

Sample size estimation and statistical power analyses

Bhavna Prajapati, Mark Dunne & Richard Armstrong

The concept of sample size and statistical power estimation is now something that Optometrists that want to perform research, whether it be in practice or in an academic institution, cannot simply hide away from. Ethics committees, journal editors and grant awarding bodies are now increasingly requesting that all research be backed up with sample size and statistical power estimation in order to justify any study and its findings.¹ This article presents a step-by-step guide of the process for determining sample size and statistical power. It builds on statistical concepts presented in earlier articles in Optometry Today by Richard Armstrong and Frank Eperjesi.²⁻⁷

Basic statistical concepts

There are several statistical concepts that must be grasped before reading this article. The first is the concept of hypothesis testing. Convention has it that any difference or effect found in an experiment has been caused by chance alone. This is referred to as the null hypothesis. Statistical analysis determines whether the null hypothesis is correct or not. If analysis indicates that the difference or effect is not likely to have occurred by chance then the null hypothesis is rejected in favour of the alternative hypothesis, stating that a real effect has occurred. Rarely will you see the terms null and alternative hypothesis used in scientific papers. Instead, a finding is described as “not statistically significant” if the null hypothesis is accepted and “statistically significant” if the alternative hypothesis is accepted. Clearly, a criterion must

be set for rejecting the null hypothesis. This is referred to as the alpha level (α). Alpha is often set at 0.05 or 5%.^{8,9} Statistical analysis is then carried out in order to calculate the probability that the difference or effect was purely due to chance. The null hypothesis is only rejected if the probability (P-value) is equal to or less than the alpha level.

This process however has two potential errors; type I and type II. A type I, or false-positive, error occurs if the null hypothesis is rejected incorrectly. There is a 5% chance of this occurring if the alpha level is set at 0.05. A type II, or false-negative, error occurs if the null hypothesis is accepted incorrectly. A beta (β) level can be chosen as protection against this type of error.

What is statistical power?

Statistical power (P) is defined as: $P = 1 - \beta$

Power is dependent on a number of factors, which will be explained later. Statistical power is conventionally set at 0.80 or 80%¹⁰ i.e. there is a 20% chance of accepting the null hypothesis in error, i.e. beta is 0.20 or 20%.

Why is statistical power important?

Sample size estimation and statistical power analyses are important for a number of reasons. Firstly, it is increasingly becoming a requirement for most research proposals, applications for ethical clearance and journal articles. Research ethics committees often ask for justification of the study based on sample size estimation and statistical power. It would not be ethically acceptable to conduct a study that would not be stringent enough to detect a real effect due to a lack of statistical power. Equally, it would not be ethically acceptable to conduct a study by recruiting thousands of participants when sufficient data could be obtained with hundreds of participants instead. Recruiting more participants than required would also be a waste of both resources and time.

How large should a sample size be?

Unfortunately there is no one simple answer to this question. As a rule, larger sample sizes have more statistical power. However, other factors need to be considered, as discussed below.

Effect size

This is the smallest difference or effect that the researcher considers to be clinically relevant. Determining the effect size can be a difficult task. In some cases it can be based on data from previous studies. A pilot study may be required for this purpose or expert clinical judgement could be sought. For circumstances where none of these options apply, Cohen¹¹ has determined

Test	Effect size	Small	Medium	Large
Difference between two means	d	0.20	0.50	0.80
Difference between many means	f	0.10	0.25	0.40
Chi-squared test	w	0.10	0.30	0.50
Pearson's correlation coefficient	ρ	0.10	0.30	0.50

Table 1

Small, medium and large effect sizes as defined by Cohen¹¹

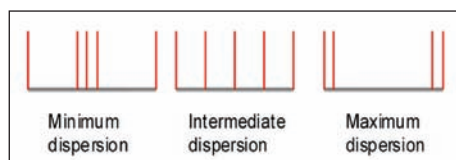


Figure 1

Dispersions for effect size calculations for F tests

standardised effect sizes described as “small”, “medium” and “large” (see Table 1). These vary for different study designs. For smaller effect sizes a larger sample size would be required.

Alpha level

For a smaller alpha level a larger sample size is needed and vice versa.

Standard deviation

Effects being investigated often involve comparing mean values measured in two or more samples. Each mean value will be associated with a standard deviation. As standard deviation increases a larger sample size is needed to achieve acceptable statistical power. Again, the standard deviations expected in a sample need to be estimated based on clinical judgement, previous (pilot) studies and/or other published literature.⁹

One or two-tailed statistical tests

There are two types of alternative hypothesis. The first is one-tailed and is appropriate when a difference in one direction is expected. For example, it might be hypothesised that sample A has a higher intraocular pressure (IOP) than sample B. The second is two-tailed and is appropriate when a difference in any direction is expected. For example, it might be hypothesised that sample A has a different IOP to sample B, but it could be higher or lower. One-tailed alternative hypotheses require smaller sample sizes. However, the use of one-tailed tests should be justified and not be used purely to reduce the sample size required.

Formulae for determining effect size

Table 2 shows how effect size is calculated for some common statistical tests. Some formulae (see equations 2-5 in Table 2) to determine effect size for the difference between many means require a prior knowledge of the dispersion of the means of each group. There are

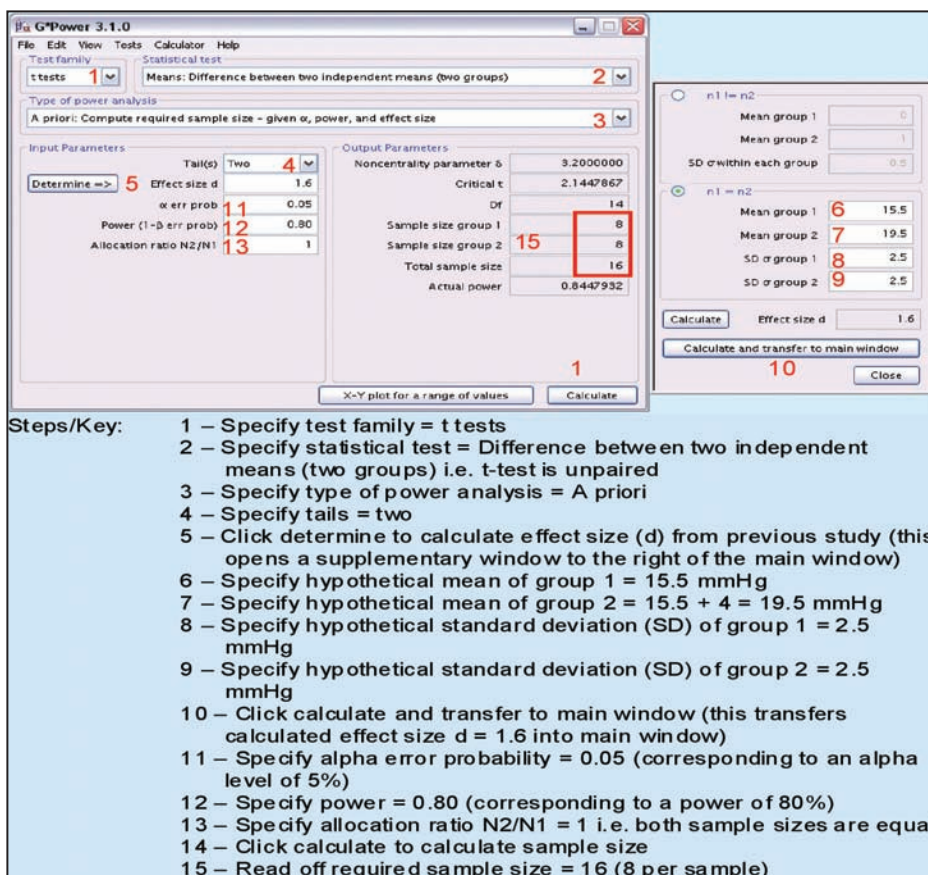


Figure 2

Computing sample size for an unpaired t-test using GPower 3

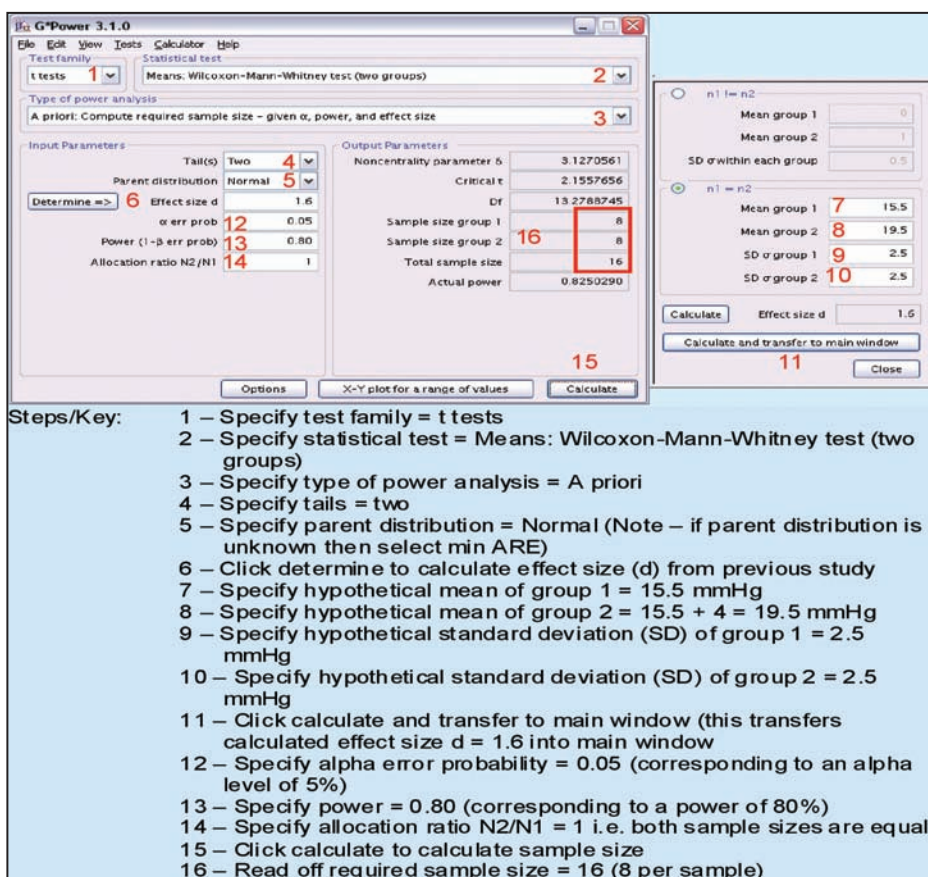


Figure 3

Computing sample size for Wilcoxon Mann Whitney U test using GPower 3

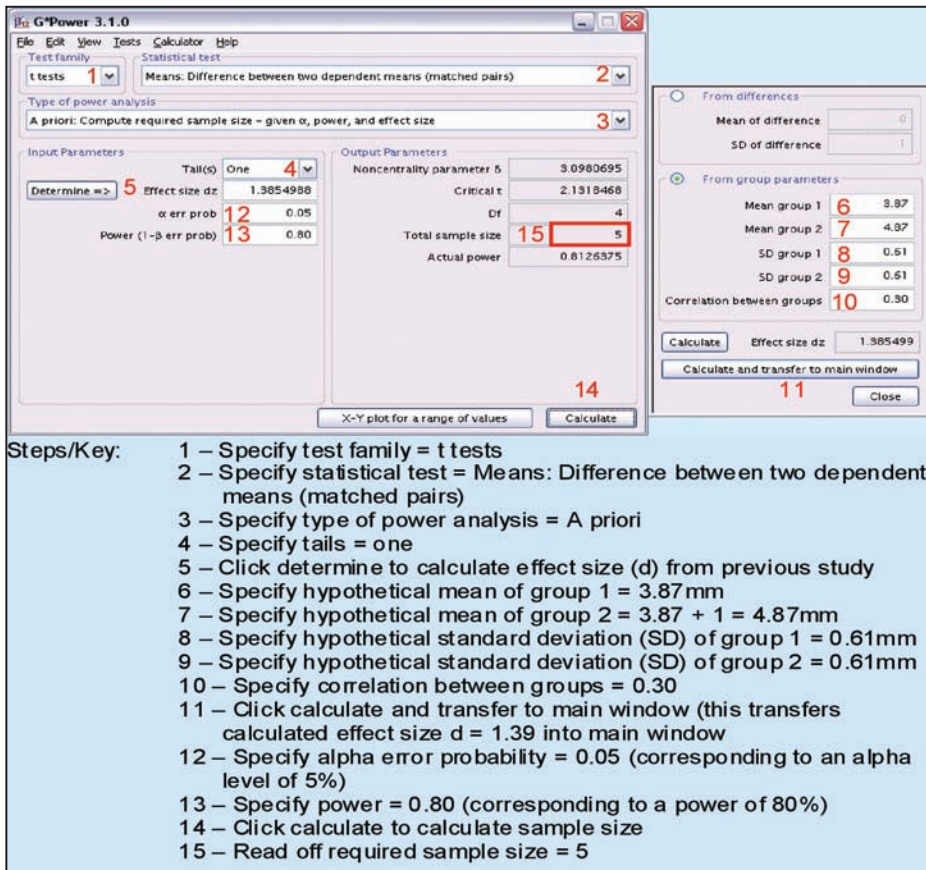


Figure 4

Computing sample size for a paired t-test using GPower 3

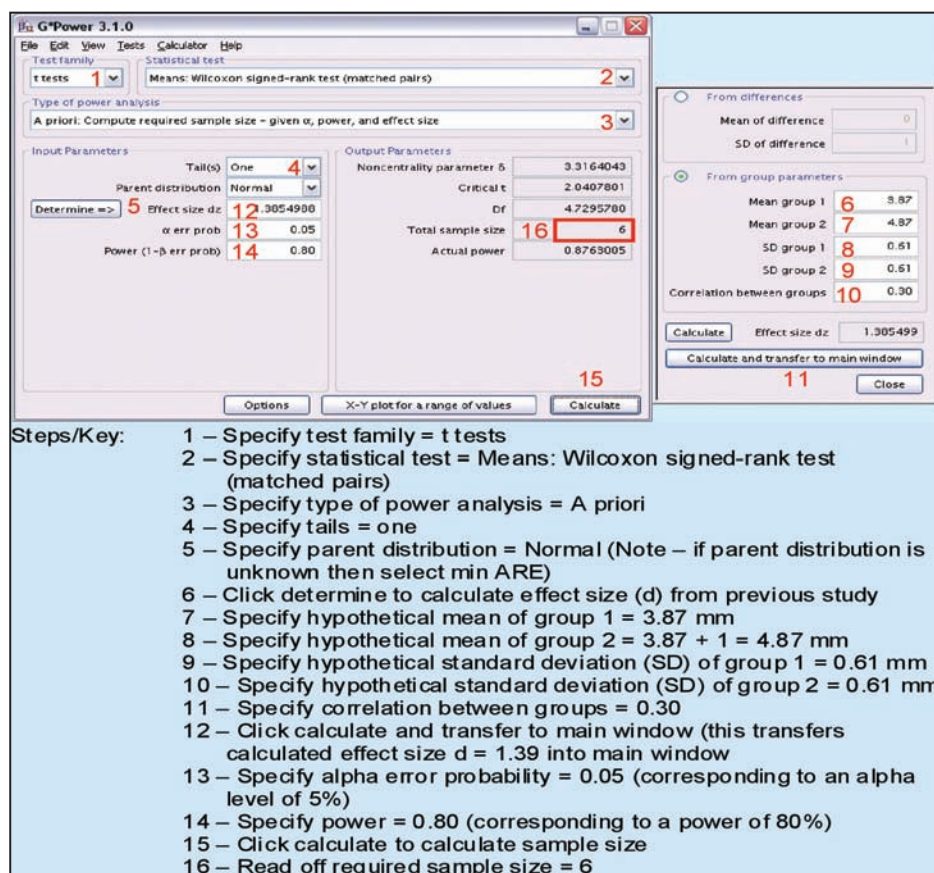


Figure 5

Computing sample size for Wilcoxon signed-ranks test using GPower 3

three types of dispersions (Figure 1). A minimum dispersion is one where there is one mean value at each extreme and the rest are clustered at the mid point. An intermediate dispersion is one where all means are equally spread out. A maximum dispersion is one where all means are clustered near the two extremes.

In some cases (see equation 6 in Table 2) the effect size f can be determined based on eta squared (η^2). This is a measure of association and is the proportion of the total variance that is attributed to an effect. Eta squared ranges from 0 to 1 and as a rule 0.01 is a small effect, 0.06 is a medium effect and 0.14 is a large effect.

Parametric versus non-parametric statistical tests

There are two major types of statistical test, parametric and non-parametric. Parametric tests are more powerful but less robust as they make assumptions about the frequency distribution of the data being analysed i.e. the data is assumed to follow a normal distribution.² Non-parametric statistical tests make no assumptions about the frequency distribution of data and this makes them more robust but less powerful.¹² It follows that larger sample sizes will be required when using less powerful non-parametric statistical tests.¹² The sample size required for a non-parametric test is determined by multiplying the sample size calculated for an equivalent parametric test by a correction factor. This correction factor is referred to as the asymptotic relative efficiency (ARE) and was first described by Pitman.¹³ The value of the ARE varies depending on the nature of the parent distribution (the distribution of the population from which the sample is drawn). For the purposes of ophthalmic research, it would be reasonable to assume that the parent distribution is normal. Table 3 shows ARE values based on a normal parent distribution for some common non-parametric tests. In total there are over 100 different formulae and methods of determining sample sizes for different statistical tests and study designs.¹⁴ They all make slightly different assumptions about the data and so may yield slightly different results. The good news is that there are also computer programs that are freely

Test		Effect size (d, f or w)	
Difference between two means (assumes equal sample sizes)		$d = \frac{\mu_1 - \mu_0}{\sigma}$	(1)
Difference between many means (assumes equal sample sizes)	Minimum dispersion	$f = d \times \sqrt{\frac{1}{2k}}$	(2)
	Intermediate dispersion	$f = d \times \frac{1}{2} \sqrt{\frac{(k+1)}{3(k-1)}}$	(3)
	Maximum dispersion (k = odd)	$f = d \times \frac{\sqrt{k^2 - 1}}{2k}$	(4)
	Maximum dispersion (k = even)	$f = d \times \frac{1}{2}$	(5)
Difference between many means (assumes equal sample sizes)		$f = \sqrt{\frac{\eta^2}{1 - \eta^2}}$	(6)
Chi-squared test		$w = \sqrt{\sum_{i=1}^r \frac{(P_{1i} - P_{0i})^2}{P_{0i}}}$	(7)
Pearson's correlation coefficient		ρ	(8)

Table 2

Formulae to determine effect sizes for common statistical tests. Note: In the case of comparing the difference between many means, d is calculated using the difference between the highest and lowest means. Key to symbols: μ_1 = mean of sample 1; μ_0 = mean of sample 2; σ = standard deviation; k = number of groups; ρ = Pearson's correlation coefficient; η^2 = eta squared; P_{1i} = the proportion in cell i under the alternative hypothesis; P_{0i} = the proportion in cell i under the null hypothesis; r = rows in chi square table; c = columns in chi square table

Parametric test	Equivalent non-parametric test	ARE
One sample t test	Wilcoxon One sample test	0.955
Paired t test	Wilcoxon Signed-ranks test	0.955
Unpaired t test	Mann Whitney U test	0.955
Pearson's correlation coefficient	Spearman and Kendal's correlation coefficient	0.910
One way ANOVA	Kruskal-Wallis test	0.955
Repeated measures ANOVA	Friedman test	$\frac{0.955k}{k+1}$

Table 3

Asymptotic relative efficiency (ARE) of some common non-parametric tests. "k" is the number of groups

available, such as GPower^{3,15} which will do all the hard work for you.

GPower 3

Types of analyses

GPower 3 is capable of computing five different types of power analyses. These

are *a priori*, post hoc, compromise, criterion and sensitivity power analysis. Of these, the *a priori* power analysis is the most relevant to sample size estimation, as it involves determining the sample size required for any specified power, alpha level and effect size.

Post hoc power analysis involves determining the level of statistical power achieved for a given sample size, effect size and alpha level. Therefore, this type of power analyses is most useful at the end of a study. Here, it is important that the clinically relevant effect size is specified and not the actual effect size found based on the results of the study.

A compromise power analysis involves determining the alpha level and statistical power based on the sample size, the effect size and the error probability ratio "q" where q = beta/alpha. This is useful in scenarios where an *a priori* power analysis yields a larger sample size than is feasible. In these circumstances, the maximum feasible sample size is specified and a compromise power analysis is used to alter the alpha level and power based on the error probability ratio.

A criterion power analysis involves computing the alpha level based on the sample size, the effect size and the statistical power level. This type of power analysis should be used as an alternative to post hoc power analysis where the control of alpha is less important than the control of beta.

A sensitivity power analysis involves determining the effect size based on the sample size, statistical power level and the alpha level. This type of power analysis can be used when critically evaluating research published by others. It allows you to determine the minimum effect size that the study was sensitive to for a certain level of power, based on the sample size recruited and the alpha level specified.

Types of tests

GPower 3 is capable of performing power analysis for over 40 different experimental designs. These are classified into five families of statistical tests; exact tests, t tests, f tests, χ^2 tests and z tests. Worked examples based on the most commonly used tests are discussed in more detail below.

Worked examples using GPower 3

Comparing two independent means

Consider an experiment designed to test if IOP was different in males compared to females. If an equal number of subjects were to be recruited in each

- Steps/Key:**
- 1 – Specify test family = t tests
 - 2 – Specify statistical test = Means: Difference between two dependent means (matched pairs)
 - 3 – Specify type of power analysis = A priori
 - 4 – Specify tails = one
 - 5 – Click determine to calculate effect size (d) from previous study
 - 6 – Specify hypothetical mean of group 1 = 3.87mm
 - 7 – Specify hypothetical mean of group 2 = 3.87 + 1 = 4.87mm
 - 8 – Specify hypothetical standard deviation (SD) of group 1 = 0.61mm
 - 9 – Specify hypothetical standard deviation (SD) of group 2 = 0.61mm
 - 10 – Specify correlation between groups = 0.30
 - 11 – Click calculate and transfer to main window (this transfers calculated effect size d = 1.39 into main window)
 - 12 – Specify alpha error probability = 0.05 (corresponding to an alpha level of 5%)
 - 13 – Specify power = 0.80 (corresponding to a power of 80%)
 - 14 – Click calculate to calculate sample size
 - 15 – Read off required sample size = 5

Figure 6

Computing sample size for a one-way ANOVA using GPower 3

- Steps/Key:**
- 1 – Specify test family = F tests
 - 2 – Specify statistical test = ANOVA: Fixed effects, special, main effects and interactions
 - 3 – Specify type of power analysis = A priori
 - 4 – Click determine to calculate effect size
 - 5 – Select Direct
 - 6 – Specify eta squared = 0.06
 - 7 – Click Calculate and transfer to main window
 - 8 – Specify alpha error probability = 0.05 (corresponding to an alpha level of 5%)
 - 9 – Specify power = 0.80 (corresponding to a power of 80%)
 - 10 – Specify numerator degrees of freedom = 1 (this is the degrees of freedom for the factor being analysed (number of levels - 1) i.e. for the gender factor this is 2 - 1 = 1)
 - 11 – Specify the number of groups = 4 (there are four possible groups in this study design; (i) males with blue irides, (ii) males with brown irides, (iii) females with blue irides and (iv) females with brown irides)
 - 12 – Click calculate to calculate sample size
 - 13 – Read off required sample size = 125

Figure 7

Computing sample size for the gender factor of a factorial ANOVA using GPower 3

group, how many subjects would be required to achieve 80% power at the 5% alpha level? This would be analysed with a t-test (parametric test). The test would also be two-tailed, since IOP could be higher or lower in males compared to females. An unpaired t-test would be used, as the two sets of IOP measurements would represent independent means, having been measured in different subjects.

Firstly, the effect size needs to be determined. A clinically relevant difference of 4 mmHg is chosen based on clinical judgement. Previous literature shows that in normal healthy eyes mean IOP is 15.5±2.5 mmHg.¹⁶ Figure 2 shows how this information is used to determine that the effect size of interest (d) is 1.6. An *a priori* analysis in GPower 3 then shows that 8 subjects would be required in each group (Figure 2).

If it were later found that the readings for IOP were not normally distributed, then a Wilcoxon Mann Whitney U test would have to be used instead. This is the non-parametric equivalent of an unpaired t-test (Figure 3).

Note that although the required sample size for both the unpaired t-test and the Wilcoxon Mann Whitney U test is identical in this case, the actual power for the unpaired t-test (0.845) is greater than the Wilcoxon Mann Whitney U test (0.825).

Comparing two dependent means

Consider an experiment investigating whether a new mydriatic drug had any affect on pupil diameter. In this study design, pupil size would be measured in a group of subjects with and without the drug instilled. The pupil sizes under each condition would represent dependent means because both sets of measurements were taken in the same subjects. A one-tailed test would also be used, as it is only feasible that the drugs will dilate the pupils. How many subjects would be required for 80% power at the 5% alpha level? Firstly, the effect size would need to be determined. A clinically relevant difference of 1mm is chosen based on clinical judgement. Current literature shows that the mean pupil diameter is 3.87mm with a standard deviation of 0.61mm.¹⁷ When looking at dependent means the correlation between the

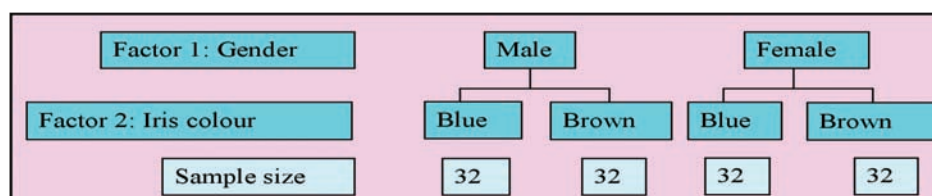


Figure 8

Sample size required for 2 (gender) x 2 (iris colour) factorial ANOVA

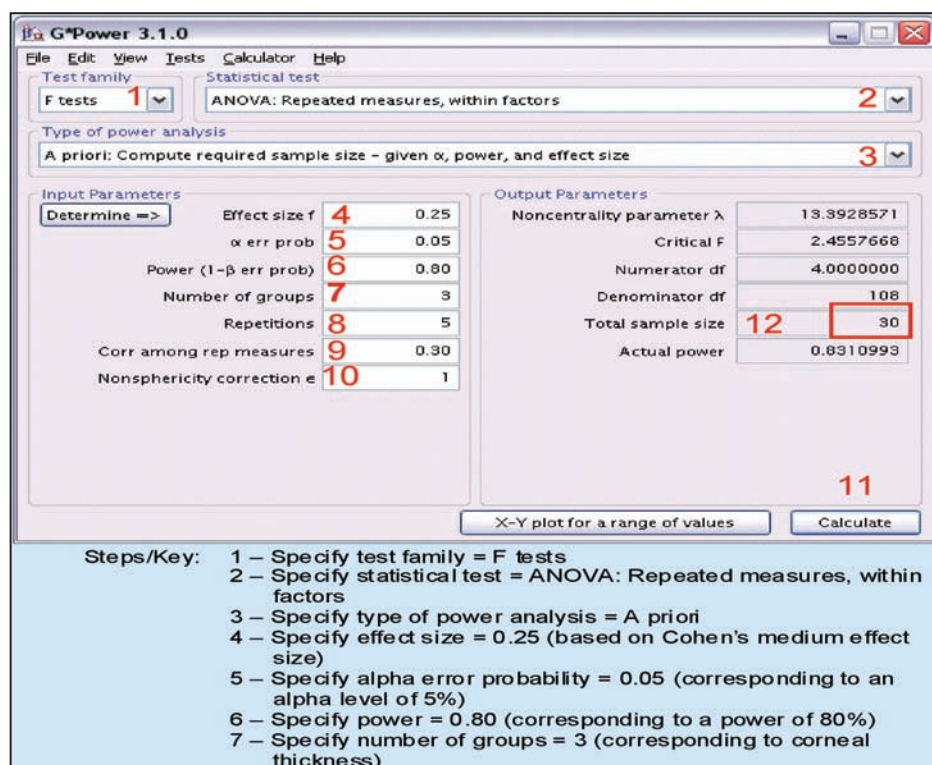


Figure 9

Computing sample size for within-subjects factor for a repeated measures ANOVA using GPower 3

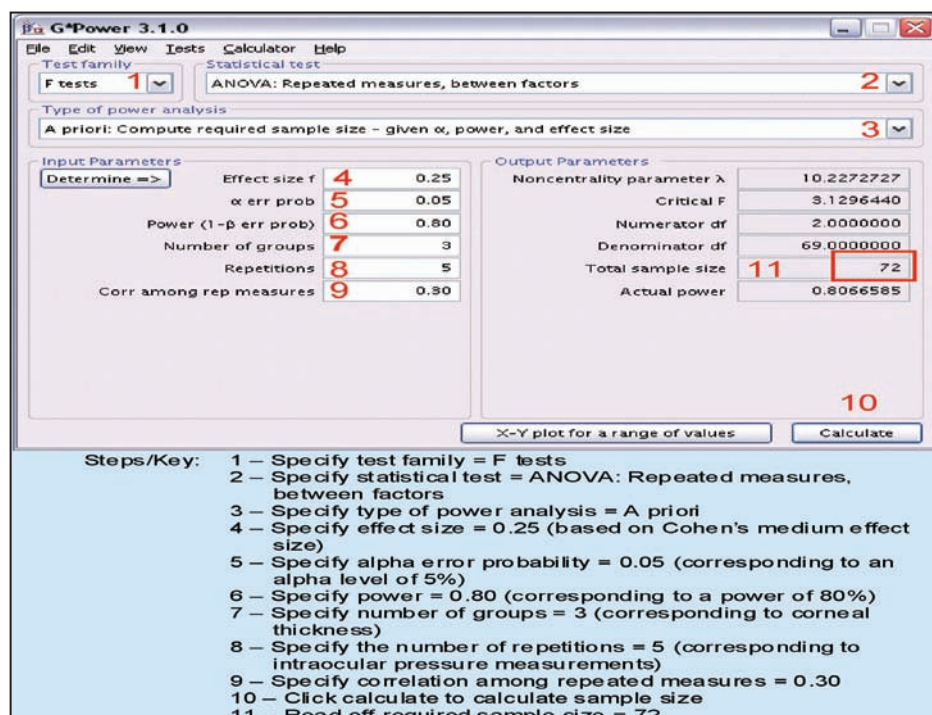


Figure 10

Computing sample size for between-subjects factor for a repeated measures ANOVA using GPower 3

two groups of measurements is also required. Let's suppose that a pilot study returned a correlation coefficient (Pearson's correlation coefficient, ρ , referred to in Table 2) of 0.30. GPower 3 shows that this results in an effect size of 1.39 and the required sample size would therefore be 5 (Figure 4).

Again, if the data were later found to violate the assumptions of parametric tests, a Wilcoxon signed-rank test would have to be used instead (the non-parametric equivalent of a paired t-test). GPower 3 shows that a sample size of 6 would be required instead (Figure 5).

Comparing many independent means

Consider an experiment designed to test if IOP varied with age. Suppose there were four age groups being considered (40-49 years, 50-59 years, 60-69 years and 70-79 years). This would be analysed with a one-way ANOVA (parametric test). How many subjects would be required for 80% power at the 5% alpha level? Firstly, the effect size needs to be determined. This is more challenging for F tests as the formulae to determine effect size requires a prior knowledge of the dispersion of the means of each group. If data from relevant previous studies are available, GPower 3 can use these means to compute the effect size by clicking the "determine" button near the effect size box. On the other hand, if relevant previous studies do not exist, the power analysis can be performed using Cohen's standard effect sizes (from Table 1).

In this example, if a small effect size is selected ($f = 0.10$), GPower 3 shows that the total sample size required would be 1096 (274 in each of the four age groups) to have 80% power at the 5% alpha level (Figure 6). This is a large sample size, which reduces to 180 (45 in each of the four age groups) if a medium effect size is selected ($f = 0.25$) or 76 (19 in each of the four age groups) if a large effect size is selected ($f = 0.40$).

Other ANOVA designs

There are many other study designs that can be used to compare the differences between more than two means. These include factorial and repeated measures ANOVAs.¹⁸ As study designs get more complicated, more assumptions are made about

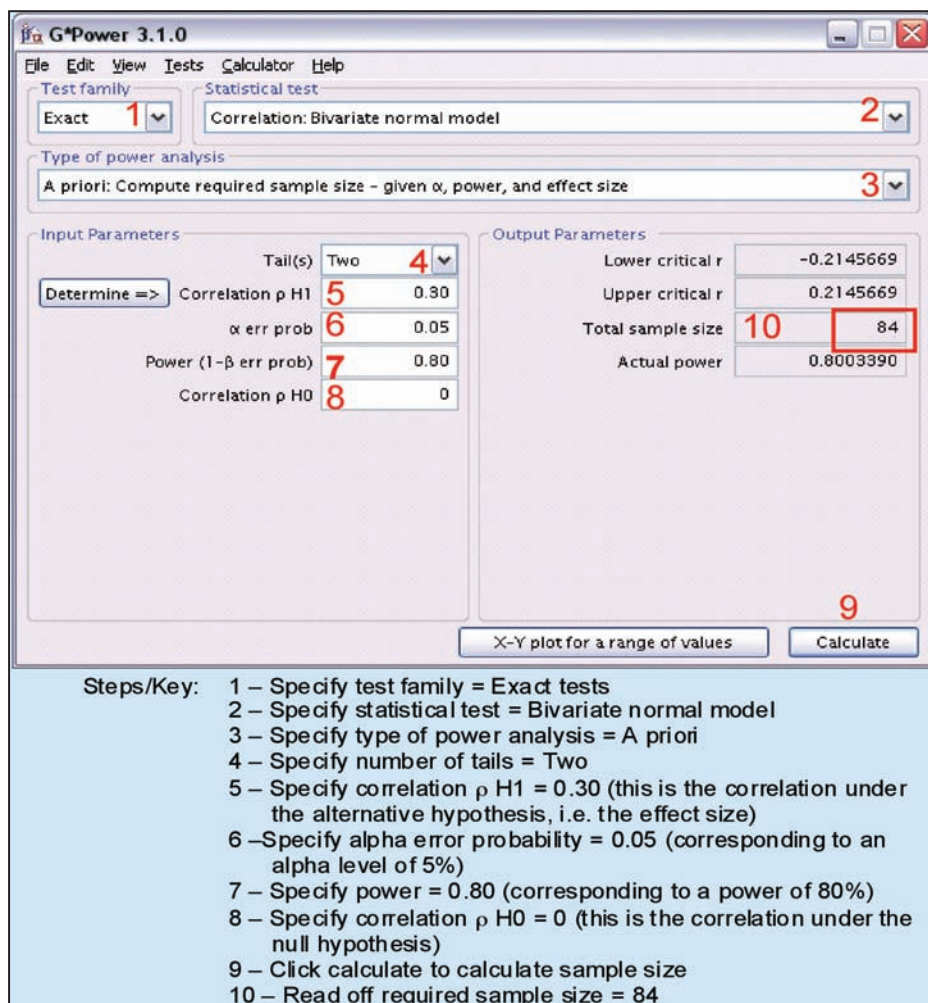


Figure 11

Computing sample size for correlations using GPower 3

the data in order to estimate the sample size required. This means that power analyses become less precise.¹⁹

Factorial ANOVA

A factorial ANOVA is used to test hypotheses about means when there are two or more independent factors in the design. It also reveals any possible interactive effects between these independent factors. A simple factorial design would be a 2 x 2 ANOVA, where there are two independent factors each with two levels. For example, consider a study designed to compare the amount of pupil dilation that results after tropicamide is administered to males and females with blue or brown irides. Gender is thus one independent factor with two levels (males and females) and iris colour is another independent factor also with two levels (blue and brown). How many subjects would be required for 80% power at the 5% alpha level? First the effect size needs

to be determined. For factorial ANOVAs GPower 3 determines effect sizes based on eta squared (see Table 2, equation 6). Let's assume that we are interested in a medium eta squared value, i.e. 0.06.

In GPower 3, power needs to be computed for each factor and each interaction in the study design individually. The final sample size is then based on the largest of the sample estimates arising from these separate analyses. GPower 3 shows that the sample size required to analyse the gender factor is 125 (Figure 7). This cannot be split equally between the four groups and so 128 would have to be recruited, i.e. 32 in each group. This sample size also applies to the iris colour factor as it has the same number of levels as the gender factor. In cases where one factor has more levels, this would require a larger sample size, and so the larger sample size should be used as the overall sample size recruited. The sample size required for the gender /

iris colour interaction can be computed in the same way. In this case, the numerator degrees of freedom (step 10 in Figure 7) is calculated as the degrees of freedom of the gender factor (i.e. 2 levels – 1 = 1) multiplied by the degrees of freedom of the iris colour factor (i.e. 2 levels – 1 = 1). The numerator degrees of freedom is therefore 1. This also results in a sample size of 125, which needs to be rounded up to 128 (32 in each of the four groups), as shown in Figure 8.

Repeated measures ANOVA

A repeated measures ANOVA is used to test hypotheses about means when there are two or more dependent factors in the design. These dependent factors are termed within-subject factors as the same subjects are used for each level of the variable. Independent factors can also be added to a repeated measures ANOVA design and are termed between-subject factors as different subjects are used for each level of the variable. A repeated measures ANOVA makes the assumption of sphericity. This means that (i) the variances of all the levels of the within-subjects factors are equal and (ii) the correlation among all repeated measures are equal. When this assumption is violated, a correction is required; this is the non-sphericity correction (ϵ).

Consider a study designed to test the repeatability of a new non-contact tonometer. Five IOP readings are taken for each subject. This is therefore a within-subjects factor as all five readings are from the same subject. The researcher is also interested in whether corneal thickness has any influence on the repeatability of the tonometer. The subjects are classified as having thin (<555 μ m), normal (556 – 587 μ m) or thick (>587 μ m) corneas. This is therefore a between-subjects factor as each level will consist of different subjects. The study design is a 3 x 5 repeated measures ANOVA as there are 3 levels of one factor (corneal thickness) and 5 levels of the other factor (IOP). How many subjects are required for 80% power at the 5% alpha level?

As for the factorial ANOVA, the sample size needs to be determined individually for each factor and each interaction. This study design also requires prior knowledge of the

correlation among repeated measures and a non-sphericity correction (ϵ). Both need to be based on data from previous studies or pilot studies. Let's assume the effect size f is medium (0.25), Pearson's correlation coefficient (ρ , see Table 2) among repeated measures is 0.30 and the non-sphericity correction is 1, i.e. all five groups of repeated measurements have equal variance and equal correlation among repeated measures. Figure 9 shows that the sample size required to analyse the within-subjects factor (IOP) is 30, i.e. 10 in each of the three groups for corneal thickness. Figure 10 shows that the sample size required to analyse the between-subjects factor (corneal thickness) is 72, i.e. 24 in each of the three groups for corneal thickness. The required sample size to analyse the corneal thickness/IOP interaction is calculated in the same way as Figure 9 but in step 2 the statistical test is replaced with ANOVA: Repeated measures, within-between interaction. This shows that a sample size of 36 is required to analyse the interaction effects, i.e. 12 in each of the three groups for corneal thickness. These calculations all generate three different sample sizes in this example. So how many subjects need to be recruited? In multi-factorial designs like this one, there are two approaches that can be adopted. Firstly, the researcher can compute sample sizes for all factors and interactions and then recruit the largest sample size generated. In this case this would be 72. However, this may not always be the best option as in some studies there will clearly be some complex interactions that may be of no interest to the researcher. Therefore, the researcher can specify which factor or interaction is the most interesting from a theoretical point of view and then only compute sample size for this factor. Post hoc power analysis can show what the resultant power would be for the remaining factors or interactions.

Correlations

Armstrong and Eperjesi⁶ described an example of a study designed to investigate the relationship

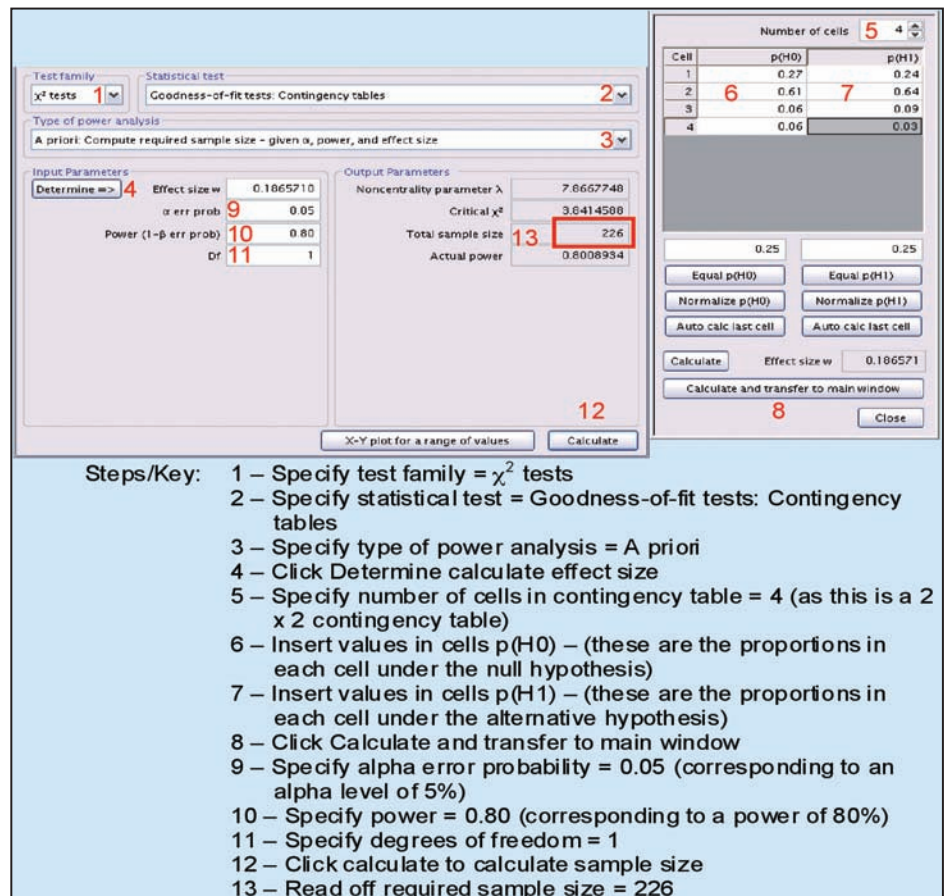


Figure 12

Computing sample size for a Chi squared test using GPower 3

between post-operative IOP and residual corneal thickness after laser refractive surgery. Pearson's correlation coefficient can be used to test this, but how many subjects are required for 80% power at the 5% alpha level? The effect size is taken simply as the magnitude of Pearson's correlation coefficient (ρ , see Table 2). For a medium effect size ($\rho = 0.30$), GPower 3 shows that 84 subjects would therefore be required (Figure 11).

Chi squared test

Consider a study designed to investigate the possible effects of smoking on age-related macular degeneration (AMD). A random sample of elderly people is drawn from the population and they are classified as smokers or non-smokers. Both of these groups are then examined to see whether there is any evidence of the presence of AMD. How big a sample size would be required for 80% power at the 5% alpha level? Firstly the effect size needs to be determined. This is based on the proportions in each cell of a 2 x 2 contingency table under the null and alternative hypothesis.

Current literature shows that 12% of the population over 60 smokes²⁰ and 33% of the elderly population have AMD.²¹ Armstrong and Eperjesi⁴ stated that although there are studies in the literature that suggest a possible connection between AMD and smoking, the results of an individual study are often inconclusive and generalisations of whether smoking is considered to be a "risk factor" for AMD are often based on combining together many studies. Armstrong²² did so and found that smokers are two to five times more likely to develop AMD. Based on this, if we hypothesise that smokers are three times more likely to develop AMD, the proportions for a 2 x 2 contingency table are shown in Table 4a for the alternative hypothesis. The null hypothesis states that there is no link between smoking and AMD and so the proportions in this case are shown in Table 4b.

Using this data, GPower 3 shows that a sample size of 226 (113 smokers aged over 60 years and 113 non-smokers aged over 60 years) would be required (Figure 12).

	Non-smokers	Smokers	Row totals
ARMD	0.24	0.09	0.33
No ARMD	0.64	0.03	0.67
Column totals	0.88	0.12	1.00

(a)

	Non-smokers	Smokers	Row totals
ARMD	0.27	0.06	0.33
No ARMD	0.61	0.06	0.67
Column totals	0.88	0.12	1.00

(b)

Table 4

Proportions in 2 x 2 chi square table under (a) the alternative hypothesis, and (b) the null hypothesis, for smoking and AMD study

A note of caution

Calculations of the type described in this article have their critics. Sample size estimation has been called a “game” of numbers²³ or “a guess masquerading as mathematics”.²⁴ The results of these mathematical analyses can be manipulated in a number of ways to suit the researcher. A large number of assumptions are made about the data and both the effect size and standard deviation have to be estimated by the researcher. There are no strict rules to help estimate these variables. Therefore, the size of these variables can easily be manipulated to produce any number the researcher desires. For example, by choosing a larger effect size or a smaller standard deviation a smaller sample size would be required for a given level of power. Norman and Streiner²⁵ have shown that only very small changes to these parameters are required in order to dramatically reduce the sample size to any desired value. They say, “the morale of the story is that a sample size calculation informs you whether you need 20 or 200 people to do the study. Anyone who takes it more literally than that, unless the data on which it is based are very good indeed, is suffering from delusion”. So, though sample size estimates and power analyses are currently in fashion, it is wise to treat them with a healthy dose of caution.

About the Authors

Bhavna Prajapati is an optometrist who works in practice and is also engaged in postgraduate research at Aston University. Mark Dunne and Richard Armstrong are lecturers on the optometry degree course at Aston University. They have also co-written the Research Methods module of the Ophthalmic Doctorate.

References

- 1 Batterham, A. M. and G. Atkinson, (2005), How big does my sample need to be? A primer on the murky world of sample size estimation, *Physical Therapy in Sport*, 6: 153-163
- 2 Armstrong, R. A. and F. Eperjesi, (2000), The use of data analysis methods in optometry: Basic methods, *Optometry Today*, 36-40
- 3 Armstrong, R. A. and F. Eperjesi, (2001), The use of data analysis methods in optometry: Comparing the difference between two groups, *Optometry Today*, 41: 34-37
- 4 Armstrong, R. A. and F. Eperjesi, (2002), Data analysis methods in optometry. Part 3 - The analysis of frequencies and proportions, *Optometry Today*, 42: 34-37
- 5 Armstrong, R. A. and F. Eperjesi, (2004), Data methods in optometry: Part 4: Introduction to analysis of variance, *Optometry Today*, 44: 33-36
- 6 Armstrong, R. A. and F. Eperjesi, (2005), Data methods in optometry. Part 5: Correlation, *Optometry Today*, 45: 34-37
- 7 Armstrong, R. A. and F. Eperjesi, (2006), Data methods in optometry: Part 6: Fitting a regression line to data, *Optometry Today*, 46: 48-51
- 8 Zodpey, S. P., (2004), Sample size and power analysis in medical research, *Indian J Dermatol Venereol Leprol*, 70: 123-8
- 9 Eng, J., (2003), Sample size estimation: how many individuals should be studied? *Radiology*, 227: 309-13
- 10 Araujo, P. and L. Froyland, (2007), Statistical power and analytical quantification, *Journal of Chromatography B*, 847: 305-308
- 11 Cohen, J., (1988), Statistical power analysis for the behavioral sciences, Lawrence Erlbaum Associates Inc.

- Publishers, New Jersey
- 12 Mumby, P. J., (2002), Statistical power of non-parametric tests: a quick guide for designing sampling strategies, *Mar Pollut Bull*, 44: 85-7
 - 13 Pitman, E. J. G., (1948), Lecture notes on nonparametric statistical inference, Columbia University
 - 14 Zodpey, S. P. and S. N. Ughade, (1999), Workshop manual: Workshop on sample size consideration in medical research, MCIAPSM, Nagpur
 - 15 Faul, F., E. Erdfelder, A. G. Lang and A. Buchner, (2007), G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences, *Behavior Research Methods*, 39: 175-191
 - 16 Gupta, D., (2005), Glaucoma diagnosis and management, Lippincott Williams & Wilkins, Philadelphia, PA
 - 17 Hsieh, Y. and F. Hu, (2007), The correlation of pupil size measure by Colvard Pupillometer and Orbscan II, *Journal of Refractive Surgery*, 23
 - 18 Armstrong, R. A., F. Eperjesi and B. Gilmartin, (2002), The application of analysis of variance (ANOVA) to different experimental designs in optometry, *Ophthalmic and Physiological Optics*, 22: 248-256
 - 19 Dattalo, P., (2008), Determining Sample Size: Balancing Power, Precision, and Practicality, Oxford University Press, New York
 - 20 ONS (2008) Office for National Statistics, General Household Survey 2006, UK
 - 21 Mukesh, B. N., P. N. Dimitrov, S. Leikin, J. J. Wang, P. Mitchell, C. McCarty and H. R. Taylor, (2004), Five year incidence of age-related maculopathy: the visual impairment project, *Ophthalmology*, 111: 1176-1182
 - 22 Armstrong, R. A., (2004), AMD and smoking: An update, *Optometry Today*, Dec 17: 44-46
 - 23 Pocock, S. J. (1996) *Advances in biometry IN ARMITAGE, P. & DAVID, H. A. (Eds.) Clinical trials: A statistician's perspective*. Wiley, Chichester
 - 24 Senn, S., (1997), Statistical issues in drug development, Wiley, Chichester
 - 25 Norman, G. R. and D. L. Streiner, (1994), *Biostatistics: The bare essentials*, Mosby-Year Book Inc, St Louis