Adv in Health Sci Educ (2010) 15:625–632
DOI 10.1007/s10459-010-9222-y

METHODOLOGIST'S CORNER

# Likert scales, levels of measurement and the "laws" of statistics

**Geoff Norman**

Received: 22 January 2010 / Accepted: 22 January 2010 / Published online: 10 February 2010


**Abstract** Reviewers of research reports frequently criticize the choice of statistical methods. While some of these criticisms are well-founded, frequently the use of various parametric methods such as analysis of variance, regression, correlation are faulted because: (a) the sample size is too small, (b) the data may not be normally distributed, or (c) The data are from Likert scales, which are ordinal, so parametric statistics cannot be used. In this paper, I dissect these arguments, and show that many studies, dating back to the 1930s consistently show that parametric statistics are robust with respect to violations of these assumptions. Hence, challenges like those above are unfounded, and parametric methods can be utilized without concern for "getting the wrong answer".

**Keywords** Likert · Statistics · Robustness · ANOVA

One recurrent frustration in conducting research in health sciences is dealing with the reviewer who decides to take issue with the statistical methods employed. Researchers do occasionally commit egregious errors, usually the multiple test phenomenon associated with data—dredging. But this is rarely the basis of reviewer's challenges. As Bacchetti (2002) has pointed out, many of these comments are unfounded or wrong, and appear to result from a review culture that encourages "overvaluation of criticism for its own sake, inappropriate statistical dogmatism", and is subject to "time pressure, and lack of rewards for good peer reviewing". Typical reviewers' comments in this genre may resemble those listed below, drawn from reviews of 5 different papers, all brought to my attention in a 2 month period:

Paper 1

> *…and in case of use of parametric tests (like t-test) I'd like to see the results of the assumption of normality of the distribution*

G. Norman (✉)
McMaster University, 1200 Main St. W., Hamilton, ON L8N3Z5, Canada
e-mail: norman@mcmaster.ca

Paper 2

*… the authors [use]analytical practices which are not supported by the type of data they have available…. Ordinal data do not support mathematical calculations such as change scores, …. the approach adopted by the authors is indefensible….*

Paper 3

*The statistical analysis of correlation …. is done with a method not suitable for non-parametric, consult with statistician.*
*The t-test performed requires that the data be normally distributed. However, the validity of these assumptions …has not been justified*
*Given the small number of participants in each group, can the authors claim statistical significance?*

Paper 4:

*The sample size is very low …. As the data was not drawn from a normal distribution due to the very low sample size, it is not possible to analyse the data using parametric tests, such as ANOVA.*

Paper 5:

*Did you complete a power analysis to determine if your N was high enough to do these tests?*
*…with the low N, not sure if you can claim significance without a power analysis to confirm; otherwise Type II error is highly possible in your results*

Some of these comments, like the proscription on the use of ANOVA with small samples, the suggestion to use power analysis to determine if sample size was large enough to do a parametric test, or the concern that a significant result still might be a Type II error, are simply wrong and reveal more about the reviewer's competence than the study design.

Others, like the various distributional assumptions or the use of parametric statistics with ordinal data, may be strictly true, but fail to account for the robustness of parametric tests, and ignore a substantial literature suggesting that parametric statistics are perfectly appropriate. Regrettably, these reviewers can find compatible company in the literature. For example, Kuzon et al. (1996) writes about the "seven deadly sins of statistical analysis". Sin 1 is using parametric statistics on ordinal data; Sin 2 relates to the assumption of normality and claims that "Before parametric statistical analysis is appropriate… the study sample must be drawn from a *normally distributed population [ital. theirs]*" and (2) the sample size must be large enough to be representative of the population".[1]

The intention of this paper is to redress the balance. One of the beauties of statistical methods is that, although they often involve heroic assumptions about the data, it seems to matter very little even when these are violated. In order to help researchers more effectively deal with challenges like those above, this paper is a review of the assumptions of

[1] Representativeness is required of all statistical tests and is fundamental to statistical inference. But it is unrelated to sample size.

various statistical methods and the problems (or more commonly the lack of problems) when the assumptions are violated.

These issues are particularly germane to educational research because so many of our studies involve rating scales of one kind or another and virtually all rating scales involve variants on the 7 point Likert scale. It does not take a lot of thought to recognize that Likert scales are ordinal. To quote a recent article in Medical Education (Jamieson 2004) "the response categories have a rank order but the intervals between values cannot be presumed equal". True—strictly speaking. The consequence is that, again according to Jamieson, "the appropriate descriptive and inferential statistics differ for ordinal and interval variables and if the wrong statistical technique is used, the researcher increases the chance of coming to the wrong conclusion". Again, true—strictly speaking. But what is left unsaid is how much it increases the chance of an erroneous conclusion. This is what statisticians call "robustness", the extent to which the test will give the right answer even when assumptions are violated. And if it doesn't increase the chance very much (or not at all), then we can press on.

It is critically important to take this next step, not simply because we want to avoid "coming to the wrong conclusion". As it turns out, parametric methods are incredibly versatile, powerful and comprehensive. Modern parametric statistical methods like factor analysis, hierarchical linear models, structural equation models are all based on an assumption of normally distributed, interval-level data. Similarly generalizability theory, is based on ANOVA that again is a parametric procedure. By contrast, rank methods like Spearman rho, Kruskal–Wallis, appear frozen in time and are used only rarely. They can handle only the simplest of designs. If Jamieson and others are right and we cannot use parametric methods on Likert scale data, and we have to prove that our data are exactly normally distributed, then we can effectively trash about 75% of our research on educational, health status and quality of life assessment (as pointed out by one editor in dismissing one of the reviewer comments above).

Well, despite the fact that Jamieson's recent paper has apparently taken the medical education world by surprise and was the most downloaded paper in Medical Education in 2004, the arguments back and forth have been going on for a very long time. I will spend some time reviewing these issues, but instead of focusing on assumptions, I will directly address the issue of robustness. I will explore the impact of three characteristics-sample size, non-normality, and ordinal-level measurement, on the use of parametric methods. The arguments and responses:

1) You can't use parametric tests in this study because the sample size is too small

This is the easiest argument to counter. The issue is not discussed in the statistics literature, and does not appear in statistics books, for one simple reason. Nowhere in the assumptions of parametric statistics is there any restriction on sample size. It is simply not true, for example, that ANOVA can only be used for large samples, and one should use a $t$ test for smaller samples. ANOVA and t tests are based on the same assumptions; for two groups the $F$ test from the ANOVA is the square of the $t$ test. Nor is it the case that below some magical sample size, one should use non-parametric statistics. Nowhere is there any evidence that non-parametric tests are more appropriate than parametric tests when sample sizes get smaller.

In fact, there is one circumstance where non-parametric tests will give an answer that can be extremely conservative (i.e. wrong). The act of dichotomizing data (for example, using final exam scores to create Pass and Fail groups and analyzing failure rates, instead of simply analyzing the actual scores), can reduce statistical power enormously. Simulations I conducted showed that if the data are reasonably continuous and reasonably "wellbehaved" (begging the issue of what is "reasonable") dichotomizing the data led to a reduction in statistical power. To do this, I began with data from two hypothetical distributions with a known separation, so that I could compute a *Z* test on the difference between means. (For example, two distributions centered on 50 and 55, with a sample size of 100, and a standard deviation of 15. I then drew a cutpoint so that each distribution was divided into 2 groups (a "pass and a "fail"). This then led to a 2 $\times$ 2 table with proportions derived from the overlap of the original distributions and the location of the cutpoint. I then computed the required sample size for a *P*-value of .05 using a standard formula. Finally, I calculated the ratio of the sample size for a significant Z test and computed the ratio. The result was a cost in sample size from 20% (when the cutpoint was on the 50th percentile) to 2,600% (when the cutpoint was at the 5th or 95th percentile). The finding is neither new nor publishable; other authors have shown similar effects (Suissa 1991; Hunter and Schmidt 1990).

Sample size is not unimportant. It may be an issue in the use of statistics for a number of reasons unrelated to the choice of test:

(a) With too small a sample, external validity is a concern. It is difficult to argue that 2 physicians or 3 nursing students are representative of anything (qualitative research notwithstanding). But this is an issue of judgment, not statistics.
(b) As we will see in the next section, when the sample size is small, there may be concern about the distributions (see next section). However, it turns out that the demarcation is about 5 per group. And the issue is not that one cannot do the test, but rather that one might begin to worry about the robustness of the test.
(c) Of course, small samples require larger effects to achieve statistical significance. But to say, as one reviewer said above, "Given the small number of participants in each group, can the authors claim statistical significance?", simply reveals a lack of understanding. If it's significant, it's significant. A small sample size makes the hurdle higher, but if you've cleared it, you're there.

2) You can't use t tests and ANOVA because the data are not normally distributed

This is likely one of the most prevalent myths. We all see the pretty bell curves used to illustrate *z* tests, *t* tests and the like in statistics books, and we learn that "parametric tests are based on the assumption of normality". Regrettably, we forget the last part of the sentence. For the standard t tests ANOVAs, and so on, it is the assumption of normality *of the distribution of means*, not of the data. The Central Limit Theorem shows that, for sample sizes greater than 5 or 10 per group, the means are approximately normally distributed regardless of the original distribution. Empirical studies of robustness of ANOVA date all the way back to Pearson (1931) who found ANOVA was robust for highly skewed non-normal distributions and sample sizes of 4, 5 and 10. Boneau (1960) looked at normal, rectangular and exponential distributions and sample sizes of 5 and 15, and showed that 17 of the 20 calculated *P*-values were between .04 and .07 for a nominal 0.05. Thus both theory and data converge on the conclusion that parametric methods 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 nonnormal and asymmetric distributions like exponentials.

3) You can't use parametric tests like ANOVA and Pearson correlations (or regression, which amounts to the same thing) because the data are ordinal and you can't assume normality.

The question, then, is how robust are Likert scales to departures from linear, normal distributions. There are actually three answers. The first, perhaps the least radical, is that expounded by Carifio and Perla (2008) in their response to Jamieson (2004). They begin, as I have, in pointing out that those who defend the logical position that parametric methods cannot be used on ordinal data ignore the many studies of robustness. But their strongest argument appears to be that while Likert questions or items may well be ordinal, Likert scales, consisting of sums across many items, will be interval. It is completely analogous to the everyday, and perfectly defensible, practice of treating the sum of correct answers on a multiple choice test, each of which is binary, as an interval scale. The problem is that they, by extension, support the “ordinalist” position for individual items, stating “Analyzing a single Likert item, it should also be noted, is a practice that should occur only rarely.” Their rejoinder can hardly be viewed as a strong refutation.

The second approach, as elaborated by Gaito (1980), is that this is not a statistics question at all. The numbers “don’t know where they came from”. What this means is that, even if conceptually a Likert scale is ordinal, to the extent that we cannot theoretically guarantee that the true distance between 1 = “Definitely disagree” and 2 = “Disagree” is the same as “4 = “No opinion” and 5 = “Moderately agree”, this is irrelevant to the analysis because the computer has no way of affirming or denying it. There are no independent observations to verify or refute the issue. And all the computer can do is draw conclusions about the numbers themselves. So if the numbers are reasonably distributed, we can make inferences about their means, differences or whatever. We cannot, strictly speaking, make further inferences about differences in the underlying, latent, characteristic reflected in the Likert numbers, but this does not invalidate conclusions about the numbers. This is almost a “*reductio ad absurbum*” argument, and appears to solve the problem by making it someone else’s, but not the statistician’s problem. After all, someone has to decide whether the analysis done on the numbers reflects the underlying constructs, and Gaito provides no support for this inference.

So let us return to the more empirical approach that has been used to investigate robustness. As we showed earlier, ANOVA and other tests of central tendency are highly robust to things like skewness and non-normality. Since an ordinal distribution amounts to some kind of nonlinear relation between the number and the latent variable, then in my view the answer to the question of robustness with respect to ordinality is essentially answered by the studies cited above showing robustness with respect to non-normality.

However, when it comes to correlation and regression, this proscription cannot be dealt with quite so easily. The nature of regression and correlation methods is that they inherently deal with variation, not central tendency (Cronbach 1957). We are no longer talking about a distribution of means. Rather, the magnitude of the correlation is sensitive to individual data at the extremes of the distribution, as these “anchor” the regression line. So, conceivably, distortions in the distribution—skewness or non-linearity—could well “give the wrong answer”.

If the Likert ratings are ordinal which in turn means that the distributions are highly skewed or have some other undesirable property, then it is a statistical issue about whether or not we can go ahead and calculate correlations or regression coefficients. It again becomes an issue of robustness. If the distributions are not normal and linear. what happens to the correlations? This time, there is no “Central Limit Theorem” to provide theoretical confidence. However, there have been a number of studies that are reassuring. Pearson (1931, 1932a, b), Dunlap (1931) and Havlicek and Peterson (1976) have all shown, using theoretical distributions, that the Pearson correlation is robust with respect to skewness and nonnormality. Havlicek and Peterson did the most extensive simulation study, looking at sample size from 5 to 60 (with 3,000–5,000 replications each), for normal, rectangular, and ordinal scales (the latter obtained by adding and subtracting numbers at random). They then computed the proportions of observed correlations within each nominal magnitude, e.g. for a nominal proportion of 0.05, the proportion of samples in this zone ranged from .046 to .053. They concluded that "The Pearson r is rather insensitive to extreme violations of the basic assumptions of normality and the type of scale".

I confirmed these results recently with some real scale data. I had available a data set from 93 patients who had completed a quality of life measure related to cough consisting of 8, 10 point scales, on two occasions (Fletcher et al. 2010). The questions were of the form:

*I have had serious health problems before my visit.*
*I have been unable to participate in activities before my visit.*

and the responses were on a 10 point scale, with gradations:

0 = no problem
2 = mild problem
4 = moderate problem
6 = severe problem
8 = very serious problem
10 = worst possible problem

Each response was made by inspecting a card that showed: (a) The number, (b) The description, (c) A graphical "ladder", and (d) a sad to happy face.

Using the data set, I computed the Pearson correlation between each of the Time 1 scale responses and each of the Time 2 responses, resulting in 64 correlations based on a sample of 93 respondents. I then calculated the Spearman correlation based on ranks derived from the 10 scale points. Finally, I then treated these 64 pairs of Spearman and Pearson correlations as raw data, and computed the regression line, predicting the Spearman correlation from the Pearson correlation. A perfect relationship would have a correlation (Pearson) of 1.0 between the calculated Pearson and Spearman correlations, a slope of 1.0 and an intercept of 0.0.

To then create more extremely ordinal data sets, I first turned the raw data into 5 point scales, by combining 0 and 1, 2 and 3, 4 and 5, 6 and 7, and 8, 9 and 10. Finally, to model a very ordinal skewed distribution, I created a new 4—point scale, where 0 = 1; 1 and 2 = 2; 3, 4, and 5 = 3; and 6, 7, 8, 9, and 10 = 4. Again I computed Pearson and Spearman correlations and looked at the relation between the two (Table 1).

For the original data, the correlation between Spearman and Pearson coefficients was 0.99, the slope was 1.001, and the intercept was −.007. Even with the severely skewed data, the correlation was still 0.987, the slope was 0.995, and the intercept was −.0003. The means of the Pearson and Spearman correlations were within 0.004 for all conditions.

For this set of observations, the Pearson correlation and the Spearman correlation based on ranks yielded virtually identical values, even in conditions of manifestly non-normal, skewed data. Now it turns out that, when you have many tied ranks, the Spearman gives slightly different answers than the Pearson, but this reflects error in the Spearman way of dealing with ties, not a problem with the Pearson correlation. The Pearson correlation like all parametric tests we have examined, is extremely robust with respect to violations of assumptions.

*4) You cannot use an intraclass correlation (or Generalizability Theory) to compute the reliability because the data are nominal/ordinal and you have to use Kappa (or Weighted Kappa)*

Although this appears to be a special case of the previous section, there is a concise answer to this particular question. Kappa was originally developed as a "Coefficient of

**Table 1** Relation between Pearson and Spearman correlations for 64 pairs based on $N = 93$ patients

| | Original 10 point scales | Collapsed 5 point scales | Transformed 4 point scales |
|---|---|---|---|
| Slope | 1.001 | 1.018 | 0.995 |
| Intercept | −0.007 | −0.013 | −0.0003 |
| Correlation | 0.990 | 0.992 | 0.987 |
| Mean Pearson | 0.529 | 0.521 | 0.485 |
| Mean Spearman | 0.523 | 0.517 | 0.488 |

agreement for nominal scales" (Cohen 1960), and in its original form was based on agreement expressed in a 2 × 2 frequency table. Cohen (1968) later generalized the formulation to "weighted kappa", to be used with ordinal data such as Likert scales, where the data would be displayed as agreement in a 7 × 7 matrix. Weighting accounted for partial agreement (Observer 1 rates it 6; Observer 2 rates it 5). Although any weighting scheme is possible, the most common is "quadratic" weights, where disagreement of 1 unit is weighted 1, of 2 is weighted 4, of 3, 9, and so forth.

Surprisingly, if one proceeds to calculate an intraclass correlation with the same 7-point scale data, the results are mathematically identical, as proven by Fleiss and Cohen (1973). And if one computes an intraclass correlation from a 2 × 2 table, using "1" when there is agreement and "0" when there is not, the unweighted kappa is identical to an ICC. Since ICCs and G theory are much more versatile (Berk 1979), handling multiple observers and multiple factors with ease this equivalence is very useful.

## Summary

Parametric statistics can be used with Likert data, with small sample sizes, with unequal variances, and with non-normal distributions, with no fear of "coming to the wrong conclusion". These findings are consistent with empirical literature dating back nearly 80 years. The controversy can cease (but likely won't).

## References

Bacchetti, P. (2002). Peer review of statistics in medical research: the other problem. *British Medical Journal, 234*, 1271–1273.

Berk, R. A. (1979). Generalizability of behavioral observations: a clarification of interobserver agreement and interobserver reliability. *American Journal of Mental Deficiency., 83*, 460–472.

Boneau, C. A. (1960). The effects of violations of assumptions underlying the *t* test. *Psychological Bulletin, 57*, 49–64.

Carifio, L., & Perla, R. (2008). Resolving the 50 year debate around using and misusing Likert scales. *Medical Education, 42*, 1150–1152.

Cohen, J. J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*, 37–46.

Cohen, J. J. (1968). Weighted Kappa; Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin, 70*, 213–220.

Cronbach, L. J. (1957). The two disciplines of scientific psychology. *American Psychologist, 12*, 671–684.

Dunlap, H. F. (1931). An empirical determination of means, standard deviations and correlation coefficients drawn form rectangular distributions. *Annals of Mathematical Statistics, 2*, 66–81.

Fleiss, J. L., & Cohen, J. J. (1973). The equivalence of weighed kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement, 33*, 613–619.

Fletcher, K. E., French, C. T., Corapi, K. M., Irwin, R. S. & Norman, G. R. (2010). Prospective measures provide more accurate assessments than retrospective measures of the minimal important difference in quality of life. *Journal of Clinical Epidemiology* (in press).

Gaito, J. (1980). Measurement scales and statistics: Resurgence of an old misconception. *Psychological Bulletin, 87*, 564–567.

Havlicek, L. L., & Peterson, N. L. (1976). Robustness of the Pearson correlation against violation of assumption. *Perceptual and Motor Skills, 43*, 1319–1334.

Hunter, J. E., & Schmidt, F. L. (1990). Dichotomozation of continuous variables: The implications for meta-analysis. *Journal of Applied Psychology, 75*, 334–349.

Jamieson, S. (2004). Likert scales: How to (ab)use them. *Medical Education, 38*, 1217–1218.

Kuzon, W. M., Urbanchek, M. G., & McCabe, S. (1996). The seven deadly sins of statistical analysis. *Annals of Plastic Surgery, 37*, 265–272.

Pearson, E. S. (1931). The analysis of variance in the case of non-normal variation. *Biometrika, 23*, 114–133.

Pearson, E. S. (1932a). The test of signficance for the correlation coefficient. *Journal of the American Statistical Association, 27*, 128–134.

Pearson, E. S. (1932b). The test of signficance for the correlation coefficient: Some further results. *Journal of the American Statistical Association, 27*, 424–426.

Suissa, S. (1991). Binary methods for continuous outcomes: a parametric alternative. *Journal of Clinical Epidemiology, 44*, 241–248.