Journal of the Korean Data & Information Science Society
2009, **20**(3), 587–592

한국데이터정보과학회지

Likelihood ratio in estimating Chi-square parameter

Mezbahur Rahman[1]

Department of Mathematics and Statistics, Minnesota State University

Received 7 January 2009, revised 25 February 2009, accepted 10 April 2009

Abstract

The most frequent use of the chi-square distribution is in the area of goodness-of-fit of a distribution. The likelihood ratio test is a commonly used test statistic as the maximum likelihood estimate in statistical inferences. The recently revised versions of the likelihood ratio test statistics are used in estimating the parameter in the chi-square distribution. The estimates are compared with the commonly used method of moments and the maximum likelihood estimate.

Keywords: Anderson-Darling test, Chi-square distribution, Cramer von-Mises, goodness-of-fit, Kolmogorov-Smirnov test, parameter estimate.

1. Introduction

The parameter estimation for the chi-square distribution is not as specifically discussed as some other commonly used distributions, such as, exponential distribution, gamma distribution, etc. The reason may be that the chi-square distribution is a special case of the gamma distribution. For the literature reviews for the parameters in a two-parameter gama distributions, the readers are referred to Dang and Weerakkody (2000), Rahman et al. (2007), and the references therein. In this paper, it is shown that the likelihood ratio test statistics can also be successfully used in estimating papameters. As the likelihood ratio tests use the distribution functions instead of the density function, might have advantages in certain situations. For example, Cheng and Amin (1983) and Cheng and Iles (1987) showed that the method of maximization of the product spacings using the distribution functions perform better than usual maximum likelihood estimates using the density functions when the support of the distribution is dependent on a parameter.

2. Motivation

In constructing frequency distribution and histogram one of the very fundamental question arises is that how many groups to be considered. Then in the very early stage of data analysis the next important question is how data is distributed. The answer to the second question

[1] Department of Mathematics and Statistics, Minnesota State University, Mankato, MN 56002, USA. E-mail: mezbahur.rahman@mnsu.edu

is often the aid of the chi-square goodness-of-fit test. But both the questions are related. In literature, these two questions are discussed extensively. Here, the attempt is made to investigate that how closely a chi-square distribution can be estimated. And to investigate that whether there is any criteria which is more stable than the others which can later be used in determining the optimum number of groups to capture the distributional properties of a data set.

3. Likelihood-ratio tests

Let X be a continuous random variable with distribution function $F(x)$, and $X_1, X_2, \ldots, X_n$ be a random sample from $F(x)$ with order statistics $X_{(1)}$, $X_{(2)}$, …, $X_{(n)}$. One may wish to test the null hypothesis

$$\begin{aligned} &H_0 : F(x) = F_0(x) \quad \text{for all } x \in (-\infty, \infty), \\ &H_1 : F(x) \neq F_0(x) \quad \text{for some } x \in (-\infty, \infty), \end{aligned}$$

where $F_0(x)$ is a hypothetical distribution function which is completely specified.

Zhang and Wu (2005) developed three different versions of likelihood-ratio tests and implemented for the tests for normality along with some power comparisons. The tests are as follows: Likelihood-ratio Kolmogorov-Smirnov statistic

$$LK = \max_{i \in \{1,2,\cdots,n\}} \left\{ (i-0.5)\log\frac{i-0.5}{nF_0(X_{(i)})} + (n-i+0.5)\log\frac{n-i+0.5}{n[1-F_0(X_{(i)})]} \right\},$$

Likelihood-ratio Cramer-von-Mises statistic

$$LC = \sum_{i=1}^{n} \left\{ \log\frac{F_0(X_{(i)})^{-1} - 1}{(n-0.5)/(i-0.75) - 1} \right\}^2,$$

and Likelihood-ratio Anderson-Darling statistic

$$LA = -\sum_{i=1}^{n} \left\{ \frac{\log F_0(X_{(i)})}{n-i+0.5} + \frac{\log[1-F_0(X_{(i)})]}{i-0.5} \right\}.$$

Zhang and Wu (2005) showed that these tests are more powerful than the traditional Kolmogorov-Smirnov test, Cramer-von-Mises test, and Anderson-Darling test.

4. Estimation in Chi-square distribution

Let us consider $X_1, X_2, \ldots, X_n$ be a random sample from the chi-square distribution having the probability density function

$$f(x;\theta) = \frac{1}{2^{\frac{\theta}{2}}\Gamma\left(\frac{\theta}{2}\right)} x^{\frac{\theta-2}{2}} e^{-\frac{x}{2}}; \quad x > 0, \quad \theta > 0,$$

where θ is the parameter, usually known as the degree of freedom.

4.1. Method of moment estimate (MME)

The method of moment estimate for θ is the sample mean as the population mean is θ, $\hat{\theta}_M = \bar{X}$. The mean of $\hat{\theta}_M$ is θ and the variance of $\hat{\theta}_M$ is $2\theta/n$.

4.2. Maximum likelihood estimate (MLE)

The maximum likelihood estimate for θ, $\hat{\theta}_L$ is found by maximizing the log-likelihood with respect to θ,

$$\ln L = \frac{\theta - 2}{2}\sum_{i=1}^{n} \ln X_i - \frac{1}{2}\sum_{i=1}^{n} \ln X_i - \frac{n\theta}{2}\ln 2 - n\ln\Gamma\left(\frac{\theta}{2}\right), \tag{4.1}$$

where 'ln' stands for the natural logarithm. To maximize (4.1), a grid search in the range of $\left\{\max\left(0, \hat{\theta}_M - 4\sqrt{2\hat{\theta}_M/n}\right), \hat{\theta}_M + 4\sqrt{2\hat{\theta}_M/n}\right\}$ can be used. Here, usual maximization involving the Di-gamma function avoided to keep consistency with the methods discussed below. The grid search interval considered using 4 standard deviations to accomodate almost all possibilities.

4.3. Likelihood-ratio Kolmogorov-Smirnov statistic (KSE)

The likelihood-ratio Kolmogorov-Smirnov statistic estimate for θ, $\hat{\theta}_K$ is found by minimizing the Likelihood-ratio Kolmogorov-Smirnov statistic

$$LK = \max_{i\in\{1,2,\cdots,n\}} \left\{(i-0.5)\log\frac{i-0.5}{nF(X_{(i)};\theta)} + (n-i+0.5)\log\frac{n-i+0.5}{n[1-F(X_{(i)};\theta)]}\right\}, \tag{4.2}$$

where

$$F(x;\theta) = \int_0^x f(t;\theta)dt = \int_0^x \frac{1}{2^{\frac{\theta}{2}}\Gamma\left(\frac{\theta}{2}\right)} t^{\frac{\theta-2}{2}} e^{-\frac{t}{2}} dt.$$

4.4. Likelihood-ratio Cramer-von-Mises statistic (CVE)

The likelihood-ratio Cramer-von-Mises statistic estimate for θ, $\hat{\theta}_C$ is found by minimizing the likelihood-ratio Cramer-von-Mises statistic

$$LC = \sum_{i=1}^{n} \left\{\log\frac{F(X_{(i)})^{-1} - 1}{(n-0.5)/(i-0.75) - 1}\right\}^2, \tag{4.3}$$

where

$$F(x;\theta) = \int_0^x f(t;\theta)dt = \int_0^x \frac{1}{2^{\frac{\theta}{2}}\Gamma\left(\frac{\theta}{2}\right)} t^{\frac{\theta-2}{2}} e^{-\frac{t}{2}} dt.$$

4.5. Likelihood-ratio Anderson-Darling statistic (ADE)

The likelihood-ratio Anderson-Darling statistic estimate for θ, $\hat{\theta}_A$ is found by minimizing the likelihood-ratio Anderson-Darling statistic

$$LA = -\sum_{i=1}^{n} \left\{ \frac{\log F(X_{(i)})}{n-i+0.5} + \frac{\log[1-F(X_{(i)})]}{i-0.5} \right\}, \tag{4.4}$$

where

$$F(x;\theta) = \int_0^x f(t;\theta)dt = \int_0^x \frac{1}{2^{\frac{\theta}{2}}\Gamma\left(\frac{\theta}{2}\right)} t^{\frac{\theta-2}{2}} e^{-\frac{t}{2}} dt.$$

5. Simulation results

Ten thousand samples are generated for three different parameter settings θ =2, 5, and 10 and for three different sample sizes $n = 10, \;\; 20, \;\; \text{and} \;\; 30$. Means (MEAN), standard deviations (SD), mean of the absolute biases (MAB), biases (BIAS), and mean squared errors (MSE) are computed and displayed in Tables 5.1-5.2. The results are analyzed in Section 7.

MATLAB software is used in all computations and the codes are readily available upon request.

6. Application

To demonstrate how these estimates can be implemented for a real life data, the following data in Table 6.1 represents failure times of machine parts from manufacturer A and are taken from *http : //v8doc.sas.com/sashtml/stat/chap29/sect44.htm* is consedered: If we

assume that this data in Table 6.1 follows a Chi-square distribution, the estimates of the parameter are $\hat{\theta}_M = 468.74$, $\hat{\theta}_L = 481.63$, $\hat{\theta}_K = 455.83$, $\hat{\theta}_C = 455.83$, and $\hat{\theta}_A = 455.83$. This data is considered only to show a life example where the estimation procedures mentioned in the paper can be performed successfully.

7. Summary and concluding remarks

From Table 5.1-5.2, it is observed that all the estimates appear to be consistent and asymptotically unbiased, i.e., as n increases the standard deviations decrease and the biases decrease. In all cases, $\hat{\theta}_M$ has the smallest biases, the highest standard deviations, and highest mean square errors (except for $\theta = 10$ and $n = 20$). In all cases, $\hat{\theta}_L$ has the smallest standard deviations and the smallest mean square error (except for $\theta = 2$ and $n = 10$). For the other three estimates, in some instances, biases are smaller than $\hat{\theta}_L$, the standard deviations and the mean square errors are always smaller than $\hat{\theta}_M$. Among the three likelihood-ratio estimates, $\hat{\theta}_A$ performed better, often having lower biases, lower standard deviations, and lower mean square errors compared to the other two estimates. The

asymptotic properties are not established in tthis paper as there is no closed form estimators

Table 5.1 Simulation results

	$\hat{\theta}_M$	$\hat{\theta}_L$	$\hat{\theta}_K$	$\hat{\theta}_C$	$\hat{\theta}_A$
		$\theta = 2$	$n = 10$		
MEAN	1.9888	2.0800	2.0521	2.0732	2.0408
SD	0.6288	0.5114	0.5403	0.5160	0.5139
MAB	0.4998	0.4031	0.4266	0.4068	0.4051
BIAS	-0.0112	0.0800	0.0521	0.0732	0.0408
MSE	0.3955	0.2679	0.2946	0.2716	0.2658
		$\theta = 2$	$n = 20$		
MEAN	1.9944	2.0392	2.0276	2.0417	2.0232
SD	0.4451	0.3576	0.3843	0.3629	0.3611
MAB	0.3546	0.2821	0.3025	0.2862	0.2849
BIAS	-0.0056	0.0392	0.0276	0.0417	0.0232
MSE	0.1981	0.1294	0.1484	0.1333	0.1310
		$\theta = 2$	$n = 30$		
MEAN	1.9990	2.0316	2.0224	2.0341	2.0228
SD	0.3660	0.2912	0.3159	0.2961	0.2957
MAB	0.2912	0.2309	0.2501	0.2352	0.2343
BIAS	-0.0010	0.0316	0.0224	0.0341	0.0228
MSE	0.1339	0.0858	0.1003	0.0889	0.0880
		$\theta = 5$	$n = 10$		
MEAN	5.0122	5.1100	5.0789	5.1029	5.0650
SD	1.0155	0.9304	0.9689	0.9409	0.9371
MAB	0.8040	0.7340	0.7626	0.7420	0.7391
BIAS	0.0122	0.1100	0.0789	0.1029	0.0650
MSE	1.0314	0.8777	0.9451	0.8959	0.8824
		$\theta = 5$	$n = 20$		
MEAN	4.9974	5.0476	5.0319	5.0516	5.0299
SD	0.7086	0.6427	0.6841	0.6543	0.6524
MAB	0.5618	0.5097	0.5419	0.5182	0.5165
BIAS	0.0122	0.1100	0.0789	0.1029	0.0650
MSE	0.5021	0.4154	0.4690	0.4307	0.4265
		$\theta = 5$	$n = 30$		
MEAN	5.0083	5.0402	5.0308	5.0437	5.0288
SD	0.5808	0.5240	0.5667	0.5325	0.5333
MAB	0.4628	0.4180	0.4499	0.4248	0.4251
BIAS	0.0083	0.0402	0.0308	0.0437	0.0288
MSE	0.3374	0.2762	0.3221	0.2855	0.2852

Table 5.2 Simulation results continued

	$\hat{\theta}_M$	$\hat{\theta}_L$	$\hat{\theta}_K$	$\hat{\theta}_C$	$\hat{\theta}_A$
		$\theta = 10$	$n = 10$		
MEAN	10.0080	10.1075	10.0772	10.0998	10.0618
SD	1.4172	1.3545	1.4042	1.3714	1.3624
MAB	1.1266	1.0779	1.1147	1.0901	1.0812
BIAS	0.0080	0.1075	0.0772	0.0998	0.0618
MSE	2.0085	1.8462	1.9777	1.8907	1.8598
		$\theta = 10$	$n = 20$		
MEAN	10.0230	10.0677	10.0571	10.0719	10.0492
SD	0.9957	0.9495	1.0081	0.9682	0.9625
MAB	0.7929	0.7543	0.8005	0.7685	0.7636
BIAS	0.0230	0.0677	0.0571	0.0719	0.0492
MSE	0.9920	0.9062	1.0195	0.9425	0.9288
		$\theta = 10$	$n = 30$		
MEAN	9.9993	10.0367	10.0290	10.0432	10.0283
SD	0.8199	0.7835	0.8323	0.7968	0.7950
MAB	0.6550	0.6268	0.6660	0.6374	0.6363
BIAS	-0.0007	0.0367	0.0290	0.0432	0.0283
MSE	0.6722	0.6152	0.6935	0.6368	0.6328

in the proposed procedures. And the asymptotic properties for the maximum likelihood estimates and the method of product spacings are well established. In this paper, simulations

Table 6.1 Failure times

620	470	260	89	388	242	103	100	39	460	284
1285	218	393	106	158	152	477	403	103	69	158
818	947	399	1274	32	12	134	660	548	381	203
871	193	531	317	85	1410	250	41	1101	32	421
32	343	376	1512	1792	47	95	76	515	72	1585
253	6	860	89	1055	537	101	385	176	11	565
164	16	1267	352	160	195	1279	356	751	500	803
560	151	24	689	1119	1733	2194	763	555	14	45
776	1									

are used to establish the asymptotic behaviors of the estimates.

8. Acknowledgements

The author would like to thank the editor for facilitating the publication of the paper. The author also would like to thank the referees for their constructive comments which improved the presentation of the paper significantly.

References

Cheng, R. C. H. and Amin, N. A. K. (1983). Estimating parameters in continuous univariate distributions with a shifted origin. *Journal of Royal Statistical Society, Series B*, **45**, 394-403.

Cheng, R. C. H. and Iles, T. C. (1987). Corrected maximum likelihood in non-regular problems. *Journal of Royal Statistical Society, Series B*, **49**, 95-101.

Dang, H. and Weerakkody, G. (2000). Bounds for the maximum likelihood estimates in two-parameter Gamma distribution. *Journal of Mathematical Analysis and Applications*, **245**, 1-6.

Rahman, M., Pearson, L. M. and Martinovic, U. R. (2007). Method of product spacings in two-parameter Gamma distribution. *Journal of Statistical Research*, **41**, 51-58.

Zhang, J. and Wu, Y. (2005). Likelihood-ratio tests for normality. *Compulational Statistics & Data Analysis*, **49**, 709-721.