Common statistical mistakes in manuscripts submitted to biomedical journals

Farrokh Habibzadeh
President, World Association of Medical Editors (WAME); Editor in Chief and Founder, The International Journal of Occupational and Environmental Medicine; Director, NIOC Health Organization Medical Education and Research Center, Shiraz, Iran; Farrokh.Habibzadeh@theijoem.com

Introduction

Statistical methods have rapidly developed over the past decades and become instrumental in data analysis of research articles, so that currently most journals ask the authors to describe in detail the statistical methods used for the analysis of their data in a separate section in the methodology. This is helpful, as it allows the internal validity of the findings presented in the article to be examined. In this review, based on more than 20 years of experience as an editor and reviewer, I will describe the most common mistakes I have encountered in manuscripts submittted to biomedical journals. I found these mistakes with more or less similar frequency in the submissions to both prestigious and small medical journals. Many of these mistakes can also be found in published articles, which means even some editors are not aware of these points.

Distribution of data

Many of the submitted manuscripts involve analysis of continuous variables like age, blood pH, and serum cholesterol level. One of the common mistakes is to treat all such variables similarly. They are presented by many authors as mean and standard deviation (SD) and compared by parametric tests like Student's t test. However, one of the most basic steps in the analysis of these data is to determine if these variables are normally distributed or not.[1] Only normally distributed variables should be presented as mean and SD; non-normally distributed variables should be presented as median and interquartile range (IQR)—the distance between the 25th and 75th percentiles.[2] Parametric tests (for example, Student's t test and one-way analysis of variance [ANOVA]) should only be used for the analysis of normally distributed variables, as normality of the distribution is one of the basic assumptions made by these tests and violation of which would lead to incorrect results.[3] Variables that do not have a normal distribution should be compared with non-parametric (or distribution-free) tests such as Mann-Whitney U test and Kruskal-Wallis.[1] But, how should we test if a variable has normal distribution? The one-sample Kolmogorov-Smirnov test is one of the most popular (non-parametric) statistical tests that can be used. However, as a useful rule of thumb (without access to the raw data, which is very useful for reviewers and editors), if the SD exceeds half of the mean value, then it is unlikely that the distribution of the variable is normal.[4]

SD vs SEM

Another common mistake in submitted manuscripts (even also in published articles) is using the standard error of the mean (SEM) instead of SD to indicate dispersion of the data. SEM is always smaller than the SD, as it is SD divided by the square root of the sample size. Some authors, inappropriately, use SEM instead of SD to imply that their measurements were less dispersed. SEM is in fact the SD of the distribution of the mean, it therefore measures the precision the mean.[1]

Assume that you measure fasting glucose in 225 healthy men and find a mean glucose level of 90 mg/dL with an SD of 15 mg/dL. Assume that the variable has a normal distribution, thus, almost 95% of the study sample (214 = 0.95×225 people) are expected to have a blood glucose between 60 (90 – 2×15) mg/dL and 120 (90 + 2×15) mg/dL (according to the characteristics of the normal distribution, that is, in a normal distribution, 95% of data are within the interval mean±2×SD). Assume that the sample was representative of the population. Then, we can state that 95% of people in the studied population have a blood glucose level between 60 and 120 mg/dL. That is of course helpful. But, suppose that our researcher studied 900 people instead of 225 and came to the same mean (90 mg/dL) and SD (15 mg/dL). What will have changed? Our statements would be exactly the same as before. Here again, based on the results obtained, we can state that 95% of people in the studied population have a blood glucose concentration between 60 and 120 mg/dL. The only difference, as most of you intuitively felt, is that when you study 900 people, the results are more precise than those obtained when you study 225 people.

Suppose that we try to do the research with 225 people 100 times; that is to take samples of 225 people 100 times, to measure their blood glucose. Then, we will have 100 means (and 100 SDs). Of course these 100 means will not be exactly equal and distributed around a number—"mean of means." This "mean of means" is the closest possible value to the true population mean. To examine the level of dispersion of these 100 means around their "mean of means," we can calculate their SD. It can be proved that SEM is a good estimation for this SD.[1,3] Fortunately, for the derivation of this SD, we do not need to run the experiment 100 times and we can simply calculate it from the SD derived in one experiment (here 15 mg/dL). As mentioned earlier, SEM for the 225-participant study is:

$$SEM = \frac{SD}{\sqrt{n}} = \frac{15}{\sqrt{225}} = \frac{15}{15} = 1\,\mathrm{mg/dL}$$

For the 900-participant study, the SEM is:

$$SEM = \frac{SD}{\sqrt{n}} = \frac{15}{\sqrt{900}} = \frac{15}{30} = 0.5\,\mathrm{mg/dL}$$

Regardless of the distribution of the variable in the sample, the distribution of the means is usually normal, thus, considering the characteristics of the normal distribution, 95% of all possible values of the mean are within almost 2×SEM around the mean value. In other words, this interval is the 95% confidence interval (95% CI) of the mean. It means that in our examples, for the 225-participant study, with a probability of 95%, the real mean of the population would be between 98 (100 – 2×1) mg/dL and 102 (100 + 2×1) mg/dL;

the 95% CI for the mean for the 900-participant study is 99 to 101 mg/dL. It is now clear that the 95% CI for the mean for 900 participants (2 = 101–99 mg/dL) is half of that for 225 participants (4 = 102–98 mg/dL)—the measurements of the mean were twice as precise in the 900-participant study compared to those made in the 225-participant study. It simply shows that to get double the precision we should quadruple (2^2) the sample size ($900=4 \times 225$).

SEM is in fact not a measure of dispersion of the studied variable. It is a measure indicating how precise the mean value is.[1,3] In scientific writing, when we want to present a measure of data dispersion, we should use SD, whereas to present how precise a mean value is measured we should present SEM. The standard error is not specific for the mean and we can calculate it as well for other statistics like odds ratio (OR), relative risk, and percentages (for example, prevalence and incidence rate). And that is why these statistics are usually reported along with their 95% CIs.[2] Do not forget that 95% CI and standard error are closely correlated and one can simply be calculated from the other. However, we usually report 95% CI rather than the standard error. The standard error, however, may be used as error bars in graphs to present the accuracy of the measurement.

Inappropriate precision in reporting statistics

The precision with which we report statistics should depend on the precision of our measurement. For example, in a research study on adults, we usually record age in years as generally measuring age more precisely has no implication clinically. In the same study, however, we may measure blood pH with two or even three digits after the decimal point, as minute changes in blood pH are associated with serious clinical implications. Statistical software programmes, however, often calculate the results with a predefined precision, say three digits after the decimal place, no matter how precisely the raw data were measured. Therefore, unless rectified, software programmes report the mean of both of the above-mentioned variables, age and blood pH, with three digits after the decimal point. In submitted manuscripts, it is not uncommon to read statements like "the mean age of patients was 37.351 years." When reporting an age in one-thousandth of a year (almost 9 hours), it means that we asked participants about the hour they were born! While, we usually only ask them about their birth year. The above mean should probably be reported as '37' or '37.4' years with no more precision. There are no consensus on the number of digits to be reported in presenting the mean and SD. While it can be shown mathematically that the mean and SD should be reported with the accuracy used in the measurement of the raw data, some authorities believe that they should be reported with one extra digit.[4,5]

A similar argument is true for percentages. The prevalence of fever in the statement "of 35 participants, 12 (34.29%) had fever" should have been written as '34%.' When increasing or decreasing one participant of 35 participants changes the percentage by almost 3%, talking about 0.29% is not reasonable. Therefore, as a rule of thumb, when the number of total participants (the denominator) is equal to or less than 100 (or, when the value of percentage [here, 34.29] exceeds the number of participants [here, 12]), we should not report any number after the decimal place. When the number of participants is equal to or less than 20, it is better not to report percentage at all, as it may be misleading.[5] Furthermore, it is better to report the 95% CI of the percentage, particularly if that is the primary outcome. Therefore, the above statement should be presented as, "of 35 participants, 12 (34%; 95% CI: 18–51%) had fever." From another perspective, when considering the width of the 95% CI, reporting the prevalence with a higher precision sounds unreasonable.

Reporting p values

In some manuscripts, authors reported p values as $p<0.05$, $p>0.05$ or p=NS. Many authorities believe that it is better to report the exact value of the p value like p=0.023, p=0.647. Previously, p values were read from statistical tables and therefore, determination of their exact value was difficult. However, currently, statistical software programs report the exact value of p. Sometimes, when the p value is very small, say 0.00001, the software that by default reports the value in only three digits after the decimal point, shows the value as '0.000' and the authors incorrectly report the value as p=0.000 or worse $p<0.000$. The p value is a probability and thus can vary from a minimum of zero to maximum of one. If the value is either one or zero, the event will happen (or not happen) for sure. In experimental research, however, we can never be sure and thus, we are practically facing p values that are more than (not equal to) zero and less than (not equal to) one. Therefore, if a software reports a p value as 0.000, the correct presentation would be $p<0.001$. As a p value is a probability, it can never be negative and thus it can never be presented as $p<0.000$. In reporting p values, it is not necessary to report more than three digits after the decimal point. Some journals may ask you to also report the statistical test used, like Pearson $\chi^2=1.796$, df=3; p=0.62.

95% confidence interval vs p value

Sometimes, manuscripts present both p value and 95% CI as statistics. For example, we may see statements like "smoking was significantly (p=0.04) associated with a higher incidence of lung cancer (OR=2.6; 95% CI: 1.3–5.2)." A p value can only indicate the probability of observing the difference by chance, when there is really no such difference in the population (type I error); it does not provide any information on the amount of the change—the so-called effect size. On the other hand, 95% CI not only tells us the effect size, but also if the difference is statistically significant (for example for OR, the difference is significant if the 95% CI does not contain 1). For the above example, the 95% CI of OR (1.3–5.2) indicates that with a probability of 95%, the risk is not less than 1.3 and is not more than 5.2 times that for non-smokers, hence, the effect size; since the 95% CI does not contain 1, it reflects that smoking has a significant effect on the incidence of lung cancer. Therefore, it is not necessary to mention both p value and 95% CI; the latter is sufficient and the statement could instead be written as "smoking was associated with a higher incidence of lung cancer (OR=2.6; 95% CI: 1.3–5.2)."

Sometimes the situation is worse; the p value contradicts the 95% CI. The statement "(OR=3.1; 95% CI: 0.97–9.91, p<0.05)" has internal inconsistency! While p is significant, the 95% CI for OR contains 1, which is impossible. Other impossible statements would be "(OR=4.3; 95% CI: 1.12–16.51; p=0.06)," where the p value is not significant but the 95% CI does not contain 1. These errors are more common in the tables of submitted (and published) manuscripts. The general trend in the use of p value vs 95% CI is to use the latter.

Calculation of the minimum sample size

In many trial reports, the number of people studied is stated but the necessary information to calculate the minimum sample size is not presented. For example, in prevalence studies, the authors usually do not provide the expected frequency of the disease and the acceptable error in the calculation of the prevalence; or in clinical trials, the authors usually fail to provide the minimum change important to them (of clinical importance), the effect size, and the expected SD in the variable. In this way, it is impossible to calculate the minimum sample size.[3]

These problems usually arise from failure to describe the study hypothesis in enough detail. For example, in many submitted manuscripts, you may read "our hypothesis is that drug X is better than drug Y for reduction of low-back pain." Whereas, a better hypothesis would be "compared to drug Y, drug X can reduce, by at least 20%, the pain score of women with mechanical low-back pain, as measured by the visual analog scale," where the study population (women with mechanical low-back pain), the outcome (drop in pain score), measurement (by visual analog scale), and the expected effect size (20%) are described.

Sometimes, we receive studies that are descriptive in nature, say studies on the prevalence of malaria in a region. In such studies, since there is generally no hypothesis, no statistical tests are necessary. Some authors, however, try to decorate such manuscripts with inappropriate use of statistical tests and p values. Another example of inappropriate use of statistical tests is when we examine all members of the population rather than a sample.

Yet another problem that is closely correlated with inappropriate sample size is the issue of distinguishing "clinical significance" from "statistical significance." Sometimes, we read manuscripts that found a statistically difference that is not clinically significant. For example, we read "the mean serum cholesterol level in the study group (189 mg/dL) was significantly (p=0.031) higher than that in the control group (187 mg/dL)." This difference, though statistically significant, is not of any clinical importance and was probably the result of the higher-than-necessary sample size studied. That is why the difference of clinical importance is considered in the calculation of the minimum sample size. Recruiting more people than necessary may result in the observation of differences that, though statistically significant, have no clinical significance. Apart from being unethical, study participants fewer than the minimum sample size may result in type II errors.[3]

Another reason why we may come to statistically significant results without a real difference existing in the population (type I error), is multiple comparisons made in the data analysis.[3] For example if we want to compare the means of five groups by comparing every two groups by Student's t test, we need to run 10 tests. Even if there is no real difference between the five studied groups, with a probability of almost 40%, we will come to a statistically significant p value. This issue will be resolved either by using the appropriate test (eg one-way ANOVA) or by correcting the cutoff value for p for multiple comparisons (say, by the use of Bonferroni's correction).

Non-significant p values

In submitted manuscripts, sometimes we encounter statements like "fasting blood sugar levels in men (97.3 mg/dL) were higher than in women (90.1 mg/dL), however, the difference was marginally significant (p=0.057)." The cutoff value of 0.05 (the probability of 1 in 20) was chosen arbitrarily by Fisher to distinguish "significant" from "non-significant" differences. There is in fact no logical rationale behind the selection of '0.05' for the cutoff value. However, when we choose the cutoff value of 0.05 (which is very common in biomedical sciences), we can no longer talk about "marginally significant," "partially significant," or …—a difference is either significant (p<0.05) or not. In the discussion of manuscripts, we sometime encounter statements like "…the difference was however not statistically significant (p=0.057). If we had recruited more people the difference might become significant." I believe this is not acceptable, as the authors presumably calculated the minimum sample size of their study and recruited the necessary participants.

If a p value is non-significant, it may be due either to the fact that there is really no difference in the population, or the study failed to pick up the real difference that existed in the population (type II error). Therefore, a non-significant p value cannot simply be interpreted as "no difference" in the population. Instead, the authors/reviewers should perform a power analysis to determine the study power and see whether the study is able to detect the difference if there were really a difference in the population.[3] If the minimum sample size was determined correctly, then we can be confident that the study power is also correct.

Conclusion

Submitted manuscripts and even some published articles contain statistical mistakes in the data analysis and presentation. Having a good command of statistics would help editors, reviewers and authors to better evaluate a study. This review touches on some of the most frequent mistakes; however, each of these mistakes should be examined in more detail.

References

1 Spatz C, Johnston JO. *Basic Statistics: Tales of Distribution*. 4th ed. California, Brooks/Cole Publishing Co., 1989.

2 Bowers D, House A, Owens D. *Understanding Clinical Papers*. New York, John Wiley & Sons, 2002.

3 Glantz SA. Primer of Biostatistics. 5th ed. New York, McGraw-Hill, 2002.

4 Lang TA, Secic M. *How to Report Statistics in Medicine: Annotated Guidelines for Authors, Editors, and Reviewers*. 2nd ed. Philadelphia, American College of Physicians, 2006.

5 Peat J, Elliot E, Baur L, Keena V. *Scientific Writing: Easy when you know how*. London, BMJ Publishing Group, 2002.