STATISTICS IN MEDICINE
Statist. Med. **19**, 369–372 (2000)

ON THE RELATIVE SAMPLE SIZE REQUIRED FOR MULTIPLE COMPARISONS

JOHN S. WITTE[1]*, ROBERT C. ELSTON[1] AND LON R. CARDON[2]

[1]*Department of Epidemiology and Biostatistics, Case Western Reserve University, 2500 MetroHealth Drive, Cleveland OH 44109-1998, U.S.A.*
[2]*Wellcome Trust Centre for Human Genetics, University of Oxford, Windmill Road, Oxford OX3 7BN, UK.*

SUMMARY

Multiple comparisons are commonly made in epidemiologic and genetic research. How to appropriately adjust for multiple comparisons remains a controversial issue. This note demonstrates, however, that large increases in the number of comparisons has a limited effect on the sample size required to maintain an experimentwise α-level. In particular, the relative sample size required increases only linearly with the logarithm of the number of comparisons made. Copyright © 2000 John Wiley & Sons, Ltd.

INTRODUCTION

Much literature exists regarding issues surrounding multiple comparisons (or tests). At one extreme, some epidemiologists argue that multiple comparisons are a non-issue, requiring no consideration.[1] At the other extreme, some genetic statisticians have proposed 'adjusting' for multiple comparisons regardless of whether or not they have actually been performed.[2,3] Furthermore, statisticians have extensively evaluated multiple comparisons issues.[4–7] It is therefore of interest to consider how increasing the number of comparisons affects the relative sample size. Here we show that, if a Bonferroni adjustment is used to allow for multiple comparisons, this increase is surprisingly small, indicating that large increases in the number of comparisons do not require equivalently larger sample sizes.

* Correspondence to: John S. Witte, Department of Epidemiology and Biostatistics Case Western Reserve University, 2500 MetroHealth Drive, Cleveland OH 44109-1998, U.S.A. E-mail: witte@darwin.cwru.edu

Contract/grant sponsor: National Cancer Institute
Contract/grant number: CA73270

Contract/grant sponsor: National Institute of General Medical Sciences
Contract/grant number: GM28356

Contract/grant sponsor: National Center for Research Resources
Contract/grant number: RR03655

CCC 0277–6715/2000/030369–04$17.50

Received March 1999
Accepted April 1999

SAMPLE SIZE VERSUS MULTIPLE COMPARISONS

Assume that we are interested in comparing (or testing) the equality of two proportions from independent samples of equal size. Further make the asymptotic assumption that the sample proportions are normally distributed. Then the number of subjects required for this comparison can be written (ignoring the necessity for n to be an integer and using a simplified formula) as

$$n = \frac{(Z_\alpha + Z_\beta)^2}{\delta^2} \tag{1}$$

where Z_x is the x-fractile of the standard normal distribution (that is, $Z_x = \Phi^{-1}(x)$ where $\Phi(t) = \int_{-\infty}^{t} \frac{1}{\sqrt{2}} e^{u/2}\, du$), α and β are the specified probabilities of type I and type II error, and δ^2 is a function of the two proportions. That is, δ is proportional to the standardized size difference between the proportions that one desires to detect

$$\delta = \frac{p_1 - p_2}{\sqrt{\{2\bar{p}(1-\bar{p})\}}}$$

where p_1 and p_2 are the two proportions and $\bar{p} = (p_1 + p_2)/2$. Similarly, if we are interested in testing the equality of two means from independent samples of normally distributed random variables, the required sample size can also be expressed in the form of equation (1) if the variances are known, or approximately so if the variances are unknown. For example, if the two means are μ_1 and μ_2, and the two variances are σ_1^2 and σ_2^2, we have

$$\delta = \frac{\mu_1 - \mu_2}{\sqrt{(\sigma_1^2 + \sigma_2^2)}}.$$

Now suppose we wish to make multiple independent comparisons with δ^2 being the same for each comparison. This might happen in epidemiology, for example, if one looks at numerous environmental exposures with homogeneous ubiquitousness. Similarly, in genetics this would occur if we type each subject at multiple independent loci and perform the same test, for example, of a specified dominant or recessive effect, at each locus assuming that the genotype frequencies remain constant across the loci.

If we wish to perform m tests and retain an experimentwise significance level α, then (using the conservative Bonferroni adjustment), the required sample size is

$$n = \frac{(Z_{\alpha/m} + Z_\beta)^2}{\delta^2}. \tag{2}$$

Figure 1 shows a plot of n against m (on a log scale) from equation (2), for $\beta = 0{\cdot}2$, $0{\cdot}1$ and $0{\cdot}05$ with $\delta^2 = (1{\cdot}645 + 0{\cdot}842)^2$, so that the sample size required is relative to the conventional (one test) values $\alpha = 0{\cdot}05$ and $\beta = 0{\cdot}2$ (using one-sided tests). From Figure 1 we see that for each value of β considered, the relation between the relative sample size and $\log_{10}$ (number of comparisons) is approximately linear with a modest slope. When $\beta = 0{\cdot}20$, the relative sample sizes required for 1, 10, 100 or 100,000 comparisons are 1, 1·9, 2·8 and 5·6, respectively. That is, increasing the number of comparisons from 1 to 100,000 results in a less than six-fold increase in relative sample size (Figure 1). When $\beta = 0{\cdot}05$, relative to $\beta = 0{\cdot}2$ the sample sizes required for 1, 10, 100 or 100,000 comparisons are 1·8, 2·9, 3·9, and 7·3, respectively. Thus, increasing the number of

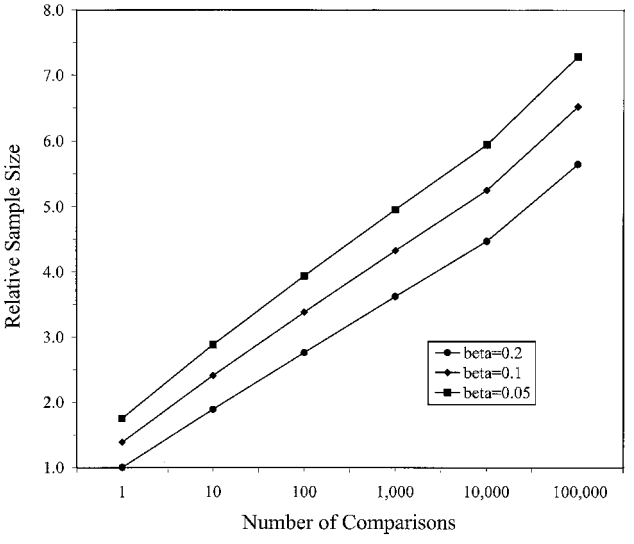

Figure 1. Relation of the relative sample size required to the number of comparisons for a Bonferroni-adjusted test to achieve a specified $(1 - \beta)$ power, relative to the sample size required for the nominal α-level = 0·05 test to achieve 80 per cent power ($\beta = 0{\cdot}20$)

comparisons from 1 to 100,000 results in a $7{\cdot}3/1{\cdot}8 =$ four-fold increase in relative sample size when $\beta = 0{\cdot}05$ (Figure 1).

DISCUSSION

This note shows that the increase in sample size required for multiple comparisons (or tests) is not as great as might perhaps have been expected. In particular, the relative sample size – allowing for a Bonferroni adjustment – is approximately linearly related to the logarithm of the number of comparisons made. In light of the rapid technological advances that enable multiple comparisons to be made, this striking result is reassuring. For example, in genetic research, the potential for undertaking comparisons across tens, or even hundreds, of thousands of single nucleotide polymorphisms (SNPs) with similar allele frequencies is rapidly approaching. In this connection, others have recently noted the limited increase in sample size,[8,9] and it was pointed out for a specific situation that increasing the number of comparisons from 1 to 500,000 requires an approximately 7·3-fold increase in sample size.[9] Of course, a crucial assumption of this result is that δ^2 is constant across comparisons; variability in δ^2, which is a function of the size of the effect to be detected and the sample gene frequencies among subjects with the different responses, may in some cases be a more important determinant of the required sample size than the actual number of comparisons to be made.[10] Furthermore, the precise relationship plotted in Figure 1 depends on the assumption of approximate normality, an assumption that can be expected to become less accurate at the extreme tails of the test statistic's distribution. Nevertheless, Figure 1 does indicate, as a first approximation, how sample size is related to the number of multiple comparisons one wishes to make.

ACKNOWLEDGEMENTS

This work was supported by U.S. Public Health Service grants CA73270 from the National Cancer Institute, GM28356 from the National Institute of General Medical Sciences and RR03655 from the National Center for Research Resources.

REFERENCES

1. Rothman, K. 'No adjustments are needed for multiple comparisons', *Epidemiology*, **1**, 43–46 (1990).
2. Lander, E. and Kruglyak, L. 'Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results', *Nature Genetics*, **11**, 241–247 (1995).
3. Witte, J. S., Elston, R. C. and Schork, N. S. 'Re: Genetic dissection of complex traits (correspondence)', *Nature Genetics*, **12**, 355–356 (1996).
4. Hochberg, Y. and Tamhane, A. *Multiple Comparison Procedures*, Wiley, New York, 1987.
5. Saville, D. J. 'Multiple comparison procedures: the practical solution', *American Statistician*, **44**, 174–180 (1990).
6. Hochberg, Y. and Benjamini, Y. 'More powerful procedures for multiple significance testing', *Statistics in Medicine*, **9**, 811–818 (1990).
7. Benjamini, Y. and Hochberg, Y. 'Controlling the false discovery rate: a practical and powerful approach to multiple testing', *Journal of the Royal Statistical Society*, **57**, 289–300 (1995).
8. Risch, N. and Merikangas, K. 'The future of genetic studies of complex human diseases', *Science*, **273**, 1516–1517 (1996).
9. Lander, E. S. 'The new genomics: global views of biology', *Science*, **274**, 536–539 (1996).
10. Elston, R. C., Idury, R. M., Cardon, L. R. and Lichter, J. B. 'The study of candidate genes in drug trials: sample size considerations', *Statistics in Medicine*, **18**, 741–751 (1999).

