Common Statistical Errors and how to Avoid them

NJ Gogtay, UM Thatte

““To consult the statistician after an experiment is finished is often merely to ask him to conduct a post mortem examination. He can perhaps say what the experiment died of.” – R. A. Fisher

Introduction

Researchers should ensure publication of work done as papers that do not see the light of the day have wasted precious time and resources of all stakeholders and will have failed to advance Evidence Based Practice. Getting the statistics right in the publication is crucial. There is evidence both from India and elsewhere about inappropriate reporting of statistics in papers. For example, Jayakaran and Yadav [2011] analyzed n=196 articles in two Pharmacology journals published from India and found that 78.1% articles had inappropriate descriptive statistics, 31.1% used the wrong statistical test for between group comparisons and only 1% reported statistical assumptions.

Researchers can commit errors at two levels, both of which can affect statistics in the manuscript. The first is errors that occur during the process of research [such as errors in planning and/or implementation of the study] and the second is those that occur during data analysis, interpretation and presentation of results [either in the form of a manuscript or podium/poster presentations]. This paper will describe errors with the latter. It is important to point out here that more often than not, errors during the planning and conduct of research cannot be easily identified by a journal editor [who only sees the final manuscript] and hence any study should be done with utmost care and attention to detail. This is because errors of omission or commission that occur during the process of research are detrimental to the most important stakeholder in research – the patient one whose quality of life we hope to improve through the research process.

Common Statistical Errors

Before we proceed to common statistical errors, let us understand why it is important that we carry out data analysis meticulously. Data analysis is a process that seeks to identify relationships, associations, differences, variance or trends that may exist within the data. The purpose is to see if the results can be generalized to the population or in other words “how true” or real these findings are. It is useful remember the following before data analysis- 1) Organize all collection forms, and material used to record the data in one place 2) Check the data for completeness and accuracy 3) Note missing data if any and decide whether or not to remove from the analysis and document this 4) Assign unique identifiers to the data 5) Feed the data into an appropriate software [Microsoft Excel, SPPS are two examples] and do a thorough quality check of the fed data by someone independent of the study.

The subsequent sections describe common statistical errors. The list given is not intended to be all encompassing but rather cover some of the more common errors made by authors and often missed by editors.

Errors in Data analysis

The choice of Parametric versus non-parametric methods and the importance of assumptions

Statistics as a discipline uses models and assumptions. Prior to applying any parametric test, the researcher needs to check if the assumption of normality is met as these tests are to be used only when the data is normally distributed. Also, it is a common myth that parametric tests are more powerful than non-parametric tests. In fact, they are powerful only to the extent that the assumptions are met, [see later for assumptions for the unpaired t-test] Else, they miss differences or relationships, which would have been otherwise picked up by non-parametric tests.

Using the wrong statistical test

Statistical tests are not only numerous, but also have similar sounding names. Each test is to be used only if certain assumptions are satisfied. For example, the student’s t-test is a widely used parametric test that is of two types- the unpaired [also called the two-sample t test] and the paired t test and data needs to be normally distributed for its use. For the unpaired t-test, in
addition, observations need to be independent and the variances in the two groups equal. This test cannot be used for multiple group comparisons. Williams found that among articles published in the American Journal of Physiology that used either the unpaired or the paired t-test, 17% used the test inappropriately for multiple comparisons.

The need to address confounding and effect modification

Confounding: This occurs when the effect or association between an exposure and outcome is distorted by the presence of a third variable. This variable is one that is linked to both the exposure and the outcome, but does not lie in the causal pathway. Confounders are viewed as the “nuisance” factor/s that distort the association one way or the other [positive or negative]. Let us understand this with an example. One study that compared pet owners versus non-pet owners found that the former had significantly lower systolic blood pressure relative to the latter despite coming from similar socioeconomic backgrounds and having similar body mass index. One possible reason is that pet owners [dog owners in particular] tend to get more exercise relative to non-pet owners and thus exercise here becomes the confounder or confounding variable.

Effect modification: This occurs when the exposure has a different effect on different groups of patients leading to differential outcomes in sub groups. Let us understand this with an example of the use of perioperative beta blockers and association of their use with mortality after non-cardiac surgery in a retrospective cohort study of 663,635 patients. The authors stratified patients based on the Revised Cardiac Risk Index (RCRI); which is a tool used to estimate a patient’s risk of perioperative cardiac complications. The Odds ratios for different levels of risk are presented in Table 1.

The table shows that when beta blockers are given to patients with high RCRI scores [2, 3 or 4 and more], there is a reduction in mortality. However, at lower scores of 1 or 0, this effect is attenuated or even lost and thus RCRI here acts as the “effect modifier”.

Sub group analyses

When two group comparisons are made, the result is an average effect of the two interventions [and the difference between them] in a heterogenous group of patients. The practicing clinician would however like to know if the better treatment is likely to work in the individual patient that he is treating. This is because each individual patient has certain characteristics that define him/her- for example gender, severity of the disease, age [young/middle aged/elderly], alcohol [presence/absence], smoking [presence/absence], diabetes and so on. The clinician may want to know that given a certain set of characteristics, what is the probability of response. This is what is essentially addressed by sub group analyses, which are defined as “analysis of treatment effects with subgroups of patients enrolled in a study/trial. The fundamental idea with these analyses is to uncover interesting relationships that could be explored further. Their disadvantage lies in the fact that 1) the differences or associations uncovered could be spuriously positive [Type 1 error] 2) They may not able to pick up a difference due to smaller numbers of patients in each group [beta error/false negative error] 3) difficult to interpret. Let us understand both their utility and difficulties with two examples.

The IRESSA Survival Evaluation in Lung Cancer (ISEL) was a phase III study that compared the efficacy of Gefitinib versus placebo in patients with refractory advanced non-small cell lung cancer (NSCLC). The study did not yield a significant difference in survival between the two groups of patients. However, when a planned sub group analysis of n = 342 patients of Asian origin was done, [n= 235 received Gefitinib and n= 107 received placebo], it was seen that Gefitinib significantly improved survival among the Asian patients [HR 0.66, 95% CI [0.48,0.91], median survival 9.5 months versus 5.5 months, p <0.01]. At the other end of the spectrum, a sub group analysis of men versus women [a meta-analysis that included n = 6 studies] concluded that the effects of use aspirin for primary prevention reduced myocardial infarction significantly in men but not in women. These findings were however not confirmed by two other meta-analyses [where trials of both primary and secondary prevention were included] and it was concluded that gender did not really affect the efficacy of aspirin. Sub group analysis at best, should be hypothesis generating and the best ones that
those that are specified a priori. Post-hoc sub group analysis when done should be transparently reported.

Understanding and addressing bias

Bias can very simply be defined as any deviation from the truth. Since the purpose of research is to arrive at the truth, researchers must have a good grasp of bias and minimizing it. Bias can be broadly classified as – 1) Selection bias [for example prevalence- incidence bias, admission rate bias and non-responder bias] and 2) Information bias [for example misclassification bias and Hawthorne effect].

Let us understand how bias can affect study results (and interpretation) with an example of a paper by Redelmeier and Singh who reported that Academy Award–winning actors and actresses lived almost 4 years longer than those who did not receive an Oscar. It was later shown that the statistical method used for the analysis actually conferred an unfair advantage to the Oscar winners, a type of bias called as “immortal time bias” and the difference in survival was just one year and not statistically significant. Another and oft quoted example is that of a study by Coren and Halpern [1991] which reported that left handers died much earlier than right handed people. This questionnaire based study did not take into account the fact that in the early part of the 20th century, many parents forced children who were naturally inclined to be left handed to use the right hand resulting in groups that were not clearly only left handed or right handed.

Presenting Data Appropriately

Quantitative data- use summary measures appropriately

When quantitative data [height, weight, blood pressure for example] are presented, both means and medians can be reported as summary measures. It is important that the mean is always accompanied by the standard deviation [SD] which gives the extent of variability seen in the data and written as mean [SD]. The standard error of mean [which is much smaller than the SD and is given by SD/√n] should not be given as it is a population parameter and not sample statistics and will always be smaller than the SD. Jaykaran and Yadav for instance showed that 78.1% papers had inappropriate descriptive statistics and use of mean ± SEM rather than mean ± SD was the most common presentation error seen. The median is used to present skewed data and when used, it must always be accompanied by the range. Alongside the median, graphical depictions of skewed data such as the box and whisker plot giving the inter-quartile range are useful visual aids that help understand variability in the data better.

Presentation of categorical data

Categorical variables may be dichotomous or binary [for example -male and female] or non-binary [mild, moderate and severe pain]. These are described as proportions of the total number of participants [along with 95% Confidence Intervals]. They can also be expressed in the form of a bar or pie chart. Often times, binary categorical data is best presented as a 2 x 2 table. This is particularly done when there are two group comparisons and helps in the calculation of several metrics such as the relative risk, odds ratio, hazard ratio. It is also useful when diagnostic tests are being evaluated for the calculation of sensitivity, specificity, positive and negative predictive values and the diagnostic odds ratio. In addition, tests such as the Chi-square test, Fisher’s exact probability test and the McNemar’s test are best understood and interpreted with the 2 x 2 table.

Outliers and their reporting

An outlier is essentially an abnormal value that lies far away from the rest of the values in the sample. Outliers are important are they can have a significant impact and alter results of the analysis dramatically. They could be true outliers, a typographical error that resulted during data entry [which needs to be corrected], or a wrong measurement. All outliers need to be carefully considered. Given that here is little consensus on how outliers are to be analyzed, it is important that are outliers are reported with honesty and where appropriate an analysis with and without the outlier be performed and reported.

Reporting only P values, not reporting the exact p value and confusing it with the effect size and not reporting Confidence Intervals

The p value that is usually set at 5% essentially tells you whether the results are consistent with being due to chance. It does not by itself provide a good measure of evidence. It must always be accompanied by the effect size [the magnitude and direction of the difference when two group comparisons are made] and the 95% CI of the difference [the confidence interval gives the range in which we expect the true population value to lie]. The p value, the effect size and the confidence intervals of the effect size must be viewed in tandem for drawing meaningful conclusions.

Interpreting Data Correctly

Correlation and Causation

A common mistake is to assume that just because we find a correlation between two variables, one causes the other. This is often described in statistical parlance as “Correlation does not imply causation”. An often quoted example in this regard is the “Correlation” of Sun Signs in Astrology with outcomes by the researchers of the Second International Study of Infarct Survival Trials Collaborative Group [ISIS-2]. Overall, the study showed a significant benefit of aspirin over
placebo \( p < 0.00001 \). However, based on the date of birth entered in the case record forms, when the researchers classified all patients as per the Sun Sign they were born under, two Sun Signs- Gemini and Libra showed no apparent benefit with aspirin while another Sun sign Capricorn, showed a nearly 50% reduction in mortality! It would be inappropriate to say that one Sun Sign appears to benefit more with aspirin than the other. Thus, these relationships should be viewed merely as associations and not cause [Sun Signs] and effect [mortality].

**Extending results Inappropriately**

Both authors and readers are often tempted to draw conclusions beyond what the data actually shows and this should not be done. For example, if a study shows that 5% of lawyers in a certain court in the country have alcohol dependence syndrome, it would be inappropriate to extrapolate this finding to all lawyers who in practice the country or all lawyers practicing worldwide.\(^1\) This finding would only be applicable to populations similar to those where the original sample was drawn from. Prior to extrapolation/generalization, it is thus important to critically appraise the representativeness of the sample.

**The challenge of small sample sizes**

All sample size calculations should be done before starting the study regardless of the number of groups being studied. A small sample size does not necessarily make the study a weak or a poor one. Rather, the ability to generalize and draw inference about the population of interest simply becomes more difficult. Also, with small sample sizes, one must be careful not to overstate the strength of evidence or go beyond what you have observed to draw overarching conclusions.

**Clinical versus Statistical significance**

Statistical significance can be mistaken by both authors and readers with clinical significance. Let us understand this with an example of blood pressure reduction after the use of two anti-hypertensive drugs. Say that Drug A produces a greater reduction in blood pressure than Drug B and the difference is 2mm Hg which is reported as \( p < 0.05 \) and the 95% CI is.\(^1,6\) This means that when Drug A is used in the population, the reduction maybe as low as 1 mm or as high as 6mm Hg, 95% of the times. Thus, whether this 2 mm of Hg [which is the average reduction] is significant enough to alter a change in prescription from Drug B to Drug A must be well thought through as this difference may really not be clinically meaningful.

**Conclusions**

In summary, it is useful to remember that the process of research and its subsequent publication is fraught with the potential for making errors and all efforts must be made to minimize if not eliminate them.

**References**