Large Sample Test for identifying the form of Probability Function using Sample Moments

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

[1]Research Scholar, [2]Assistant Professor
Department of Applied Statistics,
Telangana University, Nizamabad

***Abstract*: The main objective of the present paper is to identify the probability distributions using sample moments in large sample test. The method is applied to the exponential uni-variate continuous distribution. Also, a numerical illustration of comparison of this method with Kolmogorov-Smirnov non-parametric test. We, identify that the proposed method of large sample test and non-parametric test for identifying the form of probability distribution do agree.**

***Keywords* – Higher order Moments, Exponential Probability Distribution, Kolmogorov-Smirnov one sample test, Z-test.**

1. Introduction: In general moments are usually discussed in terms of their order. The zeroth ordered moment is the full mass, the first moment gives the center of mass (mean) of the object and the second order moment is the rotational inertia i.e., measures the variance or spread of the object. The moments beyond the second order do not have much physical interpretations but are useful in describing an object's shape parameters. The third order moment generally represent possible lack of skewness (lack of symmetry or asymmetry) of the distribution and the fourth order moment describes the shape of curve of the distribution i.e., whether the distribution forms a normal (bell shaped), platy kurtic or leptokurtic curve.

So in general at least the first four moments of any distribution are of much important for the purpose of characterizing distributions of some of the well known theoretical discrete as well as continuous. Thus, the mathematical idea is closely associated with the concept of moment in physics. Hence, mathematical concept of moments can be equally well used to characterize any positive functions, whose total area under its curve is finite and normalized to unity. This possibility has gained more importance in dealing with conventional statistical methods and data analysis in pattern classification and identification problem and also in digital image processing.

In Statistical applications it is to identify a distribution type by its first few (usually four) standardized moments particularly when the "family" to which the distribution of interest belongs is specified. The rationale for this practice probably is that the families of distributions of interest are defined as classes of solutions of differential (or difference) equations involving not more than four parameters (as in the case of Pearsonian system) or, are derived by suitably transforming a variate from such a distribution into a new variate, (as in the case of the Johnson family). In fact, the well known generalized Tukey approach (Tukey (1960), Ramberg et al. (1979) and Ramesh(1987)) to fitting distribution to data, which has its advantages in generating a distribution effectively equivalent to an observed distribution, uses only the first four moments of the distribution.

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); $\theta \in \Theta$

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

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

2.2 Assumption:

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

$f(m_1, m_2, \ldots\ldots m_r) = 0$ for all $\theta \in \Theta$

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 rth sample moment about origin (non-central moment) of the given sample data $X_1, X_2, \ldots\ldots X_n$ be denoted and defined by

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

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 ϵ Θ, we have $\sqrt{n}f(\widehat{m}_1, \widehat{m}_2, \ldots\ldots \widehat{m}_r) \to N(0, V(\theta))$,

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 $X_1, X_2, \ldots\ldots, X_n$ 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

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

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

$$\sqrt{n}f(\widehat{m}_1, \widehat{m}_2, \ldots\ldots \widehat{m}_r) = \sqrt{n}\{f(\widehat{m}_1, \widehat{m}_2, \ldots\ldots \widehat{m}_r) - f(m_1, m_2, \ldots\ldots m_r)\} \to N(0, V(\theta)),$$

with V(θ) 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, ……m_r are continuous of parameter θ. Then $V(\hat{\theta})$ is aconsistent estimator of V(θ) under null hypothesis H_0.

Let us define the test statistic,

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

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

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 Exponential Distribution:

3.1 The set up and assumptions

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

$$f(x, \theta) = \begin{cases} \theta e^{-\theta x} & ; x \geq 0 \\ = 0 & ; \text{otherwise} \end{cases}$$

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

To test the hypothesis, we consider r = 3 and the function $f(x, y, z) = 8z^2 - 36y^3 - y + 2x^2$. Further, because of m_1= the 1st moment about origin $=\frac{1}{\theta}$, m_2 = the 2nd moment about origin $=\frac{2}{\theta^2}$, m_3= the 3rd moment about origin = $\frac{6}{\theta^3}$ so that, we have $f(m_1, m_2, m_3) = 8m_3^2 - 36m_2^3 - m_2 + 2m_1^2 = 0$ for all θ > 0.

We have to estimate parameter θ by either method of moments or method of maximum likelihood estimation i.e., $\hat{\theta} = \frac{1}{\bar{x}}$ 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

$$H=\begin{bmatrix} m_2-m_1^2 & m_3-m_1m_2 & m_4-m_1m_3 \\ m_3-m_2m_1 & m_4 \ -m_2^2 & m_5-m_2m_3 \\ m_4-m_3m_1 & m_5-m_3m_2 & m_6-m_3^2 \end{bmatrix} \ldots\ldots.(5)$$

Thus the V(θ) = [$4m_1$ $-108m_2^2$-1 $16m_3$] H [$4m_1$ $-108m_2^2$-1 $16m_3$] ^{T} ……(6)

[Where H is defined in equation (5)]

Finally, the test statistic defined by (3.2) is

$$Z = \sqrt{n}(8\widehat{m}_3{}^2 - 36m_2{}^3 - m_2 + 2m_1{}^2) / \sqrt{V(\theta)} \sim N(0,1) \ \ldots..(7)$$

[Where $V(\theta)$ is defined by equation (4)]

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 an exponentially distributed population having the parameter θ i.e., X_i ~ Exp (θ). Then, our problem is to test the null hypothesis H_0: The sample observations are drawn from exponential population against an alternative hypothesis H_1: The Sample observations are not drawn from an exponential population.

To test the hypothesis, the test statistic is (From equation (7))

$$Z = \sqrt{1000}(8\widehat{m}_3{}^2 - 36\widehat{m}_2{}^3 - \widehat{m}_2 + 2\widehat{m}_1{}^2) / \sqrt{V(\theta)} \sim N(0,1) \ \ldots\ldots.. (8)$$

We have the first six non-central moments about origin as given below:

Sample moment about Origin (Non-Central)	Value
m_1	1.003
m_2	1.876
m_3	5.064
m_4	18.144
m_5	82.209
m_6	445.202

and V(θ) = 386939.2518

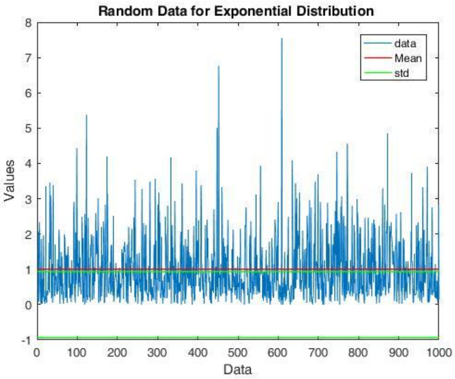
Substituting the above values in equation (5), we get

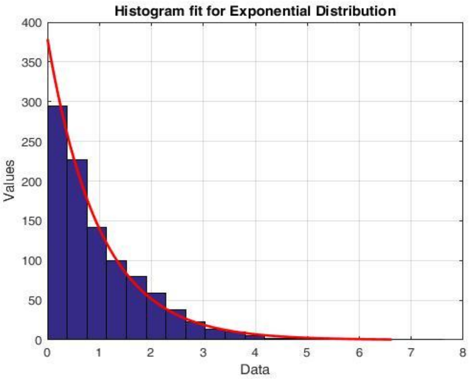
$$Z = \sqrt{1000}(8*5.06^2 - 36*1.876^3 - 1.876 + 2*1.003\text{^}2) \ / \sqrt{}386939.2518$$

Z_{cal} = -1.6469

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 have been drawn from exponential population.

The following figures shows the fit of the random data and histogram for a sample of n=1000 observations drawn from exponential population.





5. Comparing with Kolmogorov-Smirnov One Sample Test:

Consider a sample of n=1000 observations used in the above illustration and test the goodness of fit using the Kolmogorov-Smirnov test.

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

The result obtained from SPSS is given below:

One-Sample Kolmogorov-Smirnov Test

		exponential
N		1000
Exponential parameter.[a,b]	Mean	1.00257410
Most Extreme Differences	Absolute	.034
	Positive	.014
	Negative	-.034
Kolmogorov-Smirnov Z		1.068
Asymp. Sig. (2-tailed)		.204

a. Test Distribution is Exponential.

b. Calculated from data.

From above table, the null hypothesis H_0 is not significant. Thus, we conclude that the sample data is drawn from exponential population.

6. Conclusion: We, observe that proposed large sample test for identifying the form of the probability distribution by proposed method using sample moments and Kolmogorov-Smirnov one sample test do agree. The proposed method of moments can be applied to any parametric family of distributions where there exists normality in the data.

References:

[1] 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.

[2] 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.

[3] G.LI, A.PAPADOPOULOS (2002), A note on goodness of fit test using moments, “Statistica, annoLXII, n,1,2002”.

[4] Kendall, M.G. and Stuart, A. (1977): The Advanced Theory of Statistics. Vol. I. Charles Griffin &Co.

[5] Laksminarayana, J., Pandit, S.N.N., and Srinivas Rao, K. (1999): On a Bivariate Poisson Distribution, Communication in Statistics – Theory and Methods, 28(2)-pp 267-276.

[6] Bruggisser, Moritz, et al. "Retrieval of higher order statistical moments from full-waveform LiDAR data for tree species classification." Remote Sensing of Environment 196 (2017): 28-41.

[7] Arismendi, Ivan, Sherri L. Johnson, and Jason B. Dunham. "Higher-order statistical moments and a procedure that detects potentially anomalous years as two alternative methods describing alterations in continuous environmental data." Hydrology and Earth System Sciences 19.3 (2015): 1169-1180.

Dr. K. Samath Kumar

J. Purushotham