Sample size and power for comparing two or more treatment groups in clinical trials

Simon J Day, David F Graham

Abstract

Methods for determining sample size and power when comparing two groups in clinical trials are widely available. Studies comparing three or more treatments are not uncommon but are more difficult to analyse. A linear nomogram was devised to help calculate the sample size required when comparing up to five parallel groups. It may also be used retrospectively to determine the power of a study of given sample size. In two worked examples the nomogram was efficient.

Although the nomogram offers only 5% and 1% significance levels and can be used only for up to five treatment groups, this is sufficient for most researchers.

Introduction

The problem of determining sample size and power when comparing two groups has been described widely.[1-3] A simple and practical method for determining sample size in terms of power, significance, and standardised difference without tables or formulas was described by Altman.[4] This had the slight disadvantage that it underestimated sample size when the groups were small, but for the practical range of sample sizes given in the nomogram this bias was negligible. Conversely, if the nomogram was used to determine the power of a study in which the sizes of the groups were known the power specified was slightly higher than the true value. All of these nomograms, formulas, and tables apply equally to comparisons of two groups by an independent samples t test but vary in complexity.

Methods to determine sample size when the outcome is binary (for example, survival rates at five years or incidence of nausea after anaesthetic) are similar in concept but differ in detail. Formulas,[5] tables,[6] and nomograms[7,8] are available for determining sample size and power when comparing two groups. Lachin described the problem of comparing more than two groups.[9]

Clinical trials comparing three or more treatments (one of which may be a placebo) are not uncommon—for example, Lucki *et al*[10] and Banner *et al*[11] compared three groups and Rowbotham and Nimmo[12] and Tatsuta *et al*[13] compared four groups. These trials are, however, more difficult both to design[3] and to analyse.[14] The initial method of analysis should usually be analysis of variance rather than multiple t tests to reduce the chance of a type I error—that is, the chance finding of at least two of the treatments seeming significantly different when really they are not. Specific contrasts using t tests may then be useful.

We describe the use of a linear nomogram to estimate the required sample size when comparing three or more treatment groups by analysis of variance. As it is more complicated than Altman's nomogram for comparing two groups we do not routinely advocate its use for this purpose, although when used in this way it does not overestimate power when sample sizes are small. It has wider application when three or more groups are to be compared.

Methods and results

Fleiss described a numerical method for determining iteratively the power of the analysis of variance test.[15] The nomogram that we describe in the present paper uses the principles that he outlined, although with important modifications. The exact method and calculations and details of the design of these modifications will be described elsewhere.

To use the nomogram prospectively to calculate sample size an estimate of the possible mean response for each group and the expected standard deviation within each group are needed. The only calculation necessary is to evaluate a difference parameter ($\sqrt{\lambda}$), which is the standard deviation of the possible group means divided by the standard deviation of the measurements. The nomogram (figure) has three axes: the bottom horizontal axis gives the difference parameter, the left vertical axis the power, and the top horizontal axis the sample size of each group. To make it easier to use each axis is duplicated at the opposite side of the nomogram (that is, $\sqrt{\lambda}$ is at the top, power is on the right, and sample size is at the bottom). We describe how to use the nomogram to determine sample size and power in two clinical trials.

EXAMPLE 1

Hypertensive patients were to receive one of three randomised treatments. As all of the patients were to receive active drugs the researcher expected an overall reduction in diastolic blood pressure to about 90 mm Hg, which he considered to be beneficial clinically. He judged that mean diastolic pressure would fall to 100 mm Hg, 95 mm Hg, and 85 mm Hg in three groups. From previous studies he knew that the standard deviation within each group would be about 15 mm Hg, and he wanted 90% power to detect a difference between the treatments at the 1% level of significance. The difference parameter ($\sqrt{\lambda}$)=SD(100, 95, 85 mm Hg)/15 mm Hg=0·509. To determine the sample size for each group the point corresponding to $\sqrt{\lambda}$=0·509 and power=90% was plotted on the nomogram (point A, on figure). For 1% significance (α=1%) and three groups (g=3) a line was drawn from the point (○) in the lower left hand corner (point B) through point A until it reached the horizontal line labelled α=1%, g=3 (point C). A vertical line was then drawn from point C upwards to give the size of each group (point D).

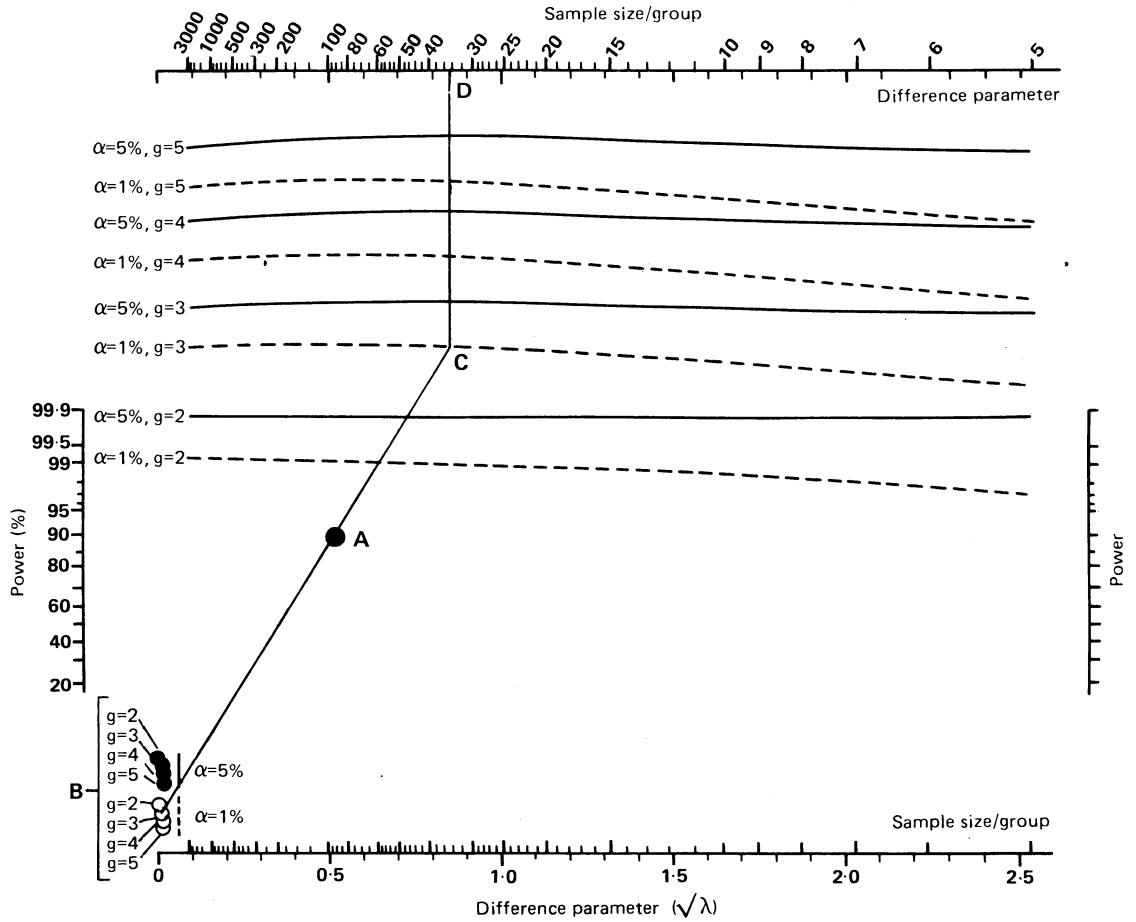
The nomogram shows that three groups of 35 subjects were required. If instead of the 1% significance level the researcher wished to work to the 5% signifi-

Department of Epidemiology and Population Sciences, London School of Hygiene and Tropical Medicine, London WC1E 7HT
Simon J Day, BSC, *statistician*

Sterling Research Group Europe, Guildford, Surrey GU1 4YS
David F Graham, MB, *clinical research director*

Correspondence and requests for reprints to: Mr Day.

Br Med J 1989;299:663-5



Nomogram for comparing up to five independent samples (g=number of groups) of continuous variable relating power, group sample size, difference parameter ($\sqrt{\lambda}$), and significance (α). Points (A-D) relate to determining sample size required to show difference between means of 85, 95, and 100 mm Hg with standard deviation 15 mm Hg ($\sqrt{\lambda}$=0·509) at 1% significance with 90% power (see text for details)

cance level point A would be the same, point C would be on the line labelled α=5%, g=3, and the size of each group would be 26. To use the nomograms retrospectively to determine power a line is drawn from point D (sample size) to whichever line is appropriate for the number of groups and significance level (point C). Point C is then joined to the appropriate point B; the power can be read off for any given value of $\sqrt{\lambda}$.

EXAMPLE 2

The two by two factorial design is a common design for multiple groups. In it the four possible combinations of two binary treatment regimens are compared. For example, Dodd *et al* compared low (5 μg/kg) and high (10 μg/kg) doses of glycopyrronium given either simultaneously with or one minute before edrophonium.[16] There were two factors: time of administration and dose. Taking the mean heart rate 10 minutes after intervention as the response variable, they tested whether the dose alone or the time of administration alone had an effect. Finally, they also tested whether the dose had a different effect if the drug was given simultaneously with edrophonium or one minute earlier (that is, whether there was a dose by time interaction). The table gives the mean responses in each of the four groups 10 minutes after intervention.

Mean (SD) heart rates (beats/min) 10 minutes after treatment in four groups of patients in trial of glycopyrronium and edrophonium[16]

Time of giving glycopyrronium	Dose of glycopyrronium		
	5 μg/kg (n=15)	10 μg/kg (n=15)	5 or 10 μg/kg (n=30)
1 Minute before edrophonium (n=15)	71·3 (12·1)	93·9 (11·4)	82·6
With edrophonium (n=15)	77·1 (14·4)	93·3 (12·4)	85·2
1 Minute before or with edrophonium (n=30)	74·2	93·6	

Overall, a dose of 5 μg/kg reduced the heart rate by 19·4 beats/min more than a dose of 10 μg/kg and giving glycopyrronium one minute before edrophonium reduced the heart rate by 2·6 beats/min more than giving the two drugs simultaneously; the standard deviation in each group was about 12 beats/min. To determine sample size (prospectively) or power (retrospectively) for comparing the two doses they applied the method for comparing two groups. In this part of the analysis the number of subjects in each group was the number of patients receiving each dose of glycopyrronium. Each of these groups, however, itself comprised two groups (patients given glycopyrronium simultaneously with or one minute before edrophonium), so the sample size in each of the four subgroups was half that specified by the nomogram. Similar comments applied to testing the effect of time of administration.

Determining the sample size for detecting the effect of interaction entailed estimating the size of the interaction. This was the difference between the effect of the time of administration on those receiving the low dose (77·1−71·3=5·8 beats/min) and those receiving the high dose (93·3−93·9=−0·6 beats/min). So the size of the interaction was 6·4 beats/min. The same result was achieved by taking the difference between the doses at both of the times of administration. Determining sample size for detecting this interaction also entailed applying the two sample problem, the sample size in each of the four groups being half that specified for comparing two groups. The power to detect the effect of dose of 19·4 beats/min at the 5% significance level was in excess of 99·9%; the power to detect the effect of time of administration of 2·6 beats/min was about 15%; the power to detect the interaction was about 33%.

An alternative to analysing the raw data is to analyse changes from baseline values by subtracting each

person's baseline heart rate from their heart rate 10 minutes after intervention. As patients were assigned to groups at random this should not affect the size of the observed effects of treatment. This analysis reduced the standard deviation to about 3 beats/min and so increased the power of the study to detect the effect of time of administration to about 92%.

Discussion

Considerations of sample size in studies of many groups are just as important as those in studies of two groups. The nomogram described in this paper allows sample size to be estimated accurately when the initial analysis is analysis of variance. If *t* tests are used subsequently for comparing specific pairs of groups sample size should be estimated to ensure that sufficient power is obtained for each pairwise comparison. When comparing pairs of treatments the sample size relative to the required precision of the effects of treatment should also be considered.[17 18]

When comparing two groups it is sensible to consider the smallest difference of clinical interest, but such a difference cannot be defined naturally among several groups. In example 1, however, it might be more relevant to consider how the mean responses differ rather than what those mean responses might be. So instead of specifying the mean responses as 100, 95, and 85 mm Hg we might consider that the second treatment reduces blood pressure by 5 mm Hg more than the first and the third by 15 mm Hg more than the first. Specifying the problem in this way would lead to the same difference parameter because the standard deviation of 0, 5, and 15 mm Hg is the same as that of 100, 95, and 85 mm Hg. In some clinical applications it may be easier and more realistic to think about effects of treatment in this relative way (that is, differences or changes from baseline values).

We believe that although the nomogram offers only 5% and 1% significance levels and can be used only for up to five treatment groups, this is sufficient for most researchers. Extending the method for other levels of significance or for more treatment groups follows easily (details may be obtained from SJD).

1 Armitage P, Berry G. *Statistical methods in medical research*. 2nd ed. Oxford: Blackwell, 1987:182.
2 Bland JM. *An introduction to medical statistics*. Oxford: Oxford Medical Publications, 1987:160.
3 Pocock SJ. *Clinical trials*. Chichester: Wiley, 1983:128, 229.
4 Altman DG. How large a sample? In: Gore SM, Altman DG, eds. *Statistics in practice*. London: BMJ, 1982:6-8.
5 Fleiss JL. *Statistical methods for rates and proportions*. Chichester: Wiley, 1981:33-49.
6 Machin D, Campbell MJ. *Statistical tables for the design and analysis of clinical trials*. Oxford: Blackwell, 1987:18-33.
7 Clark CJ, Downie CC. A method for the rapid determination of the number of patients to include in a controlled clinical trial. *Lancet* 1966;ii:1357-8.
8 Miller DK, Homan SM. Graphical aid for determining power of clinical trials involving two groups. *Br Med J* 1988;**297**:672-6.
9 Lachin JM. Sample size determination for r×c comparative trials. *Biometrics* 1977;**33**:315-24.
10 Lucki I, Rickels K, Giesecke MA, Geller A. Differential effects of the anxiolytic drugs, diazepam and buspirone, on memory function. *Br J Clin Pharmacol* 1987;**23**:207-11.
11 Banner NR, Lloyd HM, Hamilton RD, Innes JA, Guz A, Yacoub MH. Cardiopulmonary response to dynamic exercise after heart and combined heart-lung transplantation. *Br Heart J* 1989;**61**:215-23.
12 Rowbotham DJ, Nimmo WS. Effect of cisapride on morphine-induced delay in gastric emptying. *Br J Anaesth* 1987;**59**:536-9.
13 Tatsuta M, Iishi H, Yamamura H, Yamamoto R, Taniguchi H. Enhancement by tetragastrin of experimental induction of gastric epithelium in the duodenum. *Gut* 1989;**30**:311-5.
14 Clayton DG. Comparing three groups. *Applied Statistics* 1983;**32**:64-8.
15 Fleiss JL. *Design and analysis of clinical experiments*. Chichester: Wiley, 1986:371-4.
16 Dodd P, Day SJ, Goldhill DR, MacLeod DM, Withington PS, Yate PM. Glycopyrronium requirements for antagonism of the muscarinic effects of edrophonium. *Br J Anaesth* 1989;**62**:77-81.
17 McHugh RB, Le CT. Confidence estimation and the size of a clinical trial. *Controlled Clin Trials* 1984;5:157-63.
18 Day SJ. Sample sizes and confidence intervals of pre-specified size. *Lancet* 1988;ii:1427.

(Accepted 30 May 1989)

MATERIA PARAMEDICA

I saw it first, so it's mine: the struggle for priority in reporting discoveries in science

Two or more researchers may be investigating in the same topic and aiming at the same goal. The one who is first to announce his discovery is the one who gets all the credit, renown, advancement, and, in the United Kingdom, perhaps even funding. This is seemingly unfair to the other researcher, who may have been only slightly in arrears of submitting for publication. But competitive life is like that. This pertains also in the pharmaceutical industry. Two drugs with virtually identical actions are synthesised by rival companies. One of these is marketed a few months ahead of the other. It captures 90% of the market, and permanently. It is not surprising, therefore, that many researchers jealously guard their results until publication is imminent. Theft and plagiarism in research is rare but not unknown.[1] I recall getting lost in the corridors of a research institute on my way to a meeting and accidentally finding myself in an unoccupied laboratory. A moment later the incumbent of the laboratory returned. The charm and tact for which she was universally known were not now in evidence, and I was almost thrown out.

Is the striving for priority a new phenomenon? No, it is not. Claims for priority in scientific research were ardently pressed in the days of the Enlightenment.

In the seventeenth century one method adopted of staking one's claim was to publish an anagram of a statement epitomising the discovery. If no one else published an identical discovery in the ensuing months, then the anagram was unscrambled and republished, and priority was duly established at the date of first publication. Thus, when Galileo, in the summer of 1610, aimed his primitive telescope at the planet Saturn, he misinterpreted the rings projecting on either side and thought that they were two additional planets, aligned three in a row. He sent the following to various friends: SMAISMRMIL ME POETA LEAMIBUNEN UGTTAVIRAS. In November of that year he unscrambled the anagram thus: "Altissimam planetam tergeminum observavi" (submitting one v for a u), which may be translated: "I have observed that the farthest planet is triple." Later Galileo made the revolutionary discovery that the planet Venus had phases like the moon. This established that it was illuminated by the sun and confirmed the heliocentricity of the solar system. On this occasion his anagram comprised, except for two letters, proper but meaningless words, thus: HAEC IMMATURA A ME IAM FRUSTRA LEGUNTUR, OY. This was duly unscrambled as follows: "Cynthiae figuras aemulatur Mater amorum." The direct translation reads: "The Mother of Lovers rivals the shapes of Cynthia." The "mother of lovers" is Venus. "Cynthia" is Diana, the Moon Goddess. The free translation now reads: "Venus rivals the shapes of the moon." One can but imagine that the classical education of seventeenth century cosmologists was equal to the task of unravelling Galileo's statement.

The pronouncement of Hooke's law of the spring was almost the least of his many accomplishments, yet it is that by which he is known to schoolchildren. In 1676 Robert Hooke published his "law" in the following anagram: CEIIINOSSSTTUV. Having received no contenders, he unscrambled it two years later thus: "Ut tensio sic vis," which may be translated, "The power (of the spring) is as the tension (thereof)."

Readers who enjoy word games can take pleasure in constructing anagrams that comprise whole words, and which describe important discoveries made in the past century. Florey and Chain, for example, might have succinctly announced their discovery thus: "Nice dame will pine." In view of the long wait for publication experienced by some authors perhaps editors of medical and scientific journals, who are pressed for space, would consider mitigating the priority race by accepting one line anagrams. It could be fun.—BERNARD J FREEDMAN

1 Broad W, Wade N. *Betrayers of the truth*. Oxford: Oxford University Press, 1982:163-78.