Appendix 5.3.2
The Null Hypothesis, Type I / Type II Error, P-values, and Sample Sizes[1]

In designing a study hypothesis, researchers are comparing two groups—usually defined as the study group versus the standard or control group. The "null hypothesis" in clinical research means that there is no difference between the two populations being compared. The general objective of clinical trials is to prove that the study arm is indeed different from the control arm, so researchers aim *to reject the null hypothesis* in favor of the "alternative hypothesis," or the outcome that the two groups are different. For example, in a medical device trial, the null hypothesis might be that the incidence of restenosis (an unwanted outcome) for patients receiving ABC stent is not lower than it is for those receiving XYZ stent. The goal of a study is to determine if the null hypothesis can be disproved in favor of the alternative hypothesis (that restenosis occurs less in patients with ABC stent than with XYZ stent) through the performance of a test evaluating this outcome. The result of the test may be that the null hypothesis is true (that there is no difference in restenosis rates between patients receiving the different types of stents). Or, the null hypothesis may be rejected in favor of the alternative hypothesis, which indicates that there *is* a difference in restensosis rates between the two groups.

Statistical error is commonly described as the difference between anestimated or measured value and the true or theoretically correct value that is caused by random and inherently unpredictable fluctuations in the measurement apparatus, methods, and patient response to treatment.[2] The magnitude of the error depends on the amount of variation in measurement accuracy and treatment response.

Understanding statistical error is important because medical device trials are particularly vulnerable to unpredictability, for a number of different reasons:[3]

- Clinical and commercialization pathways often involve relatively small sample sizes (i.e., compared to pharmaceutical trials).
- Smaller device companies often have scarce resources, limited experience, and suboptimal methods for making critical trial estimates.
- Larger device companies may over-accelerate trial evaluations due to the rapid pace of technology turnover and decreasing length of total product lifecycles..
- Human biology and anatomy are unpredictable and affect patient response to a particular treatment.

In testing for the null hypothesis, there are two possible kinds of errors. Type I error, also known as a "false positive" or alpha error, occurs if the null hypothesis is rejected when it is actually true. Stated another way, this is the error of accepting an alternative hypothesis (the real hypothesis of interest) when the results can be attributed to chance. An easy example to understand is when a test shows that a woman is pregnant when she actually is not carrying a child. A test results in Type II error, also called a "false negative" or beta error, when the null hypothesis is not rejected when the alternative hypothesis is true. In other words, it occurs when the researchers fail to observe a

difference when, in reality, a difference exists. The following table summarizes these concepts using the example of a pregnancy.

Table 5.3.2-1 – A simple example helps illustrate the difference between type I and type II error.

		Reality	
		Null hypothesis is true	Alternative hypothesis is true
Research	Null hypothesis is true	Accurate (not pregnant)	False negative type II or beta error (pregnant but not detected)
	Alternative hypothesis is true	False positive type I or alpha error (not pregnant but incorrectly detected)	Accurate (pregnant)

Hypothesis testing is used to determine whether or not the observed difference between two groups can be explained simply through random chance. Traditionally, an acceptable level of Type I error in medical device trials is set at .05. This means that there is a 5 percent chance that the variation observed as a result of the test is due to chance. This probability is called the "level of significance" and is reported as a study's "p-value." Specifically, if the null hypothesis is true (no difference exists between the two populations being compared), then the p-value is the probability that random sampling would result in a difference as big as, or bigger than, the one observed in the sample size actually evaluated.[4] Another common convention when reporting a p-value is to select a confidence interval; typically, this is set at a value of 95 percent.

Although they are closely related, it is worth noting the difference between the p-value and the confidence interval. A confidence interval gives an estimated range of values which is likely to include an unknown value or parameter. The estimated range is calculated from a given set of sample data. If independent samples are taken repeatedly from the same set of values, and a confidence interval calculated for each sample, then a certain percentage (confidence level) of the intervals will include the unknown population parameter.[5] Confidence intervals are usually calculated so that this percentage is 95 percent. Wider intervals may indicate that more data should be collected before anything definite can be said about the value or parameter. Confidence intervals are more informative than the simple results of hypothesis tests (i.e., deciding whether to reject or accept the null hypothesis), since they provide a range of plausible values for the unknown parameter. By applying mathematical theory, confidence intervals can be inferred from small sample sizes. This method is typically used in clinical studies, as confidence intervals are established before researchers acquire large data sets. Meaningful information is predicated upon a random sample, independent observations, accurate assessment, and the assessment of the event that is truly central to the study (e.g., the probability of a life-threatening complication versus the probability of *all* complications).

Importantly, the lower the probability of Type I error in a study, the higher its likelihood of Type II error (and vice versa). When a significant difference exists in the population being studied but the test fails to find this difference (Type II error), the study is said to lack "power"[6] (power is defined as the probability that the test will reject a false null hypothesis).[7] A significant difference in clinical studies is generally considered to be a p-value of .05, meaning that if a p-value is .0500 *or less*, the two populations being compared are indeed statistically different from one another and the null hypothesis can be rejected.

Sample Size Determination

Sample size determination is a complex subject, best performed with the careful assistance of a qualified statistician. However, sample sizes generally can be determined in two ways:

1. ***Use of Confidence Intervals*** – Through this process, one determines how many subjects are needed so that the 95 percent confidence interval has a desired width. This means that if samples of the same size are drawn repeatedly from a population, and a confidence interval is calculated from each sample, then 95 percent of these intervals should contain the population mean.[8]
2. ***Use of Hypothesis Testing*** – Determine how many subjects are needed so that the study has enough power to obtain a significant difference, given the specified experimental hypothesis.
    - Usually, alpha (probability of a Type I error) is set to 0.05, but if a more significant value is desired (say 0.01), a larger sample size will be needed.
    - Beta (the probability of a Type II error) is typically set at a power of 80 (0.20) to 90 (0.10) percent. Conventionally, this is the standard imposed, but a company can choose any threshold it deems appropriate based on what is being studied. However, a threshold more lenient than 80 to 90 percent will be questioned and therefore must be justified. It will also have a higher likelihood of a failed trial.

With either method, it is necessary to make an assumption about the so-called delta (the true difference between the two groups being studied) because that delta will determine the sample size that is needed to provide the desired power. The delta is usually the smallest difference that would be clinically important, which can be hard to define.[9] Consulting prior studies will provide useful guidance. Thoughtful consideration of assumptions leading to a decision on sample size cannot be overemphasized, as these assumptions define the study endpoints from a statistical standpoint.

It is also important to consider whether the study is designed to prove non-inferiority (equivalence) or superiority, as this will impact the statistical considerations and required sample size significantly. Non-inferiority requires a smaller sample size than superiority.

Additionally, always base the study protocol on a sample size that exceeds what is statistically required in order to account for patient withdrawals, patients lost during follow-ups, and other exemptions that inevitably occur during the clinical trial. It is best to study more patients than necessary rather than to come up short during the final analysis and fail to prove the statistical endpoint based upon an insufficient sample size.

---

[1] Based on information provided from "Type I and Type II Errors," Wikipedia.org, http://en.wikipedia.org/wiki/Type_I_error#Type_I_error (March 30, 2014), unless otherwise cited.
[2] "Error," Answers.com, http://www.answers.com/topic/error (March 30, 2014).
[3] Richard Kuntz, "Data Monitoring Committees and Adaptive Clinical Trial Design, Innovation.org, http://www.innovation.org/documents/File/Adaptive_Designs_Presentations/23_Kuntz_Data_Monitoring_Committees_and_Adaptive_Clinical_Trial_Designs_Medica_Device_Considerations.pdf (March 30, 2014).
[4] Harvey Motulsky, *Intuitive Biostatistics* (Oxford University Press, 1995).
[5] "Confidence Intervals," Exploring Data Website, http://exploringdata.cqu.edu.au/conf_int.htm (April 25, 2007).
[5] Motulsky, op. cit.
[6] "Research Methods: Inferential Statistics," AllPsych Online, http://allpsych.com/researchmethods/errors.html (March 30, 2014).
[7] "Statistical Power," Wikipedia.org, http://en.wikipedia.org/wiki/Statistical_power (March 30, 2014).
[8] "Confidence Intervals," op. cit,
[9] Motulsky, op. cit.

