Review Of Research

ISSN: 2249-894X Impact Factor : 3.8014(UIF) Volume - 6 | Issue - 3 | December - 2016

IDENTIFYING THE PROBABILITY DISTRIBUTION USING HIGHER ORDER MOMENTS (HOM)

J. Purushotham[1] and Dr. K.Sampath Kumar[2]

[1]Research Scholar, Department of Applied Statistics, Telangana University, Nizamabad.

[2]Assistant Professor, Department of Applied Statistics, Telangana University, Nizamabad.

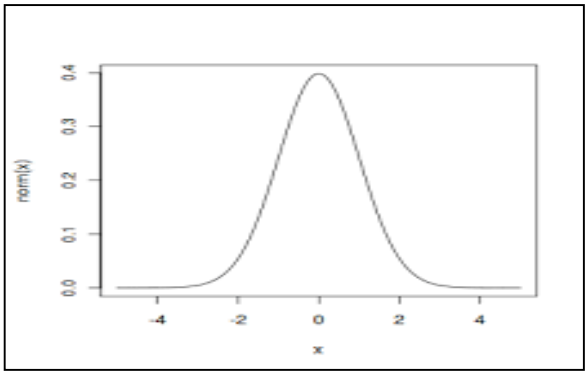
ABSTRACT:

The main objective of the present paper is to test the goodness of fit of the Poisson probability distributions using the moment based approach. Also, a numerical illustration of comparing the proposed method with the familiar Pearson-Fisher Chi-Square test is done. We, observe that the results obtained in identifying the probability distribution by both the methods are same.

KEYWORDS – Higher order Moments, Poisson distribution, Pearson-Fisher Chi-square test.

1. INTRODUCTION:

Moments are commonly used in statistics, physics and engineering. Moments are widely used in scientific areas that deal with data, random variables or stochastic process. The moment of an object measures the distribution of its mass relative to some



coordinate system. For instance, the moment of inertia about an axis characterizes the distribution, or spread, of an object's mass about that axis. This measures the tendency of the object to rotate about that axis. Moments are also found in the analysis of statistical distributions. In Statistical theory, they provide a measure of mass distribution, in terms of probability. An object's moments can also be treated as features and used to characterize the object for the purpose of identification or regain.

An important problem in statistical applications is to test whether or not an assumed probability distribution gives good fit to the data. The most commonly used method for testing goodness of fit of parametric family of distributions is the Chi-square test (Fisher, 1922, 1924).

Another general method to test the goodness of fit is to use some distance statistics such as Kolmogorov-Smirnov and Anderson-Darling statistic and others. The distance statistic tests have the advantage that they do not involve subjective selection of a partition. However, most distance tests are appropriate for testing simple hypothesis whether a set of observations are from some specified distribution function or not. But when to estimate the parameters of the distributions these methods are no longer useful. Lilliefors (1967, 1969 and 1973) studied the use of distance statistics in testing the goodness of fit of the normal, exponential family and gamma families.

In this context, it is found necessary to identify the form of the distribution from statistical analysis point of view.

2. METHODOLOGY TO TEST THE GOODNESS OF FIT

The methodology used in identifying the form of the probability distribution by the higher order moments consists of the following major steps.

2.1 The setup and assumption:

Let X_1, X_2,.....X_n are identically independently distributed (i.i.d) random variables from a

population having the probability density function (p.d.f) f(x,θ). Consider the problem of testing the null hypothesis

H0: F(x,θ) is a member of a parametric family f(x, θ); θ □ Θ

Where Θ is a subset of $\mathbb{R}^d$

Let r-th moment about origin (non-central) of the given distribution be denoted by m_r and is defined as

$$m_r = \int x^r f(x;\ \theta)\ dx\ ;\quad r = 1,2,3,\ldots\ldots$$

2.2 Assumption:

Assume that m_r i.e., r^{th} moment about origin do exists for some positive integer r and that m_1, m_2, …,m_r do satisfy the following equation

$f(m_1, m_2, \ldots\ldots m_r) = 0$ for all θ □ Θ

for some functioning f: $\mathbb{R} \rightarrow \mathbb{R}$

In general, it is very easy to find a function f satisfying above assumption. For example, for a parametric distribution which is symmetric about mean i.e., E(X-mean) = E(X-m_1) = 0, which implies that $m_3 - 3m_1m_2 + 2m_1^3 \equiv 0$. Thus, we can choose $f(x, y, z) = z - 3xy + 2x^3$.

In general, existence of function f satisfying assumption is possible, if all the moments m_1, m_2, …..,m_r depend on a common finite dimensional parameter θ.

2.3 The test procedure:

Let r^{th} sample moment about origin (non-central moment) of the given sample data X_1, X_2,…….X_n be denoted and defined by

$$\hat{m}_r = \sum_{j=1}^{n} x_j^r\ /n \quad \text{for } r = 1,2,\ldots\ldots$$

Theorem: Assume that above assumption holds and the function f(m1, m2, ……mr) is continuously differentiable.

Then, under null hypothesis H_0: F(x, θ) is a member of a parametric family f(x, θ); θ □ Θ, we have

$$\sqrt{n} f(\hat{m}_1, \hat{m}_2, \ldots\ldots \hat{m}_r) \rightarrow N(0, V(\theta)),$$

Where

$$V(\theta) = \left(\frac{\partial f(m_1, m_2, \ldots\ldots m_r)}{\partial m_1}, \ldots\ldots, \frac{\partial f(m_1, m_2, \ldots\ldots m_r)}{\partial m_r}\right) H \left(\frac{\partial f(m_1, m_2, \ldots\ldots m_r)}{\partial m_1}, \ldots\ldots, \frac{\partial f(m_1, m_2, \ldots\ldots m_r)}{\partial m_r}\right)^T \ldots (1)$$

where $H = (\Pi_{ij})_{rXr}$ with $\Pi_{ij} = m_{i+j} - m_i m_j$ for all i,j = 1,2,….r ……(2)

Proof: By the central limit theorem (i.e., suppose X_1, X_2,…….,X_n is a sequence of i.i.d random variables with mean (μ) and variance ($\sigma^2 > 0$), then as n tends to ∞ the random variable $\sqrt{n}(S_n - \mu) \sim N(0, \sigma^2)$, where $S_n = \frac{x_1 + x_2 + \cdots x_n}{n}$) and in combination with the Cramer-Wald method, the random vector

$$\sqrt{n}\{f(\hat{m}_1, \hat{m}_2, \ldots\ldots \hat{m}_r) - f(m_1, m_2, \ldots\ldots m_r)\}$$

encounters to r-variate normal distribution with mean vector (0,0,…..0) and variance matrix H (defined by (2)). This, together with the delta method gives that

$\sqrt{n}f(\hat{m}_1, \hat{m}_2, \ldots\ldots \hat{m}_r) = \sqrt{n}\{f(\hat{m}_1, \hat{m}_2, \ldots\ldots \hat{m}_r) - f(m_1, m_2, \ldots..m_r)\} \rightarrow N(0, V(\theta))$, with $V(\theta)$ is defined by (1)

Let $\hat{\theta} = \theta(X_1, X_2, \ldots\ldots, X_n)$ be a consistent estimator of parameter θ under null hypothesis H_0. Assume that $m_1, m_2, \ldots\ldots m_r$ are continuous of parameter θ. Then $V(\hat{\theta})$ is aconsistent estimator of $V(\theta)$ under null hypothesis H_0.

Let us define the test statistic,

$$Z = \sqrt{n}f(\hat{m}_1, \hat{m}_2, \ldots\ldots \hat{m}_r) / \sqrt{V(\theta)} \quad \ldots\ldots\ldots (3)$$

Then, Z follows N(0,1) (from above theorem under H_0) i.e., $Z \rightarrow N(0, 1)$ as n (sample size) $\rightarrow \infty$.

This gives the level of test of significance α to test null hypothesis H_0. Thus reject H_0 if $|Z| > Z_{\alpha/2}$, where $Z_{\alpha/2}$ is the upper α/2 percentile of the standard normal distributional value.

3. APPLICATION TO POISSON DISTRIBUTION:

Assume that $X_1, X_2, \ldots\ldots, X_n$ be a random sample drawn from an poisson population having the probability mass function (p.m.f)

$$P(x;\lambda) = \frac{e^{-\lambda}\lambda^x}{x!} \qquad \text{for all } x = 0,1,2,\ldots\ldots$$

$$= \text{other wise}$$

Consider the problem of testing the null hypothesis, H_0: the sample observations are drawn from poisson population against an alternative hypothesis H_1: the sample observations are not drawn from poisson population.

To test the hypothesis, we consider r = 3 and the function $f(x, y, z) = z - y - 2x^2 - x^3$ Further, because of m_1= the 1st moment about origin = λ; m_2 = the 2nd moment about origin = $\lambda^2 + \lambda$; m_3 = the 3rd moment about origin = $\lambda^3 + 3\lambda^2 + \lambda$ so that we have $f(m_1, m_2, m_3) = m_3 - m_2 - 2m_1^2 - m_1^3 \equiv 0$ for all $(\lambda) > 0$.

We have to estimate parameter $\theta = \lambda$ by either method of moments or method of maximum likelihood estimation i.e., $\hat{\theta} = (\hat{\lambda} = \bar{x})$ where $\bar{x}$ = Sample mean = $\frac{\sum_{i=1}^{n} x_i}{n}$

Also the V(θ) is obtained on using,

$$V(\theta) = \left(\frac{\partial f(m_1,m_2,m_3)}{\partial m_1}, \frac{\partial f(m_1,m_2,m_3)}{\partial m_2}, \frac{\partial f(m_1,m_2,m_3)}{\partial m_3}\right) H \left(\frac{\partial f(m_1,m_2,m_3)}{\partial m_1}, \frac{\partial f(m_1,m_2,m_3)}{\partial m_2}, \frac{\partial f(m_1,m_2,m_3)}{\partial m_3}\right)^T \ldots (4)$$

Where, $H = \begin{bmatrix} \Pi_{11} & \Pi_{12} & \Pi_{13} \\ \Pi_{21} & \Pi_{22} & \Pi_{23} \\ \Pi_{31} & \Pi_{32} & \Pi_{33} \end{bmatrix}$ with $\Pi_{ij} = m_{i+j} - m_i m_j$ for all i,j = 1,2,3

$$H = \begin{bmatrix} m_2 - m_1^2 & m_3 - m_1 m_2 & m_4 - m_1 m_3 \\ m_3 - m_2 m_1 & m_4 - m_2^2 & m_5 - m_2 m_3 \\ m_4 - m_3 m_1 & m_5 - m_3 m_2 & m_6 - m_3^2 \end{bmatrix} \ldots\ldots.(5)$$

On substituting the non-central moments listed above, we get

$$H = \begin{bmatrix} \lambda & 2\lambda^2 + \lambda & 3\lambda^3 + 6\lambda^2 + \lambda \\ 2\lambda^2 + \lambda & 4\lambda^3 + 6\lambda^2 + \lambda & 6\lambda^4 + 22\lambda^3 + 14\lambda^2 + \lambda \\ 3\lambda^3 + 6\lambda^2 + \lambda & 6\lambda^4 + 22\lambda^3 + 14\lambda^2 + \lambda & 9\lambda^5 + 54\lambda^4 + 85\lambda^3 + 32\lambda^2 + \lambda \end{bmatrix} \ldots(6)$$

Thus the $V(\theta) = [-4m_1 — 3m_1^2 \ -1 \ 1] H [-4m_1 — 3m_1^2 \ -1 \ 1]^T$ .....(7)

On using (3)

$$V(\theta) = [-4\lambda-3\lambda^2 \quad -1] \begin{bmatrix} \lambda & 2\lambda^2+\lambda & 3\lambda^3+6\lambda^2+\lambda \\ 2\lambda^2+\lambda & 4\lambda^3+6\lambda^2+\lambda & 6\lambda^4+22\lambda^3+14\lambda^2+\lambda \\ 3\lambda^3+6\lambda^2+\lambda & 6\lambda^4+22\lambda^3+14\lambda^2+\lambda & 9\lambda^5+54\lambda^4+85\lambda^3+32\lambda^2+\lambda \end{bmatrix} [-4\lambda-3\lambda^2 \quad -1 \quad 1]^T \quad -1$$

$$= 18\lambda^4 + 29\lambda^3 + 10\lambda^2 \ldots\ldots.(8)$$

The parameter λ in above is replaced by its estimate obtained by method of maxmimum likelihood estimation i.e., $\hat{\theta} = (\hat{\lambda} = \bar{x})$ where $\bar{x}$ = Sample mean = $\frac{\sum_{i=1}^{n} x_i}{n}$

Finally, the test statistic defined by (3.3) reduces to

$$Z = \sqrt{n}(\hat{m}_3 - \hat{m}_2 - 2\hat{m}_1^2 - \hat{m}_1^3) / \sqrt{V(\theta)} \sim N(0, 1) \ldots\ldots (9)$$

[Where $V(\theta)$ is defined by equation (7) or (8)]

and we use the normal test. This gives the level of test of significance α to test the null hypothesis H_0. Thus reject H_0 if $|Z| > Z_{\alpha/2}$, where $Z_{\alpha/2}$ is the upper α/2 percentile of the standard normal distribution otherwise accept H_0. In the next section, we illustrate numerically the above procedure.

4. NUMERICAL ILLUSTRATION:

Consider a random sample of n = 1000 observations drawn from poisson population having the parameter λ i.e., $X_i \sim P(\lambda)$. Then, our problem is to test the null hypothesis H_0: The sample observations are drawn from poisson population against an alternative hypothesis H_1: The Sample observations are not drawn from poisson population.

To test the hypothesis, the test statistic is (From equation (9))

$$Z = \sqrt{1000}(\hat{m}_3 - \hat{m}_2 - 2\hat{m}_1^2 - \hat{m}_1^3) / \sqrt{V(\theta)} \sim N(0, 1)\ldots\ldots (10)$$

The first six sample moments about origin (non-central) are listed below.

Sample moment about Origin (Non-Central)	Value
m_1	0.997
m_2	2.047
m_3	5.371
m_4	17.227
m_5	64.507
m_6	271.987

and V(θ) = 81.231119

Substituting the above values in equation (10), we get

$Z = \sqrt{1000}(5.371 - 2.047 - 2 * 0.997^2 - 0.997\text{^}3)/\sqrt{81.231119}$

Z_{cal} = 1.2103

Thus, $Z_{cal} < Z_{\alpha/2}$ (± 1.96) at α=0.05 level of significance. We accept the null hypothesis H_0 and conclude that the sample has been drawn from poisson population.

5. COMPARING WITH CHI-SQUARE TEST OF GOODNESS OF FIT:

Consider a sample of n=1000 observations used in the above illustration and test the goodness of fit using the method of Pearson-Fisher Chi-square test.

Observation (x)	0	1	2	3	4	5	6	Total
No. of times (f)	375	366	177	58	19	3	2	1000

Here our problem is to test the null hypothesis H_0: the sample observations are drawn from poisson population against an alternative hypothesis H_1: the sample observations are not drawn from poisson population.

In order to fit a poisson distribution to the given data, we take the mean (parameter) λ of the Poisson distribution equal to the mean of the given distribution i.e., λ = 0.997.

Observation (x)	Obsered frequency (oi)	$P(x) = \frac{e^{-\lambda}\lambda^x}{x!}$	Expected frequency (ei)	$(o_i - e_i)^2/e_i$
0	375	0.3689847	368.98≈369	0.09756
1	366	0.3678777	367.87≈368	0.01087
2	177	0.183387	183.38≈183	0.19672
3	58	0.060945	60.94≈61	0.14754
4	19	0.015190	15.19≈15	1.3157
5	03	0.003029	3.03≈3	
6	02	0.5033	0.5033≈1	
Total	1000		1000	1.76847

(d.f. = 7 -1-1-2 =3; one d.f. being lost because of the linear constraint $\sum o_i = \sum e_i$ 1 d.f. is lost because the parameter λ has been estimated from the given data and is then used for computing expected frequencies, 2 d.f. are lost because of pooling the last 3 expected frequencies).

Tabulated value of χ2 for 3 d.f. at 5% l.o.s. is 7.815

Conclusion: Since calculated value of χ2 = 1.76847 is less than the critical value 7.815, we accept the null hypothesis and conclude that poisson distribution is a good fit to the given data.

6. CONCLUSION:

We, observe that the proposed moment based goodness fit test and the Pearson-Fisher Chi-Square test results the same. The importance of the method is it can be applicable to any parametric family of distributions to test the goodness of fit.

REFERENCES:

1. P.J.BICKEL, K.A.DOKSUM (1979), Mathematical Statistics, llolden-Day, Inc., San Francisco.
2. R.B.D'AGOSTINO, M.A. STEPHENS (1986), Goodness of fit techniques, Marcel Dekker, New York.
3. R.A. FISHER (1922), On the interpretation of χ^2 from contingency tables, and the calculation of l, " Journal of the Royal Statistics Society", 87, pp.442-450.
4. H.W. LILLIEFORS (1967), On the Kolmogorov-Smirnov test for normality with mean and variance unknown, "Journal of the American Statistical Association", 62, pp.339-402.
5. H.W. LILLIEFORS (1969), On the Kolmogorov-Smirnov test for the exponential distributions with mean unknown, "Journal of the American Statistical Association:, 64,pp.387-389.
6. G.LI, A.PAPADOPOULOS (2002), A note on goodness of fit test using moments, "Statistica, annoLXII, n,1,2002".

7. K. Sampath kumar, Prof. V.V. Hara Gopal (2007), On distributions which differ only higher order moments.
8. Cramer (1946): Mathematical methods of Statistics, Princeton University, and Princeton University Press.
9. Hu, M.K. (1962): Visual Pattern Recognition by moment invariants IRE Trans. Info. Theory, Volume IT-8, pp-178-187.
10. Dutt, V.A.K. (1995): Multivariate and related statistical methods in pattern recognition. Un-Published Ph.D. Thesis-Osmania University, Hyderabad.

J.Purushotham
Research Scholar, Department of Applied Statistics, Telangana University, Nizamabad.

Dr. K.Sampath Kumar
Assistant Professor, Department of Applied Statistics, Telangana University, Nizamabad.