Cell PRESS
Immunity
Commentary

t Testing the Immune System

Tracey J. Lamb,[1,4] Andrea L. Graham,[2] and Aviva Petrie[3,*]
[1]Division of Parasitology, National Institute for Medical Research, The Ridgeway, Mill Hill, London NW7 1AA, UK
[2]Institutes of Evolution, Immunology and Infection Research, School of Biological Sciences, King's Buildings, University of Edinburgh, West Mains Road, Edinburgh EH9 3JT, Scotland, UK
[3]Biostatistics Unit, University College London Eastman Dental Institute, 256 Grays Inn Road, London WC1X 8LD, UK
[4]Present address: School of Biological Sciences, University of Reading, Whiteknights, Reading RG6 6AJ, UK.
*Correspondence: a.petrie@eastman.ucl.ac.uk
DOI 10.1016/j.immuni.2008.02.003

Amid the flurry of grant writing and experimentation, statistical analysis sometimes gets less attention than it requires. Here, we describe fully the considerations that should go into the employment of the statistical two-sample t test.

The biological significance of immunological data is paramount in their interpretation. Nevertheless, immunological data are variable, and because statistical science aims to make sense of variability, statistical methods are not superfluous to immunology. The most informative interpretation of experimental data emerges from a combination of biological and statistical insight.

Proper statistical analysis of results can only be achieved if the researcher has some understanding of statistical theory and of the potential that it offers when summarizing data and drawing conclusions from them. Here, we concentrate on the frequently misused two-sample t test, using an experimental data set to illustrate the appropriate application of this test in immunological research.

Units of Analysis: What Is Being Analyzed in an Experiment?

Before undertaking any statistical analysis, it is essential to be clear about what is being analyzed. The "unit of analysis" (Altman and Bland, 1997), also called the experimental unit, is the smallest unit of replication that can be assigned at random to a "treatment" (see this and other key definitions in Box 1). Take, for example, an in vivo experiment designed to assess whether a particular experimental infection upregulates interleukin-10 (IL-10) production in CD4$^+$ T cells in mice (Box S1A Experiment 1 and Figure S1). Because each animal's spleen is processed separately, the unit of analysis is the individual mouse in which the measurement of interest is the CD4$^+$ T cell IL-10 Median Fluorescence Intensity (MFI).

Often in immunological experiments, however, it is not feasible to assess mice individually, and it is necessary to pool material from several animals in order to carry out the experiment. For example, in pooling CD4$^+$ T cells from five animals to coculture with primed and unprimed dendritic cells in the wells of a tissue-culture plate, the unit of analysis in this in vitro experiment (Box S1B) is the well and not the mouse. The result of a statistical hypothesis test in such an experiment refers to one pooled cell population. Conclusions cannot be drawn about the behavior of mouse T cells in general until repeated experiments with independently derived cell populations have been conducted. Whether the unit of analysis is the mouse or the well, the correct application of the two-sample t test to the data must provide conclusions that relate to the relevant unit of analysis.

The Two-Sample t Test and the p Value

Currently, the two-sample t test (also called the "independent samples" or "unpaired" t test) is one of the most commonly used statistical hypothesis tests in immunological research. Because it is often performed with computer software, only a basic explanation of the mechanics of the test is provided (Box S2), although further details can be obtained from standard statistical textbooks (e.g., Petrie and Watson, 2006). However, to apply the t test correctly, it is important for immunologists to understand the concepts and terminology underlying hypothesis testing.

The two-sample t test compares the means of two groups. Generally, it is not possible to study the whole population of observations, so a representative sample is used to make inferences about the population. More specifically, the sample mean is used as an estimate of the true mean in the population. In the two-sample t test, the population means are compared by use of the sample mean responses of the relevant units of analysis when each unit receives one of two treatments. The result is used to assess whether any apparent difference in means reflects a real difference or is due to random variation.

With regard to the example of whether infection induces IL-10 production in CD4$^+$ T cells (Box S1A Experiment 1 and Figure S1), the two-sample t test is conducted to test the null hypothesis (the hypothesis of "no effect") that the true group-mean MFI for IL-10 from CD4$^+$ T cells in mice with and without infection is equal (i.e., on average, there is no upregulation of IL-10 in CD4$^+$ T cells in response to infection). The results of the hypothesis test include a p value, the probability of obtaining the observed results or, if the null hypothesis is true, of obtaining more extreme results. If the p value is small, then there is a poor chance of getting the observed results (here, the observed difference in means) if the null hypothesis is true, so the null hypothesis is rejected and the result is said to be statistically significant. If the p value is large, then there is no evidence to reject the null hypothesis and the result is said to be not statistically significant.

The cut-off for the p value that determines significance is called the significance level (unfortunately, this term is frequently misappropriated by researchers, who incorrectly use it to describe the p value obtained from the test) or the alpha level. Its value, chosen at the design stage of the study, is usually 0.05. This

Box 1. Useful Statistical Definitions

- 95% confidence interval (CI) for the mean: loosely defined as the range of values within which the true population mean lies, with 95% certainty. Provided the sample size is greater than about 10, it is approximately equal to the sample mean ± 2× the standard error of the mean (SEM).
- Analysis of variance (ANOVA): a general term for analyses that compare the means of three or more groups of observations.
- Average: a summary measure of central tendency, such as the arithmetic mean (usually simply called the "mean;" equal to the sum of all the observations divided by the number of observations) or the median (which is the middle observation in the ordered set).
- Nonparametric (distribution-free) test: test that does not make any assumptions about the distribution of the data.
- Normal distribution: a symmetrical, theoretical, statistical distribution with many useful properties. The mean and median of a normal distribution are equal.
- Random allocation (also called randomization): the units in the sample are randomly (i.e., by use of a method based on chance) allocated to the different treatment groups.
- Random selection: each unit of analysis (e.g., mice or wells) is selected from the population by use of a method based on chance and therefore has an equal probability of being selected.
- Robust test: the chance of making a mistake (i.e., of incorrectly either rejecting or not rejecting the null hypothesis) is hardly affected when the test's assumptions are violated.
- Standard error of the mean (SEM): a measure of precision of an estimate equal to the standard deviation divided by the square root of the number of observations in the sample.
- Unit of analysis: the smallest unit of replication that can be randomly assigned a treatment.
- Variance: a measure of the variability or spread of a data set; it is equal to the square of the standard deviation (SD), which can be thought of as a sort of average of the deviation of every observation from the mean.

means that if $p < 0.05$, the null hypothesis is rejected and the conclusion is that the treatment means are different. In the $CD4^+$ T cell in vivo example (Box S1A Experiment 1), $p = 0.008$, so the null hypothesis is rejected; i.e., on average, infection does alter the production of IL-10 in $CD4^+$ T cells compared with the naive animals (the group means for IL-10-producing $CD4^+$ T cells are 82.4 MFI and 66.0 MFI, respectively).

It is important to recognize that lacking evidence to reject the null hypothesis is not the same as accepting the null hypothesis—i.e., "absence of evidence is not evidence of absence" (Altman and Bland, 1995). There may be a real difference between the treatment means but the sample size is too small to be able to detect it as statistically significant.

The Assumptions Underlying the Two-Sample t Test

If the assumptions underlying a statistical test are not satisfied, the p value may be incorrect and/or the test may fail to detect as statistically significant a true treatment effect. Although not all erroneous p values lead to incorrect conclusions, statistical methods applied inappropriately undoubtedly increase the chance of making a mistake. The assumptions underlying the two-sample t test (Petrie and Watson, 2006) are described next.

The sample data should be randomly selected from the population. If the units of analysis are selected randomly from the population, the sample should be representative of the population about which inferences are to be made. Unfortunately, it is rarely possible in immunological studies to use random selection because there is a tendency to use the units that are available (e.g., to choose mice from the cages in which they were bred or supplied rather than from the larger population). However, random allocation of the units to the different treatment groups may be used as a substitute for random selection because the differences between treatment groups are akin to the differences between random samples. The use of random allocation avoids the bias that might arise if the different treatment groups are not balanced with regard to those factors likely to influence response. For example, docile mice may have less testosterone than aggressive mice, and testosterone has previously been reported to be immunosuppressive in some systems. Thus, if there is a preponderance of docile mice in one of the two treatment groups (say, if they were the first mice picked from the stock cage and were all allocated to the same experimental cage for treatment 1), the conclusions drawn from the two-sample t test comparing the mean $CD4^+$ T cell IL-10 MFI under two treatment regimes will be biased; the difference in the mean responses in the two treatment groups might be due to differences in amounts of testosterone rather than to differences between experimental treatments. To avoid this problem, the mice should be randomly allocated from the stock cage into the experimental cages to ensure that the docile mice are evenly distributed in the two treatment groups. In terms of in vitro work involving tissue culture plates, complete randomization on a tissue culture plate is problematic and can lead to physical plating errors. As a step toward achieving randomization, researchers could alter the order of the treatments on the plate between experiments.

The two groups of data must also be independent. Independence of the observations *between* groups should not be confused with independence of the observations *within* groups, the latter also being a requirement of the two-sample t test. In particular, there are theoretical issues regarding the independence of animals housed in the same cage (Festing and Altman, 2002). Correction of this type of nonindependence may be difficult to achieve in immunological experiments. Another issue relates to the independence of cell populations in wells. It is common for immunologists to perform hypothesis tests using the same cell population in replicate wells in vitro rather than to replicate cell preparations. Replicate wells in vitro are not independent, because the cell population being tested is the same preparation in all wells. Consider, as an example, the generation of IL-10-producing CD4$^+$ T cells by pulsed DCs in vitro (Box S1B). The desired inference from the experiment (CD4$^+$ T cells from individual mice) and the units of analysis (wells containing CD4$^+$ T cells pooled from different mice) are incongruous. Because the cells in each well are from the same pooled CD4$^+$ T cell population, the only way to draw conclusions about CD4$^+$ T cell populations in general is to a perform statistical analysis with data from separately derived cell populations (i.e., by repeating the experiments).

If there is dependence, either between or within groups, it is not appropriate to perform the two-sample t test. Different statistical tests, such as the paired t test or various forms of analysis of variance, should be considered instead (http://www.isogenic.info/index.html; Cox and Reid, 2000; Festing and Altman, 2002; Grafen and Hails, 2002; Howell, 1999; Mead, 1988).

It is also important that the data in each group are normally distributed in the population. If the results of a two-sample t test are to be valid, the data in each group should come from a population of values that approximately follows the normal distribution. A formal statistical hypothesis test for normality, such as the Anderson-Darling, Shapiro-Wilk, or Kolmogorov-Smirnov tests, is unnecessary. Instead, a visual impression of the symmetry of a histogram (Figure S1B) or box-and-whisker plot (Figure S1C) or a check that the mean and median are approximately equal is usually sufficient. By these criteria, the box plots of MFI from CD4$^+$ T cells ex vivo (Figure S1C) suggest that the data in each group are approximately normally distributed. In contrast, the distribution of the fluorescence intensity of individual IL-10-producing CD4+ T cells from a single infected animal (the histogram in Figure S1B) is clearly not Normal; rather, it is right-skewed, with a long tail to the right. The mean fluorescence intensity of 90.8 is not a satisfactory summary measure of IL-10 fluorescence intensity in the CD4$^+$ T cells in this animal because the mean is inflated by the values in the tail; the median (fluoresence intensity of 59.2) is more appropriate.

The final assumption that must be considered is whether the variability is the same in the two groups being compared (i.e., whether there is homogeneity of variance). Ideally, the variances of the observations in the two groups should be equal (i.e., the data should exhibit homogeneity of variance) when a two-sample t test is performed. However, it might be possible, depending on the software, to perform a modified test that does not rely on the assumption of equal variances.

Equality of variance is usually assessed by a test of the null hypothesis that the two population variances are equal by use of, for example, Bartlett's (Armitage et al., 2002) or Levene's test (Levene, 1960). A nonsignificant result (usually if $p > 0.05$) indicates that there is no evidence to reject the null hypothesis. In the first in vivo example (Box S1A Experiment 1 and Figure S1C), the variances of the MFI in the groups of naive and infected mice are estimated as 45.1 MFI and 255.6 MFI, respectively. Levene's test gives $p = 0.07$, which suggests that there is insufficient evidence to reject the null hypothesis that the variances of IL-10 MFI from CD4$^+$ T cells in the two groups are equal. Hence the two assumptions underlying the two-sample t test, those of normality and constant variance, are satisfied in this example.

Violations of the Assumptions Underlying the Two-Sample t Test

What happens when assumptions of normality and homogeneity of variance are violated? The two-sample t test is fairly robust to violations of the normality and constant variance assumptions (Bland, 2000). In particular, heterogeneity of variance is not so important if the data are normally distributed (Bland, 2000). Nevertheless, when the variances are very different (which could lead to a false-positive finding), when the sample sizes in the two groups are very disparate, and/or when the sample sizes are so small (frequently the case in immunological studies) that is impossible to check the assumptions, it might be better to consider alternatives to the two-sample t test.

One approach is to transform each data point mathematically (Bland and Altman, 1996; Grafen and Hails, 2002) and perform the two-sample t test on the transformed data, checking that the assumptions of this new analysis are satisfied. For example, if the logarithmic transformation (to any base, but typically to base *e* or 10) is taken of right-skewed observations (when the distribution has a long tail to the right; Figure S1B), the distribution of the transformed data will usually be approximately normal. The logarithmic transformation and other transformations can also correct for heterogeneity of variance (Petrie and Watson, 2006).

Another approach is to use a nonparametric or distribution-free test. A nonparametric or distribution-free analysis does not make any assumptions about the distribution of the data. It replaces the observed data by their ranks in the ordered set and is therefore not influenced by the few extremely large (or small) values in non-normally distributed data. A nonparametric alternative to the two-sample t test is the Mann-Whitney *U* test (equivalent to the Wilcoxon rank sum test). However, if all assumptions of the two-sample t test are met, it is better to use the t test because it has a greater ability to detect as significant a real difference between groups.

One scenario where the independence assumption could be violated is when the data in the two groups are paired rather than independent (e.g., pretreatment versus posttreatment samples). In this case a paired t test is advocated instead of the two-sample t test, provided the differences between the paired observations are approximately normally distributed. If normality is of concern, a paired t test may be performed on suitably transformed data or an appropriate nonparametric test, such as the Wilcoxon Signed Rank test or the Sign test, may be used.

All of the above considerations are summarized in the flow-chart of Figure 1, which may be used as a decision tool for choosing the most appropriate test for the comparison of two groups. A lack of consideration for the assumptions of the two-sample t test can lead to incorrect conclusions. Consider a second in vivo experiment in which the investigator wants to assess whether IL-10 is upregulated (Box S1A Experiment 2). The data generated are skewed to the right in both infected and uninfected groups of mice, and there is heterogeneity of variance (Levene's test gives $p = 0.02$). A two-sample t test performed incorrectly on these data gives $p = 0.06$, with insufficient evidence to show that, on average, IL-10 is altered. However, an appropriate

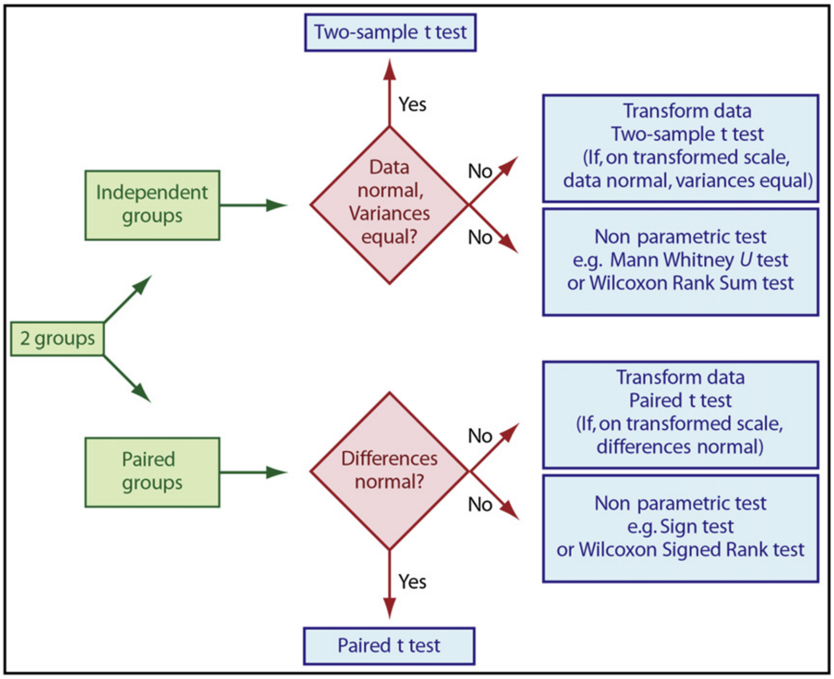

Figure 1. Flow Chart for Choosing the Appropriate Test to Compare Two Groups of Observations
Various forms of analysis of variance should be used if there are more than two groups.

analysis, such as the nonparametric Mann-Whitney *U* test performed on the raw data or the two-sample t test performed on the logarithmically transformed data (correcting for lack of normality and heterogeneous variances), gives p = 0.04 in each case, indicating that, on average, CD4$^+$ T cells do upregulate IL-10 in response to this infection. This example underlines the importance of correctly applying statistics to the analysis of immunological data.

Sample-Size Estimation and Statistical Power in the Two-Sample t Test

Sometimes, a difference in treatment means appears biologically important but the results lack statistical significance, perhaps because of an inadequate sample size and consequent low statistical power. The power of a test, which increases with larger samples, is the probability of detecting a real difference as statistically significant. A power calculation to determine the optimal sample size at the design stage of a study, before carrying out an experiment, is essential. Such a calculation is also required for most grant applications. An optimal sample size is one that is adequate to detect as statistically significant a treatment effect (e.g., a difference in means) of a given magnitude but that is not so large that it is wasteful of resources. A generally accepted view is that a test should have at least 80% power. It is not sensible to embark on a study that is believed at the outset to have a lesser chance (say, 50%) of finding a real effect statistically significant.

Calculation of the optimal sample size depends on the proposed hypothesis test (e.g., the two-sample t test) and a specification of the power, significance level, minimum treatment effect that is considered important, and variation in the data (see Box S3 and calculations for the IL-10 example in Box S4). Several different techniques can be used to determine the optimal sample size, including computer programs such as nQuery Advisor: Statistical Solutions (www.statsol.ie/nquery/nquery.htm), tables (Machin et al., 1997), relatively complex formulae (Kirkwood et al., 2003), and a diagram (Altman, 1982). In addition, Lehr (Lehr, 1992) devised simple formulae (provided in Box S3) specifically for a 5% significance level and an 80% or 90% power. The optimal sample size should be justified by providing a power statement (Box S4) that specifies the values of all the factors that are incorporated into the sample-size calculations. Then reviewers and readers of the article can assess whether the sample size used in the study is sensible.

Reporting the Results of a Two-Sample t Test

Details of how to report the results of a two-sample t test are given in Box 2 and can also be found in (Lang and Secic, 2006) and in the CONSORT statement guidelines (http://www.consort-statement.org). For the example shown in Box S1A Experiment 1 and in Figure S1, in addition to the power statement, the reporting of the results should state that there is evidence to show that the CD4$^+$ T cells from infected animals have a significantly altered IL-10

Box 2. Reporting Results from t Tests

In a complete presentation of the results of the two-sample t test the following should be included:

- The statistical test used should be named explicitly (i.e., “the two-sample t test,” not simply “a t test”).
- An indication should be given that the underlying assumptions of the test have been validated (by use of a named transformation, if necessary).
- The sample size should be justified by a power calculation and accompanied by a power statement.
- The exact p value should be given, rather than an interval estimate of it or an asterisk (e.g., * to indicate $0.01 < P < 0.05$).
- The estimated mean in each group, with its associated confidence interval, should be provided.
- An estimate of the difference in means, indicating the magnitude of the treatment effect, should be given with its associated confidence interval.

(mean = 82.4 MFI, 95% CI 61.2 to 70.8 MFI) compared with naive animals (mean = 66.0 MFI, 95% CI 71.0 to 93.9 MFI): estimated difference in means is 16.4 MFI, 95% CI 4.9 to 27.9 MFI, p = 0.008.

Multiple Testing and the Analysis of Variance

The two-sample t test is used to compare the means of two groups. However, when it is of interest to compare the means in three treatment groups, A, B and C (e.g., A = infected, B = sham-injected, and C = vaccinated), it is inappropriate to perform all pairwise two-sample t tests (i.e., A versus B, A versus C, and B versus C). This is because as more tests are performed, it is more likely that a statistically significant result will occur on the basis of chance alone (Bender and Lange, 2001; Grafen and Hails, 2002).

The correct procedure in these circumstances is to start by performing a one-way analysis of variance (ANOVA), a global test of the null hypothesis that all (three) group means are equal. If the result of this ANOVA is not significant (typically if $p > 0.05$), no further testing is required. However, if the one-way ANOVA produces a significant result with $p < 0.05$, this implies that at least two of the group means are different, and it is necessary to find out where any differences lie. This is achieved by performing post hoc tests (such as those attributed to Bonferroni, Scheffé, Dunnett, and Tukey) that compare the means of all relevant pairs of groups, but each test adjusts the p value to take into account the multiple comparisons and thereby avoids a spuriously significant result.

There are many different forms of ANOVA, the simplest being the one-way ANOVA, which can be thought of as an extension to the two-sample t test when more than two group means are to be compared. More complicated forms (Edwards, 1993; Lindman, 1992; Weber and Skillings, 2000) should be employed when other factors need to be taken into consideration, such as the cages in which experimental animals are housed, the wells in which cell preparations are placed, and the replications of an experiment.

Conclusion

Failure to take stock of the statistical as well as the biological significance of data can ultimately be a waste of time and money and, where animals or patients are involved, can also raise ethical issues (Festing and Altman, 2002). At present, however, errors in violation of both statistical science and editorial policy appear to be common in the primary immunological literature. For example, many articles contain quantitative data from which conclusions are drawn about treatment effects in the absence of inferential statistical analysis, and others include significant p values without any information about the units of analysis, the sample sizes, the tests used, or their validity (Olsen, 2003). Unfortunately, these errors can lead to incorrect conclusions, especially for p values very close to 0.05 (i.e., of marginal statistical significance). There appears to be much room for improvement of statistical practice in immunology.

Giving proper consideration to the use of the two-sample t test, as outlined in this article, will improve the evaluation of treatment effects in immunological data and reduce the chances of drawing erroneous conclusions. Reviewers of manuscripts also need to be statistically informed, in order to judge whether the amount of variability in experimental data is acceptable (Altman, 1998) and the use of the two-sample t test is justified. Indeed, authors, reviewers, and editors alike must take steps to improve the rigor with which scientists t test the immune system.

SUPPLEMENTAL DATA

One figure and four boxes can be found online with this article at http://www.immunity.com/cgi/content/full/28/3/288/DC1.

ACKNOWLEDGMENTS

We thank J. Langhorne, A. Read, A.-M. Sponaas, A. Gelfand, A. Boonstra, G. Kassiotis, D.B. Guiliano, and S. Kock for insightful discussions and/or critical appraisal of early drafts of the manuscript. Finally, we thank the British Medical Research Council and the Biotechnology and Biological Sciences Research Council for funding. Data were provided by T.L., and the article was written by all three coauthors.

REFERENCES

Altman, D.G. (1982). How large a sample size? In Statistics in Practice: Articles Published in the British Medical Journal, S.M. Gore and D.G. Altman, eds. (London: British Medical Association).

Altman, D.G. (1998). Statistical reviewing for medical journals. Stat. Med. *17*, 2661–2674.

Altman, D.G., and Bland, J.M. (1995). Absence of evidence is not evidence of absence. BMJ *311*, 485.

Altman, D.G., and Bland, J.M. (1997). Statistics notes. Units of analysis. BMJ *314*, 1874.

Armitage, P., Berry, G., and Matthews, J.N.S. (2002). Statistical Methods in Medical Research, Fourth Edition (Oxford: Blackwell Science).

Bender, R., and Lange, S. (2001). Adjusting for multiple testing–when and how? J. Clin. Epidemiol. *54*, 343–349.

Bland, J.M., and Altman, D.G. (1996). Transforming data. BMJ *312*, 770.

Bland, M. (2000). An Introduction to Medical Statistics, Third Edition (Oxford: Oxford University Press).

Cox, D.R., and Reid, N. (2000). The theory of the design of experiments (Boca Raton, FL: Chapman & Hall/CRC).

Edwards, L.K. (1993). Applied analysis of variance in behavioral science (New York: M. Dekker).

Festing, M.F., and Altman, D.G. (2002). Guidelines for the Design and Statistical Analysis of Experiments Using Laboratory Animals. ILAR J. *43*, 244–258.

Grafen, A., and Hails, R. (2002). Modern Statistics for the Life Sciences, First Edition (New York: Oxford University Press Inc.).

Howell, D.C. (1999). Fundamental Statistics for the Behavioral Sciences, Fourth Edition (Pacific Grove, CA: Brooks/Cole Pub. Co).

Kirkwood, B.R., Sterne, J.A.C., and Kirkwood, B.R. (2003). Essential Medical Statistics, Second Edition (Malden, MA: Blackwell Science).

Lang, T.A., and Secic, M. (2006). How to Report Statistics in Medicine: Annotated Guidelines for Authors, Editors, and Reviewers, Second Edition (Philadelphia, PA: American College of Physicians).

Lehr, R. (1992). Sixteen S-squared over D-squared: A relation for crude sample size estimates. Stat. Med. *11*, 1099–1102.

Levene, H. (1960). Robust tests for equality of variances. In Contributions to Probability and Statistics. Essays in honor of H. Hotelling, I. Olkin, S.G. Ghurye, W. Hoeffding, W.G. Madow, and H.B. Mann, eds. (Menlo Park, CA: Stanford University Press), pp. 278–292.

Lindman, H.R. (1992). Analysis of Variance in Experimental Designs (New York: Springer-Verlag).

Machin, D., Campbell, M.J., Fayers, P.M., and Pinol, A.P.Y. (1997). Sample Size Tables for Clinical Studies, Second Edition (Oxford: Blackwell Science).

Mead, R. (1988). The Design of Experiments: Statistical Principles for Practical Applications (Cambridge: Cambridge University Press).

Olsen, C.H. (2003). Review of the use of statistics in infection and immunity. Infect. Immun. *71*, 6689–6692.

Petrie, A., and Watson, P.F. (2006). Statistics for Veterinary and Animal Science, 2nd ed. (Oxford: Blackwell).

Weber, D., and Skillings, J.H. (2000). A First Course in the Design of Experiments: A Linear Models Approach (Boca Raton, FL; London: CRC Press).