome rate for the control group) becomes an act of guesswork that often turns out to be erroneous. If the error leads to a huge overestimate of sample size, the trial becomes excessively expensive.

d. *The future uninformed choice of  $p_1$ .* The estimate of a single value of  $p_1$  has no real clinical precision. What is usually needed, instead, is a series of  $p_1$  values—one for each of the cogent clinical strata<sup>2</sup> of patients subjected to therapy. If the data of large-scale randomized therapeutic trials are not analyzed with a cogent

\*For my education in these concepts and for other helpful comments on this text, I am indebted to several clinical and statistical colleagues: Donald Archibald, Robert Deupree, Michael Gent, Walter Spitzer, and Carolyn Wells. Their aid is gratefully acknowledged here, while they are also absolved of responsibility for the contents.

clinical stratification, however, the results of a current trial cannot provide a good estimate of  $p_1$  values for use in future trials. Today's expensive, unproductive therapeutic trial may thus be followed by tomorrow's.

*e. The arbitrary choices of  $\alpha$  and  $\beta$ .* Despite the elaborate reasoning that has been discussed for choosing  $\alpha$  and  $\beta$ , their values are seldom selected in the abstract intellectual manner described here. What often happens is that the statistician and investigator decide on the size of  $\Delta$ . The magnitude of the sample is then chosen to fit the two requirements (1) that the selected number of patients can actually be obtained for the trial and (2) that their recruitment and investigation can be funded. The values of  $\alpha$  and  $\beta$  are then adjusted to fit this number and a suitable mathematical rationale is then developed for presentation to the granting agency.

*f. The arbitrary choice of  $\Delta$ .* As the difference that indicates 'clinical significance', the magnitude of  $\Delta$  is often the most crucial issue in planning and evaluating the research. Despite this importance, the scope of  $\Delta$  is not only constrained by the univariate restrictions noted earlier, but it also gets chosen arbitrarily. Judgments about the proper size of  $\Delta$  have received almost no concentrated attention via symposia, workshops, or other conclaves of experts assembled to adjudicate matters of clinical importance. In the absence of established standards, the clinical investigator, on being badgered by the statistician to choose a  $\Delta$  so that sample size can be calculated, picks what seems like a reasonable value. This value is tossed into the formula, using  $z_\alpha$ ,  $z_\beta$ , etc. If the sample size that emerges is unfeasible,  $\Delta$  gets adjusted accordingly, and so do  $\alpha$  and  $\beta$ , until  $n$  comes out right.

In some brave new world of the future, when clinicians begin to insist that large-scale therapeutic trials be truly clinical investigations as well as elaborate exercises in mathematics, better solutions may be developed for these clinical and scientific problems. In the meantime, clinical investigators can take comfort in knowing about the panacea-like marvels offered by modern statistical methods for determining  $\alpha$ -error,  $\beta$ -error, and sample size. Even if we don't

know what we're doing and even if we can't specify it, repeat it, or make good clinical sense out of it, we can still calculate the required populational numbers and determine the probabilistic uncertainties going in both logical directions. Not since the days of alchemy have scientists been able to rely on such dazzling transmutations.

### References

1. Cornfield, J.: The University Group Diabetes Program. A further statistical analysis of the mortality findings, *J. A. M. A.* **217**:1676-1687, 1971.
2. Feinstein, A. R.: Clinical biostatistics. The purposes of prognostic stratification, *CLIN. PHARMACOL. THER.* **13**:285-297, 1972.
3. Feinstein, A. R., and Ramshaw, W. A.: A procedure for rapid mental calculation of the fourfold chi-square test, *J. Chron. Dis.* **25**:551-553, 1972.
4. Feinstein, A. R.: Clinical biostatistics. XXIII. The role of randomization in sampling, testing, allocation, and credulous idolatry (Part 2), *CLIN. PHARMACOL. THER.* **14**:898-915, 1973.
5. Feinstein, A. R.: Clinical biostatistics. XXXI. On the sensitivity, specificity, and discrimination of diagnostic tests, *CLIN. PHARMACOL. THER.* **17**:104-116, 1975.
6. Feinstein, A. R.: Clinical biostatistics. XXXII. Biologic dependency, 'hypothesis testing', unilateral probabilities, and other issues in scientific direction vs. statistical duplexity, *CLIN. PHARMACOL. THER.* **17**:499-513, 1975.
7. Fleiss, J. L.: Statistical methods for rates and proportions, New York, 1973, John Wiley & Sons, Inc.
8. Halperin, M., Rogot, E., Gurian, J., and Ederer, F.: Sample sizes for medical trials with special reference to long-term therapy, *J. Chron. Dis.* **21**:13-24, 1968.
9. Kramer, M., and Greenhouse, S. W.: Determination of sample size and selection of cases. *in* Cole, J. O., and Gerard, R. W., editors: Psychopharmacology: Problems in evaluation, National Academy of Sciences, National Research Council, 1959, pp. 356-371.
10. Neyman, J., and Pearson, E. S.: On the use and interpretation of certain test criteria for the purposes of statistical inference, *Biometrika* **20A**:175 and 263, 1928.
11. Pasternack, B. S.: Sample sizes for clinical trials designed for patient accrual by cohorts, *J. Chron. Dis.* **25**:673-681, 1972.
12. Sackett, D. L., Spitzer, W. O., Gent, M., and Roberts, R. S.: The Burlington randomized trial of the nurse practitioner: Health outcomes

- of patients, *Ann. Intern. Med.* **80**:137-142, 1974.
13. Schlesselman, J. J.: Planning a longitudinal study. I. Sample size determination, *J. Chron. Dis.* **26**:535-560, 1973.
  14. Schneiderman, M. A.: The proper size of a clinical trial: "Grandma's strudel" method, *J. New Drugs* **4**:3-11, 1964.
  15. Spitzer, W. O., Sackett, D. L., Sibley, J. C., Roberts, R. S., Gent, M., Kergin, D. J., Hackett, B. C., and Olynich, A.: The Burlington randomized trial of the nurse practitioner, *N. Engl. J. Med.* **290**:251-256, 1974.
  16. Spitzer, W. O., Feinstein, A. R., and Sackett, D. L.: What is a health care trial? *J. A. M. A.* **233**:161-163, 1975.
  17. University Group Diabetes Program. A study of the effects of hypoglycemic agents on vascular complications in patients with adult-onset diabetes. I. Design, methods and baseline results; and II. Mortality results, *Diabetes* **19** (Suppl. 2):747-830, 1970.