Eye (2009) 23, 1589–1597

www.nature.com/eye

LABORATORY STUDY

The estimation of approximate sample size requirements necessary for clinical and epidemiological studies in vision sciences

EA Goodall[1], J Moore[2] and T Moore[3]

Abstract

Purpose **The aim of this paper is to provide an overview of sample size estimations for the most frequent type of group studies that result in continuous, binary and ordered categorical outcomes.**

Methods **The theory behind power and sample size calculations is explained using the basic probability concepts that underpin the most frequently used statistical significance tests.**

Results **Simple formulae and tables are presented for the estimation of sample sizes necessary for efficient and effective clinical and epidemiological trials. These may be used without recourse to sophisticated and complex computer software packages. Mathematical complexity is kept to a minimum. Examples and applications from the vision sciences are specifically highlighted.**

Conclusions **The paper highlights, with practical examples, the concepts and computations necessary to make sample size estimations accessible to all eye professionals involved in research, diagnostic and statutory work.**

Eye (2009) **23,** 1589–1597; doi:10.1038/eye.2009.105; published online 15 May 2009

Keywords: power; sample size; ethics; odds ratio; ordinal data; clinical difference

Introduction

In recent years, Research Ethics Committees have paid increasing attention to the numbers required for the conduct of efficient research proposals from universities, official government research institutes, and private industrial organisations. Grant proposal forms also frequently include a section on the justification for sample size numbers proposed for clinical and epidemiological trials. Many research publications incorporating quantitative and qualitative data also demand consideration of appropriate sample sizes for adequate analysis and reporting.

A substantive number of papers[1–5] have been published with the objective of elucidating the calculations required for sample size estimation. The quoted references form a small proportion of the literature on the subject. Such work describes the basis of the necessary computations that rely on the fundamental concepts of probability, statistical significance, clinically meaningful differences, and the subtle concept of statistical power, the probability of detecting such differences. The latter is intimately and intricately linked to sample size estimation.

Although the subject area has been comprehensively dealt with in many research applications, there has tended to be a deficit of attention paid to it in some vision science research. This paper has the objective of elucidating the underlying concepts of power calculations and their specific application to vision science. Examples of these will be provided along with tables and formulae, which should prove to be of benefit to all eye-related professionals in the conduct of their research. Three main types of data will be considered, namely continuous, binary, and ordered categorical data. This paper will focus on studies in which two independent groups are to be compared. A second paper will concentrate on more sophisticated trials in which dependent groups are to be compared.

[1]Department of Statistics and Operational Research, School of Mathematics and Physics, The Queen's University of Belfast, Belfast, Northern Ireland, UK

[2]Mater Hospital, Belfast Health and Social Care Trust, Belfast, Northern Ireland, UK

[3]Centre for Molecular Biosciences, The University of Ulster, Northern Ireland, UK

Correspondence:
EA Goodall, School of Mathematics and Physics, The Queen's University of Belfast, Belfast BT7 1NN, Northern Ireland, UK
Tel: +44 0289 077 6580;
Fax: +44 0287 032 4375.
E-mail: edwardagoodall@hotmail.com

Received: 4 September 2008
Accepted in revised form: 11 April 2009
Published online: 15 May 2009

Proprietary interest: None

Materials and methods

Basic assumptions

It is assumed that researchers who are required to perform a power calculation are familiar with the basic measures of descriptive statistics such as mean (or average), median (or middle value), and standard deviation. The former two are used to summarise a set of data and are often referred to as measures of central tendency. The latter statistic provides an estimate of how closely (or not) data are spread about the mean. A combination of the two is frequently used to characterise a set of data that follows a normal distribution in which the mean and median approximately coincide. In this case, a graphical summary of the data is bell shaped. For such a distribution, about 95% of data are contained within 1.96 standard deviations of the mean. In the comparison of a set of means (each with standard deviation s, and based on n values each), the variance between means is reduced by a factor of n. The associated standard deviation, called the standard error, is estimated by $s/\sqrt{n}$. For a normal distribution of means, $\sim$95% should be contained within 1.96 standard errors of the overall mean. The boundaries of the relevant interval are called confidence limits or critical values.

Our next assumption is that researchers are acquainted with the basic notions of probability theory and, in particular, that the probability, p, of an event outcome lies between 0 and 1. There are many excellent texts on basic statistical theory. For revision purposes and the use of statistical software, Goodall[6] provides an introduction that focuses on applications in biomedical science without straying too far into the underpinning mathematical complexity.

Statistical significance

The concept of probability lies at the heart of statistical significance tests. It may be worthwhile to delineate the basic theory in which textbooks often use the simple experiment of tossing a coin. Such a trial has two possible outcomes, a head or a tail. If we tossed a coin 10 times and obtained 10 heads, we might ask ourselves whether the coin was really fair (that is, the probability of obtaining a head was the same as the probability of obtaining a tail and both were equal to 0.5). We might now think after carrying out our experiment that the coin was biased towards heads and that the probability of obtaining a head was higher than 0.5.

The above situation is similar to what happens in the design and analysis of clinical trials in a variety of applications in biomedicine. We have a hypothesis that we wish to test, for example, is one drug treatment really more effective than another? A hypothesis called the null hypothesis (often abbreviated to H_0) is set up that there is no difference. We then proceed to test this hypothesis by conducting an experiment to assess the impact of the drug on two sets of patients. If there is a major difference between the average results for each set, we would suspect that the null hypothesis is incorrect and that the alternative hypothesis, (H_A, one drug is superior) is true. However, it is often difficult to know where to draw the line in making a decision. This is where probability theory is important. In the coin tossing experiment, the probability of obtaining the observed result is $(1/2)^{10}\ldots(1/1024)$ that is, <0.001 if the null hypothesis is true ($P<0.001$). In this case, we would tend to favour the alternative hypothesis because the observed result was so unusual and unlikely. A scientist would now say that the results are very highly significant and reject the null hypothesis. A convention has grown up that if $P<0.05$, the results are said to be significant. With $P<0.01$, the results are said to be highly significant. Many favour the strategy of simply stating the probability of obtaining the experimental results found and allowing readers of their report to make up their own minds.

Summary

The null hypothesis (H_0) is that which we are happy to accept in the absence of any definite evidence to the contrary. It should say something precise about the population. Typically, it might say that the population is as it is claimed to be, or that the population has not changed.

The alternative hypothesis (H_A) is that for which we are seeking evidence. We shall accept it only if the evidence is reasonably conclusive.

We devise some rule so that, on the basis of the evidence available, we can make a decision either to accept H_0 or to reject H_0. We might make the correct decision (to accept H_0 when it is actually true, or to reject H_0 when it is actually false) but there are two possible types of error.

To reject H_0 when it is actually true.

This is called *Type 1 error.*

The probability of making this error is often denoted by α, and is called the significance level of the test.

To accept H_0 when it is actually false.

This is called *Type 2 error.*

The probability of making this error is often denoted by β.

The power of the test is $(1-\beta)$ that is, the probability of rejecting H_0 when it is false.

The tests that we shall consider are designed so as to have a given level of significance ($P<0.05$). Power should be at least 80%, preferably 90%. It is closely connected to the sample size of an experiment or survey.

If H_0 is rejected, this is a positive result. We can be reasonably confident that the alternative hypothesis H_A is true.

If H_0 is accepted, this does not mean that we are confident that H_0 is true. It simply means that the null hypothesis H_0 is plausible in the light of the evidence available, and cannot confidently be rejected.

Clinically important difference

A subject of confusion to some researchers is the definition of an outcome difference that is considered to be of practical importance. This is often confused with a statistically significant difference. However, the two are quite different. In fact, the most important consideration in the estimation of sample size is a realistic assessment of the minimum level of difference between two groups that it would be worthwhile to detect. Researchers should spend some time thinking about this at the beginning of their work as it can have a dramatic impact on the sample size numbers required. Well-designed research work should result in the detection of an important difference (assuming it actually exists) between groups that is also statistically significant that is, it incorporated a large enough sample size to be reasonably certain that the difference is not purely due to chance.

Tests for statistical significance

A number of standard tests have been developed to assess statistical significance between two groups of data. These are based on the theory discussed earlier. The tests vary depending on the type of data under investigation. They may be summarized as follows:(1) continuous, (2) binary, and (3) ordered categorical. It may be useful to review the main ones in current use.

t-*test*

In a two-group comparative study in which the outcome measure is a continuous variable that is approximately normally distributed (for example, blood pressure or reduction in level of astigmatism), the two sample *t*-test is the usual test of choice for the analysis of results. The appropriate test statistic is calculated as follows

$$t = \frac{(\bar{X}_1 - \bar{X}_2)}{s\sqrt{\frac{2}{n}}} \qquad (1)$$

The numerator estimates the true difference between the means of the two groups under investigation and the denominator is its standard error. The number in each group is n and s is obtained by taking the square root of the average of the two sample standard deviations squared. This formula assumes that there are equal sample numbers in each group. If there are not, then the term $(2/n)$ is replaced by $((1/n_1)+(1/n_2))$, where n_1 is the number in the first group and n_2 is the number in the second group. This situation sometimes pertains because of dropouts from a trial. The standard deviations are also appropriately weighted. However, to achieve fairness and balance, researchers usually aim to have equal numbers in each group before a trial commences. Cambridge Statistical Tables[7] provide critical values (at 5% ($P<0.05$), 1% ($P<0.01$), and 0.1% ($P<0.001$) levels of significance) to assess whether t is statistically significant. Estimates of t that are larger in magnitude than the critical values indicate statistical significance (that is, the difference between the means was unlikely to have arisen by chance). Note that, with a very large value of n, statistical significance can be achieved for a difference in means of no clinical importance. Standard statistical packages such as SPSS (Chicago, IL, USA), Minitab (PA, USA), SAS (NC, USA) and STATA (TX, USA) will also produce the *P*-value for assessment. As discussed earlier, if the difference between means is also considered to be clinically meaningful, a researcher may have discovered something of interest for further investigation. Throughout this paper, it will be assumed that differences between means could arise in either direction and that the tests are two sided.

Binomial test and Chi-square (χ^2) test

A variable that can only take two values is referred to as binary. For example, an eye treatment may result in success or failure or a symptom of disease may be present or absent. Variables such as these are said to follow a binomial distribution rather than a normal distribution. In the comparison of the two different eye treatments, the proportion of successes, (p), in each group is calculated and the two proportions compared using a binomial test. The philosophy is the same as in the earlier section with the test statistic calculated as follows

$$z = \frac{(p_A - p_B)}{s\sqrt{\frac{2}{n}}} \qquad (2)$$

The numerator estimates the true difference between the two proportions and the denominator is its standard error. The latter may be estimated by calculating the average of the two proportions, p, and letting $s = \sqrt{p(1-p)}$ (Appendix 1). (Cambridge Statistical Tables[7] again provide the relevant critical value). The

assessment of whether the difference between the two proportions arose by chance (or not) is also frequently performed using a Chi-square (χ^2) test that provides approximately the same *P*-value to reach a decision on statistical significance. Computer packages such as SPSS and Minitab also produce the correct *P*-value. Results are often expressed as odds ratios (OR). These are calculated by computing the odds of success for each of the two groups that is, the proportion of successes is divided by the proportion of failures. The smaller odds is then usually divided into the larger one, enabling a researcher to make a statement such as 'the odds of success for one eye treatment was three times that of the other'. This approach proves to be particularly useful when more than two proportions are to be compared simultaneously, for example, three levels of application of a drug treatment. In these trials, the logarithm of OR turns out to be more mathematically tractable. This statistic is frequently used in observational studies in which a risk factor is under investigation in a case–control study. A classic case is the linkage of smoking to lung cancer.

Mann–Whitney U-*test*

Many research studies result in an outcome measure that is defined on an ordered categorical scale. For example, a patient's subjective response to whether a surgical eye treatment was successful (or not) might be assessed using a Likert scale (strongly agree (1), agree (2), disagree (3), and strongly disagree (4)). In this situation, where two groups are to be compared, the usual test of choice is the Mann–Whitney *U*-test with allowance made for ties.[8] This is a distribution-free test that compares the sums of ranks and usually attempts to answer the question 'Are the medians of the two groups statistically significantly different?' (see Appendix 2 for more detail).

Power analyses

Suppose that we have a two-group comparative study in which the response is measured on a continuous scale. Further, we wish the sample size in each group to be sufficiently large to be 90% certain (power = 0.90) that a clinically important difference, d, will be detected at the 5% ($P<0.05$) level of significance. The appropriate test statistic is given by equation (1). Assuming a normal distribution, t needs to be at least 1.96 to achieve significance. The least difference between the two means (LSD) that will be statistically significant is given by rewriting equation (1) as

$$\mathrm{LSD} = 1.96 \times s\sqrt{\frac{2}{n}} \tag{3}$$

However, to be 90% certain of detecting a difference, d (that is, the clinically important difference defined earlier), we must also satisfy a second equation

$$\frac{\mathrm{LSD} - d}{s\sqrt{\frac{2}{n}}} = -1.28 \tag{4}$$

The value 1.28 is the critical point corresponding to the ninetieth percentile of a normal distribution. Rewriting equation (4) we have

$$\mathrm{LSD} = d - 1.28s\sqrt{\frac{2}{n}} \tag{5}$$

Equating the right-hand sides of (3) and (5), then solving for n, we obtain

$$\begin{aligned} n &= \frac{2(1.96 + 1.28)^2 s^2}{d^2} \\ &\doteq 21\left(\frac{s}{d}\right)^2 \end{aligned} \tag{6}$$

For 80% power, the 1.28 value reduces to 0.84, the eightieth percentile of a normal distribution, and

$$n \doteq 16\left(\frac{s}{d}\right)^2 \tag{7}$$

We can see that the lower power leads to a reduced sample size requirement. Also note that, for 90% power, if $s = d$ then we need 21 samples in each group. However, if the standard deviation, s, is twice the clinically important difference, d, then the sample size for each group increases dramatically to 84. For 80% power, the corresponding numbers required in each group are 16 and 64. Before a trial commences, s may be estimated from an earlier pilot study or from the research literature.

For binary data, where two proportions, p_A and p_B, are to be compared, first calculate $p = (p_A + p_B)/2$. If this value is not too small (greater than 0.10 say), estimate s by $\sqrt{p(1-p)}$ as in Section 2 and insert it into (6) and (7) to estimate the approximate sample sizes required in each group. Recall also that the clinically important difference, d, is defined by the researcher before the trial begins. For binary data, the numbers required are usually much higher than for continuous data. For example, if $p_A = 0.5$ and $d = 0.1$, p may be computed to be 0.55 and thus $s = 0.50$.

For 90% power,

$$\begin{aligned} n &= 21\left(\frac{0.5}{0.1}\right)^2 \\ &= 525 \text{ in each group} \end{aligned}$$

For 80% power, $n = 400$. These are slight overestimates as we are making assumptions that underlying normal distributions exist. If we wish to be more exact, correction

factors can be applied in both the continuous and binary cases. Tables 1 and 2 provide corrected values. The simple estimates provided above do not differ too radically from the latter and various scenarios can be investigated using only a calculator.

When the outcome or response variable from a trial is measured on an ordered categorical scale (for example, 1, 2, 3, 4, and 5), the power calculation for the appropriate Mann–Whitney *U*-test is not as straightforward as in the earlier two sections. However, by making a number of reasonable assumptions, we can arrive at a good approximation of the sample sizes required in each independent group.

The mathematics involved uses the concept of the OR to estimate the clinically meaningful difference, which the researcher thinks it is important to detect. As a guideline, if an eye treatment usually results in 40% success and researchers wish to use a new treatment in which a 30% superiority of 70% is claimed, they seek an OR of

$$\left[\frac{\left(\frac{0.7}{0.3}\right)}{\left(\frac{0.4}{0.6}\right)}\right] = 3.50$$

For one of the groups, the proportion of cases expected in each category of the scale has also to be specified. If we consider the earlier example with five categories, it is usually reasonable to assume that the mean proportions in each category are approximately equal and that the clinically meaningful difference, d, will be consistent. If we have two groups called A and B, then denote the proportions expected in group A by pA1, pA2, pA3, pA4, and pA5 with similar nomenclature for group B. If we denote the cumulative probabilities by CA1, CA2, CA3, CA4, and CA5, then CA1 = pA1, CA2 = (pA1 + pA2), and so on. The OR is the probability of a patient being in a given category or lower in one group compared with the other. For category 1, it is estimated by

$$\mathrm{OR1} = \frac{\left[\left(\frac{\mathrm{CA1}}{(1-\mathrm{CA1})}\right)\right]}{\left[\left(\frac{\mathrm{CB1}}{1-\mathrm{CB1}}\right)\right]}$$

OR2, OR3, and OR4 can be similarly calculated. The assumption that these ORs are all approximately equal justifies the use of the Mann–Whitney *U*-test as the best one to use. The combination of this assumption with the earlier assumption that the mean proportions in each group are roughly equal leads to a formula for the numbers required in each group.

For 90% power, this is given by

$$n = \frac{6[1.96 + 1.28]^2}{\left[\log_e \mathrm{OR}\right]^2}$$

For 80% power, the 1.28 reduces to 0.84. As usual, we have assumed a 5% level of statistical significance. If we assume an OR of 2, then $\log_e 2.0 = 0.693$. For 90% power, we require $((6 \times 10.5)/(0.693)^2) = 131$ in each group. For 80% power, 98 in each group are required.

The above estimate of the number, n, required in each group is approximate. A correction factor should be applied where the number of categories is small. The estimated n should be multiplied by $(1-(1/k^2))^{-1}$, where k is the number of categories. It is not really necessary when k is greater than five. For $k = 2$, 3, 4, and 5, it is easily shown that the correction factors are 1.333, 1.125, 1.067, and 1.042, respectively. Whitehead[4] has shown that there is only a relatively small increase in power to be obtained by increasing the number of categories beyond

Table 1 Sample sizes required per group at the two-sided 5% significance level for different values of (s/d) and power

(s/d)	*Power (1−β)*			
	95	*90*	*80*	*50*
2.50	164	133	100	49
2.00	105	86	64	32
1.50	70	48	36	18
1.25	42	34	26	13
1.00	27	22	17	9
0.50	9	6	5	4

Table 2 Sample sizes to detect a difference in two proportions, pA and pB, at a 5% significance level with 80% power

	pB															
pA	*0.15*	*0.20*	*0.25*	*0.30*	*0.35*	*0.40*	*0.45*	*0.50*	*0.55*	*0.60*	*0.65*	*0.70*	*0.75*	*0.80*	*0.85*	*0.90*
0.10	686	199	100	62	43	32	25	20	16	14	11	10	8	7	6	5
0.15		906	250	121	73	49	36	27	22	17	14	12	10	8	7	6
0.20			1094	294	138	82	54	39	29	23	18	15	12	10	8	7
0.25				1251	329	152	89	58	41	31	24	19	15	12	10	8
0.30					1377	356	163	93	61	42	31	24	19	15	12	10
0.35						1471	376	170	96	62	43	31	24	18	14	11
0.40							1534	388	173	97	62	42	31	23	17	14
0.45								1565	392	173	96	61	41	29	22	16

Table 3 Numbers required in each group of ordered categorical data (for five categories) and for varying odds ratios and powers

Odds ratio	*Power*			
	50	*80*	*90*	*95*
1.5	146	299	400	494
2	50	102	137	169
3	20	41	54	67
4	13	26	34	43
5	9	19	25	31
10	5	9	12	15

five. Thus, any saving in the number of patients to be recruited is minimal. Table 3 provides the numbers required for different powers and ORs when the number of categories chosen is five.

A special case of the ordered categorical data scenario is when the number of categories is equal to two. The situation here is much the same as for binary data analysis. Suppose, for example, that researchers wish to show that a new eye treatment increases success rates from 40 to 60%, corresponding to an OR of 2.24. For 80% power and 5% significance, the numbers required in each group, using the computation of this section, is given by

$$n = \frac{(1.333)\times(47)}{0.65} = 96$$

This number is approximately the same as if the analysis had been performed using the binomial test for comparing two proportions.

Results and examples

Continuous data

In a paper by Wong *et al*,[9] the authors evaluated the difference in the duration and power of phacoemulsification required between the *in situ* fracture variation of divide and conquer (DC) nucleofractis as described by Shepherd[10] and the phaco chop (PC) technique, a variation of divide and conquer nucleofractis in which nuclei are divided into fragments with a phaco chopper instrument without the use of central sculpting, as described by Koch.[11] They found that PC required significantly less ($P<0.001$) phaco time in minutes (mean = 1.2, standard deviation = 0.79, $n=62$) than DC (mean = 2.4, standard deviation = 0.74, $n=55$). Other researchers wish to investigate whether they can identify comparable improved outcomes and, in the preparation of their Research Proposal, are asked to perform a power calculation to assess the sample numbers required in each experimental group. They are also asked what they wish to achieve and they state that a clinically meaningful difference between the means of the two outcome groups is 0.5 min. A standard deviation, s, for the computation is taken from the earlier study as 0.75, from the square root of the weighted averages of the two reported sample standard deviations. Assuming continuous data that are approximately normally distributed, it is envisaged that the t-test will be used to assess the results. The sample size (n) required for each group can be estimated from the relevant method delineated earlier. Thus, for 80% power and 5% significance,

$$n = \frac{16\times(0.75)^2}{(0.50)^2} = 36$$

For 90% power, $n=47$. It is also prudent to allow for 'dropouts' after initial recruitment of patients. This level may be taken as around 10–15%. Allowing for the latter and 80% power, the researchers need to seek a total recruitment size of 84 patients. The (d/s) value, in this case ($(0.5/0.75)=0.67$), is usually referred to as the standardised difference. Some Ethics Committees and Research Grant fund holders may wish to see several scenarios presented, that is, varying standardised difference ratios with a range of clinically meaningful differences, d, being defined.

Binary data

Our next example will use a paper by Sullivan *et al*.[12] The purpose of the research was to evaluate the imaging characteristics of a cohort of patients with ocular adnexal lymphoproliferative disease (OALD). One of the major results reported was that positron emission tomography (PET) upstaged 71% of patients with systemic lymphoproliferative involvement, having a higher sensitivity than computed tomography (CT) in detecting distant disease (86 *vs* 72%). In statistical studies, sensitivity is defined as the proportion of true positives correctly identified by a clinical test or procedure.

Many studies require the comparison of two sensitivities, such as the one quoted above, in which the outcomes are essentially binary in nature. Suppose that other researchers want to conduct a similar type of trial, perhaps with an enhancement of the earlier procedures used, and proceed to define a clinically meaningful difference between the two relevant proportions as being of the order of 20%. The null hypothesis is stated that both procedures yield sensitivities of 70% and the alternative hypothesis is that there is a 20% difference between the two. They are allowing for the possibility that the difference could be in either direction (for example, two-sided) but really expect the newer

procedure to be superior to a standard one. As before, a Research Ethics Committee demands a power calculation before approving the research. The actual test of choice here should be a variation on the χ^2 test called McNemar's test,[13] as it is planned to assess two procedures on the same set of patients. This involves the analysis of correlated data and will be dealt with in a future paper. For the purposes of this example, it will be assumed that we wish to compare a proportion of 0.7 with one that may be 0.2 better (or worse) using a straightforward test of proportions or χ^2 test. In this case, $d = 0.20$ and $s = \sqrt{p(1-p)}$ where p can be taken as about 0.8 (that is, $p = (0.7 + 0.9)/2$) and thus $s = 0.40$. For a power of 80%, the sample size required for each group is estimated by

$$n = 16\frac{(0.4)^2}{(0.2)^2} = 64$$

For 90% power, $n = 84$. Allow for 'dropouts' as before and the researchers should satisfy an Ethics Committee.

If the difference between two procedures is suspected to be in the other direction (for example, 0.7 *vs* 0.5), the sample size for each group will need to be considerably higher as p will now equal 0.60 and s will be 0.49. For a power of 80%, the sample size for each group is then given by

$$n = 16\frac{(0.49)^2}{(0.2)^2} = 96$$

Another example is provided by the Macular Photocoagulation Study.[14] In research work on age-related macular degeneration, 0.58 (98/169) of an untreated group and 0.49 (86/174) of a group treated with krypton photocoagulation had lost six or more lines of visual acuity in the study eye after 3 years of follow-up. If the original objective was to detect a difference of 0.10 in the two groups, with the null hypothesis that there was no difference (that is, both about 0.60), then for the earlier power calculation, $p = 0.55$ and $s = 0.50$. For a power of 80%, the sample size required for each group is estimated by

$$n = 16\frac{(0.5)^2}{(0.10)^2} = 400$$

Ordered categorical data

A research paper by McMonnies[15] delineates the use of a detailed questionnaire to elucidate the diagnosis and level of severity in patients with presenting signs of 'dry eye'. This approach results in a patient being scored on an ordered categorical scale of 1–25. Suppose that a researcher wishes to initiate a study in which the scale scores will be compressed into four categories indicating the relative severity of the condition with a four corresponding to very severe, three to severe, two to moderate, and one to mild. The researcher wishes to use this scale to assess the claim of a pharmaceutical company that their new eye ointment produces improvements in treatment that are twice as good as those of a competitor. From earlier analysis of retrospective results, this claim is interpreted as meaning that the proportion of patients still recording a very severe condition after treatment will be 0.5 in one group and 0.25 in the group using the new eye ointment, equating to an expected OR in the latter's favour of about three. This assumption is extrapolated across the categories as defined earlier in our consideration of ordered categorical data. As usual, an Ethics Committee is requested to consider a Research Proposal including a power calculation to estimate the sample size required in each eye treatment group. For 80% power, this can be estimated from

$$\begin{aligned} n &= \frac{47}{(\log_e \mathrm{OR})^2} \\ &= \frac{47}{(\log_e 3)^2} \\ &= \frac{47}{(1.207)} \doteq 39 \end{aligned}$$

Multiplying by the correction factor, 1.067, for four categories, this yields a sample size of 42 per group.

The Mann–Whitney *U*-test is often used on data that are not ordered on a categorical scale (as before), but an initial analysis of continuous data has shown a significant deviation from normality. An example might be data measurements of visual acuity, which, even on the logMAR scale, may not satisfy normality assumptions. For the purposes of power calculations, it could be considered as approximately categorical in nature for example, categories correspond to best corrected visual acuities of 0–0.3, >0.3–0.6, and so on. This approach is often taken when different types of lens are to be compared with the objective of showing that a newer type leads to significantly better improvements in eyesight. Note, however, that this strategy may lead to greater sample sizes being demanded. In general, a Mann–Whitney *U*-test is less powerful than a *t*-test, except perhaps in cases where outliers in the data exist. The former test is less sensitive to the existence of these as medians, rather than means, are being investigated.

Discussion

Awareness of the need for power calculations to estimate sample sizes has become more widespread in recent

years. Their requirement has been increasingly demanded by the Ethics Committees of Hospitals, Universities, and by those committees formally established by regional and national government action. Research Grant fund holders also inevitably necessitate the production of a power analysis to justify expenditure on valuable recurrent and capital resources.

Although it is advisable to consult a statistician at the onset of planning of a clinical or epidemiological trial, many organisations have limited access to such a service as it is often shared by a large number of medical research staff across a wide band of disciplines. Thus, the routine availability of advice may be constricted. When collaboration is instigated, it helps if both sides can speak the language of the other. It is therefore useful if researchers have an acquaintance with the nomenclature required for good statistical design and analysis.

Power calculations require some knowledge of the meaning of statistical significance level, probability, and the concept of a clinically meaningful difference or effect size. The three main types of numerical scale used in quantitative research are continuous, binary, and ordered categorical. The efficient computation of sample size needed in a two-group comparison trial is dominated by the correct choice of scale and the relevant statistical test. In turn, the choice of test leads to the most appropriate type of power calculation.

It has been the authors' experience that there has been, in some areas of the vision sciences, minimal contact and collaboration between statisticians and researchers. This paper has attempted to bridge a gap by conducting a review of the statistical theory required for the computation of adequate sample sizes needed for research trials. It is hoped that the paper will provide guidelines and form the basis for fruitful advice to those who are engaged in ophthalmologic research but who, through no fault of their own, have had little training in the use of statistical analysis. For the sake of simplicity, the paper has concentrated on power analyses for two group comparisons in which the outcomes are assumed to be independent. A future paper will concentrate on the more complex challenges faced when groups of correlated data are to be compared (for example, different tests on the same set of patients).

Although the demand for power calculations seems unnecessarily bureaucratic to some researchers, it nevertheless presents an opportunity to think in more depth about the design of a study, its aims and objectives, and how the final statistical analysis will be conducted. Many are surprised by the actual numbers required for the efficient conduct of their research. However, this challenge inevitably arises from situations in which the standard deviation of a set of data exceeds the clinical difference which it is considered important to detect. Large effect sizes combined with a comparatively low standard deviation will, in turn, reduce the sample size required for the research.

Acknowledgements

We wish to acknowledge the seminal work of Sir Ronald Aylmer Fisher[16] and his classic text on Statistical Methods for Research Workers.

References

1 Campbell MJ, Julious SA, Altman DG. Estimating sample sizes for binary, ordered categorical, and continuous outcomes in two group comparisons. *Br Med J* 1995; **311**: 1145–1148.
2 Daly L. Confidence intervals and sample sizes: don't throw out all your old sample size tables. *Br Med J* 1991; **302**: 333–336.
3 Lehr R. Sixteen s squared over d squared: a relation for crude sample size estimates. *Stat Med* 1992; **11**: 1099–1102.
4 Whitehead J. Sample size calculations for ordered categorical data. *Stat Med* 1993; **12**: 2257–2272.
5 Fayers PM, Machin D. Sample size: how many patients are necessary? *Br J Cancer* 1995; **72**: 1–9.
6 Goodall EA. *Statistics for Biomedical Science*, 1st ed. Blueberry Publishing: UK, 2007.
7 Lindley DV, Scott WF. *New Cambridge Statistical Tables*, 2nd ed. Cambridge University Press: UK, 1984.
8 Mann HB, Whitney DR. On a test of whether one of 2 random variables is stochastically larger than the other. *Ann Math Stat* 1947; **18**: 50–60.
9 Wong T, Hingorani M, Lee V. Phacoemulsification time and power requirements in phaco chop and divide and conquer nucleofractis technique. *J Cataract Refract Surg* 2000; **26**(9): 1374–1378.
10 Shepherd JR. *In situ* fracture. *J Cataract Refract Surg* 1990; **16**: 436–440.
11 Koch PS, Katzen LE. Stop and chop phacoemulsification. *J Cataract Refract Surg* 1994; **20**: 566–570.
12 Sullivan TJ, Valenzuela AA. Imaging features of ocular adnexal lymphoproliferative disease. *Eye* 2006; **20**: 1189–1195.
13 McNemar Q. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* 1947; **12**: 153–157.
14 Macular Photocoagulation Study Group. Krypton laser photocoagulation for neovascular lesions of age-related macular degeneration. Results of a randomized clinical trial. *Arch Ophthalmol* 1990; **108**(6): 816–824.
15 McMonnies CW. Key questions in a dry eye history. *J Am Optom Assoc* 1986; **57**: 512–517.
16 Fisher RA. *Statistical Methods for Research Workers*, 14th ed. Oliver and Boyd Ltd: Edinburgh, 1970.

Appendix 1

Binomial distribution

If a sterling coin is tossed, there are two outcomes, a head (H) or a tail (T). If the coin is fair, the probability, p, of each is $(1/2)(=0.5)$. If a coin is tossed 100 times ($n=100$), on average, we should obtain 50 heads. Mean $=np=$ $(100)2)=50$. In practice, we will not obtain exactly 50. A variable X with a Binomial distribution is known to have a variance of $np(1-p)$. The standard deviation is thus $\sqrt{np(1-p)}$. For the above example, the standard deviation is five. For large n, the binomial may be approximated by a normal distribution and we can make statements such as 95% confidence limits for the mean are mean $(np)\pm 1.96\sqrt{np(1-p)}$. For the above example, the limits would be approximately $40\rightarrow 60$ $[50\pm 10]$. If we obtained values outside these limits, we would suspect that the coin was not fair.

Suppose we are investigating the sensitivity of a test that is, proportion p of test positives (X) from a sample of n true positives

$$\begin{aligned}\text{variance}(p) &= \text{variance}\left(\frac{X}{n}\right)\\ &= \frac{1}{n^2}\text{variance}(X)\\ &= \frac{1}{n^2}n(p)(1-p)\\ &= \frac{p(1-p)}{n}\\ \text{Hence standard error} &= \sqrt{\frac{p(1-p)}{n}}\end{aligned}$$

Appendix 2

The Mann–Whitney U-test

This is a distribution-free ranking test that asks the question 'Are the medians of two sets of data the same?'

1. Suppose there are two samples, with values $x_1, x_2, x_3, \ldots, x_{N_x}$ and $y_1, y_2, y_3, \ldots, y_{N_y}$. Rank them in order together. This gives a sequence such as

$$x\,y\,y\,x\,x\,y\,x$$

2. For each x value, count the number of y values that come after it. Thus, in the above example, the first x precedes three y values, the second one, the third one, and the fourth none.

3. Form the total $3+1+1+0$ and call it U_x, the number of times an x precedes a y. In the same way, find U_y—here $3+3+1=7$. Check that

$$U_x + U_y = N_x N_y$$

Under the null hypothesis that the averages are the same, one expects $U_x=U_y=(1/2)N_xN_y$, as each x value will on average have half the y sample behind it and the other half in front. If the medians are significantly different, say x is ahead of y, then the x values will precede more than their fair share of y values, and U_x will be greater than U_y. For small samples, the significance is given by tables. For large samples, the normal approximation can be used, with the mean of U_x equal to $(1/2)N_xN_y$, and variance $(1/12)N_xN_y(N_x+N_y)$.