Practical Guidelines on the Required Sample Size in a Two Group Multivariate Mean Comparison

Timo Bechger

Abstract

This paper concerns the sample size required for two group multivariate mean comparison. Its main purpose is to provide guidelines on the sample size that is required (a) for sufficient power of the T test, and (b) for the sample means to be reasonably close to the population values. To facilitate the choice of the expected difference the well-known indices for effect size by Cohen (1977) and Stevens (1980) are expressed in terms of the common language effect size measure proposed by McGraw and Wong (1992). Cohen's effect size measures was used to simplify the relation between the precision of the estimate and sample size. Examples are given.

Faculteit der Psychologie, Vakgroep Psychonomie,
De Vrije Universiteit, De Boelelaan 1111/1115, 1081 HV, Amsterdam.
E-mail: TM.Bechger@psy.vu.nl

Introduction

In the planning stages of experimental or quasi-experimental studies one should choose an appropriate sample size. One criterion is that the sample should be large enough for statistical tests to have sufficient statistical power. A second criterion is that each sample be large enough to obtain estimates that are reasonably close to the population values. Although both statistical power and the precision of the estimates increase with sample size, they need not require the same sample size.

This paper concerns two group mean comparison. Its main purpose is to provide practical guidelines on sample size requirements for sufficient power of the t test, as well as for sample means to be reasonably close to the population values. To illustrate these guidelines we apply them to the reading literacy study that was recently conducted by the international association for the evaluation of education (e.g. Elley, 1994; Bechger, 1997).

Univariate Comparison of Means

If two samples of independent observations have been drawn from normally distributed populations with unknown common dispersion, the t test is used to test the equality of the means in two populations. Cohen (1977) defines a standardized measure of population effect size d as the difference in mean values divided by the common population standard deviation, i.e., $d = (\mu_1 - \mu_2) / \sigma$. d indicates by how many standard deviation units the population means are separated.

Table 1: *Group size required for power 0.80 and 0.90, α =0.05 (two sided) with the univariate t test. Adapted from Table 2.4.1 in Cohen (1977).*

Effect size:	d = 0.2		d = 0.5		d = 0.8	
Power:	0.80	0.90	0.80	0.90	0.80	0.90
Common group size:	393	526	64	85	26	34

With an a prior estimate of the effect size, the sample size required to detect the difference with specific power can be looked up in tables provided by Cohen (1977). As a rough rule-of-thumb, Cohen (1977) suggests that an effect size around 0.2 is small, an effect size around 0.5 is medium and an effect size > 0.8 is large. Table 1 shows the group sizes required to detect these effects with adequate power (e.g., at least 80%).

McGraw and Wong (1992) suggest that the predictive value of the difference is easier to interpret than Cohen's effect size. The predictive value of the significant difference may be investigated by randomly pairing members from both populations and counting how many times a subject from the first group is really different from the subject in the other group. A less cumbersome procedure starts by assuming that the difference has a normal distribution with mean $(\mu_1 - \mu_2) = \Delta$ and variance $2\sigma^2$. Let $P(X_1 - X_2 > 0)$ denote the probability that the first randomly chosen subject has a higher score than the subject from the other population. This probability equals:

$$P\left(Z > \frac{0 - \Delta}{\sqrt{2}\,\sigma}\right) = P\left(Z < \frac{\Delta}{\sqrt{2}\,\sigma}\right) \qquad (1)$$

If the cumulative probability associated with the z-value is 0.50, $\Delta/\sqrt{2} = 0$, and the results have no predictive value. McGraw and Wong call this probability the *common language effect size measure.*[1]

If one expresses the common language effect size measure in terms of Cohen's effect size one would know immediately the relation between the common language effect size measure, sample size and power. Suppose, for instance, that 70% is judged to be the minimal predictive effect size.

$$P\left(Z < \frac{\Delta}{\sqrt{2}\,\sigma}\right) = 0.70. \qquad (2)$$

Equation 2 implies that $\Delta/\sqrt{2}$ is about 0.50 and

---

[1] Mishra et al. (1986) call (1) an overlapping coefficient. They were probably the first to propose this effect size measure.

$$\frac{\Delta}{\sigma} = 0.50\sqrt{2}, \qquad (3)$$

where Δ/σ is Cohen's effect size measure. $0.50\sqrt{2} \approx 0.71$, which is Cohen's effect size corresponding to a predictive effect of 0.70. A further look at the tables for the standard normal distribution shows that a mean difference larger than 2 standard deviations corresponds to an almost perfect separation of the groups.

Multivariate Comparison of Means

Hotelling (1931) offered the T^2 test as a multivariate generalization of the univariate t test for independent samples, i.e., a test of the hypothesis that two groups have equal means on p variables. A multivariate measure of overall effect size can be calculated as:

$$D^2 = (\mu_1 - \mu_2)' \mathbf{S}^{-1} (\mu_1 - \mu_2), \qquad (4)$$

where μ_1 and μ_2 denote the sample mean vectors in the first and second group, respectively, $(..)^t$ denotes transposition, and $\mathbf{S}^{-1}$ the inverse of the common within-group covariance matrix.[2] D^2 is a natural squared generalization of the univariate effect size, where the means have been replaced by mean vectors and the standard deviation by its squared multivariate generalization of within-group variability.

The power of the T^2 test can be determined using the tables provided by Cohen (1977, Table 8.3.1- 8.3.33). For this purpose, Cohen's effect size f may be calculated as (cf. Stevens, 1980):

$$f = \frac{|D|}{\sqrt{2}\sqrt{p+1}} \qquad (5)$$

---

[2] Note that $T^2 = D^2$ times $(n_1 n_2)/(n_1+n_2)$, where n_1 and n_2 denote the sample size in group 1 and group 2, respectively (Stevens, 1986, p. 140).

If no estimate of D^2 can be obtained one may use the conventions suggested by Stevens (1980). These are: $D^2 = 0.25$ for a small effect, $D^2 = 0.64$ or 1.00 for a medium effect and $D^2 = 2.25$ for a large effect. In Table 2, I summarized some group sizes required for adequate power of the T^2 test.

Table 2: *Sample sizes required for adequate power (0.80, 0.90) of Hotelling 's T^2 test given $\alpha = 0.05$. Cohen 's (1977) Tables 8.4.4 to 8.4.5 were used entering the Table with $u = p$ and f as calculated with Equation 5.*

	D^2= 0.25		D^2= 0.64		D^2= 1.00		D^2= 2.25	
nr. of variables								
2	78	101	31	40	20	26	9	12
4	96	124	38	49	24	31	11	14
6	110	140	43	55	28	35	13	16
8	121	153	48	60	31	39	14	17
power	0.80	0.90	0.80	0.90	0.80	0.90	0.80	0.90

It is concluded that even for small effects, i.e., $D^2 = 0.25$, power is good when there are more than 153 subjects in each group.

Multivariate significance implies that there is a linear combination of the variables (the discriminant function), which is significantly separating the groups (Flury and Riedwyl, 1988). Discriminant analysis may be performed to determine the weights for the linear combinations of the variables that maximally discriminate between the two groups. The significance test for a two-group discriminant analysis is Hotelling's T^2. Hence, the power analysis is the same as well as the sample size requirements.

How large should D^2 be in order for a difference in mean vectors to have any predictive value ? Morrison (1988, p. 235) proves that if the discriminant function is used to classify subjects, assuming equal prior probabilities of group membership, the probability of misclassification equals :

$$P(Z < - D/2). \qquad (6)$$

If we accept a minimum predictive value of 70%, the probability of misclassification must be less then 0.30, which means that D^2 should be equal to or larger than 1.00 to have sufficient predictive value, i.e., a medium effect.

The Sample Size that is Required to Estimate Means

This section concerns the size of the sample that is required to obtain in each single group a sample mean which is close enough to the population value. The following formula may be used as a first approximation to the required sample size (Barnett, 1974, Section 2.5):

$$n \approx S^2 / (\frac{k}{z_\alpha})^2 . \qquad (7)$$

S^2 is the population variance, k denotes the tolerable difference between the population mean and the sample mean and z_α the value of the standard normal distribution with probability α, where α denotes the risk of obtaining an absolute difference between the sample mean and the population mean greater than k.

The population variance is usually unknown. To remove S^2 from the formula, k may be expressed in terms of standard deviation units, i.e., $k = c\,S$, where c is a constant that must be chosen to represent a reasonable difference. Formula 7 now reduces to:

$$n \approx \frac{z_\alpha^2}{c^2} . \qquad (8)$$

If we apply Cohen's rule of thumb in this context, c takes the values 0.2, 0.5 and 0.8. Given $\alpha = 0.05$, the relationship between the tolerable difference and the required sample size n turns out to be very simple.

$$\text{If } k = \begin{cases} 0.2\,S \Rightarrow n \cong 96 \\ 0.5\,S \Rightarrow n \cong 15 \\ 0.8\,S \Rightarrow n \cong 6 \end{cases} \qquad (9)$$

This means that at least 96 observations are required for a 'small' tolerable difference, given 5% probability of not obtaining such tolerance. 96 is a rather small sample in this context, which suggests that other values of c may be more appropriate in this situation. In the international reading literacy study, for instance, 0.1S was considered an acceptable precision, which corresponds to an effective sample size of about 400 subjects in each of the participating nations (Elley, 1994; Rust, 1995).

Conclusion

In practice, methodologists or statisticians are often asked for a single indication of the minimal sample size. In addition, they are required to justify their choice in terms that are acceptable from a substantive point of view. To this aim, we used effect size measures to provide guidelines on sample size requirements for sufficient power and for precise estimation. We demonstrated that the CL effect size measure can be related to sample size and statistical power, and that it can be generalized to the multivariate case. We also demonstrated that Cohen's rules of thumb may be used to obtain a simple relation between precision of estimation and sample size.

The formulae presented above are based on simple random sampling from a very large population. When a complex sampling design is employed the estimate of the sample size may be updated by multiplying the estimated sample size by the so-called *design effect* (e.g. Cochran, 1977, p. 21). In the reading literacy study, the design effects on the mean literacy scores ranged from 5.9 to 10.1 (Rust, 1995). The true sample size, corresponding to an effective sample size of 400, must therefore lie between (5.9 × 400=) 2360 and (10.1 × 400=) 4040. Samples of this size were indeed gathered by the majority of the countries that participated in the study (e.g. Postlethwaite and Ross, 1992, Table 2.2). Even samples between 590 and 1010 would have been acceptable according to Cohen's rule of thumb.

References

Barnett, V. (1974). Elements of sampling theory. The English University Press.

Bechger, T. M. (1997). Methodological aspects of educational comparison: The case of reading literacy. TT-Publications: Amsterdam.

Cochran, W. G. (1977). Sampling Techniques. (3rd Ed.). John Wiley and Sons: New-York

Cohen, J. (1977). Statistical Power Analysis for the Behavioral Sciences. Revised Edition. Academic Press, inc.

Elley, W.B. (1994). The IEA Study of Reading Literacy: Achievement and Instruction in Thirty-Two School Systems. Pergamon Press.

Flury, B, & Riedwyl, H. (1988). Multivariate statistics: A practical approach. London: Chapman and Hall

Hotteling, H. (1951). A generalized T test and measure of multivariate dispersion. Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability. pp. 111-184, McGraw-Hill Book Company, New-York.

McGraw, K. O. , & Wong, S. P. (1992). A common language effect size measure. Psychological Bulletin, 111, 361-365.

Mishra, S.N., Shah, A.K., & Lefante, J.J. (1986). Overlapping coefficients: The generalized t-approach. Communications in Statistics, 15: 123-128.

Morrison, D. F. (1988, 8th printing). Multivariate statistical methods. MacGraw-Hill International Editions.

Postlethwaite, T. N. , & Ross, K. N. (1992). Effective schools in reading. Implications for educational planners. The IEA.

Rust, K. (1995). Issues in Sampling for international comparative studies in Education: The case of the IEA reading literacy study. In, Methodological Issues in comparative educational studies: The case of the IEA reading literacy study. U.S. department of Education.

Stevens, J. P. (1980). Power of the Multivariate Analysis of Variance Test. Psychological Bulletin, 3, 728-737.

Acknowledgement

The author thanks dr. Jacob Cohen, dr. Han van der Maas, Prof. dr. Joop Hox, and an anonymous reviewer for their critical comments and encouragement.

Ontvangen: 13-08-1996
Geaccepteerd: 16-06-1997