Journal of Diagnostic Medical Sonography

http://jdm.sagepub.com

# Statistics Refresher II: Choice of Sample Size

H. J. Khamis

*Journal of Diagnostic Medical Sonography* 1988; 4; 176
DOI: 10.1177/875647938800400402

The online version of this article can be found at:
http://jdm.sagepub.com/cgi/content/abstract/4/4/176

Published by:
SAGE
http://www.sagepublications.com

On behalf of:
SDMS
Society of Diagnostic Medical Sonography

*JDMS* 4:176–183, July/August 1988

# Statistics Refresher II

## Choice of Sample Size

H. J. KHAMIS, PHD*

**One of the important considerations in the planning of any experiment involving statistical analysis of data is the sample size to be used. The appropriate sample size for a given experiment depends upon the hypotheses being tested, the kind of statistical test being applied, the degree of accuracy required by the researcher, and other factors. Generally, it is not appropriate to make an intuitive guess at what the appropriate sample size should be; nor is it always true that "the larger the sample size, the better."**

**Key words: Level of significance, power, type I error, type II error, alpha, beta.**

One of the most frequent questions received by the Statistical Consulting Center at Wright State University is: "How large a sample do I need to take?" The answer to this question carries with it important consequences, especially if the experimentation involves human or animal subjects. For instance, as a member of the Laboratory Animals Utilization Committee, the author has constantly emphasized the importance of the sample size question in experimentation with animals; that using too many animals or using too few animals is ill advised. The recommendation is to use just that number of animals that will provide statistically valid results. The subjugation of any more animals to experimentation would increase the accuracy of the resulting scientific conclusions beyond that which is practically useful or needed and, hence, would be morally unjustified and unacceptable. The subjugation of any fewer animals to experimentation would fail to achieve the level of statistical validity required for confidence in the resulting scientific conclusions, thereby negating the effect of those animals used in the experiment; this, too, is unacceptable. Of course, these remarks also can be made about experiments concerning human subjects.

The use of the phrase "statistical validity" in the above discussion needs to be carefully explained. This will be done in the following section where the general technique of sample size determination is discussed. Then, sample size formulas will be presented for each of the most common hypothesis tests. Finally, examples will be provided in order to illustrate the use of the formulas. It might be helpful to review the JDMS article by Khamis[1] before reading on. In each of the hypothesis tests discussed below, the way in which the test is conducted, and the derivation of the P value, can be determined by referring to a basic statistics textbook.

From the *Statistical Consulting Center, Department of Mathematics and Statistics, Wright State University Dayton, Ohio.

Correspondence: H. J. Khamis, PhD, Statistical Consulting Center, Department of Mathematics and Statistics, Wright State University, Dayton, OH 45435.

## SAMPLE SIZE DETERMINATION

Generally, there are two kinds of sample size considerations to be discussed. The first considera-

tion concerns the level of significance that the researcher has chosen for his/her test of hypothesis. Recall from Khamis[1] that the level of significance for a test of hypothesis is a value chosen by the experimenter so that if the $P$ value of the test of hypothesis is smaller, then the null hypothesis is rejected, and if the $P$ value is larger, then the null hypothesis is not rejected. The level of significance is denoted by the Greek letter $\alpha$ (alpha). Hence, we have the following rule:

$P$-value $< \alpha \Rightarrow$ reject the null hypothesis
$P$-value $\geq \alpha \Rightarrow$ do not reject the null hypothesis

The definition of the level of significance, $\alpha$, is that it is the probability of making the mistake of rejecting the null hypothesis when it is actually true. Such a mistake is called the type I error, and $\alpha$ is called the type I error rate. Of course, we want the probability of making a type I error, $\alpha$, to be as small as possible. Most researchers set $\alpha$ to be 0.05, so that there is at most a 5% chance of making a type I error in conducting their test of hypothesis. For each kind of hypothesis test, there is a certain sample size needed in order to insure that the type I error rate does not exceed $\alpha$. If this is the only error rate that the researcher is interested in maintaining, then the test is said to be statistically valid when this sample size is achieved. The sample size needed for statistical validity (in the case of type I error protection) in each of the simpler, commonly used hypothesis tests will be provided below.

The second sample size consideration concerns the only other kind of error that can be made in a test of hypothesis, the type II error. Recall that in a test of hypothesis, the null hypothesis specifies a particular value for a parameter being tested. The alternative hypothesis specifies that the value of the parameter is something other than that specified by the null hypothesis. Of course, if the true value of the parameter differs from the value specified by the null hypothesis, then we would want to reject the null hypothesis. Failure to reject the null hypothesis when it is actually false is a mistake referred to as the type II error. The probability of committing a type II error in a test of hypothesis is denoted by the Greek letter $\beta$ (beta), called the type II error rate. Now, the calculation of $\beta$ depends upon how much the true parameter value differs from the value specified under the null hypothesis. Suppose we denote by **d** the difference between the true value of the parameter and the value specified under the null hypothesis. Note that this value is not known to the experimenter. If the value of **d** is large, then you would certainly want to reject the null hypothesis; i.e., you would want the probability of a type II error, $\beta$, to be small. The probability of detecting a difference as large as **d** (i.e., the probability of rejecting the null hypothesis when the true parameter value differs from the value specified under the null hypothesis by as much as **d**) is $1 - \beta$; this is called the *power* of the test.

Hence, the experimenter may decide on a value of **d** such that if the true parameter value deviates from the value specified under the null hypothesis any more than **d**, then a rejection of the null hypothesis occurs with high probability, $1 - \beta$. The value typically chosen for $1 - \beta$ by researchers ranges anywhere from 0.75 to 0.95 (so that $\beta$ ranges between 0.05 and 0.25).

The determination of the sample size, **n**, which maintains a type II error rate, $\beta$, for a given deviation, **d**, is generally quite complicated; formulas exist only for the simpler hypothesis tests. Typically, when a researcher wishes to conduct a test of hypothesis, he or she determines that sample size which maintains the type I error rate at $\alpha$. If this is the only error against which it is important to guard, then this sample size provides statistical validity. However, if there is a value of **d** for which it is particularly important to the researcher that the null hypothesis be rejected, then that sample size that maintains the type II error rate at $\beta$ for that value of **d** is computed, and it becomes the sample size required for statistical validity. The following example is presented in order to bring some of these concepts and details into focus.

## EXAMPLE: IS THE COIN FAIR OR NOT?

Recall that in Khamis,[1] we dealt with the problem of deciding whether a particular coin was fair. Statistically, the problem is posed in terms of a null hypothesis, that the proportion of heads occurring in a large number of tosses (theoretically, an infinite number of tosses) of the coin is 1/2, and an alternative hypothesis, that the proportion of heads is not 1/2. Notationally, the hypotheses are written:

$H_O$: $P = 1/2$ (null hypothesis)
$H_A$: $P = 1/2$ (alternative hypothesis)

where $P$ is the true proportion of heads for the coin being tested. Note that the null hypothesis, $H_O$, is a claim that the coin is fair, and the alternative hypothesis, $H_A$, is a claim that the coin is biased.

Now, suppose we set $\alpha = 0.05$, so that the probability of incorrectly rejecting $H_O$ (type I error rate) is at most 5%. This means we want our test procedure to guarantee that there is no more than a 5% chance of rejecting a true null hypothesis, and hence, wrongfully accusing Slim (who owns the coin) of offering the use of a biased coin. We now need to know what sample size is needed in order to maintain the type I error rate at $\alpha = .05$. A simple but conservative rule for the required sample size is that **n** must satisfy

$$n \geq 5/P_O(1 - P_O),$$

where $P_O$ is the value of $P$ specified under the null hypothesis, which in this example is $P_O = 1/2$. In this case, $n \geq 20$. So, if the coin is tossed at least 20 times, then the test of hypothesis ensures that there is at most a 5% type I error rate.

We now have the sample size required in order for our test procedure to protect against the mistake of rejecting a true null hypothesis, but we have no protection against the mistake of failing to reject a false null hypothesis (concluding that the coin is fair when it really isn't). Suppose we decide that if the true proportion of heads differs from 1/2 (the proportion specified under the null hypothesis) by more than 1/5, then the null hypothesis should be rejected with a high probability, say 90%. That is, we want to detect a difference of 1/5 with 90% probability ($d = 1/5$ and $1 - \beta = 0.9$)

A conservative formula that provides the minimum sample size required to satisfy these conditions is

$$n = (Z_{\alpha/2} + Z_\beta)^2/4d^2,$$

where $Z_k$ is the $(1 - k)$ ·100th percentile of the *standard normal distribution*. Percentiles of the standard normal distribution are provided in most every statistics textbook; for the most common values of **k** these percentiles are provided in Table 1.

TABLE 1. (1-k) ·100th Percentiles of the Standard Normal Distribution for Common Values of k

| k | 0.01 | 0.025 | 0.05 | 0.10 | 0.15 | 0.20 | 0.25 |
|---|---|---|---|---|---|---|---|
| $Z_k$ | 2.326 | 1.960 | 1.645 | 1.282 | 1.036 | 0.842 | 0.675 |

For our example, we have $\alpha = 0.05$, $\beta = 0.10$, and $d = 1/5$. From Table 1, $Z_{\alpha/2} = Z_{0.025} = 1.96$ and $Z_\beta = Z_{0.1} = 1.282$. Substituting these values into the above formula we get $n = 66$, after rounding up. Consequently, the test of hypothesis based on 66 tosses of Slim's coin is designed to ensure that there is, at most, a 5% chance of committing a type I error, and, at most, a 10% chance of committing a type II error when the true proportion of heads for the coin differs from 1/2 by more than 1/5. That is, there is, at most, a 5% chance that you will wrongly accuse Slim of using a biased coin, and, at most, a 10% chance that you will conclude that the coin is fair ($P = 1/2$) when in fact $P < 0.3$ or $P > 0.7$.

Note in the above sample size formula that if **d** is reduced, **n** becomes larger. For instance, if we wish to detect a difference of 1/10 between the true proportion of heads and the hypothesized proportion with probability 0.90, then we get $n = 263$. This should make good intuitive sense—if you want to detect a smaller difference with the same high probability, you need to work harder; namely, take a larger sample. Also note that if $\alpha$ or $\beta$ or both are reduced then **n** increases ($Z_k$ increases as **k** decreases [Table 1]). Again, this is intuitively reasonable: in order to reduce the probability of making an error (type I or type II), a larger sample size must be taken. Generally speaking, the purpose of using a large sample size in the first place is to reduce the chance of making an incorrect decision in the test of hypothesis. Sample size formulas will now be given for some of the commonly occurring hypothesis tests.

## SAMPLE SIZE FORMULAS

Each of the hypothesis tests discussed in Khamis[1] will now be looked at from the point of view of sample size considerations. In each case, the sample size required for keeping the type I error rate below $\alpha$ will be given. Where possible, the sample size required for keeping the type II error rate below $\beta$ for a specified value of **d** will be given.

### Test of Hypothesis Concerning a Proportion

The conservative rule for determining the sample size required for maintaining the type I error rate at $\alpha$ is that **n** satisfy

$$n \geq 5/P_O(1-P_O).$$

When this sample size is met, then the "large sample test for a binomial proportion" is a statistically valid test. This test is based on normal distribution theory and is presented in basic statistics textbooks.

For example, suppose that an advertisement claims that 80% of the cases of abdominal or muscular involvement in pulmonary tuberculosis can be successfully delineated by ultrasonography. In order to test this claim, several patients having this condition will be evaluated with ultrasonography and the number of correct diagnoses will be determined. The hypotheses being tested are:

$$H_O: P = 0.8$$
$$H_A: P \neq 0.8$$

where $P$ is the true proportion of correct diagnoses resulting from ultrasonography. In this example, $P_O = 0.8$, and the minimum sample size required for a statistically valid test of hypothesis is $n = 5/(0.8)(0.2) = 31.25$, or $n = 32$ after rounding up. Therefore, the chance of making a type I error in conducting this test of hypothesis will not exceed $\alpha$ provided $n \geq 32$.

The conservative rule for the sample size required in order to detect a difference of **d** between the true proportion and the hypothesized proportion with probability $1 - \beta$ is

$$n = (Z_{\alpha/2} + Z_\beta)^2/4d^2.$$

In the above example, if we wish to detect a one-tenth difference between $P$ and 0.8 with 80% probability when the level of significance is $\alpha = 0.1$ ($d = 1/10$, $\beta = 0.2$ and $\alpha = 0.1$), then $n = (1.645 + 0.842)^2/4\cdot(1/10)^2 = 154.6$, and $n = 155$ subjects would need to be sampled.

The above test is referred to as a *two-tail test of hypothesis*. An *upper-tail test of hypothesis* is a test in which the null hypothesis is rejected only if there is sufficient evidence to indicate that the true parameter value exceeds the value specified under the null hypothesis; for example,

$$H_O: P = 0.8$$
$$H_A: P > 0.8.$$

A *lower-tail test of hypothesis* involves a rejection of the null hypothesis when the true parameter value is less than the null hypothesis value; for example,

$$H_O: P = 0.8$$
$$H_A: P < 0.8.$$

In both of the latter two instances the test is referred to as a *one-tail test of hypothesis*.

For a one-tail test of hypothesis, the sample size requirement for type II error protection becomes

$$n = (Z_\alpha + Z_\beta)^2\, 4d^2.$$

In the above example, we might be interested in a lower-tail test because we wish to know if there is sufficient evidence to show that $P$ is less than the claimed 0.8. In fact, suppose that if $P < 0.7$ then we want to reject $H_O$ with 80% probability. The sample size required to satisfy these conditions is

$$n = (1.282 + 0.842)^2/4\cdot(1/10)^2 = 112.8,$$
or $n = 113$ after rounding up.

In terms of the type I error rate, if the sample size required for statistical validity, $n \geq 5/P_O(1 - P_O)$, cannot be met, then a different test of hypothesis must be used, one involving the binomial probability function. This "small-sample test for a binomial proportion" is described in some statistics textbooks, such as Chapter 7 in Rosner.[2]

## Test of Hypothesis Concerning a Mean

In this case it is desired to test that the mean, denoted by the Greek letter $\mu$ (mu), of some random variable under study is equal to a hypothesized value, say $\mu_O$. Notationally, the two-tail set of hypotheses would be written

$$H_O: \mu = \mu_O$$
$$H_A: \mu \neq \mu_O$$

If the sample size is at least 30,

$$n \geq 30,$$

then the "large-sample test for a population mean," or "z-test," can be conducted, and the type I error rate will not exceed $\alpha$. The sample size required in order to detect a difference of **d** units between the true mean of the variable, $\mu$, and the hypothesized mean, $\mu_O$, with probability $1 - \beta$ is

$$n \geq (Z_{\alpha/2} + Z_{\beta})^2\sigma^2/d^2,$$

where the Greek letter $\sigma$ (sigma) is the *standard deviation* of the variable under study, and the square of $\sigma$, $\sigma^2$, is the *variance*.

In order to use this formula, we need to obtain an estimate of $\sigma^2$. One way to do this is to draw a random sample of size $n_1 \geq 30$ and compute the *sample variance* $s^2$,

$$s^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2/(n-1).$$

Then $s^2$ can be used to estimate $\sigma^2$, and the required sample size, **n**, can be computed from the above formula. Since $n_1 \geq 30$ subjects have already been sampled, only $n - n_1$ more subjects need to be sampled, for $n > n_1$. Another way to estimate $\sigma^2$ that does not require the drawing of a sample is the following. Estimate the largest value that the study variable can be, $x_{max}$, and estimate the smallest value that the study variable can be, say $x_{min}$. Then estimate $\sigma^2$ by

$$\sigma^2 \approx (x_{max} - x_{min})^2/16.$$

In a one-tail hypothesis test, the sample size required to detect a deviation of **d** units with probability $1 - \beta$ is

$$n \geq (Z_{\alpha} + Z_{\beta})^2\sigma^2/d^2.$$

If the sample size, **n**, is smaller than 30, then the z-test is no longer a statistically valid test. However, if the sample comes from a set of data that are bell-shaped (normally distributed), then the "t-test" is statistically valid. That is, the t-test controls for the type I error rate in this case.

Finally, if the sample size, **n**, is smaller than 30 *and* the data do not come from data that are bell-shaped (nonnormal data), then "nonparametric test procedures" must be used. Most basic statistics textbooks include a chapter on nonparametric procedures, *eg*, Rosner,[2] Chapter 9.

Consider the hypothesis that the mean velocity of umbilical venous blood flow for pregnant women in their 33rd week is 15 cm/sec:

$$H_O: \mu = 15$$
$$H_A: \mu \neq 15$$

where $\mu$ represents the true mean velocity. A type I error rate of, say $\alpha = 0.05$, is ensured if a random sample of size at least 30 is used. Suppose we wish to detect a 1 cm/sec difference in velocity from the hypothesized value of $\mu = 15$ with 95% probability. A random sample of $n_1 = 30$ normal pregnant women rendered a sample standard deviation of $s = 2.1$ (see Chen, et al.[3]). So, we have $\alpha = 0.05$, $\beta = 0.05$, $d = 1$ and $\sigma \approx 2.1$. Then the required sample size is $n \geq (1.960 + 1.645)^2 \cdot (2.1)^2/1 = 57.3$, or $n = 58$ is the smallest number of subjects that need to be sampled in order to satisfy the above conditions. Since $n_1 = 30$ subjects have already been sampled, only 28 more subjects need to be sampled.

## Test of Hypothesis Concerning a Variance

Here we are interested in testing that the variance of the study variable, $\sigma^2$, is equal to some hypothesized value, $\sigma_O^2$:

$$H_O: \sigma^2 = \sigma_o^2$$
$$H_A: \sigma^2 \neq \sigma_o^2$$

If the data come from a normal distribution then the chi-squared test is a statistically valid test for any sample size $n \geq 2$. If the data do not come from a normal distribution, then the sample size must be at least 30,

$$n \geq 30,$$

in which case there is an approximate test that is statistically valid e.g., Miller and Freund,[4] Chapter 8).

In order to determine the sample size required to detect a deviation between $\sigma$ and $\sigma_O$ of, say, $\frac{\sigma}{\sigma_o} = d$ units (note that here the deviation is expressed in terms of a ratio between the true and hypothesized

standard deviations) with probability $1 - \beta$, one must refer to a table such as that provided in Gill[5] (Figures A.15.7–A.15.12). These figures contain *operating characteristic* (OC) curves which, for given one-sided and two-sided hypothesis tests and given values of $\alpha$, $\beta$, and **d**, provide the necessary sample size, **n**. For those cases, such as the chi-squared test, where a convenient sample size formula is not available for type II error rate protection, the OC curve is the next best resource.

### Test of Hypothesis Concerning Two Means

We are interested in testing whether two means, say $\mu_1$, and $\mu_2$, are the same:

$$H_O: \mu_1 - \mu_2 = 0$$
$$H_A: \mu_1 - \mu_2 \neq 0$$

A random sample is taken for each of the two variables under study. If the sample size in each case is at least 30 ($n_1 \geq 30$ and $n_2 \geq 30$) then the "large-sample test for the difference between two means," sometimes called the "two-sample z-test" is used.

For the case $n_1 = n_2$, the sample size required in order to detect a difference between $\mu_1$ and $\mu_2$, of **d** units or more with probability $1 - \beta$ is

$$n \geq (Z_{\alpha/2} + Z_\beta)^2(\sigma_1^2 + \sigma_2^2)/d^2.$$

For the one-tail test we have

$$n \geq (Z_\alpha + Z_\beta)^2(\sigma_1^2 + \sigma_2^2)/d^2.$$

The variances, $\sigma_1^2$ and $\sigma_2^2$, need to be estimated in order to use these formulas, just as for the one-sample case above.

For the case in which $n_1 < 30$ or $n_2 < 30$, but for which the data come from normal populations, other techniques must be used. If the variances can be determined to be approximately equal ($\sigma_1^2 \approx \sigma_2^2$), then the "pooled-sample t-test," sometimes called the "two-sample t-test" is used. If the variances are not equal, an approximate test procedure called the "Satterthwaite procedure" is used. Of course, the correct procedure must be used in order to ensure that the type I error rate does not exceed $\alpha$. In the case of small samples ($n_1 < 30$ or $n_2 < 30$) there are tables available for determining the correct sample size to insure a given type II error rate (e.g., Gill,[5] section 1.6.6).

Finally, if $n_1 < 30$ or $n_2 < 30$ and the data do not come from normal populations, then nonparametric procedures must be used.

### Test of Hypothesis Concerning Correlation

We wish to test that the correlation coefficient denoted by the Greek letter $\rho$ (rho) is equal to some hypothesized value, say $\rho_O$:

$$H_O: \rho = \rho_o$$
$$H_A: \rho \neq \rho_o.$$

If the sample size is $n \geq 4$ and both random variables correspond to data from normal distributions, then the test based on the "Fisher Z-transformation" provides a statistically valid test (e.g., Rosner,[2] Chapter 11). If at least one of the random variables does not correspond to data coming from a normal distribution then a nonparametric procedure must be used.

### Test of Hypothesis Concerning Two Proportions

Consider a proposal to replace a standard diagnostic modality (such as angiography) for studying vascular disease with color flow doppler. In order to compare the two diagnostic modalities statistically, we have the following set of hypotheses:

$$H_O: P_1 - P_2 = O$$
$$H_A: P_1 - P_2 \neq O.$$

where $P_1$ = proportion of correct diagnoses using angiography, and $P_2$ = proportion of correct diagnoses using color flow doppler.

The data are collected by obtaining a random sample of, say, $n_1$ subjects who have the condition being studied and using angiography for identification of the condition, and independently obtaining a random sample of $n_2$ subjects with the condition and using color flow doppler.

The layout of the data will appear in a $2 \times 2$ table, called a contingency table, as shown in Table 2.

In Table 2, **a** and **c** are the numbers of subjects correctly diagnosed by angiography and color flow doppler, respectively, and **b** and **d** are the numbers of subjects incorrectly diagnosed by angiography

TABLE 2. 2 × 2 Contingency Table

| Diagnostic Modality | Diagnosis | | |
|---|---|---|---|
| | Correct | Incorrect | Sample Size |
| Angiography | a | b | $n_1$ |
| Color flow doppler | c | d | $n_2$ |

and color flow doppler, respectively. Note that $n_1 = a + b$, $n_2 = c + d$, $P_1$ is estimated by $a/n_1$ and $P_2$ is estimated by $c/n_2$. The question, then, is how large should the sample be?

A sample but conservative sample size rule that maintains the type I error rate at $\alpha$ is that each frequency count be at least five:

$$a \geq 5, b \geq 5, c \geq 5 \text{ and } d \geq 5$$

If this sample size requirement is met then the "chi-squared test for equality of proportions" is a statistically valid test. This test appears in basic statistics textbooks, *eg*, Rosner,[2] Chapter 10.

Suppose it is decided to replace angiography with color flow doppler as the primary modality for studying this condition only if the success rate of color flow doppler, $p_2$, is at least 15 percentage points higher than that for angiography, $p_1$. Suppose further that it is known that the success rate of angiography is approximately 70%; $p_1 \approx 0.7$. Then we wish to reject $H_O$ with high probability, $1 - \beta$, if $p_2 \geq 0.85$. A recommendation that is made for *clinical trials* of this type is to set $\beta$ to $4\alpha$ (see Fleiss,[6] section 3.3). So, if we use $\alpha = 0.05$, then $\beta = 0.20$ and $1 - \beta = 0.80$. Then, using the values $\alpha = 0.05$, $1 - \beta = 0.80$, $p_1 = 0.70$ and $p_2 = 0.85$, Table A.3 in Fleiss[6] can be used to obtain the sample sizes required to satisfy the above conditions for the case $n_1 = n_2$. In this example, we get $n_1 = n_2 = 134$. So, 268 subjects must be randomly sampled, 134 are randomly assigned to color flow doppler and the remainder to angiography.

## CONCLUSION

Recall that in the five-step procedure for a test of hypothesis presented in Khamis,[1] the first step is to set up the null and alternative hypotheses. The second step is to choose the type I error rate, $\alpha$, that you wish the test to ensure. In the conduct of the test a *P* value is derived, and the decision about whether the null hypothesis should be rejected is based on the following rule (step 5):

$$P \text{ value} < \alpha \Rightarrow \text{reject } H_O$$
$$P \text{ value} \geq \alpha \Rightarrow \text{do not reject } H_O.$$

For each of the tests discussed above, the conduct of the test and the derivation of the *P* value (step 4) are provided in most all basic statistics books.

Step 3 is the collection of the data. We may now insert a step between steps 2 and 3; say, step 2.5: determine the sample size needed in order to make the test statistically valid. In particular, the sample size must be large enough so that the type I error rate does not exceed $\alpha$; and, in those cases where it is of importance, the sample size must be large enough so that the type II error rate does not exceed $\beta$ for specific alternative parameter values of interest. On the other hand, assuming that the investigator wishes to detect only differences that are of practical importance, and not merely differences of any magnitude, he or she should not employ sample sizes that are larger than that needed to guard against the type II error.

Notice in the above treatment that there is more emphasis on protection against the type I error than on protection against the type II error. The reason for this is that the null and alternative hypotheses are typically designed so that the type I error is a more serious error than the type II error. For any test of hypothesis, the sample size for type I error protection must be achieved for statistical validity. If this is the only sample size that is considered (i.e., type II error protection is not considered), then a rejection of $H_O$ after conducting the test of hypothesis means that you can conclude with a high degree of confidence that $H_O$ is false, but a failure to reject $H_O$ means that no conclusion can be made regarding the truth or falsehood of $H_O$ based on the data.

Many of the sample size requirements discussed above are based on the famous *Central Limit Theorem*. For many hypothesis tests, when the sample size exceeds 30, certain statistics involved in the test are approximately normally distributed according to the Central Limit Theorem, and hence normal distribution theory can be used to obtain the appropriate *P* value for the test.

Most of the sample size formulas given above are conservative in the sense that the statistical validity of the test will be maintained even with somewhat

smaller sample sizes, depending upon certain characteristics of the data. For more details about this, it is recommended that you consult a statistician. In fact, it is generally recommended that you go over your research protocol, including sample size considerations, with a statistical consultant before collecting any data.

## ACKNOWLEDGMENT

The author wishes to thank Roger Warner, RDMS, of Miami Valley Hospital, Dayton, Ohio for helpful comments and suggestions in the writing of the manuscript.

## REFERENCES

1. Khamis HJ: Statistics Refresher: tests of hypothesis and diagnostic test evaluation. *JDMS* 1987;3(3):123–129.
2. Rosner B: *Fundamentals of Biostatistics*, 2nd edition. Boston, MA: Duxbury Press, 1986.
3. Chen H-Y, Lu C-C, Cheng Y-T, Hsieh F-J, Lin J-Y: Antenatal measurement of fetal umbilical venous flow by pulsed Doppler and B-mode ultrasonography. *J Ultrasound Med* 1986;5(6):319–321.
4. Miller I, Freund FE: *Probability and Statistics for Engineers*, 2nd edition. Englewood Cliffs, NJ: Prentice-Hall, 1977.
5. Gill JL: *Design and Analysis of Experiments in the Animal and Medical Sciences*. Ames, IA: The Iowa State University Press, 1978.
6. Fleiss JL: *Statistical Methods for Rates and Proportions*, 2nd edition. New York: John Wiley and Sons, 1981.