Sample Size Specification Guidelines for NASA Human Research Studies

Statistical requirements for most grant or major study proposals:
Statistical planning plays an important role in virtually all scientific research. It plays a particularly valuable role in the design of experiments, including specification of sample size(s), and also in the analysis of outcomes that address primary aims and hypotheses. As a result, PIs are highly encouraged to recruit statisticians as Co-Is, so that they can apply their skills to help with study design, and the plan to collect, analyze, and interpret data to produce high-quality research proposal.

This document gives particular emphasis to the problem of how to arrive at and justify experiment sample size(s). Much of what follows reflects the basic principles cited by Russell Lenth (Lenth, 2001) in his article in The American Statistician; however planning of statistical analyses methodology is not discussed here. Some new ideas (Mudge, et al., 2012) are included that could be applicable to dealing with NASA's limited availability of research subjects, although other non-traditional sample size justification methods (ex. Bacchetti, et al., 2011) may also apply. These recommendations are necessarily general, and may not be universally applicable. Nevertheless, these guidelines are intended to clarify an understanding among PIs, grant reviewers and NASA pertaining to sample-size issues for NASA human research studies.

Position Statement: Clinically or Scientifically Meaningful Outcome
If a study is worth conducting, there must be a high expectation of it producing some meaningful and measurable outcome.
The basic assumption is that research conducted in support of NASA's Human Research Program (HRP), should be targeted at detecting meaningful changes/effects, for example, the ability of a novel intervention to reduce negative consequences of spaceflight on the human by XX %, relative to current standards. In some cases, where the treatment or outcomes of interest are completely novel, it is acceptable to propose research aimed at measuring the effects, as long as an *a-priori* level of measurement accuracy is proposed, and a compelling justification of the importance of making the observations is made.

More specifically, most of these studies will fall in one of two broad categories:

1. Inferential (i.e., hypothesis-testing) studies: These are studies in which subjects are recruited to participate in a study designed to make inferences about the effects of a novel countermeasure to the population (e.g., all current and future astronauts), based on sample data. The best studies use random assignment to treatment group(s) or control(s), however this is not always possible or feasible (e.g., astronauts often choose to participate in a novel intervention or as controls). These studies are generally conducted in disciplines where there is sufficient knowledge about the negative effects of spaceflight and the gaps associated with current countermeasures, such that novel countermeasures may be tested.
2. Characterization/Pilot research: These are studies in which data are collected from research participants with the primary goal of measuring the effects of a set of predictors (e.g., spaceflight, subject characteristics, nutrition) on one or more outcomes (e.g., fluid distribution, strength, BMD, neural activation). One distinction of these studies from Clinical Trial-Type studies is that the predictors are not systematically altered or

manipulated by the researchers; observations of both predictors and outcomes are made without intervention. These studies are generally conducted in areas where little is currently known about phenomena, and/or where observational/pilot data are needed.

Hypothesis-driven studies:

The following recommendations are relevant for studies in which you will be collecting data that includes primary outcomes (dependent variables) and predictor variable(s) (e.g., group, gender, treatment, age, BMI) for the purpose of making inference (testing hypotheses) about a target population (e.g., astronauts). Giving careful thought to the design of experiments with respect to sample size(s) is critical for the HRP because traditional (i.e., "big" N) assumptions about how sample sizes are calculated are often not realistic for NASA ISS or analog studies where small N is usually unavoidable. These guidelines are designed to help PIs address this challenging issue.

Describe your Experimental Design
Statistical reviewers need to understand what groups and/or treatments you are planning to compare, frequency of measurement, important control variables, and other relevant aspects of your experimental design. Sample size justification is necessarily linked to your statistical analysis plan, which in turn, depends on all of these experimental design parameters.

Identify your Primary outcomes
While you will probably collect a large amount of data, not all of it is directly relevant for addressing your primary aims and hypotheses. Some of your data will be collected for secondary analyses. However a sample size determination should be based on your primary outcomes (i.e., those that are specifically referenced in your hypotheses). Some specific recommendations include:

1. Choose your primary outcome(s) carefully. In general, clinically relevant measurements that are reliable and valid, with low "noise" variability provide more statistical power to detect differences. These may be more expensive to collect, but the tradeoff in sample size requirements may well justify the additional cost.
2. Focus on one or two **key** outcomes for the purpose of sample size calculations. Choose the critical measurement that discipline experts would expect to see in a final manuscript. It is not necessary that you power your study based on every outcome measurement—one or two generally suffice.
3. Provide a description of how your primary outcome(s) is distributed, and its anticipated variability based on pilot data or published work. Is it normally distributed in the population? Or is it typically skewed? This helps to determine what statistical analysis will ultimately be needed, but it also impacts sample size calculations. Note that if you are considering a repeated-measures design (e.g., pre/post) it is the anticipated variability and distribution of pre-post differences that drives sample size calculations.

Effect size must be well articulated for sample size determination
Sample sizes are determined, in part, on the probability of detecting an effect, and so the magnitude of an effect is a critical piece of the puzzle. Effect size is defined as the magnitude of a treatment effect, relative to the expected variability of the data used to estimate it. While the effect size together with sample size determines power, the components of the effect size calculation should be based on pilot studies or previous information in the literature. It is the

responsibility of the PI to propose a study (including sample size) that is designed to have a reasonable likelihood of detecting an effect that has some clinical or scientific meaning. For example, if an operationally feasible sample size (i.e., small N) is determined with the unfounded assumption that a novel countermeasure will be overwhelmingly successful; the overly optimistic effect size specification may be called into question. On the other hand, if a study detects a statistically significant effect that has no operational, scientific, or clinical impact, then the practical significance of the study may be legitimately questioned. In summary, finding a statistically significant difference is not, in and of itself, meaningful. This point cannot be overstated. NASA cares most about research that informs knowledge gaps in meaningful ways.

Assumptions about the relative weight given to Type I and II errors must be articulated
Good experimental design, adequate sample size, and sound theory development typically results in a decision (reject the null or not) that makes the "correct" inference about a population from sample data. Nevertheless, there are two types of errors that the hypothesis testing paradigm can encounter. Type I errors result when a null hypothesis is rejected based on sample data, when the effect is just due to chance variation in the sample. Type II errors result when the sample data analysis fails to reach statistical significance for rejecting the null, when in truth there is an effect in the population.

Historically, academia has emphasized avoiding a Type I error (α, falsely rejecting the null hypothesis) more than a Type II error (β, failing to detect a real effect). As a result, scientists are trained to avoid the Type I error at the expense of potential Type II errors. Of course, one cannot know for sure whether the decision to accept or reject the null hypothesis is "correct" at the time of making the decision. Only repeated studies and time would accumulate enough evidence to help researchers retrospectively appreciate whether or not they made a correct decision to reject (or not) the null hypothesis.

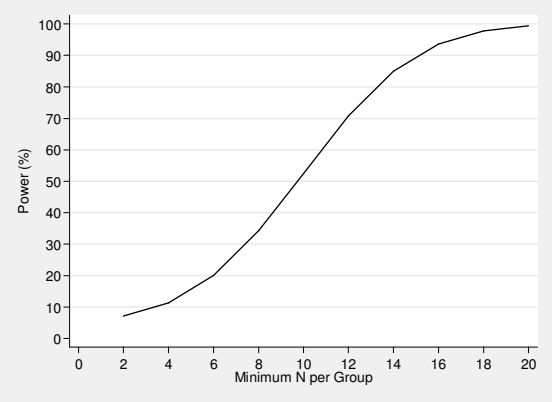
Avoiding Type I errors is accomplished by setting alpha (the probability of a Type I error) very low (e.g., 0.05 or 0.01), regardless of statistical power ($1-\beta$). For large-n studies, this is usually not a problem because large-n studies also tend to provide sufficient statistical power that results in reasonably low probability of a Type II error.

Small samples, however, often result in under-powered studies—studies that have a low probability of detecting an effect that exists (i.e., have a higher likelihood of Type II errors). Type II errors may lead to failing to identify a promising countermeasure or clinically important effect. This can be a serious problem if it leads scientists and NASA away from potentially effective countermeasures.

Because ISS research almost always involves small samples, and sometimes missing data as well as operational constraints, NASA researchers must carefully contemplate the costs of both Type I and II errors in their sample size consideration and statistical analysis plan. PIs should weigh the consequences of making Type I versus Type II errors in their proposed research and clearly defend their position. As a reference point, it may be useful to appreciate that the implied traditional approach to this balance gives Type I errors roughly four times the weight of Type II errors. This is true because alpha levels (probability for Type I error) are typically set at 0.05, and power to detect effects is usually set to .80 (so probability of Type II error = .20). The

default 4:1 weighting of errors may or may not be appropriate depending on each individual study. Using weights closer to equal may also be appropriate, particularly where novel study designs, treatments, or ideas are being tested.
Since ISS research inherently involves small sample sizes that inflate Type II error rates due to low statistical power, the HRP supports the position that if PIs cannot defend a relative weight of Type I versus II errors other than equal, they should consider an equal-weighted error risk posture. There is recent statistical thinking (Mudge, et al. (2012)) that is consistent with this position.

Proposing a range of reasonable sample sizes for a proposed study
Traditional grant proposal solicitations require the PI to propose and defend one specific sample size (e.g., n per group) so that there is reasonable certainty that when the study is complete, the null hypothesis will be rejected, leading evidence in support of the PI's alternative hypotheses. In order to better appreciate the effect of sample size, we highly recommend that PIs provide a graph or table showing the relationship between sample size and statistical power, in addition to a specific sample size request (see example right). This provides NASA and reviewers a better understanding of the tradeoff between sample size and power to detect an effect. It also provides flexibility for NASA in terms of funding levels so that the PI may be granted funds at a lower (or higher) level, depending on this relationship, competing proposals, and other factors.



Characterization studies (i.e. descriptive, pilot):

Unlike inferential studies (see above), which are designed to determine whether or not there is a clinically meaningful effect, characterization studies are designed to quantify the extent of an effect without knowing a priori knowledge of the effect size. Thus, the following recommendations are relevant for studies in which the primary goals are to measure and/or characterize observed effects, perhaps as a pre-curser to a future more refined inferential study.

Describe your Study Design and Outcomes
The same principles for describing study design and selection of outcomes for inferential studies apply here (see above).

Margin of Error/Level of Precision
In order to understand the impact of sample size in characterization studies it is necessary to specify a level of precision that would allow clinicians, scientists, or other content experts to draw meaningful conclusions from a study's descriptive results. Precision is typically specified in terms of the margin of error (ME), defined as the half-width of a 95% confidence interval. For example, one might want to estimate a mean change in systolic blood pressure in response to microgravity exposure with a ME of 12 mm Hg, representing 10% of a clinically accepted norm of 120 mm Hg. Note that the ME is something that the researcher establishes based on content

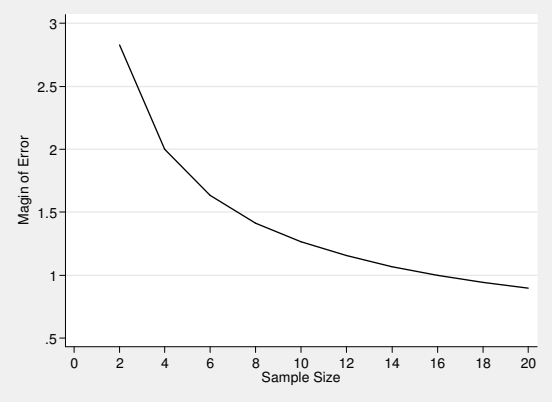
expertise—not statistical analysis. Also note that in this example it was not necessary to know or anticipate the actual mean change to arrive at a meaningful value for the ME. By contrast, if this were an inferential study, the mean change would have to be specified to arrive at an effect size, needed for sample size determination.

Relating precision to sample size

In general, the ME is related to the standard error of an estimated effect and proportional to the reciprocal square root of the sample size. More specifically, depending on the type of statistical analysis used to estimate effects, there are well-established statistical methods for precisely relating sample size to margin of error.

Proposing a range of reasonable sample sizes for a proposed study

Consistent with the above recommendations for inferential studies, NASA HRP recommends that PI's provide a graph or table showing the relationship between sample size and ME along with a specific sample size request (see example right). This provides NASA and reviewers a better understanding of the tradeoff between sample size and the ability to accurately characterize effects. It also provides flexibility for NASA in terms of funding levels so that the PI may be granted funds at a lower (or higher) level, depending on this relationship, competing proposals, and other factors.



References

Lenth, R. V. (2001), ``Some Practical Guidelines for Effective Sample Size Determination,'' *The American Statistician,* 55, 187-193.

Hoenig, John M. and Heisey, Dennis M. (2001), ``The Abuse of Power: The Pervasive Fallacy of Power Calculations for Data Analysis,'' *The American Statistician,* 55, 19-24.

Mudge JF, Baker LF, Edge CB, Houlahan JE (2012), Setting an Optimal α That Minimizes Errors in Null Hypothesis Significance Tests. PLoS ONE 7(2): e32734. doi:10.1371/journal.pone.0032734

National Research Council (2001), Small Clinical Trials: Issues and Challenges. Washington, DC: The National Academies Press.

Bacchetti, P., Deeks, S., and Mc Cune, J. (2011), "Breaking Free of Sample Size Dogma to Perform Innovative Translational Research" *Sci Transl Med 15 June 2011 3:87ps24. [DOI:10.1126/scitranslmed.3001628]*