An RP-Aided Pearson-Fisher Chi-square Approach to Testing Goodness-of-fit for Location-scale Distributions

Jie Li, Jiajuan Liang, Xiaoling Peng[1]

BNU-HKBU United International College, China

Abstract

The classical Pearson-Fisher chi-square test is a general approach to testing goodness-of-fit for univariate data. There is a considerable amount of discussion on how to effectively apply this test to practical goodness-of-fit problems in the literature. There are two major types of discussions: the number of grouping intervals and the selection of grouping intervals. In this paper, we apply the idea of representative points (RP) of statistical distributions to the selection of grouping intervals for the classical Pearson-Fisher chi-square test. We focus on the location-scale distributions due to their simple distributional structure. Special attention is paid to the three types of location-scale distributions: normal, Laplace, and the logistic distributions. Monte Carlo simulation shows that the RP-aided Pearson-Fisher test outperforms the traditional equiprobable Pearson-Fisher test in most cases under study. Some examples with real data are illustrated for practical application.

AMS classification: 62G10, 62H15

Keywords: Goodness of fit test; Location-scale distributions; Pearson-Fisher test; Representative points

1. Introduction

Since the first paper Pearson (1900) proposed the idea of testing goodness-of-fit, there have been a rich source of literature and a huge amount of discussion on how to construct the Pearson chi-square test and how to effectively apply it to practical prob-

[1]Corresponding author, BNU-HKBU United International College, TangJiaWan, 2000 Jintong Road, Zhuhai 519087, China; Email: xlpeng@uic.edu.cn

lems. Fisher (1924) may be the first paper that pointed out a clear construction of the Pearson test when involving estimated parameters in goodness-of-fit problems. As a result, the classical Pearson chi-square test is generally called Pearson-Fisher chi-square test. Because there exists some uncertainty in constructing Pearson-Fisher chi-square test, statisticians never stop exploring the various possibilities for improving its performance since 1924. It has been almost a century when statisticians brought with a relatively clear view in applying the Pearson-Fisher chi-square test to various applications. Most up-to-date and comprehensive discussions and guidance with rich references are summarized in D'Agostino and Stephens (1986), Greenwood and Nikulin (1996), Voinov, Nikulin, and Balakrishnan (2013), and Witkov and Zenget (2019). The original Pearson's (1910) chi-square test for goodness-of-fit is to test if an identically distributed (i.i.d.) sample $\{x_1, \ldots, x_n\} \in R^1$ (R^1 stands for the one-dimensional Euclidean space) can be considered to come from a population $\mathcal{X}$ with an absolutely continuous distribution function $F(x)$ ($x \in R^1$):

The asymptotic distribution of the Pearson-Fisher statistic (7) is generally not an exact chi-square distribution $\chi^2(k-s-1)$ if the parameter estimator $\hat{\boldsymbol{\theta}}_n$ is obtained from raw data (such as the maximum likelihood estimator) instead of grouped data. This was first discovered by Chernoff and Lehmann (1954) for the case of fixed grouping intervals, and later it was generalized by Moore (1971) and Chibisov (1971) for random grouping cells. For location-scale distributions with $s = 2$ parameters (θ_1, θ_2) [θ_1 stands for the location parameter and $\theta_2 > 0$ for the scale parameter) and the MLE for $\boldsymbol{\theta} = (\theta_1, \theta_2)'$ being the form of $(\bar{x}_n, s_n^2)$ [$\bar{x}_n$ stands for the sample mean and s_n^2 for the sample variance], and the grouping intervals taking the form of $\hat{\theta}_1 + c_i\sqrt{\hat{\theta}_2}$ ($i = 1, \ldots, k$), the asymptotic distribution of the Pearson-Fisher statistic (7) can be expressed by

$$\chi^2_{PF}(\hat{\boldsymbol{\theta}}_n) \to \chi^2(k-3) + \lambda_1 Z_1^2 + \lambda_2 Z_2^2, \tag{8}$$

where $0 < \lambda_j < 1$ ($j = 1, 2$) is a number which is independent of the two parameters (θ_1, θ_2) (see Theorem 2.5 of Greenwood and Nikulin, 1996, page 26), and Z_j's are i.i.d. standard normal $N(0, 1)$ variates that are independent of the first term $\chi^2(k-3)$. Watson (1958) gave an explicit formula for computing λ_j for the normal distribution $N(\mu, \sigma^2)$ as the null hypothesis in (1).

Based on the above property (8) and the fact that the optimal choice for selecting the grouping points c_i's in the grouping intervals $\hat{\theta}_1 + c_i\sqrt{\hat{\theta}_2}$ ($i = 1, \ldots, k$) is still an

open problem. In this paper, we propose using the statistical representative points (RP for short) for the family of location-scale distributions to replace the grouping points c_i's, and carry out a comprehensive Monte Carlo study on the comparison of power performance between our RP-approach and the traditional equiprobable approach to constructing the Pearson-Fisher (PF for short) statistic (7). Section 2 gives a simple introduction to representative points for a continuous probability distribution and the RP-approach to constructing the Pearson-Fisher statistic. Section 3 presents the simulation results on an approximate power comparison between the RP-approach and the traditional equiprobable PF-approach to testing goodness-of-fit for three types of location-scale distributions: the normal distribution, the Laplace distribution, and the logistic distribution. Section 4 demonstrates a simple application of the RP-Pearson-Fisher test and its comparison with some existing methods. Some concluding remarks are summarized in the last section.

2. Constructing the Pearson-Fisher Test with RP-intervals

Let X be a continuous random variable with a probability density function (pdf for short) $f(x)$ ($x \in R^1$). A set of points $\{R_1, \ldots, R_m\}$ satisfying $-\infty < R_1 < \ldots < R_m < +\infty$ is called the set of representative points of X if the mean square error (MSE)

$$MSE(R_1, \ldots, R_m) = \frac{1}{\sigma^2} \int_{-\infty}^{\infty} \min_{1 \le i \le m} (x - R_i)^2 f(x) dx \tag{9}$$

reaches its minimum, where $\sigma^2 = Var(X)$. This definition was first given by Fang and He (1982). Such a set of points is also called the MSE representative points or simply called RP in this paper. It is also called a set of principal points in some references (Flury, 1990). A comprehensive discussion on principal points and their relationship to some important statistical concepts are given by Tarpey and Flury (1996). To find

the RP for any continuous distribution, one can rewrite equation (9) as

$$\begin{aligned} MSE(R_1,\ldots,R_m) &= \frac{1}{\sigma^2}\int_{-\infty}^{\frac{R_1+R_2}{2}}(x-R_1)^2 f(x)dx \\ &+ \frac{1}{\sigma^2}\int_{\frac{R_1+R_2}{2}}^{\frac{R_2+R_3}{2}}(x-R_2)^2 f(x)dx \\ &+ \ldots \\ &+ \frac{1}{\sigma^2}\int_{\frac{R_{m-1}+R_m}{2}}^{+\infty}(x-R_m)^2 f(x)dx. \end{aligned} \tag{10}$$

The RP can be obtained by doing the partial derivative with regard to each variable in $\{R_1,\ldots,R_m\}$ and solve a set of nonlinear equations. Fang and He (1982) derived the RP for standard normal $N(0,1)$ for $m = 1,\ldots,31$. Theoretically, RP from any continuous distribution can be found by using the algorithm given by Zoppé (1995). But for any location-scale distribution with a pdf of the form $\frac{1}{\sigma}f\left(\frac{x-\mu}{\sigma}\right)$ ($\sigma > 0$ and $f(\cdot)$ is a known function), based on the result in Matsuuraa, Kurata, and Tarpey (1995), one only needs to find the RP for the "standard distribution" $f(\cdot)$ and then use

$$\hat{R}_j = \hat{\mu} + R_j\hat{\sigma}, \quad j = 1,\ldots,m \tag{11}$$

to estimate the the RP for $\frac{1}{\sigma}f\left(\frac{x-\mu}{\sigma}\right)$, where $\{R_1,\ldots,R_m\}$ are the RP for the "standard distribution" $f(\cdot)$, for example,

$$\begin{aligned} f(x) &= \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right),\ x\in R^1,\ \text{for normal } N(\mu,\sigma^2), \\ f(x) &= \frac{1}{2}\exp(-|x|),\ x\in R^1,\ \text{for Laplace distribution}, \\ f(x) &= \frac{\exp(-x)}{1+\exp(-x)},\ x\in R^1,\ \text{for logistic distribution}, \end{aligned} \tag{12}$$

and $\hat{\mu}$ and $\hat{\sigma}$ are (asymptotic) unbiased estimates for the location parameter μ and the scale parameter σ, respectively. The RP-approach to constructing the the Pearson-Fisher test for goodness-of-fit of the location-scale distribution $\frac{1}{\sigma}f\left(\frac{x-\mu}{\sigma}\right)$ is based on the grouping intervals:

$$\begin{aligned} \hat{J}_1 = \left(-\infty, \frac{\hat{R}_1+\hat{R}_2}{2}\right),\ \hat{J}_2 = \left[\frac{\hat{R}_1+\hat{R}_2}{2}, \frac{\hat{R}_2+\hat{R}_3}{2}\right),\ldots, \\ \hat{J}_{m-1} = \left[\frac{\hat{R}_{m-2}+\hat{R}_{m-1}}{2}, \frac{\hat{R}_{m-1}+\hat{R}_m}{2}\right),\ \hat{J}_m = \left[\frac{\hat{R}_{m-1}+\hat{R}_m}{2}, +\infty\right). \end{aligned} \tag{13}$$

where $\{\hat{R}_j : \ j = 1, \ldots, m\}$ are given by (11). The equiprobable grouping points (Mann and Wald, 1942) $\{a_i : \ i = 1, \ldots, m-1\}$ for the "standard distribution" $f(\cdot)$ are given by

$$\int_{-\infty}^{a_1} f(x)dx = \frac{1}{m}, \ \int_{a_1}^{a_2} f(x)dx = \frac{1}{m}, \ \ldots, \ \int_{a_{m-1}}^{+\infty} f(x)dx = \frac{1}{m}. \tag{14}$$

Then the PF-approach to constructing the the Pearson-Fisher test for goodness-of-fit of the location-scale distribution $\frac{1}{\sigma} f\left(\frac{x-\mu}{\sigma}\right)$ is based on the grouping intervals:

$$\hat{I}_1 = (-\infty, \hat{a}_1), \ \hat{I}_2 = [\hat{a}_1, \hat{a}_2), \ldots, \hat{I}_m = [\hat{a}_{m-1}, +\infty), \tag{15}$$

where

$$\hat{a}_j = \hat{\mu} + a_j \hat{\sigma}, \quad j = 1, \ldots, m-1. \tag{16}$$

The following section will present the power comparison between the RP-approach and the equiprobable PF-approach.

3. Monte Carlo Studies on Testing Goodness-of-fit for Location-scale Distributions

In this section, some monte carlo studies are designed to study the performance of the above RP based chi-square goodness-of-fit test. Three popular location-scale distributions including normal distribution, logistic distribution and Laplace distribution are employed as null hypothesis of the goodness-of-fit tests. Various alternative distribtuinos were considered so the power of our method can be compared among three alternative types: symmetric fat-tailed distribution, symmetric thin-tailed distribution and asymmetric distribution.

In each study, range of sample size is adjusted so that we can observe the complete change of test power. Meanwhile, according to our previous simulation results, for normality test we set the number of intervals $m = 10, 20, 30$; meanwhile when testing the other two distributions, the number of intervals $m = 6, 8, 10$. All simulations are performed in R 3.6.1.

3.1. Normality test

Here we use three methods to determine the MSE representative points and their intervals based on asymptotic distributions under null hypothesis.

Method 1. Estimating two normal parameters using MLE

Let $\{x_1, \ldots, x_n\}$ be an i.i.d. sample from a population with an unknown distribution function $F(x)$ $(x \in R^1)$. Testing normal goodness-of-fit is to test the hypothesis

$$H_0 : F(x) \in N(\mu, \sigma^2) \quad \text{versus} \quad H_1 : F(x) \notin N(\mu, \sigma^2), \tag{17}$$

where μ and $\sigma > 0$ are unknown mean and standard deviation of the normal distribution. Because normal $N(\mu, \sigma^2)$ is a location-scale distribution family, we can employ the RP-approach to construct the Pearson-Fisher test using the grouping intervals (13) with

$$\hat{\mu} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad \hat{\sigma} = s_n \text{ with } s_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

and $\{R_1, \ldots, R_m\}$ being the m RP from the standard normal $N(0, 1)$ that is given in Fang and He (1982).

Method 2. Estimating one normal parameter using MLE

Apply the Helmert transformation (Mardia, 1980) to the original sample $\{x_1, \ldots, x_n\}$:

$$y_j = \frac{x_1 + \ldots + x_j - jx_{j+1}}{\sqrt{j(j+1)}}, \quad j = 1, \ldots, n-1. \tag{18}$$

If $\{x_1, \ldots, x_n\}$ is an i.i.d. sample from $N(\mu, \sigma^2)$, then $y_j : \ j = 1, \ldots, n-1$ is an i.i.d. sample from $N(0, \sigma^2)$. As a result, hypothesis (5) is transferred to

$$H_0 : \ \{y_j : \ j = 1, \ldots, n-1\} \sim N(0, \sigma^2) \tag{19}$$

versus the alternative that hypothesis (19) is not true.

Method 3. Testing normality based on the Studentized transformation

Let

$$T_i = \frac{x_i - \bar{x}}{v_n}, \quad v_n^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2, \tag{20}$$

for $i = 1, \ldots, n$. Under the normal assumption in (5), $\{T_1, \ldots, T_n\}$ are approximately

Table 1. Empirical type I error rates for the RP-χ^2 test for normal $N(0,1)$

	n	30	60	90	120	150	180	210	240	270	300
m = 10	RP-χ^2 (1)	0.0481	0.0497	0.0491	0.049	0.0495	0.0504	0.0522	0.05	0.055	0.0528
	RP-χ^2 (2)	0.0517	0.0527	0.0531	0.0472	0.052	0.0497	0.0544	0.0513	0.0535	0.0528
	RP-χ^2 (3)	0.0166	0.0159	0.016	0.0163	0.0177	0.02	0.0182	0.0182	0.0187	0.0178
	FP χ^2	0.0541	0.0548	0.0549	0.0558	0.0511	0.0586	0.056	0.0588	0.056	0.0556
m = 20	RP-χ^2 (1)	0.0441	0.0471	0.0474	0.0493	0.0506	0.0517	0.0483	0.052	0.0505	0.0509
	RP-χ^2 (2)	0.0559	0.0505	0.0517	0.0489	0.0507	0.0522	0.0496	0.0507	0.053	0.0536
	RP-χ^2 (3)	0.0158	0.0191	0.0203	0.0239	0.0241	0.0255	0.024	0.0258	0.0257	0.0253
	FP-χ^2	0.054	0.0503	0.0541	0.053	0.0553	0.0518	0.0506	0.0522	0.0543	0.0534
m = 30	RP-χ^2 (1)	0.049	0.0525	0.0437	0.0477	0.0494	0.0493	0.0503	0.0535	0.0511	0.0478
	RP-χ^2 (2)	0.0582	0.0551	0.0562	0.0495	0.0479	0.0521	0.05	0.0514	0.0522	0.0551
	RP-χ^2 (3)	0.021	0.0281	0.0227	0.025	0.0291	0.0244	0.0244	0.0305	0.029	0.0258
	FP-χ^2	0.0501	0.0535	0.0492	0.0501	0.0549	0.0526	0.0476	0.0512	0.048	0.0478

i.i.d. and have a Student's t-distribution $t(n-1)$ for large sample size n. As a result, hypothesis (5) is transferred to

$$H_0: \ \{T_i: \ i = 1, \ldots, n\} \sim t(n-1) \tag{21}$$

versus the alternative that hypothesis (21) is not true. With $t(n-1)$ being the null distribution for $\{T_1, \ldots, T_n\}$ in (21), we do not need to estimate the normal parameters. The PF-statistic (7) reduces to asymptotic chi-square distribution $\chi^2(k-1)$ if sample data are grouped into k intervals. The RP-χ^2 test for (21) is based on the RP for $t(n-1)$ given by Zhou and Wang (2016).

The empirical type I error rates for the above three methods based on 2,000 replications of simulation is summarized in Table 1, where the null distribution is chosen as $N(0,1)$ for a simple demonstration, the asymptotic null distribution is chosen as $\chi^2(m-3)$ for method 1, $\chi^2(m-2)$ for method 2, and $\chi^2(m-1)$ for method 3, respectively. The term $\lambda_1 Z_1^2 + \lambda_2 Z_2^2$ in (8) is dropped for computational convenience. Table 1 shows that ~~the type I error rates were well controlled by three methods using the RP-χ^2 test (8) and its asymptotic null distribution $\chi^2(m-3)$ for method 1, $\chi^2(m-2)$ for method 2, and $\chi^2(m-1)$ for method 3, with $\lambda_1 = \lambda_2 = 0$ for all three cases. It can be seen that~~ method 1 and method 2 control their type I error rates quite well at the significance level $\alpha = 0.05$ while method 3 seems to under estimate the type I error rate.

The performance is studied by selecting the following alternative distributions, where 1-3 are symmetric fat-tailed distributions, 4-5 are symmetric thin-tailed distributions and 6-8 are asymmetric distributions.

1) $t(5)$: Student's t-distribution with degrees of freedom 5;

2) Laplace(0,1): Standard Laplace (or double exponential) distribution with pdf

$$f(x) = \frac{1}{2}\exp(-|x|), \quad x \in R^1.$$

3) Standard Cauchy distribution with pdf

$$f(x) = \frac{1}{\pi(1+x^2)}, \quad x \in R^1.$$

4) Kotz(N, r, s): Kotz-type distribution with pdf

$$f(x) = Cx^{2(N-1)}\exp\left(-r|x|^{2s}\right), \quad C \text{ is a normalizing constant,}$$

with $N = 2.5$, $r = 0.5$, $s = 2$, and $N = 4.5$, $r = 0.5$, $s = 4$, respectively.

5) β-generalized normal distribution with a pdf

$$f(x) = \frac{\beta r^{1/\beta}}{2\Gamma(1/\beta)}\exp\left(-r|x|^{\beta}\right), \quad x \in R^1.$$

and $\beta = 5$, $r = 1$.

6) Shifted chi-square distribution: $X - E(X)$ with $X \sim \chi^2(2)$.

7) Log-normal distribution: $X - E(X)$ with $\log(X) \sim N(0, 1)$.

8) Shifted Weibull distribution(λ, k): $X - E(X)$ where $E(X)$ stands for expected value with X having a pdf

$$f(x;\lambda,k) = \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1}\exp\left\{-\left(\frac{x}{\lambda}\right)^{k}\right\}, \quad x \geq 0$$

with $\lambda = 3$ and $k = 2$.

These continuous probability distributions can be easily generated by using MATLAB internal functions or writing MATLAB codes (available upon request). The empirical power from each of these distributions is presented in the Figures 1-3. It can be observed that the RP chi-square test method 1 (light green line) significantly improves the traditional equiprobable chi-square test (red line) for almost all cases.

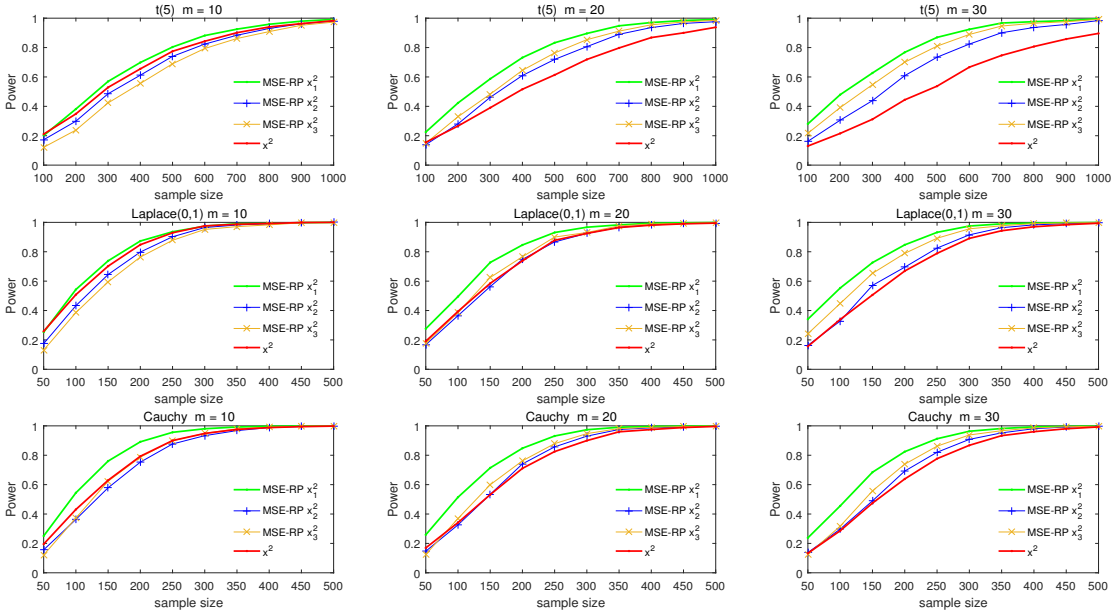

Figure 1. Power comparison in testing normal distribution vs. symmetric fat-tailed distributions

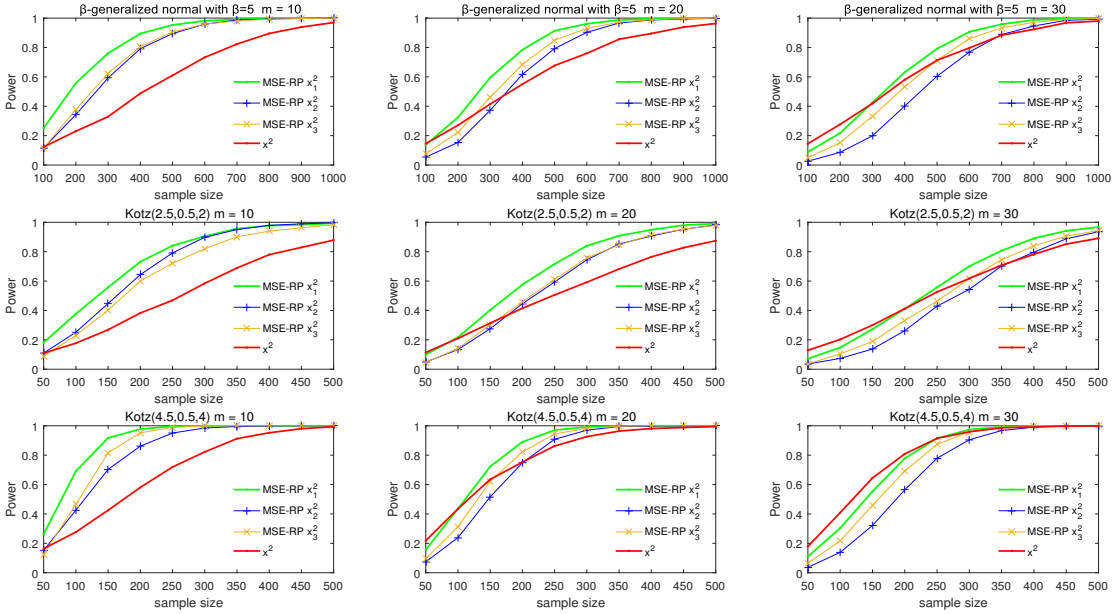

Figure 2. Power comparison in testing normal distribution vs. symmetric thin-tailed distributions

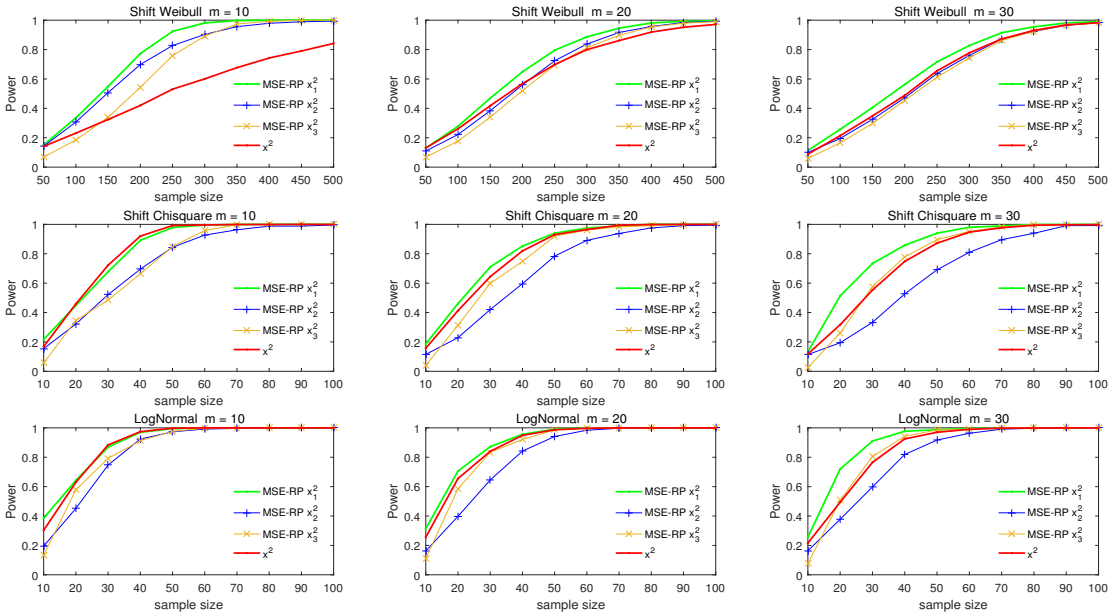

Figure 3. Power comparison in testing normal distribution vs. asymmetric distributions

As shown in Figures 1-3, RP-based method 1 outperforms the other two RP-based methods and FP method in in most normality tests. From Figure 2 we notice that when the data comes from symmetric thin-tailed distributions, the number of intervals should not be too many, as it will lead to the decline of test efficiency. At this time, the superiority of our method is not obvious.

Although methods 2 & 3 are also RP-interval goodness-of-fit approaches, the power of the two tests is not ideal due to different parameter estimation of null distribution. Therefore, these two methods are not taken into account in the later simulation and comparisons.

3.2. Testing goodness-of-fit for the Laplace distribution

A random variable X is said to have a location-scale Laplace distributions if its pdf has the form of $\frac{1}{\sigma} f\left(\frac{x-\mu}{\sigma}\right)$ with

$$f(x) = \frac{1}{2}\exp(-|x|),\ x \in R^1.$$

It is known that median$(X) = E(X) = \mu$ and Var$(X) = 2\sigma^2$. Using $\alpha = 0.05$, the empirical type I error rates for the standard Laplace distribution ($\mu = 0$ and $\sigma = 1$) were obtained from 10,000 simulation runs and summarized in Table 2, where the RP for the Laplace distribution is generated by using the same algorithm as in Fang and He (1982). A limited number of RP from the standard Laplace distribution are given in the Appendix A.

Table 2. Empirical type I error rates for laplace La$(0, 1)$

	n	30	60	90	120	150	180	210	240	270	300
m = 6	RP-χ^2	0.0824	0.0412	0.0694	0.0542	0.0598	0.0662	0.0574	0.0598	0.0602	0.0672
	FP-χ^2	0.0774	0.0722	0.077	0.0806	0.0802	0.0796	0.083	0.084	0.0814	0.0764
m = 8	RP-χ^2	0.0214	0.08	0.0934	0.056	0.0514	0.0506	0.0606	0.059	0.0822	0.0746
	FP-χ^2	0.0592	0.0626	0.0668	0.0692	0.0688	0.0752	0.0656	0.0642	0.0724	0.0712
m = 10	RP-χ^2	0.0126	0.0292	0.0548	0.073	0.0948	0.1198	0.1042	0.0924	0.0788	0.068
	FP-χ^2	0.0586	0.0558	0.0632	0.0604	0.0666	0.065	0.059	0.0656	0.0718	0.0594

The outcome on power performance and a comparison between the RP-invertal (denoted as RP-χ^2) chi-square and the traditional equiprobable chi-square (denoted as FP-χ^2) is presented in Figures 4-7.

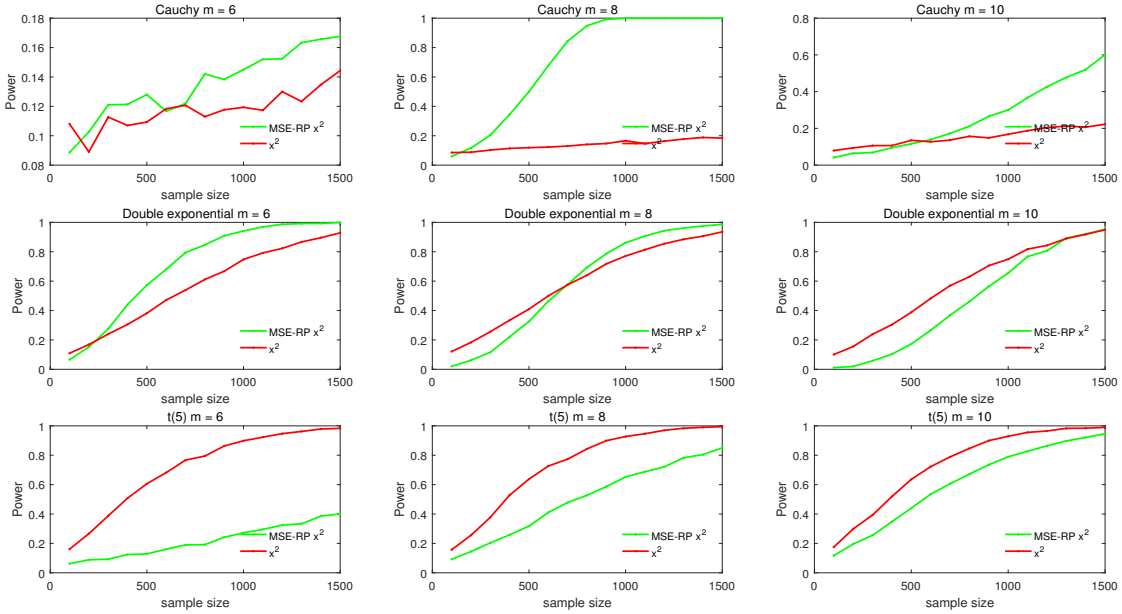

Figure 4. Power comparison in testing Laplace distribution VS symmetric fat tail distributions

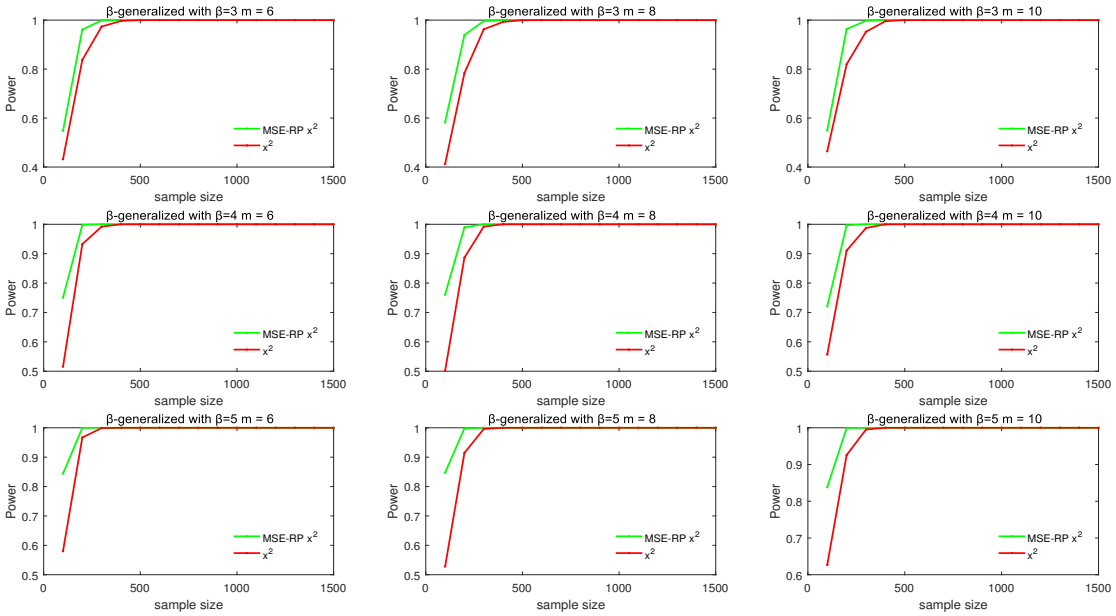

Figure 5. Power comparison in testing Laplace distribution VS symmetric thin tail distributions

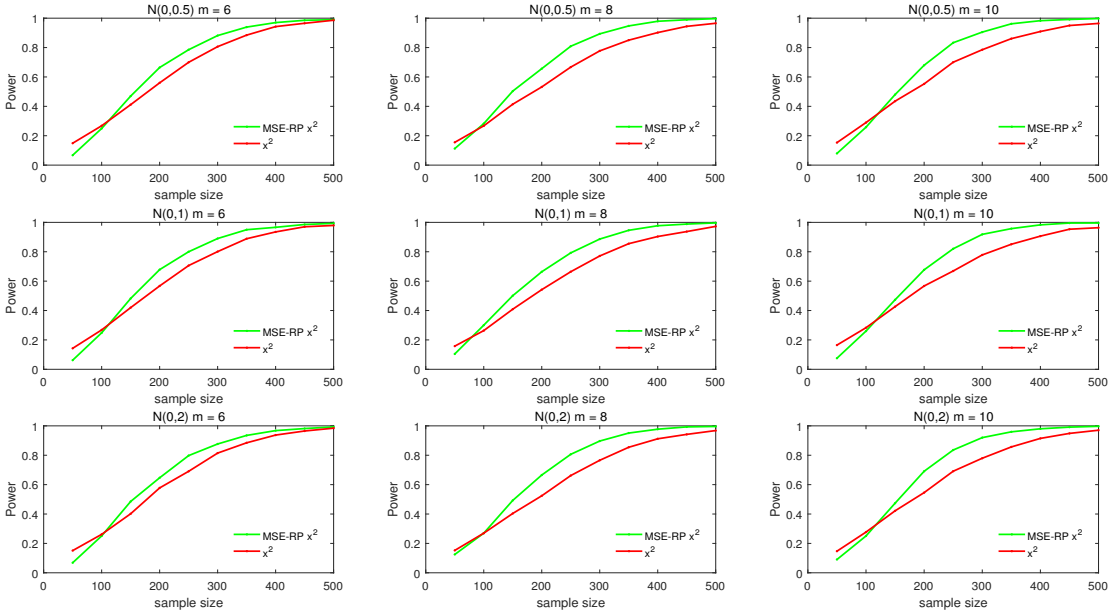

Figure 6. Power comparison in testing Laplace distribution VS asymmetric distributions

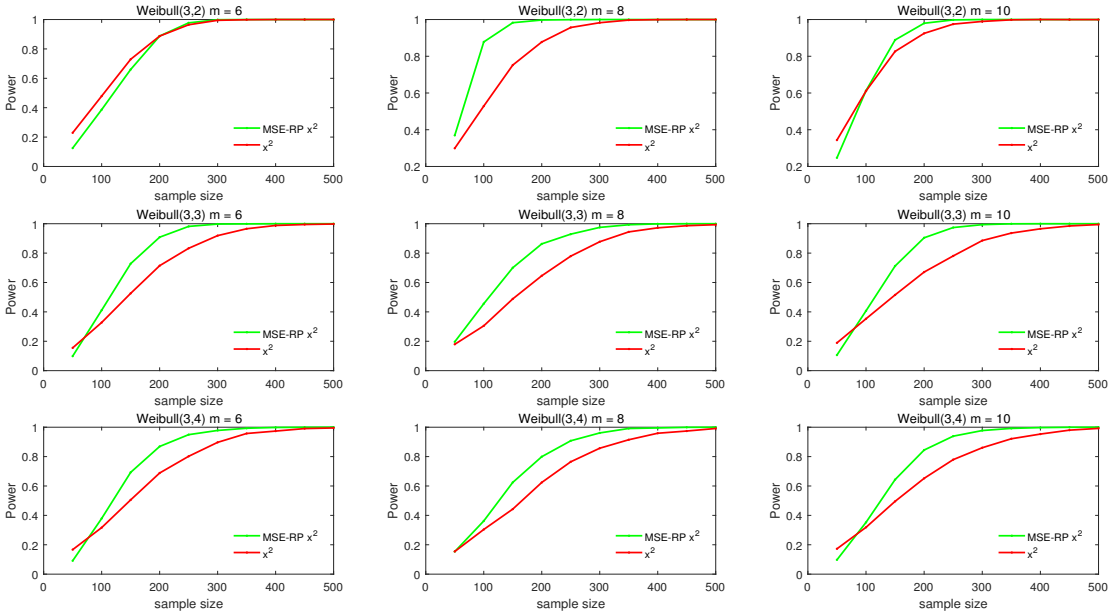

Figure 7. Power comparison in testing Laplace distribution VS asymmetric distributions

The Monte Carlo Study on testing goodness-of-fit for the Laplace distribution is presented in Figures 4-7. Compare with the KS test and AD test, both two chis-square

method perform better. When the alternative distribution is symmetric, the power is similar. But for the asymmetric distribution, the MSE method has better power which means this method has advantage in this kind of distribution.

3.3. Testing goodness-of-fit for the Logistic distribution

A random variable X is said to have a location-scale logistic distributions if its pdf has the form of $\frac{1}{\sigma} f\left(\frac{x-\mu}{\sigma}\right)$ with

$$f(x) = \frac{\exp(-x)}{1+\exp(-x)},\ x \in R^1.$$

It is known that median$(X) = E(X) = \mu$ and Var$(X) = \pi^2\sigma^2/3$. The empirical type I error rates for the standard logistic distribution ($\mu = 0$ and $\sigma = 1$) are summarized in Table 3, where the RP for the logistic distribution is generated by using the same algorithm as in Fang and He (1982) and a limited number of RP from the standard logistic distribution is given in the Appendix.

Table 3. Empirical type I error rates for the MSE RP-χ^2 test for logistic Log$(0, 1)$

	n	30	60	90	120	150	180	210	240	270	300
m = 6	RP-χ^2	0.045	0.0312	0.0442	0.027	0.038	0.0384	0.0352	0.04	0.0402	0.037
	FP-χ^2	0.0434	0.0488	0.0526	0.0476	0.0448	0.0468	0.0482	0.0524	0.0498	0.0516
m = 8	RP-χ^2	0.0182	0.0768	0.0534	0.0348	0.0288	0.0334	0.0432	0.0504	0.047	0.0448
	FP-χ^2	0.0398	0.0386	0.0414	0.048	0.041	0.0434	0.0426	0.046	0.0428	0.0416
m = 10	RP-χ^2	0.006	0.027	0.0486	0.0676	0.091	0.0938	0.0718	0.059	0.0446	0.0398
	FP-χ^2	0.0356	0.0368	0.0412	0.0424	0.0402	0.0358	0.043	0.0422	0.0394	0.0396

The outcome on power performance and a comparison between the RP chi-square and the traditional equiprobable chi-square is presented in Figures 8-12.

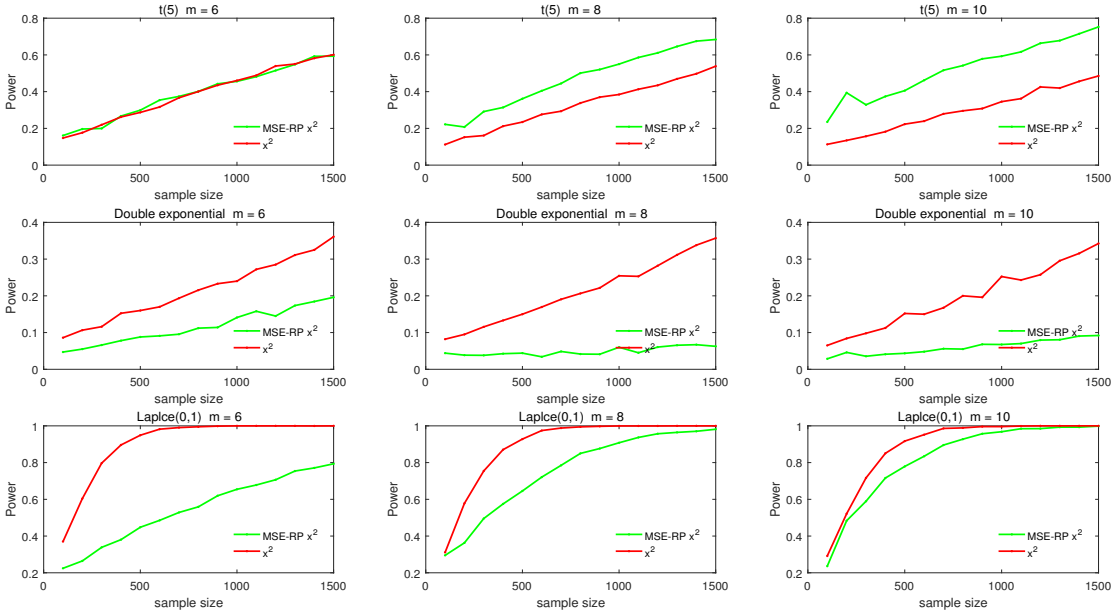

Figure 8. Power comparison in testing Logistic distribution VS symmetric fat tail distributions

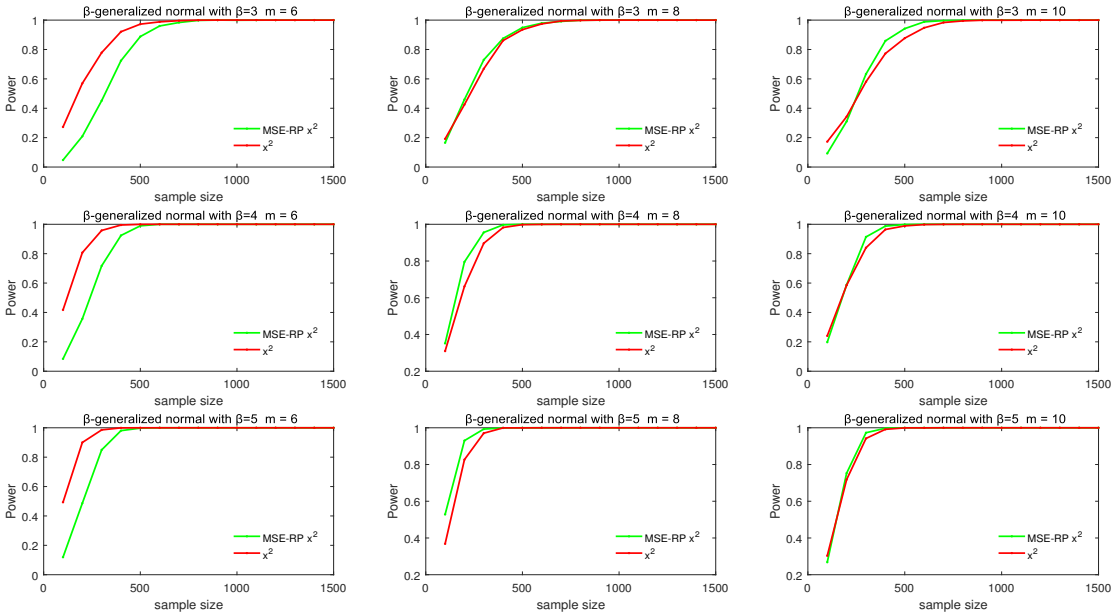

Figure 9. Power comparison in testing Logistic distribution VS symmetric thin tail distributions

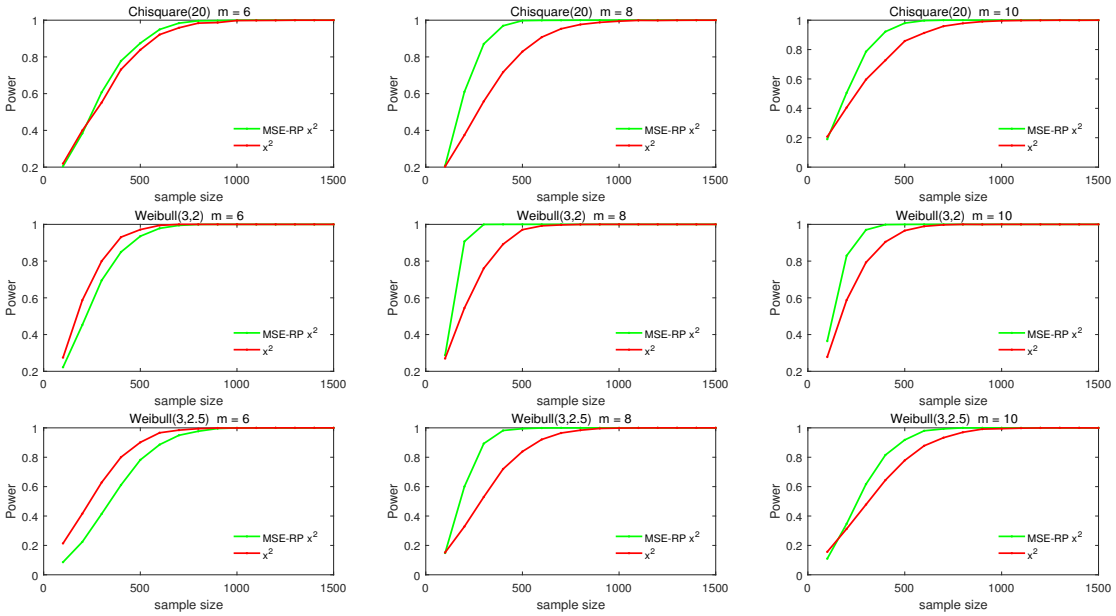

Figure 10. Power comparison in testing Logistic distribution VS symmetric thin tail distributions

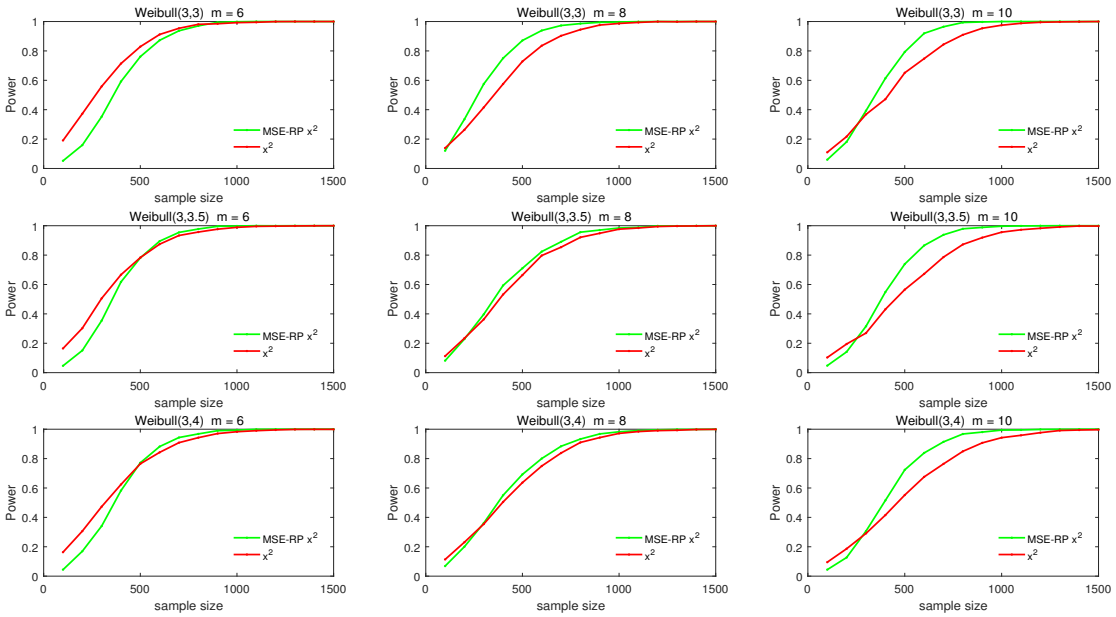

Figure 11. Power comparison in testing Logistic distribution VS asymmetric distributions

The Monte Carlo Study on testing goodness-of-fit for the Laplace distribution is presented in Figures 8-12. Figure 8 presented the fat tails distribution reveals that the

Table 4. P-values of testing normal distribution in Example 4.1

	m = 6	m = 8	m = 10
RP-χ^2	0.6316	0.8765	0.9553
FP-χ^2	0.7812	0.8145	0.9577

power doesn't has a clear pattern. Figures 9-12 presented the thin tails distribution and asymmetric distribution which reveals that the MSE method has better power than the traditional chi-square method. Hence this method has advantage with the symmetric thin-tailed distribution and asymmetric distribution.

4. Three Illustrative Examples

In this section, three real data sets were employed to illustrate the efficacy of our method in testing normal distribution, logistic distribution and laplace distribution, respectively.

Example 4.1: testing for normal distribution. Listed below gives 29 values of Cavendish's determinations ofthe density of earth (ref)".

5.5,5.55,5.57,5.34,5.42,5.3,5.61,5.36,5.53,5.79,5.47,5.75,4.88,5.29,5.62,
5.10,5.63,5.68,5.07,5.58,5.29,5.27,5.34,5.85,5.26,5.65,5.44,5.39,5.46

Example 4.2: testing logistic distribution. Here we use the Representative method and original x^2- test to test a real data which come from the M.H Tahir (2014). According to this paper, the data follow a McDonald Log-Logistic distribution but not a Logistic distribution. Hence we can use our method to test its conclusion.

Table 5. P-values for testing logistic distribution in Example 4.2

	m = 6	m = 8	m = 10
RP-χ^2	0.0049	0.0951	0.0398
FP-χ^2	0.0941	0.1297	0.1625

From the above result we can see that when we set the proper classification number, then our method can reject the hypothesis that the data following a Logistic distribution while the original method could not.

Table 6. P-values for testing logistic distribution in Example 4.3

	m = 6	m = 8	m = 10
RP-χ^2	0.0966	0.0314	0.0824
RP-χ^2	0.3557	0.0403	0.9534

Example 4.3: testing laplace distribution. Here we use the Representative method and original x^2- test to test a real data which come from the Pedro (2000). In this reference, it provide a real data of Yarn breaking strength which is a Laplace distribution:

62,66,78,79,80,84,84,85,85,86,86,87,88,88,89,89,91,91,91,91,92,92,92,92,93,94,94,94,95,
95,95,96,96,96,96,96,97,97,97,97,97,97,98,98,98,98,98,98,98,99,99,99,99,99,100,100,100,
100,100,101,101,101,101,102,102,102,102,102,102,102,103,103,103,103,104,104,104,104,
104,104,105,105,106,107,107,109,110,111,111,111,114,115,117,122,132,132,137,137,138

The parameters of the laplace distribution are estimated by based on above data

$$\hat{\mu} = \frac{1}{2}(x_{n/2} + x_{n/2+1}) = 99,\ \hat{b} = \frac{1}{2}\sum_{i=1}^{n} |x_i - \hat{\mu}| = 8.32$$

Then we apply the two method to this data, and get the following result:
From Table 6 we can see that both methods are fail to reject the null hypothesis which suggest that the data of Yarn breaking strength do follow a laplace distribution.

5. Concluding Remarks

References

Birch, M. W. (1964). A new proof of the Pearson-Fisher theorem. *Ann. Math. Statist.*, **35**, 817-824.

Chernoff, H. and Lehmann, E. L. (1954). The use of maximum likelihood estimates in tests for goodness of fit. *Ann. Math. Statist.*, **25**, 579-589.

Chibisov, D. M. (1971). Some chi-squared type criteria for continuous distribution. *Theory of Probability and its Applications*, **16**, 4-20.

Cox, D. R. (1957). Note on Grouping. *J. Amer. Statist. Assoc.*, **52**, 543-547.

D"Agostino, R. B. and Stephens, M. A (1986). *Goodness-of-fit Techniques*, Statistics: Textbooks and Monographs. Marcel Dekker, New York.

Dahiya, R. C., Gurland, J. (1972). Pearson chi-squared test of fit with random intervals. *Biometrika*, **59**, 147-153.

Fang, K. T. and He, S. D. (1982). The problem of selecting a given number of representative points in a normal population and a generalized Mills ratio. *Stanford Technical Report* No. 327.

Fisher, R. A. (1924). The conditions under which χ^2 measures the discrepancy between observation and hypothesis. *J. R. Statist. Soc.*, **87**, 442-450.

Flury, B. A. (1990). Principal points. *Biometrika*, **77**, 33-41.

Greenwood, P. E. and Nikulin, M. S. (1996). *A Guide to Chi-Square Testing*. John Wiley & Sons.

Hamdan, M. A. (1963). The number and width of classes in the chi-square test. *J. Amer. Statist. Assoc.*, **58**, 678-689.

Koehler,K. J. and Larntz, K. (1980). An empirical investigation of goodness-of-fit statistics for sparse multinomials. *J. Amer. Statist. Assoc.*, **75**, 336-344.

Mann, H. B. and Wald, A. (1942). On the choice of the number of class intervals in the application of the chi-square test. *Ann. Math. Statist.*, **13**, 306-317.

Mardia, K. V. (1980). Tests of univariate and multivariate normality. Krishnaiah, P. R. ed. *Handbook of Statistics*, **Vol. 1**, North-Holland Publishing Company, 279-320.

Matsuuraa, S., Kurata, H., and Tarpey, T. (1995). Optimal estimators of principal points for minimizing expected mean squared distance. *J. Statist. Plann. Infer.*, **167**, 102-122.

Moore, D. S. (1971). A chi-square statistic with random cell boundaries. *Ann. Math. Statist.*, **42**, 147-156.

Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine*, **50**, 157-175.

Rayner, J. C. W. and Best, D. J. (1982). The choice of class probabilities and number of classes for the simple χ^2 goodness of fit test. *Sankhya: The Indian Journal of*

Statistics (Series B), **44**, 28-38.

Tarpey, P. and Flury, B. A. (1996). Self-consistency: a fundamental concept in statistics. *Statistical Science*, **11**, 229-243.

Tate, M. W. and Hyer, L. A. (1973). Inaccuracy of the χ^2 test of goodness of fit when expected frequencies are small. *J. Amer. Statist. Assoc.*, **68**, 836-848.

Voinov, V., Nikulin, M. S., and Balakrishnan, N. (2013). *Chi-Squared Goodness of Fit Tests with Applications*. Academic Press.

Witkov, C. and Zenget, K. (2019). *Chi-Squared Data Analysis and Model Testing for Beginners*. Oxford University Press.

Zhou, M. and Wang, W. (2016). Representative points in t distribution and its application. *Acta Mathematica Applicatae Sinica*, **39** (4), 620-640 (In Chinese)

Zoppé, A. (1995). Principal points of univariate continuous distributions. *Statistics and Computing*, **5**, 127-132.

Appendix A. Representative points of Laplace distribution

Table 3.3: MSE representative point of Laplace distribution $m = [2, 30]$

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
m = 2	1														
m = 3	2														
m = 4	0.593625	2.593623													
m = 5	1.18725	3.187247													
m = 6	0.423954	1.611202	3.611202												
m = 7	0.847907	2.035156	4.035156												
m = 8	0.330079	1.177986	2.365234	4.365234											
m = 9	0.660158	1.508065	2.695313	4.695313											
m = 10	0.270351	0.930508	1.778415	2.965664	4.965664										
m = 11	0.540702	1.200859	2.048766	3.236015	5.236015										
m = 12	0.228969	0.769671	1.429828	2.277735	3.464983	5.464983									
m = 13	0.457938	0.99864	1.658797	2.506704	3.693952	5.693952									
m = 14	0.198593	0.656531	1.197233	1.85739	2.705298	3.892546	5.892547								
m = 15	0.397187	0.855125	1.395826	2.055984	2.903891	4.091139	6.09114								
m = 16	0.175343	0.572529	1.030467	1.571169	2.231326	3.079233	4.266482	6.266482							
m = 17	0.350685	0.747872	1.20581	1.746511	2.406669	3.254576	4.441825	6.441825							
m = 18	0.156971	0.507656	0.904843	1.362781	1.903483	2.56364	3.411547	4.598796	6.598796						
m = 19	0.313942	0.664627	1.061814	1.519752	2.060454	2.720611	3.568518	4.755767	6.755767						
m = 20	0.142087	0.456029	0.806715	1.203901	1.661839	2.202541	2.862698	3.710606	4.897854	6.897854					
m = 21	0.284175	0.598117	0.948802	1.345988	1.803927	2.344628	3.004786	3.852693	5.039941	7.039942					
m = 22	0.129783	0.413958	0.7279	1.078585	1.475772	1.93371	2.474412	3.134569	3.982476	5.169725	7.169725				
m = 23	0.259567	0.543741	0.857684	1.208369	1.605555	2.063493	2.604195	3.264353	4.11226	5.299508	7.299509				
m = 24	0.119442	0.379009	0.663183	0.977125	1.327811	1.724997	2.182935	2.723637	3.383794	4.231701	5.418949	7.418949			
m = 25	0.238884	0.498451	0.782625	1.096567	1.447253	1.844439	2.302377	2.843079	3.503236	4.351143	5.538391	7.538391			
m = 26	0.110628	0.349512	0.609078	0.893253	1.207195	1.55788	1.955067	2.413005	2.953707	3.613864	4.461771	5.64902	7.649019		
m = 27	0.221255	0.460139	0.719706	1.003881	1.317823	1.668508	2.065695	2.523633	3.064334	3.724492	4.572399	5.759647	7.759647		
m = 28	0.103026	0.324281	0.563165	0.822732	1.106906	1.420848	1.771534	2.16872	2.626658	3.16736	3.827517	4.675424	5.862673	7.862672	
m = 29	0.206051	0.427306	0.66619	0.925757	1.209932	1.523874	1.874559	2.271746	2.729684	3.270385	3.930543	4.77845	5.965698	7.965698	
m = 30	0.096401	0.302452	0.523708	0.762592	1.022159	1.306333	1.620275	1.970961	2.368147	2.826085	3.366786	4.026944	4.874851	6.062099	8.062098

Appendix B. Representative points of Logistic distribution

Table 3.2: MSE representative point of Logistic distribution $m = [2, 30]$

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
m = 2	1.386294														
m = 3	2.289109														
m = 4	0.828806	2.971407													
m = 5	1.441413	3.522743													
m = 6	0.593142	1.935422	3.986331												
m = 7	1.063273	2.352669	4.386679												
m = 8	0.462204	1.457016	2.715349	4.739209											
m = 9	0.844945	1.798062	3.036863	5.054269											
m = 10	0.378741	1.174069	2.100135	3.325994	5.339073										
m = 11	0.701885	1.464318	2.371981	3.588916	5.599										
m = 12	0.320846	0.985263	1.724873	2.619542	3.830112	5.838035									
m = 13	0.600618	1.23866	1.961876	2.847095	4.053007	6.059358									
m = 14	0.278327	0.849698	1.468539	2.179641	3.057813	4.260237	6.265418								
m = 15	0.525061	1.07491	1.679384	2.381321	3.254122	4.453871	6.458134								
m = 16	0.245768	0.747384	1.280975	1.874461	2.569329	3.437968	4.635643	6.639209							
m = 17	0.46648	0.950242	1.471282	2.056201	2.74552	3.610875	4.806904	6.809894							
m = 18	0.220034	0.66731	1.137187	1.648364	2.226499	2.911403	3.77414	4.968861	6.971418						
m = 19	0.419711	0.851957	1.310844	1.814145	2.386831	3.068172	3.928789	5.122445	7.124636						
m = 20	0.199182	0.602878	1.023158	1.473214	1.970144	2.5384	3.216835	4.07572	5.268507	7.270406					
m = 21	0.381495	0.772375	1.182986	1.625853	2.117567	2.682181	3.35822	4.215673	5.407738	7.409385					
m = 22	0.181942	0.54988	0.930357	1.333056	1.770003	2.257403	2.818994	3.49304	4.349301	5.540765	7.542206				
m = 23	0.349677	0.706565	1.07849	1.474642	1.906667	2.390465	2.949526	3.621901	4.47716	5.668122	7.669392				
m = 24	0.16745	0.505505	0.853274	1.218092	1.608774	2.036668	2.517435	3.074361	3.745324	4.599738	5.790275	7.791407			
m = 25	0.322773	0.651203	0.991366	1.350221	1.736292	2.160692	2.638885	3.193997	3.863756	4.717447	5.907617	7.908617			
m = 26	0.155097	0.467795	0.788174	1.121932	1.475743	1.857897	2.279319	2.755316	3.308875	3.977604	4.830677	6.020535	8.021428		
m = 27	0.299719	0.603963	0.91754	1.245861	1.59537	1.974173	2.393042	2.867151	3.419373	4.087216	4.939755	6.129345	8.130148		
m = 28	0.14445	0.435346	0.73243	1.040215	1.363885	1.709702	2.08562	2.502284	2.974762	3.525828	4.192902	5.044979	6.234336	8.235061	
m = 29	0.279745	0.563166	0.854137	1.156953	1.476617	1.819244	2.192663	2.607415	3.078475	3.628535	4.29494	5.146616	6.335772	8.336433	
m = 30	0.135157	0.407125	0.68414	0.969852	1.268377	1.584571	1.924425	2.295669	2.708752	3.178572	3.727751	4.39357	5.24489	6.433863	8.434449