Stat Note 8

In the eighth of a series of articles about statistics for biologists, **Anthony Hilton** and **Richard Armstrong** discuss:

Statistical power and sample size

There are two important questions that should be asked about any experiment. First, before the experiment is carried out, what sample size (N) would it be appropriate to use in a given situation? Second, what is the strength or 'power' (P') of an experiment that has been conducted, *i.e.*, what difference between two or more groups was the experiment actually capable of detecting?

The second question is of particular interest because an experiment in which a non-significant difference is reported confirms the null hypothesis (H_0) that no difference exists between the groups. This may not mean, however, that the hypothesis should actually be rejected because the experiment may have been too small to detect the 'true' difference. In any hypothesis test, the statistical test, *e.g.*, a 't' or 'F' test, indicates the probability of a result if H_0 were actually true and therefore, if that probability is less than 5% ($P < 0.05$), H_0 is usually rejected. The ability of an experiment to reject the hypothesis depends on a number of factors including the probability chosen to reject H_0 (usually set at 0.05), the variability of the measurements, the sample size since larger values of N lead to more accurate estimates of statistical parameters, and the effect size, i.e., the size of the actual effect in the population, larger effects being easier to detect. Statistical software is now available to calculate P' and to estimate N in a variety of circumstances and it is therefore important to understand the value and limitations of this information. This Statnote discusses statistical power and sample size as it relates to the comparison of the means of two or more independent groups using 't' tests or analysis of variance (ANOVA).

How to calculate sample size for comparing two independent treatments

In Statnote 2 (Hilton & Armstrong 2005) we described an experiment to investigate the efficacy of a novel media supplement in promoting the development of cell biomass. Essentially two sets of 25, 10-litre fermentation vessels were filled with identical growth media with the exception that the media in one of the vessels was supplemented with 10ml of the novel compound under investigation. The vessels were then inoculated with a culture of *Bacterium 'x'* and the fermentation allowed to proceed until all the available nutrients had been exhausted and growth had ceased. The dry weight of cells was measured in each flask. A good question might be how many flasks should actually have been used in this experiment?

As a first step, decide on a value δ that represents the size of difference between the media with and without supplement that is regarded as important and which the experiment is designed to detect. If the true difference is as large as δ, then the experiment should have a high probability of detecting this difference, *i.e.*, the test should have a high P' when the true difference is δ. Levels of P' = 0.8 (80%) or 0.9 (90%) are commonly used whereas levels of 0.95 or 0.99 can be set, but are often associated with substantial sample sizes. To determine N for two independent treatments, the following data are required:

1. δ the size of the difference to be detected
2. The desired probability of obtaining a significant result if the true difference is δ (Z_β)
3. (Z_β) obtained from 'z' tables
4. The significance level of the test (Z_α) usually P = 0.05)
5. The population standard deviation σ usually estimated from previous experiments

The formula for calculating sample size is:

$N = (Z_\alpha + Z_\beta)^2\ 2\sigma^2/\delta^2$------------1.

A worked example using this formula is shown in Table 1 and suggests that given the parameters listed, the investigator should have used N = 36 in each group to have had an 80% chance of detecting a difference of 10 units. Note that Z_α is based on a two-tail probability but Z_β is always based on a one-sided test (Norman & Streiner, 1993). This is because the tails of the two distributions representing the two media overlap on only one side.

What are the implications of sample size calculations?

This procedure is designed to protect the investigator against finding a non-significant result and reporting that the data are consistent with H_0 when in fact the experiment was too small. This suggests that a sample size calculation should always be carried out in the planning stage of an experiment. However, in reality, sample sizes are usually constrained by expense, time, or availability of human subjects for research and quite often a sample size calculation will result in an unrealistic N. Microbiologists would be surprised at the number of samples required to detect modest differences between two groups given the level of variability often encountered in experiments. Hence, sample size calculations may be an interesting adjunct to a study and may provide an approximate guide to N but should not be taken too seriously (Norman & Streiner, 1993). In addition, increasing sample size is only one method of increasing P'. Reducing the variability between replicate samples by using more homogenous groups or the use of experimental designs such as a paired or randomised block design and which eliminate certain sources of variability may also increase P'.

Table 1. Examples of sample size (N) and power (P') calculation for comparing two independent treatments

A) Sample size calculation (N)

Difference to be detected δ = 10 units

Standard deviation σ = 15 units

Significance of test P = 0.05, Z_α (from Z table at P = 0.05) = 1.96 (two-tail test)

Power of test say P = 0.80 and therefore P of not demonstrating an effect = 0.20

Z_β = (from Z table at P = 0.20) = 0.84 (one-tail test)

$(Z_\alpha + Z_\beta)^2\ 2\sigma^2/\delta^2$ = 3528/1000 = 35.28, say 36 per group

B) Power calculation (P')

Suppose in above example, the experiment had been carried out with 36 per group but the standard deviation had been 20 units not 15:

$Z_\beta = (\sqrt{N}.\ \delta/\ \sqrt{2}.\ \sigma) - Z_\alpha = 0.17$.

Hence, P of ***not*** demonstrating an effect = 0.43 (from Z table) and therefore, experiment has a P' = 0.57 (57%) of demonstrating a difference of 10 units

Calculation of P'

Sample size calculations also contain a useful corollary, calculation of the strength or power (P') of an experiment to detect a specific difference. This type of calculation is very useful in experiments that have failed to detect a difference the investigator *thought* was present. In such circumstances, it is useful to ask whether the experiment had sufficient P' to detect the anticipated difference. To calculate P' of an experiment equation 1 is rearranged to give Z_β:

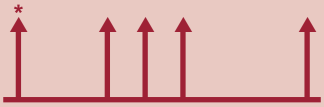
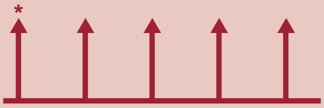
$Z_\beta = (\sqrt{N}.\delta/\sqrt{2}.\sigma) - Z_\alpha$ ------------- 2.

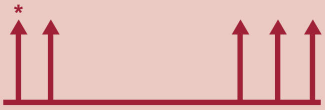
A worked example utilising this equation is given in Table 1. Suppose that the experiment described in the previous section had been conducted with a sample size of N = 36 but that the σ was actually 20 and not 10 units. The value of Z_β has fallen to 0.17 corresponding to a probability of not demonstrating an effect of P = 0.43. Hence, the probability of detecting a difference between the two means of 10 units has fallen to 57% and hence, P' would have been too low for this experiment to have had much chance of success.

Power and sample size in other designs

The equations used for calculating P' and N differ depending on the experimental design, *e.g.*, in a 'paired' design (Hilton & Armstrong, 2005) or when comparing two proportions (Katz 1997). Statistical software is available for calculating P' and N in most circumstances and although the equations may differ, the principles described in this Statnote remain the same. However, the situation becomes more complicated if there are more than two groups in a study and if the data are analysed by analysis of variance (ANOVA) (Armstrong & Hilton, 2004).

Figure 1. Adjustment to the effect size for calculation of sample size in a one-way analysis of variance (ANOVA) (K = number of groups, * = control group)

Distribution of means	Formulae
A. 	d x $\sqrt{1/2K}$
B. 	d x 0.5 $\sqrt{(K + 1)/3(K-1)}$

C. 	K = odd: d x $\sqrt{(K^2 - 1)/2K}$
	K = even: d x 1

Sample size and power in ANOVA

Calculation of P' is more complex when several group means are involved because the difference between the means may be distributed in various ways (Figure 1). An important statistic when several means are present is the effect size $d = \delta/\sigma$ where δ is the difference between the highest and lowest mean (Norman & Streiner, 1994). For example, if there are five groups (K = 5, a control and four treatments), one treatment may have a large effect while the remaining three may have similar but lesser effects (scenario A). In scenario B, the treatment means are spread more or less evenly and in scenario C, three treatments have large but similar effects and one has little effect and is therefore similar to the control. The essential approach is that 'd' is transformed into the effect size for ANOVA by multiplying by a formula which varies depending on the distribution of means. Various scenarios and sample formulae are illustrated in Figure 1 and how the calculations are made is shown in Table 2.

More complex experimental designs

In more complex experimental designs where there are many treatments or if a factorial arrangement of treatments is present, calculation of N by these methods becomes less useful. A more relevant concept is to consider the number of degrees of freedom (DF) associated with the error term of the ANOVA. In the general case, in a one-way design (Armstrong & Hilton, 2004) if there are 'p' treatments and N observations in each group, the error term will have p(N-1) DF and the greater the value of N, the greater the DF of the error term and the more precise and reliable the error estimate will be. A change of 1DF has a large effect on 't' or 'F' when DF < 10 but the effect is quite small when DF > 20. Hence, it is good practice to have at least 15 DF for the error term and this figure will be dependent on both the number of treatments and N. In factorial designs (Armstrong & Hilton 2004), with different factors or variables in the experiment, the presence of factorial combinations of treatments leads to *internal replication* and therefore such experiments can often be carried out using much smaller sample sizes. The principles underlying factorial experiments will be discussed in more detail in a future Statnote.

Table 2. Sample size calculation for a one-way analysis of variance (ANOVA)

Difference to be detected (largest mean – smallest mean) = δ
Assume individual means (K groups) equally distributed
Standard deviation = σ

1. Calculate effect size $d = \delta/\sigma$
2. Adjustment to formula (from Fig 1): effect size for ANOVA = (f) = $\sqrt{d}$ x 0.5 (K + 1/3(K – 1)
3. Look up 'f' in Table I (Norman & Streiner, 1993) to give sample size having chosen Z_α and Z_β

Conclusions

Statistical software is now commonly available to calculate P' and N for most experimental designs. In many circumstances, however, sample size is constrained by lack of time, cost, and in research involving human subjects, the problems of recruiting suitable individuals. In addition, the calculation of N is often based on erroneous assumptions about σ and therefore such estimates are often inaccurate. At best, we would suggest that such calculations provide only a very rough guide of how to proceed in an experiment. Nevertheless, calculation of P' is very useful especially in experiments that have failed to detect a difference which the experimenter thought was present. We would recommend that P' should always be calculated in these circumstances to determine whether the experiment was actually too small to test H_0 adequately.

references

- Armstrong R A & Hilton A (2004) The use of analysis of variance (ANOVA) in applied microbiology. *Microbiologist* Vol 5: No. **4**, pps 18 - 21.
- Hilton A & Armstrong R A (2005) Statnote 2: Describing the normal. *Microbiologist* Vol 6: No.**3**, pps 30 - 33.
- Hilton A & Armstrong R A (2005) Statnote 3: Testing the difference between two groups. *Microbiologist* Vol 6: No.**4**, pps 30 - 32.
- Katz D L (1997) Epidemiology, Biostatistics, and Preventive medicine review. W B Saunders, Philadelphia.
- Norman R N & Streiner D L (1993) Biostatistics: The Bare Essentials. Mosby, Toronto.
- Snedecor G W & Cochran W G (1980) Statistical Methods, 7th Ed. Iowa State University Press, Ames Iowa.

Dr Anthony* Hilton and Dr Richard Armstrong**
*Pharmaceutical Sciences and **Vision Sciences, Aston University, Birmingham, UK