Please find below information on how sample sizes are determined. This document assumes that the reader has little or no statistical knowledge. The goal of the document is to make clear the issues associated with sample size determination and why investigators often need help with this process.

Determination of Sample Size

Determining the sample size for a study is a crucial component of study design. The goal is to include sufficient numbers of subjects so that statistically significant results can be detected. Using too few subjects results in wasted time, effort, research dollars, and animal lives, and yields statistically inconclusive results. Statistically inconclusive findings make it difficult to determine whether a particular treatment or intervention was effective and to identify directions for future studies. Studies with insufficient subjects also may result in potentially important research advances that go undetected. In statistical language, these studies are referred to as "under-powered." That is, the probability that they will detect an existing treatment effect is lower than optimal (see **Parameters for Sample Size Determination**).

Using too many subjects may result in statistically significant conclusions and clear future study directions. However, if the same answer could have been obtained with fewer subjects, then time, effort, research dollars, and animal lives also have been wasted. In statistical language, these studies are referred to as "over-powered." That is, the probability that they will detect a treatment effect is higher than optimal (see **Parameters for Sample Size Determination**).

Using the appropriate number of subjects optimizes the probability that a study will yield interpretable results and minimizes research waste. From a statistical perspective, studies with the optimal number of subjects have sufficient -- neither too much nor too little -- statistical "power" to detect findings.

Under federal regulations, one of the responsibilities of the Institutional Animal Care and Use Committee (IACUC) is to ensure that study sample sizes have been rigorously determined. One of the roles of the Data Management Services (DMS) Statistical Consulting group is to assist investigators with these determinations.

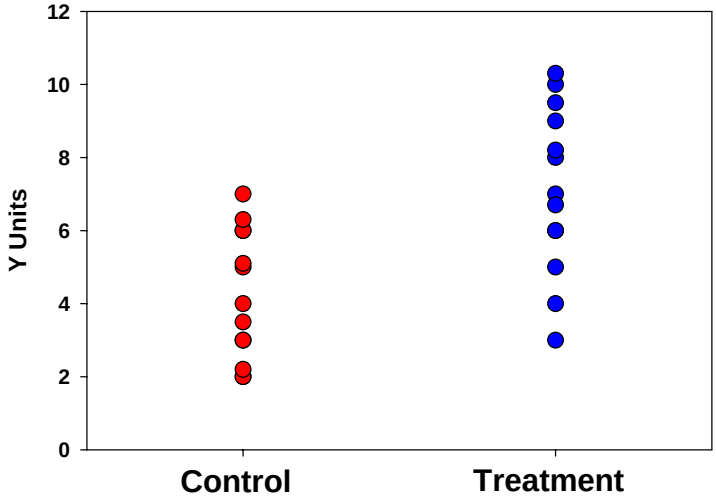
The Role of Variability

In a perfect research environment in which measurement devices were errorless, subjects were identical and exhibited identical responses to a treatment, and treatments were implemented flawlessly and consistently, there would be no variability in responses. All subjects treated the same would manifest exactly the same response. In such a world – without variability -- there would be no need for statistical analysis because whether a treatment altered responses or not would be a certainty. This ideal world, however, does not exist. In the absence of absolute consistency – that is, in the presence of variability – uncertainty exists about whether or not a treatment altered responses. Statistical analyses

address this uncertainty. (As an aside, because statistical analyses require the presence of at least some variability, it is not possible to use statistical approaches when only one subject is used per treatment group.)

The measurement of almost any attribute reveals variability. For example, it is not surprising that the body weights of animals of the same age and sex are not exactly the same. Nor is it surprising that the viral titers or immune parameters of animals that were infected with the same agent at the same time with the same dose also exhibit some variability. In the simplest experimental design, containing a control group and a treatment group, no investigator is surprised to find that the control group values exhibit some variability and that the treatment group values also exhibit some variability (see **Figure 1**). Keeping this variability in mind, the investigator is more interested, however, in whether the treatment group values are generally higher than the control group values.

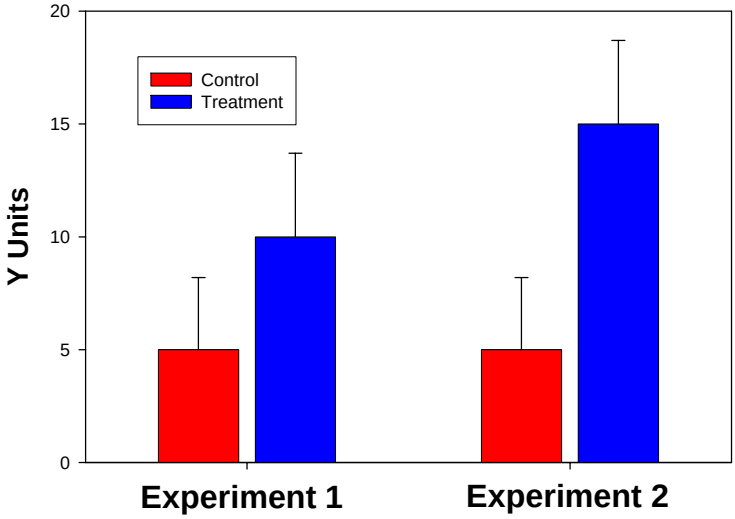
Figure 1: Variability Within Groups



The problem for the investigator is: given the variability that exists among subjects treated the same (within the control group and within the experimental group), is the difference between the two groups consistent enough to be certain that the treatment had an effect?

Holding variability within groups constant, the larger the difference between group means, the more certainty the investigator has that a treatment worked. **Figure 2** depicts this situation. On the left side of the figure the difference between the group means is five units. The variability within each group is indicated visually by error bars that represent the standard error of the mean (sem) – a way of quantifying variability among subjects treated the same. On the right side of the figure, the difference

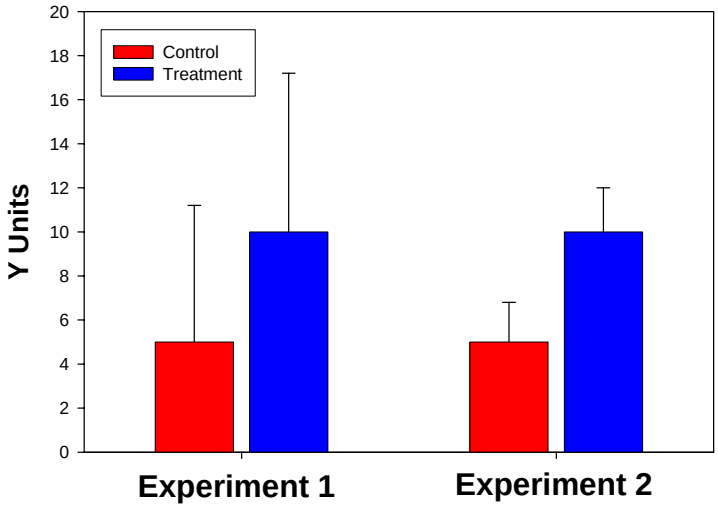
Figure 2: Within-Group Variability Constant Difference Between Group Means Varied



aday)

between the group means is 10 units. Note that the within-group variability – the error bars – are exactly the same. Intuitively, the data on the right side of the figure reveal more certainty that the treatment worked than do the data on the left side of the figure.

Figure 3: Within-Group Variability Varied Difference Between Group Means Constant



What if the difference between group means is the same, but the within-group variability differs?

Figure 3 illustrates this situation. On both sides of the figure, the mean difference between treatment groups is five units. On the left side of the figure, however, the error bars are much larger than on the right side of the figure. In this hypothetical situation, there is much more within-group variability on the left side than on the right side. Intuitively, the data on the right side of the figure indicate more certainty that the treatment worked – because the subjects responded with more consistency -- than do the data on the left side of the figure.

The statistician comes to the same conclusion but expresses it in somewhat different terms. The ultimate purpose of most studies is to use a sample (a subgroup) to make inferences about a population (the larger group of interest). When data exhibit large amounts of within-group variability relative to treatment variability, then any generalizations made to the population must be made with uncertainty. In other words, the reliability with which the sample can be used to make inferences about the population is less than when within-group variability is relatively small.

Statistical Analysis and Variability

Many statistical analyses grapple with this problem – given that we know that subjects will vary in their responses to the same treatment, are the observed differences between treatment groups consistent enough to state with relative certainty that the treatment worked?

The statistician conceptualizes the problem in terms of variability. Within a particular study, there are two major influences on the variability of measured responses: 1) the treatment, and 2) error. The treatment contributes to the variability of measured responses, if it was effective, by systematically increasing or decreasing them. Error contributes to the variability of measured responses in several ways. It is important to

note that the term "error" does not indicate that mistakes were made in the study. "Error" is the term used to refer to all of the influences other those that result from the treatment that could alter measured responses. Error includes, therefore, the inconsistency inherent in measurements obtained with a measurement device or technique that is not perfect, procedural differences in how the same treatment was administered to subjects, and inherent differences among subjects that are not related to the treatment. Error is considered a non-systematic influence on responses because it can increase or decrease them. The total variability in responses in a particular study can be divided into these two components: 1) variability that is associated with or that is the result of the treatment and, 2) variability that is not the result of the treatment or error variability.

Many statistical analyses address the same question: is the variability associated with the treatment large enough relative to the variability associated with error to be relatively certain that the treatment worked?

Notice that this is the same question that was stated above using different terminology. The intuitive grasp that the situation in the right side of **Figure 3** reflects more certainty about the treatment effectiveness than the situation on the left side of the figure illustrates this point.

Also note that the absolute size of treatment variability and error variability is not important – only their relative relationship. A useful analogy is a signal-to-noise ratio. The treatment variability is the signal; the error variability is the noise. Noisy data – data that exhibit a great deal of within-group variability – require that the signal – the treatment variability -- be strong in order to be detected.

Parameters for Sample Size Determination

Sample size determinations depend on four parameters. These parameters are: 1) the desired level of statistical power, 2) the p level, 3) treatment variability, and 4) error variability.

Statistical power refers to the probability that a treatment effect will be detected if it is there. By convention, power is generally set at about 0.80, or an 80% probability that a treatment effect will be detected if present. When a study is under-powered, it has less than an 80% chance of detecting an existing treatment effect. When it is over-powered, it has a greater than 80% chance of detecting a treatment effect.

P level refers to the probability of detecting a statistically significant difference that is the result of chance – not the result of the treatment. In other words, the p level determines the probability of obtaining an erroneously significant result. In statistical language, this error is called Type I error. By convention, p levels generally are set at 0.05, or a 5% probability that a significant difference will occur by chance.

Two of the four parameters – power and p level -- are pre-determined. The other two parameters – treatment variability and error variability – must be estimated in order to complete the sample size determination. Treatment and error variability can be estimated in three ways.

1) Pilot Studies: The most accurate determination of sample size is obtained when the investigator has collected relevant data from which an estimate of treatment variability and an estimate of error variability can be made. These data generally are obtained in a pilot or small-scale preliminary study. Note that the results of a pilot study do not have to be statistically significant in order for the data to be used to estimate treatment and error variability. This procedure is the best way to determine sample size.

2) Relevant Literature: Another means of making treatment and error variability estimates is to use the relevant scientific literature. Estimates could be made from the published work of investigators who have conducted similar studies or who have addressed related questions. This is the second-best way to determine sample size.

3) Rule-of-Thumb Estimates: The third means of making variability estimates is to use rough approximations or rules-of-thumb that are accepted in a particular field in the absence of data or published work. This procedure is, by far, the least accurate means of determining sample size, but sometimes must be used in the absence of data and relevant literature.

In general, if the variability associated with the treatment is large relative to the error variability, then relatively few subjects will be required to obtain statistically significant results. Conversely, if the variability associated with the treatment is small relative to the error variability, then relatively more subjects will be required to obtain statistically significant results.