Journal of Surgical Research **126,** 149–159 (2005)
doi:10.1016/j.jss.2004.12.013

Statistical Power and Estimation of the Number of Required Subjects for a Study Based on the *t*-Test: A Surgeon's Primer

Edward H. Livingston, M.D.,*,[1] and Laura Cassidy, Ph.D.†

**Division of Gastrointestinal and Endocrine Surgery, University of Texas Southwestern School of Medicine and the Veterans Administration North Texas Health Care System, Dallas, Texas; and †Graduate School of Public Health, University of Pittsburgh, Pittsburgh, Pennsylvania*

Submitted for publication October 22, 2004

The underlying concepts for calculating the power of a statistical test elude most investigators. Understanding them helps to know how the various factors contributing to statistical power factor into study design when calculating the required number of subjects to enter into a study. Most journals and funding agencies now require a justification for the number of subjects enrolled into a study and investigators must present the principals of powers calculations used to justify these numbers. For these reasons, knowing how statistical power is determined is essential for researchers in the modern era. The number of subjects required for study entry, depends on the following four concepts: 1) The magnitude of the hypothesized effect (i.e., how far apart the two sample means are expected to differ by); 2) the underlying variability of the outcomes measured (standard deviation); 3) the level of significance desired (e.g., α = 0.05); 4) the amount of power desired (typically 0.8). If the sample standard deviations are small or the means are expected to be very different then smaller numbers of subjects are required to ensure avoidance of type 1 and 2 errors. This review provides the derivation of the sample size equation for continuous variables when the statistical analysis will be the Student's *t*-test. We also provide graphical illustrations of how and why these equations are derived.

Key Words: ***t*-test; normal distribution; Z-statistic; study design**

INTRODUCTION

Frequently asked these days is how many subjects are really needed for a study [1]. Calculations for answering this question are not intuitively obvious making the determination of adequate sample size a mysterious process usually relegated to the local statistician. Most computerized statistical packages include sample size calculators allowing investigators to perform these assessments on their own. Because computers will provide answers even if they are the wrong ones, it is important for surgical investigators to understand the basic tenets of sample size calculation so that they can ensure that computerized algorithms are appropriately used.

Previous statistical reviews published in the *Journal of Surgical Research* summarized concepts necessary for the understanding of sample size determination. The basis for these calculations includes knowledge of data classification, measures of central tendency, the characterization of data sets [2], the fundamental concepts of group comparisons and determination of statistically significant differences [3]. Statistics are all about probabilities, using a sample to make inferences about a population and minimizing the risk of making erroneous conclusions regarding that population. As such, there are several potential errors that must be avoided that are summarized in Table 1.

Table 2 illustrates these relationships. Type 1 error occurs when a screening test returns a negative result when a patient has a disease. Type 2 error occurs when a screening test returns a positive result when a patient does not have a disease. In designing experiments, we attempt to minimize both types of errors but minimization of type 1 error is most important. The consequences of not establishing a diagnosis in a pa-

[1] To whom correspondence and reprint requests should be addressed at Gastrointestinal and Endocrine Surgery, UT Southwestern Medical Center, 5323 Harry Hines Blvd Room E7-126, Dallas, TX 75390-9156. E-mail: edward.livingston@utsouthwestern.edu.

0022-4804/05 $30.00

TABLE 1

Types of Statistical Error

Symbol	Defintion	Alternate definition	Implication
α	Probability of rejecting H_0 when it is true	Probability of an observed difference resulting from chance alone	Type 1 error
β	Probability of accepting H_0 when it is false	Probability of concluding that no difference exists when one is present	Type 2 error
$1-\beta$	Probability of rejecting H_0 when it is false	Probability of detecting a statistically significant difference if one exists	Power

Note. Errors are the risk of falsely accepting or rejecting a null hypothesis when it is true or false. H_0-the null hypothesis.

tient with some disease are more significant than falsely believing a person has a disease that, in reality, they do not have.

Statistical writings are replete with double negatives and confusing verbiage. Surgeons and other biologists may be intimidated by statistical language resulting in a poor understanding of statistical concepts and tests. Particularly striking is the basic tenet of significance or hypothesis testing: The null hypothesis. It is represented by $\mathbf{H_0}$ and is defined as the assumption that no statistically significant differences exist between important properties describing groups being compared. Alternatively, $\mathbf{H_1}$, the alternative hypothesis represents the assumption that the measured entities characterizing the two groups are indeed different.

Confusion results from the application of double negatives. We seek to prove that the null hypothesis, i.e., that no statistically significant differences exist, is false. It is easier to state that we are seeking to find differences between groups when they exist. That view is more intuitively obvious and easier to reconcile. However, it is important to recognize the null hypothesis's meaning given that it is ubiquitous in statistical writings.

When statistically significant differences are calculated, an arbitrary α value is set. Statistical convention sets this value at 0.05. In other words, there is a less than 5% probability that observed differences between groups occur because of chance alone rather than a true difference between the groups. The α value establishes the risk of type 1 error, or the risk of falsely concluding that differences between groups exist when in fact none do.

Type 2, or β error, is the possibility of concluding that no statistically significant difference exists when, in fact, the groups being compared really are different. The statistical power of a test is defined by $1-\beta$ or the probability that when a test concludes that there is a difference between the groups being compared that the test result is correct. For example, if the β is 0.1 then there is a 10% chance that two groups really were different when a statistical test suggests that the mean values for the properties describing the groups were not different. $1-\beta = 0.9$ such that the tests power is 0.9. This means that the statistical test has a 90% probability of being correct if it concludes that there is a difference between groups when a difference really exists. Implicit in this is that if the test finds no difference between the mean values describing the groups properties, there is a 90% chance that there really is no statistically significant difference between the groups.

THE LONG FORGOTTEN EPIC BATTLE BETWEEN STATISTICAL GIANTS

A bitter, ferocious argument smoldered over the course of decades early in the last century between those responsible for developing the concepts of statistical significance testing and hypothesis evaluation [4]. The story starts with Karl Pearson (1857–1936) considered being one of the founders of statistics. He was responsible for the Pearson correlation coefficient, the χ^2 test, linear regression and other fundamental concepts of statistics. He founded and headed the Department of Statistics at University College in London. That department was split into two to accommodate two other giants in statistics, Egon Pearson (1895–1980), Karl's son and Ron Fisher (1890–1962). Fisher was most prolific, publishing on average one paper

TABLE 2

Further Examples of Error Types as they Pertain to Diagnostic Tests

	Actual situation (disease)	
	Positive	Negative
Screening test		
Positive	Correct	Type 2 error
Negative	Type 1 error	Correct

Note. Type 1 errors occur when a diagnostic test is negative but a patient actually has a disease. This is considered to be more serious than type 2 errors being that establishing the diagnosis of a medical problem may be missed. Type 2 error occur when a test is falsely positive, i.e., that a test is positive when the patient does not have a disease. Under these circumstances, the positive test will result in further evaluation that will, hopefully, reveal that the disease was not actually present.

every 2 months and was responsible for many concepts guiding research today. Most importantly, he developed the idea of significance testing and P values. Fisher's analytic approach was inductive being that he believed that experimental observations were representative samples of a larger universe or population of observations for features that characterize groups being compared experimentally. Experimental data were considered as a sample from a larger population with various assumptions being made regarding how representative the sample was of the overall population. Fisher's concept was that when groups were compared the probability (P) that there was no statistically significant difference between them, i.e., the null hypothesis was true, could be estimated. The smaller the P, the more likely that the groups were different. Importantly, Fisher did not advocate a threshold P value for decision-making regarding a yes or no decision for statistical significance. Fisher believed that the strength of evidence for group difference rested with the P value and the investigator had to decide if the differences were important or not. Legend has it that the origin of the famed $P < 0.05$ threshold emanated from Karl Pearson's refusal to publish (as editor of *Biometrika*) Ron Fisher's table of statistical probabilities in their full form [5]. Fisher was forced to truncate the tables into categories of $P = 0.05$ and 0.01. Given the availability of these values in the published tables, these probability levels became the standard thresholds for statistical analysis.

Fisher was completely opposed to the concept of a yes or no decision regarding significance testing. He believed that the P values should be assessed in the contect of the data as a continuous variable. The smaller the P value, the greater likelihood that groups were truly difference [6]. A major weakness of Fishers approach was that when groups do not differ, i.e., the null hypothesis is proven correct, definite conclusions regarding the groups could not be made. Fisher passionately believed that conclusions were valid only when they proved differences between groups. When not different further experimentation was necessary. With this philosophy data are viewed as representative of some larger universe of findings and that there is no prior knowledge of regarding the system under study.

Fisher's colleague at the University of London, Egon Pearson developed a completely different view of data analysis. Together with Jerzy Neyman (1894–1981), Pearson developed the concept of hypothesis, rather than significance testing [7]. The intent was to establish a mechanism for making decisions about groups being compared and quantitating the costs incurred for making these decisions. This had great practical importance for industry being that when using statistics to analyze business systems, experiments were designed that was costly to perform and had to result in useful decisions regarding industrial processes. Fisher's significance testing was more theoretical and less practical and ignored the cost of performing an analysis. In this regard Fisher's system was considered more applicable to scientific research than to industrial applications. Neyman and Pearson's system was deductive, in contrast to Fisher's inductive analysis, in the sense that data were produced under predefined circumstances with predetermined constraints for analyzing them. They introduced the concept of α, the level of significance associated with type 1 error. They also introduced β, the counterpart for type 2 error. From this $1-\beta$ or the experiments power could be determined. Now the cost of an experiment in terms of trading off type 1 and 2 errors could be established when designing an experiment. Additionally, their system was a decision-making scheme for accepting or rejecting hypothesis. An important difference between this and Fisher's significance testing was that Fisher did not provide for making decisions regarding groups, only for assessing the relative strength to accepting or rejecting null hypotheses. Neyman and Pearson provide yes or no decisions regarding hypotheses but not any assessment of the strength of the differences between groups. In this regard, P values and α levels are not the same and should not be confused. At the heart of the Neyman-Pearson approach is the simultaneous consideration of the null and an alternative hypothesis whereas Fisher believed that alternative hypotheses had no validity.

Table 3 summarizes the differences between Fisher significance analysis and Neyman-Pearson hypothesis testing. Modern statistics has adopted the latter approach but, commonly, significance testing remains pervasive and often blended into discussions of hypothesis testing. Some of the most rancorous discussions and life-long disputes between Fisher and the others resulted from these differences in opinion regarding the appropriate form of data analysis. Fisher complained that Karl Pearson restrained his ability to publish his ideas. Fisher and Neyman-Pearson had public battles waged in open discussions and the published literature. These battle were waged over the course of four decades with neither side fully acknowledging the others validity. Only Student, i.e., William Sealy Gosset, was able to mediate between parties. Although intensely loyal to Fisher for legitimizing his concepts of small-sample effects resulting in the development of the Student's t-test, even Student noted that the Neyman-Pearson approach to hypothesis testing was more effective that Fisher's significance tests.

STATISTICAL POWER

Power is a measure of a statistical tests ability to detect differences. The importance of this is that when no statistically significant differences are found be-

TABLE 3

Difference between Significance and Hypothesis Testing

Fisher significance test	Neyman-Pearson hypothesis test
Inductive: From specific to general	Deductive: From general to specific
Inductive Inference: Interpretation of the strength of evidence in data	Inductive behavior: Make decisions based on data
P-value: Data-based random variable	α: Pre-assigned fixed value
Property of data	Property of test
Short-run: Applies to any single experiment or study	Long-run: Applies only to ongoing identical repetitions of original experiment or study, not to any given study
Hypothetical infinite population	Clearly defined population

Note. Significance testing only enables the investigator to characterize the strength of the rejection of a null hypothesis. When the null hypothesis cannot be rejected, i.e., no statistically significant differences are found between the groups being compared, no conclusions can be drawn. Thus, one can never definitively conclude that the groups being compared are, in fact, the same. By simultaneous assessment of a null and alternaive hypothesis, hypothesis testing allows for conclusions that the null hypothesis is accepted when the alternate hypothesis is rejected providing some degree of assurance that the groups being compared are statistically the same. Adapted from [5].

tween groups, i.e., the null hypothesis cannot be rejected, there is a quantifiable degree of assurance that the groups are indeed not different. For hypothesis testing, an α level is fixed, establishing a threshold for rejecting the null hypothesis. This process defines what the allowable type 1 error will be and is typically set at 0.05. Next the β level is established, fixing the allowable type 2 error rate will be. Recall that β, or type 2 error, represents the likelihood of accepting the null hypothesis when it is in fact false. Under these circumstances the groups are believed to be the same when in fact they are different. Being less critical than type 1 error, allowable type 2 error rates are usually fixed at 0.2. Power is $1-\beta$, usually 0.8, and represents the likelihood of rejecting the null hypothesis when it is false. This represents the probability of the testing procedure correctly concluding that the groups are different when they are indeed different.

For continuous, normally distributed data, the test most commonly used to determine the statistical significance between observed differences between groups is the Student's t-test [3]. The basic equation for this test is:

$$t = \frac{\mu_1 - \mu_2}{\sigma\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$$

t is the t-statistic whose numerical value is a measure of how different group means are. The t-statistic provides a measure of how extreme a statistical estimate is. The larger the t-value, the greater the strength of evidence is that two group means are different. μ_1 and μ_2 represent the mean values from the sample being studied.[1] Obviously, the larger the difference between means the larger the t-value and, therefore, the greater the probability that the groups differ. σ is the standard deviation for the two groups. The Student's t-test is predicated on both groups having equal variance, i.e., the same standard deviation. The number of data points in each group is represented by n. This term accounts for the uncertainty in determining the true mean from small samples. Smaller group sizes will result in lower t-values and, therefore, less likelihood that statistically significant differences exist.

Rearranging the equation yields the following proportionality:

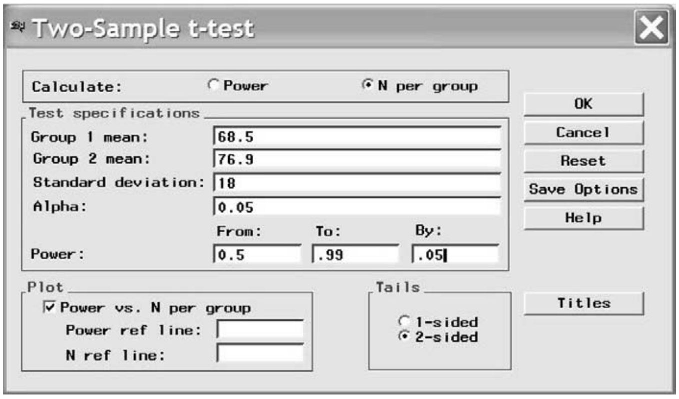
$$n \propto \left[\frac{t\sigma}{\Delta\mu}\right]^2$$

which means that to avoid type 2 error, the number of subjects required for a study will be dependent on the t-value for that study, which in turn, is dependent on the α-level that has been established. The required number will also be proportional to the study populations' standard deviation and inversely proportional to the difference in mean values for the two groups. Thus, one of the parameters most computer programs require for calculating statistical significance is the α-level that has been established for the study. In most cases this is 0.05. The number of subjects will be proportional to the standard deviation, σ. When the standard deviation is larger, there will be a need for larger study groups. Similarly, if the means between groups are relatively close to one another then the study groups will also need to be larger to ensure that the study has sufficient power to ensure that if no statistically significant differences between groups are found then there is little likelihood that a difference truly exists. Sample size is a balancing act and in general, the more variability in the data (σ.), the larger the sample size, the smaller the difference being detected ($\Delta\mu$), the larger the sample size and/or the smaller the probability of an observed difference resulting from chance alone (α-level), the larger the sample size. From this relationship, one can see that establishing the α-level and having some idea of the population's standard deviation and the size of the difference between means of the groups is needed for calculating the number of subjects required for a study.

SIZE EFFECTS OF THE SAMPLE, VARIANCE, AND MEAN DIFFERENCE

To illustrate power calculations consider the following example: We desire to know if men are, on average, heavier than women in the U.S. population. To determine this we obtained body weight measurements from 31,132 individuals that underwent physical examinations as part of the Third National Health and Nutrition Examination Survey (NHANES III) [8]. We avoided the confounding effects of children and patients with disease by excluding those with weight less than 40 kg. When this was done the mean ± SD weight for 11,035 women was 68.5 ± 17.8 kg and for 9,709 men was 76.9 ± 18.1 kg. A t-test revealed that men's weight is statistically significantly different than women, with a P value of <0.0001. Using a two-tailed test, we can only make the inference that the means are different, however, based on the values of the means we can state that men average weight is significantly greater than woman's. The assumption that the standard deviations for men and women are approximately equal was evident, ensuring the t-test was a valid method. Using the SAS statistical package (Fig. 1), a power analysis was performed yielding the output displayed in Fig. 2.

Given the body weights we assessed, one would need 98 patients per group to ensure a 90% probability that statistical testing will reveal that weights were different if indeed they really are. Similarly, if the two populations whose weights are being compared are really not different, there is a 90% chance that the testing procedure will result in a conclusion that they are not different, i.e., the null hypothesis is not rejected. If there were 37 patients per group, then there would only be a 50% probability that the statistical testing procedure would result in rejecting the null hypothesis, i.e., concluding that there was a statistically significant difference between the groups.



Calculate: ○ Power ● N per group

Test specifications

Group 1 mean: 68.5

Group 2 mean: 76.9

Standard deviation: 18

Alpha: 0.05

Power: From: 0.5 To: .99 By: .05

Plot

☑ Power vs. N per group

Power ref line:

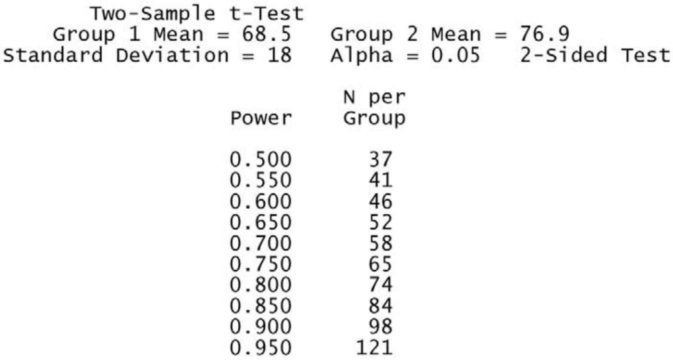
N ref line:

Tails

○ 1-sided

● 2-sided

FIG. 1. SAS input for power analysis. This form typifies that found in most statistical packages calling for entry of the means and standard deviations to be assessed. The SAS package can provide the statistical power or the number needed per group in a study to achieve the desired type 1 and 2 error rates.



```
              Two-Sample t-Test
      Group 1 Mean = 68.5    Group 2 Mean = 76.9
 Standard Deviation = 18     Alpha = 0.05    2-Sided Test

                              N per
                    Power     Group

                    0.500       37
                    0.550       41
                    0.600       46
                    0.650       52
                    0.700       58
                    0.750       65
                    0.800       74
                    0.850       84
                    0.900       98
                    0.950      121
```

FIG. 2. SAS output for power analysis. Achieving greater power is dependent on the sample size with increasingly large samples resulting in greater statistical power. When designing experiments the desired power is established ahead of time and represents a compromise between assurance that the resulting statistical analysis accurately represents the groups characteristics behaviors against the cost of performing the study attributable to the number of required subjects. For most instances, selection of 80% power results in an adequate compromise.

For the entire population of body weights the distribution is displayed in Fig. 3. The y-axis represents the frequency with which any individual weight occurs in the population. The x-axis displays the weights. Women are represented in by the black-colored curve and men by the one that is gray. The curves are reasonably broad that pictorially demonstrates the somewhat large standard deviations. However, the peaks are far apart from one another. Thus, there is little doubt that men are significantly heavier than women.

What do the curves look like when there is less power? As seen in Fig. 4, a random selection of 90 patients were selected from the overall population. As can been seen, the smaller sample size results in a frequency distribution that is much less bell-shaped (Gaussian in statistical terms) than the curves obtained from the overall population. There is much more variation in men's weight manifested by the wider frequency distribution. Statistically, this is manifested by a larger standard deviation. Despite the large variation in the body weights, the mean values between men and women remain very different. This illustrates that even when sample size's become small, if the means are very different, there is a high likelihood that significant differences will be found.

What is the effect of group size when the means are relatively close together? Using the same population we asked the question if the high-density lipoprotein (HDL) levels were different between women and men.

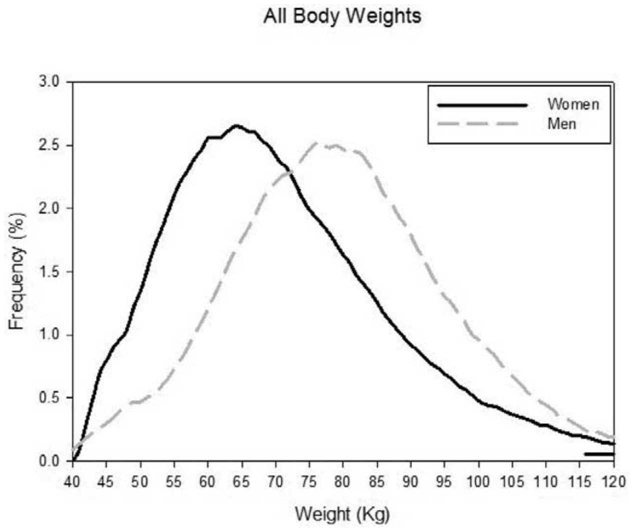


FIG. 3. Body weights for the entire NHANESIII population. Assuming that this dataset represents the entire universe of body weights, the groups are different because their mean values are different.

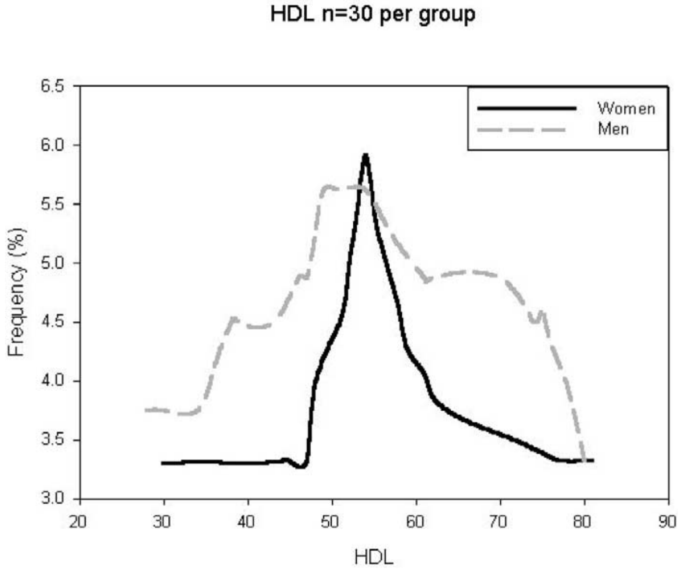


FIG. 5. HDL for a sample of the NHANESIII population.

Selecting 30 patients in each group yielded the probability distribution shown in Fig. 5. For this sample the mean ± SD for women was 53.4 ± 12.9 and for men it was 47.4 ± 11.9. The *t*-test for significant differences between these two groups was 0.09, i.e., they were not significantly different at the $\alpha = 0.05$ level. However, the statistical power was only 0.40 suggesting that given the wide scatter in the data and the relatively close values for the group's means, this sample could not reliably exclude that the groups were not different. In other words, based on this sample, there is not enough power to detect a statistically significant difference and one cannot definitively say that the groups were not statistically different based on the *t*-test. What happens when we examine the entire population of 2,700 patients that had lipids drawn in NHANESIII? Figure 6 shows the frequency distribution for the entire population. It is evident that the two groups do differ with women having higher HDLs than men. The mean ± SD HDL for women was 54.2 ± 15.2 and for men it was 47.4 ± 14.1.

These figures and statistics demonstrate the importance of having sufficient numbers of subjects in a study before concluding that no statistically significant differences exist. Thus, when designing studies it is crucially important to account for the anticipated mean differences and standard deviations of the measured samples to perform a power calculation. Unless adequate numbers of individuals are

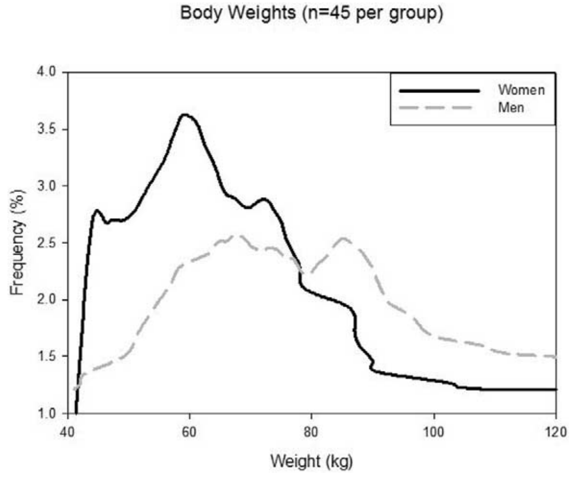


FIG. 4. Body weights for a sample of the NHANESIII population. With smaller sample sizes the peaks are less distinct.

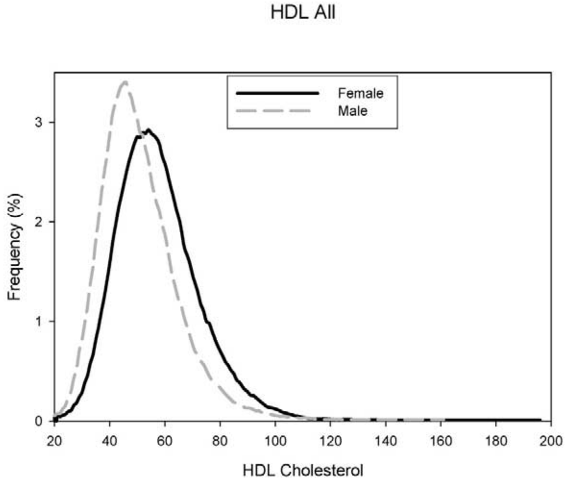


FIG. 6. HDL for the entire NHANESIII population. Assuming that these curves represent the entire universe of High Density Lipoprotein (HDL) measurements, it is evident that females have higher HDL levels than males being that the mean values for the two genders are clearly different.

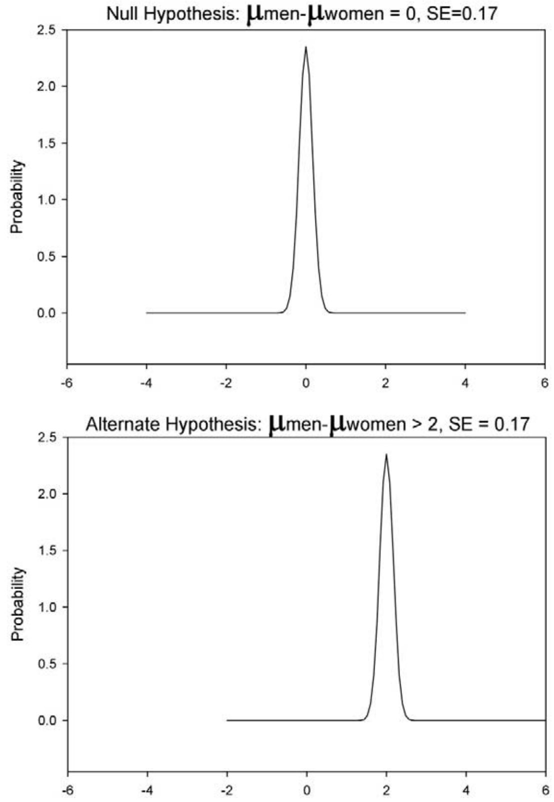
included, apparent lack of significant differences between groups may be unreliable. One cannot conclude that differences do not exist unless the study has sufficient statistical power to detect clinically significant differences.

DERIVING THE EQUATION FOR POWER CALCULATION

Although statistical nomenclature is confusing, it does serve a purpose. We have noted that use of double negatives, such as failing to reject the null hypothesis, is conceptually difficult for the non-statistician. Confusion results from verbiage such as “failing to reject the null hypothesis” rather than simply stating we accept the null hypothesis. However, statistics is used to produce estimates and make inferences about populations using smaller samples and thus must always consider events that happen because of chance. Graphical presentation of these notions clarifies why statisticians view the world in terms of null hypotheses and will show how power equations are derived.

When comparing the means of two groups we rely on statistical tests to determine if the groups are different or not. We start with the null hypothesis and assume that the groups are not different. We need to develop a mathematical equation to represent this concept. This is done by subtracting the two sample means, which should equate to zero if the means are the same. From a statistical perspective, one must account for the fact that we may not know what the mean value is exactly because of sampling issues. As was pointed out earlier in this series, the central limit theorem demonstrates that the smaller a sample size is, the less likely one is to know exactly what the overall population means is. The genius of Student, i.e., William Sealy Gossett of the Guinness brewery, was to characterize the uncertainty of knowing the true population mean value from a sample and develop a statistical test that incorporated all these concepts [3]. Thus, when we hypothesize that the means from two samples are the same, we must account for the uncertainty in not knowing the true population means. When we subtract the two, there will not be a discrete number but, rather a distribution of values reflective of the uncertainty of the true mean values. In the figures below, we present the null hypothesis as a Gaussian distribution (bell shaped, normal distribution) centered around zero. We assume that the subtracted mean values are zero but the size and shape of the bell shaped null hypothesis distribution is determined by the population variance and the uncertainty in knowing values of true population means actually based on the sampling distribution. Both of these are mathematically represented by the standard error of the mean (SEM). Recall that the SEM is the standard deviation divided by the square root of the sample size *n*. As sample sizes decrease, there is greater uncertainty in knowing exactly what the population mean is from the sample reflected in larger SEM values.

Evaluation of the null hypothesis is used to determine if groups are really different. The alternative hypothesis, $\mathbf{H_A}$, states that the groups differ by some value. Just as the null hypothesis is used to determine if the groups are different, the same backward thinking applies to the alternative hypothesis: We ask if the groups differ by some amount to statistically assure ourselves that they may be the same. Assessment of $\mathbf{H_A}$ provides with a statistical analyses probability of type 2 error and its power. These concepts are illustrated in Fig. 7. It is apparent from the above figures that when evaluating all 20,000 body weight measurements, there is a narrow distribution of error when estimating the difference between mean values. If we ask the question if men weigh at least >2 kg then women, the answer is an


Null Hypothesis: μmen-μwomen = 0, SE=0.17

Alternate Hypothesis: μmen-μwomen > 2, SE = 0.17

FIG. 7. Distribution of the null and alternate hypothesis for the patient weight data presented above. The alternate hypothesis was defined as observing a difference between the groups of larger than 2 kg. For the entire sample of measured body weights, the observed SEM was approximately 0.17. The equation for these plots is Z = $(x-\mu)$/SEM where Z represents the area under the curve and corresponds to the probability of μ, x is the value represented along the x-axis, μ, the mean value, and SEM. The plots represent the standard normal distribution such that 95% of the area under the curve occurs at 1.65 × SEM.

unequivocal yes. There is no overlap between the distribution of possible mean values at x = 2 with no overlap at x = 0. What happens when the sample size is smaller? In the example we used above, a sample of 45 individuals from each group was obtained. This resulted in a SEM of 2.5. This is demonstrated in Fig. 8.

The curves in Fig. 8 represent the null and alternate hypothesis for a sample of 45 males and 45 females taken from the NHANESIII population comparing male and female body weights. Note that with the small sample ($n = 45$) of the large (approximately 10,000) population, the uncertainty of knowing exactly where the mean is located is manifested by a much wider curve than that shown in Fig. 7. Consequently, there is much greater overlap between the null and alternate hypothesis curves for this smaller sample. The upper curve represents the null hypothesis, i.e., that there are no statistically significant differences between the samples and, when subtracted, the result is zero. The vertical line represents the critical value, which is 1.65 × the SEM, or 4.2 for this example. To the right of the critical value the area under the curve = 0.05, is the α region and is delineated in red. Thus, if the difference between the group means exceeds 4.1 there is a greater than 95% probability that the observed difference is real and not because of chance alone.

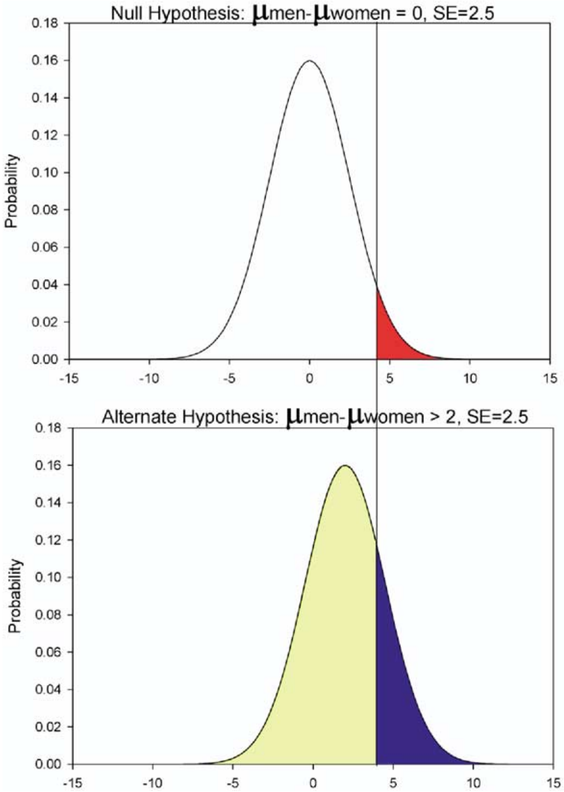


FIG. 8. Null and alternate hypothesis plots. Red shading represents the α region of the null hypothesis curve. The yellow, the β region of the alternative hypothesis. The blue portion of the alternative hypothesis defines $1-\beta$ or the power region. The vertical line is defined as the critical value. The critical value defines the accept/reject region, By convention it is usually established at the $\alpha = 0.05$ region, i.e., where the area under the probability curve is 0.05 (0.025 for two-tailed tests). To the right of the critical value we reject the null hypothesis and to the left we accept it. As described in the legend for Fig. 7, the 95%/5% region is at 1.65 × SEM. In this figure that corresponds to 4.1. Thus, the critical value for this example is 4.1.

The lower figure represents the alternate hypothesis that the means are at least 2 kg greater than one another. The yellow region to the left of the critical value is the β region and for this case is 88% of the area under the curve. The blue region to the right of the critical value on the alternate hypothesis represents the statistical power and in this case is 12%. Thus, for this example there is a 12% likelihood that statistical testing will result in rejection of the null hypothesis. In other words, there is a 12% chance that a statistical test will find that the two groups being compared will be different. From a practical perspective, this also means that if no statistically significant differences are found between the groups, the investigator cannot reliably conclude that no statistically significant differences exist between the groups. For a larger difference in the alternative hypothesis, the curve shifts to the right as is demonstrated in Fig. 9.

From these curves one can see how statistical power may be increased. If the means are very different the alternate hypothesis moves to the right increasing the area under the power region. Greater sample sizes narrows the shape of these curves such that there will be less overlap between the curves as was observed in Fig. 7. From these illustrations one can also derive equations for calculating sample sizes. First, some simplification: We assume that these curves are normally distributed. An equation describing them is:

$$Z = \frac{x - \mu}{\sigma\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$$

Where Z is a number characteristic for normal curves and can be obtained from standard tables. σ represents the standard deviation, which we assume to be equal for the two groups being compared and n is the sample size. x is a point along the x-axis and μ is the sample mean.

We use the null hypothesis to test whether or not two sample means are different or not. We assume that they are the same and if they turn out to be different we reject the null hypothesis if the means differ by a value exceeding that associated with the 5% probability region on the upper curve in Fig. 9. This is represented by the red region on the rightward tail of the probability curve. The line passing through the x-axis at this location is called the critical value. Any value of z greater than x_c will be associated with groups that

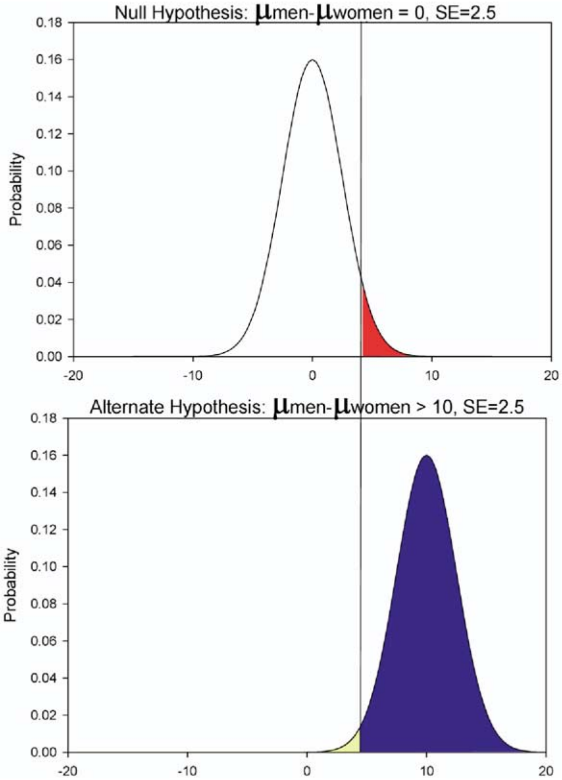


FIG. 9. An alternate hypothesis of 10 shifts the curves to the right. The yellow region, i.e., the β region, is much smaller meaning that there is a much lower probability of concluding that no difference exists when in fact it does. The blue region, which represents statistical power, is much larger.

have a greater than 95% chance of being truly different. For the null hypothesis, $\Delta\mu = 0$ thus:

$$Z_\alpha = \frac{x_c}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$$

meaning that the z value for the null hypothesis relating to the probability of avoiding α error is equal to the critical x value divided by the SEM.

The alternate hypothesis is stated to assure ourselves that we do not falsely conclude that no statistically significant differences exist when in fact they do. We assumed that no differences existed in stating the null hypothesis to prove that groups are really different. We do the opposite to assure ourselves that we do not falsely assume that there are not differences by assuming they are there. We start by making some arbitrary guess at how far apart the differences should be such that they are important.

For Fig. 9, we assumed that the means of two groups being compared must differ by more than 10. We know what x_c should be from the null hypothesis. A line is drawn through x_c on the alternate hypothesis curve dividing it into two regions. To the left of the line is the β region and is depicted in yellow. To the right is $1-\beta$, or the power region and is colored blue. For the alternate hypothesis:

$$Z_\beta = \frac{\mu - x_c}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$$

As demonstrated in Fig. 9, the critical region is to the left of the mean and, therefore subtracted from it. Making some substitutions and rearrangements:

$$\left(Z_\beta\, X\, \sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\right) = \mu - \left(Z_\alpha\, X\, \sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\right)$$

Rearranging again:

$$\left(Z_\beta + Z_\alpha\right) = \frac{\mu}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$$

and assuming that the groups being compared are the same size, i.e., $n_1 = n_2$:

$$\left(Z_\beta + Z_\alpha\right) = \frac{\mu}{\sigma\sqrt{\frac{2}{n}}}$$

or:

$$\left(Z_\beta + Z_\alpha\right) = \frac{\mu\sqrt{n}}{\sqrt{2}\sigma}$$

Squaring both sides and rearranging yields the equation used to calculate sample size for each group:

$$n = 2\left(Z_\beta + Z_\alpha\right)^2\left(\frac{\sigma}{\mu}\right)^2 \qquad \text{(Eq.1)}$$

Z_β is selected such that the critical region lies to the left of that portion of the tail corresponding to the probability that we avoid β error. Typically this is selected as 80 or 90%. Given our definitions, μ represents the expected difference between the means for the two groups being compared. n refers to the number of subjects in each group such that the total number of required observations will be $2n$ when there is two groups.

From these graphs and Eq. 1 it is evident that the number of subjects required for a study varies proportionately to the α and β levels selected. Larger standard deviations of the samples will increase the needed number by the square of σ such that small increases in the data's scatter require great increases in the sample size to avoid type 2 error. Conversely, if the two sample means are very far apart, the number of subjects required is much smaller. Although this equation is used to determine sample sizes for two groups of continuous data, similar equations have been derived for all statistical tests based on the assumptions regarding the distributions of those tests.

Often times, statistical texts refer to the effect size. This is μ/σ or what fraction of the standard deviation the difference between the expected mean values will be. From Eq. 1, we can see that the number of subjects required for a study is proportional to the squared inverse of the effect size.

Figure 10 demonstrates the impact effect size has on the required number of subjects for a study. The red line represents this effect when α is 0.05 and β is 0.8, the most common set of parameters used for these calculations. When more statistical power is desired, i.e., β is increased to 0.9, more subjects are required for the same effect sizes. The figure demonstrates the inverse square relationship between the required sample size and effect size. When the difference between means is one-half a standard deviation, one needs 32 subjects per group for $\alpha = 0.05$ and $\beta = 0.2$. For this same set of α and β, one requires 16 subjects per group if the mean difference is one standard deviation or 252 per group if the mean difference is one-fourth of the standard deviation. This figure illustrates the powerful effect on power and sample size the mean difference and standard deviation have. When designing experiments, the goal is to obtain the largest possible effect size with the smallest investment in the number of subjects studied. Studies can be optimized by looking for endpoints that have the largest possible difference between groups or by selecting groups that are homogenous as possible resulting in the smallest possible standard deviation.

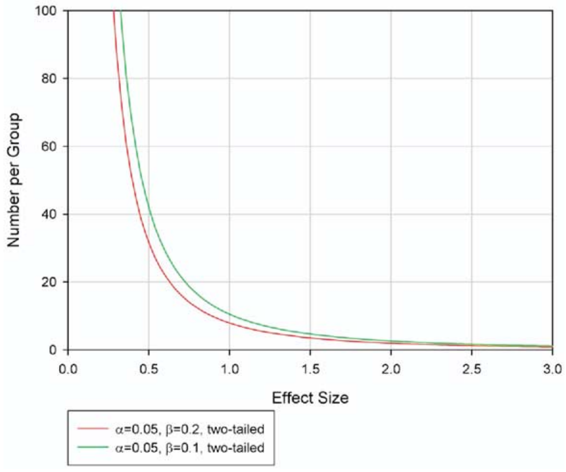


FIG. 10. Impact of effect size on the number of subjects required for each group. Effect size is defined as the difference between group means divided by the standard deviation. Given that effect size is given in terms of a fraction of the standard deviation it can provide a general classification for observed differences as a function of the data's dispersion. By convention, small effects are defined as approximately 20% of the standard deviation, medium 50% and large 80%.

SUMMARY

The calculation of sample size depends primarily on four factors:

1. Magnitude of the hypothesized effect
2. Underlying variability of the outcome measurements of interest
3. Power
4. Pre-determined level of significance.

If a target level of power is chosen (e.g., 80%), and assumptions can be made regarding the size of the true effect and the underlying variability, then one can compute the required sample size.

The magnitude of the hypothesized effect or the "minimal clinical significant difference" can be based on pilot studies, previously published literature, or even a researchers clinical experience. The underlying variability may be more difficult to determine, however, reasonable estimates may be obtainable. Ultimately, we must also consider whether the sample size estimate is feasible. There may be monetary or time constraints. Often a range of sample sizes based on different power levels and error bounds can be useful.

In conclusion, statistical power and sample size depend on the degree of assurance one selects for avoiding type 1 error (α level) and type 2 error (β level). The number required increases proportionately to the square of the standard deviation and inversely to the square of the expected difference between the means of the two groups being compared.

NOTES

1. Statistical writings always use x(bar) to denote sample means and μ to denote population means. Similarly, the population standard deviation is denoted as σ and as s when the standard deviation is derived from a sample of the population.
2. The following Web-based resources were used: URLs to sample size references (http://www.graphpad.com/index.cfm?cmd=library.page&pageID=19&categoryID=4); Tutorial regarding sample size determination for various types of data and experimental designs (http://obssr.od.nih.gov/Conf_Wkshp/RCT03/Lectures/Catellier_Sample_Size.pdf); Listing of web-based statistical resources (http://members.aol.com/johnp71/javastat.html#Power).

REFERENCES

1. Detsky, A. S., and Sackett, D. L. When was a “negative” clinical trial big enough? How many patients you needed depends on what you found. *Arch. Intern. Med.* **145:** 709, 1985.
2. Livingston, E. H. The mean and standard deviation: What does it all mean? *J. Surg. Res.* **119:** 117, 2004.
3. Livingston, E. H. Who was student and why do we care so much about his t-test? *J. Surg. Res.* **118:** 58, 2004.
4. Nickerson, R. S. Null hypothesis significance testing: A review of an old and continuing controversy. *Psychol. Methods* **5:** 241, 2000.
5. Hubbard, R. Blurring the distinctions between p’s and alpha’s in psychological research. *Theory Psychol.* **14:** 295, 2004.
6. Alder, H. L., and Roessler, E. B. *Introduction to Probability and Statistics.* San Francisco: W.H. Freeman Co., 1977.
7. Neyman, J. Frequentist probability and frequentist statistics. *Synthese* **36:** 97, 1977.
8. Plan and Operation of the Third National Health and Nutrition Examination Survey, 1988–1994. 1994. Hyattsville, MD, National Center for Health Statistics.