Practical Assessment, Research, and Evaluation

Volume 27 Article 6

2022

Power Properties of Ordinal Regression Models for Likert Type Data

Ulf Olsson
Swedish University of Agricultural Sciences

Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 27 Number 6, April 2022 ISSN 1531-7714

Power Properties of Ordinal Regression Models for Likert Type Data

Ulf Olsson, *Swedish University of Agricultural Sciences*

We discuss analysis of 5-grade Likert type data in the two-sample case. Analysis using two-sample *t* tests, nonparametric Wilcoxon tests, and ordinal regression methods, are compared using simulated data based on an ordinal regression paradigm. One thousand pairs of samples of size n=10 and n=30 were generated, with three different degrees of skewness. For all sample sizes and degrees of skewness, the ordinal probit model has highest power. This is not surprising since the data was generated with this model in mind. Slightly more surprising is that the *t* test has higher power than the Wilcoxon test in all studied situations, even for skewed data. For n=30, the differences between the methods are small.

Introduction

Likert type data are often obtained in questionnaires. The respondent would answer some opinion type question by selecting among alternatives such as

Strongly disagree	Disagree	Undecided	Agree	Strongly Agree
1	2	3	4	5

The answer is often coded before storing it in computer files for analysis.

Data of this type is called Likert type data. There is an inherent order among the alternatives, but the distance between, e.g. strongly disagree and Disagree is not necessarily the same as the distance between Agree and strongly agree. Thus, the scale is not equidistant. The measurement scale is *ordinal*. Likert data are common in subject-matter areas such as education, psychology, political science and public health.

There are several options for statistical analysis of Likert type data, in the two-sample case:

- You can ignore the ordinal nature of the data and pretend that the data is numeric and normally distributed. Then, parametric methods such as regression analysis and t tests are used.
- You can analyze the data using non-parametric methods like the Wilcoxon test.
- You can analyze the data using generalized linear models for ordinal data, so called ordinal regression methods (McCullagh and Nelder, 1989), using a probit or a logistic link function.

Some other approaches to analysis of Likert data have been suggested. One option is to make the data binary by coding, for example, [1, 2, 3] as 0, and [4, 5] as 1, and then using binary logistic regression on the coded data. This approach would discard some of the information in the data and is sensitive to the choice of cutting point.

There is some debate on the use of parametric vs. nonparametric methods for Likert data. For example, Boone and Boone (2012) argue against use of parametric methods for single Likert items. On the other hand, Norman (2010) claims that "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" (p. 631). As a contribution to this debate, it may be of interest to compare the power properties of the different tests.

For large samples, the Central limit theorem (see e.g., Sen and Singer, 1993) suggests that most of the tests will perform reasonably well. Also, Bhattacharya and Sengupta (2021) discuss large-sample methods for Likert data. Since the large-sample properties of Likert data are relatively well known, the purpose of this paper is to examine how ordinal regression methods compare to the t test and to the Wilcoxon test, in terms of power, for small (n=10) or moderately small (n=30) data sets.

Earlier Research

Several authors have made comparisons between the *t* test and the Wilcoxon test. Some examples are Neave and Granger (1968); Blair and Higgins (1980); MacDonald (1999). These authors make comparisons based on simulated continuous distributions. It is shown that the *t* test has slightly higher power if the distribution is normal, but that the Wilcoxon test to be preferred for skewed continuous distributions.

De Winter and Dodou (2010) made a simulation study to compare the *t* test to the Wilcoxon test for Likert items. They simulated data from fourteen populations and concluded that, except for a few extreme cases, the two methods have similar power.

The Wilcoxon test

The Wilcoxon test, also called the Mann-Whitney-Wilcoxon *U* test, is used to test the hypothesis that two samples come from the same population. The test is performed by ranking the data and calculating the sums of the ranks for each group. If several observations share the same value ("ties") they are assigned the average rank. Denote the sum of the ranks for group 1 with R_1, and the sample size for group 1 with n_1. The test statistic is calculated as

$$U = R_1 - \frac{n_1(n_1+1)}{2}$$

In large samples, the significance of U can be assessed based on a normal approximation. If there are ties in the data, some modifications to the formulas for the variance are needed; see e.g. Lehmann (1975). In small samples, tables of the distribution of U, or appropriate computer routines, are used.

Ordinal regression models

Ordinal regression methods are a special case of generalized linear models (McCullagh and Nelder, 1989). One way to motivate ordinal regression models is to assume that the observed data, i.e. the manifested opinion Y, is generated from an underlying (latent) variable X as

If $X \leq t_1$ then Y=1

If $t_1 < X \leq t_2$ then Y=2

If $t_2 < X \leq t_3$ then Y=3

If $t_3 < X \leq t_4$ then Y=4

If $t_4 \leq X$ then Y=5

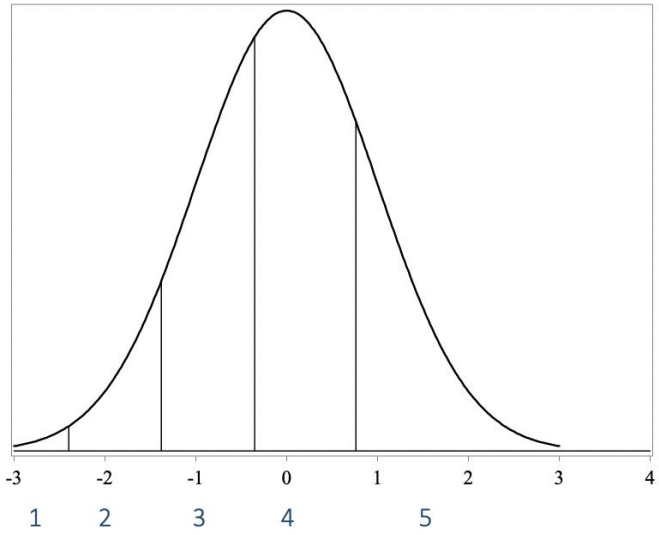
The values t_1 to t_4 are called thresholds. This model is illustrated in Figure 1. In this graph, the latent distribution is assumed to be Normal and the thresholds are computed from the "Moderately skewed" observed distribution used in the simulations; see below.

Alternatively, the model can be motivated using a proportional odds argument. Sample estimates of cumulative probabilities of type $P(Y \leq y)$ are modeled as functions of the independent variables using binary logit or probit regression. It is assumed that the intercepts are different for different y, but that regression coefficients are equal.

It can be shown (McCullagh and Nelder, 1989) that these two approaches are mathematically equivalent. The cumulative logit model corresponds to a model with a logistic latent variable while the cumulative

probit model, used in this paper, corresponds to a model with a normally distributed latent variable.

Figure 1. Latent distribution that generates the observed values Y=1, 2, 3, 4, or 5



Simulation Study

In this paper, we compare the power of the ordinal regression model with the power of the *t* test and the Wilcoxon test using simulated "pseudo-Likert" data. The simulations are based on the ordinal regression paradigm assuming an underlying (latent) normal distribution; see Figure 1. For this reason, most of the analyses in this paper are made using a multinomial probit model, which will serve as a baseline for the other methods.

According to our experience, Likert data may be skewed but are rarely multi-modal. One option for generating "pseudo-Likert" data is to use a discrete probability distribution that is flexible enough to permit skewed data. The binomial distribution has these properties.

We generated data for two groups, denoted as control group and experimental group. Data for the control group were generated from a binomial distribution with N=4 and a known value of P. Since binomial data with N=4 can take on the five values (0, 1, 2, 3, 4), the value 1 was added to make the range into 1 to 5 instead of 0 to 4. Data were generated using a symmetric distribution (P=0.5); a moderately skewed distribution (P=0.7); and a skewed distribution (P=0.9). We believe that this approach emulates many types of Likert data that may occur in practice. This approach is similar to the one used by Olsson (1979). The distributions are presented in Table 1.

The thresholds in Figure 1 were placed at the percentage points of a standard Normal distribution that correspond to the probabilities in Table 1, for the chosen value of P. Figure 1 illustrates the position of the thresholds for the moderately skewed case (P=0.7).

These thresholds were applied on a second Normal distribution, for the experimental group. It has mean value d and standard deviation 1. This distribution was used to calculate the probability distribution of Y for this group; see Figure 2.

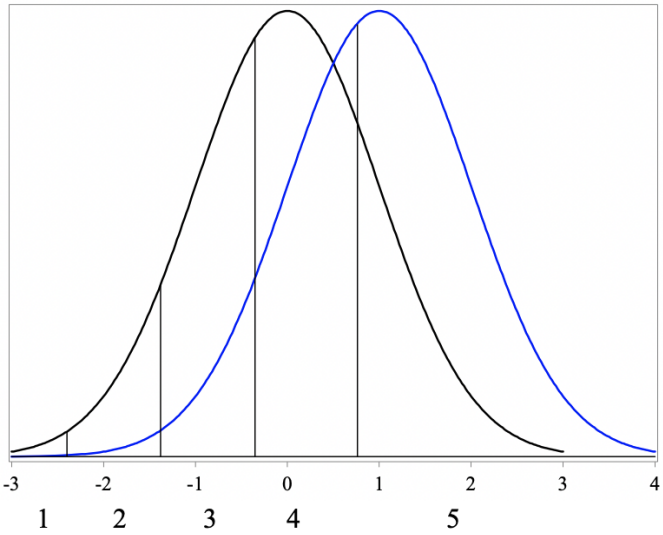
Thus, the setup includes two latent normal distributions, one with mean value 0 and one with mean value d.

Table 1: Probability distributions for Y used in the simulations, for the control group.

Y	P=0.5 Symmetric	P=0.7 Moderately skewed	P=0.9 Skewed
1	0.06	0.01	0.00
2	0.25	0.08	0.00
3	0.38	0.26	0.05
4	0.25	0.41	0.29
5	0.06	0.24	0.66

For each value of *P*, *d* and *n*, one thousand samples were generated. Each sample contained n observations from each of the two groups. The mean difference d between the two groups was varied in six steps from 0 to C, where C was determined such that the plots of the empirical power functions would end at a power of about 99%. C was calculated as 5×SE(*d*), where SE(*d*) is the standard error of the difference between the two mean values.

Figure 2: Latent distribution for the control group (black) and the experimental group (blue). Thresholds, common for both distributions, are indicated by black vertical lines. In this illustration, the means for the groups, on the latent variable scale, are 0 and 1, respectively, so the difference in mean value is d=1.



The sample sizes used were *n*=10 and *n*=30. Data were generated and analyzed using SAS (2018) software. Each pair of samples was analyzed in three ways:

- Using a pooled *t* test of the observed scores Y, assuming equal variances. The SAS Ttest procedure was used.
- Using a Wilcoxon test of the ranks. The *p* value was computed in an exact way and not using the large-sample approximation. The Npar1way procedure in SAS was used.
- Using an ordinal regression model in the Genmod procedure in SAS. A multinomial distribution and a cumulative probit link was used.

For each analysis, the *p* value was classified as significant if $p<0.05$. The proportion of significant results, i.e., the empirical power of the test, was tabulated and plotted against the mean difference *d* for the latent variable.

Results

Empirical power functions for the different values of *P* and *d* are given in Table 2 for *n*=10, and Table 3 for *n*=30. The corresponding graphs are presented in Figures 3 to 8.

Table 2: Estimated power values at for the three tests at different values of *d*, for *n*=10. *d* is the difference in mean value between the two groups, i.e., between the two latent distributions.

	P=0.5			P=0.7			P=0.9		
d	t test	ordinal	Wil-coxon	t test	ordinal	Wil-coxon	t test	ordinal	Wil-coxon
0.00000	0.048	0.067	0.041	0.041	0.077	0.018	0.040	0.076	0.019
0.44721	0.248	0.285	0.196	0.117	0.184	0.079	0.129	0.178	0.075
0.89443	0.576	0.619	0.488	0.451	0.523	0.348	0.455	0.515	0.366
1.34164	0.812	0.844	0.727	0.866	0.903	0.834	0.840	0.878	0.811
1.78885	0.947	0.955	0.932	0.990	0.996	0.987	0.992	0.996	0.987
2.23607	0.995	0.997	0.994	0.999	1.000	0.999	0.995	1.000	0.996

Practical Assessment, Research & Evaluation, Vol 27 No 6

Table 3: Estimated power values at for the three tests at different values of *d*, for *n*=30. *d* is the difference in mean value between the two groups, i.e. between the two latent distributions.

	P=0.5			P=0.7			P=0.9		
d	t test	ordinal	Wil-coxon	t test	ordinal	Wil-coxon	t test	ordinal	Wil-coxon
0.00000	0.069	0.076	0.061	0.067	0.081	0.068	0.055	0.062	0.043
0.25820	0.356	0.365	0.325	0.383	0.443	0.375	0.148	0.155	0.129
0.51640	0.825	0.827	0.761	0.746	0.792	0.719	0.435	0.460	0.413
0.77460	0.955	0.956	0.918	0.950	0.962	0.942	0.789	0.810	0.778
1.03280	0.988	0.991	0.977	0.995	0.999	0.996	0.969	0.973	0.971
1.29099	1.000	1.000	0.999	1.000	1.000	1.000	0.998	0.999	0.999

Figure 3: Empirical power functions for the three tests, for symmetric data, *n*=10.

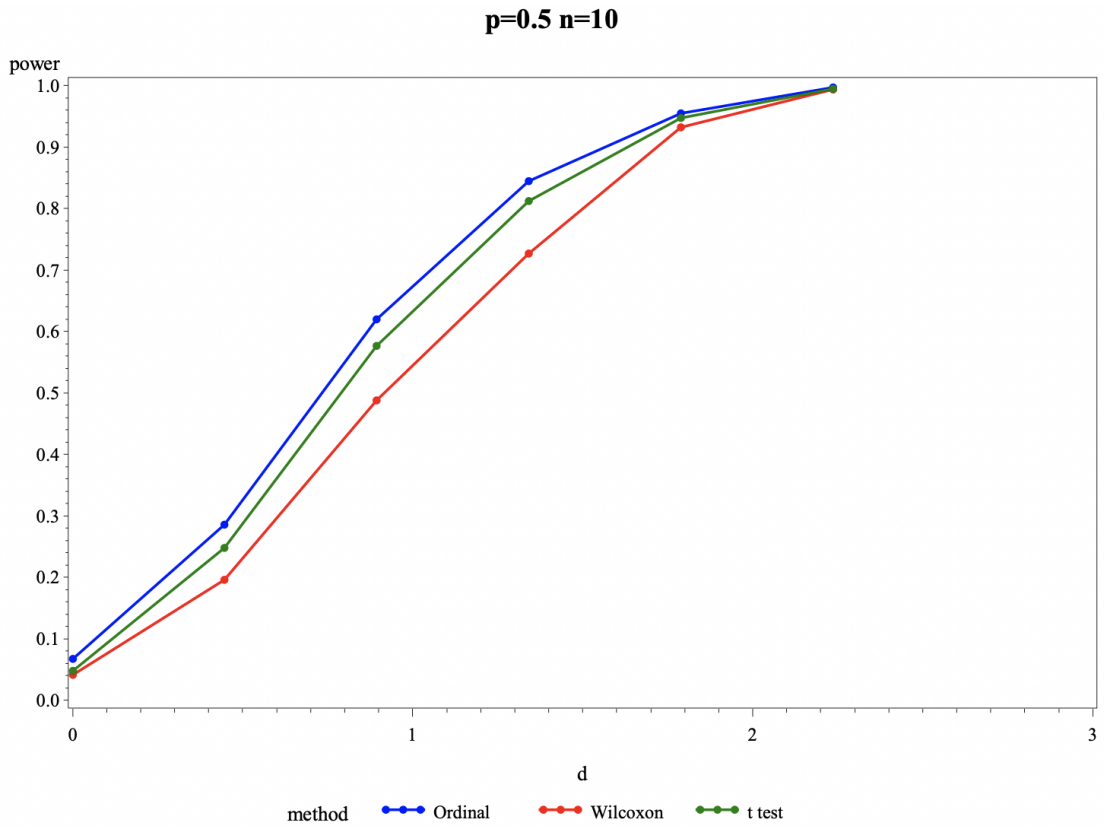


Figure 4: Empirical power functions for the three tests, for moderately skewed data, *n*=10.

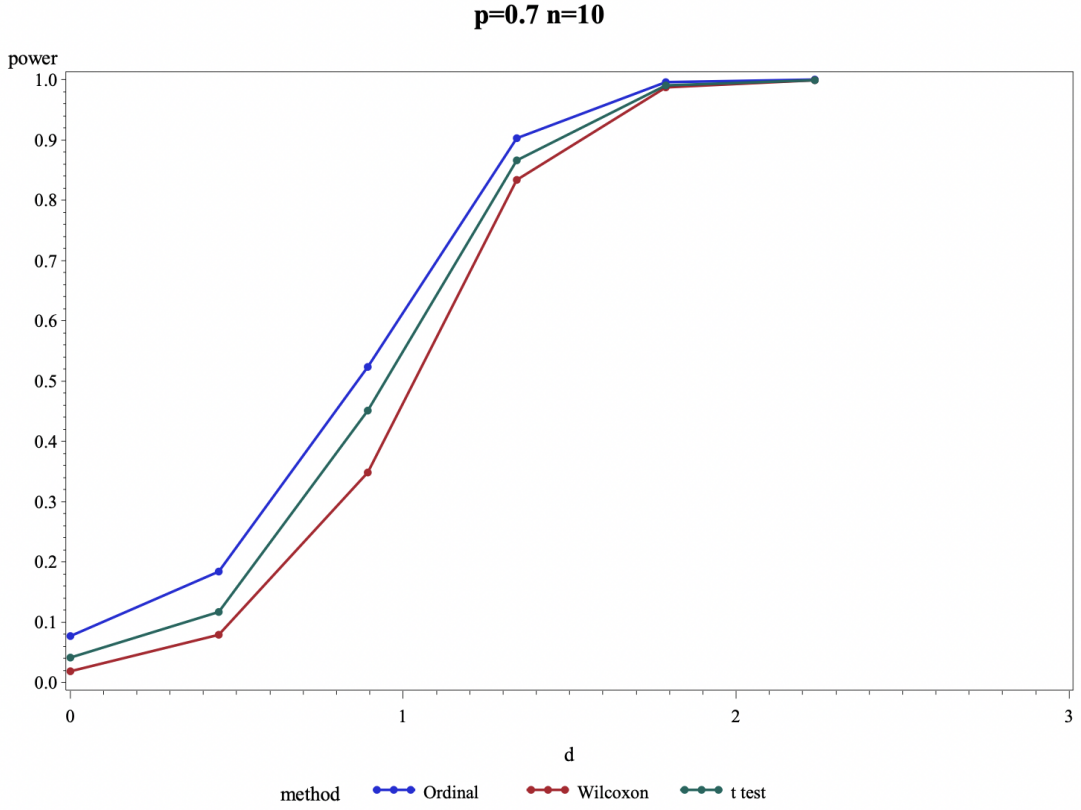


Figure 5: Empirical power functions for the three tests, for skewed data, *n*=10.

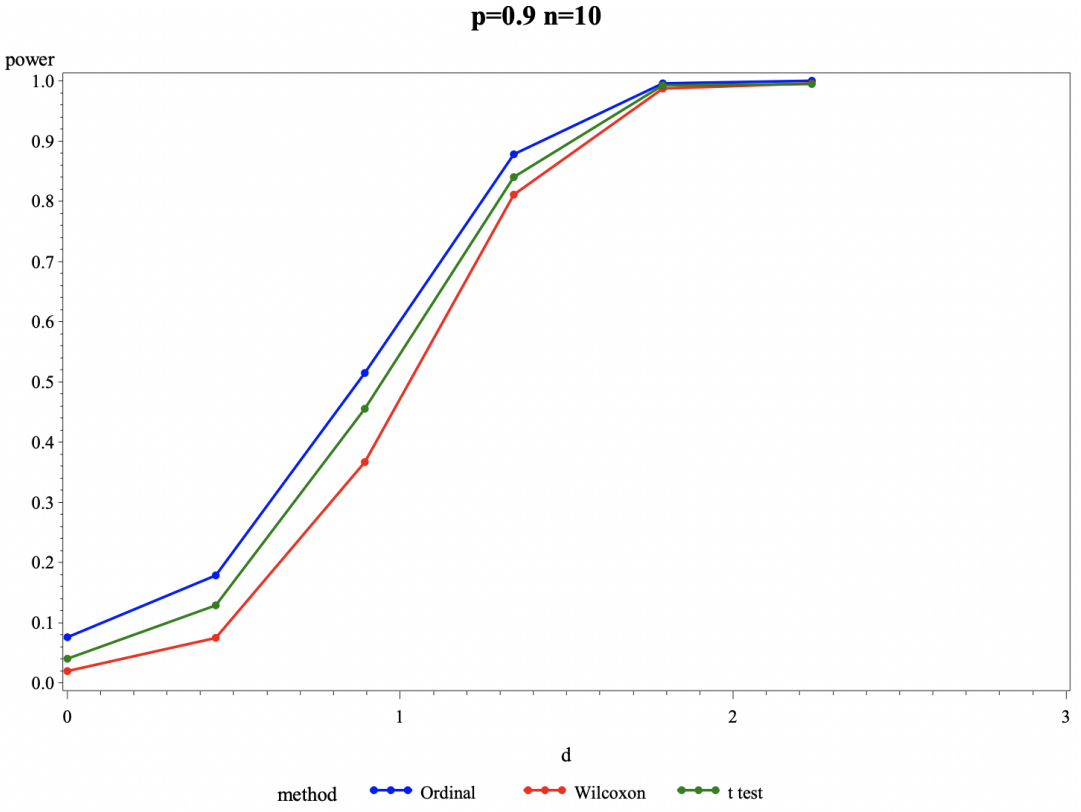


Figure 6: Empirical power functions for the three tests, for symmetric data, *n*=30.

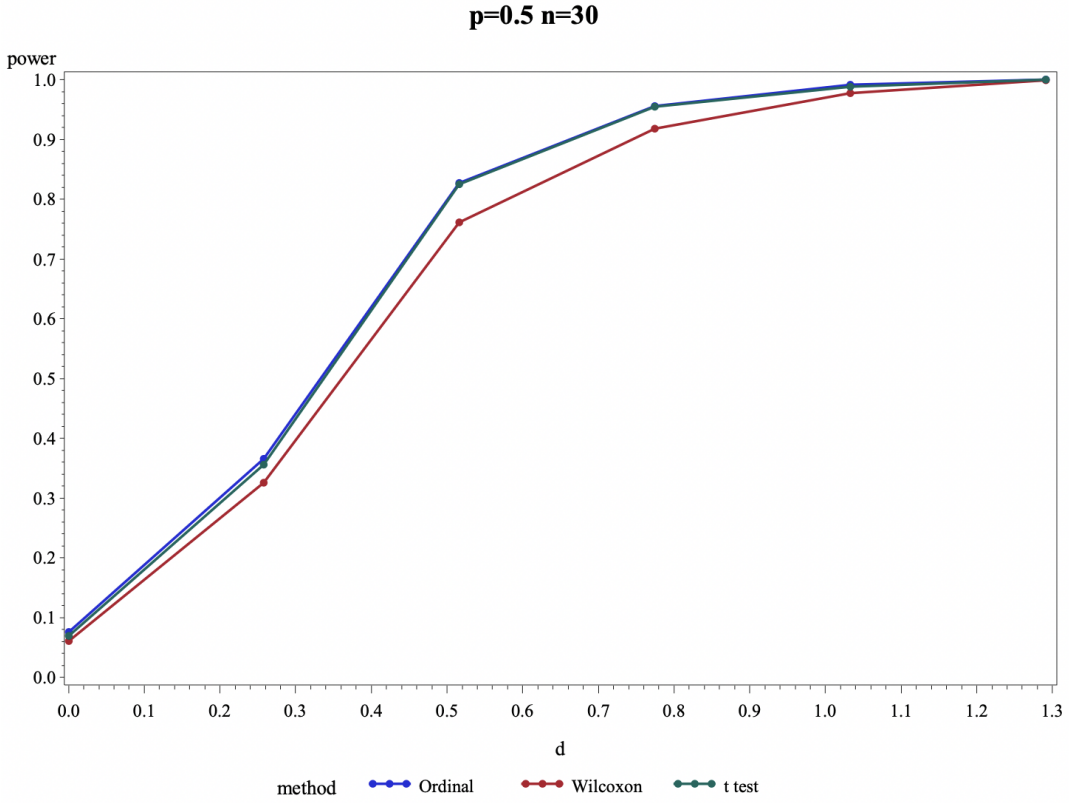


Figure 7: Empirical power functions for the three tests, for moderately skewed data, *n*=30.

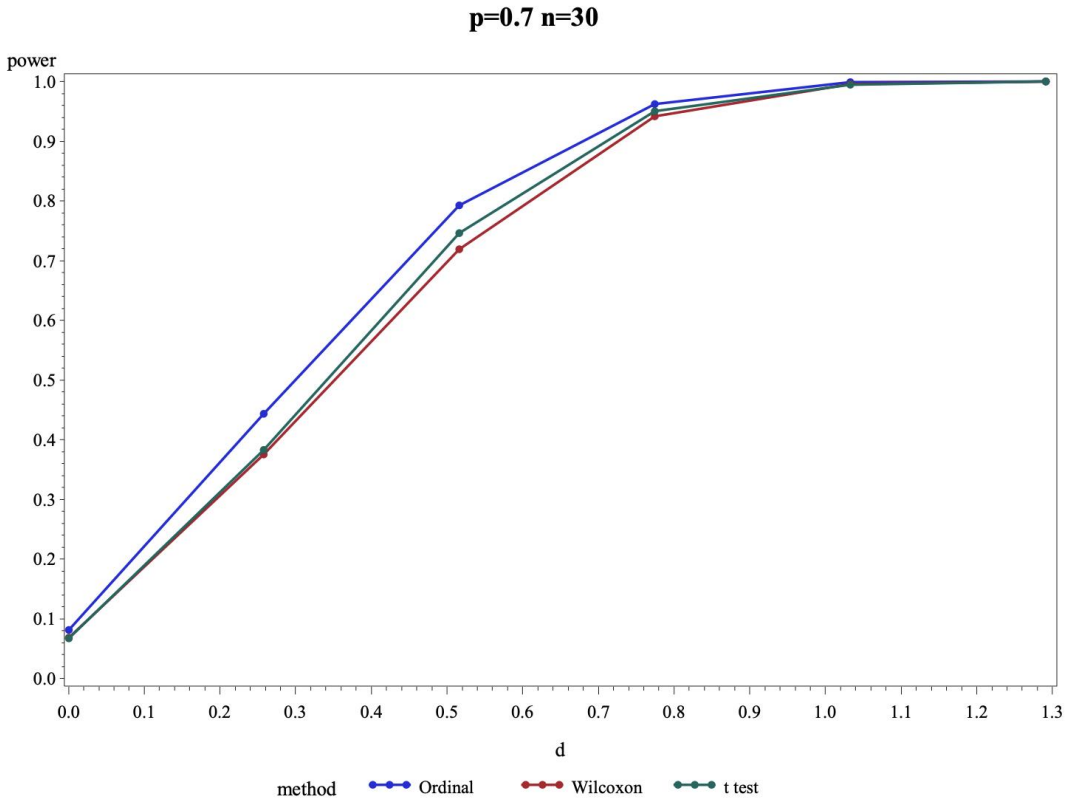
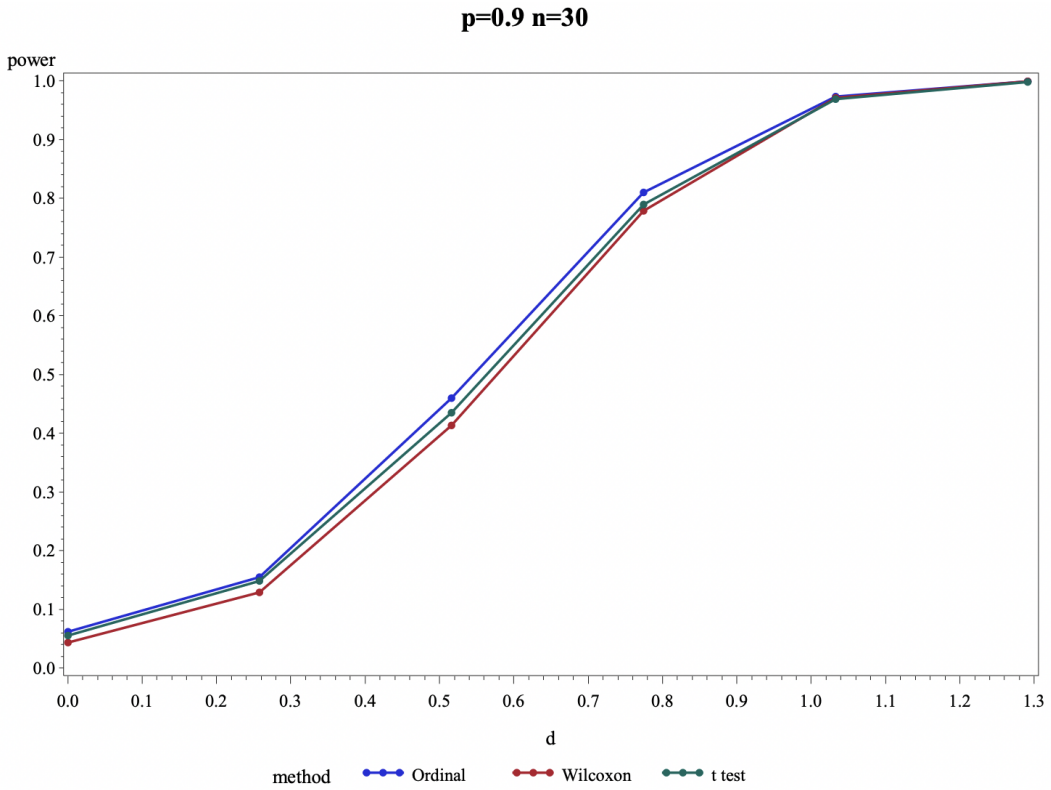


Figure 8: Empirical power functions for the three tests, for skewed data, n=30.



Discussion

The validity of our results depends on whether our simulated data can be regarded as "similar to" real Likert data. Although the study is limited in size, the fact that the results are consistent across experimental conditions suggest that the following conclusions are warranted.

The pattern is similar for all values of n and P: the ordinal probit model has the highest power, followed by the t test. The success of the ordinal probit model is not surprising, since the data were generated to agree with that model. Slightly more surprising is that the t test has higher power than the Wilcoxon test, even for highly skewed data.

Our simulations do not suggest that the ordinal probit model is always superior to t-tests or Wilcoxon tests for Likert data. It does suggest, however, that ordinal probit models work well in situations where the underlying assumptions are fulfilled, even for rather small data sets. The t test has higher power than the Wilcoxon test in all studied situations, even for skewed data.

As a comparison, some of the analyses were repeated using an ordinal logistic model instead of the probit model. This did not change the general results, since the differences in power between the probit and the logit models were minute.

The differences between methods become smaller when the sample sizes increase. For large samples, the choice of method is of minor importance. For smaller samples, our results suggest that the Wilcoxon test does not work well for Likert type data.

References

Bhattacharya, T. and Sengupta, A, (2021). *Large-sample tests for comparing Likert-type scale data, Communications in Statistics - Theory and Methods.*

Blair, R. C., & Higgins, J. J. (1980): A comparison of the power of Wilcoxon's rank-sum statistic to that

of Student's t statistic under various nonnormal distributions. *Journal of Educational Statistics, 5*, 309–335.

Boone H. N., Jr. and Boone, D. A. (2012). Analyzing Likert Data. *Journal of Extension, 50*, 2.

Lehmann E. L. (1975). *Nonparametrics: Statistical methods based on ranks*. Oakland, CA: Holden-Day Inc.

MacDonald, P. (1999). Power, Type I, and Type III error rates of parametric and nonparametric statistical tests. *The Journal of Experimental Education, 67*, 367–379.

McCullagh, P. and Nelder, J. A. (1989). *Generalized linear models*. London: Chapman and Hall.

Neave, H. R., & Granger, C. W. J. (1968). A Monte Carlo study comparing various two-sample tests for differences in mean. *Technometrics, 10*, 509–522.

Norman, G. (2010). Likert scales, levels of measurement and the "laws" of statistics. *Adv in Health Sci Educ 15*, 625–632.

Olsson, U. (1979). On the robustness of factor analysis against crude classification of the observations. *Multivariate Behavioral Research, 14*, 485-500.

SAS Institute Inc. (2017). *SAS/Stat User's Guide. Version 9*.4. Cary, N. C.: SAS Institute Inc.

de Winter, J. and Dodou, D. (2010). Five-point Likert Items: t test versus Mann–Whitney–Wilcoxon. *Practical Assessment, Research and Evaluation, 15* (11).

Citation:

Olsson, U. (2022). Power Properties of Ordinal Regression Models for Likert Type Data. *Practical Assessment, Research & Evaluation*, 27(6). Available online: https://scholarworks.umass.edu/pare/vol27/iss1/6/

Corresponding Author:

Ulf Olsson
Swedish University of Agricultural Sciences
Ultuna, Uppsala, Sweden

E-mail: Ulf.Olsson@slu.se