Journal of Selçuk University Natural and Applied Science
Online ISSN: 2147-3781
www.josunas.org

Some applications of Kullback-Leibler and Jeffreys' divergences in multinomial populations

Atıf EVREN[a*]

[a]*Yildiz Technical Univ., Faculty of Sciences and Literature, Dept. of Statistics, Esenler, 34210, Istanbul, Turkey*

Abstract

Some of the entropy measures proposed are Shannon entropy, Rényi entropy, and Tsallis entropy. The limit of Rényi divergence is relative entropy (or Kullback-Leibler divergence) which is a measure of discrepancy between two statistical hypotheses or two probability distributions. Jeffreys' divergence is a measure of difficulty of making discrimination between two probability distributions. These divergence measures are related to some chi-square distributions asymptotically such that they can be used in some hypothesis tests. In this study, I try to show that entropy based statistics like Kullback-Leibler divergence and Jeffreys' divergence can be used in some statistical hypothesis tests for multinomial populations by some examples.

Keywords: Goodness of fit; Jeffreys' divergence; Kullback-Leibler divergence; Shannon entropy.

Introduction

Statistical entropy can be evaluated as a measure of uncertainty before a statistical experiment takes place (Khinchin, 1957). Information gained means a decrease in uncertainty (Topsoe, 2001). Therefore, in some sense, entropy can be viewed as the amount of information that can be gathered through statistical experimentation (Rényi, 2007). Secondly, entropy is used to determine the degree of variability of a probability distribution, especially when the random variable is qualitative. In such cases, entropy measures can be used as well as other qualitative variation statistics like coefficient of variability, index of diversity etc. These are summarized by Wilcox in detail (Wilcox, 1967).

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

* Corresponding author. Tel.: +90 533 345 49 73 ; fax: +90 212 383 44 16; e-mail: aevren@yildiz.edu.tr

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 (Press, 2003).

Relative entropy is very popular in statistics especially in model selection, goodness of fit testing, Bayesian modeling, Bayesian asymptotics and parameter estimation based on minimum distance (DasGupta, 2008). Relative entropy (or Kullback-Leibler divergence) is a measure of discrepancy between two probability distributions, one of which is probably a distribution stated by the null hypothesis while the other being an empirical one, based on data at hand. Mutual information is a measure of information that one random variable contains about the other. 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 be used in analysing associations between categoric variables.

1. Entropy and Statistical Measures of Uncertainty

Some of the entropy measures proposed for discrete cases are as follows:

1. $Shannon\ Entropy$: $H_s(p) = -\sum p_i \log(p_i)$ (1)
2. $Tsallis\ entropy$: $H_\alpha(p) = \frac{1-\sum p_i^\alpha}{\alpha-1}, \alpha > 0, \alpha \neq 1$(Teixeria et al, 2011) (2)
3. $Rényi\ Entropy$: $H_R(p) = \frac{log \sum p_i^\alpha}{1-\alpha}, \alpha > 0, \alpha \neq 1$(Rao, 1984) (3)

2. Rényi's Divergence

Rényi's order-alpha divergence of **q** from **p** is defined as

$$D_R(pq) = \frac{1}{\alpha-1} log \sum \frac{p(x)^\alpha}{q(x)^{\alpha-1}} \quad (4)$$

It has the following properties:

- $D_R(pq) \geq 0; for\ all\ p\ and\ q, \alpha > 0$
- $D_R(pq) = 0;\ if\ p = q$
- $lim_{\alpha \to 1} D_R(pq) = D_{KL}(pq)$ (Pets, 2010)

Here $D_R(pq)$ and $D_{KL}(pq)$ represents Rényi divergence and Kullback-Leibler (K-L) divergence between two probability distributions **p** and **q**, respectively. Note 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 (Hero et al., 2002). Rényi entropies are important in echology and statistics as indices of diversity .

3. Further Generalizations and φ Divergences

Various types of generalizations can still be proposed. For a detailed exposition of these generalizations, one can consult Cichocki and Amari (2010). One type of generalization, namely, φ divergence as introduced by Czissár is defined as

$$D_{\varphi}(pq) = \sum q(x)\varphi\left(\frac{p(x)}{q(x)}\right) \tag{5}$$

with a convex function φ defined for x>0 (Pets, 2010).

3.1. Kullback-Leibler divergence

The limit of Rényi divergence ($\alpha \to 1$) is relative entropy. For discrete cases

$$D_{KL}(pq) = \sum_x p(x) log\left(\frac{p(x)}{q(x)}\right) \tag{6}$$

This measure is also known as cross entropy. K-L divergence 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**. It is always nonnegative and does not fulfill all requirements of metric functions. It is in general not commutative (Principe 2010).

$$D_{KL}(pq) \neq D_{KL}(qp) \tag{7}$$

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

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

$$D_J(pq) = D_{KL}(pq) + D_{KL}(qp) \tag{8}$$

$D_J(pq)$ is a measure of difficulty of making a discrimination between **p** and **q**. It has all properties of a metric function except triangular inequality .

3.3. Asymptotic properties of Kullback-Leibler and Jeffreys' divergences

Suppose

$$\hat{I} = D_{KL}(pq) \tag{9}$$

$$\hat{J} = D_J(pq) \tag{10}$$

The following statistic based on a sample of **n** continuous observations

$$2n\hat{I} = 2n\left(\int f(x,\theta)\log\frac{f(x,\theta)}{f(x,\theta_2)}d\lambda(x)\right)_{\theta=\hat{\theta}} \tag{11}$$

(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. $\hat{\theta}$ is also assumed to be a consistent, asymptotically multivariate normal, efficient random vector estimator of θ (Kullback, 1996).

3.4. Some modifications

If instead of a single sample, we have **r** independent samples of size n_i ; i=1,2,…,r and each with a consistent , asymptotically multivariate normal, efficient estimator of $\widehat{\theta_i} = (\widehat{\theta_{i1}}, \widehat{\theta_{i2}}, \dots, \widehat{\theta_{ik}})$, i=1, 2,…,r, then under regularity conditions,

$$2\sum_{i=1}^{r} n_i\widehat{I_i} = 2\sum_{i=1}^{r} n_i\left[\int f(x,\theta_i)log\frac{f(x,\theta_i)}{f(x,\theta)}d\lambda(x)\right]_{\theta_i=\widehat{\theta_i}} \tag{12}$$

is asymptotically distributed as chi-square with **rk** degrees of freedom (Kullback, 1996). If the null hypothesis that **r** samples are from the same population whose functional form is known, but with unspecified parameters

$$2\sum_{i=1}^{r} n_i \widehat{I_i} = 2\sum_{i=1}^{r} n_i \left[\int f(x,\theta_i) log \frac{f(x,\theta_i)}{f(x,\theta)} d\lambda(x)\right]_{\substack{\theta_i=\widehat{\theta_i} \\ \theta=\widehat{\theta}}} \quad (13)$$

is asymptotically distributed as chi-square with **(r-1)k** degrees of freedom, where n_i is the number of independent observations in the **ith** sample, $\widehat{\theta_i}$ is a consistent, asymptotically multivariate normal, efficient estimator of the **k** parameters for the **ith** sample, and $n\hat{\theta} = n_1\widehat{\theta_1} + \cdots + n_r\widehat{\theta_r}$ and $n = n_1 + \cdots + n_r$.

When the null hypothesis is not true, (11), (12) and (13) converge in probability to an indefinitely large number and that the large-sample distribution may be approximated by a distribution related to the noncentral chi-square distribution with a large noncentrality parameter and the same number of degrees of freedom as the chi-square distribution under the null hypothesis.

The statistic

$$n\hat{J} = n\left(\int \left(f(x,\theta) - f(x,\theta_2)\right) \log \frac{f(x,\theta)}{f(x,\theta_2)} d\lambda(x)\right)_{\theta=\hat{\theta}} \quad (14)$$

is a chi-square variate with **k** degrees of freedom, asymptotically when the sample is from the population specified by $f(x;\theta_2)$

$$\sum_{i=1}^{r} n_i \left[\int (f(x;\theta_i) - f(x;\theta)) log \frac{f(x;\theta_i)}{f(x;\theta)} d\lambda(x)\right]_{\theta_i=\widehat{\theta_i}} \quad (15)$$

is asymptotically distributed as chi-square with **rk** degrees of freedom if the **r** samples are from the population specified by $f(x;\theta)$. On the other hand, the following statistic

$$\sum_{i=1}^{r} n_i \left[\int (f(x;\theta_i) - f(x;\theta)) log \frac{f(x;\theta_i)}{f(x;\theta)} d\lambda(x)\right]_{\substack{\theta_i=\widehat{\theta_i} \\ \theta=\widehat{\theta}}} \quad (16)$$

is asymptotically distributed as chi-square as **(r-1)k** degrees of freedom. For two samples,

$$\frac{n_1 n_2}{n_1+n_2}\left[\int (f(x;\theta_1)-f(x;\theta_2))log\frac{f(x;\theta_1)}{f(x;\theta_2)}d\lambda(x)\right]_{\substack{\theta_1=\widehat{\theta_1}\\ \theta_2=\widehat{\theta_2}}} \quad (17)$$

is asymptotically distributed as chi-square with **k** degrees of freedom when the two independent samples are from the same population with unspecified vector parameter $\hat{\theta}$. These tests are consistent, the power tends to 1 for large samples. The behavior of the estimates of the divergence when the null hypothesis is not true is similar to that of the estimates of the discriminant information (Kullback, 1996).

4. Multinomial Populations

Suppose two simple hypotheses, H_1 and H_2 specify the probabilities of two hypothetical c- valued populations (c categories) $H_i: p_{i1}, p_{i2}, \dots, p_{ic}; i = 1,2, \dots, n$

The mean information per observation from H_1, for discriminating for H_1 against H_2 is

$$I(1:2) = \sum_{j=1}^{c} p_{1j} log\frac{p_{1j}}{p_{2j}} \quad (18)$$

The mean information per observation from H_2 , for discriminating for H_2 against H_1 is

$$I(2:1) = \sum_{j=1}^{c} p_{2j} log\frac{p_{2j}}{p_{1j}} \quad (19)$$

The divergence between H_1 and H_2, a measure of difficulty of discriminating between them is J(1:2)=I(1:2)+I(2:1)

$$J(1:2) = \sum_{j=1}^{c} (p_{2j}-p_{1j}) log\frac{p_{2j}}{p_{1j}} \quad (20)$$

The mean discrimination information and divergence for a random sample of **n** independent observations O_n, are

$$I(1:2;O_n) = n\sum_{j=1}^{c} p_{1j} log\frac{p_{1j}}{p_{2j}} \quad (21)$$

$$I(2:1;O_n) = n\sum_{j=1}^{c} p_{2j} log\frac{p_{2j}}{p_{1j}} \quad (22)$$

$$J(1:2;O_n) = n\sum_{j=1}^{c} (p_{2j}-p_{1j}) log\frac{p_{2j}}{p_{1j}} \quad (23)$$

5. Applications

1) Suppose that we roll a die 120 times to check whether it is fair or not and we have also observed the frequencies summarized by the following table:

Table 5. 1

Outcome	Observed frequency (o_i)	Expected frequency (e_i)
1	20	20
2	30	20
3	18	20
4	22	20
5	20	20
6	10	20

In other words, our null and alternative hypotheses are as follows:

$H_1: p_1 = p_2 = \cdots = p_6 = 1/6$

$H_2: Not\ all\ p_i's\ are\ \ equal\ to\ eachother. (i = 1,2, \ldots ,6)$

Our ordinary chi-square and I(1:2); and J(1:2) statistics are computed as below:

Table 5.2

Outcome	o_i	e_i	$(o_i - e_i)^2/e_i$	p_i	q_i	$p_i log_2(\frac{p_i}{q_i}))$	$(p_i - q_i) log_2(\frac{p_i}{q_i}))$)
1	20	20	0	0,166	0,166	0	0
2	30	20	5	0,166	0,25	-0,097	0,0487
3	18	20	0,2	0,166	0,15	0,0253	0,0025
4	22	20	0,2	0,166	0,183	-0,0229	0,0022
5	20	20	0	0,166	0,166	0	0
6	10	20	5	0,166	0,083	0,166	0,0833
Total	120	120	10,4	1	1	0,0715	0,1369

and the summarizing statistics are found to be;

- Chi-square=10,4 ;
- I(1:2)=2*120* 0,0715=17,18 ;
- J(1:2)=120*0,1369=16,428.
- The critical chi-square value (with 5 degrees of freedom, and α=0,10) is 11,070,

So with 90% confidence I(1:2) and J(1:2) statistics do not support the null hypothesis that the die is well-balanced (fair) while ordinary chi-square statistic supporting this claim. It is important to note that especially, when the null hypothesis is false, entropy-based statistics tend to reject the null hypothesis more frequently than the ordinary chi-square statistics.

2) As a second example, we used the data of Turkish Ministry of Health for the distribution of medical doctors according to their titles, employed either in university hospitals or in other hospitals. By comparing these figures we try to investigate any significant difference among the proportions of each title from two categories.

Table 5.3 Distribution of Some Health Care Personnel by Sectors and Titles, 2010 (TC Sağlık Bakanlığı Sağlık İstatistikleri Yıllığı 2010)

	Ministry of Health	University	Total
Specialist Physicians	31527	11843	43370
General Practitioners	33229	262	33491
Medical Residents	7679	13340	21019
Total	72435	25445	97880

The distributions of medical doctors according to their titles in either hospital category constitute two multinomial distributions having three categories. The null hypothesis is that exactly the same proportion of special physicians; general practitioners, and medical residents are employed in both of the categories. The alternative hypothesis claims that at least one of the proportions is significantly different from its counterpart. Dividing each element in each column of Table 5.3 by corresponding totals given at the end of each column, one can estimate the probabilities as given below:

Table 5.4

Probabilities	p_i	q_i
Specialist Physicians	0,435245392	0,465435253
General Practitioners	0,458742321	0,010296718
Medical Residents	0,106012287	0,524268029
Total	1	1

Before going into hypothesis testing procedure, one can roughly deduce that these two distributions differ each other considerably by a quick visual check. For instance, the percent of medical doctors in the hospitals of Ministry of Health as general practitioners is 45% while only 1% of the medical doctors are general practitioners in university hospitals. We obtain a common probability scheme representing both of the distributions of each category. How we estimate the common probabilities is as follows:

$$\begin{bmatrix} Estimated \\ Probabilities \\ (Pest) \end{bmatrix} = \frac{72435}{97880}\begin{bmatrix} \mathbf{0,435245392} \\ \mathbf{0,458742321} \\ \mathbf{0,106012287} \end{bmatrix} + \frac{25445}{97880}\begin{bmatrix} \mathbf{0,465435253} \\ \mathbf{0,010296718} \\ \mathbf{0,524268029} \end{bmatrix}$$

$$\begin{bmatrix} Estimated \\ Probabilities \\ (Pest) \end{bmatrix} = \begin{bmatrix} \mathbf{0,443093584} \\ \mathbf{0,342163874} \\ \mathbf{0,214742542} \end{bmatrix}$$

Estimated probabilities are just the weighted averages of the corresponding probabilities from each category. The weights are calculated by dividing the total number of medical doctors from one category to the total number of medical doctors employed in either way. In this case, the null hypothesis states that these two vectors of probabilities (p_i and q_i probabilities) are taken from the population whose probability distribution is "P_{est}". To compute some entropy-based statistics (those mentioned in (12), (15) and (17) the following tables will be useful:

Table 5.5

	p_i	q_i	P_{est}	$p_i log_2(\frac{p_i}{P_{est}})$	$q_i log_2(\frac{q_i}{P_{est}})$
	0,435245	0,465435	0,443094	-0,0112	0,033
	0,458742	0,010297	0,342164	0,194	-0,052
	0,106012	0,524268	0,214743	-0,107	0,675
TOTAL	1	1	1	0,0748	0,656

Table 5.6

	p_i	q_i	P_{est}	$(p_i - P_{est}) log_2(\frac{p_i}{P_{est}})$	$(q_i - P_{est}) log_2(\frac{q_i}{P_{est}})$	$(p_i - q_i) log_2(\frac{p_i}{q_i})$
	0,435	0,4654	0,443	0,0002	0,0015	0,0029
	0,459	0,0103	0,342	0,0493	1,6773	2,456
	0,106	0,5243	0,215	0,1107	0,398	0,964
TOTAL	1	1	1	0,1602	2,0775	3,423

The summarizing statistics are

- Modified K-L divergence (12) = 44233,78
- 1st Modified J- divergence (16) = 64470,71
- 2nd Modified J-divergence (17) = 64470,71 .

In addition, ordinary chi-square statistic is found to be 26445,95. Although there are some differences in these goodness of fit statistics, all take very high values and the null hypothesis is rejected definitely.

Conclusion

In this exposition, I tried to discuss that some entropy-based statistics can be used generally in goodness of fit tests, and particularly in goodness of fit tests on multinomial populations. Within the limitations of this study and due to some artificial and real data I have, I conclude that entropy-based statistics have given similar results to traditional chi-square statistic. From the second example, it can be stated that entropy-based statistics have reached larger values than the ordinary chi-square statistic especially when the null hypothesis is false. Indeed, this is a fact emphasized in literature (Powers & Xie, 2000). In addition, one can consult Hoel (1971) to see the close relation between relative entropy and ordinary chi-square statistic asymptotically. In other words, Kullback-Leibler divergence or relative entropy and Jeffreys' divergence are closely related to "likelihood principle" such that these concepts should be taken into consideration beyond the limitations of information theory.

References

Cichocki, A., Amari, S., Families of Alpha-Beta and Gamma-Divergences: Flexible and Robust Measures of Similarities, Entropy, 2010, 12, 1532-1568

DasGupta, A., Asymptotic Theory of Statistics and Probability, Springer Texts in Statistics, 20-21, 2008

Hero, A.O. Ma, B., Michel, O., Gorman, J., Alpha-Divergence for Classification, Indexing and Retrieval (Revised 2), Communications and Signal Processing Laboratory, Technical Report CSPL-328, May 2001 (revised Dec 2002), http:// www.eechs.umich.edu/~hero, 2002

Hoel, P.G., Introduction to Mathematical Statistics, Fourth edition, Wiley International Edition, 365-367 (1971).

Khinchin, A.I., Mathematical Foundations of Information Theory, Dover Publications, NY, 7, 1957

Kullback, S., Information Theory and Statistics, Dover Publications, NY , 5-100, 1996

Pets, D., From f-divergence to Quantum quasi-entropies and their use, Entropy, 2010, 12, 304-325

Powers, D.A., Xie,Y., Statistical Methods for Categorical Data Analysis, Academic Press, 106 (2000)

Press, S.J., Subjective and Objective Bayesian Statistics, Principles, Models and Applications, Wiley-Interscience, Second Edition, 72-244, 2003

Principe, J.C. Information Theoretical Learning Renyi's Entropy and Kernel Perspectives, Springer, 47-102, 2010

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

Renyi, A., Foundations of Probability, Dover Publications, NY, 23, 211, 2007

Teixeria, A., Matos,A., Souto, A., Antunes, L., Entropy Measures vs. Kolmogorov Complexity, Entropy, 2011, 13, 595-611

Topsøe, Basic Concepts, Identities and Inequalities –the Toolkit of Information Theory, Entropy, 2001, 3, 162-90

Wilcox, A.R, "Indices of Qualitative Variation", 1967, Oak Ridge National Laboratory; ORNL-TM-1919

Internet Resources

http://en.wikipedia.org/wiki/R%C3%A9nyi_entropy

http://www.saglik.gov.tr/TR/dosya/1-72577/h/saglikistatistikleriyilligi2010.pdf