Adv in Health Sci Educ (2011) 16:291–296
DOI 10.1007/s10459-011-9283-6

INVITED PAPERS ON METHODOLOGY

Alphas, betas and skewy distributions: two ways of getting the wrong answer

Peter Fayers

Received: 14 September 2010 / Accepted: 13 February 2011 / Published online: 12 March 2011

Abstract Although many parametric statistical tests are considered to be robust, as recently shown in Methodologist's Corner, it still pays to be circumspect about the assumptions underlying statistical tests. In this paper I show that robustness mainly refers to α, the type-I error. If the underlying distribution of data is ignored there can be a major penalty in terms of the β, the type-II error, representing a large increase in false negative rate or, equivalently, a severe loss of power of the test.

Keywords Statistics · Robustness · Type-I error · Type-II error · t test · ANOVA

I greatly enjoyed reading Geoff Norman's recent article on Likert scales and the "laws" of statistics (Norman 2010), and emphatically agree with most of the views that he so lucidly expressed. Perhaps that is not surprising, since I believe I was the editor that he cited as dismissing a reviewer's comments about non-normality of ordinal scales as being inappropriate and "it would be unreasonable to single out one paper for rejection on these grounds—the same criticism could be levelled at any number of publications". The purpose of this present note is to question whether the "laws" of statistics can always be dismissed, or whether best practice demands a more cautious approach.

Much of the article focused on the issues of analysing ordinal Likert data as if it is continuous. Here, the assumption is that there is an underlying continuous distribution, but we observe data grouped into discrete categories. As Norman observes, there is loss of power if the data is dichotomised. More commonly, we have four- or five-point or longer Likert scales, and these can indeed safely be regarded as continuous data for many purposes. However, I am very concerned about the somewhat extreme dismissal of *all*

P. Fayers (✉)
Institute of Applied Health Sciences, School of Medicine and Dentistry, University of Aberdeen, Aberdeen, UK
e-mail: p.fayers@abdn.ac.uk

P. Fayers
Department of Cancer Research & Molecular Medicine, Norwegian University of Science and Technology (NTNU), Trondheim, Norway

concerns about non-normality, for example the statement that for “examining differences between means, for sample sizes greater than 5, do not require the assumption of normality and will yield nearly correct answers even for manifestly non-normal and asymmetric distributions like exponentials”. The controversy maybe should cease, as Norman declares, but the statement that “Parametric statistics can be used … with unequal variances, and with non-normal distributions, with no fear of coming to the wrong conclusion'' is not entirely correct, as we shall show.

Although his article is excellent, I would suggest that three critical words were omitted from several crucial statements. This has in fact also been the case throughout much of the history of empirical literature on this topic over the past 80 years. Thus parametric tests are described as being “robust”. What does “robust” mean? Norman defines it as “extent to which the test will give the right answer even when assumptions are violated.” But there are two types of error, false positives and false negatives. The right answer is either a true positive or a true negative, and both forms of error are, well, not “the right answer”. Most of the literature does not state it, but implicitly assumes that robustness only relates to type-I errors. Thus a clearer statement is that t tests and ANOVA are “robust *against type-I errors*”. This of course accords with the enthusiasm that many researchers have in obtaining “significant” *p* values.

The aim of this article is to show that type-II errors can be substantially increased if non-normality is ignored. So a natural precursor question is, do type-II errors matter? Is an increase in type-II error important, or is it simply the type-I error that matters? Some reasons against dismissing type-II errors are:

1. The power of a test is the probability of avoiding a type-II error, also known as the false negative rate. Thus 90% power corresponds to a 10% type-II error rate, or a one-in-ten risk of failing to detect a difference when there really is a true effect. Why lose power and increase the chance of false negatives, when it is easy to select a more appropriate method of statistical analysis?
2. Some argue that type-II errors, which represent a loss of power, can be addressed by increasing the sample size. That is true, but is it really sensible to recruit perhaps as many as 50% extra subjects (and therefore seek 50% extra funding), simply because one cannot be bothered to use the most appropriate statistical analysis? This approach is too absurd for further discussion.
3. It is well known that “no evidence of an effect is not the same as evidence of no effect”. Or, it should be well known. However, many readers assume that if a study has a sample size that claims to provide 90% power to detect an effect, then lack of significance is at least an indication that the effects are likely to be small. Readers might form a very different assessment if they knew that the real power is reduced because an inappropriate analysis ignored non-normality, perhaps for example resulting in an effective power of 70%, which is far less than the nominal 90% that was claimed.
4. Some studies of assessment scales try to demonstrate that two scales are equivalent. Maybe a newer scale can be shorter, containing fewer test items. Or perhaps one wishes to compare this year’s test examination against last year’s version. In statistical terms, these are equivalence studies. An increase in false negatives means we are more likely to declare “no significant difference” and misleadingly conclude that the scales are by default equivalent.

Hence type-II errors should not be ignored. Why is it that most texts ignore this aspect of robustness? Why are these type-II errors more affected than type-I errors? How large can the increase in type-II errors be? Wilcox (2005) eloquently answers the first two questions thus:

"For many years conventional wisdom held that standard analysis of variance methods are robust, and this point of view continues to dominate applied research. In what sense is it correct? What many early studies found was that if two groups are *identical,* meaning that they have equal distributions, Students t test and more generally the ANOVA F test are robust to non-normality, in the sense that the actual probability of a type I error would be close to the nominal level. Many took this to mean that the F test is robust when groups differ. In terms of power some studies seemed to confirm this by focusing on standardised differences among the means…. What these studies failed to take into account is that small shifts away from normality towards a heavy-tailed distribution, [increases the standard deviation and therefore] lowers the standardised difference, and this can mask power problems associated with Student's t test. The important point is that for a given difference between the means modern methods can have substantially more power."

What this means is that methods such as the t test are only robust against non-normality in the sense that the false-positive rate remains approximately "correct". Thus, if the result from a t test is found to be "significant", it most likely is. But the type II error, the probability of a false negative, can be substantially increased (lower power). We can readily illustrate this. One of the distributions frequently encountered in medical and other health studies is the log-normal distribution, which is asymmetric and has a long tail ("skew") to the right. Examples of log-normal distribution are less frequent in educational settings, although times frequently follow this shape: these include time taken by students to complete an examination, or the time taken to learn a new procedure. However, our principal aim in using the lognormal distribution is simply that it provides a convenient illustration of a moderately skew distribution. Figure 1 shows a log-normal distribution, with the equivalent normal distribution (the log-normal distribution is formed by taking exponentials of observations from normal distribution).

As shown on the diagram, the estimate of the mean is pulled to the right by the relatively few very high observations, making the median a better estimate of central tendency. However, several less obvious problems also arise when data are asymmetrically

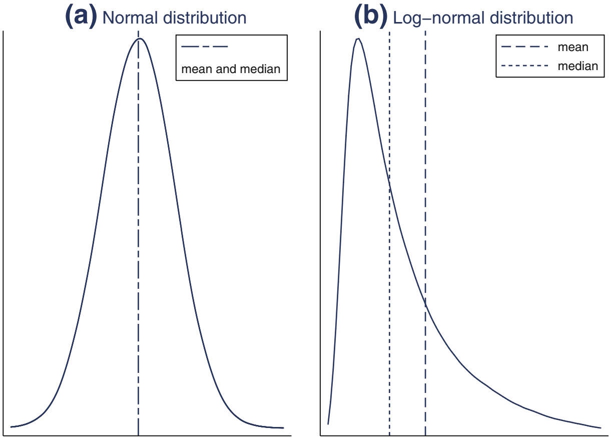


Fig. 1 Log-normal distribution and normal distributions, showing how the mean value is pulled towards the *right* tail if the data is skewed to the *right*

distributed like this. Firstly, the estimate of the standard deviation (SD) becomes inflated by the small proportion of observations in the drawn-out upper tail. Hence the estimated standard error (SE) associated with the mean values is increased, as is the SE of the difference between the means. Thus we might anticipate a smaller value for the t-statistic, leading to increased type-II errors. Secondly, theory tells us that the distribution of the sample variance may deviate substantially from a chi-squared distribution and therefore, especially if the sample size is small, the t test can give misleading results. Thirdly, for a distribution such as the log-normal, the SD is partly a function of the mean value; this implies that (when comparing two groups as in a t test) if there really is a true difference in mean values, then the SDs will also differ—perhaps substantially. For all these reasons we can expect that the results of a t test will be misleading if data follows a log-normal distribution. However, alternative methods are so readily available. In this particular case, as the name log-normal implies, we can obviously use a logarithmic transformation which rescales the data into a well-behaved symmetrical normal distribution; whenever right-skew data is observed, one natural thought should be: "Are the data log-normal, because if so we can use a logarithmic transformation and avoid all complications." In general, a logarithmic transformation always helps to stabilise right-skew data. However, a more general alternative, applicable to data arising from any distribution, is to use a distribution-free test (commonly mis-termed "nonparametric" test) such as the Wilcoxon test. I illustrate the consequences of these methods by simple computer simulations.

Computer-generated random numbers were used to produce data that followed a log-normal distribution. Technically, this was accomplished by first generating random observations from a normal distribution with mean 0 and standard deviation 1, and then using an exponential transformation. A constant was added to the first 50% of these observations, creating a group of observations with an increased mean value. This increase was measured in terms of effect sizes, where the effect size is the mean difference expressed as a multiple of the standard deviation. For example, to simulate a study with a sample size of 100 observations per group, 200 log-normal data items were generated and the first 100 increased. Thus, for example, to produce an effect size of 0.5 which is generally regarded as an effect of moderate magnitude, a value of 0.5 would be added to the normal values because the standard deviation had been set at 1.0. A t test was then applied. A Wilcoxon two-sample signed-rank test was also used, and in addition a logarithmic transformation was applied before a second t test. This was repeated some 30,000 times, to obtain a reasonably precise estimate of the proportion of times that such an effect size, in a study of this magnitude, would result in a difference being found significant ($p < 0.05$). The whole exercise was repeated over and again, for varying numbers of observations in the two comparison groups, and with varying effect sizes The effect sizes considered ranged from 0 (no difference in mean values), through 0.2, 0.5, 0.8 and 1.0, but results are only displayed for an effect size of 0.5 (Fig. 2) and zero (Fig. 3). Thus Fig. 2 shows the power to detect an effect size of 0.5, for varying sample sizes. It should be noted that when the effect size is zero (Fig. 3), there is no difference between the two simulated groups and so the proportion of p values deemed "significant" will represent false positives—that is, in a robust and reliable significance test, on average 5% of results ought to be declared "significant, $p < 0.05$".

For an effect size of half-a-standard deviation, Fig. 2 shows that the straightforward t test is greatly inferior to the more appropriate nonparametric Wilcoxon two-sample test, and that the latter is virtually as good as a test that takes full account of the log-normal distribution (t test on logarithms of the values). This is exactly what we would expect from theory. Inspection of the plot shows that, for any level of power, to obtain comparable

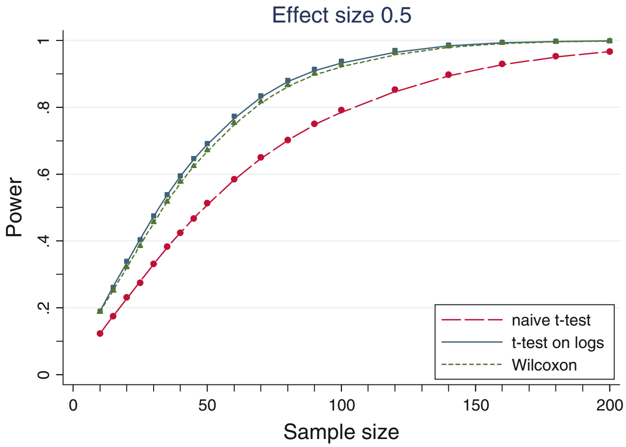
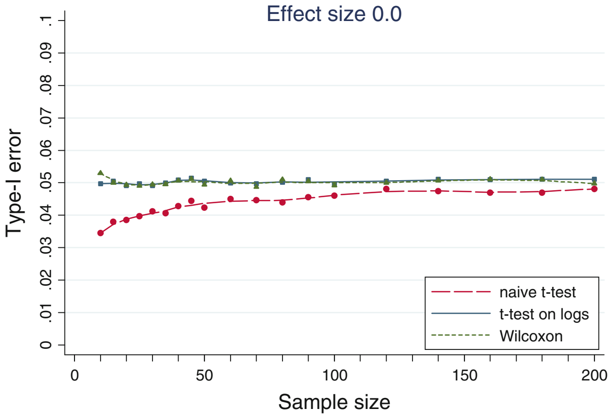


Fig. 2 The relation between power and sample size, for an effect size of 0.5, when a t test, a t test on logarithms and a Wilcoxon rank test are applied to log-normal data. The sample sizes shown are the number of subjects in each of the two groups

Fig. 3 The relation between type-I error and sample size when a t test, a t test on logarithms and a Wilcoxon rank test are applied to log-normal data. Here, the effect size is 0.0 implying that the null hypothesis of "no difference" is true, because the type-I error is "the probability of falsely rejecting the null hypothesis when it is true". The sample sizes shown are the number of subjects in each of the two groups



power using a naïve t test that disregards the distribution would require an increase of about 50% in sample size. Or, conversely, for a sample size of 64 in each group the power is 80% as is expected from theory, but a naïve test would have a power nearer 60%. These are very large and unacceptable losses of power. Broadly similar results (not shown here) were found for other effect sizes.

When the effect size is zero, the null hypothesis is by definition true, and so the type-I error is displayed (Fig. 3). Again, the nonparametric test is close to the optimum test using logarithms of the observations. However, the nominal 0.05 p value derived from the naïve t test deviates quite a bit—for samples of less than about 40 subjects, it is in fact closer to an over-cautious 0.04.

Therefore we see that there can be major losses of power and even some distortion of the p value if the t test is applied to data from a skew distribution. Similar problems arise with ANOVA and many other statistical tests that are founded on the theory of normal distributions. Since it is so easy to use a nonparametric test, why not do so whenever possible and whenever there is reasonable doubt about non-normality and in particular skewness of the distribution?

Finally, it should be noted that when the sample size is small it may be difficult to identify the distribution that underlies the data. Then one may be blissfully unaware that the data deviates markedly from a normal distribution. This leads many statisticians to recommend that in small samples, unless one is confident about normality, distribution-free nonparametric methods should be applied for safety. However, it should be noted that what matters is the population distribution, not the distribution within the sample that has been drawn, and so in the case of small-sized samples the investigator should whenever possible make use of external or prior knowledge about the distribution of the data. Of course nonparametric methods have limitations—it is difficult to generalise nonparametric analyses for more complex designs or for estimation and modelling; frequently, preferable alternatives will include appropriate transformations, general linear models or ordered logistic regression.

So, thank you Geoff, a much needed critique and I thoroughly support the aim of the editorial. However, when the distribution of data is asymmetric and markedly non-normal, best practice should involve considering whether the simplest methods of analysis are optimal. Yes, it is usually appropriate for Likert scales, as widely used in health science education, to be analysed using parametric statistics—but not invariably so, and simple analyses may again be suboptimal if major asymmetry in the underlying distribution is ignored. Please, a degree of caution and circumspection is called for.

References

Norman, G. R. (2010). Likert scales, levels of measurement and the ''laws'' of statistics. *Advances in Health Sciences Education, 15*, 625–632.

Wilcox, R. R. (2005). *Introduction to Robust Estimation and Hypothesis Testing* (2nd ed.). Burlington MA: Elsevier Academic Press. ISBN 978-0-12-751542-7.