Statistics for the reader: what to ask before believing the results

Peter T. Choi MD FRCPC

IN this refresher course, I will review some basic questions one should ask regarding a study's use of statistics (Table I). Other aspects of research methodology relevant to a study's validity and generalizability will not be discussed. The reader is referred to other texts for further details on those aspects of research design and interpretation.[1–3]

TABLE I Basic questions to ask regarding a study's data analysis

1. Was the study's sample size sufficient?
   a. What was the research question?
   b. Was the sample size based on the primary outcome of interest?

2. Were the statistics appropriate?
   a. Was the measured variable expected to have a normal distribution?
   b. What were the measures of centrality and spread?
   c. What inferential tests were used?

3. How were common issues handled when they occurred?
   a. Was there adjustment of the *P*-value when multiple comparisons were made?
   b. Were dropouts, withdrawals, and outliers handled appropriately?
   c. When no difference was found between groups, was there a power analysis, description of the smallest detectable difference based on the study's sample size, or calculation of the required sample based on the observed frequencies?
   d. When no event was observed, was there a description of the maximum possible frequency based on the study's sample size?

Is the study big enough?

The credibility of a study's results depends on the extent of bias and random error that exists in the study. Bias leads to systematic deviations from the truth; methodological factors that affect internal validity influence the magnitude of bias. Random error relates to chance; sample size influences the magnitude of random error.

The sample size is influenced by four factors. An adequate sample size is one that can measure the anticipated frequency (for a study with a single cohort) or detect a realistic and clinically important difference (for a study comparing two or more cohorts) for the primary outcome of interest with minimal probabilities of positive or negative results being due to chance (i.e., false positive and false negative results). To determine the adequacy of the sample size, the reader must be able to identify the study's research question and the probabilities of false positive and false negative results on which the sample size calculation was made.

If a study has been written clearly, the reader should be able to identify the following elements of a research question: the population of interest, the intervention(s) or risk factor(s) depending on the nature of the study, the control(s) if the study compares two or more cohorts, and the outcome(s) of interest. The interventions and controls enable the reader to determine the appropriateness of the anticipated frequency or the difference to be detected. The outcomes enable the reader to determine whether the sample size is based on the primary outcome (ideally), on a surrogate for the primary outcome (if the primary outcome is difficult to measure or exceedingly rare), or on other outcomes (less appropriately).

The probabilities of false positive and false negative results are usually set by convention. The probability of a false positive result is the alpha level, which is usually set at 5%. This value is based on the fact that two standard deviations encompass 95% of the measurements of any variable with a normal ("bell curve") distribution. On rare occasions, alpha may be set at 1%. (Three standard deviations encompass 99% of the measurements in a normal distribution). The probability of a false negative result is the beta level.

From the Department of Anesthesia and the Vancouver Coastal Health Research Institute, University of British Columbia, Vancouver, British Columbia, Canada.

Address correspondence to: Dr. Peter Choi, Department of Anesthesia, University of British Columbia, Room 3200, Third Floor, JP Pavilion, 910 West Tenth Avenue, Vancouver, British Columbia V5Z 4E3, Canada. Phone: 604-875-4111; Fax: 604-875-5344; E-mail: pchoi@vanhosp.bc.ca

TABLE II Common statistical tests

Purpose	*Parametric test*	*Nonparametric test*	*Example*
Comparison of two independent sets of non-continuous data	Chi-square test, logistic regression	Fisher's exact test	Comparison of laryngoscopic grade obtained from two intubation techniques
Comparison of > two independent sets of non-continuous data	Chi-square test, logistic regression	Kruskal-Wallis test	As above, but with three intubation techniques
Comparison of two independent sets of continuous data	Unpaired t test	Mann-Whitney U test	Comparison of extubation times between two anesthetic regimens
Comparison of > two independent sets of continous data	ANOVA	Kruskal-Wallis test	As above, but with three anesthetic regimens
Comparison of two paired sets of continous data	Paired t test	Wilcoxon matched paired test	Comparison of pain intensity before and after an analgesic
Comparison of > two repeated measurements	Repeated measures ANOVA		Comparison of pain intensity at hourly intervals
Determine association between sets of non-continuous data	Chi-square, logistic regression		Determine relationship between gender and postoperative nausea
Determine association between sets ofcontinuous data	Pearson's rank correlation	Spearman's rank correlation	Determine relationship between preoperative hemoglobin and amount of transfusion

Typically, one refers to the power (1 - beta), which is usually set at 80% or 90%, instead of beta.

Most readers will be content to believe the sample size if the above factors have been addressed adequately. For those individuals who wish to verify the calculations, sample size formulae can be found in general biostatistical textbooks.[4] For studies involving two or more cohorts, PS version 2.1.31 (Vanderbilt University Medical Center, Nashville, USA),[5] a freeware sample size and power calculator, may be downloaded from www.mc.vanderbilt.edu/prevmed/ps/.

Are the statistics appropriate?

Non-statisticians often complain that no two statisticians will use the same statistical test when posed with a statistical question. Although this view is highly exaggerated, there is often more than one appropriate statistical test or method for data analysis. The key points to consider when choosing a statistical test include the purpose of the analysis (e.g., to make comparisons, to assess relationships, or to test for interactions), the type of data (e.g., categorical or continuous), the study design, and the number of cohorts.[6] The details of choosing a statistical test are beyond the scope of this refresher course. Some common statistical tests are described in Table II; readers are referred to other texts for additional details.[3,6,7] Although a reader may not have the expertise to determine the appropriateness of a specific analysis plan, most of the statistical errors in anesthesia studies can be recognized (and avoided) without extensive statistical knowledge.[8–10]

In general, the most common statistical errors relate to a failure to recognize the data's distribution, which affects the choices of descriptive statistics (to describe the data's centrality and spread) and inferential statistics (to compare two or more sets of data). Data can have a binomial distribution, a normal ("bell curve") distribution, a skewed (asymmetric) distribution, or other less common distributions (such as a bimodal one). I will focus on normal and skewed distributions, which tend to be more common.

The normal distribution is commonly seen for many biological phenomena (e.g., systolic blood pressure in a random sample). The common statistics of centrality, mean (arithmetic average), median (middle observation), and mode (most frequent observation), will be identical and located at the peak of the bell curve. In contrast, the values of those three statistics will differ in skewed distributions. Skewed distributions can be left-

skewed (also called negatively skewed), with the tail to the left of the peak (e.g., age distribution of individuals with perioperative myocardial infarction since children suffer infrequently from this problem), or right-skewed (also called positively skewed), with the tail to the right of the peak (e.g., postoperative hospital length of stay for a low-risk surgical procedure). In skewed distributions, the mode will describe the value at the peak of the curve, the mean will be the statistic of centrality that has a value closest to the tail, and the median will lie between the mode and the mean. To remember this easily, note that mean, median, and mode are in alphabetical order going from the tail to the peak. For a skewed distribution, the median is the most appropriate statistic of centrality.

Statistics that describe the spread of the data include range (difference between largest and smallest observation), interquartile range (central 50% of a distribution), standard deviation (SD, a summary measure of the differences of each observation from the mean of all observations), and variance (square of the SD). The standard deviation and the interquartile range are usually the most appropriate statistics to describe the spread in normal and skewed distributions respectively.

Inferential tests can be divided into parametric and non-parametric tests. Parametric tests make a number of assumptions about the data, including an assumption the data is distributed in a particular distribution - usually the normal distribution. Non-parametric tests do not make such assumptions.[3] Traditionally, parametric tests were believed to be more powerful in their ability to detect a statistical difference compared to non-parametric tests; however, the difference in power between the two types of tests may not be significant.[11,12] When the data appear to be distributed in a non-normal fashion, the reader should determine whether the investigators used non-parametric tests, which are appropriate, or parametric tests. A common error is the use of parametric statistics with data that violate the assumptions of those tests. The risk of a false positive result will increase, especially if the distribution of data deviates greatly from a normal distribution. If parametric tests are to be employed appropriately, non-normally distributed data should be mathematically transformed (by using the reciprocal, square root, or logarithm of the variable) to yield a normal distribution.

How were common issues addressed when they occurred?

A reader should expect to see a description of the methods used to address some common issues that occur in a study. Issues relating to multiple comparisons and dropouts / outliers should be addressed before the study and are described in the study's methods section; issues relating to negative results and zero events are usually addressed after the study is completed and may be described either in the study's methods or discussion sections.

Issues that should have been addressed before the study started

When a *P*-value of < 0.05 demonstrates a statistically significant difference, the chance of a false positive result is one in 20 for a single comparison. If multiple comparisons are made amongst variables that may be related to each other (e.g., comparisons of multiple demographic factors between two groups, comparisons of an outcome over multiple time points between two groups, or pairwise comparisons of more than two groups), each comparison has a one in 20 chance of resulting in a false positive result. Without adjustment of the *P*-value, one should not be surprised to see a "positive" result if 20 comparisons were made. In situations of multiple comparisons, adjustment of the *P*-value is needed to ensure that the chance of a false positive result is one in 20 for all comparisons. There are a number of methods to adjust for multiple comparisons. One common and conservative method is the Bonferroni correction, in which the *P*-value needed to demonstrate a statistically significant difference in one comparison is $< 0.05/n$, where n is the number of comparisons to be made.

During a study, subjects may drop out or withdraw before completing the study. Whether the total number of subjects should include or exclude dropouts/withdrawals will depend on the nature of the research question, the intervention, and the outcome of the study. For example, if the question is "What is the postoperative pain intensity of patients receiving analgesia from a properly sited epidural catheter?" one may be justified in excluding subjects with dislodged epidurals or patchy sensory blockade. However, if the question is "What is the postoperative pain intensity of patients receiving epidural analgesia?" one should include all subjects with epidural catheters. In situations when the decision to include or exclude dropouts/withdrawals is debatable, including all subjects in the analysis (i.e. intention-to-treat analysis) is more conservative. One may perform post-hoc sensitivity analyses to determine whether exclusion of dropouts affects the conclusion or not.

Similarly, outliers or extreme values can occur due to extremes of the norm (e.g., height in a basketball player), idiosyncracies in a subject (e.g., genetic varia-

tions of a metabolic pathway), or to errors in the collection or manipulation of the data (e.g., transcription errors). The first two should be included in the data analysis; the third should not.

Issues that should have been addressed at the end of the study

A study with a presumably adequate sample size may still find no difference (a "negative result") between groups. Failure to detect a difference is not the same as the absence of a difference (equivalence). If the sample size was calculated based on a formula for equivalence, then one may conclude that the compared groups were statistically equivalent with respect to the outcome on which the sample size was based. If the anticipated difference in the frequency of the outcome between the compared groups was so small that any lesser difference would not be clinically significant, one may conclude that the groups are sufficiently similar, clinically, with respect to the outcome on which the sample size was based. If the negative result is not due to either of the two situations above, the authors should calculate the power of the study to detect the anticipated difference based on the observed frequency of the outcome in the control group (power analysis), describe the smallest detectable difference based on the study's sample size and the observed frequency of the outcome in the control group, or calculate the number of additional subjects required to conclude that the observed difference is statistically significant. For readers who wish to determine whether a negative study has an adequate sample size to detect a clinically important difference, sample size nomograms[13] and freeware power calculators[5] are available.

Another issue is the null result or zero event. This is especially common with measurements of rare events. As with negative results, absence of an event is not the same as a zero probability of its occurrence. The authors should provide an estimate of the maximum frequency of the outcome (i.e., the upper 95% confidence limit of $0/n$). For readers who wish to calculate the frequency, the formula is:

$$\text{maximum frequency} = 1 - 0.05^{1/n}$$

where n is the number of subjects.[14,15] For sample sizes with more than 30 subjects, the equation can be simplified to:

$$\text{maximum frequency} = 3/n$$

where n is the number of subjects.[14,15]

Summary

Along with issues related to study design, errors in the data analysis plan can threaten the validity of results. Readers (and authors) should check for common statistical errors that may bias results or invalidate conclusions.

Acknowledgements

I thank Boris Sobolev, PhD, and Henry Sung, BSc(Math) MD, for their comments of earlier drafts of this manuscript and all of my former and current residents, graduate students, and fellows, whose questions on research design and statistics have improved my own understanding of this field.

References

1 *Guyatt G, Rennie D.* Users' Guides to the Medical Literature. Chicago: AMA Press; 2002.
2 *Fletcher RH, Fletcher SW, Wagner EH.* Clinical Epidemiology: The Essentials, 3rd ed. Baltimore: Williams & Wilkins; 1996.
3 *Greenhalgh T.* How to Read a Paper: the Basics of Evidence Based Medicine, 2nd ed. London: BMJ; 2000.
4 Appendix A. Calculating the required sample size: formulae. *In*: Pereira-Maxwell F (Ed.). A-Z of Medical Statistics. A Companion for Critical Appraisal. London: Oxford University Press; 1998: 87–8.
5 *Dupont WD, Plummer WD.* PS: Power and sample size. 16 Oct 2003 [cited 5 Jan 2005]. Available from URL; http://www.mc.vanderbilt.edu/prevmed/ps/.
6 Appendix B. Choosing the appropriate statistical test. *In*: Pereira-Maxwell F (Ed.). A-Z of Medical Statistics. A Companion for Critical Appraisal. London: Oxford University Press; 1998: 89–91.
7 *Swinscow TD, Campbell MJ.* Statistics at Square One, 10th ed. London: BMJ Books; 2002.
8 *Avram MJ, Shanks CA, Dykes MH, Ronai AK, Stiers WM.* Statistical methods in anesthesia articles: An evaluation of two American journals during two six-month periods. Anesth Analg 1985; 64: 607–11.
9 *Goodman NW, Hughes AO.* Statistical awareness of research workers in British anaesthesia. Br J Anaesth 1992; 68: 321–4.
10 *Goodman NW, Powell CG.* Could do better: statistics in anaesthesia research. Br J Anaesth 1998; 80: 712–4.
11 *Dexter F.* Analysis of statistical tests to compare doses of analgesics among groups. Anesthesiology 1994; 81:610–5.
12 *Delucchi KL, Bostrom AG.* Small sample longitudinal clinical trials with missing data: a comparison of analytic methods. Psychol Methods 1999; 4: 158–72.
13 *Young MJ, Bresnitz EA, Strom BL.* Sample size nomograms for interpreting negative clinical studies. Ann Intern Med 1983; 99: 248–51.
14 *Hanley JA, Lippman-Hand A.* If nothing goes wrong,

is everything all right? Interpreting zero denominators. JAMA 1983; 249: 1743–5.

15 *Ho AM, Dion PW, Karmakar MJ, Lee A.* Estimating with confidence the risk of rare adverse events, including those with observed rates of zero. Reg Anesth Pain Med 2002; 27: 207–10.