Pergamon

Marine Pollution Bulletin Vol. 38, No. 6, pp. 439–447, 1999

Printed in Great Britain

PII: S0025-326X(99)00048-X

0025-326X/99 $ – see front matter

VIEWPOINT

How Large should my Sample be? Some Quick Guides to Sample Size and the Power of Tests

CHARLES R. C. SHEPPARD

Department of Biological Sciences, University of Warwick, Coventry CV4 7AL, UK

This article provides a simple explanation of how power analysis can help us determine how large a sample should be. It covers basic examples of three commonly used tests: *t*-test, χ^2 test and analysis of variance. It presents several graphs which can be used to obtain important critical values, such as required sample size for a particular task, and the power (or ability) of a chosen statistical test to actually determine the task required of it.

Keywords: power; statistics; sample size; probability.

Introduction

In marine environmental science we use power analysis much less frequently than we should. In some fields such as medicine, its use is almost mandatory for determining the sample sizes needed in, say, drug trials, but it would be equally valuable in environmental assessments, pollution monitoring and several other forms of environmental work. The lack of use of power analysis may partly be due to a poor understanding of what it is, what it can tell us and of how it works. Nevertheless, it is important because it gives us the means to calculate the sample size we need to detect a given change, and it can tell us the probability that our sampling regime or experiment can actually detect an effect if one exists.

The number of possible permutations involving sample size and power of a test is enormous, but the following provides some generic examples. Of course, researchers should still conduct their own, more specific analyses for a particular experiment or set of samples or results. There are available today several excellent software packages which make the analysis so straightforward that there is now no excuse for not doing so (see Thomas and Krebs, 1997). Further, some of them are available at no cost.

What is Power Analysis?

To express it at its simplest, assume we are performing a *t*-test to examine whether the means of two samples are the same or different. The two samples may be from a test site and a control site, or they may be samples from the same site taken before and after an event such as a pollution incident. Power analysis shows us whether the experiment is likely to be capable of detecting an environmentally important (or predetermined) difference in the mean values of our two samples.

If our test shows that there is a significant difference between our two samples, then usually we need look no further. If, with the level of confidence we have decided upon, such as $p < 0.05$, the two samples are 'different', then that is all we need to know (but see later). But what if the test showed no significant difference? This could mean that there actually was no underlying difference, and this is what is commonly concluded. However, it could also mean that our samples were too few to detect the difference, the difference between the means of the two samples (the effect size) might have been real but small, or, that the variance in each sample was too large, so our *t*-test could not have shown it. Power analysis shows us, in effect, the probability that the *t*-test could have shown a difference were there to have been one in reality.

In more statistical language, if we measure a large *t* value, we invariably state something to the effect of 'the means of the two populations are significantly different, $p < 0.05$' (or $p < 0.01$, or other value). What we mean is that we have rejected the null hypothesis (H_0) which stated (even if we did not explicitly spell it out) that there is no difference between the two populations. If the *p* value calculated is smaller than 0.05 it means that there is less than a 5% chance that the observed 'significant difference' could have occurred by chance alone. On this basis we conclude that there is a significant difference between the two samples. There remains, of course a 5%

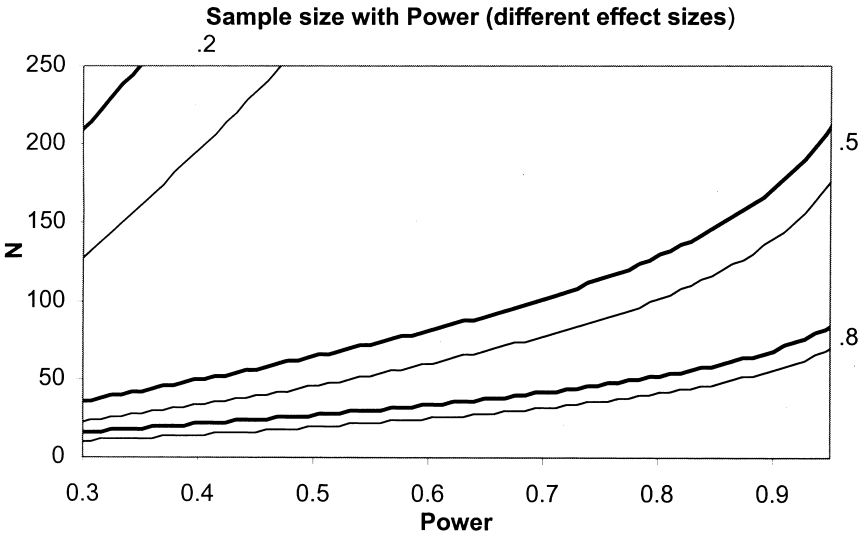

Fig. 1 *t*-test curves. Change in sample size (N) with power for three different effect sizes: small (0.2), medium (0.5) and large (0.8). Thick lines are for two-tailed tests, thin lines are for one-tailed tests. α for all curves was 0.05.

chance that the difference recorded was a fluke, that the test showed a difference even if in reality there was none. Indeed, when we use an α of 0.05 and perform one hundred similar experiments using pairs of samples where we somehow know that there really is no difference between them, the p value from the statistical test will fall below this α about 5 times. We make an error in these five cases, when we declare that there was a significant difference between the two samples, but a certain amount of uncertainty is generally inevitable and the price of doing any statistics in the first place. This is a 'Type I error'. We would expect to make such an error 5% of the time. We could reduce the chance of making such an error if we use a more strict α value, and by requiring $p < 0.01$ instead of $p < 0.05$. With this 'cut-off' there would be only a 1% chance of making the error. The t value obtained needs to be larger to achieve this greater certainty, but if it does, we would be 99% certain that two samples which were not different in reality would not appear to be. This is understood by most who use these simple statistics.

There is, however, a complementary problem. In our t-test, we may compute a t value which is low, which says in effect that p is not less than our cut-off level of 0.05. If this happens, we accept the null hypothesis H_0 which says that there is no significant difference between our two samples. We say that there is no difference between the two samples. But if, in nature, there really was

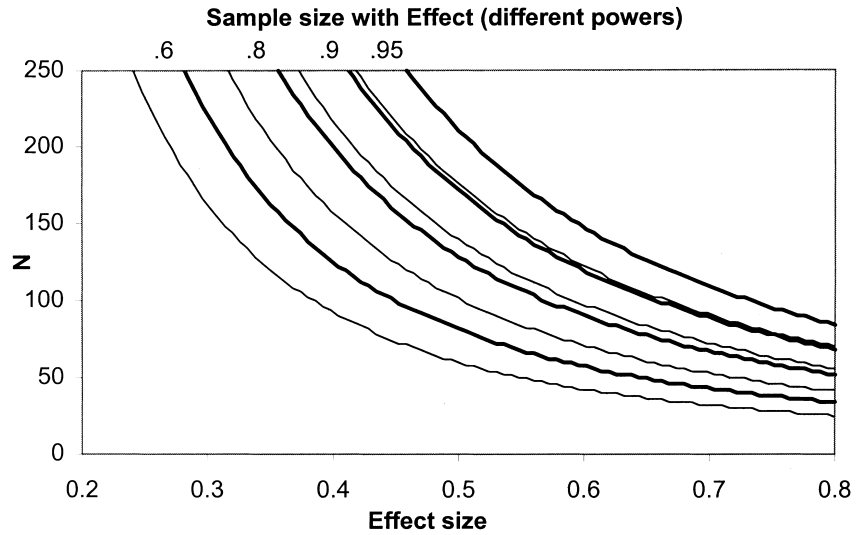

Fig. 2 *t*-test curves. Change in sample size (N) with Effect size for several different powers (0.6, 0.8, 0.9 and 0.95). Thick lines are for two-tailed tests, thin lines are for one-tailed tests. α for all curves was 0.05.

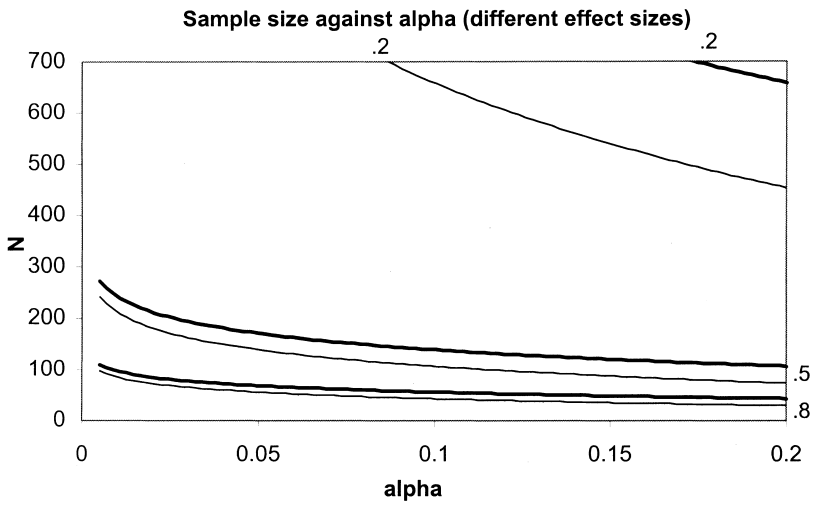

Fig. 3 *t*-test curves. Change in sample size (N) with α for different effect sizes: small (0.2), medium (0.5) and large (0.8). Thick lines are for two-tailed tests, thin lines are for one-tailed tests. Power in this set was 0.9 throughout. Note that the y-axis on this graph shows up to $N = 700$ compared with $N = 250$ in the previous two graphs.

TABLE 1

Example of effect size in *t*-tests. If mean value of first sample = 100, then mean value of second sample is 110, 125, 150 ... etc. as down the rows. Or, if the mean value in first sample is 10, then the rows indicate mean values in second samples of 11, 12.5, 15, etc. The SD within each group is shown as a multiple of the first sample mean. Thus with a sample mean = 100, columns are for SDs of 10, 25, 50 ... 500. Or, if the mean = 10, then SDs would be 1, 2.5, 5 ... 50. N.B. A small effect is 0.2, medium effect is 0.5 and a large effect is 0.8 or greater (Cohen, 1988).

Mean of sample 2 relative to sample 1	SD (× 1st mean)							
	0.1	0.25	0.5	0.75	1.0	1.5	2.0	5.0
×1.10	1	0.4	0.2	0.133	0.1	0.07	0.05	0.02
×1.25	>1	1	0.5	0.333	0.25	0.167	0.125	0.05
×1.50	>1	>1	1	0.667	0.5	0.333	0.25	0.1
×1.75	>1	>1	>1	1	0.75	0.5	0.375	0.15
×2.00	>1	>1	>1	>1	1	0.667	0.5	0.2
×5.00	>1	>1	>1	>1	>1	>1	>1	0.8

a difference, we have made the opposite kind of error, a Type II error. We have accepted the null hypothesis when there really was a difference. The probability of making this type of error is called β. Statistical power is $1 - \beta$ or the probability of correctly rejecting the null hypothesis, or it is the probability of being right when stating that there is no difference between the two samples.

The importance of sufficient power

If the sampling regime was too small, our test may not be adequate to detect a difference that in reality is there. The smaller our sample, or the smaller the true difference if it exists, the greater is the probability of accepting the null hypothesis in error, and this probability is increased with increasing variance in the samples. The importance of this to our experiments or surveys should be obvious. If, for example, there is a small fall in the mean diversity in an important site, or if there is a rise in the contaminant level, our *t*-test may lead us to accept H_0 and say that there is no significant difference or no deterioration when in reality there is. On this basis, no action in terms of contaminant clean-up, or site protection is likely to be the result, especially if that action would be expensive. Consequences of this are very clearly expressed by Peterman and M'Gonigle (1992) in the context of pollution regulation: "... the monitoring programme may not be a reliable source of information because it is likely (with probability = 1 − power) to have failed to detect such an effect, even if it was present. In that case, little confidence should be placed in results from the monitoring programme." In other words, with inadequate power, the work was likely to have been not only a waste of time, but misleading. They go on: "These concepts of type I and type II error emphasise an asymmetry in our current approaches to environmental regulation." They point out that an industry lawyer may legitimately ask a marine scientist how the latter can be sure that an effect which has been discovered is not in reality due to some other natural event rather than being caused by the industry, or he may legitimately question whether the effect is really there at all; that the scientist

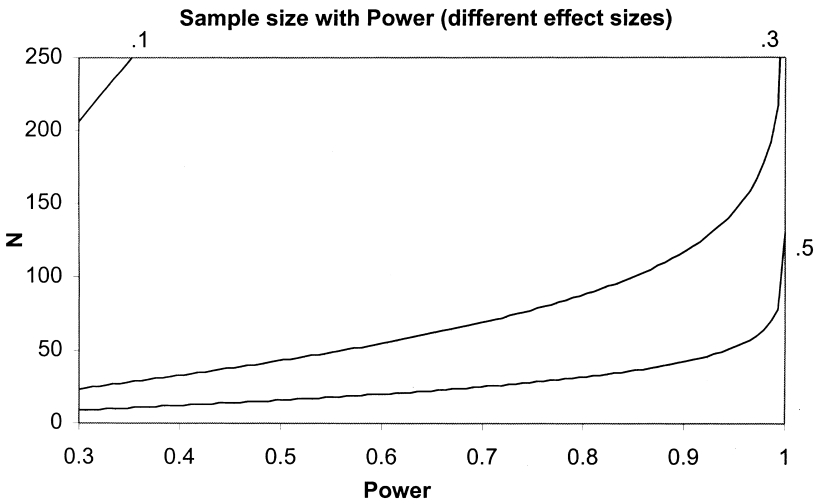

Fig. 4 χ^2 test curves. Change in sample size (N) with power for three different effect sizes, small (0.1), medium (0.3) and large (0.5). α for all curves was 0.05.

has perhaps made a type I error. But: "While this is a legitimate question, there is an equally legitimate one concerning type II errors *that is almost never asked* when the biologist or industry *fails* to reject the null hypothesis of no effect. That is, 'How do you know that the absence of a statistically significant effect in a monitoring programme or experiment is not just due to small sample size or sampling variation, which tend to reduce the chances of detecting an effect that is actually present?' " (Peterman and M'Gonigle, 1992, italics in original).

The commonest factors which make a test unable to detect a change include too few samples, too small a difference between the two means, and a large variation in the values making up the means. Thus it is when a test does *not* show a difference that power analysis is especially important. It is needed to show whether or not the test *could* have shown a difference where difference existed in reality. The ideal of course is to use power analysis before designing an experiment or sampling programme, but it can equally be used in a post-hoc way, where the data available include the means and standard deviations of the samples already obtained, and of course the numbers of samples used, so that the probability of a correct rejection of H_0 can be estimated.

A balance of probabilities

Understandably, there is a trade-off between the chances of making a type I and a type II error. In an experiment, if we wish to be extremely sure of demonstrating an effect, we could set α at a low value, say 0.01 or even smaller. If our t-test still gives us a $p < 0.01$ then we are even more confident that there is a difference between the two samples. But with a small α we have also inadvertently increased the chance of saying that there is no difference in cases when in fact there is. There

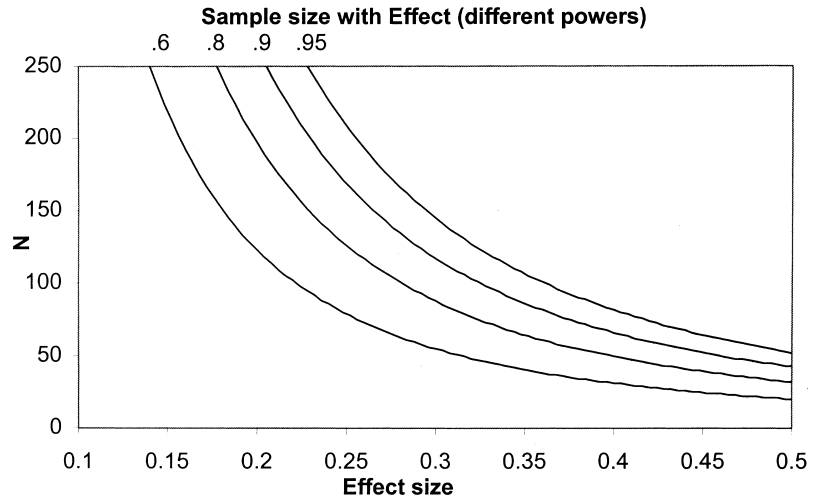

Fig. 5 χ^2 test curves. Change in sample size (N) with Effect size for several different powers (0.6, 0.8, 0.9 and 0.95). α for all curves was 0.05.

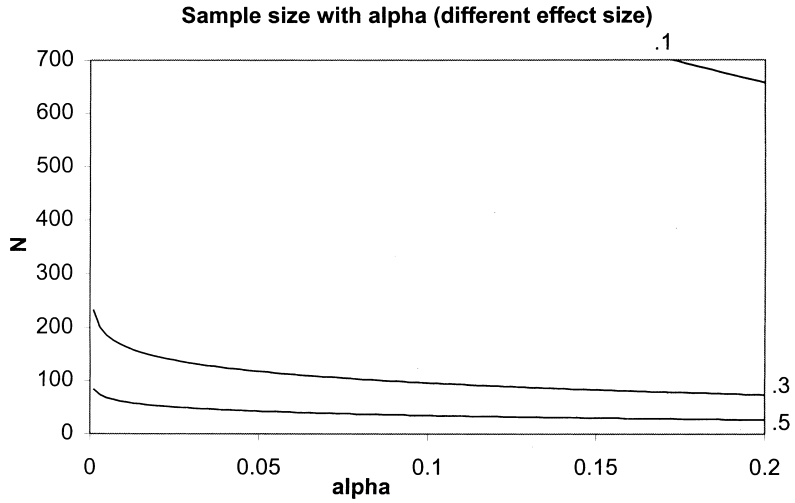

Fig. 6 χ^2 test curves. Change in sample size (N) with α for different effect sizes: small (0.1), medium (0.3) and large (0.5). Power in this set was 0.9 throughout. Note that the y-axis on this graph shows up to $N = 700$ compared with $N = 250$ in the previous two graphs.

TABLE 2

Example of effect size in χ^2 tests. Effect in this example is measured as difference from a 2×2 table whose four cells all have a probability of 0.25. For simplicity, two cells are kept as 0.25, so only two change (one increasing and one decreasing). N.B. A small effect is 0.1, medium effect is 0.3 and a large effect is 0.5 or greater (Cohen, 1988).

	Differences from the H(0) of 0.25, 0.25, 0.25, 0.25						
	0.03	0.04	0.05	0.1	0.15	0.2	0.25
Cells of the table	0.25	0.25	0.25	0.25	0.25	0.25	0.25
	0.25	0.25	0.25	0.25	0.25	0.25	0.25
	0.22	0.21	0.2	0.15	0.1	0.05	0
	0.28	0.29	0.3	0.35	0.4	0.45	0.5
Effect size	0.08	0.11	0.14	0.28	0.42	0.57	0.71

is a trade off, but knowing this, we can design our sampling programme to include, for example, a greater number of samples, to reduce the probability of saying that there is no proven difference when in nature there was. As Peterman and M'Gonigle (1992) point out, in many past evaluations of licensed pesticides statistical power was low, so that substances which were shown not to have caused effects in experiments were assumed to be safe, but in fact turned out to be harmful.

Methods and definitions

Data in all tables and sets of graphs were calculated using *G*Power. For details see Erdfelder *et al*. (1996) or http://www.psychologie.uni-trier.de:8000/projects/gpower.html from where the software is available. The latter has extensive on-line documentation, including algorithms used. This (free) software was specifically selected from alternatives to illustrate further that there is now little obstacle to use of power analysis! Output data were imported into Excel for producing the graphs shown here.

Several 'varieties' of each test exist, (e.g. one-tailed and two-tailed t-tests, different grid sizes in the case of χ^2 , and numerous kinds of analysis of variance). In all cases, simple options were used here, since the present purpose is to illustrate. In the case of Anovar, the simplest variety where two sets of samples are compared, is closely analogous to the t-test and is mathematically equivalent to it (Sokal and Rolf, 1981) so in this case a four group analysis of variance is shown.

The emphasis in the following is on sample size, so all graphs have N as the y-axis. This is shown against either power, effect size or α.

A definition of effect size is needed. Effect size is a measure of the difference between the two samples (or between the empirical data and the theoretical expected values as given by a null hypothesis). A difference could be the depression of a diversity index compared to a control site, or it could be the increase in pesticide quantity measured in tissue. However, effect size is usually not a simple difference between, for example, two means. For a t-test, effect size is affected consider-

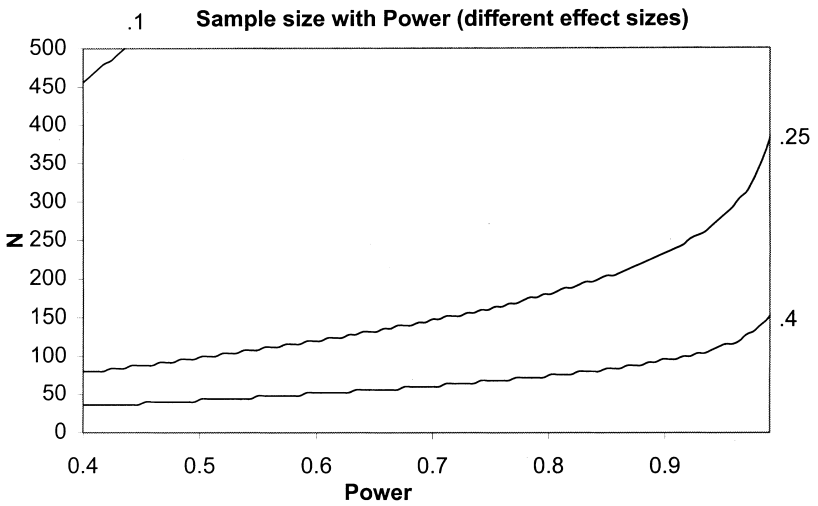

Fig. 7 Anovar curves. Change in sample size (N) with power for three different effect sizes, small (0.1), medium (0.25) and large (0.4). α for all curves was 0.05.

ably by the variation in the samples making up the mean values. Table 1 shows how effect changes as a function of both the difference between the mean values of the two compared samples and their standard deviation. As the difference between the means increases, so effect size increases, for a given standard deviation. For a particular pair of means on the other hand, as standard deviation increases, the effect size becomes smaller. Thus, to detect a given effect size, as standard deviation increases, sample size needs to increase. Table 2 shows analogous data for χ^2 tests. Here a basic, or reference, 2×2 table is used in which all cells have equal probabilities of 0.25. The 'test' which is compared against this is one in which two cells change, one increasing and one decreasing (the other two cells remaining fixed at 0.25 for simplicity). As the difference from the equal probability condition increases, so does effect size, and this is shown in the table. Of course, if all four cells rather than just two were varied from the 0.25 reference value, effect size, which is measuring the difference from the equal probability condition, would increase further. Table 3 shows effect sizes for a simple analysis of variance with four groups. Its structure is broadly similar to that of t-tests in Table 1. Each row of the table, however, refers to a set of four groups whose four means differ by various increments starting with a mean of 10. Thus in the row showing an increment of 1, the means of the four groups would be 10, 11, 12 and 13, and the row showing an increment of 5 shows the data for a four group Anovar whose means are, for example, 10, 15, 20 and 25.

It should also be noted that a statistically significant test statistic in a sample does not necessarily imply a scientifically important effect in the underlying population. With a huge sample size, for example, even a very small effect with no biological importance whatever could generate statistically significant sample statistics. Thus power and sample size are important in cases of significant results too. To make sure that a negligible effect does not generate statistically significant sample results, α could be 0.01 or even 0.001 (Erdfelder, pers. comm.). The software used may have a means of helping determining optimum α levels; G*Power offers a

TABLE 3

Example of effect size in Analysis of Variance with a simple four group test. The 'Increments in' column shows the range of means in the four groups used in each test, based on a first group mean of 10. Thus an increment = 0.5 is the case where the four means are 10, 10.5, 11 and 11.5 and, with an increment = 2, means are 10, 12, 14, and 16. The SD (within each group) is shown as a multiple of the first sample mean. Thus with the first group mean of 10, the columns show data for within group SDs of 1, 2.5, 5, 7.5, 10 ... 50. (The table requires that the four groups have equal sample sizes; keeping the means and SDs equal but changing the relative sample sizes, even if total N remains the same, also changes effect size. Further, the increments in column 1 are based on a first group mean of 10; increments of 1 based on a mean in the first group of, say 23, giving group means of 23,24,25,26, would again give different effect sizes.) N.B. A small effect in Anovar is 0.1, a medium effect is 0.25 and a large effect is 0.4 or greater (Cohen, 1988).

Increments in:	SD (× 1st mean)								
	0.1	0.25	0.5	0.75	1	1.5	2	3	5
0.5	0.56	0.22	0.11	0.07	0.06	0.04	0.03	0.02	
1	> 1	0.44	0.22	0.15	0.11	0.04	0.06	0.04	
2	> 1	0.89	0.45	0.30	0.22	0.15	0.11	0.07	
3	> 1	> 1	0.67	0.45	0.33	0.22	0.17	0.11	0.07
4	> 1	> 1	0.89	0.60	0.45	0.30	0.22	0.15	0.09
5	> 1	> 1	> 1	0.75	0.56	0.37	0.28	0.19	0.11

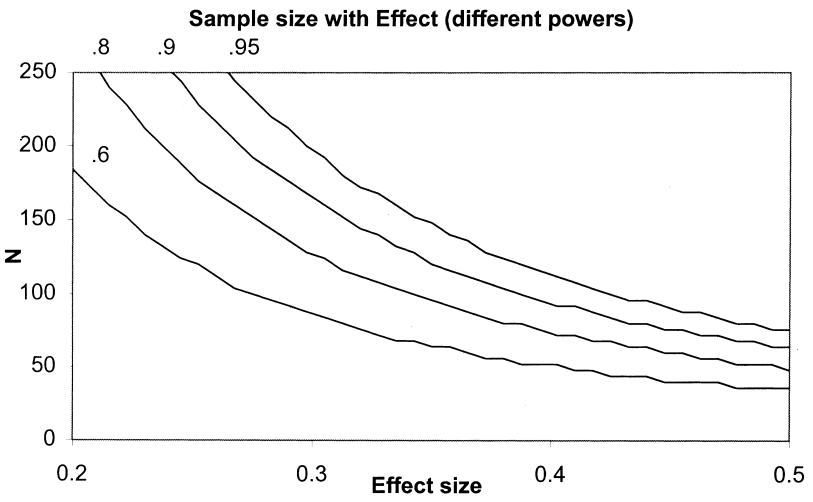

Fig. 8 Anovar curves. Change in sample size (N) with Effect size for several different powers (0.6, 0.8, 0.9 and 0.95). α for all curves was 0.05.

'compromise power analysis' which does this, for example.

Cohen (1988) has simplified the infinite range of effect sizes by conveniently defining three levels: small medium and large effects. For t-tests, χ^2 and Anovar the actual values themselves for small, medium and large effects are different (Table 4). In the examples of graphs shown, plots are limited to the small, medium and large effect sizes as appropriate for each test. It should be remembered that the meaning of 'small', 'medium' and 'large' are likely to differ between disciplines, and may not directly correspond to a perceived 'degree of importance'. The levels determined by Cohen (1988) for social sciences may not easily correspond to levels in, say, contaminant levels or diversity values.

The parameter β may be set to 0.05, like α, but applications commonly 'relax' this somewhat so that larger values are used (giving lower power since power $= 1 - \beta$). Where necessary, β in the following is set to 0.1 so that power is 0.9.

Taking the three main variables of sample size (N), Power and Effect size, the graphs show how each covaries under different conditions, for t-tests, simple χ^2 tests and a simple Anovar in turn. Many simple values, such as sample size needed in certain situations, can therefore be determined by simply measuring off on these graphs. With G*Power, N is the total number of samples, not the number in each of the groups. The curves in each set assume that the number of samples is identical in each group. If groups contain different sample sizes, the G*Power (or other) software should be referred to, from which power values may be obtained for any configuration of sample sizes.

The point should be made that all the sample size and power analysis methods employed here make the assumption that samples are both representative and independent. The first point is usually addressed by the sampling methodology, while the second is a function of the sampling and any correlations that might exist between the data.

TABLE 4

Small, medium and large effect sizes for t-tests, χ^2 and Anovar (Cohen, 1988). Note that these levels were chosen for a social sciences context and 'large', 'medium' and 'small' may not be equivalent in different biological or ecological contexts.

	Small	Medium	Large
t-test	0.2	0.5	0.8
χ^2	0.1	0.3	0.5
Anovar	0.1	0.25	0.4

t-tests

Figures 1–3 show result curves for one-tailed tests (thin lines) and two-tailed tests (thick lines). Figure 1 shows how N changes with power, using different effect sizes, Fig. 2 shows the value of N required against effect size, for powers of 0.6–0.95, while Fig. 3 shows the required N, for large medium and small effect sizes, for different levels of α. Table 1 shows a set of examples of how effect size varies with the difference between the two means being compared, and the standard deviation of the replicate samples making up the means.

χ^2 tests

Figures 4–6 show the equivalent χ^2 results based on a simple 2×2 table. Fig. 4 shows how N changes with power, using different effect sizes, Fig. 5 shows the value of N required against effect size, for powers of 0.6–0.95, while Fig. 6 shows the required N, for large medium and small effect sizes, for different levels of α. Table 2 gives an example of how effect varies with divergence from a reference table in which all four cells have the same probability of 0.25.

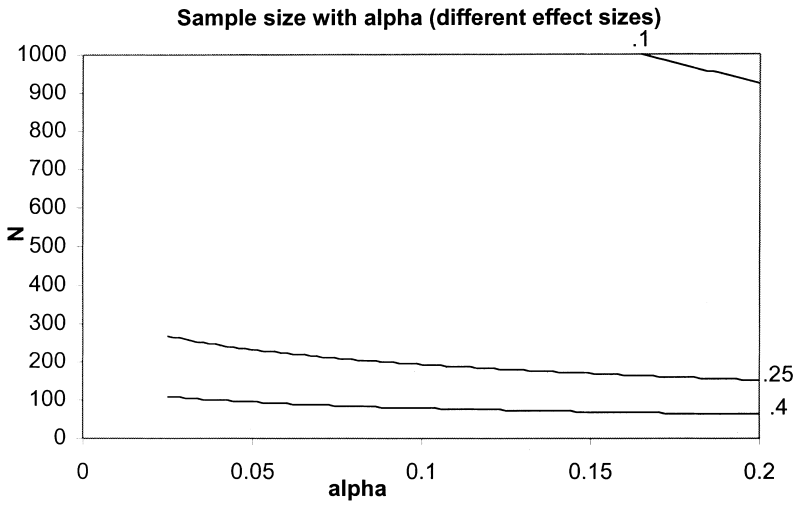

Fig. 9 Anovar curves. Change in sample size (N) with α for different effect sizes: small (0.1), medium (0.25) and large (0.4). Power in this set was 0.9 throughout. Note that the y-axis on this graph shows up to $N = 700$ compared with $N = 250$ in the previous two graphs.

Anovar

Figures 7–9 show results for a four sample Anovar. Figure 7 shows how N changes with power, using different effect sizes, Fig. 8 shows the value of N required against effect size, for powers of 0.6–0.95, while Fig. 9 shows the required N, for large medium and small effect sizes, for different levels of α. Table 3 gives an example of how effect size changes with increasing differences between the means of the four groups in a test, and with within-group standard deviation.

Conclusions

It may be surprising to many to see how large samples need to be in order to be 'adequate' in the sense of having sufficient power to detect the event sought. If the effect sought is a small one, sample size increases enormously. This illustrates, quite simply, the difficulty of reliably being able to measure small changes to an ecosystem, or small changes in a level of a toxic substance, for example. Given the costs often associated with both biological and chemical sampling, and sample processing, small sample sizes may not be surprising, but the fact remains that without consideration of the results of a power analysis, our sample sizes have often been, and may continue to be, too small. The curvi-linear nature of the lines of these figures show that sample size, for example, increases very rapidly for the achievement of greater certainty, or for the detection of a smaller effect, so the result usually must be that a trade-off is selected. If cost determines that samples will be few, it must be recognised that any resulting declaration of no-effect may contain very little certainty.

If, with a small sample size, a 'significant' difference is shown to exist then there is no problem. In this sense there is no minimum sample size which is *required*. It is when the result shows otherwise, then a conclusion stating that there is no significant difference between two samples can be claimed only with a sometimes surprisingly low level of certainty.

There is of course no absolute size of sample which is 'best'. In statistics we tend to think in terms of absolutes, driven by the apparent importance of the almost magic values of 0.05 and perhaps 0.01. These, however, are only convenient 'cut-off' values, a point which is sometimes forgotten. It may be tempting to declare a 'critical number' of samples as being absolutely necessary, but since this would depend on too many other variables, such as effect size and power wanted, there would be too many critical numbers and so this is unlikely to materialise.

A very large effect needs fewer samples for its detection than does a small effect, but of course, the biological or ecological importance of a change does not necessarily correlate in a linear way with the effect size. For example, the amount of a contaminant in body tissue might increase smoothly and linearly along some gradient of coastline, but the biological effect of this increase may not be linear; the increasing contaminant may have little biological effect at first until some threshold is reached, then a little more of it may have a marked impact on the population. Measures of a diversity index of an ecosystem also, still commonly employed by the more simplistic kinds of environmental work (Gray, 1999) likewise may not correlate well or at all with even a linear gradient of impact or pollution. Effect size is a statistical number, while the biological effect on an ecosystem or animal may change slowly or rapidly, or importantly or unimportantly with the environmental gradient being investigated.

It is still the case that many technically sound pieces of work are being done which are not accepted for

publication (see Sheppard, 1998a,b). Reasons are varied, but one is that conclusions may be reached which are of the sort that there is 'no significant difference', while in fact there was no way that the piece of work could have detected any difference even if it were there. This simply wastes effort. Possibly more important is the large body of work done for unrefered output such as environmental assessments, many of which are used as a basis for decisions. It is here that conclusions may be made, and commonly have been, which are quite invalid for the reasons outlined above (see Buhl-Mortensen, 1996; Gray, 1996).

This article does not attempt to cover all the kinds of information that can be obtained from a power analysis. It merely demonstrates the point of it, giving graphs which can be used as generic aids in experimental design. One important omission here, for example, is the *post-hoc* power analysis. In this, we have already done the sampling and we have the results of the analyses. It is too late to go out and collect more samples – the site may even have already been changed by, for example, a coastal development, so that the 'pre-development' number of samples could not be increased or repeated in any event. In a *post-hoc* power analysis, we can, however, still determine the certainty or confidence that can now be placed on those results.

I am very grateful to Dr. Edgar Erdfelder of the University of Bonn, one of the authors of the software used here, for his most helpful comments on the text. Similarly, Dr. Chris O'Callahan in the University of Warwick and Prof. John Gray at the University of Oslo provided several helpful suggestions for clarifying the text.

Buhl-Mortensen, L. (1996) Type II statistical errors in environmental science and the precautionary principle. *Marine Pollution Bulletin* **32**, 528–531.

Cohen, J. (1988) *Statistical Power Analysis for the Behavioural Sciences* 2nd ed. Hillsdale, NJ.

Erdfelder, E., Faul, F. and Buchner, A. (1996) GPOWER: a general power analysis program. *Behaviour Research Methods, Instruments & Computers* **28**, 1–11.

Gray, J. (1996) Environmental Science and the precautionary approach revisited. *Marine Pollution Bulletin* **32**, 532–534.

Gray, J. (1999) The status of science today. *Marine Pollution Bulletin* (in press).

Peterman, R. M. and M'Gonigle, M. (1992) Statistical power analysis and the precautionary principle. *Marine Pollution Bulletin* **24**, 231–234.

Sheppard, C. R. C. (1998a) Patterns in manuscripts submitted to Marine Pollution Bulletin: 1. Subject matter. *Marine Pollution Bulletin* **36**, 316–317.

Sheppard, C. R. C. (1998b) Manuscripts submitted. (2) Regional variations. *Marine Pollution Bulletin* **36**, 402–403.

Sokal, R. R. and Rolf, F. J. (1981) *Biometry*, 2nd ed. Freeman, New York.

Thomas, L. and Krebs, C. J. (1997) A review of statistical power analysis software. *Bulletin Ecological Society of America* **78**, 126–139.