# Small Samples: Does Size Matter?

*Andrew John Anderson and Algis Jonas Vingrys*

"Only five subjects in a scientific study? I trust this is a typographical error. . . . "[1] In all scientific studies, investigators must consider how large a sample should be to reflect the population from which it was drawn. Some studies are designed to quantify the magnitude of a particular parameter in the population (e.g., average flicker sensitivity)[2] or to compare parameters between different populations (e.g., treated and control groups), and in these cases power analyses are accepted methods for determining how large a sample should be.[3] However, there are other types of studies in which investigators demonstrate new effects within a system but do not explicitly quantify population parameters. Many of the psychophysical and neurophysiological studies reported in major journals fit this latter category. Typically, these studies use small numbers of subjects and show that all the subjects tested demonstrate the investigated effect—for example, two rhesus monkeys[4] or two human observers with rod dysfunction,[5] three human observers,[6] four rats,[7] five human observers.[8] However, the method for determining the number of subjects is rarely, if ever, stated. How can these small sample sizes be reconciled with other studies investigating novel effects that use markedly larger sample sizes (e.g., 23 human subjects,[9] 40 human subjects[10])?

It could be argued that studies using small sample sizes are not meant to quantify general performance within a population but merely to document the existence of an effect, and so the number of subjects is less important. However, the fact that investigators bother to perform replications in such studies implies a wish to demonstrate that their findings are not aberrant and should be taken as representing the performance of the population at large. Why, therefore, is the ability of these studies to predict the population's performance not considered? Can an author justify the extra costs (in time and money) in testing four subjects, when he or she may just as well test only two (or even one)?

This issue becomes even more important when considering that large subpopulations can exist within a population. An obvious case is gender. A naive investigator could perform an experiment on three randomly selected subjects and arrive at the conclusion that all people are female. Although such an example may seem ridiculous, it highlights the effects that sampling artifacts can have, especially when subpopulations exist. Therefore, the question that begs consideration is: what sample size is required to ensure, to a specified confidence, that the results are indicative of the general population?

We will consider the situation in which the presence of a previously undocumented effect is to be investigated. The following assumptions are made:

1. Using a particular experimental paradigm, or set of paradigms, the effect is either present or absent; that is, equivocal results are not found.
2. In the group of subjects tested, all subjects show the effect (which we will term "serial successes"). The number of serial successes is therefore equal to the sample size, $N$.
3. The group of subjects is randomly chosen from a selectively normal population.

If assumption 1 is taken to be correct, then the probability of the effect being present can be described by a binomial distribution. Even if the effect is, in fact, part of a continuum, it will typically be rendered binomial by some criterion based on statistical testing (that is, findings are either significant or nonsignificant). For example, a study may investigate the effect of exercise on pulse rate. Although pulse rates represent a continuum (as might the effects of exercise), subjects will either show significantly altered rates or not. In a well-designed study, it is likely that the presence of the effect in each subject will be confirmed using a number of experimental paradigms and rigorous statistical analysis.

Assumption 2 is reasonable and realistic, given that the majority of studies using small sample numbers report serial successes. The situation in which subjects who do not show the effect are present is necessarily more complex and will not be discussed, except to say that any departure within a small sample necessitates a more thorough investigation with enlarged sample numbers.

Assumption 3 needs further consideration. The term selectively normal is used, because many studies have selection criteria for their subjects (e.g., criteria for general health, color vision, visual acuity). As such, subjects are not sampled from the entire population, but from a criterion-determined subpopulation (a selectively normal population). However, it is important to note that samples are often a more narrow subset than stated. Selection from undergraduate or postgraduate students, for example, will result in an overrepresentation of young, educated, myopic subjects, even if age, educational status, and refractive error are not specified as selection criteria. Similar sampling artifacts can unwittingly manifest in animal studies as well.[11]

If we accept these underlying assumptions, then $\theta$ can be used to describe the proportion of the selectively normal population that shows the effect being investigated. For any number of serial successes ($N$) in the sample group, this result is always consistent with $\theta = 1$—that is, the entire population shows the effect. This defines the upper limit on the population proportion, $\theta$. What is more important is to find the smallest population proportion that is consistent with the observed number of serial successes. Taking the common statistical criterion of $P = 0.05$, then the lower limit for $\theta$ provides the minimum population proportion for the effect, with a 95% confidence, given a number of serial successes, $N$. Stated another way, if the population proportion were any smaller than the lower limit on $\theta$, there would be a greater than 1 in 20 chance that, in $N$ subjects, the effect would not be shown (that is, a failure would be present).

From the Department of Optometry and Vision Sciences, The University of Melbourne, Carlton, Victoria, Australia.

Submitted for publication August 9, 2000; revised October 23, 2000 and January 3, 2001; accepted February 15, 2001.

Corresponding author: Algis Jonas Vingrys, Department of Optometry and Vision Sciences, The University of Melbourne 3010, Keppel and Cardigan Streets, Carlton, Victoria 3053, Australia.
a.vingrys@optometry.unimelb.edu.au

**TABLE 1.** Minimum Population Proportions Consistent with *N* Serial Successes, Given Statistical Criteria of $P = 0.10$, 0.05, and 0.01

| Successes ($N$) | % Minimum Population Proportions ($\theta_{min}$) | | |
|---|---|---|---|
| | $\theta_{min}$ ($P = 0.10$) | $\theta_{min}$ ($P = 0.05$) | $\theta_{min}$ ($P = 0.01$) |
| 1 | 10 | 5 | 1 |
| 2 | 32 | 22 | 10 |
| 3 | 46 | 37 | 22 |
| 4 | 56 | 47 | 32 |
| 5 | 63 | 55 | 40 |
| 6 | 68 | 61 | 46 |
| 7 | 72 | 65 | 52 |
| 8 | 75 | 69 | 56 |
| 9 | 77 | 72 | 60 |
| 10 | 79 | 74 | 63 |
| 22 | 90 | | |
| 29 | | 90 | |
| 44 | | | 90 |
| 45 | 95 | | |
| 59 | | 95 | |
| 90 | | | 95 |

The following equation describes the range of values $\theta$ can take:

$$\theta^N \geq 0.05$$

where $\theta$ is the population proportion (as a fraction), $N$ is the number of serial successes (and is equivalent to the sample size), and 0.05 is the level of confidence (1 in 20). The equation is derived from that given by Clopper and Pearson[12] for the calculation of binomial distribution confidence limits. Solving for the minimum value of $\theta$ ($\theta_{min}$, as a percentage) gives the column headed $\theta_{min}$ ($P = 0.05$) in Table 1.

What should the criterion for $\theta_{min}$ be? For an unknown effect, a useful starting point is that an effect must be present in the majority of the population if it is to be classified as "normal"; that is, $\theta_{min}$ must be at least 50%. Using this assumption (as well assumptions 1–3) a sample size $N = 5$, all showing the effect, is required to confidently ($P = 0.05$) say that the population proportion for the effect is greater than 50%. The sample size must be increased if subjects who do not show the effect are present (that is, serial successes are not achieved). For completeness, Table 1 also lists the relationship between $\theta_{min}$ and sample size for $P = 0.10$ and $P = 0.01$. Using these criteria, sample sizes of four and seven, respectively, are required to be consistent with a population proportion of at least 50%.

To provide more confident estimates of the population proportion, much larger numbers are needed. For example, to be confident ($P = 0.05$) that the population proportion is at least 95%, 59 subjects showing the effect would be required. Such studies, however, are rarely performed. Instead, it is more common for data to be collected on a smaller sample, whose size is determined by a power analysis and mean values for the magnitude of the effect compared with conventional statistical analyses (e.g., *t*-tests). It should be noted, however, that these latter types of analyses determine whether a significant effect exists in the population on average and provide no estimate of the population proportion, $\theta$. Such analyses may be successfully used on small-sample-size psychophysical data.[13]

It should also be noted that a study may not be designed to quantify the performance of a normal population, but that of a disease group instead.[5] The model outlined herein is identical, however, except that the predicted values for $\theta_{min}$ now relate to the population of observers with a particular disease, instead of the normal population.

It is possible that the model can be improved. Often, an investigated effect is shown to be dependent on, or correlate with, a previously documented effect. In such cases, the estimated population proportion of this previously documented effect provides additional information about the population proportion of the investigated effect, and so a more confident estimation of $\theta$ may be made than that given in Table 1. As such, it may be possible to use reduced numbers of subjects to clarify aspects of documented "normal" effects. However, there are also instances in which the outcomes of similar experiments differ between authors. In such cases, the estimated population proportion of the previously documented effect provides additional knowledge that reduces our confidence in our estimation of $\theta$. It should be emphasized, however, that the reliability of such previous studies depends on the number of subjects investigated and the soundness of the studies' experimental designs.

It is possible that some form of Bayesian logic could be used to combine the results of previous small-sample-size studies with new studies, in a way similar to that proposed for clinical decision making.[14] Until the validity of such a model has been established for the type of data discussed in this article, the approach outlined herein provides a starting point for determining the general applicability of studies making use of small sample sizes. Despite criticisms,[1] a sample size of five may well be useful in scientific research.

In summary, the model outlined allows predictions to be made from experimental data obtained from limited numbers of samples. Our approach is appropriate for studies documenting the presence of an effect in each of a small number of subjects and allows inferences to be made regarding the proportion of the population expected to show the same effect. As such, the model may be usefully employed in small-sample-size psychophysical investigations, so that the general applicability of results may be predicted. In addition, the model may be used to estimate the number of subjects needed to determine, to a desired statistical confidence, the prevalence of an effect. Our approach is not applicable to analyzing the magnitude of a particular effect within a population, however; conventional power analyses and statistical testing are available for this task.

## *References*


1. Norris E. Downsized sample. *New Scientist*. November 1999:60.
2. Tyler CW. Two processes control variations in flicker sensitivity over the life span. *J Opt Soc Am A*. 1989;6:481–490.
3. Cohen J. *Statistical Power Analysis for the Behavioural Sciences*. New York: Academic Press; 1969:1–16.
4. Fuster JM, Bodner M, Kroger JK. Cross-modal and cross-temporal association in neurons of frontal cortex. *Nature*. 2000;404:347–351.
5. Hansen RM, Fulton AB. Background adaptation in children with a history of mild retinopathy of prematurity. *Invest Ophthalmol Vis Sci*. 2000;41:320–324.
6. Freeman TCA, Fowler TA. Unequal retinal and extra-retinal motion signals produce different perceived slants of moving surfaces. *Vision Res*. 2000;40:1857–1868.
7. Laubach M, Wessberg J, Nicolelis MAL. Cortical ensembles activity increasingly predicts behaviour outcomes during learning of a motor task. *Nature*. 2000;405:567–571.
8. Braun C, Schweizer R, Elbert T, Birbaumer N, Taub E. Differential activation in somatosensory cortex for different discrimination tasks. *J Neurosci*. 2000;20:446–450.
9. Blog MG, Kersten D, Hurlbert AC. Perception of three-dimensional shape influences colour perception through mutual illumination. *Nature*. 1999;402:877–879.

10. Bonato F, Cataliotti J. The effects of figure/ground, perceived area and target saliency on the luminosity threshold. *Perception Psychophys*. 2000;62:341–349.
11. Ward GE, Wainwright PE. The contribution of animal models to understanding the role of fats in infant nutrition. In: Huang Y, Sinclair AJ, eds. *Lipids in Infant Nutrition*. Champaign, IL: American Oil Chemists Society Press; 1998:39–62.
12. Clopper CJ, Pearson ES. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*. 1934;26:404–413.
13. Anderson AJ, Vingrys AJ. Interactions between flicker thresholds and luminance pedestals. *Vision Res*. 2000;40:2579–2588.
14. Aspinall P, Hill AR. Clinical inferences and decisions. I: diagnosis and Bayes' theorem. *Ophthalmic Physiol Opt*. 1983;3:295–304.

## New Developments in Vision Research

Written for a broad audience, the articles in this column succinctly and provocatively review a rapidly changing area of visual science that shows progress and holds potential. Authors and topics are chosen by the Editor-in-Chief in collaboration with the Editorial Board.

To avoid bias, the Editor-in-Chief subjects these articles to the same rigorous peer review process to which all other *IOVS* articles are subjected. Space and reference limitations are imposed on the authors.

The purpose of this series is not the recognition of individual scientists, nor exhaustive review of a subject, but the stimulation of interest in a new research area.

*Editor-in-Chief*