On the Topic of Sample Size

(and Representative Sampling)

Bill Ross

CEO, Sigma Science Inc.

What is sample size?

Sample size is the number of customers, survey respondents, components, sub-assemblies or other *experimental units* included in a study (e.g., market research, surveys, experimental designs). One of the most common questions statisticians get is “what is the appropriate sample size?” Sample size calculations are described in most conventional statistical textbooks and the equations vary with each situation (e.g., known or unknown variance). The wide range of formulas used for specific situations and designs makes it difficult to decide which method to use. Moreover, these calculations are sensitive to errors because small differences in selected parameters (e.g., size of effect to be detected (Δ), variance (σ^2), α & β risks and power) can lead to large differences in the sample size. An example sample size equation for the number of samples for each treatment of a designed experiment:

$$N = [(z_\alpha + z_\beta)^2] [\sigma^2/\Delta^2]$$

In most cases there is NOT have enough information to calculate the theoretical sample size (N), and therefore estimates may be inadequate estimates. Arbitrary selection of minimum sample sizes without understanding the situation may be both costly and inefficient. Also making decisions on a sample size of N=1 (or less) is also unlikely to be effective.

“Life is short and the world is immense…The essence of science is intelligent sampling, not sitting in single place and trying to get every last one.”

Stephen Jay Gould

Sampling

The most important question regarding sampling is not how many, but how <u>representative</u> are the samples. Representative of, for example:

1. the population of consumers you hope to sell to (e.g., school-age children, homeowners, contractors),
2. the noise (the conditions under which your product will be subject to in the hands of the customer) you hope to be robust to (e.g., environmental factors, raw material variation, installation),
3. current and future conditions (e.g., new ink formulations, lot-to-lot variation of the incoming materials, ambient conditions)

The selection of representative samples has a significant effect on inference space (the area over which you draw conclusion). In consumer research, selecting the appropriate subjects for the study has a significant bearing on who you may be able to sell your product to. In research and design engineering, selecting the appropriate noise to include in the experiment has a significant impact on whether your products will be robust. In logistics, selecting the supplier to provide components may have a significant effect on incoming materials variation.

Representative

To increase the chances our samples are representative there are two approaches[1]:

1. Enumerative: Increase sample size and acquire the samples randomly (to remove sampling bias). This is the approach used when knowledge about the population is poor (ironically, this is when an actual sample size calculation may be beneficial for guidance, but there is a lack of information to make the calculation meaningful), or
2. Analytical: Choose samples based on hypotheses about underlying causal relationships. This requires an understanding of the population being sampled or at least enough of an understanding to develop reasonable explanations as to why variation exists.

For example, let's say there is interest in understanding variation occurring day-to-day (as there is an hypothesis raw material lots effect product performance and lots change day-to-day) in our manufacturing processes, this would require samples taken over multiple days. An **infinite** sample in one day would provide NO insight to day-to-day variation.

By and large sample size determination is an enumerative exercise, descriptive statistics used to draw conclusions on that which already exists (e.g., comparison of two types of product, your product vs. a competitors). Sample size equations are NOT intended to increase your confidence in your ability to extrapolate conclusions into the future (e.g., inferential statistics used for predicting the performance of your product in the future). They are meant to increase confidence in conclusions about what already exists.

[1] Deming, W. Edwards (1975), "On Probability As a Basis For Action", *The American Statistician*, 29(4), 1975, p. 146-152

“To describe a big data set, you only need a representative sample. How big is big enough? Enough to make it representative, which requires understanding the problem you're trying to answer.”

David Hand

Reasons

There are a number of reasons for sample size consideration, to name a few:

1. Required by Regulatory Agencies and certain publications,
2. Required to make specific claims about your products (e.g., reliability),
3. Determine and perhaps reduce α errors,
4. Improve ones confidence in the findings of a study,
5. Avoid bias in the conclusions of the study,
6. Reduce measurement error

For consumer research, it is naturally neither practical nor feasible to study the whole population. Hence, a subset of participants is selected from the population, which is less in number (size) but chosen so they adequately represent the population of interest and so true inferences about the population can be made from the results obtained. This subset of individuals is known as the “sample.” In a statistical context, the “population” is defined as the complete set of individuals. The “sample” is a subset of individuals with specific characteristics (e.g., age, demographics) you want to use to study your product (e.g., children between ages 8-15, contractors building spec. homes). The product may be studied for performance, feature likes and dislikes (e.g., sensory perceptions, reliability, ease of use). Thus a “sample” is a portion, piece, or segment that is meant to be representative of the whole.

For product development, research and design, it is imperative samples are acquired over conditions that represent **reality now and in the future**. For example: How the product will be manufactured, what raw materials will be used, how it will be distributed, stored and how it will be used IN THE HANDS OF THE CUSTOMER! These are not sample size questions, nor in most cases do we have enough information to make sample size calculations. These require analytical thinking to acquire representative samples and thereby predict the product performance IN THE FUTURE.

Conclusion

Determining the appropriate sample size is completely dependent on the situation. In some situations (studying a poorly understood population), huge sample sizes will be needed where in others (fractional factorial screening designs) a relatively small number of samples may be adequate. The important concept: diagnose and understand the situation, critically think and choose a REPRESENTATIVE sample to build confidence in conclusions you draw from evaluation of the samples.