Some Improved χ^2-test for Normality Based on Representative Points

J. Li*

Supervisor: Dr. X. L. Peng

Statistics Program, Division of Science and Technology, BNU-HKBU United International College

** Corresponding student author. Tel: +86-17727650623, E-mail: m730005024@mail.uic.edu.cn*

Abstract

The goodness-of-fit test, a hypothesis test to see how well sample data fit a given statistical distribution, has been one of the fundamental problems in Statistics for decades. The Pearson x^2 test is among the oldest known goodness-of-fit tests, it has been in use in many fields ever since. The performance, that is the testing power, of the x^2 test depends crucially on the way the data is grouped. In this thesis, a natural partition scheme based on representative points was introduced to improve the traditional x^2 test. By using this new partition scheme, three methods for normality test were proposed. These three new methods for normality test have show good performance compare with the traditional method. Some Monte Carlo Studies and an illustrative example all show that the improved x^2-tests for normality based on representative points do increase the power of the test w.r.t. various alternative distributions.

Keywords: MSE Representative point, x^2-tests , Simulation

Introduction

In many applications, we need to know or verify the statistical distribution from which the data comes from. Therefore the goodness-of-fit test, a hypothesis test to see how well sample data fit a given statistical distribution, has been one of the fundamental and most studied problems in Statistics for decades. Some discussions of existing goodness-of-fit tests were given by D'Agostini and Stephens (1986), Raynoe et al. (2009) and Zhang (2002). Let X be a random variable with distribution function $F(x)$, the goodness-of-fit test test the following hypothesis:

$$H_0 : F(x) = F_0(x)$$
$$H_a : F(x) \neq F_0(x)$$

where F_0 can be any statistical distribution. When F_0 refers to normal distribution, the corresponding test is called normality test. The application of normality test is very extensive, one common application of it in regression analysis is to test the normality of residuals. And some other applications refer to the normality test will be given in the following content.

Since the representative points are selected to represent the distribution, it is expected that the partition of the support of X determined by representative points will be a good partition that retain as much information as the distribution. In this work, a natural partition scheme based on representative points was introduced to improve the traditional χ^2 test. By using this new partition scheme, three methods for normality test were proposed.And in the fourth part, we will used MATLAB to simulate the testing power of these three method comparing with the original χ^2-tesing.

Classical Pearson's x^2-test for goodness-of-fit

Let $\{x_1, \ldots, x_n\}$ be a set of i.i.d. sample from a population with distribution function $F(x)$ $(x \in \mathcal{R})$. We want to test the null hypothesis

$$H_0 : \ F(x) \sim N(\mu, \sigma^2)$$

Pearson's χ^2-statistic (Pearson, 1933) is defined by

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i}, \qquad p_i = \int_{I_i} dF(x)$$

asymptotic null distribution of χ^2 is given by Fisher $\chi^2 \xrightarrow{\mathcal{D}} \chi^2(k-3),\ n \to \infty.$

The p-value for χ^2-test is approximately computed by

$$P(\chi^2, \nu) = K \int_{\chi^2}^{\infty} z^{\frac{\nu}{2}-1} \exp(-\frac{z}{2})\, dz, \ \text{with}\ \nu = k-3,\ K = \left[2^{\frac{\nu}{2}} \Gamma(\frac{\nu}{2})\right]^{-1}$$

However, there're still exist two uncertanties of determine the n_i and p_i.

- The first one is determine the value of k which is the partitions number of the
- second uncertainty which is determine the way to divide the continuous distribution into discrete multinomial distribution under the give number of k.

Pearson's x^2-test base on MSE representative points

Introduction of Representative points

Fang and He (1984) discussed the problem of selecting a specified number of MSE representative poitns from a normal distribution. This set of points could be retain as much information of the population as possible by changing a continuous distribution into a discrete multinomial distribution. Let X be a continuous random variable with pdf $f(x)$, consider the set of points $\{R_1, \ldots, R_m\}$, where $-\infty \leq R_1 \leq R_2 \leq \ldots \leq R_m \leq \infty$. The points $\{R_1, \ldots, R_m\}$ are called MSE representative points of X when

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

Three new methods base on RPs

Method 1

Let $z_1, ..., z_m$ be the m representative points of the standard noraml distribution $N(0,1)$. Then as we need to apply χ^2 test to verify if the sample $x_1, ..., x_m$ is from the normal distribution with unknown parameter μ and σ^2.Then the representative points $R_1, ..., R_m$ for $N(\mu, \sigma^2)$ can be estimate by

$$R_s = \hat{\mu} + \hat{\sigma} z_s$$

$$\chi^2 \xrightarrow{\mathcal{D}} \chi^2(m-3),\ n \to \infty$$

Method 2

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

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

$$R_s = \hat{\sigma} z_s \qquad \chi^2 \xrightarrow{\mathcal{D}} \chi^2(m-2),\ n \to \infty$$

Method 3

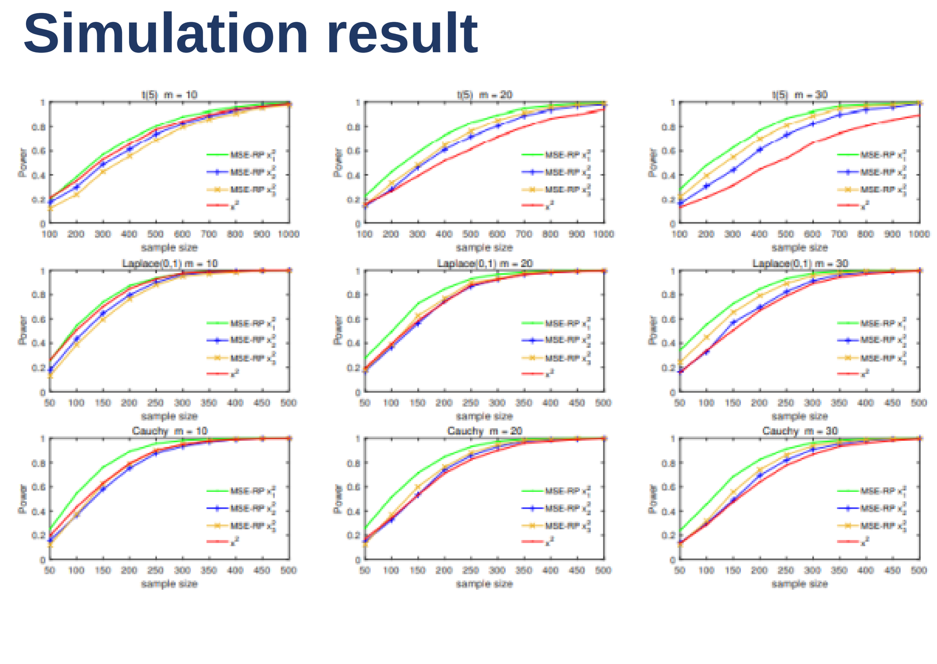
Apply the Studentized transformation

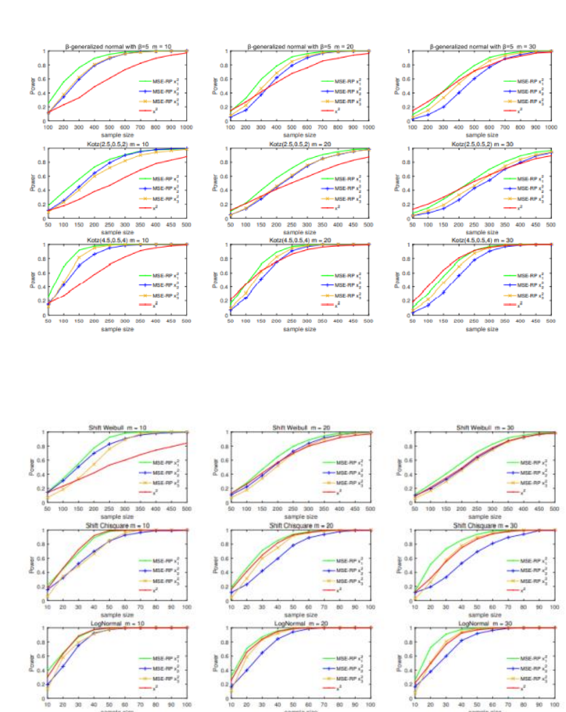
$$T_i = \frac{x_i - \bar{x}}{s_n}, \quad s_n^2 = \frac{1}{n} \sum_{l=1}^{n} (x_l - \bar{x})^2,\ \bar{x} = \frac{1}{n} \sum_{l=1}^{n} x_l,$$

for $i = 1, \ldots, n$. Under hypothesis (1), $\{T_1, \ldots, T_n\}$ are approximately i.i.d. and have a Student's t-distribution $t(n-1)$. with $t(n-1)$ as the null distribution and $\{T_1, \ldots, T_n\}$ in (10) being the set of transformed approximate i.i.d. sample:

$$\chi^2 \xrightarrow{\mathcal{D}} \chi^2(m-1),\ n \to \infty$$

Simulation result





References

[1] Kai-Tai Fang, Shu-Dong He(1996). How to choose a given number of representative points in a normal population, ACTA MATHEMATICAR APPLICATAE SINICA, vol.7 No.3.
[2] Ming Zhou, Wen-Jun Wang(2016). Representative points in t distribution and it's application, ACTA MATHEMATICAR APPLICATAE SINICA, vol.39 No.4.
[3] H.B Mann, A. Wald(1942). On the choice of the numberof class intervals in the application of the chi square test Columbia University
[4] Cox, D. R(1957). Note on GroupingJournal of the American Statistical Association,52 280, pp. 543-547
[5] Max, J(1960). Quantizing for minimum distortion IRE Transformation theory,6,pp. 7-12.
[6] Kenneth J.Koehler, Kinley Larntz(1978).An empirical investifation of goodness-of-fit test for sparse multinomials, Technical Report,No. 327.
[7] D'Agostini, R.B. and Stephens, M.A(1986). Goodness-of-fit techniques, statistics: textbooks and monographs. Marcel Dekker, New York.
[8] Raynor, J.C., Thas, O. and Best, D.J(2009). . Smooth tests of goodness of fit using RWiley and Sons
[9] Lian-Hua Liu, Wen-Qiang Luo(2013). Comparison of effectiveness of goodness of fit test, Journal of Yangtze University, vol.10 No.4.
[10] Gang-Yong Zhang, Lu-Ning Ruan(2011).Random simulation based on Monte-Carlo Comparison of several normal testing methods, statistics and decision, vol.331 No.7.
[11] Gang-Yong Zhang, Lu-Ning Ruan(2011).Random simulation based on Monte-Carlo Comparison of several normal testing methods, statistics and decision, vol.331 No.7.
[12] Wolfgang Rolke, · Cristian Gutierrez Gongora(2020).A chi-square goodness-of-fit test for continuous distributions against a known alternative, Computational Statistics, Vol. 81, No. 12