# Power analysis for testing two independent groups of likert-type data

Zun-xiong Liu[1,a]& Hao Chen[2,b*]

[1]School of Information Engineering, East China Jiaotong University, P.R.China

[2] School of Information Engineering, East China Jiaotong University, P.R.China

[a]email:Darrent.liu@gmail.com, [b]email:chenhao806@126.com, [*]corresponding author

**Keywords:** two-sample problem, nonparametric test, power of test

**Abstract.** In the one-sample location problem, it is tested whether the center of the whole is equal to a known value, otherwise whether there are significant differences between two samples is in consideration on pratical situation. In practical problems, there are many simulations where two general parameters are compared, instead it is tested whether the center of the whole is equal to a known value in the one-sample location problem. The aim of this article is to determine the goodness-of-fit of three different nonparametric tests, which being two sample rank test, Smirnov test ( two-sample Kolmogorov-Smirnov test) and two-sample Cramér-von Mises test. In the meantime the efficacies of their respective comparative analyses are also tested to choose their own two-sample test methods. Simulation results indicate that neither of the tests is the best for each sample distribution, but in most instances, the Cramer-von Mises test performs best. Moreover the Kolmogorov-Smirnov test is better than the Mann-Whitney test in term of distribution of samples, sample size and effect size .

## Introduction

Categorical data is widely used in educational, psychological, and economic and social life. Such as in social attitude surveys and heart tests Likert-type data are used. This article discusses a specific link between the existences of two independent samples, whether to obey the same specific data distribution or not [2]. We chose to test two independent samples by comparison.

Measurement on a continuous scale is sometimes not available, In particular for those variables concerning feelings, attitudes, or opinions. Therefore, researchers create rating instruments according to ordered categories. Thus, one can describe feelings, attitudes or opinions. Rensis Likert's dissertation created a new attitude-scaling technique from a survey of student attitudes.

We have chosen to compare by testing for two independent samples. The statistical methods in this study: the Mann-Whitney test, Kolmogorov-Smirnov test [1] and Cramér-von Mises test [3], will be examined for the robustness and statistical power in different circumstances by the simulated Likert-type data sets. In order to find the appropriate test for each condition, the empirical Type I error rate of all tests will be compared [4].

In the following section the alternative distributions are defined and the simulation and linear interpolation technique used to approximate the powers are discussed. The number of categories is restricted to 10 to enable reasonable examples of the different-shaped alternative distributions. The results of the power studies are presented as follows. A summary of the most powerful test statistic(s) for each of the specified alternative distributions is included in the Conclusions section.

Note: This article is supported by the National Nature Science Foundation of China (No.71361009); Social Science Foundation of the State Education Ministry (No.13YJC63019 ).

 

**Proposed method**

**A. Mann-Whitney test**

The Mann-Whitney measures the probability that observations from two recordings are taken from continuous distributions with equal median.

Using R function wilcox.test{stats}:Performs one and two-sample Wilcoxon tests on vectors of data; the latter is also known as 'Mann-Whitney' test.

Data: The data contains two independent samples, the capacity of a group of n, $X_1$, $X_2$, ..., $X_n$; another group capacity m, $Y_1$, $Y_2$, ..., $Y_m$. N = n + m.

Test statistic:

$$T = \frac{\sum_{i=1}^{n} R(X_i) - n\frac{N+1}{2}}{\sqrt{\frac{nm}{N(N-1)}\sum_{i=1}^{N} R_i^2 - \frac{nm(N+1)^2}{4(N-1)}}}$$

Exactly equal to the value of the two samples into a knot, R () represents the rank assigned to each sample.

**B. Two-sample Kolmogorov-Smirnov test**

The two-sample Kolmogorov-Smirnov test measures the probability that observations from two recordings are taken from the same continuous distribution, measuring the distance between empirical distributions.

Using R Package 'dgof': This package contains a proposed revision to the stats::ks.test() function and the associated ks.test.Rd help page. With one minor exception, it does not change the existing behavior of ks.test(), and it adds features necessary for doing one-sample tests with hypothesized discrete distributions.

Data: The data contains two independent samples, the capacity of a group of n, $X_1$, $X_2$, ..., $X_n$; another group capacity m, $Y_1$, $Y_2$, ..., $Y_m$.

Test statistic: $S_1$ (x) is the empirical distribution function of a sample $X_1$, $X_2$, ..., $X_n$, while $S_2$ (x) is the empirical distribution function of a sample $Y_1$, $Y_2$, ..., $Y_n$.

Define Test statistic $T_1$ as the maximum vertical distance between two empirical distribution functions as follows:

$$T_1 = \sup_x |S_1(x) - S_2(x)|$$

**C. Two-sample Cramér-von Mises test**

The two-sample Cramér-von Mises test measures the probability that observations from two recordings are taken from the same continuous distribution, measuring the goodness-of-fit between empirical distributions.

Using R Package 'cramer': Perform Cramér-test for two-sample-problem. Both unvaried and multivariate data is possible. For calculation of the critical value Monte-Carlo-bootstrap-methods and eigenvalue-methods are available. For the bootstrap access ordinary and permutation methods can be chosen as well as the number of bootstrap-replicates taken.

Data: The data contains two independent samples, the capacity of a group of n, $X_1$, $X_2$, ..., $X_n$; another group capacity m, $Y_1$, $Y_2$, ..., $Y_m$.

Test statistic:

$$T_2 = \frac{mn}{(m+n)^2} \sum_{\substack{x=X_i \\ x=Y_j}} [S_1(x) - S_2(x)]^2$$

Where $S_1(x)$ is the empirical distribution function of a sample $X_1$, $X_{2,\ldots,}X_n$, while $S_2(x)$ is the empirical distribution function of a sample $Y_1$, $Y_2$, ..., $Y_n$. The number m, n corresponds to the size of each sample.

**Simulation experiments of power**

**A. Distribution of samples**

Use the R software simulation to simulate the identification of the Type1 error comparison, as well as test the effectiveness of each distribution. This selection of five-point Likert scale simulation data distribution, the probability distribution of each distribution is as follows [5][6]:

Table 1: Five marginal distributions for the 5-point response scale

| 5-point scale | Uniform | Moderate Skew | Highly Skew | Symmetric | Bimodal |
|---|---|---|---|---|---|
| 1 | 0.2000 | 0.2400 | 0.6561 | 0.0625 | 0.3276 |
| 2 | 0.2000 | 0.4117 | 0.2906 | 0.2500 | 0.1471 |
| 3 | 0.2000 | 0.2636 | 0.0496 | 0.3750 | 0.0506 |
| 4 | 0.2000 | 0.0766 | 0.0046 | 0.2500 | 0.1471 |
| 5 | 0.2000 | 0.0081 | 0.0001 | 0.0625 | 0.3276 |

**B. Sample size**

Calculate using the Monte Carlo simulation method of goodness-of-fit test for each power. The total of twelve sample sizes was examined for the difference of testing groups. The sample sizes chosen for this study were as follows (10, 10), (10, 30), (10, 50), (30, 30), (30, 50), (30, 100), (50, 50), (50, 100), (50, 300), (100, 100), (100, 300), and (300, 300).

**C. Power and levels of effect size**

Statistical power is defined as the probability of rejecting the null hypothesis given the alternative hypothesis is true. In order to evaluate the statistical power of the tests we need to specify the effect size. The effect size refers to the magnitude of the effect of the alternative hypothesis. If the effect size is large enough, the alternative hypothesis will be true and the null hypothesis of equality is false. Therefore, there is a real difference between both testing groups. In this study the effect sizes of 0.30, and 0.50 will be examined.

### D. Significance level

In this study, we define the significance level( α ) as 0.05. At nominal α =0.05, an observed Type I error rate within 25% of this rate, i.e. ., from 0.0375 to 0.0625, is considered robust.

### Results of experiments

Figures of the power of three tests are as follows:

### A. Uniform distribution

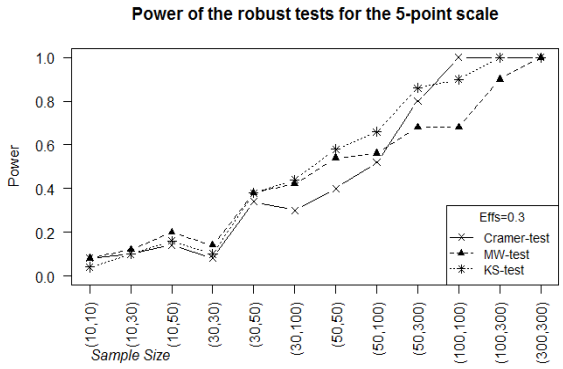


Fig. 1 Uniform distribution(effs=0.03)

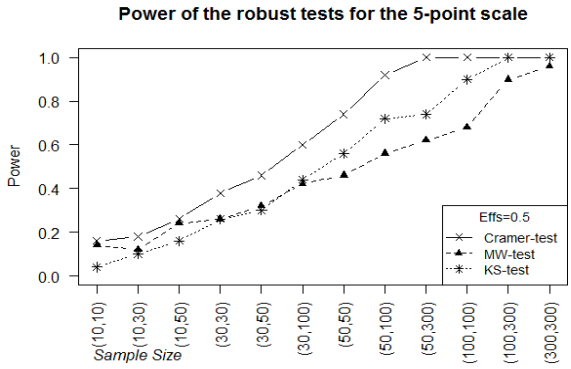


Fig. 2 Uniform distribution(effs=0.05)

Define the effect size as 0.3. For the smaller sizes Figure 1 shows that the Cramér-von Mises test performs worst when the both sample sizes are below 50. As the sample sizes grow bigger, there is a great upward trend in terms of power for the Cramér-von Mises test, followed by the KS test, at last the Mann-Whitney test.

When the effect size is 0.5, power curve of three tests corresponds to three parallel lines approximately in Figure 2, the power of Cramér-von Mises test is highest, then KS test and finally Mann-Whitney test.

### B. Moderate Skew distribution

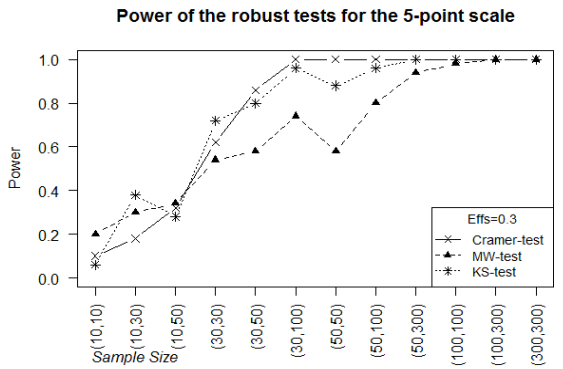


Fig. 3 Moderate skew distribution(effs=0.3)

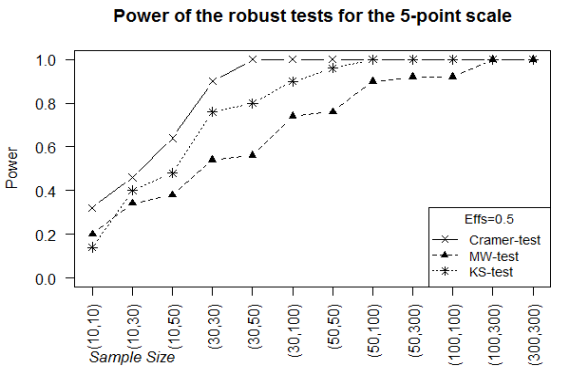


Fig. 4 Highly skew distribution(effs=0.5)

When the effect size equals 0.3, Figure 3 shows that the curves of the three tests seem to be a disorder as the sample size choose the first four selections(the sample size is (10, 10), (10, 30), (10,50)and (30, 30)).When the sample size grows bigger ,the Cramér-von Mises test shows well than the other two tests, the Mann-Whitney test performs worst.

When the effect size is 0.5, obviously the power of Cramér-von Mises test is highest, then KS test and finally Mann-Whitney test.

## C. Highly Skew distribution

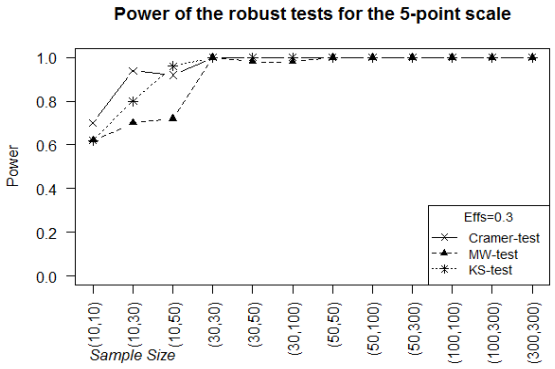


Fig.5 Highly skew distribution(effs=0.3)

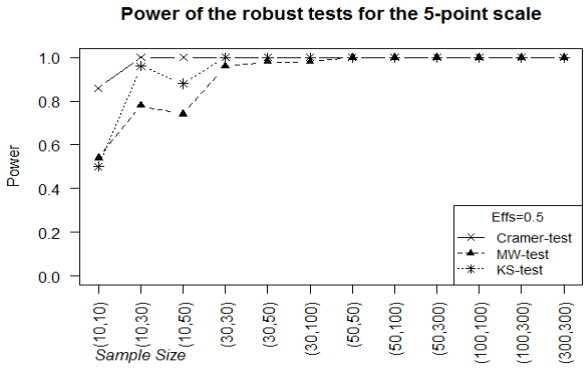


Fig.6 Highly skew distribution(effs=0.5)

By simulation studies, Figure 5 and Figure 6 show that the statistic power of the Cramér-von Mises test is superior to the KS test and Mann-Whitney test both under the effect size=0.3 and 0.5. The power of KS test ranks second among these three tests.

## D. Symmetric distribution

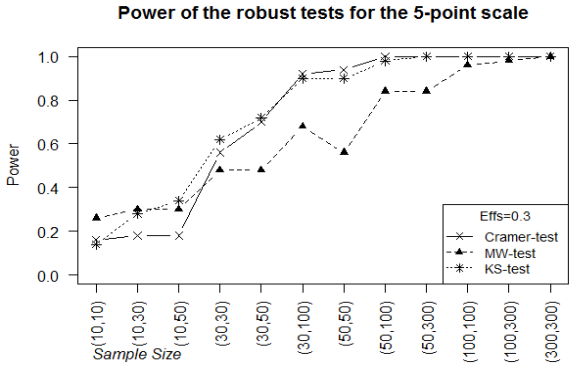


Fig.7 Symmetric distribution(effs=0.3)

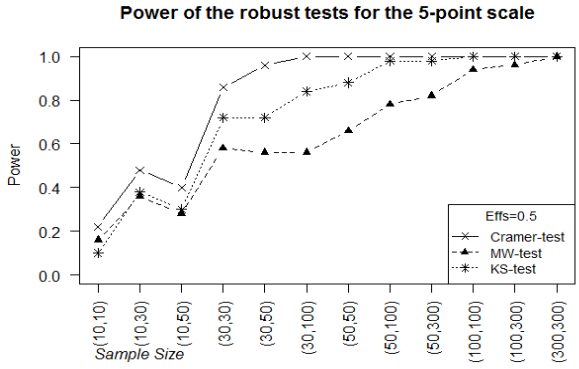


Fig.8 Symmetric distribution(effs=0.5)

In Figure 7, choosing small sample size, such as (10,10) and (10,50),the statistic power of Cramér-von Mises test is lower than the KS test, even lower than the Mann-Whitney test. But as the sample size increases, the power of Cramér-von Mises test returns to normal, to become the highest.

In Figure 8, CVM test contains the efficacy of the highest, KS test, followed, finally, Mann-Whitney test.

## E. Bimodal distribution

We note that Figure 4.5.1 show that the Cramer-von Mises performs the best under effect size=0.3 by simulation studies. We can find that this pattern can also be observed in the estimation of the effect size=0.5,which specified in Figure 4.5.2. When the effect size is 0.3, the statistic power of Cramér-von Mises test is the lowest, also lower than the Mann-Whitney test. When the effect size

increases to 0.5, this situation slows, but does not change the fundamental problem. We can draw that the Cramér-von Mises test is not fit for bimodal distribution.

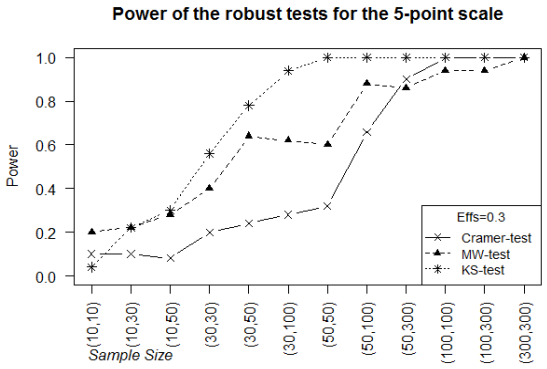


Fig.9 Bimodal distribution(effs=0.3)

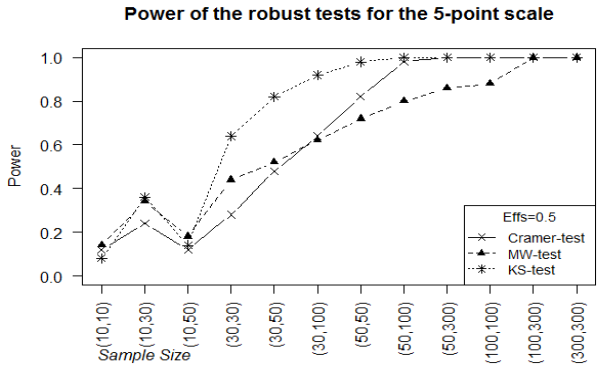


Fig.10 Bimodal distribution(effs=0.5)

## Conclusion and Recommendations

This study obviously indicates that the statistic power will be increased when effect size and sample size are increased. It also shows that none of these three tests can realistically recommended to the applied econometrician as having higher power for all situations. When the effect size is 0.3 or 0.5, the two-sample Cramér-von Mises test seems to perform well than them. But when the alternative distribution is bimodal distribution, the Cramér-von Mises test performs badly. Indeed the difference between Cramér-von Mises test and KS test is subtle, they examine ergonomics similar capacities in this regard considering goodness-of-fit, but the former seems to better good use of data. Undoubtedly, the Mann-Whitney test is improper for the goodness-of-fit for likert-type data.

The results and the summary in the table below can at least give the applied econometrician some guide to the choice of alternative goodness-of-fit test statistics with respect to power.

Table 2: General summary of the power of three tests

| Alternative distribution | Comparison of General Ranking of power |
|---|---|
| Uniform | CVM>KS>MW |
| Moderate Skew | CVM>KS>MW |
| Highly Skew | CVM>KS>MW |
| Symmetric | CVM>KS>MW |
| Bimodal | KS>MW>CVM |

Repeat simulation iterations should be increased, while the value of the effect size of 0.1 or 0.7 can be selected for deeper study.

## References


[1] Choulakian, V. Lockhart , R.A. and Stephens, M.A. (1994). *Cramér-von Mises statistics for discrete distributions*. The Canadian Journal of Statistics, 22,125-137.

[2] Michical Steele and Janet Chaseling, (2006). *Powers of Discrete Goodness-of-Fit Test Statistics for a Uniform Null Against a Selection of Alternative Distributions*. Communications in Statistics—Simulation and Computation, 35,1067–1075.

[3] Pettitt, A.N. and Stephens, M.A. (1977). *The Kolmogorov-Smirnov goodness-of-fit statistic with discrete and grouped data*. Technometrics, 19 205-210.

[4] W.J.Conover. *Practical nonparametric statistics,*Wiley, John & Sons, Incorporated,1998.

[5] Steele, M. *The power of categorical goodness-of-fit test statistics*.[D] Griffith University, Brisbane, Australia.2002.

[6] Gibbons, J. D., & Chakraborti, S. *Nonparametric statistical inference* (3rd ed.). New York: M. Dekker.1992.