ON SOME PROPERTIES OF GOODNESS OF FIT MEASURES BASED ON STATISTICAL ENTROPY

Atif Evren[1] & Elif Tuna[2]
[1,2] Yildiz Technical University Faculty of Sciences and Literature , Department of Statistics
Davutpasa, Esenler, 34210, Istanbul Turkey

ABSTRACT
Goodness of fit tests can be categorized under several ways. One categorization may be due to the differences between observed and expected frequencies. The other categorization may be based upon the differences between some values of distribution functions. Still the other one may be based upon the divergence of one distribution from the other. Some widely used and well known divergences like Kullback-Leibler divergence or Jeffreys divergence are based on entropy concepts. In this study, we compare some basic goodness of fit tests in terms of their statistical properties with some applications.

Keywords: *Measures for goodness of fit, likelihood ratio, power divergence statistic, Kullback-Leibler divergence, Jeffreys' divergence, Hellinger distance, Bhattacharya divergence.*

1. INTRODUCTION

Boltzmann may be the first scientist who emphasized the probabilistic meaning of thermodynamical entropy. For him, the entropy of a physical system is a measure of disorder related to it [35]. For probability distributions, on the other hand, the general idea is that observing a random variable whose value is known is uninformative. Entropy is zero if there is unit probability at a single point, whereas if the distribution is widely dispersed over a large number of individually small probabilities, the entropy is high [10].
Statistical entropy has some conflicting explanations so that sometimes it measures two complementary conceptions like information and lack of information. Shannon relates it with positive information, while Brillouin associates it with lack of information and ignorance [3].

2. SOME APPLICATIONS OF ENTROPY IN STATISTICS.

2.1. Entropy as a Measure of Qualitative variation

Entropy is used to determine the degree of variability of a probability distribution, especially when the random variable is qualitative. For qualitative distributions, it is impossible to calculate arithmetic mean as a measure of central tendency. Hence, mode is used as the average value. Similarly, standard deviation and variance can not be calculated as a measure of variation in turn. In such cases, entropy measures as well as other qualitative variation statistics like coefficient of variation, index of diversity etc. as summarized by Wilcox [47] can be used.

2.2. Relative Entropy as a Measure of Goodness of Fit

Relative entropy (or Kullback-Leibler divergence) is a measure of discrepancy between two statistical hypotheses or two probability distributions so that it can serve as a measure of goodness of fit of any distribution to the other.
Besides, entropy and entropy concepts are related to some goodness of fit statistics of any statistical or econometric model to the data at hand. Suppose the likelihood function is given as $l(\theta / y) = f(y / \theta)$ where y is observed. One can find that the expected likelihood is the negative of entropy associated with the sampling density $f(y / \theta)$.This introduces entropy as an information criterion in statistics and tells us that $\hat{\theta}$ is associated with minimizing the disorder or lack of information about the sampling model. General information criterion (GIC), Akaike's information criterion (AIC) and Bayesian information criterion (BIC) are closely related to entropy by the statistic $\log l(\hat{\theta} / y)$ [24]. For example, suppose that some data are generated by an unknown process f. We consider two alternatives or competing models g_1 and g_2 to represent f. If we knew f, then we could find the information loss using g_1 to represent f by Kullback-Leibler divergence between g_1 and f. Similarly, information loss from using g_2 to represent f would be calculated by Kullback-Leibler divergence between g_2 and f. Here it should be noted that Kullback-Leibler divergence is relative entropy. At the same time, the quantity of information

lost using g_2 instead of g_1 to represent f, can also be estimated by AIC [50] . In other words, for large samples AIC relies on Kullback-Leibler information as averaged entropy. [5] .

2.3. Entropy-based Divergence Measures in Evaluating Sample Information

In Bayesian statistics, evaluating the value of sample information by comparing the entropies of prior and posterior distributions can be realized similarly. These evaluations can be especially easier for the cases in which prior and posterior distributions are in conjugate nature. Besides Kullback-Leibler divergence is closely related to Fisher information [48]. Fisher Information is the curvature (second derivative) of the Kullback-Leibler divergence of the distribution $f(X;\theta)$ from the true distribution $f(X;\theta_0)$ [51]

2.4. Mutual Information as a Measure of Association between Qualitative Variables

Mutual information is a measure of information that one random variable contains about the other. In other words, it is the reduction in the uncertainty of one random variable due to the knowledge of the other. Therefore entropy-based measures like mutual information can also be used in analysing associations and correlations between variables.

2.5. Entropy-based Statistics and Some Concepts of Hypothesis Testing

Entropy-based measures are closely related to some statistical hypothesis testing concepts like type-I and type-II errors [21]. Besides, one can consult Kalbfleisch [18] to study some of the applications of entropy concepts in hypothesis testing, since Kullback-Leibler divergence is the negative logarithm of average multinomial loglikelihood [40].

2.6. Entropy-based Statistics in Multivariate Comparisons

Jeffreys divergence is the symmetric version of relative entropy (or Kullback-Leibler divergence). Jeffreys divergence between two multivariate normal distributions can be employed in some multivariate problems, since it is closely related to Hotelling's T^2 [21].

2.7. Entropy-based Statistics in Residual Analysis on Linear Regression Models.

Entropy-based divergence measures between two multivariate normal distributions can also be helpful in analyzing residuals obtained by linear regression models (especially in calculating Cook's distances).

2.8. Maximum Entropy as the Principle of Insufficient Reason.

If an event can be produced by a number of n different causes, the probabilities of these causes given that the event has already occurred, are supposed to be equal to each other. This is Laplace's well-known principle of insufficient reason. His principle suggested that in the absence of any reason, all the causes should be taken equally likely, a priori. By following Laplace, one can easily estimate the probability of observing a head if a coin is flipped once as 0.5. In this case the principle of insufficient reasoning coincides with the principle of maximum entropy! On the other hand, for some Bayesian considerations, some scientists, may prefer to select a prior density that expresses the idea of knowing little (or a maximum entropy, minimum information prior distribution). Therefore prior distribution and entropy are related to each other [32]. If we use the maximum entropy principle to assign sampling distributions, this automatically generates the distributions with the most desirable properties from the standpoint of inference in either sampling theory or Bayesian theory [17].

2.9. Entropy-based Parameter Estimation

In some optimization problems like the method of least squares, the general approach is to define a loss function and then try to find the estimator giving the minimum loss. In a similar manner, Kullback-Leibler divergence or Jeffreys divergence can be taken into consideration as a loss function. In such cases, the parameter that minimizes Kullback-Leibler or Jeffreys divergence can be investigated. Besides, some information criteria have been developed based on Kullback-Leibler divergence for Bayesian model selection [41]. One can find some applications of entropy-based parameter estimation in hydrology in Singh[39].

2.10. Sufficient Statistics and Entropy Concepts

In many of the estimation problems we summarize the information in the sample $x_1, x_2, \ldots, x_n$ by a function (or a statistic). Nontechnically speaking, such a function that tells us as much about the parameter(s) of the distribution as the sample itself is said to be sufficient. On the other hand, "a more technical definition" says that a statistic

$S = s(x_1, x_2, ..., x_n)$ is sufficient if the conditional distribution of the sample given the value of the statistic does not depend on a parameter (or parameters) of the distribution. The idea is that if you know the value of sufficient statistic, then the sample values themselves are not needed and can tell you nothing more about parameter(s) [26]. This is the same as the condition $I(\theta; X) = I(\theta; T(X))$ Here, θ is the parameter of the parent distribution from which the sample X comes, and T(X) is a sufficient statistic. Hence sufficient statistics preserve mutual information. The relationship of mutual information and sufficiency is due to Kullback. [21]

3. METRIC FUNCTIONS AND DIVERGENCE

In mathematics, a metric (distance) function, $\rho(x, y)$ satisfies the following requirements for all $(x, y) \in \Re^n$:

(i) $\rho(x, y) \geq 0$; (ii) $\rho(x, y) = \rho(y, x)$; (iii) $\rho(x, y) = 0$ only for x=y ;

(iv) $\rho(x, y) + \rho(y, z) \geq \rho(x, z)$ (triangular inequality) [9]

Some measures proposed for determining the goodness of fit, do not possess all properties summarized from (i) to (iv). Therefore they are rather called as "divergence measures".

4. CLASSIFICATION OF GOODNESS OF FIT TESTS

In statistics, goodness of a fit is the closeness of agreement between a set of observations and a hypothetical model that is currently suggested [27]. Loosely speaking, there are three types of goodness of fit families (although some overlaps are possible) widely used in statistical applications:

i) Goodness of fit measures derived from empirical probability distribution functions. Kolmogorov-Smirnov statistic, Cramér-Von-Mises statistic and Anderson-Darling statistic belong to this family[23].

ii) Goodness of fit measures based on the discrepancies between observed and expected frequencies. Likelihood ratio statistic and chi-square goodness-of-fit statistics belong to this category.

iii) Measures derived from entropy and relative entropy concepts. Among them, Kullback-Leibler divergence, Jeffreys divergence are the most popular representatives. Indeed these divergences can be studied with their close relations with Shannon entropy, Renyi entropy, and Tsallis entropy. There is a rapidly-growing literature on different divergence measures obtained by different parametrizations. It should be emphasized that among these divergence families, ϕ (or f) divergence deserves special attention. Cressie and Read have shown that some measures from some of these categories can also be derived through different parametrizations of power divergence statistic.

4.1. Testing Goodness of Fit by Comparing Empirical Distribution Functions

The tests in this category are harder to work with numerically, but they tend to have good power relative to many alternatives. The empirical distribution function (EDF) tests are generally more powerful than the chi-squared goodness of fit test because they make more direct use of the individual observations [2]. Some measures in this category are Kolmogorov-Smirnov, Cramér-Von Mises, and Anderson-Darling statistics.

4.1.1. Kolmogorov-Smirnov test

Kolmogorov-Smirnov (K-S) test which is more appropriate for continuous distributions is suitable for small samples. K-S test requires that the null distribution function $F_0(x)$ be completely specified such that the functional form as well as parameters should be known [28].

Suppose $x_{(1)}, x_{(2)}, ..., x_{(n)}$ are the realizations of order statistics $X_{(1)}, X_{(2)}, ..., X_{(n)}$ from $F_0(x)$. The sample or empirical distribution function is defined as

$$F_n(x) = \begin{cases} 0, & x < x_{(1)} \\ \dfrac{j}{n}, & x_{(j)} \leq x < x_{(j+1)},\ j = 1,2,...,n-1 \\ 1, & x \geq x_{(n)} \end{cases} \quad (1)$$

It can be shown that

$$P(F_n(x) = \frac{j}{n}) = \binom{n}{j} (F_0(x))^j (1 - F_0(x))^{n-j}, \quad j = 0,1,...,n \quad (2)$$

with the mean and variance

$$E(F_n(x)) = F_0(x) \tag{3}$$

$$Var(F_n(x)) = \frac{F_0(x)(1-F_0(x))}{n} \tag{4}$$

respectively. The sample distribution function $F_n(x)$ converges in probability to $F_0(x)$ for all real x. Hence for large n (sample size), $|F_n(x) - F_0(x)|$ should be small. Besides the sampling distribution of $F_n(x)$ is asymptotically normal. The random variable D, known as K-S statistic

$$D = \sup|F_n(x) - F_0(x)| \tag{5}$$

is used in choosing one of the competing alternatives

$$H_0 : F_n(x) = F_0(x) \quad \text{for } -\infty < x < \infty, \tag{6}$$
$$H_1 : F_n(x) \neq F_0(x)$$

D is distribution-free and its sampling distribution does not depend upon $F_0(x)$ provided $F_0(x)$ is continuous. In addition, test procedure is exact for any n [28]. If $F_0(x)$ is discrete, the same methodology can still be followed to reach approximate conclusions [26]. In this case the null distribution is no longer exact [37]. The limiting distribution of D is

$$\lim_{n\to\infty} P(D > \frac{z}{\sqrt{n}}) = 2\sum_{k=1}^{\infty}(-1)^{k-1}\exp(-2k^2z^2) \tag{7}$$

The approximation to (7) when z>1 yields the following [25]:

$$P(D > \frac{z}{\sqrt{n}}) \cong 2\exp(-2z^2) \tag{8}$$

When the sample size is considerably low, K-S test is preferred rather than an ordinary chi-square test [8]. K-S test is consistent so that its power against any $H_1 : F_n(x) \neq F_0(x)$ tends to 1 as $n \to \infty$ [22]

4.1.2. Cramér-Von Mises (C-M) statistic

If two distributions are continuous, C-M statistic is defined as below:

$$\omega^2 = \int_{-\infty}^{\infty}\left(F_n(x) - F_0(x)\right)^2 \psi(x)\, dF_0(x) \tag{9}$$

Here $\psi(x)$ is a non-negative weighting function. C-M and Anderson&Darling accepted it as $\psi(x) = 1$ and $\psi(x) = [F(x)(1-F(x))]^{-1}$, respectively [42]. Here $F_n(x) = j/n$, where j is the number of sample values that are less than or equal to x. We have

$$\omega^2 = \frac{1}{12n^2} + \frac{1}{n}\sum_{j=1}^{n}\left(F_n(x_{(j)}) - \frac{2j-1}{2n}\right)^2 \tag{10}$$

$$E(\omega^2) = \frac{1}{6n} \tag{11}$$

$$Var(\omega^2) = \frac{4n-3}{180n^3} \tag{12}$$

Smirnov has shown that as $n \to \infty$; $n\omega^2$ has a limiting non-normal distribution [7]. This test has been adapted for use with discrete random variables, for cases where parameters have to be estimated from the data, and for comparing two samples [46]. One can refer to Serfling [38] for studying asymptotic distribution theory of C-M test statistic.

4.1.3. Anderson-Darling (A-D) test

A-D test is a modification of the C-M test. The test statistic A^2 is

$$A^2 = -\frac{1}{n}\sum_{j=1}^{n}(2j-1)\left[\ln\left\{F_n(x_{(j)})\right\} + \ln\left\{1-F_n\left(x_{(j)}\right)\right\}\right] - n \tag{13}$$

C-M and A-D tests are independent of $F_0(x)$. They both tend to be more powerful than K-S test. A-D test, in particular, is more sensitive to departures from $F_0(x)$ in the tails of the distribution than either K-S test or C-M test since the factor $\psi(x) = \left[F(x)(1-F(x))\right]^{-1} \to \infty$ as $x \to \pm\infty$ [22].

4.2 Testing Goodness of Fit by the Discrepancies between Observed and Expected Frequencies

4.2.1. Likelihood ratio and chi-square statistics

Let X is from k disjoint categories with probabilities $p_1, p_2, \ldots, p_k$ with respective frequencies $n_1, n_2, \ldots, n_k$

The likelihood function is

$$L(p) = {p_1}^{n_1}{p_2}^{n_2}\ldots{p_k}^{n_k} \tag{14}$$

Taking logarithms and differentiating with respect to p_i will yield the maximum likelihood estimate of p_i (which is denoted by $\hat{p}_i$)

$$\hat{p}_i = \frac{n_i}{n} \quad i = 1,2,\ldots,k \tag{15}$$

Now consider the likelihood ratio test for testing the hypothesis

$$H_0 : p_i = p_{i0} \quad i = 1,2,\ldots,k \tag{16}$$

Then it follows that the likelihood ratio is given by

$$\lambda = \frac{L_0(\hat{p})}{L(\hat{p})} = \frac{{p_{10}}^{n_1}{p_{20}}^{n_2}\ldots{p_{k0}}^{n_k}}{\left(\frac{n_1}{n}\right)^{n_1}\left(\frac{n_2}{n}\right)^{n_2}\ldots\left(\frac{n_k}{n}\right)^{n_k}} \tag{17}$$

The large sample approximate value of $-2\log(\lambda)$ is given by

$$-2\log(\lambda) = -2\sum_{i=1}^{k} n_i \log\left(\frac{e_i}{n_i}\right) \tag{18}$$

$$-2\log(\lambda) \sim \sum_{i=1}^{k}\frac{(n_i - e_i)^2}{e_i} \tag{19}$$

where $e_i = np_{i0}$ (20)

This statistic in (19) was proposed by Pearson to measure the goodness of fit. For large n values this statistic fits a chi-square distribution with k-1 degrees of freedom [16].

$$\sum_{i=1}^{k}\frac{(n_i - e_i)^2}{e_i} \sim \chi^2_{k-1} \tag{21}$$

Instead of using (21) as a measure of fit, some statisticians prefer G^2 statistic defined as

$$G^2 = -2\log(\lambda) \tag{22}$$

This statistic is called the likelihood ratio chi-squared statistic. A serious disadvantage in using (18) is that the approximation is usually poor while the approximation in (21) holds up for smaller value of n [8]. Stuart et al. [42] indicate that

$$G^2 = \chi^2(1 + O_p(n^{-1/2})) \tag{23}$$

Agresti emphasizes that when H_0 is true and n is increasing, χ^2 and G^2 increase steadily but do not have similar values even for very large n values [1]. The use of G^2 statistic as a goodness of fit measure is criticized in literature because G^2 is designed to detect any departure between a model and observed data. In this sense, the

likelihood ratio test often rejects acceptable models [31]. To apply a chi-square test in the general case, it is necessary that the sample size n and the class numbers n_i be sufficiently large. Practically it is sufficient that n ~50-60 and n_i ~5-8 [42]. In practice, it is common to "pool" the end classes of the distribution until the n_i's reach a satisfactory size. On the other side, too much pooling reduces the chance of rejecting H_0, if it really should be rejected [19]. If the χ^2 test is based on a very large number of observations, we can almost be certain that the tail area corresponding to the observed chi-square statistic will be very small. For this reason, a very small tail area should not be regarded as evidence against H_0 without further analysis [12]. Besides, if we do not reject H_0 we cannot conclude that the model is correct. We may have failed to reject simply because the test did not have enough power [48]. With small samples the test is not very sensitive [15].

4.2.2. Freeman-Tukey T statistic

If X is a Poisson variable with expectation λ, then for $\lambda > 1$
It is defined as the transformed variable Z, given by

$$Z = \sqrt{X} + \sqrt{X+1} - \sqrt{4\lambda + 1} \tag{24}$$

has an approximate standard normal distribution. Based on this transformation, as proposed in a 1950 paper by M.F. Freeman, the following is a goodness of fit statistic for discrete cases

$$T = \sum_{i=1}^{k} \left(\sqrt{o_i} + \sqrt{o_i + 1} - \sqrt{4e_i + 1} \right)^2 \tag{25}$$

which fits a χ^2_{k-s-1} asymptotically where s is the number of parameters [13].

4.2.3. Power-Divergence (P-D) statistic

Another goodness of fit statistic was proposed by Cressie and Read. It is defined as

$$P - D = \frac{2}{\lambda(\lambda+1)} \sum o_i \left[\left(\frac{o_i}{e_i} \right)^{\lambda} - 1 \right] \qquad -\infty < \lambda < \infty \tag{26}$$

The importance of this statistic lies in its generality. For example for $\lambda = 1$ one can obtain χ^2, and for $\lambda \to 0$ one can get G^2.Besides, for $\lambda = -\frac{1}{2}$ Freeman-Tukey statistic, and for $\lambda \to -1$ Kullback's minimum discriminating information, $\lambda = -2$, Neyman modified chi-squared statistic can be derived [1]. Cressie and Read showed that under H_0, all different parametrizations of λ yield equivalent results asymptotically [43]. Note that one can obtain Cressie-Read statistic by substituting $\lambda = 2/3$ in (26) [29].

4.3. Entropy and Statistical Measures of Uncertainty

Statistical entropy can be evaluated as a measure of uncertainty before a statistical experiment takes place [20]. Information gained means a decrease in uncertainty [45]. Therefore, in some sense, entropy can be viewed as the amount of information that can be gathered through statistical experimentation [36]
Some of the entropy measures proposed for discrete cases are as follows:

$$H_s(p) = -\sum p_i \log p_i \quad \text{(Shannon entropy(1948))} \tag{27}$$

$$H_\alpha(p) = \frac{1 - \sum p_i^\alpha}{\alpha - 1}, \alpha > 0, \alpha \neq 1 \quad \text{(Havrda\&Charvát(1967) and Tsallis(1988))[44]} \tag{28}$$

$$H_R(p) = \frac{\log \sum p_i^\alpha}{1 - \alpha}, \alpha > 0, \alpha \neq 1 \quad \text{(Rényi (1961))[34]} \tag{29}$$

For continuous distributions, summation operators are replaced by integration operators.

4.3.1. Rényi's divergence

Rényi's order-α divergence of q(x) from p(x) is defined as

$$D_R(p\,//\,q) = \frac{1}{\alpha - 1}\log \int_{-\infty}^{\infty} p(x)\left(\frac{p(x)}{q(x)}\right)^{\alpha-1} dx \tag{30}$$

which has the following properties:

i) $D_R(p\,//\,q) \geq 0; \forall p,q,\alpha > 0$; ii) $D_R(p\,//\,q) = 0 \quad iff \quad p(x) = q(x) \quad \forall x \in \Re$;

iii) $\lim_{\alpha \to 1} D_R(p\,//\,q) = D_{KL}(p\,//\,q)$ [25].

Here $D_R(p\,//\,q)$ and $D_{KL}(p\,//\,q)$ represents Rényi divergence and Kullback-Leibler (K-L) divergence between two probability distributions p(x) and q(x) respectively. Note also that if q is a kind of uniform distribution such that $q(x) = 1 \quad for \quad 0 < x < 1$ then Rényi's divergence reduces to Rényi entropy given in (29) [15].

Rényi entropies are important in echology and statistics as indices of diversity [53].

4.3.2. Kullback-Leibler information and relative entropy

The limit of Rényi divergence when $\alpha \to 1$ is relative entropy. Relative entropy or Kullback-Leibler divergence is a measure of discrepancy between p and q. For discrete cases

$$D_{KL}(p\,//\,q) = \sum_x p(x)\log\frac{p(x)}{q(x)} \tag{31}$$

This measure is also known as cross entropy or information theoretical divergence of p from q [35]. Several explanations on (31) are possible. First, K-L divergence is called the gain of information resulting from the replacement of a priori distribution p by a posteriori distribution q [34]. Secondly, it can be evaluated as the measure of inefficiency of assuming that the distribution is q when the true distribution is p. K-L divergence can also be evaluated as the mean information for discrimination in favor of p against q per observation from p[1] [21].

K-L divergence is always nonnegative and does not fulfill all requirements of metric functions. For example it is in general not commutative;

$$D_{KL}(p\,//\,q) \neq D_{KL}(q\,//\,p) \tag{32}$$

4.3.3. Jeffreys' divergence as a symmetric version of Kullback-Leibler divergence

Jeffreys' divergence (or sometimes J-divergence[11]) is defined to be

$$D_J(p\,//\,q) = D_{KL}(p\,//\,q) + D_{KL}(q\,//\,p) \tag{33}$$

$D_J(p\,//\,q)$ is a measure of difficulty of making a discrimination between p and q. Note that $D_J(p\,//\,q)$ is symmetric with respect to p and q. It has all properties of a metric except triangular inequality and is therefore not called a distance [21].

4.3.4. Jensen-Shannon divergence

A symmetric and finite measure is Jensen-Shannon divergence which is defined in terms of K-L divergences.

$$D_{JS}(p\,//\,q) = \frac{1}{2}D_{KL}(p\,//\,m) + \frac{1}{2}D_{KL}(q\,//\,m) \tag{34}$$

where

$$m(x) = \frac{1}{2}p(x) + \frac{1}{2}q(x) \text{ [4]} \tag{35}$$

4.3.5. Asymptotic properties of $D_{KL}(p\,//\,q)$ and $D_J(p\,//\,q)$

Suppose $\hat{I} = D_{KL}(p\,//\,q)$ (36)

[1] Similarly $D_{KL}(q\,//\,p)$ can be seen as the mean information from q in favor of q against p.

$$\hat{J} = D_J(p // q) \tag{37}$$

The following statistic based on a sample of n continuous observations

$$2n\hat{I} = 2n\left(\int f(x,\theta) \log \frac{f(x,\theta)}{f(x,\theta_2)} d\lambda(x) \right)_{\theta = \hat{\theta}} \tag{38}$$

(under the assumption $\theta \neq \theta_2$) fits asymptotically a chi-square distribution with k (the elements of parameter vector) degrees of freedom. Here $f(x,\theta)$ is a general multivariate probability distribution having multiple parameters. $\hat{\theta}$. It is also assumed to be a consistent, asymptotically multivariate normal, efficient random vector estimator of θ. The statistic

$$n\hat{J} = n\left(\int (f(x,\theta) - f(x,\theta_2)) \log \frac{f(x,\theta)}{f(x,\theta_2)} d\lambda(x) \right)_{\theta = \hat{\theta}} \tag{39}$$

is a chi-square variate with k degrees of freedom, asymptotically. These tests are consistent, the power tends to 1 for large samples [21].

4.3.6. **Bhattacharyya divergence**

$$\rho. = \int_{-\infty}^{\infty} \sqrt{p(x)q(x)} dx \tag{40}$$

is defined as Bhattacharyya coefficient. The quantity ρ is also known as the affinity between p and q. Bhattacharyya divergence between p and q is defined to be

$$D_B(p // q) = -\ln(\rho) \tag{41}$$

Although D_B is frequently called as a "distance" in literature, it does not obey the triangle inequality [52]. For that reason, we will rather call it as a "divergence". Note that $0 \le \rho \le 1$ [17]. Therefore one can conclude that $0 \le D_B(p // q) < \infty$. Note also that

$$D_B(p // q) = \frac{1}{2} D_R(p // q) \quad for \quad \alpha = 1/2 \text{ in (30)} \tag{42}$$

The generalized Bhattacharya divergence or Chernoff divergence is defined by

$$D_{GB}(p // q) = -\ln\left(\int (p(x))^{1-s} (q(x))^s dx \right) \quad 0 < s < 1 \tag{43}$$

4.3.7. Hellinger distance

Hellinger distance between two univariate and continuous probability distributions p and q is

$$D_H(p // q) = \left[\int \left(\sqrt{p(x)} - \sqrt{q(x)} \right)^2 dx \right]^{(1/2)} {}^2 \tag{44}$$

$$D_H(p // q) = \sqrt{2(1-\rho)} \tag{45}$$

Compared with the Kullback-Leibler and Renyi's divergences, Hellinger distance avoids stability problems when the denominator probability density function is zero [33].

---

[2] Sometimes Hellinger distance in literature is rather defined as $D_H{}^2(p // q) = \frac{1}{2}\left[\int \left(\sqrt{p(x)} - \sqrt{q(x)} \right)^2 dx \right]$

If one uses this formulation then Hellinger distance is limited between 0 and 1. The maximum distance 1 is achieved when p assigns probability zero to every set to which q assigns a positive probability, and vice versa. Otherwise by (44) it is limited between 0 and $\sqrt{2}$.

For a multinomial population with c categories whose cell frequencies and probabilities are $n_1, n_2, ..., n_c$ and $p_1, p_2, ..., p_c$ respectively, where $\sum_{i=1}^{c} n_i = n$, Ferguson [14] states that a transformed chi-square statistic ${\chi_H}^2$ has the following form

$${\chi_H}^2 = 4n\sum_{i=1}^{c}\left(\sqrt{n_i / n} - \sqrt{p_i}\right)^2 \tag{46}$$

Therefore ${\chi_H}^2$ is known as Hellinger χ^2 because of its relation to Hellinger distance in (44).

Bhattacharya divergence and Hellinger distance are related to each other. In other words,

$$D_H(p // q) = \sqrt{2(1-\exp(-D_B(p // q)} \tag{47}$$

Compared with the K-L and Renyi's divergences, Hellinger's distance has the advantage of being a difference of probability density functions so it avoids stability problems when the denominator probability density function is zero. DasGupta [11] also points out that

$$D_H(p // q) \le \sqrt{D_{KL}(p // q)} \tag{48}$$

4.3.8. Further generalizations and ϕ divergences

Various types of generalizations can still be offered. For a detailed exposition of these generalizations, one can consult Cichocki and Amari [6] . At this point, only one type of generalization, namely, ϕ divergence will be taken into consideration. ϕ divergence as introduced by Czissár[30] is defined as

$$D_f(p // q) = \sum q(x) f\left(\frac{p(x)}{q(x)}\right) \tag{49}$$

with a convex function f defined for x>0. For the convex function f(x)=xlog(x) the relative entropy is obtained. If $f(x) = |x-1|$, one can obtain total variation or variational distance as

$$D_f(p // q) = \sum_x |p(x) - q(x)| \tag{50}$$

For $f(x) = \left(1-\sqrt{x}\right)^2$, one can find the square of Hellinger distance as

$$D_f(p // q) = 2\left(1 - \sum_x \sqrt{p(x)q(x)}\right) = 2(1-\rho) = D_H^2(p // q) \tag{51}$$

From a statistical point of view, the most important family of ϕ divergences is perhaps the family studied by Cressie and Read; the power divergence family, given by

$$D_{\phi(\lambda)}(p // q) = \frac{1}{\lambda(\lambda+1)}\left[\sum_x p(x)\left[\frac{p(x)}{q(x)}\right]^{\lambda} - 1\right] \quad \text{for } -\infty < \lambda < \infty \tag{52}$$

Note the similarity between (52) and (26). The power divergence family is undefined for $\lambda = -1$ or $\lambda = 0$. However these cases are defined as below:

$$\lim_{\lambda \to 0} D_{\phi(\lambda)}(p // q) = D_{KL}(p // q) \tag{53}$$

$$\lim_{\lambda \to -1} D_{\phi(\lambda)}(p // q) = D_{KL}(q // p) \tag{54}$$

5. APPLICATION

We have tried to determine the degree of correlation of some chi-square statistics (either based or not based on entropy measures). At the first part, we have simulated some data from an arbitrarily chosen categorical (nominal or ordinal) distribution. In other words, we have intended the compare performances of goodness of fit statistics when the null hypothesis is true. Here it should be emphasized that, the skewness and kurtosis of data are immaterial since all chi-square statistics focus mainly on the differences between observed and expected frequencies. At the second part we compared the performances of such statistics when the observed and expected frequencies differ on a scale

varying from moderate cases to extreme cases. Although various other distributions can be considered, the parent distribution we have considered is a discrete uniform distribution. It is assumed as below:

Table 1.1: Parent distribution

Groups	Probabilities	Groups	Probabilities
1	0.1	6	0.1
2	0.1	7	0.1
3	0.1	8	0.1
4	0.1	9	0.1
5	0.1	10	0.1

Then we have generated 30 samples from this population. Here the sample size of each simulation is 100. Our reason for restricting the sample size at a level of 100 lies in the fact that all chi-square tests focus on the differences between observed and expected frequencies. Therefore simulating 1000 or 10000 times instead of 100, especially when the null hypothesis is true, would not be necessary. Note that in the following table what is meant by Kullback-Leibler chi-square, Jeffreys chi-square and Hellinger chi-square are the statistics calculated by (38), (39) and (46) respectively. Note too that Cressie-Read statistic is a special case of power-divergence statistic with $\lambda = 2/3$. At the end of each simulation we have calculated the goodness of fit statistics as follows:

Table 1.2: Goodness of fit results when H0 is true.

Simulation no	Chi-square	G-square	Freman-Tukey	Cressie-Read stat.	Kullback-Leibler Chi-square	Jeffreys Chi-Square	Hellinger Chi-Square
1	7.2	7.19	6.88	7.17	7.36	7.28	7.25
2	11	11.25	10.88	11.03	11.96	11.61	11.54
3	13.4	15.67	16.50	13.90	21.64	18.66	17.98
4	8.6	9.01	8.81	8.7	9.78	9.39	9.34
5	9.2	8.91	8.41	9.07	8.89	8.90	8.87
6	8.2	8.04	7.64	8.12	8.13	8.08	8.05
7	8	8.74	8.75	8.19	10.14	9.44	9.34
8	7.4	9.09	9.63	7.85	12.24	10.67	10.42
9	16	17.55	17.66	16.36	20.98	19.26	18.96
10	9	9.85	9.82	9.23	11.21	10.53	10.46
11	5.8	5.62	5.32	5.73	5.57	5.60	5.58
12	8.4	8.98	8.85	8.56	9.92	9.45	9.40
13	5.6	6.62	6.85	5.88	8.23	7.42	7.33
14	6.8	6.39	5.95	6.64	6.13	6.26	6.24
15	8.4	8.17	7.73	8.29	8.18	8.17	8.14
16	13.8	14.29	13.95	13.88	15.67	14.98	14.85
17	11.6	10.77	10.00	11.27	10.40	10.58	10.53
18	8.8	9.03	8.76	8.84	9.61	9.32	9.27
19	13.6	13.46	12.84	13.49	13.90	13.68	13.60
20	5.6	5.93	5.83	5.69	6.44	6.18	6.16
21	11.8	12.21	11.85	11.89	13.07	12.64	12.58
22	3.4	3.54	3.46	3.44	3.73	3.63	3.62
23	6.8	6.61	6.26	6.72	6.61	6.61	6.59
24	3.4	3.12	2.91	3.30	2.92	3.02	3.01
25	5.6	5.65	5.43	5.60	5.82	5.74	5.72
26	3.4	3.41	3.29	3.40	3.47	3.44	3.43
27	4.8	4.78	4.58	4.78	4.84	4.81	4.80
28	17.2	21.15	22.46	18.22	29.31	25.23	24.47
29	15.6	17.23	17.29	16.02	20.25	18.74	18.53
30	6.2	6.261	6.01	6.20	6.42	6.34	6.33

From the table given above we can see that all the chi-square statistics usually agree. This situation can also be investigated by the simplified correlation matrix of these statistics.

Table 1.3: Correlations between different goodness of fit statistics when H0 is true.

	G-square	Freman-Tukey	Cressie-Read statistic	Kullback-Leibler Chi-Square	Jeffreys Chi-Square	Hellinger Chi-Square
Chi-square	0.98	0.96	0.99	0.93	0.95	0.96
G-square		0.99	0.99	0.98	0.99	0.99
Freman-Tukey			0.98	0.99	0.99	1
Cressie-Read stat.				0.94	0.97	0.97
Kullb-Leibler Chi-Square					0.99	0.99
Jeffreys Chi-Square						1

Secondly, we tried to determine the degree of goodness of fit of a distribution to the other. In other words, we tried to determine the degree of association of different goodness of fit statistics when observed distribution is different from expected (or hypothesized) distribution to some extent**.** (By "to some extent" we mean the discrepancy between the two distributions vary from moderate departures to extreme departures.) What we have obtained is given in Table 2.1.

Table2.1 Goodness of fit results when observed distribution is different from expected (or hypothesized) distribution to some extent.

Trials	Chi-square	G-square	Freman-Tukey	Cressie-Read statistic	Kullback-Leibler Chi-Square	Jeffreys Chi-Square	Hellinger Chi-Square
1	22	21	19.93	21.46	21.85	21.42	21.19
2	27	30.35	31.82	27.52	43.92	37.14	35.02
3	38.4	33.37	30.49	36.26	31.94	32.66	32.27
4	34	34.95	34.56	33.87	41.18	38.06	37.19
5	11.8	11.63	11.1	11.68	12.01	11.82	11.75
6	60	62.31	63.21	59.5	82.14	72.23	68.97
7	14	15.11	15.15	14.22	17.97	16.54	16.26
8	729	360.45	305.2	538.92	370.3	365.38	330.98
9	44.3	49.55	42.48	47.04	49.33	45.76	44.7
10	2	2.01	1.96	2	2.03	2.02	2.02
11	18.4	15.96	14.44	17.44	14.67	15.32	15.2
12	20.8	22.18	22.12	21.04	26.11	24.15	23.75
13	32.4	32.46	31.71	32.02	36.81	34.64	34
14	88	84.79	85.22	84.39	113.1	98.94	93.35
15	224.6	178.12	170.89	199.32	223.97	201.05	187.06
16	111.8	107.87	108.85	107.15	146.75	127.31	119.56
17	28	30.03	31.14	28	42.83	36.43	34.26
18	44.6	45.16	45.26	43.92	58.19	51.68	49.3
19	38.2	35.49	33.32	36.91	36.22	35.85	35.42
20	23	21.02	19.46	22.17	20.46	20.74	20.58
21	24.2	22.5	21.01	23.45	22.37	22.43	22.25
22	18.4	19.21	18.72	18.58	20.93	20.07	19.94
23	39.8	39.16	39.13	38.69	50.87	45.01	42.69
24	33.4	31.51	29.73	32.46	32.39	31.95	31.61
25	32.4	31.38	30.15	31.68	34.08	32.73	32.21
26	42.2	40.77	39.07	41.24	43.94	42.35	41.72
27	33	36.15	36.48	33.64	43.99	40.07	39.32
28	22.2	20.14	18.6	21.34	19.49	19.82	19.66
29	31.8	29.79	28.22	30.75	31.28	30.53	30.07
30	53	50.2	48.98	51.03	60.88	55.54	53.1

To clarify what was summarized in Table 2.1, we should emphasize that the first raw of Table 2.1 displays the values of various fit measures when the hypothesized and observed distributions are as below:

Table 2.2: The expected and observed frequencies for the first raw of Table 2.1

Groups	expected frequencies(ei)	observed frequencies
1	10	21
2	10	5
3	10	7
4	10	11
5	10	4
6	10	12
7	10	12
8	10	10
9	10	6
10	10	12
Total	100	100

It should also be noted that at each trial we tried to determine the degree of fit of a distribution to the other with 10 groups. Here it will be superfluous to construct tables like Table 2.2 . Therefore we summarized the values of different goodness of fit statistics by the help of Table 2.1 and also by the table given below:

Table 2.3: Correlations between different goodness of fit statistics when observed and expected frequencies are differed to some extent.

	G-square	Freman-Tukey	Cressie-Read statistic	Kullback-Leibler Chi-Square	Jeffreys Chi-Square	Hellinger Chi-Square
Chi-square	0.97	0.95	0.99	0.94	0.96	0.95
G-square		0.99	0.99	0.99	0.99	0.99
Freman-Tukey			0.97	0.99	1	1
Cressie-Read stat.				0.96	0.97	0.97
Kullback-Leibler Chi-Square					0.99	0.99
Jeffreys Chi-Square						1

Again we see state that the goodness of fit measures above are closely correlated to each other.

6. CONCLUSION

The concepts of entropy with their close relations with thermodynamics have some appealing aspects for the statisticians not only because of some philosophical concerns. Rather, asymptotic distributions of some entropy based measures like Kullback-Leibler divergence and Jeffreys divergence have been studied in detail in literature. In other words, these divergences or statistics are linked to the chi-square statistic for a contingency table by likelihood function as briefly mentioned in (18). In our first two examples, we wanted to show the close relations of these statistics with other goodness of fit statistics like, χ^2, G^2, power-divergence statistic, Cressie-Read statistic, and Hellinger chi-square. We limited ourselves and focused only on categorical (nominal or ordinal) distributions. For that reason goodness of fit statistics arised from distribution functions (i.e., Kolmogorov-Smirnov, Cramér-Von Mises, Anderson-Darling test statistics) were not taken into consideration since they required continuous probability distributions. Although we did not care much about asymptotic properties, these two statistics, namely Kullback-Leibler divergence and Jeffreys divergence were not worse than the other chi-square statistics in the applications. Infact we found that all these statistics were highly correlated with each other. This is true for the stuation that the null hypothesis (the distributions are identical) is true as well as for the situations that the observed and expected distributions may differ from each other considerably.

7. REFERENCES

[1]. AGRESTI, A., *Categorical Data Analysis*, Wiley Interscience (Second Edition), Hoboken, New Jersey, 24, 112,395-396 (2002).

[2]. BAIN, J.L., ENGELHARDT,M., *Introduction to Probability and Mathematical Statistics*, Second edition, Duxbury Classic Series, 457-462 (1992) .

[3]. BRISSAUD, J.P., *The Meanings of Entropy*, Entropy, 7[1],68-96 (2005).

[4]. BUDKA, M.,GABRYS,B.,MUSIAL, K, *On Accuracy of PDF Divergence Estimators and Their Applicability to Representative Data Sampling*, Entropy, 13,1229-1266 (2011).

[5]. BURNHAM, K.P., ANDERSON,D.R., *Kullback-Leibler Information as a Basis for Strong Inference in Echological Studies*, Wildlife Research, 28, 111-119 (2001).

[6]. CICHOCKI,A.,AMARI,S., *Families of Alpha-Beta and Gamma-Divergences:Flexible and Robust Measures of Similarities*, Entropy, 2010, 12, 1532-1568 (2010).

[7]. CRAMER ,H., *Mathematical Methods of Statistics*, Princeton University Press, Nineteenth Printing and First Paper Printing 1999, 416-452 (1999).

[8]. CONOVER, W.J., *Practical Nonparametric Statistics*, Wiley Series in Probability and Statistics, Third Edition, 259-430 (1999).

[9]. COVER, T.M.; THOMAS, J.A., *Elements of Information Theory*, Wiley Interscience (Second Edition), Hoboken, New Jersey, 45 (2006).

[10]. COX, D.R., *Principles of Statistical Inference*, Cambridge University Press, 76 (2006).

[11]. DASGUPTA, A., *Asymptotic Theory of Statistics and Probability*, Springer Texts in Statistics, 20-21 (2008).

[12]. DE GROOT, M.H., SCHERVISH, M.J., *Probability and Statistics*, Addison Wesley, International Edition, 3rd edition., 540 (2002).

[13]. EVERITT,B.S., *The Cambridge Dictionary of Statistics* , Cambridge University Press (Third Edition), Cambridge (2006).

[14]. FERGUSON, T.S., *A Course in Large Sample Theory*, Texts in Statistical Science, Chapman&Hall, 59 (2002).

[15]. HERO, A.O.MA,B.,MICHEL,O.,GORMAN,J.(2002), Alpha-Divergence for Classification,Indexing and Retrieval (Revised 2), *Communications and Signal Processing Labaratory*, *Technical Report CSPL-328*, May 2001 (revised Dec 2002), http:// www.eechs.umich.edu/~hero, (2002).

[16]. HOEL, P.G. , *Introduction to Mathematical Statistics*, Fourth edition,Wiley International Edition, 365-367 (1971).

[17]. JAYNES, E.T., *Probability Theory The Logic of Science*; Cambridge University Press, 365-52, (2005).

[18]. KALBFLEISCH, J.G., *Probability and Statistical Inference*, *Volume 2: Statistical Inference*, Second edition, Springer Texts in Statistics, Springer-Verlag, 156-174 (1985)

[19]. KEEPING, E.S.(1995), *Introduction to Statistical Inference*, Dover Publications, NY ,252-254 (1995).

[20]. KHINCHIN, A.I., *Mathematical Foundations of Information Theory*, Dover Publications, NY, 7 (1957).

[21]. KULLBACK, S., *Information Theory and Statistics*, Dover Publications, NY , 5-100 (1996).

[22]. LEHMANN, E.L., *Elements of Large-Sample Theory*, Springer Texts in Statistics, Spriner, 342-346 (1999).

[23]. LEHMANN,E.L.,ROMANO,J.P., *Testing Statistical Hypotheses*,Third Edition, Springer Texts in Statistics, Springer, 527-630 (2005).

[24]. LEONARD,T., HSU,J.S.J. *Bayesian Methods,An Analysis for Statisticians and Interdisciplinary Researchers*, Cambridge Series in Statistical and Probabilistic Mathematics, 8 (2001).

[25]. LINDGREN, B.W. , *Statistical Theory*, Fourth Edition, Chapman&Hall/CRC, 485 (1993).

[26]. MOOD, A.M.,GRAYBILL,F.A.,BOES,D.C., Introduction to the Theory of Statistics,Third edition, McGraw-Hill International Editions, Statistics series, 299-301 (1974).

[27]. NELSON, D., *Penguin Dictionary of Mathematics*, Penguin Reference Library (2008).

[28]. PANIK,M.J., *Advanced Statistics from an Elementary Point of View*,Elsevier Academic Press, 621-625 (2005).

[29]. PARDO, L, S*tatistical Inference Based on Divergence Measures*, Chapman&Hall/CRC, Taylor&Francis Group, 7-115 (2006).

[30]. PETS, D. , *From f-divergence to Quantum quasi-entropies and their use*, Entropy, 12, 304-325 (2010).

[31]. POWERS, D.A.,XIE,Y., *Statistical Methods for Categorical Data Analysis*, Academic Press, 106 (2000).

[32]. PRESS, S.J., *Subjective and Objective Bayesian Statistics, Principles,Models and Applications*, Wiley-Interscience, Second Edtion, 72-244 (2003).

[33]. PRINCIPE,J.C., *Information Theoretic Learning*, Renyi's Entropy and Kernel Perspectives, Springer, 81-85 (2010).

[34]. RAO,C.R., *Convexity Properties of Entropy Functions and Analysis of Diversity*, Inequalities in Statistics and Probability, IMS Lecture Notes-Monograph Series Vol. 5 (1984), 67-77 (1982).
[35]. RENYI, A, *Probability Theory*, Dover Publications, NY, 554 (2007).
[36]. RENYI, A., *Foundations of Probability*, Dover Publications,NY, 23, 211 (2007).
[37]. ROHATGI, V.K., *Statistical Inference*, Dover Publications, NY, 757-758 (2003).
[38]. SERFLING, R.J., *Approximation Theorems of Mathematical Statistics*, Wiley Series in Probability and Statistics,Wiley, 64 (2002).
[39]. SINGH, V.P., *Entropy-Based Parameter Estimation in Hydrology*, Water Science and Technology Library,Volume 30 Kluwer Academic Publishers, (1999).
[40]. SHLENS,J , *Notes on Kullback-Leibler Divergence and Likelihood Theory*, Systems Neurobiology Laboratory, Salk Institute for Biological Studies, La Jolla, CA 92037, www.snl.salk.edu (2007).
[41]. SOOFI,E.S.,*Information Theory and Bayesian Statistics*, in Bayesian Analysis and Econometrics, edited by Donald A. Berry , Katryn M. Chaloner, and John K. Geweke, John Wiley&Sons, Inc., 180 (1996).
[42]. SVESHNIKOV, A.A., *Problems in Probability Theory, Mathematical Statistics and Theory of Random Functions*, Dover Publications, NY, 301 (1968).
[43]. STUART,A. ORD,K. ARNOLD,S., *Kendall's Advanced Theory of Statistics,* Volume 2-A, Classical Inference&The Linear Model, Arnold, 389-420 (1999).
[44]. TEIXERIA, A.,MATOS,A.,SOUTO,A.,ANTUNES,L., *Entropy Measures vs. Kolmogorov Complexity*, Entropy, 13, 595-611 (2011).
[45]. TOPSØE, *Basic Concepts, Identities and Inequalities –the Toolkit of Information Theory*, Entropy, 3, 162-90 (2001).
[46]. UPTON, G.; COOK, I., *Oxford Dictionary of Statistics*, Oxford University Press (Second edition), NY (2006).
[47]. WILCOX, A.R,"Indices of Qualitative Variation", Oak Ridge National Laboratory; ORNL-TM-1919 (1967).
[48]. WILKS, S.S., *Mathematical Statistics*, Toppan Printing Company Ltd.,Japan, 409-410 (1962).
[49]. WASSERMAN, L., *All of Statistics*, A Concise Course in Statistical Inference, Springer Texts in Statistics, Springer, 169 (2004).

Internet Resources

[50] http://en.wikipedia.org/wiki/Akaike_information_criterion
[51]http://en.wikipedia.org/wiki/Fisher_information
[52]http://en.wikipedia.org/wiki/Bhattacharyya_distance
[53]http://en.wikipedia.org/wiki/R%C3%A9nyi_entropy