Journal of Orthopaedic & Sports Physical Therapy
2001;31(6):307–315

Understanding Statistical Power

Barbara J. Norton, PT, PhD[1]
Michael J Strube, PhD[2]

This article provides an introduction to power analysis so that readers have a basis for understanding the importance of statistical power when planning research and interpreting the results. A simple hypothetical study is used as the context for discussion. The concepts of false findings and missed findings are introduced as a way of thinking about type I and type II errors. The primary factors that affect power are described and examples are provided. Finally, examples are presented to demonstrate 2 uses of power analysis, 1 for prospectively estimating the sample size needed to insure finding effects of a known magnitude in a study and 1 for retrospectively estimating power to gauge the likelihood that an effect was missed. *J Orthop Sports Phys Ther 2001;31:307–315.*

Key Words: *effect size, power analysis*

Readers of the *Journal* may be accustomed to seeing authors report a P value. If $P < 0.05$, then the authors consider their results *statistically significant* and conclude that their research hypothesis is likely to be true. Conversely, when a result is *not* statistically significant, the authors may conclude that their research hypothesis is likely to be false. Should we agree with them? How do we know if the research hypothesis is true or if it is false?

HYPOTHETICAL STUDY

Before answering these questions, we will present an example of a research hypothesis and a simple study based on that hypothesis. This example will provide context for the points that we will make later. Although we will limit our discussion to the simplest of designs, our comments extend to more complex designs as well (for more extensive treatment of complex designs, see Cohen[2] and Cohen & Cohen[3]).

Our research hypothesis is that there will be a difference in range of motion (ROM) at the knee depending upon whether subjects receive contract-relax exercises or passive stretching exercises. In conducting our hypothetical study, we would begin by randomly selecting a subset of subjects from a predefined population of individuals presumed to have limited ROM. Ideally, this would be a random sample so that the results of our study would generalize to other people with similar ROM limitations. Then we would randomly assign each subject to 1of 2 treatment groups. Randomly assigning subjects to treatment groups is necessary in order to make the groups similar before the study begins; without random assignment (eg, letting subjects pick their own groups), there is the possibility of bias, that is, the groups could be different before we begin treatment. In our study, subjects in treatment group 1 would receive contract-relax exercises and subjects in treatment group 2 would receive passive stretching exercises.

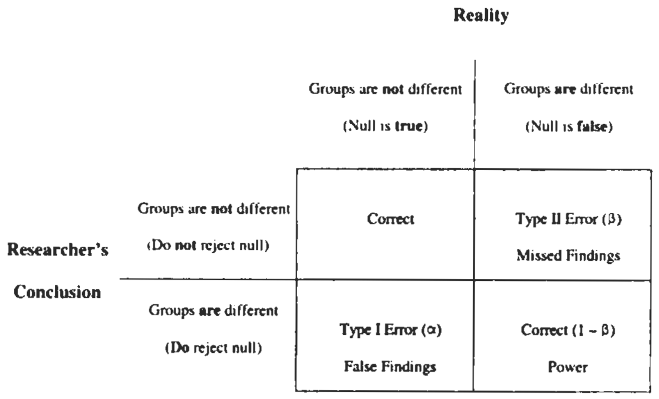
After the exercise program was completed, we would measure the knee ROM of each subject, calculate the average and standard deviation of the ROM values for each group, and calculate the value of a statistic called the t ratio. The t ratio provides a way of gauging how large the mean difference between the groups is relative to the variability within the groups. Appendix 1 summarizes the t test in the form of a signal-to-noise ratio. Appendix 2 defines statistical significance.

Missed Findings?

Now consider the question of whether we can believe a result that is *not* statistically significant. How do we know that the researchers did not *miss* finding support for their research hypothesis, which is actually true, because the signal was weak relative to the

SPECIAL REPORT

[1] *Program in Physical Therapy, Washington University, St. Louis, Mo.*
[2] *Department of Psychology, Washington University, St. Louis, Mo.*
Send correspondence to Barbara J. Norton, Program in Physical Therapy, Washington University, 4444 Forest Park Boulevard, Box 8502, St. Louis, MO 63108-2212. E-mail: nortonb@msnotes.wustl.edu



		Reality	
		Groups are **not** different (Null is **true**)	Groups **are** different (Null is **false**)
Researcher's Conclusion	Groups are **not** different (Do **not** reject null)	Correct	Type II Error (β) Missed Findings
	Groups **are** different (Do reject null)	Type I Error (α) False Findings	Correct (1 - β) Power

FIGURE 1. Schematic representation of conditions associated with correct and incorrect conclusions.

noise? Our question can be answered by turning to the concept of *statistical power.* One simple definition of the power of a statistical test is the probability that the test will yield a statistically significant result when the research hypothesis, in reality, is true. In other words, power is the ability to detect a difference in knee ROM between treatment groups when a difference really exists. Stated in terms of the null hypothesis, power is the probability of correctly rejecting a false null hypothesis.

Examining Types of Errors

A common way to represent the possible outcomes of research associated with the results of statistical tests under different sets of conditions is provided in Figure 1. The term *reality* refers to the actual state of affairs in the population regarding the effect of the treatment being studied. The term *researcher's conclusion* refers to the researcher's inference about the population, based on the results of studying a small subset, or sample, of the population. In Figure 1, the same possibilities are listed for both *reality* and the *researcher's conclusion.* In the case of our hypothetical study, the possibilities for *reality* and for the *researcher's conclusion* regarding the effect of the treatments are not different, meaning that contract-relax exercises and passive stretching exercises are equally effective, and different, meaning that one type of exercise is more effective than the other.

The key to understanding type I and type II errors is to examine the consequence of each combination of conditions, as represented by the 4 shaded cells of Figure 1. For example, the top left cell represents the outcome for the combination whereby the researcher concludes that the 2 treatments are not different and, in fact, they are not different. Because the 2 conditions are the same, the researcher's conclusion is considered correct. Note that for the lower right cell, the conditions for both *reality* and the *researcher's conclusion* are likewise the same, that is, the researcher concludes that the treatments are different and, in fact, the treatments really are different. Once again, the researcher's conclusion is considered correct because it is in agreement with reality.

This is not true for the 2 remaining cells because, in both cases, the researcher's conclusion deviates from reality. Figure 1 indicates that a type I error occurs when the researcher erroneously concludes that there *is* a difference between treatments, and a type II error occurs when the researcher erroneously concludes that there is *not* a difference between treatments. No one actually knows what is true in terms of the *reality* part of Figure 1. If we did, we would not need to conduct the study. Instead, we make informed judgments about *reality* that are based upon the research study. We try to design this judgment task in such a way that we keep the probability of making mistakes acceptably low. The symbols α (alpha) and β (beta) represent the probability of making the respective types of error. The α level is also known as the criterion of acceptability and is set by the researcher to protect against a type I error.

POWER

One definition for statistical power is the probability of *not* making a type II error. Based upon Figure 1, we know that the probability of making a type II error is β. We also know that the researcher has only 2 possible choices for a conclusion: groups are **not** different or groups **are** different. Therefore, the total probability for the 2 choices is 1. Accordingly, if the probability of making a type II error is β, then the probability of *not* making a type II error is $1 - \beta$. Thus, the expression $1 - \beta$ may be considered a mathematical definition of power. Having considered the elements of Figure 1 in some detail, a second definition for power may be easier to understand: power is the probability that the test will correctly yield a statistically significant result (ie, we reject the null hypothesis) when the null hypothesis, in reality, is false.

Recall that our research hypothesis is posed as a difference between the 2 types of stretching exercises and then focus on the bottom right cell of Figure 1. Note that the combination of conditions represented by the cell is groups **are** different for both *reality* and *researcher's conclusion.* A difference in reality is consistent with the null hypothesis being false, and the researcher's conclusion of groups **are** different would have been based upon a statistically significant test of differences between the groups who received the 2 different types of stretching exercises (ie, rejection of the null hypothesis). The probability associated with the bottom right cell is $1 - \beta$, or power. Power represents the probability of making a correct decision, and it is advisable to optimize the amount of power in a study.

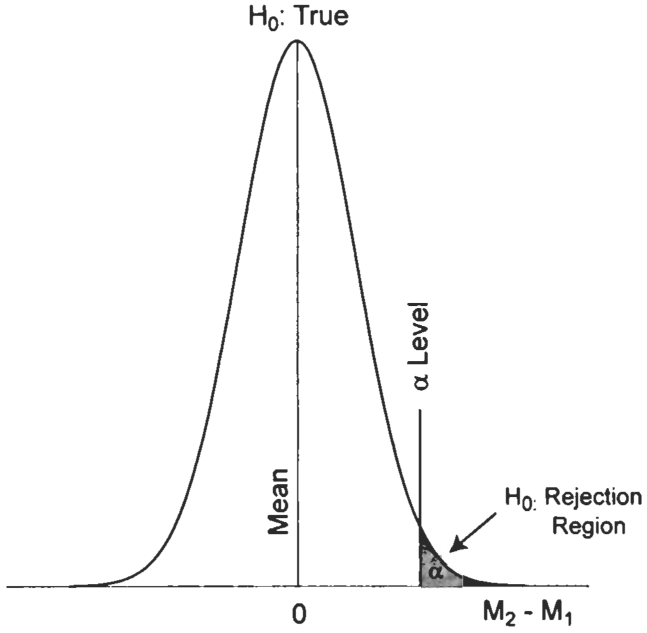


FIGURE 2. Sampling distribution under the assumption of a true null hypothesis (H_0). M_1 and M_2 represent group means. Alpha (α) level represents the criterion of acceptability and α is the area associated with a type I error.

FACTORS AFFECTING POWER

The 4 primary factors that affect the power of a statistical test are α level, difference between group means, variability among subjects, and sample size. Figures 2–5 aid our explanation of the relationship between power and each of these 4 factors. If we measured ROM on everyone in our stretching exercise study, we could plot the distribution of ROM scores, and the shape of the curve should look something like Figure 2. In this case, however, Figure 2 represents a special type of distribution known as a *sampling distribution* of mean differences under the assumption that the null hypothesis is true (ie, groups are *not* different in reality).

We can begin to understand the difference between a distribution of scores and a sampling distribution by thinking about the source of the data for each. As we just noted, the data for constructing a distribution of scores can be obtained by measuring ROM for every subject in the study. What about the sampling distribution of differences between means? Conceptually, we would randomly select *N* subjects from a population, randomly assign each subject to 1 of 2 groups, measure ROM on every subject, calculate the mean for each group, and then find the difference between the means. The value of the difference between the 2 means would represent 1 data point for constructing our sampling distribution. We would get additional data by repeating the process of sampling, testing, and calculating the difference between the means. After we repeated the process many times, we would have enough data to construct a distribution of differences between the means of 2 groups, as in Figure 2.

Having described a conceptual process for generating a sampling distribution, we will focus on some key points regarding the process and the distribution. One of the most important points to note about the process we described for generating the sampling distribution is that we would *not* have provided any type of treatment to either group of subjects; all we would have done was randomize and measure. Consequently, we would have no explanation for any differences between the means of the groups other than chance or variation that would ordinarily occur. In some instances, one group's scores would have exceeded the other group's scores, but it would have been just a matter of chance and not due to anything that was systematically done to one group and not to the other. On average, however, there should be no difference between the groups. The mean of the distribution should be zero and the peak of the distribution should be at the mean because there would have been no difference for most of the samples. The fact that Figure 2 represents a bell-shaped curve and not just a vertical line at zero implies that, in some cases, there *were* differences between the means of the 2 groups and that the size of the differences ranged from relatively small to relatively large. Thus, the sampling distribution represents the expected distribution of differences between means when the null hypothesis is true, or as is sometimes stated, the differences that would be expected to occur due to chance alone. The shape of the distribution tells us that as the mean differences deviate from the expected difference of zero, they become more rare. That matches our intuition. This distribution is useful because it tells us what to expect in the way of "noise." For a signal to be detected (ie, for us to see the effect of our treatment clearly), it must stand out against the noise or the variation in differences between means that occurs anyway. The idea of viewing a statistical test as a signal-to-noise ratio is summarized in Appendix 1.

Alpha (α) Level

In a sampling distribution, we know the magnitude of differences between means that can occur even if the null hypothesis (H_0) is true. We can use this information to decide whether the difference between the means in our study is unusually large under the assumption that the null hypothesis is true. If our obtained difference were rare under the null hypothesis—if it is quite distant from the mean of the sampling distribution—then we would reject the null hypothesis. But what do we mean by "rare?" In this case, rare is defined by the α level. Note in Figure 2 that we have included a vertical line, labeled α level.

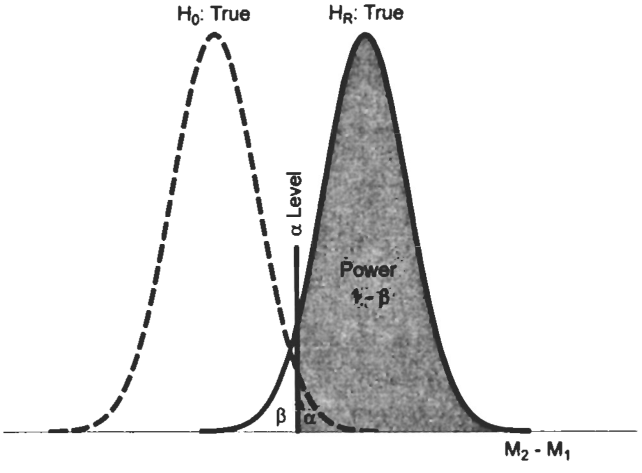


FIGURE 3. Pair of sampling distributions. The distribution on the left is based on the assumption that the null hypothesis (H$_0$) is true in reality; the distribution on the right is based on the assumption that the research hypothesis (H$_R$) is true in reality. Alpha (α) level represents the criterion of acceptability and α is the area associated with a type I error. β represents the area associated with a type II error.

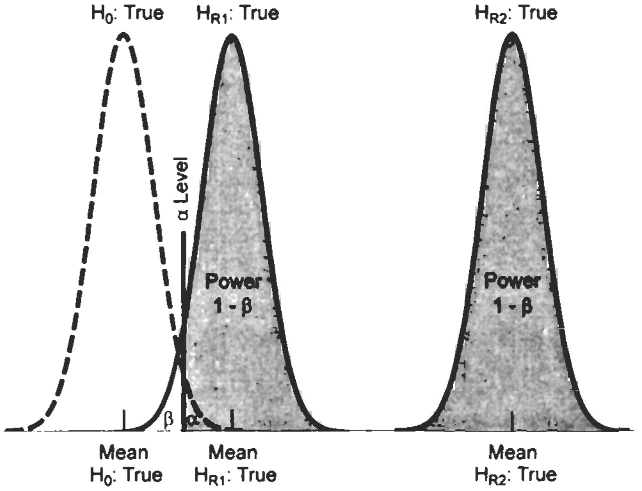


FIGURE 4. Set of 3 sampling distributions. The distribution on the left is based upon the assumption that the null hypothesis (H$_0$) is true in reality; the 2 on the right are based upon the assumption that the research hypotheses (H$_{R1}$ and H$_{R2}$) are true in reality. Alpha (α) level represents the criterion of acceptability and α is the area associated with a type I error.

This is the same α level described in Appendix 2 that protects against type I errors. Using Figure 2, we can see that the concept of $P = 0.05$ is related to the fact that 5% of the area under the curve is to the right of the vertical line (also known as the rejection region). Recalling that the magnitude of differences between the means of 2 groups ($M_2 - M_1$) is represented on the x-axis, we know that if the difference between 2 means is larger than the difference coincident with the α level, then there is a 5% or less chance that the difference was simply due to chance. Therefore, we would consider the result statistically significant (something rare or out of the ordinary). In other words, if our obtained result would have occurred under the null hypothesis with a probability of 0.05 or less, then we call it rare enough to reject the null hypothesis. By rejecting the null hypothesis because our result falls in the rejection region, we implicitly find the alternative, our research hypothesis, to be more plausible.

The α level is one of the factors that affects power (Figure 3). Once again, we have included the sampling distribution of the difference between the means under the assumption that the *null* hypothesis (H_0) is true; it is represented by the curve on the left of Figure 3 (dashed line). We have added a second sampling distribution on the right (solid line) to represent the differences between the means under the assumption that the *research* hypothesis (H_R) is true. The second distribution would have been constructed in the same way that we previously described for the null hypothesis; however, the mean of the distribution would reflect the fact that the treatments do produce differences in ROM. Furthermore, the distribution would be bell-shaped because every study would not produce precisely the same difference. Sometimes the mean difference in knee ROM between groups would be larger, sometimes smaller, and the more the difference deviates from the expected difference, the less likely or frequent the difference becomes.

Look carefully at Figure 3 and notice 3 areas of interest delineated by the vertical line: α, β, and power (1 − β). It is now possible to appreciate the effect of the α level on power by thinking about what changes if the vertically oriented α level line is moved horizontally. If the line is moved to the left, then the area labeled α becomes larger, the area labeled β becomes smaller, and, most importantly, the area labeled 1 − β becomes larger. The net effect is an *increase* in power. Conversely, if the α level line is moved to the right, α becomes smaller, β becomes larger, and the 1 − β area becomes smaller. The net effect is a *decrease* in power. Although the α level affects power, researchers typically do not set the α level any higher than 0.05 because they do not want to increase power at the expense of increasing the likelihood of a type I error and contributing *false findings* to the research literature (Figure 1).

Difference Between Group Means

The second factor that affects power is the difference between the means of the 2 groups. In studies like our hypothetical example, the magnitude of the difference between the means of the 2 groups at the end of the study is, in fact, the "signal" of interest in the study. We will use Figure 4 to help explain how this affects power. Figure 4 is essentially the same as Figure 3, except that we have added a third distribution. On the left, we still have the original sampling distribution (dashed line) of the means under the assumption that the null hypothesis (H_0) is true, but on the right we now have 2 different sampling distributions (solid lines) of the means under the assumption that the research hypothesis (H_R) is true. Each represents a different assumption about the difference between our 2 treatments. In the distribution on the far right, the treatments are clearly more different in their typical effect than is true for the distribution in the middle. This is represented by the horizontal displacement of the distributions, and it is easily seen by comparing the vertical lines designating the means of the distributions. How is power affected by this difference in the magnitude of the treatment effect? Intuition tells us that a larger difference between the means should be easier to detect and distinguish from the null hypothesis distribution. Our intuition is correct, as we can appreciate by once again considering the areas designated α, β, and $1 - \beta$. As the distance between the null hypothesis and research hypothesis distributions increases, the α area does not change, but the β area decreases with a corresponding increase in the $1 - \beta$ area. Thus, as the difference in reality between the treatments increases, power also increases.

Variability Among Subjects

The third factor that affects power is variability among subjects. The concept of variability arises from the fact that human beings are not identical. For example, we are not all the same height or weight, and we do not all have the same ROM at the knee. In all of the curves we have used to represent distributions, the width of the curve is a function of the variability among subjects. The wider the curve, the greater the variability; conversely, the narrower the curve, the less the variability. But how does variability affect power? Notice in Figure 5 that the difference between group means, as represented by the distance between the 2 means, is the *same* for the pair of distributions on the top as for the pair of distributions on the bottom. Notice also that the distributions on the bottom are wider than those on the top. There is more overlap of the distributions on the bottom than for those on the top, and as the overlap increases, so does the area designated as β. The net effect is that $1 - \beta$, our index of power, or our ability to detect a real difference, decreases as variability increases. This makes sense because the variability reflects noise, and with more noise it is more difficult to detect the signal. Actually, the variability in the figures is the variability of the differences between the means, not the variability of the individual subjects; however, this sampling variability is a function of the variability of the subjects, and so the conclusions about the effect of variability on power hold. If the samples on which means are based are quite variable, then the means, too, will vary considerably and so, too, will the differences between the group means.

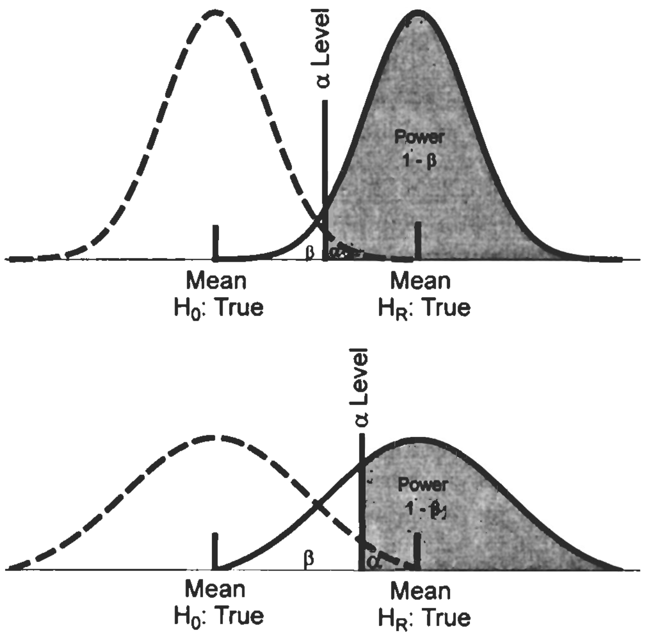


FIGURE 5. Set of 4 sampling distributions. The means for the 2 distributions on the left are identical, as are the means for the 2 distributions on the right. Note, however, the difference in the width of the distributions; the variability is greater for the 2 lower distributions than for the 2 upper distributions, and the size of the area labeled β is larger for the bottom pair than for the top pair of distributions. Thus, $1 - \beta$, or power, is greater for the top pair of distributions than for the bottom pair of distributions.

Sample Size

The last major determinant of power that we will discuss is sample size. Practically speaking, the easiest way to increase power is to increase sample size (ie, the number of subjects included in the study). Intuitively, that sounds right—the more people we have, the more precise the results, and the easier it should be to detect a signal amid the noise. One way to understand this concept is to consider how sample size affects the confidence we have in a single mean. For example, suppose that we want to know the typical knee ROM for adults. We could select a random sample of 10 adults from the population, measure their ROM, and calculate the average. How close would this average be to the average of all adults in the population? It might be close, but it might be far off. By chance alone, we might have measured a relatively young and healthy sample, and the average ROM might be a bit higher than is generally true for the population at large. If we repeated this exercise many times, we could build a distribution of means, all from random samples of 10. This would be a sampling distribution (similar to those in Figures 2–5) that would tell us how much variability in the means we could expect just by chance alone. We might expect a large variance with a small sample size and a high likelihood that one random sample could be quite different from another.

Now suppose that we measure again, but we increase our sample size to 1000. The odds that any one sample will be that much different from another are low. Random samples are more representative of their populations as sample size increases, which is why readers trust opinion polls based on a sample of 1000 more than they trust opinion polls based on samples of 10. The effect on the sampling distribution of increasing sample size is to decrease the variability of the sample means. Because each large sample is not that much different from other large samples, the means do not vary much either. Sample size affects power by influencing the variability of the sampling distribution of mean differences.

In summary, power is greatest when the α level, difference between group means, and sample size are large and the variability among subjects is small. Although greater power is achieved with a larger, rather than a smaller, α level, recall that the likelihood of making a type I error is also increased when α is set at a high level. Assuming researchers do not want to risk contributing *false findings* to the literature, they will seldom set the α level above 0.05. Instead, they will attempt to insure a *large* mean difference by selecting 2 treatment conditions that are likely to produce substantial differences, minimize variability among subjects by narrowly defining the population of interest, carefully controlling the measurement setting, and using a *large* sample size.

EFFECT SIZE

The difference between group means and variability are often combined into one number that is called an *effect size*. There are quite a number of ways to define effect size,[4–6,9] but most represent some way of standardizing the effect magnitude so that effect sizes from different studies can be compared. Standardizing the effect size is important if you want to compare the results of different studies in which the same basic question is addressed but in which the scaling properties of the measurements used are different. The most typical way to standardize a difference between group means is to divide it by the standard deviation within the groups. The result is a difference between group means represented in standard deviation units, just as a *z* score[8] represents an individual score in standard deviation units. Recalling that the variability (ie, the standard deviation) arises by chance, or noise within the groups, we can see that an effect size estimate is really a standardized signal-to-noise ratio. *T* ratios (Appendix 1) should not be confused with effect sizes. They both represent signal-to-noise ratios, but they are calculated in different ways and are used for different purposes. An effect size simply tells how large the mean difference is relative to ordinary variation. The *t* ratio takes that idea one step further and allows us to decide if the difference is improbable, given ordinary variation. There is a *P* value associated with the *t* ratio, but an effect size is typically used as a descriptive measure of difference.

TABLE. Factors affecting power.

Alpha (α) level	The criterion level of acceptability for a type I error; typically expressed as a probability value (eg, $\alpha = 0.05$).
Effect size (*d*)	An index used to express the size of the difference between group means, relative to ordinary within-group variation (eg, $d = (M_1 - M_2)/s$, where *d* is effect size, M_1 and M_2 are group means, and *s* is the pooled standard deviation of both groups).
Sample size (*N*)	The total number of subjects included in the analysis of the data.

Ordinarily, the α level is set by convention and the effect size is determined by past research or experience. This leaves sample size as the most common way to increase power and leads to one of the most common questions in research design: "How large a sample is large enough?" A technique called power analysis can be used to help answer that question.

POWER ANALYSIS

Power analysis is a technique based upon the interrelationships among power, α level, effect size, and sample size. The relationships among these factors are such that values for any 3 of the factors can be used to estimate the last. For example, if you are planning a study and you know the α level, effect size, and desired power level, you can estimate the number of subjects required. Although the example just cited is prospective in nature, power analysis can also be used retrospectively.

Assume you read an article and the results indicated that there was *no* significant difference between the treatments. You would want to know if there was enough power in the study to warrant the conclusion; that is, you would want to make sure that the researchers did not just *miss* finding a real difference. You could use information from the article about the α level, effect size, and sample size to estimate the level of power present in the study (Table). Much of the work of power analysis is performed by referring to tables,[2] using simple formulas,[3] or relying on software,[1] but first we need to gather the values for all of the relevant indices.

Estimating Sample Size

First we will consider an example of using power analysis to estimate the sample size (*N*) required to yield statistically significant results from a study. The value for the α level is generally set at 0.01 or 0.05. For our example, we will use $\alpha = 0.05$. Since we have not yet conducted the study, we do not know how large of a difference in ROM will be produced by giving contract-relax exercise to one group and passive stretching exercise to the other group. For the sake of our example, assume that a similar study had been conducted and that the authors reported means of 100 and 92, respectively, for the 2 treatment groups, and a standard deviation of 5. Given the data reported, we could calculate a standardized index of effect size (*d*) using the formula $d = (M_1 - M_2)/s$, where M_1 is the mean of group 1, M_2 is the mean of group 2, and *s* is the pooled standard deviation of both groups. In our example, *d* would be equal to $(100 - 92)/10$ or $d = 0.80$. If no data are available, we might choose to use one of the conventional values for effect size proposed by Cohen.[2] According to Cohen, values of 0.80, 0.50, and 0.20 are considered large, medium, and small effect sizes, respectively. Note, though, that relying on such crude benchmarks should be avoided if information specific to your research area is available. If we expected very little difference in ROM between the 2 exercise groups, we would be wise to select a small effect size. Conversely, if we expected a large difference in ROM between the 2 exercise groups, we would select the large effect size. Next, we would need to specify the level of power considered acceptable for our study. Recall that power can be defined as $1 - \beta$ where β represents the probability of making a type II error (ie, a *missed* finding). When choosing a power level, the researchers must decide how willing they are to make a type II error. By convention, the minimum power value typically used is 0.80. Now that we have all of the indices required, we would use the appropriate table[2(p54)] and learn that the number of subjects required for our study is 20 in each group. If we were to obtain that sample size, we would have an 80% probability of rejecting the null hypothesis if the null hypothesis is in fact incorrect and if our assumptions about effect size were correct.

Estimating Power

When a study has already been completed, power analysis typically answers whether or not the level of power present in the study was adequate to justify the conclusions. Because power analyses make sure that we do not *miss* findings, the adequacy of power is important to consider when authors report results that are *not* statistically significant and they use the results to support a view that the treatments studied are either not differentially effective or are ineffective.

In our ROM study, what happens to power if the effect size and sample size are both smaller in a new study than in our original, hypothetical study? If the researchers used only 8 subjects in each group instead of 20 subjects and the effect size was only 0.50 instead of 0.80, then at the same α level of 0.05, the power level would be 0.25 instead of 0.80. Given the new set of indices, it is much more likely that a type II error would occur and we would erroneously conclude there was no difference between the treatments; unless, of course, we decided not to accept the result due to inadequate power.

SUMMARY

We cannot know if the research hypothesis is true or false with absolute certainty, but we can attempt to reduce the probability of being wrong. By setting the α level at 0.05 or less, researchers protect against type I errors with a specified degree of certainty. In so doing, researchers limit the contribution of *false findings* to the literature when the null hypothesis is true. By maximizing the power in a study, researchers protect against type II errors with a specified degree of certainty and decrease the likelihood that they will *miss* important findings. There are, however, practical issues that occasionally limit the amount of power that can be achieved in a study. Although power can be increased by increasing the α level, researchers generally will not sacrifice protection against type I errors just to reduce the likelihood of a type II error. Instead, they will increase effect size and sample size. The problem with increasing effect size is that comparisons among some types of treatment may never be expected to yield large differences. Even if large differences can be produced, they may not be meaningful if their production relied on procedures not likely to be encountered in practice. Stated differently, there is a difference between statistical significance and practical significance. In their zeal to produce statistically significant results, researchers should not resort to tactics that reduce the practical significance of their findings. The search for statistical significance should not become an absurd end in itself. Consider Thompson's view:[10] "Statistical significance testing can involve a tautological logic in which tired researchers, having collected data on hundreds of subjects, then conduct a statistical test to evaluate whether there were a lot of subjects, which the researchers already know, because they collected the data and know they are tired."[(p436)]

Maximizing sample size may be either too costly, in terms of time and money, or impossible, if the population of interest is relatively small. Researchers are challenged to design a study that affords the optimal balance of all the factors that affect power. Readers are challenged to assess whether the researchers adequately protected against type I and type II errors and, thereby, insured against both *false findings* and *missed findings*.

REFERENCES

1. Borenstein M, Cohen J, Rothstein H. *Power and Precision: Power Analysis and Confidence Intervals.* St. Paul, Minn: Assessment Systems Corporation; 1999.

2. Cohen J. *Statistical Power Analysis for the Behavioral Sciences.* 2nd ed. Mahwah, NJ: Erlbaum; 1988.
3. Cohen J, Cohen P. *Multiple Regression/Correlation Analysis for the Behavioral Sciences.* 2nd ed. Mahwah, NJ: Erlbaum; 1983.
4. Cortina JM, Nouri H. *Effect size for ANOVA designs.* Thousand Oaks, Calif: Sage; 2000.
5. Hedges LV, Olkin I. *Statistical Methods for Meta-analysis.* Orlando, Fla: Academic Press; 1985.
6. Hunter JE, Schmidt FL. *Methods of Meta-analysis: Correcting Error and Bias in Research Findings.* Newbury Park, Calif: Sage; 1990.
7. Nickerson RS. Null hypothesis significance testing: a review of an old and continuing controversy. *Psychological Methods.* 2000;5:241–301.
8. Portney LG, Watkins MP. *Foundations of Clinical Research: Applications to Practice.* Norwalk, Conn: Appleton & Lange; 1993.
9. Rosenthal R. *Meta-analytic Procedures for Social Research.* Newbury Park, Calif: Sage; 1984.
10. Thompson B. Two and one-half decades of leadership in measurement and evaluation. *Journal of Counseling and Development.* 1992;70:434–438.

APPENDIX 1

Signal-to-Noise Ratio

Readers can think of the difference between the group means (the numerator of the *t* ratio) as a *signal* that you are trying to detect and the variability within groups (the denominator) as the *noise* that makes it difficult to detect the signal. Noise in the data (eg, excessive variability in range of motion [ROM] scores) limits your ability to detect the signal. In the case of our experimental data, noise could arise even though all of the subjects in a group received the identical treatment because there could still be random differences in ROM among the subjects. The *t* test can be thought of as a signal-to-noise ratio. If we want to be able to detect the signal, then we want the signal to be strong (ie, large between-group difference) and the noise to be weak (ie, small within-group variability), that is, we would want the *t* ratio to be as large as possible. A large *t* ratio tells us that the difference in knee ROM between the groups is greater than what occurs within the groups by chance alone.

The next question to answer is, "How large of a difference between groups is large enough?" To answer the question, we would refer to a table of critical values for the *t* ratio and learn whether the value we obtained for *t* is likely to have occurred by chance or whether our signal (mean difference between groups) was large enough relative to the noise (within-group variability) to make the treatment effect noticeable or distinct. If the probability assigned to the *t* ratio we obtained in our experiment were less than 0.05, then we would conclude that the group difference is unusual relative to what might ordinarily be expected to occur by chance (and thus the low *P* value). Accordingly, we would conclude that there *is* a difference in knee ROM depending upon the type of exercises. If the *P* value was greater than 0.05, then we would conclude that there is no statistically significant difference in the effect of the 2 types of exercise for increasing ROM. The difference we obtained is consistent with the variability expected to occur anyway.

APPENDIX 2

Statistical Significance, Alpha Level, and Null Hypothesis

The acceptable level of a type I error is known as the alpha (α) level and specifies how much of a chance the researchers are willing to take that they will falsely conclude that one type of exercise treatment is better than the other. Most researchers are unwilling to take more than a 5% chance of making this type of mistake, so they will set the α level at a probability value of $P = 0.05$. As noted in the first paragraph of this article, authors report *P* values along with the results of statistical analyses. The *P* value they report is the probability that they would make a mistake if they concluded that the research hypothesis was true when, in fact, there really is no difference between the groups (ie, there really is no signal, just noise). Once they know the *P* value for their result, they can compare the obtained *P* value to the criterion α level. At this point, the rule for deciding is really quite simple: if $P < \alpha$, then reject the *null hypothesis* and conclude that the research hypothesis is more plausible. Up to this point in our discussion, we have focused on the research hypothesis. By convention, however, a statistical test is on the *null hypothesis.* Logically speaking, a research hypothesis must be tested indirectly by trying to disconfirm the null hypothesis. An example of a null hypothesis is: "If the treatments are equally effective, then there will be no difference in range of motion when the study is over." If there are range of motion differences at the conclusion of the study, then authors will reject the null hypothesis in favor of the more plausible alternative that one treatment is more effective than the other treatment. Nickerson[7] provides a comprehensive but readable treatment of null hypothesis testing, describing the logic behind it and the interpretational difficulties that have made it a controversial procedure. In our hypothetical study, assume we set the α level at the conventional level of $P = 0.05$ and that the P value derived from our results was 0.03. Because 0.03 is less than 0.05, we would conclude that the difference between the groups is too large to have occurred just by chance. The more plausible conclusion is that the difference was produced by the intervention and that the 2 types of stretching exercises are not equally effective interventions for increasing range of motion. We can never be completely sure that we have not made a type I error. A difference as large or larger than we found *could* have occurred just by chance (ie, just due to the variability that occurs ordinarily), but it is not likely ($P = 0.03$).