1-1991

# A simple, rapid and reliable method for selecting or assessing the number of replicates for animal experiments

William E. Berndtson
*University of New Hampshire*, Bill.Berndtson@unh.edu

# A simple, rapid and reliable method for selecting or assessing the number of replicates for animal experiments

# A SIMPLE, RAPID AND RELIABLE METHOD FOR SELECTING OR ASSESSING THE NUMBER OF REPLICATES FOR ANIMAL EXPERIMENTS[1]

W. E. Berndtson

University of New Hampshire[2], Durham 03824

**ABSTRACT**

A simple approach was developed for determining the number of replicates needed per treatment group to provide experiments of known power and sensitivity, where power equals the probability that a treatment effect would not go undetected if an effect existed and sensitivity equals the minimal treatment response that will be detectable. This approach, in turn, was used to construct reference tables, applicable across scientific disciplines, from which researchers may read replication requirements directly with ease, speed and reliability. To use the tables, one need only furnish a reliable estimate of the coefficient of variability expected among replicates, which may be obtained from prior observations on similar populations. The tabular data also enable a rapid, reliable assessment of the actual power and sensitivity of completed experiments, such as those contained within the published literature.

Key Words: Experimental Design, Replication, Variance, Heterogeneity, Animal Research

J. Anim. Sci. 1991. 69:67–76

## Introduction

A number of procedures have been developed that enable researchers to estimate in advance the replication needed to provide an experiment of any chosen power and sensitivity (Tang, 1938; Cochran and Cox, 1957; Harter, 1957; Steel and Torrie, 1960; Gill, 1978, 1989; Remington and Schork, 1985; Berndtson et al., 1989). Power is defined herein as the probability that a treatment effect will not go undetected if it exists, and sensitivity equals the smallest treatment response that will be detectable. Each approach requires an estimate of expected variance among replicates, which typically is derived from similar, completed studies.

Although these approaches can be very useful, they have been utilized infrequently. Most researchers appear to choose replication arbitrarily based on the cost or availability of replicates, convenience or tradition. Many researchers are unaware of the potential shortcomings when experiments are arbitrarily replicated or, alternatively, may either be unfamiliar with established procedures for estimating replication requirements or find them too inconvenient for routine use. The present effort was undertaken to develop a simple, rapid and reliable procedure for estimating the replication needed to provide experiments of any chosen power and sensitivity and(or) for determining the power and sensitivity of completed research that would be applicable across species and scientific disciplines.

## Methods

The procedure used to determine replication requirements is an extension of that of Tang (1938), as cited by Steel and Torrie (1960). Tang's approach is represented by the following equation:

$$R \geq 2(t_0 + t_1)^2 \ s^2/d^2$$

[1]Scientific Contribution No. 1674 from the New Hampshire Agric. Exp. Sta.

[2]Dept. of Anim. and Nutr. Sci.

Received April 23, 1990.
Accepted July 6, 1990.

where R equals the number of replicates needed per treatment group, $t_0$ equals the Student's *t*-value associated with Type I error, $t_1$ represents the Student's *t*-value associated with Type II error ($t_1$ equals the tabular *t*-value for the probability 2[1–*P*], where *P* represents the probability of detecting a difference of stated size if it exists), $s^2$ equals the error mean square from an actual experiment, and d represents the difference between treatments one wishes to be able to detect (in actual measurement units).

Several modifications of Tang's approach were introduced. The most important was the substitution of the simple variance among control subjects for the error mean square in the equation. This was a valid change that had been used previously for estimating replication requirements from single, untreated, animal population data (Berndtson and Thompson, 1990). The use of the simple control variance was desirable for several reasons. First, the suggestion that the error mean square be derived from completed experiments (Steel and Torrie, 1960) is potentially misleading; it could be regarded as implying that an experiment involving the same or similar treatments must be completed. Actually, the error mean square simply provides an estimate of the variance among replicates within treatments (i.e., variability among replicates after that attributable to treatments, measurements within replicates, etc., has been partitioned out). Although treatments can and often do alter group means, most statistical analyses (such as the analysis of variance) require that all treatment groups share a common variance (i.e., although group means may differ, the actual variances must be similar; Steel and Torrie, 1960). Because data must satisfy the assumption of a common variance, the variance among control subjects provides a valid estimate for use in Tang's equation. (Note: procedures for transforming data, etc., when the assumption is not validated are discussed elsewhere; Steel and Torrie, 1960.) In addition to eliminating the potential, perceived need for a nearly identical completed study, this modification has rendered estimates of among-replicate variance much more accessible. For example, whereas error mean squares are rarely cited within the published literature, the variance within control populations can be calculated rapidly from the standard deviation or standard error of control means, which usually are published.

The second modification consisted of transforming Tang's equation to read as follows:

$$s^2/d^2 \leq R/[2(t_0 + t_1)^2]$$

so that those terms normally taken from completed studies appeared to the left. It should be noted that the $s^2/d^2$ ratio reflects the size of the response one wishes to detect (d) in relation to the normal variance ($s^2$) or standard deviation (s) of the population, and that this ratio may be determined either with actual or relative values. For example, if one wished to detect a treatment difference equal to one standard deviation from the control mean, the $s^2/d^2$ ratio would equal 1.0. In this example, a ratio of 1.0 would be obtained whether values of s and d were expressed in absolute units (grams, millimeters, etc.) or as a percentage of the control mean. Similarly, if one wished to detect a treatment response equal to one-half of one standard deviation of the control mean, the $s^2/d^2$ ratio would equal 4.0. Accordingly, it was possible to calculate the $s^2/d^2$ ratios that would apply for experiments of any stated population variance and desired sensitivity; $s^2/d^2$ ratios were calculated for experimental populations for which the coefficient of variability (CV or standard deviation as a percentage of the mean) ranged from 1 to 100% and for which it was desirable to detect treatment responses ranging from 5 to 100% of the control mean.

Next, R values appropriate for each $s^2/d^2$ ratio were determined empirically. Successive values of R and the associated values of $t_0$ and $t_1$ were substituted into the equation until the minimal value of R that satisfied the equation was found. Note that for this exercise it was necessary to first select the number of treatment groups (i.e., levels of treatment in the planned study) to be provided, because this influences the degrees of freedom associated with the values of *t*. It also was necessary to select the desired Type I and Type II error probabilities. For the computations herein, experiments were chosen with two treatments, a Type I error probability of $P < .05$ and experimental power (related to Type II error) of 80, 90 and 95%.

From the calculations, tables were constructed from which replication requirements may be read directly for experiments with any sample population coefficient of variability and experimental sensitivity combination. Al-

though animals are used as the units of replication in all illustrations to follow, it should be noted that replication frequently takes other forms (e.g., as during in vitro studies, when pens or groups of animals are used as the experimental unit, etc.) for which the tabular values are equally appropriate. A distinction must be made, nonetheless, between replicates vs measurements within replicates.

## Results and Discussion

### *Applications*

*Choosing Replication.* Tables 1 to 3 contain replication (animal numbers) requirements for experiments of 80, 90 or 95% power, respectively. To illustrate their use, assume that one wanted to assess the effect(s) of an experimental treatment on the daily sperm production (DSP) of young beef bulls. Assume further that the investigator would like to be at least 90% certain that a treatment causing a 10% change in DSP would be detected (i.e., would be declared statistically significant), if such a change existed. To use the tables in planning the experiment, an estimate of the CV for DSP (i.e., the characteristic or end point to be evaluated in the study) among young beef bulls would be needed. For this illustration, an estimate will be calculated from data for a group of 34 yearling Hereford or Angus bulls, for which the mean ± SE DSP was 3.79 ± .21 billion (Berndtson and Igboeli, 1989). Because the SE equals $\sqrt{s^2/n}$, in this example .21 = $\sqrt{s^2/34}$, and s = 1.22. The CV = $(s/\bar{x})100$. In this example, the CV = (1.22/3.79)100 = 32.2%. Interpolating from Table 2 we find that, for a population with a CV of 32.2%, 215 bulls would be needed per treatment group to provide a 90% chance that an existing 10% treatment response would be detected and declared statistically significant. A simple two-treatment experiment would require a total of approximately 430 bulls. Note that repeat

TABLE 1. REPLICATES NEEDED PER TREATMENT GROUP FOR EXPERIMENTS OF 80% POWER AT $P < .05$[a]

| | Difference from control to be detected, % | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| CV, % | 5 | 10 | 15 | 20 | 25 | 30 | 35 | 40 | 45 | 50 | 60 | 70 | 80 | 90 | 100 |
| 1 | 3 | 2 | | | | | | | | | | | | | |
| 2 | 4 | 3 | 2 | | | | | | | | | | | | |
| 3 | 7 | 3 | 3 | 2 | | | | | | | | | | | |
| 4 | 12 | 4 | 3 | 3 | 2 | | | | | | | | | | |
| 5 | 17 | 6 | 4 | 3 | 3 | 2 | | | | | | | | | |
| 6 | 24 | 7 | 4 | 3 | 3 | 3 | 2 | | | | | | | | |
| 7 | 32 | 9 | 5 | 4 | 3 | 3 | 3 | 2 | | | | | | | |
| 8 | 42 | 12 | 6 | 4 | 3 | 3 | 3 | 3 | 2 | | | | | | |
| 9 | 52 | 14 | 7 | 5 | 4 | 3 | 3 | 3 | 3 | 2 | | | | | |
| 10 | 63 | 17 | 9 | 6 | 4 | 4 | 3 | 3 | 3 | 3 | 2 | | | | |
| 12 | 91 | 24 | 12 | 7 | 5 | 4 | 4 | 3 | 3 | 3 | 3 | 2 | | | |
| 14 | 124 | 32 | 15 | 9 | 7 | 5 | 4 | 4 | 3 | 3 | 3 | 3 | 2 | | |
| 16 | 161 | 42 | 19 | 12 | 8 | 6 | 5 | 4 | 4 | 3 | 3 | 3 | 3 | 2 | |
| 18 | 204 | 52 | 24 | 14 | 10 | 7 | 6 | 5 | 4 | 4 | 3 | 3 | 3 | 3 | 2 |
| 20 | 252 | 63 | 29 | 17 | 12 | 9 | 7 | 6 | 5 | 4 | 4 | 3 | 3 | 3 | 3 |
| 25 | 393 | 99 | 45 | 26 | 17 | 12 | 10 | 8 | 6 | 6 | 4 | 4 | 3 | 3 | 3 |
| 30 | 566 | 142 | 63 | 37 | 24 | 17 | 13 | 10 | 9 | 7 | 6 | 5 | 4 | 4 | 3 |
| 35 | 770 | 193 | 86 | 50 | 32 | 23 | 17 | 14 | 11 | 9 | 7 | 6 | 5 | 4 | 4 |
| 40 | 1,005 | 252 | 112 | 63 | 42 | 29 | 22 | 17 | 14 | 12 | 9 | 7 | 6 | 5 | 4 |
| 45 | 1,272 | 318 | 142 | 80 | 52 | 37 | 27 | 21 | 17 | 14 | 10 | 8 | 7 | 6 | 5 |
| 50 | 1,571 | 393 | 175 | 99 | 63 | 45 | 34 | 26 | 21 | 17 | 12 | 10 | 8 | 6 | 6 |
| 60 | 2,262 | 566 | 252 | 142 | 91 | 63 | 48 | 37 | 29 | 24 | 17 | 13 | 10 | 9 | 7 |
| 70 | 3,078 | 770 | 342 | 193 | 124 | 86 | 63 | 50 | 40 | 32 | 23 | 17 | 14 | 11 | 9 |
| 80 | 4,020 | 1,005 | 447 | 252 | 161 | 112 | 83 | 63 | 51 | 42 | 29 | 22 | 17 | 14 | 12 |
| 90 | 5,088 | 1,272 | 566 | 318 | 204 | 142 | 104 | 80 | 63 | 52 | 37 | 27 | 21 | 17 | 14 |
| 100 | 6,281 | 1,571 | 698 | 393 | 252 | 175 | 129 | 99 | 78 | 63 | 45 | 34 | 26 | 21 | 17 |

[a]For two-tailed tests with two-treatment experiments. For experiments with a one-tiled test, the replication shown would provide an experiment of 90% power at $P < .025$.

measurements on individual animals do not constitute replicates.

*Controlling Type II Error Probability.* A study involving 430 bulls would be extremely uncommon. Yet, the SE of .21 billion represented only 5.5% of the mean (3.79 billion) for this characteristic. Because such variability is not uncommon within biological disciplines, it is likely that the power and sensitivity of many experiments have been overestimated. In this regard, most investigators appear very cautious in avoiding Type I error, which results when an investigator erroneously concludes that a treatment had an effect; most require less than a 5% probability of Type I error (i.e., $P < .05$) when declaring a treatment effect. In contrast, if one cannot maintain less than a 5% chance of error, it is customary to regard the treatment(s) as non-significant ($P > .05$). The latter usually is interpreted as evidence that the treatment was without effect, although one rarely knows the probability that an actual treatment effect may have been missed (i.e., the probability of Type II error). Note that a statement that a treatment was without effect ($P > .05$) does not quantify Type II error, but it simply indicates that the investigator could not be ≥ 95% certain that the treatment had an effect. Because a Type II error generally is not evaluated when data from a completed study are analyzed, one of the major applications of data in Tables 1 to 3 is to permit experimental power, which is a function of the Type II error probability, to be weighted during the planning of an experiment. With this information, researchers can design an experiment in which the probability of Type II error is restricted to a level they consider acceptable. Alternatively, researchers should, at a minimum, acquire a useful awareness of the potential for such error in the experiment as it is finally implemented.

*Assessing the Power and Sensitivity of Completed, Published Research.* A third application of the data in Tables 1 to 3 is for retrospective determinations of the power and sensitivity of other completed research, such as that within the published literature. Let us suppose that the hypothetical study proposed

TABLE 2. REPLICATES NEEDED PER TREATMENT GROUP FOR EXPERIMENTS OF 90% POWER AT $P < .05$[a]

| | Difference from control to be detected, % | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| CV, % | 5 | 10 | 15 | 20 | 25 | 30 | 35 | 40 | 45 | 50 | 60 | 70 | 80 | 90 | 100 |
| 1 | 3 | 2 | | | | | | | | | | | | | |
| 2 | 5 | 3 | 2 | | | | | | | | | | | | |
| 3 | 9 | 4 | 3 | 2 | | | | | | | | | | | |
| 4 | 15 | 5 | 4 | 3 | 2 | 2 | | | | | | | | | |
| 5 | 23 | 7 | 4 | 3 | 3 | 3 | 2 | | | | | | | | |
| 6 | 33 | 9 | 5 | 4 | 3 | 3 | 3 | 2 | | | | | | | |
| 7 | 43 | 12 | 6 | 4 | 3 | 3 | 3 | 3 | 2 | | | | | | |
| 8 | 55 | 15 | 8 | 5 | 4 | 3 | 3 | 3 | 3 | 2 | | | | | |
| 9 | 69 | 19 | 9 | 6 | 4 | 4 | 3 | 3 | 3 | 3 | 2 | | | | |
| 10 | 85 | 23 | 11 | 7 | 5 | 4 | 4 | 3 | 3 | 3 | 3 | 2 | | | |
| 12 | 122 | 32 | 15 | 9 | 6 | 5 | 4 | 4 | 3 | 3 | 3 | 3 | 2 | | |
| 14 | 165 | 43 | 20 | 12 | 8 | 6 | 5 | 4 | 4 | 3 | 3 | 3 | 3 | 2 | |
| 16 | 216 | 55 | 25 | 15 | 10 | 8 | 6 | 5 | 4 | 4 | 3 | 3 | 3 | 3 | 2 |
| 18 | 273 | 69 | 32 | 19 | 12 | 9 | 7 | 6 | 5 | 4 | 4 | 3 | 3 | 3 | 3 |
| 20 | 337 | 85 | 39 | 23 | 15 | 11 | 8 | 7 | 6 | 5 | 4 | 4 | 3 | 3 | 3 |
| 25 | 526 | 132 | 60 | 34 | 23 | 16 | 12 | 10 | 8 | 7 | 5 | 4 | 4 | 3 | 3 |
| 30 | 757 | 190 | 85 | 49 | 32 | 23 | 17 | 13 | 11 | 9 | 7 | 6 | 5 | 4 | 4 |
| 35 | 1,030 | 258 | 115 | 65 | 43 | 30 | 23 | 18 | 14 | 12 | 9 | 7 | 6 | 5 | 4 |
| 40 | 1,346 | 337 | 150 | 85 | 55 | 39 | 29 | 23 | 18 | 15 | 11 | 8 | 7 | 6 | 5 |
| 45 | 1,703 | 426 | 190 | 107 | 69 | 49 | 36 | 28 | 23 | 19 | 13 | 10 | 8 | 7 | 6 |
| 50 | 2,103 | 526 | 234 | 132 | 85 | 60 | 45 | 34 | 28 | 23 | 16 | 12 | 10 | 8 | 7 |
| 60 | 3,027 | 757 | 337 | 190 | 122 | 85 | 62 | 49 | 39 | 32 | 23 | 17 | 13 | 11 | 9 |
| 70 | 4,121 | 1,030 | 458 | 258 | 165 | 115 | 85 | 65 | 52 | 43 | 30 | 23 | 18 | 14 | 12 |
| 80 | 5,382 | 1,346 | 598 | 337 | 216 | 150 | 110 | 85 | 67 | 55 | 39 | 29 | 23 | 18 | 15 |
| 90 | 6,811 | 1,703 | 757 | 426 | 273 | 190 | 139 | 107 | 85 | 69 | 49 | 36 | 28 | 23 | 19 |
| 100 | 8,409 | 2,103 | 935 | 526 | 337 | 234 | 172 | 132 | 104 | 85 | 60 | 45 | 34 | 28 | 23 |

[a]For two-tailed tests with two-treatment experiments. For experiments with a one-tailed test, the replication shown would provide an experiment of 95% power at $P < .025$.

earlier had been conducted with 20 young beef bulls per treatment and that treatment effects on DSP were non-significant ($P > .05$). Because such a study would be well-replicated by conventional standards, the results might be regarded by many as definitive evidence that the experimental treatment was without effect on DSP. Assuming that the CV for DSP among bulls was 30% (a more accurate CV could be calculated from the actual data from this experiment), it is clear that a 25 to 30% treatment response would have been needed for 80% certainty of statistical significance with 20 bulls per treatment group (Table 1). Corresponding differences of 30 to 35 and 35 to 40% would have been needed for experiments of 90 and 95% power, respectively (Tables 2 and 3). From such determinations, it is obvious that rather substantial treatment responses could go undetected in an experiment. The non-significant finding in this example may be due to the true absence of a treatment response or may simply reflect inadequate experimental power and sensitivity. Lacking justification for favoring one of these possibilities over the other, such a finding must be regarded as inconclusive and should be interpreted with appropriate caution. From this example, it is clear that the power, sensitivity and potential for Type II errors in completed studies can be assessed with ease and reliability via the use of data in Tables 1 to 3, and the value of such evaluations should be evident.

### *Special Considerations*

*Number of Treatment Groups.* Application of the tabular data for the uses indicated requires several important considerations. First, to calculate the data within these tables it was assumed that the experiments would be conducted with only two treatment groups. If all other factors are held constant, the inclusion of additional treatment groups will increase error degrees of freedom. For studies with error degrees of freedom below infinity, this could decrease the number of replicates needed

TABLE 3. REPLICATES NEEDED PER TREATMENT GROUP FOR EXPERIMENTS OF 95% POWER AT $P < .05$[a]

| | Difference from control to be detected, % | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| CV, % | 5 | 10 | 15 | 20 | 25 | 30 | 35 | 40 | 45 | 50 | 60 | 70 | 80 | 90 | 100 |
| 1 | 3 | 3 | | | | | | | | | | | | | |
| 2 | 6 | 3 | 2 | 2 | | | | | | | | | | | |
| 3 | 11 | 4 | 3 | 3 | 2 | | | | | | | | | | |
| 4 | 18 | 6 | 4 | 3 | 3 | 2 | 2 | | | | | | | | |
| 5 | 28 | 8 | 5 | 4 | 3 | 3 | 3 | 2 | | | | | | | |
| 6 | 39 | 11 | 6 | 4 | 3 | 3 | 3 | 3 | 2 | 2 | | | | | |
| 7 | 53 | 14 | 7 | 5 | 4 | 3 | 3 | 3 | 3 | 3 | | | | | |
| 8 | 67 | 18 | 9 | 6 | 4 | 4 | 3 | 3 | 3 | 3 | 2 | | | | |
| 9 | 85 | 23 | 11 | 7 | 5 | 4 | 4 | 3 | 3 | 3 | 3 | 2 | | | |
| 10 | 104 | 28 | 13 | 8 | 6 | 5 | 4 | 4 | 3 | 3 | 3 | 3 | 2 | | |
| 12 | 150 | 39 | 18 | 11 | 8 | 6 | 5 | 4 | 4 | 3 | 3 | 3 | 3 | 2 | 2 |
| 14 | 204 | 53 | 24 | 14 | 10 | 7 | 6 | 5 | 4 | 4 | 3 | 3 | 3 | 3 | 3 |
| 16 | 267 | 67 | 31 | 18 | 12 | 9 | 7 | 6 | 5 | 4 | 4 | 3 | 3 | 3 | 3 |
| 18 | 337 | 85 | 39 | 23 | 15 | 11 | 9 | 7 | 6 | 5 | 4 | 4 | 3 | 3 | 3 |
| 20 | 416 | 104 | 48 | 28 | 18 | 13 | 10 | 8 | 7 | 6 | 5 | 4 | 4 | 3 | 3 |
| 25 | 650 | 163 | 73 | 42 | 28 | 20 | 15 | 12 | 10 | 8 | 6 | 5 | 4 | 4 | 4 |
| 30 | 936 | 234 | 104 | 60 | 39 | 28 | 21 | 16 | 13 | 11 | 8 | 6 | 5 | 5 | 4 |
| 35 | 1,274 | 319 | 142 | 80 | 53 | 37 | 28 | 22 | 17 | 14 | 11 | 8 | 7 | 6 | 5 |
| 40 | 1,664 | 416 | 185 | 104 | 67 | 48 | 36 | 28 | 22 | 18 | 13 | 10 | 8 | 7 | 6 |
| 45 | 2,106 | 527 | 234 | 132 | 85 | 60 | 45 | 35 | 28 | 23 | 16 | 12 | 10 | 8 | 7 |
| 50 | 2,600 | 650 | 289 | 163 | 104 | 73 | 55 | 42 | 34 | 28 | 20 | 15 | 12 | 10 | 8 |
| 60 | 3,743 | 936 | 416 | 234 | 150 | 104 | 77 | 60 | 48 | 39 | 28 | 21 | 16 | 13 | 11 |
| 70 | 5,095 | 1,274 | 567 | 319 | 204 | 142 | 104 | 80 | 63 | 53 | 37 | 28 | 22 | 17 | 14 |
| 80 | 6,654 | 1,664 | 740 | 416 | 267 | 185 | 136 | 104 | 83 | 67 | 48 | 36 | 28 | 22 | 18 |
| 90 | 8,421 | 2,106 | 936 | 527 | 337 | 234 | 172 | 132 | 104 | 85 | 60 | 45 | 35 | 28 | 23 |
| 100 | 10,397 | 2,600 | 1,156 | 650 | 416 | 289 | 213 | 163 | 129 | 104 | 73 | 55 | 42 | 34 | 28 |

[a]For two-tailed tests with two-treatment experiments. For experiments with a one-tailed test, the replication shown would provide an experiment of 97.5% power at $P < .025$.

TABLE 4. NUMBER OF TREATMENT GROUPS ENABLING THE NUMBER OF REPLICATES PER TREATMENT GROUP TO BE REDUCED WITHOUT DECREASING THE SENSITIVITY OF EXPERIMENTS

| Reduction in replication[a] | | Power of experiment, % | | |
|---|---|---|---|---|
| From | To | 80 | 90 | 95 |
| | | No. of trt. groups | | |
| 3 | 2 | 13 | 13 | 11 |
| 4 | 3 | 14 | 14 | 11 |
| 5 | 4 | 18 | 17 | 12 |
| 6 | 5 | 23 | 21 | 13 |
| 7 | 6 | –[b] | 22 | 14 |
| 8 | 7 | | –[b] | 16 |
| 9 | 8 | | | 16 |
| 10 | 9 | | | 15 |
| ≥11 | | | | –[b] |

[a]Reduction from values given in Tables 1 to 3 for two-treatment experiments.

[b]Inclusion of >2 treatment groups does not enable a reduction in the number of replicates per treatment group.

per group. Accordingly, replication requirements also were calculated as for generating Tables 1 to 3 but for experiments with ≥ 3 treatment groups. The impact of treatment number, which is summarized in Table 4, was determined to be minimal. For example, reduction from 3 to 2, 4 to 3, 5 to 4 or 6 to 5 replicates per treatment group in experiments of 80% power would require increases in the number of treatment groups from 2 to 13, 14, 18 and 23, respectively, whereas requirements for ≥ 7 replicates per treatment group per two-treatment experiment are not altered even by increasing the number of treatment groups to infinity (Table 4). Therefore, when a *t*-test will be used to detect differences among means, the data in Tables 1 to 3 should be appropriate for most experiments of larger size. Some experiments are designed to permit planned comparisons among multiple treatment means via orthogonal contrasts, orthogonal polynomials, and so on (Steel and Torrie, 1960). It has been estimated that such experiments may enable replication requirements to be reduced by as much as 20 to 30% below those when a simple two-treatment comparison of means is utilized (Gill, 1989).

*One- vs Two-Tailed Tests.* Tables 1 to 3 also apply specifically to studies involving two-tailed tests. If one were certain that a treatment could only elicit a positive or a negative effect, a one-tailed test could be used. Within limits, this would reduce replication requirements. However, the outcome of most studies is unknown in advance; despite occasional expectations concerning the nature of a potential response, deviations from expectations are not uncommon. Tables were constructed for two-tailed tests because these usually are the most appropriate. As indicated among the footnotes, however, data in Tables 1 to 3 are equivalent to those for experiments with one-tailed tests of 90, 95 and 97.5% power at $P < .025$, respectively.

*Reliability of the CV Supplied by the User.* The data in Tables 1 to 3 are based on well-established, fundamental statistical probabilities (Student's *t*, etc.). However, the accuracy of estimates taken from these tables will be no greater than that of the estimated CV provided by the user. Because the actual variability (CV, etc.) among replicates is never known in advance, this must be estimated from previous experiences of the investigator or others. Several factors will determine the accuracy and appropriateness of such estimates. One of these will be the similarity between the planned population and that from which the estimate is taken. For example, one would expect a greater CV for milk production of cows if this were determined from all cows in a herd, as opposed to only those cows of a single breed after adjustment for age, lactation number, season of calving, and so on.

The reliability of the user-supplied CV will also depend on the number of observations on which the mean and standard deviation used for its calculation are based; values based on a small sample size may be grossly over- or underestimated. As a safeguard against the selection of an unreasonable estimate, it is recommended that users examine CV for similar populations from several different studies, laboratories, etc. Also, one might determine the likely range of CV from the upper and lower confidence interval values for the mean or the standard deviation used in its calculation. Alternatively, one is not restricted to the control CV when using these tables. The appropriateness of either the control variance or the error mean square as the variance estimate ($s^2$) in Tang's equation was described previously, and it should be noted that the hypothetical $s^2/d^2$ ratios entered into Tang's equation for computing the tabular data (Tables 1 to 3) were, only by definition, based on

control data. Were $s^2$ to be defined as the pooled variance estimate, the $s^2/d^2$ ratios and resulting tabular data would remain unchanged. Thus, data in Tables 1 to 3 are equally appropriate when a CV estimate is based on the pooled variance estimate. However, the reader is cautioned that even pooled variance estimates may be subject to large potential estimation errors. Moreover, to calculate the CV from a pooled variance estimate one still must furnish an estimate of the mean (i.e., the CV equals the standard deviation as a percentage of the mean). Clearly the grand mean for the study, which would have nearly the same degrees of freedom as the error mean square, would not be suitable for this calculation, because the grand mean is subject to the effects of experimental treatments. Because the control mean would seem most appropriate for this calculation, the use of a pooled-variance estimate will increase the degrees of freedom for only one of the two statistics (e.g., standard deviation and mean) needed to calculate the CV.

If one were limited to estimating the expected population variance from the data of a single experiment, use of a pooled variance estimate (if available) would offer some advantage. Although a large body of published information may be consulted for most replicates (e.g., species of animals, etc.), the examination of control CV from several studies should more than compensate for the limited degrees of freedom associated with control data within any individual study. Such an approach also would provide an added safeguard against the possibility that a CV from a single study might be unique to that study, laboratory, strain of animal, etc.

Other factors to be considered in selecting a reasonable CV are the technical precision of individual measurements and the number of measurements per replicate. In some (and perhaps most) circumstances, limited sampling and(or) technical error (e.g., assay variability, etc.) will inflate the true variability among replicates. Alternatively, imprecision associated with crude measurements, subjective scoring, and so on, may serve to mask actual differences among replicates. Once again, the most appropriate estimate of the CV will be based on experiences with measurements of equivalent technical precision, sampling regimens, etc. A technique for simultaneously weighting and(or) optimizing the number of replicates and observations per replicate for experiments of known power and sensitivity has been presented elsewhere (Berndtson, 1989, 1990; Berndtson et al., 1989; Berndtson and Thompson, 1990). Readers should appreciate that replication requirements usually do decrease (sometimes tremendously) as one increases within-replicate sampling, but that a point is reached beyond which further sampling fails to enhance the power and sensitivity of the experiment (Berndtson, 1989, 1990; Berndtson et al., 1989; Gill, 1989; Berndtson and Thompson, 1990). Therefore, replication must be distinguished from sampling within replicates; one must not assume, for example, that the power and sensitivity of an experiment could be maintained by doubling the number of observations per replicate and decreasing to one-half the number of replicates per treatment group.

The appropriateness of any available CV also must include consideration of experimental design and(or) procedural factors aimed at reducing the impact of inherent variability among replicates. For example, a study was conducted recently that involved two treatment groups (i.e., control and treated) replicated with seven bulls each (Berndtson and Igboeli, 1988). Had bulls been assigned on a purely random basis at the initiation of the treatment period, the risk of Type II error would have been unacceptable; for several of the characteristics evaluated, statistical significance would have been unlikely unless the treatment produced 40 to 50% differences from the control (Berndtson, 1990). However, extensive pretreatment measurements taken on each animal were used to adjust for inherent, among-animal variability. In this instance, the pretreatment mean was subtracted from all post-treatment observations on a within-bull basis, and the grand mean for the experiment was added to the resulting difference. By compensating for pretreatment differences in this way, a substantial increase in sensitivity was realized for some measurements (Berndtson, 1990). Various statistical procedures, also aimed at partitioning out variability unattributable to actual treatment effects, have been developed (e.g., analysis of covariance, assignment of replicates with common characteristics to blocks or pairs, etc.). Such procedures may permit a reduction in replication or an increase in experimental power and sensitivity (Kastenbaum et al., 1970b). Before adopting such

approaches, one should weigh the benefits from partitioning out extraneous sources of variability against the loss of degrees of freedom in the error term (e.g., one will lose degrees of freedom in the error term equal to one less than the number of blocks, etc.). In general, such approaches are more likely to be helpful if inherent differences among replicates are great, but they may be counterproductive if replicates are homogeneous. The most important point relative to this issue is that to use data in Tables 1 to 3 as a reliable guide, one must determine and consider the experimental protocol and(or) design both in the experiment being planned and in that from which the CV is to be estimated. If, for example, we wished to conduct a study similar to that cited above, in which pretreatment data will be used to adjust for inherent differences within our sample population, it would be most appropriate to use an estimate of the CV among bulls based on post-treatment, adjusted control data.

*Replication Requirements as a Basis for Judging the Relative Sensitivity of End Points.* Some researchers might find it tempting to select the best end point(s) for an experiment (i.e., the one(s) most capable of detecting a treatment response) by comparing the replication needed with each end point to detect changes of equivalent magnitude. The problem with this approach is that treatments rarely affect all characteristics equally. An end point that might seem insensitive when judged by relative replication requirements might be quite sensitive if it is particularly responsive to the treatment under study. As one example, we (Berndtson et al., 1989) reported recently that the number of rabbits needed per treatment was about one-fifth as great (for equivalent power and sensitivity) when treatments were assessed via seminiferous tubular diameter vs the number of spermatids per seminiferous tubular cross-section (a measure of sperm production). However, in one experiment in which both end points were assessed in rabbits exposed to a chemical agent, the latter was depressed much more severely. In fact, we estimated that a dosage of this chemical producing a 5% change in tubular diameter, which would be detectable with 70 rabbits per treatment (90% power at $P < .05$), would be associated with a 21% change in spermatids per cross-section and detectable in an experiment of equal power with only 19 rabbits per treatment (Berndtson et al., 1989). Clearly, to predict the relative sensitivity of specific end points, one must consider the relative degree to which each is likely to change in response to treatment.

*Advantages Relative to Alternative Approaches.* Although several approaches are available by which one may estimate replication needed in a planned experiment, they have been used infrequently. Among possible reasons for this are 1) a lack of awareness or concern by most investigators regarding potential deficiencies associated with conventional replication practices and 2) the inconvenience of the computations normally required. The tabular data and the sample illustrations with actual data presented herein were intended to address both of these deficiencies. To use Tables 1 to 3, investigators need only furnish an appropriate CV for the population of interest, which can be determined in less than one minute via a simple hand calculator once one has identified either the population mean and standard deviation or the mean, standard error and number of replicates per mean. After the CV has been estimated, one can read directly the replication needed to provide experiments of any power and sensitivity combination within most reasonable ranges of interest. The ease and speed of access should render the approach attractive not only during the planning of experiments, but also for routine use by referees or by other investigators who might find it beneficial to determine the actual power and sensitivity of completed studies.

Tang's procedure could be used as originally proposed by any investigator seeking information such as that available herein, but at considerable inconvenience. First, to use Tang's equation, an estimate of the error mean square must be determined. Use of the error mean square has the advantage of providing more degrees of freedom than are associated with the simple variance among control subjects. However, error mean squares may be difficult to determine from information normally included within the published literature. This seriously hampers access to information that might be needed in planning an experiment involving a species, cell culture system, experimental end point, etc., for which the investigator lacks prior personal experience or data. Also, it precludes one's ability to independently determine the power and sensitivity of studies within the published literature.

Others might circumvent this limitation by using the control variance in lieu of the error mean square, as was done herein, but the appropriateness of this modification may not be generally recognized (Berndtson and Thompson, 1990). Second, before one can select appropriate values of Student's *t* for Tang's equation, the error degrees of freedom in the planned experiment must be determined. The latter will depend, in part, on the number of replicates per treatment group, which constitutes the unknown variable Tang's equation is designed to determine. Thus, the approach is empirical. One must estimate the number of replicates that will be needed, determine the error degrees of freedom that this would provide, look up the corresponding values of *t* and solve for R (i.e., the number of replicates per treatment) in the equation. If the value of R equals the original estimate, additional calculations are unnecessary. Should the value of R differ from the original estimate, a revised error degrees of freedom must be used to select the *t*-values, and the calculations must be repeated until all terms in the equation are satisfied. The entire process must be repeated for every power and sensitivity combination and for each experimental end point of interest. Although the specific equations advanced by various biometricians differ, all seem to require similar trial-and-error calculations (Cochran and Cox, 1957; Steel and Torrie, 1960; Berndtson et al., 1989).

Statistically based procedures have been used to construct tables or graphs of replication requirements that are applicable for specific end points. These have served a useful purpose and in several instances offer a distinct advantage by simultaneously addressing the interaction of replication and the level of sampling within replicates (Seidel and Foote, 1973; Berndtson, 1989, 1990; Berndtson et al., 1989; Berndtson and Thompson, 1990). However, their intended use was limited to the specific end points for which they were calculated. Others also have constructed tables or graphs that, like Tables 1 to 3, are applicable across scientific disciplines (Cochran and Cox, 1957; Harter, 1957; Kastenbaum et al., 1970a,b; Gill, 1978, 1989; Kraemer and Thiemann, 1987). Unfortunately, application of these materials is limited by accessibility (older texts, out of print, etc.), by the range of power, sensitivity and population variances over which they extend, and(or) are based only on one-tailed tests. Some authors have chosen graphic presentation of this information that, although useful in illustrating trends, often entails the potential for rather large interpolation errors. It must be emphasized that interpolation per se is not objectionable; it is unlikely that the actual variance within a population will be exactly as estimated in advance of the experiment. Thus, such data and Tables 1 to 3 are intended as an approximate guide. However, variances differ considerably among sample populations, end points examined, etc. For example, the CV encountered for various end points examined within recent large- and small-animal experiments in the author's laboratory have ranged from 3.9 to 59% (Berndtson and Igboeli, 1989; Berndtson and Thompson, 1990), and CV of considerably greater magnitude are commonplace within the biological sciences. To be widely applicable, it seemed important that the present data encompass a wide range of population variances and also be inclusive of both conventional levels of replication and those needed for highly sensitive experiments. Whereas this was accomplished with three tables of moderate size, graphic presentation of data on a scale encompassing a few to several thousand replicates per treatment would lead to serious interpolation difficulties. This may be the reason that many previously published graphs cover a narrow range of population variances uncharacteristic of those within many large-animal populations.

### Implications

Each researcher must determine the levels of statistical sensitivity most appropriate for his or her research aware that factors such as cost and feasibility will influence all decisions. However, the present data should be useful in identifying the number of replicates needed to meet the investigators requirements. Although it is customary to focus on Type I error, the general lack of attention to Type II error has been questioned. In the author's opinion, all research should represent a search for the truth; investigators should strive to avoid errors of either type. In many instances it will be impossible or impractical to provide as much replication as one might desire. Nonetheless, the use of the tables provided should enable researchers to determine both the power and sensitivity of the resulting experiment and to

acquire insight into the level of caution appropriate when interpreting non-significant findings. By facilitating the assessment of Type II error probabilities, these data should advance the critical review of previously completed, published research.

## Literature Cited


Berndtson, W. E. 1989. Sampling intensities and replication requirements for the detection of treatment effects on testicular function in bulls and stallions. A statistical assessment. J. Anim. Sci. 67:213.

Berndtson, W. E. 1990. Replication requirements and number of ejaculates needed for assessing treatment effects on sperm output and seminal characteristics of electroejaculated Holstein bulls. J. Anim. Sci. 68:709.

Berndtson, W. E. and G. Igboeli. 1988. Spermatogenesis, sperm output and seminal quality of Holstein bulls electroejaculated after administration of oxytocin. J. Reprod. Fertil. 82:467.

Berndtson, W. E. and G. Igboeli. 1989. Numbers of Sertoli cells, quantitative rates of sperm production, and the efficiency of spermatogenesis in relation to the daily sperm output and seminal quality of young beef bulls. Am. J. Vet. Res. 50:1193.

Berndtson, W. E., C. Neefus, R. H. Foote and R. P. Amann. 1989. Optimal replication for histometric analyses of testicular function in rats or rabbits. Fundam. Appl. Toxicol. 12:291.

Berndtson, W. E. and T. L. Thompson. 1990. Age as a factor influencing the power and sensitivity of experiments for assessing body weight, testis size and spermatogenesis in rats. J. Androl. 11:325.

Cochran, W. G. and G. M. Cox. 1957. Experimental Designs. John Wiley, New York.

Gill, J. L. 1978. Design and Analysis of Experiments in the Animal and Medical Sciences. Vol. 3. Appendices. Iowa State Univ. Press, Ames.

Gill, J. L. 1989. Statistical aspects of design and analysis of experiments with animals in pens. J. Anim. Breed. Genet. 106:321.

Harter, H. L. 1957. Error rates and sample sizes for range tests in multiple comparisons. Biometrics 13:511.

Kastenbaum, M. A., B. G. Hoel and K. O. Brown. 1970a. Sample size requirements: one way analysis of variance. Biometrika 57:421.

Kastenbaum, M. A., B. G. Hoel and K. O. Brown. 1970b. Sample size requirements: randomized block designs. Biometrika 57:573.

Kraemer, H. C. and S. Thiemann. 1987. How Many Subjects? Statistical Power Analysis in Research. Sage Publications, Newbury Park, CA.

Remington, R. D. and M. A. Schork. 1985. Statistics with Applications to the Biological and Health Sciences (2nd Ed.). Prentice-Hall, Englewood Cliffs, NJ.

Seidel, G. E., Jr. and R. H. Foote. 1973. Variance components of semen criteria from bulls ejaculated frequently and their use in experimental design. J. Dairy Sci. 56:399.

Steel, R.G.D. and J. H. Torrie. 1960. Principles and Procedures of Statistics. McGraw-Hill Book Co., New York.

Tang, P. C. 1938. The power function of the analysis of variance tests with tables and illustrations of their use. Stat. Res. Memoirs 2:126 (As cited by Steel and Torrie, 1960).