Journal of Administrative Sciences and Policy Studies
December 2015, Vol. 3, No. 2, pp. 75-90
ISSN: 2372-5109 (Print), 2372-5117 (Online)

Published by American Research Institute for Policy Development
DOI: 10.15640/jasps.v3n2a5
URL: http://dx.doi.org/10.15640/jasps.v3n2a5

On Some Relative Entropy Statistics

Erhan Ustaoglu[1] & Atif Evren[2]

Abstract

Statistical entropy is a measure of dispersion or spread of a random variable. Especially when the random variable is nominal, classical measures of dispersion like standard deviation can not be computed. In such cases, measures of variation, including entropy-based statistics;computed by using cell frequencies of a distributionmust be used. The asymptotic properties of entropy statistics have long been studied in literature. Relative entropy plays an important role in evaluating the degree of fit. In other words, relative entropy is a measure of goodness fit of an empirical distribution to a theoretical or hypothesized distribution. In this study for some frequently-used probability distributions,some relative entropy measures are derived by exploiting additivity property of Kullback-Leibler divergence and Jeffreys divergence. Their asymptotic properties under certain assumptions have been discussed. In the end, by some applications, the close relation between relative entropy statistics and other classical test statistics have been emphasized.

Keywords: Relative entropy, Kullback-Leibler divergence, Jeffreys divergence, mutual information, asymptotic properties of relative entropy

Introduction

Statistical entropy can be evaluated as ameasure of unpredictability of the outcome of a statistical experiment. The more predictable the outcome of an experiment, the less will be the uncertainty and soforth the entropy. After the experiment (or observation) is carried out, the uncertainty is not present. So in some sense, entropy is a measure of information that one can get through statistical experimentation (Renyi, p23).

[1]Marmara University, Faculty of Administrative Sciences, Department of Management, Information Sciences, Bahçelievler, 34180 Istanbul , Turkey. E-mail: erhan.ustaoglu@gmail.com, tel: + 90 530 343 82 99
[2]Yildiz Technical University, Faculty of Sciences and Literature, Department of Statistics Davutpasa, Esenler, 34210, Istanbul, Turkey. E-mail: aevren@yildiz.edu.tr, tel: + 90 533 3454973

Among various entropy measures proposed in literature, Shannon, Rényi and Tsallis entropies gained popularity recently. Yet, Shannon entropy; as the limiting form of Rényi and Tsallis entropies;may be the most outstanding or privileged one due to the simplicity it provides for further mathematical work. For a general discussion of basic concepts and applications of entropy, one may refer to Khinchin (1957), Reza(1994), Ullah (1996) and Cover & Thomas (2006).

1. Entropy for discrete cases

Let the discrete random variable X takes on the values $x_1, x_2, \dots, x_K$ with respective probabilities $p_1, p_2, \dots, p_K$ on some sample space S. Shannon entropy is defined as

$$H = -\sum_{i=1}^{K} p_i log p_i \quad (1.1)$$

For practical considerations it is customary to take the base of the logarithm as 2. So the entropy of X can be evaluated as the minimum average number of bits required to represent the outcome X (Garcia, p169).

2. Bivariate Distributions

Computing the entropies of bivariate distributions are straight forward. Let the random variables X and Y assume the values $x_1, x_2, \dots, x_n \ and \ y_1, y_2, \dots, y_m$ correspondingly. The joint probability function is $P_{X,Y}(x, y)$. For simplicity, we assume that both of the random variables are discrete. Then the Shannon entropy of this joint probabilistic scheme is

$$H(X,Y) = -\sum_{i=1}^{n}\sum_{j=1}^{m} P_{X,Y}(x_i, y_j) log\, P_{X,Y}(x_i, y_j) \quad (2.1)$$

If X and Y are independent,

$$H(X,Y) = H(X) + H(Y) \quad (2.2)$$

Stating that the joint Shannon entropy of two independently distributed random variables is merely the sum of marginal entropies. The extension to n independently distributed random variables $X_1, X_2, \dots, X_n$ is also simple:

$$H(X_1, X_2, \dots, X_n) = \sum_{i=1}^{n} H(X_i) \quad (2.3)$$

This property is also known as the additivity property of Shannon entropy for independent random variables.For dependent variables,

$$H(X, Y) = -\sum_{i=1}^{n} \sum_{j=1}^{m} P_X(x_i) P_Y(y_j / x_i) log P_X(x_i) P_Y(y_j / x_i) \quad (2.4)$$

Here $P_Y(y_j / x_i)$represents the conditional probability of $Y = y_j$given $X = x_i$. Manipulating algebraically a little bit yields

$$H(X, Y) = H(X) + H(Y / X) \quad (2.5)$$

since

$$H(Y / x_i) = -\sum_{j=1}^{m} P_Y(y_j / x_i) (log P_Y(y_j / x_i) \quad (2.6)$$

and

$$H(Y / X) = \sum_{i=1}^{n} P_X(x_i) H(Y / x_i) \quad (2.7)$$

represent conditional entropy of Y (given $= x_i$) and average conditional entropy of Y, respectively. In this case joint entropy is the sum of the entropy of a marginal distribution and the average conditional entropy. This situation implies additivity in terms of marginal and conditional entropies.

3. A measure for mutual information

Suppose the following statistic is defined as

$$I(X = x_i, Y = y_j) = log \frac{P(x_i, y_j)}{P(x_i) P(y_j)} \quad (3.1)$$

It is possible to evaluate this measure as the amount of information that the event

$X = x_i$conveys about the event $Y = y_j$. Note that if these two events are independent then this quantity will necessarily be zero. Yet an average measure that the random variable X conveys about Y may be more convenient. The measure proposed by Shannon for this purpose is

$$I(X,Y) = \sum_{i=1}^{n}\sum_{j=1}^{m} P(x_i,y_j) log \frac{P(x_i,y_j)}{P(x_i)P(y_j)} \quad (3.2)$$

This quantity is called mutual information or relative entropy. Note that for independent random variables this quantity is zero as expected a priori. The following equations can be derived for the relations between mutual information and various entropy measures:

$$I(X,Y) = H(X) + H(Y) - H(X,Y) \quad (3.3)$$
$$I(X,Y) = H(X) - H(X/Y) \quad (3.4)$$
$$I(X,Y) = H(Y) - H(Y/X) \quad (3.5)$$

Whenever X and Y are independent,average conditional entropies are simply equal to marginal entropies. Therefore mutual information will be zero. On the contrary, when X and Y are dependent, marginal entropies will be equal to joint entropy. For this reason, for this example, mutual information is equal to joint entropy. Although the minimum value that mutual information can take is 0, it does not seem to have an upper bound all the time. Especially when the two variables are dependent, mutual information can be infinitely large. Hence some modifications to mutual information are proposed which are as follows:

$$C_1(X,Y) = \frac{I(X,Y)}{H(Y)} \quad (3.6)$$
$$C_2(X,Y) = \frac{I(X,Y)}{H(X)} \quad (3.7)$$
$$C_3(X,Y) = 1 - \frac{H(Y/X)}{H(Y)} \quad (3.8)$$
$$C_4(X,Y) = 1 - \frac{H(X/Y)}{H(X)} \quad (3.9)$$

(3.6) and (3.7) and similarly (3.8) and (3.9) are not necessarily equal. A symmetric one is

$$R(X,Y) = \frac{I(X,Y)}{H(X)+H(Y)} \quad (3.10)$$

Here R is the coefficient of redundancy. Whenever two variables are independent, R is zero. Whenever they are totally dependent, R is equal to ½ implying that the one of the variables is redundant to analyse uncertainty.

3.1. Mutual Information in bivariate normal distribution

Suppose (X,Y) fits a bivariate normal distribution. The entropy of this joint distribution can be found as

$$H(X,Y) = \ln (2\pi\sigma_X\sigma_Y e\sqrt{1-\rho^2}) \quad (3.1.1)$$

The marginal entropies of X and Y are as below:

$$H(X) = \ln (\sigma_X\sqrt{2\pi e}) \quad (3.1.2)$$
$$H(Y) = \ln (\sigma_Y\sqrt{2\pi e}) \quad (3.1.3)$$

Similarly, the mutual information is

$$I(X,Y) = -\frac{1}{2}\ln (1-\rho^2) \quad (3.1.4)$$

It is interesting to see that if the correlation coefficient is zero, mutual information is also zero. In addition, as the absolute value of correlation coefficient increases, the amount of mutual information increases as can be expected intuitively.

3.2. Kullback-Leibler information and relative entropy

Relative entropy is a measure of divergence between two distributions. For discrete cases relative entropy or Kullback-Leibler divergence between p and q is defined as

$$D_{KL}(p \parallel q) = \sum p log\left(\frac{p}{q}\right) \quad (3.2.1)$$

Note that for continuous distributions summation operator in (3.2.1) is simply replaced by integration operator.At the first sight, it may seem strange to qualify both of the measures defined in (3.2) and (3.2.1) as "relative entropy". But relative entropy can also be viewed as the divergence between two hypotheses;$H_0$and H_1. To be more specific for (3.2), the null hypothesis is for the independence of two variables, and hence the null and the alternative hypotheses are

$H_0: P_{X,Y}(x,y) = P_X(x)P_Y(y) for\ all\ (x,y) \in \mathcal{R}^2$

$H_1: P_{X,Y}(x,y) \neq P_X(x)P_Y(y)$

For (3.2.1),by the same reasoning, they are

$H_0: The\ probability\ function\ is\ p$

$H_1: The\ probability\ function\ is\ q \neq p$

For the first formulation given above, we can alternatively state that $I(X,Y) = D_{KL}\left(P_{X,Y}(x,y) \parallel P_X(x)P_Y(y)\right)$.
Mutual information is the Kullback-Leibler divergence between a joint distribution and a hypothesized joint distribution under independence assumption. For the second expression, q may often be any empirical distribution derived or formulated from p by sampling. In that case, relative entropy is a statistical tool for checking goodness of fit.

The reason for using the term "divergence" rather than "distance" is that Kullback-Leibler measure is neither symmetric nor obeys triangle inequality. Therefore Kullback-Leibler measure is not a metric function. It can also be stated that relative entropy is a special case of Rényi divergence (Ullah, p146).Rényi's order-alpha divergence of q from p is defined as

$D_R(p \parallel q) = \frac{1}{\alpha-1} log \sum \frac{p^\alpha}{q^{\alpha-1}}$(3.2.2)

Note that if p is equal to q, this quantity is equal to zero. When $\alpha \to 1$

$$\lim_{\alpha\to 1}\frac{1}{\alpha-1}log\sum\frac{p^{\alpha}}{q^{\alpha-1}}=\frac{0}{0}$$

By L'Hospital's rule

$$\lim_{\alpha\to 1} D_R(p \parallel q) = \lim_{\alpha\to 1}\frac{\frac{d}{d\alpha}\left(log\sum\frac{p^{\alpha}}{q^{\alpha-1}}\right)}{\frac{d}{d\alpha}(\alpha-1)} = \sum(plogp - plogq) = D_{KL}(p \parallel q) \quad (3.2.3)$$

3.3. Jeffreys Divergence as a symmetric version of Kullback-Leibler divergence.

Another statistic which is closely-related to mutual information is Jeffreys divergence. It is a symmetrical version of Kullback-Leibler divergence and defined as follows:

$$D_J(p \parallel q) = D_{KL}(p \parallel q) + D_{KL}(q \parallel p) \quad (3.3.1)$$

For two discrete probability distributions p and q, it is defined as

$$D_J(p \parallel q) = \sum(p-q)log\left(\frac{p}{q}\right) \quad (3.3.2)$$

Jeffreys divergence does not satisfy the conditions of being a metric function either.Itdoes not fulfill the condition of triangle inequality(Kullback,p6).

3.4. Kullback-Leibler and Jeffreys Divergences for some selected probability distributions

Some Kullback-Leibler and Jeffreys measures for some frequently-used probability distributions can directly be derived from (3.2.1) and (3.3.2). They are as follows:

1)For two Bernoulli distributions having parameters p_1 and p_2 respectively,

$$D_{KL}(p \parallel q) = (1-p_1)log\left[\frac{(1-p_1)}{(1-p_2)}\right] + p_1 log\left[\frac{p_1}{p_2}\right] \quad (3.4.1)$$

$$D_J(p \parallel q) = (p_1-p_2)\left[log\left(\frac{1-p_2}{p_2}\right) - log\left(\frac{1-p_1}{p_1}\right)\right] \quad (3.4.2)$$

2)For two binomial distributions having parameters n(common) and p_1 and p_2,

$$D_{KL}(p \parallel q) = n\left[(1-p_1)log\left[\frac{(1-p_1)}{(1-p_2)}\right] + p_1 log\left[\frac{p_1}{p_2}\right]\right] \quad (3.4.3)$$

$$D_J(p \parallel q) = n\left[(p_1-p_2)\left[log\left(\frac{1-p_2}{p_2}\right) - log\left(\frac{1-p_1}{p_1}\right)\right]\right] \quad (3.4.4)$$

Here it should be noted that mutual information (Kullback-Leibler measure) and Jeffreys measure are additive for independent observations (Kullback, pp 12-26). For this reason (3.4.3) and (3.4.4) are simply formulated by multiplying the quantities on the right-hand sides of the equations (3.4.1) and (3.4.2) by "n", since a binomial experiment consists of n independent and identical Bernoulli replicates.

3) For two geometric distributions having parameters p_1 and p_2 respectively,

$$D_{KL}(p \parallel q) = \left(\frac{1-p_1}{p_1}\right) log\left(\frac{1-p_1}{1-p_2}\right) + log\left(\frac{p_1}{p_2}\right) \quad (3.4.5)$$

$$D_J(p \parallel q) = \left[\frac{1-p_1}{p_1} + \frac{1-p_2}{p_2}\right] log\left(\frac{1-p_1}{1-p_2}\right) + 2log\left(\frac{p_1}{p_2}\right) \quad (3.4.6)$$

4) For two negative binomial distributions having parameters r(common) and p_1 and p_2,

$$D_{KL}(p \parallel q) = r\left[\left(\frac{1-p_1}{p_1}\right) log\left(\frac{1-p_1}{1-p_2}\right) + log\left(\frac{p_1}{p_2}\right)\right] \quad (3.4.7)$$

$$D_J(p \parallel q) = r\left[\left[\frac{1-p_1}{p_1} + \frac{1-p_2}{p_2}\right] log\left(\frac{1-p_1}{1-p_2}\right) + 2log\left(\frac{p_1}{p_2}\right)\right] \quad (3.4.8)$$

Again it should be noted that (3.4.7) and (3.4.8) are formulated by multiplying the right-hand sides of (3.4.5) and (3.4.6) by "r".

5)For two Poisson distributions with respective parameters λ_1 and λ_2 ,

$$D_{KL}(p \parallel q) = (\lambda_2 - \lambda_1) + \lambda_1 log\left(\frac{\lambda_1}{\lambda_2}\right) \quad (3.4.9)$$

$$D_J(p \parallel q) = (\lambda_1 - \lambda_2) log\left(\frac{\lambda_1}{\lambda_2}\right) \quad (3.4.10)$$

4)For two normal distributions with parameters (μ_X, σ_X^2) and (μ_Y, σ_Y^2)

$$D_{KL}(f_1 \parallel f_2) = \frac{1}{2}\left[2ln\left(\frac{\sigma_Y}{\sigma_X}\right) + \frac{\sigma_X^2}{\sigma_Y^2} + \left(\frac{\mu_X-\mu_Y}{\sigma_Y}\right)^2 - 1\right] \quad (3.4.11)$$

$$D_J(f_1 \parallel f_2) = \frac{(\sigma_X^2+\sigma_Y^2)}{2\sigma_X^2\sigma_Y^2}[(\sigma_X^2 - \sigma_Y^2) + (\mu_X - \mu_Y)^2] \quad (3.4.12)$$

5) For two exponentially distributed random variables having parameters λ_1 and λ_2

$$D_{KL}(f_1 \parallel f_2) = \frac{(\lambda_2-\lambda_1)}{\lambda_1} + log\left(\frac{\lambda_1}{\lambda_2}\right) \quad (3.4.13)$$

$$D_J(f_1 \parallel f_2) = \frac{(\lambda_2-\lambda_1)^2}{\lambda_2\lambda_1} \quad (3.4.14)$$

6)For two gamma distributions having the probability densities as

$$f_1(x) = \frac{\lambda_1^r x^{r-1} e^{-\lambda_1 x}}{\Gamma(\lambda_1)} \quad \text{and} \quad f_2(x) = \frac{\lambda_2^r x^{r-1} e^{-\lambda_2 x}}{\Gamma(\lambda_2)}$$

$$D_{KL}(f_1 \parallel f_2) = r\left[\frac{(\lambda_2-\lambda_1)}{\lambda_1} + log\left(\frac{\lambda_1}{\lambda_2}\right)\right] \quad (3.4.15)$$

$$D_J(f_1 \parallel f_2) = r\frac{(\lambda_2-\lambda_1)^2}{\lambda_2\lambda_1} \quad (3.4.16)$$

Note too that (3.4.15) and (3.4.16) are obtained by multiplying the expressions on the right-hand sides of (3.4.13) and (3.4.14) by "r" due to the additivity propery of Kullback-Leibler and Jeffreys divergences for independent observations.

3.5. Asymptotic Properties of Kullback-Leibler and Jeffreys divergences

Under some regularity conditions, based on n continuous observations from a population whose probability function is f, $2nD_{KL}(f \parallel \hat{f}) = 2n\int \widehat{f(x)} log\frac{\widehat{f(x)}}{f(x)}dx$ fits asymptotically a chi-square distribution with k-degrees of freedom,where k is the number of parameters of the probability function f.$\hat{f}$ is the estimated probability density function based on sample information. Or as a matter of choice, one can use the test statistic

$nD_J(f \parallel \hat{f}) = 2n\int\left(\widehat{f(x)} - f(x)\right) log\frac{\widehat{f(x)}}{f(x)}dx$ which fits asymptotically a chi-square distribution with k-degrees of freedom (Kullback,pp 97-102).Therefore these two statistics can also be used in some hypothesis tests alternatively. Note that for discrete cases integration operatorsare simply replaced by summation operators.

4. Applications

First we assume a binomial distribution whose parameter (success probability) p is under consideration.Based on sample information (n=100), the sample proportion$\hat{p}$ has been estimated.Here we study three cases;

i) $H_0: p = 0.25 \quad versus H_1: p \neq 0.25$
ii)$H_0: p = 0.50 \quad versus H_1: p \neq 0.50$
iii)$H_0: p = 0.75 \quad versus H_1: p \neq 0.75$

If we suppose the population mean is 0.25 and consider various sample proportions between 0 and 1, we may get the following diagram for classical Z-square scores, Kullback-Leibler divergences calculated for $(p_0, \hat{p})$. The horizontal scale represents sample proportion.

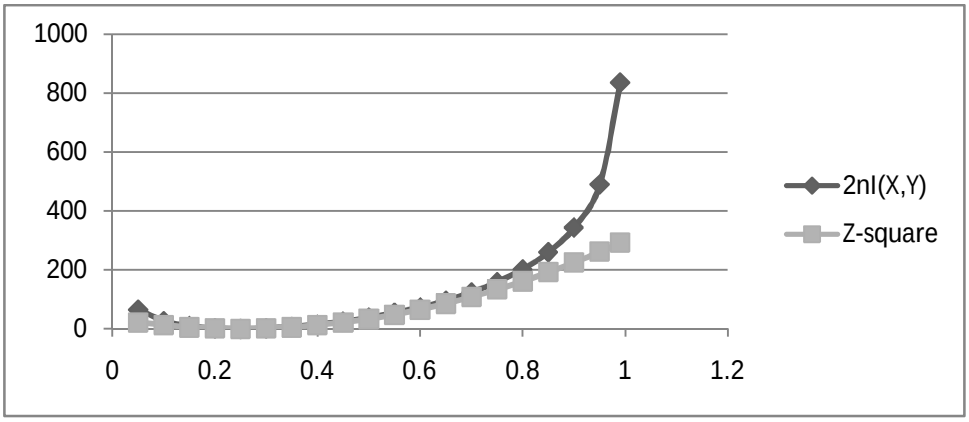

Figure 4.1. Kullback-Leibler divergence and classical Z-square scores for testing the proportion of binomial distribution

It should be noted that Z-square is computed by thefollowing formula:

$$Z^2 = \left(\frac{\hat{p} - E(\hat{p})}{S(\hat{p})}\right)^2 = \left(\frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}\right)^2$$

The square of a standard normal variable fits a chi-square distribution with one-degree of freedom. 2nI(X,Y) also follows achi-square distribution with 1-degree of freedom.We simply have calculated these two statistics to show that they behave similarly. Note that Pearson linear correlation coefficient between these two statistics is 0.923 for this example. Finally we also note that there is not a special reason for prefering 2nI(X,Y) rather than nJ(X,Y),since both of them produce very similar results.For p=0.50 and for p=0.75 we have observed very similar results. Linear correlation coefficients are found to be 0.918 and 0.946 respectively. We observe that generally2nI(X,Y) is more sensitive than ordinary Z-square statisticespecially in case of extreme deviations from the mean. Therefore we can say that for larger deviations from the mean, the power of 2nI(X,Y) (or nJ(X,Y)) test is higher.

As a second illustration, we have considered standard normal distribution. The horizontal axis of the next diagram represents populationmean. By assuming the level of significance 0.05and performing a one sided test for simplicity, we have produced the following diagram:

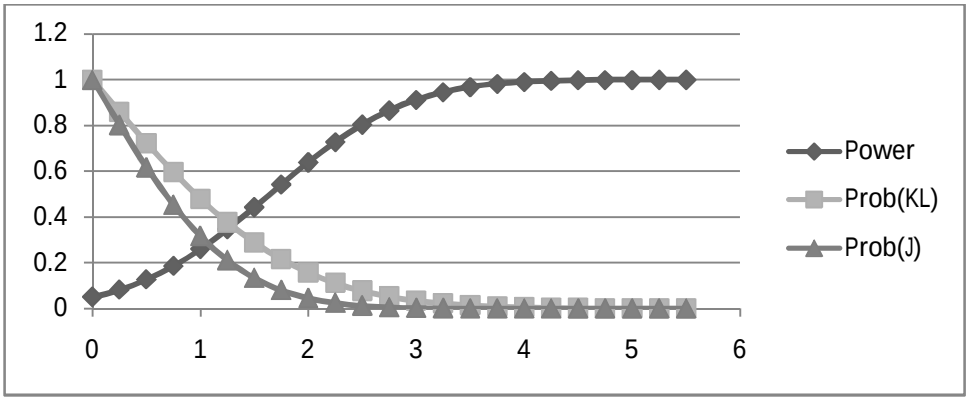

Figure 4.2. P-values for Kullback-Leibler and Jeffreys divergences and the power function of a standard normal variate (alpha=0.05)

Here Prob(KL) and Prob(J) are p-values for Kullback-Leibler and Jeffreys divergences under the null hypothesis that the mean is zero. It is important to note that as the power of a classical Z-test increases, p-values for Kullback-Leibler and Jeffreys divergences decrease also;which isa phenomenon that we expect to observe intuitively. Because as the power increases, we are more confident in rejecting a null hypothesis.

Similarly as p-values decrease, we are more confident to reject a null-hypothesis. We have repeated this procedure twice more for two normal distributions having a common mean of zero and standard deviations 2 and 5 respectively. We have observed a similar tendency.

Finally,we havestudied the sampling properties of 2nI(X,Y) and nJ(X,Y) statistics in testing the parameter of a Poisson distribution. We have simulated 12 different scenarios from Possion distributions by Microsoft Excel. Lambda values have been taken to be 1,5,15 and 30. We have observed samples consisting of 30, 50 and 100 items. Then we have repeated these experiments 1000 times to get a better understanding of sampling propertiesof classical Z-square, 2nI(X,Y) and nJ(X,Y) statistics. Z-square statistic is calculated by the following:

$$Z^2 = \left(\frac{\hat{\lambda} - E(\hat{\lambda})}{Std.Error(\hat{\lambda})}\right)^2 = \left(\frac{\hat{\lambda} - \lambda}{\sqrt{\lambda/n}}\right)^2$$

2nI(X,Y) and nJ(X,Y) are calculated by (3.4.9) and (3.4.10). $\lambda_1$and λ_2 are population and sample means, respectively. The following summarizesthe results of 1000 different simulations.

Simulation	lambda	sample size	Average Z-square	Average2nI(X,Y)	Average nJ(X,Y)
1	1	30	0.948	0.974	0.963
2	1	50	0.989	1.001	0.995
3	1	100	1.045	1.054	1.051
4	5	30	1.121	1.106	1.109
5	5	50	0.958	0.957	0.957
6	5	100	1.027	1.023	1.023
7	15	30	0.948	0.955	0.953
8	15	50	0,992	0.991	0.991
9	15	100	1.16	1.159	1.159
10	30	30	0.984	0.988	0.987
11	30	50	0.96	0.961	0.961
12	30	100	1.021	1.021	1.021

Table 4.1. Avareges of three test statistics

The following table suggests that there is no significant difference between the coefficients of variations ofall the three statistics.

Simulation	Coef. of variation(Z-square)	Coef.of variation(2nI(X,Y))	Coef. of variation(nJ(X,Y)
1	1.481	1.499	1.472
2	1.503	1.502	1.489
3	1.368	1.376	1.368
4	1.332	1.32	1.319
5	1.401	1.404	1.401
6	1.346	1.337	1.339
7	1,432	1.456	1.449
8	1.466	1.463	1.463
9	1.368	1.367	1.367
10	1.451	1.467	1.462
11	1.522	1.526	1.525
12	1.397	1.397	1.397

Table 4.2. Coefficients of variations of three test statistics

It should be noted that all these three statistics are highly and positively correlated. For all 12 simulations,the minimum correlation is found to be 0.944, whereas the maximum correlation is 0.999. Again this tendency can be investigated by Table 3.

	Correlation between Z-square and 2nI(X,Y)	Correlation between Z-square and nJ(X,Y)	Correlation between 2nI(X,Y) and nJ(X,Y)
1	0.944	0.968	0.996
2	0.966	0.981	0.997
3	0.984	0.991	0.999
4	0.989	0.993	0.999
5	0.994	0.996	0.999
6	0.997	0.998	0.999
7	0.996	0.998	0.999
8	0.997	0.998	0.999
9	0.998	0.999	0.999
10	0.998	0.999	0.999
11	0.998	0.999	0.999
12	0.999	0.999	0.999

Table 4.3.Correlations

Finally, we give frequency distributions of these three statistics for the case n=100 and $\lambda = 30$.

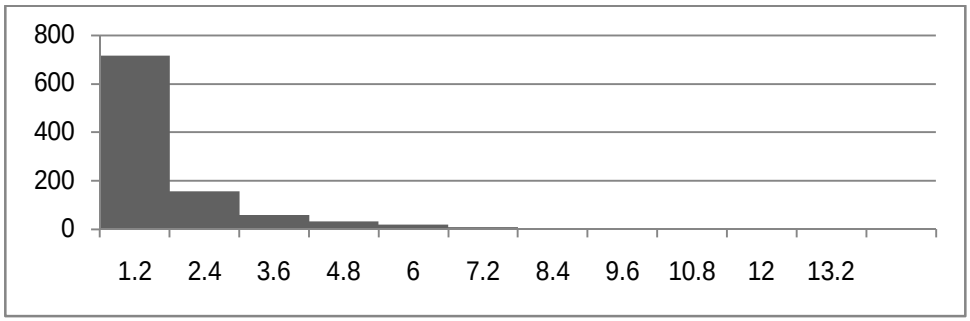

Figure 4.3. Frequency Distribution of Z-square

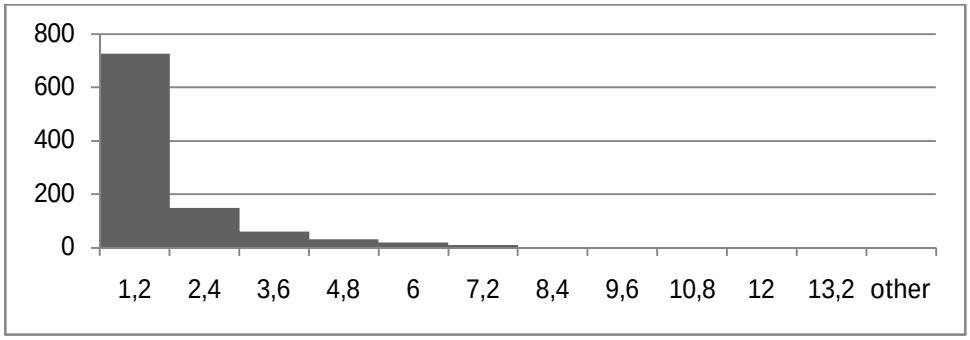

Figure 4.4. Frequency distribution of 2nI(X,Y)

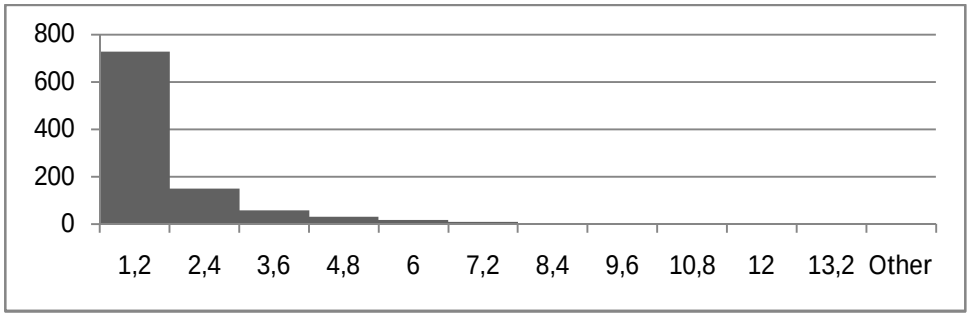

Figure 4.5. Frequency distribution of nJ(X,Y)

Obviously, all three figures suggest that all of the three statistics fit somechi-square distributions.

Discussion

For larger sample sizes, multinomial distributions can be approximated well by multivariate normal distributions. The measures of uncertainty or entropy; in discrete cases; depend on either class frequencies or some functional forms of class frequencies arising from multinomial distributions. Therefore the connection between entropy statistics and normality seems straightforward by multivariate central limit theoremsfor larger sample sizes. For a review on asymptotic normality of entropy measures, one can refer to Esteban & Morales (1995), Pardo (2006) and Zhang (2013).

On the other hand, relative entropy statistics may also be used in goodness of fit testing as well as classical test statistics like Z scores, Z-square scores and classical chi-square statistics. We have observed that in testing any claim on population proportion of binomial distribution, Kullback-Leibler and Jeffreys test statistics have produced consistent results with those of classical Z and Z-square statistics. In addition, we have observed that the power of Kullback-Leibler and Jeffreys test statistics are higher especially in case of extreme deviations from the null hypothesis.Secondly, we have observed that Kullback-Leibler and Jeffreys formalism is totally in agreement with the basic concepts of classical hypothesis testing methodology like p-values, power of a test, etc. Then,we have observed thatrelative entropy statisticsfit some chi-square distributions asymptotically.As a final illustration, under different assumptions on the parameter of Poisson distribution and sample size, we have applied these two statistics in testing the mean of a Poisson distribution to check their asymptotic nature. These two statistics have behaved like ordinary chi-square statistics.

The general trend in literature is a measure-theoretical approach on entropy and relative entropy issues. Yet Kullback-Leibler and Jeffreysdivergences are based on classical likelihood methodology.Therefore our final emphasis is on the existence of a larger set of statistical testing problems which can alternatively be studied by relative entropy methods.

References

COVER, T.M.; THOMAS, J.A.(2006) Elements of Information Theory, Wiley Interscience (Second Edition), Hoboken, New Jersey

GARCIA, A.L.(1994), Probability and Random Processes for Electrical Engineering, Addison-Wesley Longman (Second Edition)

KHINCHIN, A.I.(1957),Mathematical Foundations of Information Theory, Dover Publications

KULLBACK, Solomon.(1996) ; Information Theory and Statistics, Dover Publications, NY

RENYİ, A.(2007), Foundations of Probability, Dover Publications,NY

REZA, ,F. M.(1994) ; An Introduction to Information Theory, Dover Publications, NY

ESTEBAN, M.D., MORALES, D. (1995), *A summary on entropy statistics, Kybernetika*, Vol. 31(1995), No.4, 337-346

PARDO, L. (2006), Statistical Inference Measures Based on Divergence Measures, Chapman&Hall/CRC, 99.

ULLAH, A.(1996), *Entropy, Divergence and Distance Measures with Econometric Applications*, Journal of Statistical Planning and Inference, 49(1996), 137-162

ZHANG, Xing (2013), *Asymptotic Normality of Entropy Estimators, The University of North Carolina at Charlotte*
http://math.uncc.edu/sites/math.uncc.edu/files/2013_03.pdf