

## Invited Opinion

# Statistical Significance and Statistical Power in Hypothesis Testing

Richard L. Lieber

*Division of Orthopaedics and Rehabilitation, Veterans Administration Medical Center and University of California, San Diego, CA, U.S.A.*

---

**Summary:** Experimental design requires estimation of the sample size required to produce a meaningful conclusion. Often, experimental results are performed with sample sizes which are inappropriate to adequately support the conclusions made. In this paper, two factors which are involved in sample size estimation are detailed—namely type I ( $\alpha$ ) and type II ( $\beta$ ) error. Type I error can be considered a “false positive” result while type II error can be considered a “false negative” result. Obviously, both types of error should be avoided. The choice of values for  $\alpha$  and  $\beta$  is based on an investigator’s understanding of the experimental system, not on arbitrary statistical rules. Examples relating to the choice of  $\alpha$  and  $\beta$  are presented, along with a series of suggestions for use in experimental design. **Key Words:** Experimental Design—Type I error—Type II error—Sample size—Statistics.

---

Statistical methods are often used in scientific settings to determine whether experimental results are true in general, i.e., whether results obtained from an experimental sample can be generalized to the population. We are familiar with the phrase “the results were significant ( $p < 0.05$ ) . . .” in the scientific literature. For some experimentalists, if the results are not “significant,” they are not considered worth reporting. To others, statistical analysis seems to be a burdensome numbers game that is not necessarily related to scientific quality. In fact, it is incorrect either to “live and die” by the  $p$  value or to avoid the use of statistical analysis because of potential abuse. With the advent of modern statistical software packages (1),  $p$  values are ob-

tained automatically when one performs a “canned” analysis. Unfortunately, the ease of the analysis is also its danger in that the experimenter is not required to understand the meaning of the results. The purpose of this review is to describe the meaning of the  $p$  value, to contrast the  $p$  value with statistical power, and to describe, by example, the relationship between the  $p$  value and the two types of statistical error. [It should be noted that others have debated whether the  $p$  value should even be used at all in the scientific literature (2,3,6,7)!]

## HYPOTHESIS TESTING

After performing an experiment, one is often interested in determining whether a “treatment” has had an effect. For example, one might wish to determine whether isometric exercise has a strengthening effect on muscles. Typically, this type of experiment is performed with an experimental group (a group receiving isometric strength training treat-

---

Received February 28, 1989; accepted September 7, 1989.

Address correspondence and reprint requests to Dr. R. L. Lieber at Division of Orthopaedics and Rehabilitation (V-151), V.A. Medical Center and U.C. San Diego School of Medicine, 3350 La Jolla Village Drive, San Diego, CA 92161, U.S.A.

ment) and a control group (a group not receiving strength training). At the conclusion of the experiment, the average strength of the experimental and control groups may be compared statistically to determine whether strength training had an effect on muscle strength. The statistical analysis and the resulting "*p*" value refers directly to the veracity of what is called "the null hypothesis."

### THE NULL HYPOTHESIS

The null hypothesis states that there is no (null) effect of treatment. In the present example, the null hypothesis states that training has no effect on strength, or that the strength of the experimental group is equal to the strength of the control group. In the present example, we have a number of choices related to the null hypothesis (Table 1). Obviously, the null hypothesis can be either true or false. Additionally, we can choose to accept or reject the null hypothesis. This results in four potential decisions, two of which are correct and two of which are incorrect (Table 1). Suppose, for example, that the null hypothesis is true, i.e., there is no difference in strength between experimental and control groups. If we accept the null hypothesis, we have made the correct decision. If we reject the null hypothesis, we have made an incorrect decision. We have committed what is known in statistics as type I error, rejecting a true null hypothesis. This can be viewed in more clinical terms as a "false positive" (Table 2). Thus, in our experiment, type I error concludes that there is a significant effect of strength training when, in fact, there is not. The alternate possibility is that the null hypothesis is false, i.e., isometric training has a significant effect on muscle strength. If we reject the null hypothesis, we have again made the correct decision. If we accept the null hypothesis, we have made an incorrect decision, committing what is known as type II error, accepting a false null hypothesis. This can be viewed in clinical terms as a "false negative" (Ta-

TABLE 1. Statistical errors related to the null hypothesis

	Null hypothesis	
	Accepted	Rejected
Null hypothesis		
True	Correct decision	Type I error
False	Type II error	Correct decision

TABLE 2. Interpretation and control of statistical error

Condition	Greek symbol	Meaning	Controlled using
Type I error	$\alpha$	False positive	Significance level
Type II error	$\beta$	False negative	Statistical power

ble 2). We conclude that there is no effect of training when, in fact, there is.

Of course, we would like to commit as few errors as possible. We prefer not to commit either type I or type II error, but it should be clearly pointed out that the *p* value is directly related only to type I error, i.e., the *p* value is simply the probability (denoted  $\alpha$ ) of committing type I error in a given experiment (Table 2). When we state that "the results are significant ( $p < 0.05$ ) . . .," we are saying that we are potentially committing type I error less than 5% of the time. The problem with this automatic use of  $p < 0.05$  as the level for statistical significance is that many times (especially in clinical situations) it is not acceptable to commit type I error 5% of the time, whereas in other cases we might be willing to commit type I error a greater percentage of the time (see below). The significance level should actually be determined based on its meaning in the context of the experiment performed.

In order to decrease the probability of committing type II error (denoted  $\beta$ ), we must design our experiment with sufficient statistical power.

### STATISTICAL POWER AND THE CHOICE OF SAMPLE SIZE

While we are familiar with setting limits for type I error (by choosing a critical *p* value), we are not as familiar with limiting type II error. However, as discussed below, controlling type II error can be equally or more important than type I error. Many of us have observed presentations where a small sample size was used (e.g.,  $n = 3$ ), statistical analysis was performed, and a *p* value was obtained that was greater than 0.05. The speaker concluded that the treatment had no effect. Immediately, a protestor stated that the sample size was not large enough to demonstrate the difference. We may also have observed the situation where an individual performed an experiment with, say, 10 individuals per sample, obtained a *p* value of around 0.07, and was then encouraged to add a few more individuals

to the sample in order to achieve statistical significance! In another setting, we may have observed a scientist performing an experiment with a small sample size comparing a "new" technique to some "standard" technique. The scientist, based on a high  $p$  value, concluded that there was no significant difference between the "new" and "standard" methods and that the new method should be used. All of these situations can arise when one has not considered statistical power in the experimental design. Interestingly, a review of 71 "negative" randomized clinical trials (2) concluded that over one-half of the "negative" results were simply a result of a lack of sufficient statistical power.

Statistical power is simply  $1 - \beta$ , the logical negative of type II error. If type II error is analogous to a false negative, i.e., accepting a false null hypothesis, then power is the probability of not committing a false negative. In other words, we want to be sure that if we obtain a  $p$  value greater than 0.05, we are not incorrectly accepting a false null hypothesis. We want to be sure we are not committing type II error. In the example stated above, we may wish to design the experiment with a power of 95%. In that case, we would be 95% sure that if isometric strength training had an effect (the null hypothesis were false), we would not falsely conclude that it did not.

### DESIGNING AN EXPERIMENT OF A GIVEN POWER

Several methods (graphs, tables, equations) have been developed that allow the experimenter to set the significance level (the critical  $p$  value), the statistical power, and then to determine the sample size required to achieve that design (4,5). An example of an equation used in such a calculation is shown below:

$$n = 2 \cdot \left( \frac{\sigma}{\delta} \right)^2 \cdot (t_{\alpha, \nu} + t_{2(1-P), \nu})^2 \quad (1)$$

where  $n$  = the sample size,  $\sigma$  = the population standard deviation,  $\delta$  = the difference that is desired to detect,  $\alpha$  = significance level (probability of type I error),  $\nu$  = the degrees of freedom,  $t_{\alpha, \nu}$  = the  $t$  value corresponding to  $\alpha$  and  $\nu$ , and  $P$  = the desired statistical power. Note that as the variability of the population increases (i.e., as  $\sigma$  increases), the number of observations required also increases. This is one reason it is important to minimize all

controllable sources of error in an experiment. Also note that as the magnitude of the treatment effect decreases (i.e., as  $\delta$  decreases), the sample size also increases. This is another way of saying that if an experimenter wishes to demonstrate a very small difference (a very small treatment effect), a large sample size is required. Conversely, if treatment effects are very large, a relatively small sample may be adequate. It should be mentioned that the actual values of  $\sigma$  and  $\delta$  need not be known. Only their ratio need be estimated. In these terms, demonstrating a treatment effect that is approximately the size of the standard deviation requires a much larger sample than demonstrating a treatment effect that is several times the standard deviation.

The procedure used to calculate sample size using Eq. (1) is an iterative one. We begin with some information about the experiment, make a first guess at sample size, calculate the expected sample size, make a new, better guess at sample size, and calculate the new expected sample size.

An example of such a process is taken from a study by Wenger et al., who studied the treatment of flexible flatfoot in children (8). Before performing the study, the investigators wished to determine the number of subjects required to determine whether three different treatment methods were effective. The experimental design included one control group and three experimental groups. The investigators measured radiographic angles of the foot before and after treatment. Based on their previous experience with radiographic angle measurements on other children, they knew that the standard deviation of the general population ( $\sigma$ ) was approximately  $4.5^\circ$ . In their clinical judgment, they considered an improvement ( $\delta$ ) in the radiographic angle of  $4^\circ$  to represent a significant treatment effect. The null hypothesis in this experiment was that treatment had no effect on radiographic angle. Type I error would conclude that the treatment had an effect when, in fact, it did not. Type II error would conclude that treatment had no effect when, in fact, it did. The investigators decided to accept a type I error frequency of 5% (a critical  $p$  value, or significance level,  $\alpha$ , of 0.05) and wished to make the power of the statistical test 90% ( $P = 0.9$ ,  $\beta = 0.1$ ).

In order to calculate the required sample size given this problem, we first make a rough guess at sample size, say  $n = 10$ . We must then calculate the degrees of freedom, using the following equation:

$$\eta = a(n - 1) \quad (2)$$

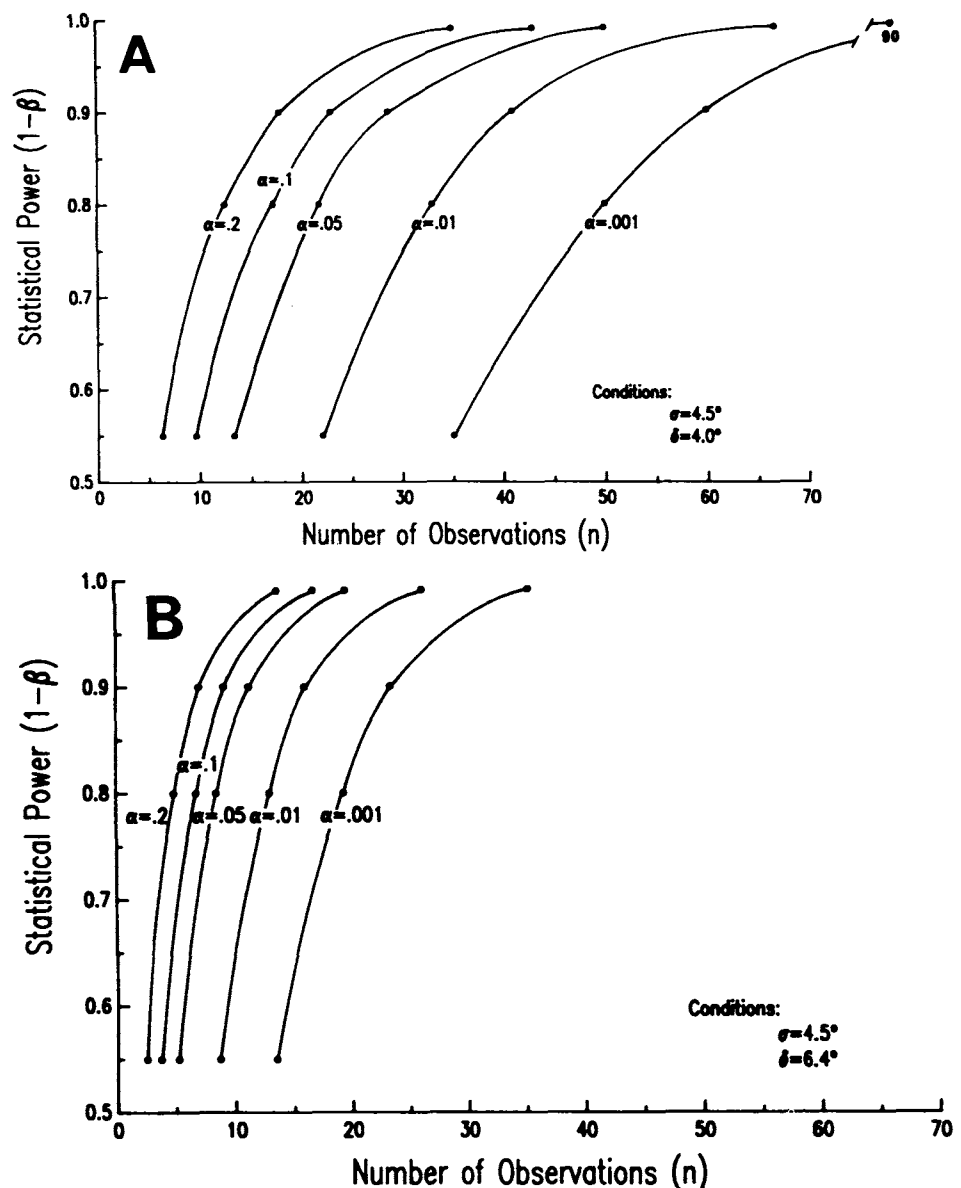
where  $\nu$  = degrees of freedom,  $a$  = the number of groups, and  $n$  = the number of independent observations per group.

Because we have four groups, the degrees of freedom is  $4(10-1) = 36$ . We obtain from a statistical table (4) the  $t$  value corresponding to a significance level of 0.05, 36 degrees of freedom and a significance level of 0.2, 36 degrees of freedom. The corresponding  $t$  values are 2.028 and 1.308, respectively. We enter these values into Eq. (1) and solve for  $n$ , obtaining  $n = 28.3$ . Based on this calculation, we now refine our sample size estimate to  $n = 30$ .

We now repeat the calculations using  $n = 30$ . The degrees of freedom is 4 (30 - 1), or 116. The appropriate  $t$  values are 1.980 and 1.289, respectively. We recalculate the sample size and  $n = 27.1$ . Thus, as we refined our guess, sample size converged on a particular number. We would probably decide to perform the experiment with 30 individuals per group. This may require actually entering about 35 individuals per group to allow for attrition.

A graph of the relationship between statistical power, significance level, and sample size is shown in Fig. 1. Note that as either power increases or

**FIG. 1.** Graphical relationship between statistical power, significance level, and sample size using Eq. (1) **(A)** Calculations for  $\sigma = 4.5^\circ$  and  $\delta = 4.0^\circ$  to represent the case where the anticipated treatment effect is relatively small. **(B)** Calculations for  $\sigma = 4.5^\circ$  and  $\delta = 6.4^\circ$  to represent the case where the anticipated treatment effect is relatively large. Note that, in general, for a large treatment effect, the required sample size is smaller.



significance level decreases, sample size increases (Figs. 1A and 1B). Note also that if a relatively small treatment effect is anticipated ( $\delta = 4^\circ$ , Fig. 1A), a larger sample size is required than if the anticipated treatment effect is large ( $\delta = 6.4^\circ$ , Fig. 1B). For example, suppose that the anticipated treatment magnitude were  $4^\circ$  (Fig. 1A) and the desired statistical power was 90%. Under these conditions, a sample size of 29 would be required to achieve a significance level of  $\alpha = 0.05$ , whereas a sample size of 41 would be required to achieve a significance level of  $\alpha = 0.01$ .

In summary, in designing this experiment, we specified the type I error, the acceptable probability that we will commit a false positive. We also established the type II error by specifying the power. We then computed sample size, given the experimental variability and our anticipated treatment effect. Having specified both type I and type II error, interpretation of the data is straightforward. If our  $p$  value exceeds 0.05, we conclude that the treatment has no effect. We can be sure that if it is greater than 0.05, it is so, not because we have too few samples, but because the null hypothesis is indeed false.

### NONSTANDARD $p$ VALUES

A survey of the scientific literature, especially the literature related to biology and medicine, reveals that an overwhelming majority of investigators set the critical  $p$  value to 0.05. It should be obvious based on the previous discussion that there is nothing "magic" about a  $p$  value of 0.05. The  $p$  value of 0.05 simply indicates that we are willing to commit type I error 5% of the time. However, there may be situations where the investigator is not willing to commit type I error 5% of the time or even 1% of the time. In such cases, the critical  $p$  value should be adjusted accordingly.

An understanding of the basis for selection of a critical  $p$  value and statistical power is especially important in clinical science. For example, in an experiment that attempts to demonstrate a significant improvement in bone strength using a particular surgical procedure, if the critical  $p$  value is 0.05, the investigator is willing to conclude incorrectly 5% of the time that the surgical procedure has an effect even if it actually has no effect. The surgical procedure represents a risk to the patient and it also represents expense. In such a case, the investigator

may only be willing to commit type I error 1% of the time or a fraction of a percent of the time. In such a case, a critical  $p$  value of 0.05 may be much too high.

It should also be noted that at times, type II error may be more important to an investigator than type I error. For example, suppose that an experimental drug were administered to treat a patient's blood pressure. In this case, type I error would indicate that the drug had an effect when in fact it did not. The detriment to the patient is that they would take a drug that had no effect. While this could represent an expense, it may not represent a medical or scientific problem (assuming the drug had no side effect). However, suppose type II error were committed in the same study. Type II error would indicate that the drug had no effect when in fact it had an effect. In this case, an effective drug would be withheld from the patient, which could represent a large problem. It may be that in this example, the power of the test should be 99.9%, while the critical  $p$  value should only be 0.1. The interpretation of the meaning of the  $p$  value is therefore paramount in selecting its value and in guarding against cookbook application of statistical methods.

### SUMMARY

The results of parametric statistical methods (e.g.,  $t$  tests, analysis of variance, regression) refer implicitly to the null hypothesis of the parameters tested. It is critical for the experimenter to determine, before the experiment is performed, the acceptable levels of type I and type II error. This can only be done if the individual understands the meaning of these errors in the context of his/her experiment. Type I error is controlled by setting the significance level while type II error is controlled by designing the experiment with sufficient statistical power. The following suggestions serve as a guide in this aspect of experimental design: (a) Write out in words the null hypothesis for your experiment. (b) Write out in words the meaning of type I error in your experiment. Select the significance level in light of the meaning of type I error in your experiment. (c) Write out in words the meaning of type II error in your experiment. Select the intended power of your experiment in light of the meaning of type II error. (d) Estimate your population SD by obtaining pilot data. If pilot data are unavailable, estimate the

SD based on similar experiments that you have performed or, if you have no data, use appropriate literature values. Insure that experimental variability is minimized so that the SD represents the actual population variability (not experimental error) enabling you to use as few samples as possible. (e) Determine the magnitude of the anticipated treatment effect and use Eq. (1) to estimate iteratively the required sample size. Alternatively, use a statistical table or graphical means to estimate sample size.

**Acknowledgment:** The author would like to acknowledge the faculty, residents, and graduate students of the UCSD School of Medicine for their participation in the statistics course from which these reviews were taken. This work was supported by the Veterans Administration and NIH Grants AR35192 and AR40050.

## REFERENCES

1. Dixon WJ: *BMDP Statistical Software*, Los Angeles, University of California Press, 1983
2. Freiman JA, Chalmers TC, Smith H Jr, Kuebler RR: The importance of beta, the type II error and sample size in the design and interpretation of the randomized control trial. Survey of 71 "negative" trials. *N Engl J Med* 299:690-695, 1978
3. Poole C: Beyond the confidence interval. *Am J Publ Health* 77:195-199, 1987
4. Rohlf FJ, Sokal RR: *Statistical Tables*, 2nd edition, San Francisco, W.H. Freeman and Company, 1981
5. Sokal RR, Rohlf FJ: *Biometry*, 2nd edition, San Francisco, W.H. Freeman and Company, 1981
6. Thompson WD: Statistical criteria in the interpretation of epidemiologic data. *Am J Publ Health* 77:191-194, 1987
7. Thompson WD: On the comparison of effects. *Am J Publ Health* 77:491-493, 1987
8. Wenger DR, Maulden D, Speck G, Morgan D, Lieber RL: Effect of corrective shoes and inserts on flexible flat foot in children—a prospective randomized trial. *J Bone Joint Surg (Am)* 71:800-810, 1989