TEXAS A&M UNIVERSITY
COLLEGE STATION, TEXAS 77843-3143

Department of STATISTICS
Statistical Interdisciplinary
Research Laboratory
E415EP@TAMVM1.BITNET

Emanuel Parzen
Distinguished Professor
Phone 409-845-3188
Fax 409-845-3144

GOODNESS OF FIT TESTS AND ENTROPY

Emanuel Parzen

Department of Statistics

Texas A&M University

Technical Report No. #103

May 1990

Texas A&M Research Foundation

Project No. 6547

'Functional Statistical Data Analysis and Modeling'

Sponsored by the U. S. Army Research Office

Professor Emanuel Parzen, Principal Investigator

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER: ARO 27574.2-MA	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle): GOODNESS OF FIT TESTS AND ENTROPY		5. TYPE OF REPORT & PERIOD COVERED: Technical
		6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s): Emanuel Parzen		8. CONTRACT OR GRANT NUMBER(s): DAAL03-90-G-0069
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Texas A&M University, Institute of Statistics, College Station, TX 77843		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS: U. S. Army Research Office, Post Office Box 12211, Research Triangle Park, NC 27709		12. REPORT DATE: May 1990
		13. NUMBER OF PAGES: 13
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report): Unclassified
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

NA

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)
Goodness of Fit; Entropy; Moran's Statistic; Information divergence; Gap Estimators, Autoregressive Estimation; Shapiro Wilk Statistic

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)
This paper discusses the unifying role of entorpy statistics and concepts in developing goodness of fit tests for a parametric model F(x;) for a continuous distribution function F(x), given a random sample from the distribution F. Statistics discussed are those by Moran (extended by Cheng and Stephens), Vasicek and Dudewicz & van der Meulen (based on gap estimators of quantile density function), Parzen (autoregressive estimators of quantile density functions), and shapiro and Wilk. They are given unified formulations as entropy difference statistics. Their 95% significance levels for sample sizes 20 and 50 are comapred and shown to increase as amount of "smoothing" decreases.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

S/N 0102-LF-014-6601

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

GOODNESS OF FIT TESTS AND ENTROPY

by Emanuel Parzen

Department of Statistics, Texas A&M University[1]

Dedicated to the memory of Paruchuri R. Krishnaiah

Abstract: This paper discusses the unifying role of entropy statistics and concepts in developing goodness of fit tests for a parametric model $F(x;\theta)$ for a continuous distribution function $F(x)$, given a random sample from the distribution F. Statistics discussed are those introduced by Moran (extended by Cheng and Stephens), Vasicek and Dudewicz & van der Meulen (based on gap estimators of quantile density function), Parzen (autoregressive estimators of quantile density functions), and Shapiro and Wilk. They are given unified formulations as entropy difference statistics. Their 95% significance levels for sample sizes 20 and 50 are compared and shown to increase as amount of "smoothing" decreases.

1. Introduction to Entropy. This paper discusses the unifying role of entropy concepts in testing goodness of fit of a random sample of a continuous random variable X. The problem is to test the fit of a parametric model $F(x,\theta)$, θ a vector of parameters, to the true distribution function $F(x) = \text{Prob}[X \leq x]$ of X with probability density function $f(x) = F'(x)$.

The true quantile function of X is $Q(u) = F^{-1}(u)$. Quantile density function is $q(u) = Q'(u) = 1/fQ(u)$; the density quantile function is $fQ(u) = f(Q(u))$. The entropy H of a random variable X is

$$H(f) = \int_{-\infty}^{\infty} \{-\log f(x)\} f(x) dx$$
$$= \int_0^1 \log q(u) = H(q).$$

In general, $H(q)$ can be any real number. But if $q(u)$ integrates to 1, corresponding to

[1]Research supported by the U.S.Army Research Office

a random variable on the unit interval, then neg-entropy $-H(q)$ is non-negative. The entropy statistics for goodness of fit are constructed to be non-negative.

The quantile density function $q(u)$ is in general a non-integrable function with a large dynamic range. We always assume that $\log q(u)$ is integrable, which means that X has finite entropy.

2. Moran's statistic. Assume that the random sample consists of distinct observations with order statistics denoted $X(1;n) < \ldots < X(n;n)$. The probability integral transform $Y = F(X;\theta)$, for a specified value of θ, has order statistics denoted $Y(1;n) < \ldots < Y(n;n)$. Let $Y(0;n) = 0, Y(n+1;n) = 1$. Let $D_i(\theta) = Y(i;n) - Y(i-1;n)$, $i = 1, \ldots, n+1$. Cheng and Stephens (1989) define Moran's statistic to be

$$M(\theta) = \sum_{i=1}^{n+1} \{-\log D_i(\theta)\}$$

They study the asymptotic distribution of $M(\theta)$ when θ is the true parameter value, and when θ is replaced by an efficient estimator $\hat{\theta}$. They illustrate the usefulness of Moran's statistic by an example of real data where $M(\theta)$ correctly rejects the hypothesis that X is normal, in contrast to more traditional empirical distribution function statistics such as the Kolmogorov-Smirnov and Cramer-von Mises statistics which accept the hypothesis of normality for the sample tested. Our aim in this paper is to provide a variety of alternatives to Moran's statistic by expressing it as an entropy statistic and to discuss how to generate entropy statistics.

Our first step is to normalize Moran's statistic by giving it a new definition; define

$$\begin{aligned} \tilde{M}(\theta) &= (1/(n+1)) \sum_{i=1}^{n+1} \{-\log d_i(\theta)\} \\ &= \int_0^1 \{-\log \tilde{d}(u;\theta)\} du \end{aligned}$$

defining for $i = 1, \ldots, n+1$

$$d_i(\theta) = (n+1)\{Y(i;n) - Y(i-1;n)\} = (n+1)D_i(\theta),$$

$$\tilde{d}(u;\theta) = d_i(\theta), (i-1)/(n+1) < u < i/(n+1).$$

The quantile function of $Y = F(X;\theta)$ is $D(u;\theta) = F(Q(u);\theta)$; it can be estimated by

$$\tilde{D}(u;\theta) = \int_0^u \tilde{d}(t;\theta)dt,$$

as well as $F(\tilde{Q}(u);\theta)$, where $\tilde{Q}(u)$ is the sample quantile function of the X sample. An estimator of the quantile density function

$$d(u;\theta) = D'(u;\theta) = f(Q(u);\theta)/fQ(u)$$

is $\tilde{d}(u;\theta)$. We call $d(u;\theta)$ a *comparison density* function, denoted $d(u;F(x),F(x;\theta))$.

Moran's statistic $\tilde{M}(\theta)$ is an estimator of $M(\theta) = -H(d(u;\theta))$, the neg-entropy $-H(Y)$ of $Y = F(X,\theta)$. When θ is the true parameter value Y is uniform and $H(Y) = 0$; Cheng and Stephens (1989) show that $\tilde{M}(\theta)$ is asymptotically normal with mean $\gamma = .57722$, Euler's constant. Therefore one may want to consider an unbiased entropy statistic $M^*(\theta) = \tilde{M}(\theta) - .57722$, which when θ is the true parameter value is asymptotically normal with mean zero and variance

$$\text{VAR}[M^*(\theta)] = (1/(n+1))(\frac{\pi^2}{6} - 1)$$

Cheng and Stephens (1989) use small sample corrections of this asymptotic distribution theory to compute significance levels of $M^*(\theta)$; for example, for $n = 20$, Prob$[M^*(\theta) \leq .48] = .95$. We note that $\tilde{M}(\theta)$ uses a least smooth estimator of $d(u;\theta)$, and one should consider other entropy statistics of goodness of fit generated by

$$\hat{M}(\theta) = \int_0^1 \{-\log \hat{d}(u;\theta)\}du$$

where $\hat{d}(u;\theta)$ is a smooth estimator of $d(u;\theta)$ of the form discussed in the sequel.

3. Kullback information divergence. The non-negative quantity $M(\theta)$ being estimated by $\tilde{M}(\theta)$ is the neg-entropy of $Y = F(X;\theta)$. It also can be identified to equal

$$I(f;f(.;\theta)) = \int_{-\infty}^{\infty} \{-\log(f(x;\theta)/f(x))\}f(x)dx$$

the Kullback information divergence btween the true distribution function $F(x)$ and the parametric model $F(x;\theta)$, since

$$M(\theta) = \int_0^1 \{-\log(f(Q(u);\theta)/fQ(u))\}du.$$

Define the sample distribution function $F^{\sim}(x)$ =fraction of random sample $\leq x$, with symbolic probability density $f^{\sim}$. Moran's statistic $M^{\sim}(\theta^{\wedge})$, where $\theta^{\wedge}$ is an efficient parameter estimator, can be regarded as an estimator of $I^{\sim\wedge} = I(f^{\sim}; f(.;\theta^{\wedge}))$, the information divergence between the data and the optimal parametric model. Other entropy statistics are obtained by alternative estimators of $I^{\sim\wedge}$, the sample to model information divergence.

Information divergence $I(f; f(.;\theta))$ can be expressed

$$I(f; f(\cdot;\theta)) = H(f; f(\cdot;\theta)) - H(f)$$

defining cross-entropy

$$H(f; f(.,\theta)) = \int_{-\infty}^{\infty} \{-\log f(x;\theta)\} f(x)dx = \int_{-\infty}^{\infty} \{-\log f(x;\theta)\}\, dF(x).$$

Cross-entropy is related to maximum likelihood estimation. Define the sample cross-entropy between the parametric model $F(x;\theta)$ and the sample distribution function $F^{\sim}(x)$ by

$$H(f^{\sim}; f((.;\theta)) = -E^{\sim}[\log f(X;\theta)] = -(1/n)\sum_{t=1}^{n} \log f(X(t);\theta).$$

The maximum likelihood estimator $\theta^{\wedge}$ is the minimum sample cross-entropy estimator. Define for any statistic $T(x)$

$$E^{\sim}[T(x)] = (1/n)\sum_{t=1}^{n} T(X(t)), E_{\theta}[T(x)] = \int_{-\infty}^{\infty} T(x) f(x;\theta)dx.$$

A model $f(x;\theta)$ is said to obey an exponential model if

$$\log f(x;\theta) = \sum_{j=1}^{k} \theta_j T_j(x) - \Psi(\theta).$$

The maximum likelihood estimator of an exponential model can be shown to be method of moments estimator; $\theta\hat{}$ is the value of θ satisfying

$$E\tilde{}[T_j(x)] = E_\theta[T_j(x)].$$

Further the minimum sample cross-entropy equals the entropy of $f(.;\theta\hat{})$:

$$H(f\tilde{};f(.;\theta\hat{}) = H(f(.;\theta\hat{})) = \Psi(\theta\hat{}) - \sum_{j=1}^{k} \theta_j\hat{} E_{\theta\hat{}}[T_j(x)].$$

4. Entropy difference goodness of fit statistics for exponential models. An important conclusion can now be formulated. When the parametric model is an exponential model, the natural entropy statistic to test goodness of fit given by the sample to model information divergence $I\tilde{}\hat{}$ can be expressed as an entropy difference statistic

$$I\tilde{}\hat{} = H(f(\cdot;\theta\hat{})) - H(f\tilde{})$$

and can be estimated by $H(f(\cdot;\theta\hat{})) - H\hat{}(f)$ where $H\hat{}(f)$ is an estimator of $H(f)$. Since $H(f(.;\theta\hat{}))$ can be interpreted as the "maximum entropy" we obtain a "non-negative statistic" by the entropy difference statistics to test goodness of fit of a parametric model. Note $H(f(.;\theta\hat{}))$ is an estimator evaluated under the assumption that f obeys the null hypothesis of belonging to the parametric family $f(x;\theta)$, and $H\hat{}(f)$ is a non-parametric evaluation based on a smooth non-parametric estimator of the true density f.

5. Gap estimators. A basic approach to estimators $H\hat{}(f)$ is to use the entropy formula

$$H(f) = \int_0^1 \log q(u)du$$

in terms of the quantile density function $q(u)$. Many approaches are available to form estimators $q\hat{}(u)$ and thus estimators $H\hat{}(f)$ of the entropy of the true probability density f. The earliest approach considered by researchers is equivalent to

$$H\hat{}(f) = H(q_\nu\hat{}) = (1/(n-2\nu)) \sum_{j=\nu}^{n-\nu} \log q_\nu\hat{}(j/(n+1))$$

where for $j = \nu + 1, \ldots, n - \nu$

$$q_\nu\hat{}(j/(n+1) = ((n+1)/2\nu)\{X(j+\nu;n) - X(j-\nu;n)\}$$

is an estimator of the quantile density $q(u)$ of X at $u = j/(n+1)$. We call these estimators gap (of order ν) estimators; they were introduced and studied for $\nu = 1, 2, 3, 4, 5$ by Vasicek (1977) to test normality and Dudewicz and van der Muelen (1981) to test uniformity.

Normality is an example of a location-scale parametric model

$$Q(u) = \mu + \sigma Q_0(u)$$

where μ and σ are parameters to be estimated and $Q_0(u)$ is a known standard distribution (for normality, $Q_0(u) = \Phi^{-1}(u)$, the inverse of the standard normal distribution function). For a location-scale parametric model

$$H(f(.;\theta\hat{})) = \log \sigma\hat{} + H(f_0).$$

For a normal distribution, $H(f_0) = .5\{1 + \log 2\pi\}$ and $\sigma\hat{}$ is the sample standard deviation. Vasicek (1977) entropy statistic for testing normality can be expressed

$$\Delta_V = \log \sigma\hat{} + H(f_o) - H(q_\nu\hat{}).$$

6. Autoregressive estimators. An alternative approach to estimating $q(u)$ when one desires a goodness of fit test of a location scale parametric model is to estimate the density $d(u)$, $0 < u < 1$, defined by

$$d(u) = (1/\sigma_0) f_0 Q_0(u) q(u),$$

where $\sigma_0 = \int_0^1 f_0 Q_0(u) q(u) du$. We call $d(u)$ a *didi* (divided difference) density, or weighted spacings density, denoted $dd(u; F(x), F_o(x))$. They provide an alternative to $Q - Q_o$ plots.

Notice that the neg-entropy of d satisfies

$$\begin{aligned} -H(d) &= \int_0^1 \{-\log d(u)\} du \\ &= \log \sigma_0 + H(f_0) - H(q). \end{aligned}$$

Therefore $-H(d)$ is an entropy difference, and an estimator $-H(\hat{d})$ provides in one stroke an entropy difference statistic for goodness of fit!

We assume that $d(u)$, $1/d(u)$, $\log d(u)$ are integrable functions. Estimating $d(u)$ rather than $q(u)$ can be regarded as a process of preflattening the function to be estimated. We currently prefer estimation of $d(u)$ by kernel estimators, using boundary kernels to compensate for end effects at 0 and 1, or by maximum entropy estimation using exponential models for $d(u)$.

In Parzen (1979) we introduced the autoregressive method of estimating $d(u)$ which has other close connections to entropy statistics for goodness of fit. Raw estimators $\tilde{d}(u)$ and $\sigma_0\tilde{}$ are formed by replacing $q(u)$ by a least smooth gap estimator $q_2\tilde{}(u)$. Smooth estimators $d_m\hat{}(u)$ are formed by the autoregressive method.

From estimators $\tilde{\rho}(v)$ of the pseudo-correlations

$$\rho(v) = \int_0^1 e^{2\pi i u v} d(u) du, \quad v = 0, \pm 1, \ldots, \pm m$$

one estimates (using suitable Yule-Walker equations) the coefficients of the autoregressive order m approximator

$$\hat{d}_m(u) = K_m\hat{}|1 + \alpha_m\hat{}(1)e^{2\pi i u} + \ldots + \alpha_m\hat{}(m)e^{2\pi i u m}|^{-2}$$

to the raw density $\tilde{d}(u)$. The coefficient $K_m\hat{}$ plays an important role in entropy calculations since

$$\int_0^1 -\log d_m\hat{}(u) du = -\log K_m\hat{}$$

can be regarded as an estimator of $\int_0^1 -\log \hat{d}(u) du$, and thus is an entropy difference statistic for goodness of fit.

7. Entropy difference interpretation of Shapiro Wilk statistic. To test the hypothesis $H_0 : X$ is $N(\mu, \sigma^2)$, a test statistic W of Shapiro-Wilk type is of the form

$$W = \sigma^*/\hat{\sigma}$$

where $\sigma\hat{}$ is the sample standard deviation and

$$\sigma^* = \sum_{j=1}^{n} \Phi^{-1}\left(\frac{j-0.5}{n}\right) X(j;n) \div \left\{\sum_{j=1}^{n} |\Phi^{-1}\left(\frac{j-0.5}{n}\right)|^2\right\}^{1/2}$$

is an asymptotically efficient estimator of σ based on linear combinations of the order statistics $X(j;n)$ of the random sample. The first step in the entropy interpretation of W is to consider instead the statistic

$$-\log W = \log \sigma\hat{} - \log \sigma^* = H(f(\cdot;\sigma\hat{})) - H(f(\cdot;\sigma^*)).$$

$-\log W$ is an entropy difference statistic, but it compares two parametric estimators of entropy based on two approaches to estimating parameters which are both efficient under the null hypothesis of normality.

8. Comparison of 95% significance levels for small samples. Significance levels for the entropy-difference statisic $\Delta_W = -\log W$ are obtainable from tables of the W statistic [for example, Filliben (1975)]. An example of 95% significance levels (for accepting normality) are

$$\Delta_W \leq 0.05, \quad \text{for sample size } n = 20;$$

$$\Delta_W \leq 0.023, \quad \text{for sample size } n = 50.$$

The various entropy difference statistics can be compared by their significance levels (see Table). Significance levels of autoregressive goodness of fit statistics $-\log K_m\hat{}$ have been derived by a very approximate Monte Carlo simulation (in the case of testing for normality). An open research problem is investigation of an Akaike-type criterion for accepting the null hypothesis that X is $F_0((x-\mu)/\sigma)$, such as:

$$(2m/n) + \log K_m\hat{} \geq 0 \text{ for } m = 1, 2, \ldots.$$

Significance levels of Vasicek (1977) statistic Δ_V, defined in section 5, are based on Monte Carlo simulation of normal; significance levels of similar Dudewicz-van der Muelen (1981) $\Delta_V = -H(q_V\hat{})$ to test uniformity statistic are based on Monte Carlo simulation of uniform.

One can conjecture a *relation between gap order* 2ν and *autoregressive order* m for the corresponding estimators to have similar distributions and therefore similar significance levels:

$$(2\nu)m = n = \text{ sample size}$$

To understand what this conjecture is alleging note that for $n = 20$, $m = 4$ is similar to $2\nu = 6$; for $n = 50$, $m = 6$ is similar to $2\nu = 8$. When one uses gap estimators of $q(u)$, and thus of entropy, one has the problem of determining the order 2ν. One may be able to more easily develop criteria for determining the order m of autoregressive estimators of $q(u)$.

Our modified Moran statistic $M^*(\theta)$ can be compared by noting that it has 95% significance level .48 for $n = 20$. Significance level appears to increase as amount of smoothing of "hidden" density $d(u)$ decreases. Investigating this phenomenon is a good topic for future research.

Table. 95% SIGNIFICANCE LEVELS FOR ENTROPY DIFFERENCE STATISTICS. Accept $H_0 : X$ is $N(\mu, \sigma^2)$ for some μ and σ if entropy difference is less than threshold given.

Sample Size n	$-\log W$ Shapiro-Wilk	$-\log K_m\hat{}$ Autoregressive order m Monte Carlo 5% level (rough approximation 2m/n)					Δ_V H(gap estimator $q_\nu\hat{}(u)$) Vasicek test for normality (Dudewicz-van der Muelen) test $U[0,1]$				
		$m=1$	$m=2$	$m=3$	$m=4$	$m=5$	$\nu=5$	$\nu=4$	$\nu=3$	$\nu=2$	$\nu=1$
20	.05	.141	.235	.299	.378	.398		.40	.40	.43	.61
		(.10)	(.20)	(.30)	(.40)	(.50)		(.43	.43	.47	.66)
50	.023	.045	.081	.126	.153	.176	.21	.21	.23		
		(.04)	(.08)	(.12)	(.26)	(.20)	(.22	.22	.24)		

REFERENCES

Cheng, R. C. H. and Stephens, M. A. (1989). A goodness of fit test using Moran's statistic with estimated parameters. *Biometrika*, 76, 385–392.

Dudwicz, E. J. and Van der Muelen, E. C. (1981). Entropy-based tests of uniformity, *Journal of the American Statistical Association*, *76*, 967–974.

Filliben, J. J. (1975). The probability plot correlation coefficient test for normality, *Technometrics*, *17*, 111–117.

Parzen, E. (1979). Nonparametric statistical data modeling. *Journal of the American Statistical Association*, *74*, 105–131.

Parzen, E. (1982). Maximum entropy interpretation of autoregressive spectral densities. *Statistics and Probability Letters*, 1, 2–6.

Shapiro, S. S. and Francis, R. S. (1972). Approximate analysis of variance test for normality. *J. American Statistical Association*, *67*, 215-216.

Shapiro, S. S. and Wilk, M. B. (1968). An analysis of variance test for normality, *Biometrika*, *52*, 591–611.

Shapiro, S. S., Wilk, M. B. and Chen, H. J. (1968). A comparative study of various tests for nomality, *J. American Statistical Association*, *63*, 1343–1372.