The Annals of Probability
1986, Vol. 14, No. 1, 336–342

ENTROPY AND THE CENTRAL LIMIT THEOREM[1]

BY ANDREW R. BARRON

Stanford University

A strengthened central limit theorem for densities is established showing monotone convergence in the sense of relative entropy.

1. Introduction. The probability density function $f_n(x)$ for the standardized sum of i.i.d. random variables with finite variance is shown to converge to the normal density function $\phi(x)$ in the sense of relative entropy: $\int f_n \log f_n/\phi \to 0$ provided the relative entropy is finite for some n. Furthermore, the relative entropy decreases along the powers of two subsequence $n_k = 2^k$. The classic results of L^1 convergence of the densities (Prohorov, 1952) and convergence in distribution follow as corollaries via the inequality $(\int |f_n - \phi|)^2 \leq 2 \int f_n \log f_n/\phi$. The proof of our result does not involve the usual Fourier transform technique, but follows instead from fundamental properties of Shannon entropy and Fisher information.

Motivations for showing convergence of the relative entropy $D_n = \int f_n \log f_n/\phi$ are well known. For tests of hypotheses (such as ϕ verses f_n), the relative entropy is the exponent in the probability of error [Stein's lemma, see Chernoff (1956)]. From information theory, the relative entropy D_n is the least upper bound to the redundancy (excess average description length) of the Shannon code based on the normal distribution when describing quantizations of samples from f_n. Our characterization of the central limit theorem resembles the second law of thermodynamics. Indeed, the decrease of the relative entropy D_n to zero is equivalent to the increase of the entropy $H_n = -\int f_n \log f_n$ to the entropy of the normal.

Linnik (1959) used the information measures of Shannon and Fisher in a proof of convergence in distribution. Rényi (1970, page 601) states that Linnik established convergence of $\int f_n \log f_n/\phi$ to zero. A reading of Linnik reveals that convergence was established only for densities of truncated random variables smoothed by the addition of independent normal random variables. We show that $D_n \to 0$, provided it is finite for some n. No smoothness conditions are required of the density f_n for this convergence to hold.

Recently Brown (1982) gave an elegant proof of convergence in distribution based on the decrease of Fisher informations. We extend Brown's argument to show that the Fisher informations converge to the reciprocal of the variance (as suggested by the Cramér–Rao bound). The link between Fisher information and relative entropy is the unexpected identity given below.

Received March 1984; revised April 1985.

[1]This work was partially supported by NSF Contract ECS 82-11568.

AMS 1980 *subject classifications.* Primary 60F05; secondary 94A17, 62B10.

Key words and phrases. Central limit theorem, local limit theorem, entropy, Fisher information, convolution inequalities.

Institute of Mathematical Statistics is collaborating with JSTOR to digitize, preserve, and extend access to *The Annals of Probability.* www.jstor.org

2. Entropy and information. Let X be any random variable with finite variance. The relative entropy D is defined as follows. If X has a density function $f(x)$, then $D(X) = \int f(x)\log f(x)/\phi(x)\,dx$ where ϕ is the normal density with the same mean and variance as f; otherwise $D(X) = \infty$. The Shannon entropy $H(X) = -\int f \log f$ satisfies $H = (1/2)\log 2\pi e\sigma^2 - D$ where σ^2 is the variance. By concavity of the logarithm, D is nonnegative and equals zero only if $f = \phi$ a.e. Consequently, the normal has maximum entropy for a given variance.

Let Y be a random variable with continuously differentiable density $g(y)$ and finite variance σ^2. Define the standardized Fisher information

$$J(Y) = \sigma^2 E\big(\rho(Y) - \rho_\phi(Y)\big)^2$$

where $\rho = g'/g$ is the score function for Y and $\rho_\phi = \phi'/\phi$ is the (linear) score function for the normal with the same mean and variance as Y. The Fisher information $I(Y) = E\rho^2(Y)$ satisfies $I = (J + 1)/\sigma^2$. Since $J \geq 0$ with equality only if $g = \phi$, the normal has minimum Fisher information for a given variance (whence the Cramér–Rao inequality $I \geq 1/\sigma^2$). The standardized informations D and J are translation and scale invariant.

LEMMA 1. *Entropy is an integral of Fisher informations. Let X be any random variable with finite variance, then*

$$D(X) = \int_0^1 J(\sqrt{t}\,X + \sqrt{1-t}\,Z)\frac{dt}{2t}, \tag{2.1}$$

where Z is an independent normal random variable with the same mean and variance as X.

Equation (2.1) is derived in Section 4 from the corresponding differential equation $(d/dt)D(Y_t) = J(Y_t)/2t$ where $Y_t = \sqrt{t}\,X + \sqrt{1-t}\,Z$ and $0 \leq t < 1$.

Two basic convolution inequalities are needed in the proofs. Let X and Y be random variables as above and suppose that the density for Y has a bounded derivative. Let V be any random variable independent of X and Y. Then convolution increases entropy and decreases Fisher information:

$$H(X + V) \geq H(X) \quad \text{and} \quad I(Y + V) \leq I(Y). \tag{2.2}$$

Equality holds if and only if V is almost surely constant. These inequalities follow from the convexity of the functions $x \log x$ and x^2 and are well known.

More specialized convolution inequalities will also be needed. Let Y_1 and Y_2 be independent random variables having densities with bounded derivatives and let $\alpha_i \geq 0$, $\alpha_1 + \alpha_2 = 1$, then $I(\sqrt{\alpha_1}\,Y_1 + \sqrt{\alpha_2}\,Y_2) \leq \alpha_1 I(Y_1) + \alpha_2 I(Y_2)$ [see Stam (1959) and Blachman (1965)]. Hence if Y_1 and Y_2 have the same variance,

$$J\big(\sqrt{\alpha_1}\,Y_1 + \sqrt{\alpha_2}\,Y_2\big) \leq \alpha_1 J(Y_1) + \alpha_2 J(Y_2). \tag{2.3}$$

Equality holds if and only if Y_1 and Y_2 are normal. An immediate consequence of Lemma 1 is the inequality $D(\sqrt{\alpha_1}\,X_1 + \sqrt{\alpha_2}\,X_2) \leq \alpha_1 D(X_1) + \alpha_2 D(X_2)$ for any independent random variables X_1, X_2 having the same finite variance. This

inequality for D may also be deduced from Shannon's entropy power inequality (Shannon, 1948, Stam, 1959, and Blachman, 1965).

3. Strengthened central limit theorem. Let $X_1, X_2, \ldots$ be independent and identically distributed (i.i.d.) random variables with mean zero and variance σ^2 and let $S_n = \sum_{i=1}^n X_i/\sqrt{n}$ be the standardized sum. Let Z be an independent normal random variable with mean zero and variance σ^2.

The following lemma states convergence of the Fisher informations.

LEMMA 2. *Let S_n be the standardized sum and let $S_n' = \sqrt{t}\,S_n + \sqrt{1-t}\,Z$ for fixed $0 \le t < 1$. Then $nJ(S_n')$ is a subadditive sequence, that is $(p+q)J(S_{p+q}') \le pJ(S_p') + qJ(S_q')$. In particular, $J(S_{2n}') \le J(S_n')$. Furthermore, the standardized Fisher information converges to zero*

$$\lim_{n\to\infty} J(S_n') = 0. \tag{3.1}$$

Equivalently, the sequence of Fisher informations $I(S_n')$ converges to the Cramér–Rao bound $1/\sigma^2$.

The subadditivity follows from the convolution inequality (2.3). Brown (1982) used $I(S_n') \ge I(S_{2n}')$ to conclude $I(S_n') - I(S_{2n}') \to 0$ and from this obtained convergence of S_n in distribution. Inequality (2.2) implies that the sequence $I(S_n')$ is bounded. The subadditivity and boundedness implies that the limit exists and equals the infimum (Gallager, 1968, page 112). Brown conjectured but did not obtain $\lim I(S_n') = 1/\sigma^2$. In Section 4 we use Brown's argument plus uniform integrability to complete the proof.

The main result of this note follows. The density function f_n for the standardized sum S_n converges to the normal density ϕ in the sense of relative entropy.

THEOREM. *Let S_n be the standardized sum. Then $nD(S_n)$ is a subadditive sequence. In particular $D(S_{2n}) \le D(S_n)$. Furthermore, the relative entropy converges to zero*

$$\lim_{n\to\infty} D(S_n) = 0 \tag{3.2}$$

if and only if $D(S_n)$ is finite for some n. Equivalently, the entropy $H(S_n)$ converges to the normal entropy $(1/2)\log 2\pi e\sigma^2$, provided the entropy is finite for some n.

PROOF. The limit exists: Using Lemma 1 the inequalities for D follow directly from the inequalities for J. Thus $nD(S_n)$ is subadditive. In particular $D(S_{mp}) \le D(S_p)$. Let p be such that $D(S_p) \le \inf_n D(S_n) + \varepsilon$. Write $n = mp + r$ where the remainder r is less than p. Using inequality (2.2) we find $D(S_n) \le D(S_{mp}) - (1/2)\log(1 - r/n)$ which is less than $D(S_p) - (1/2)\log(1 - p/n)$. Letting $n \to \infty$ then $\varepsilon \to 0$ yields $\lim D(S_n) = \inf D(S_n)$.

The limit is zero: From Lemma 1 we have $D(S_n) = \int_0^1 J(S_n')\,dt/2t$. Consider the powers of two subsequence $n_k = 2^k$. From Lemma 2, $J(S_{n_k}') \downarrow 0$ and hence

$D(S_{n_k}) \downarrow 0$ by the monotone convergence theorem, provided $D(S_n)$ is finite for some n. The entire sequence has the same limit as the subsequence, hence $\lim D(S_n) = 0$. □

COROLLARY. *Suppose the entropy $H(S_n)$ is finite for some n. Then S_n has a density function f_n which converges to ϕ in the L^1 sense,*

$$\lim \int |f_n(x) - \phi(x)|\, dx = 0. \tag{3.3}$$

Furthermore, the quadratic $\log \phi(S_n)$ is a consistent approximation of the log-likelihood $\log f_n(S_n)$. Indeed the difference converges in L^1 (and hence in probability),

$$\lim E|\log f_n(S_n) - \log \phi(S_n)| = 0. \tag{3.4}$$

PROOF. Immediate from the inequalities $(\int |p - q|)^2 \leq 2D$ and $\int p|\log p/q| \leq D + (2D)^{1/2}$ for any probability densities p and q where $D = \int p \log p/q$. The first inequality is due to Csiszár (1967) and Kullback (1967). The second inequality follows from $\int p(\log p/q)^- = \int_A p \log q/p \leq P(A)\log Q(A)/P(A) \leq Q(A) - P(A) = (1/2)\int |p - q|$, where $A = \{x\colon q(x) > p(x)\}$ and P and Q are the distributions corresponding to the densities p and q. Similar bounds, but with a constant larger than 2, were given by Pinsker (1964). □

REMARKS. These conclusions are stronger than the classic result of convergence in distribution which states that $\lim |F_n(A) - \Phi(A)| = 0$ for fixed sets A with boundary measure zero. Indeed, the L^1 convergence of the densities is equivalent to the uniform setwise convergence of the distributions, $\lim_n \sup_A |F_n(A) - \Phi(A)| = 0$ where the supremum is over all Borel sets A. [Hence $\lim |F_n(A_n) - \Phi(A_n)| = 0$ for arbitrarily varying sets A_n.] Another characterization of convergence in distribution is that

$$\lim Eh(S_n) = Eh(Z) \tag{3.5}$$

for all bounded uniformly continuous functions h. A consequence of convergence in relative entropy is that (3.5) holds for any measurable function h for which $Ee^{\alpha h(Z)}$ is finite for all α in some neighborhood of zero [see Csiszár (1975)]. In particular, (3.5) holds for functions $h(x)$ bounded by some multiple of $x^2 + 1$.

4. Verification of the details

PROOF OF LEMMA 1. We show that $D(X) = \int_0^1 J(Y_t)\, dt/2t$ where $Y_t = \sqrt{t}\, X + \sqrt{1 - t}\, Z$. The differential equation $(d/dt)D(Y_t) = J(Y_t)/2t$ for $0 < t < 1$ follows by change of variables from de Bruijn's identity $(d/d\tau)H(X + Z_\tau) = I(X + Z_\tau)/2$ (Stam, 1959 and Blackman, 1965) where Z_τ is an independent normal random variable with mean zero and variance τ. For $\tau > 0$, de Bruijn's identity holds without conditions on the distribution of X other than finite variance. [Stam (1959) asserts that de Bruijn's identity holds for random variables having a strictly positive differentiable density with finite Fisher information. We show

this is enough to conclude validity for any random variable X. Simply fix a such that $0 < a < \tau$ and define $X_a = X + Z_a$. Then X_a satisfies Stam's conditions, so $(d/d\tau)H(X_a + Z_{\tau-a}) = I(X_a + Z_{\tau-a})/2$ which reduces to de Bruijn's identity for arbitrary X. For a direct proof involving several exchanges of differentiation and expectation, each justified by the mean value and dominated convergence theorems, see Barron (1984).]

Inequality (2.2) implies the integrand $J(Y_t)/t$ is bounded by $1/1 - t$. Hence we may integrate the derivative on $[a, b]$ (where $0 < a < b < 1$) to obtain

$$D(Y_b) - D(Y_a) = \int_a^b J(Y_t)\, dt/2t. \tag{4.1}$$

Now $D(Y_a) \le \log 1/\sqrt{1-a}$ from inequality (2.2), thus $\lim_{a\to 0} D(Y_a) = 0$. Also $D(Y_b) \le D(X) + \log 1/\sqrt{b}$, so $\limsup_{b\to 1} D(Y_b) \le D(X)$. Note that $\lim_{b\to 1} Y_b = X$ in probability and hence in distribution, so $\liminf_{b\to 1} D(Y_b) \ge D(X)$ by the lower semicontinuity of the relative entropy. Thus $\lim_{b\to 1} D(Y_b) = D(X)$. If the integral on $(0, 1)$ is finite, then letting $a \to 0$ and $b \to 1$ in (4.1) we obtain the desired result $D(X) = \int_0^1 J(Y_t)\, dt/2t$. If the integral is infinite, then by Fatou's lemma D is also infinite. This completes the proof of Lemma 1. □

To prove Lemma 2, we extend the arguments of Brown (1982). First we state his contribution.

BROWN'S RESULT. Fix $\tau > 0$ and set $Y_k = S_{2^k} + Z_\tau$. Let $\rho_k = g_k'/g_k$ be the score function for Y_k. Note that $Y_{k+1} = (Y_k + Y_k')/\sqrt{2}$ where Y_k' is an independent copy of Y_k. The score for Y_{k+1} is the conditional expectation: $\rho_{k+1}(Y_{k+1}) = E[(\rho_k(Y_k) + \rho_k(Y_k'))/\sqrt{2}\,|\,Y_{k+1}]$. Hence the score of the sum $\rho_{k+1}(Y_{k+1})$ provides the best estimate of the sum of the scores in the sense of minimum mean-squared error. The corresponding pythagorean relation yields $I(Y_k) - I(Y_{k+1}) = E[(\rho_k(Y_k) + \rho_k(Y_k'))/\sqrt{2} - \rho_{k+1}(Y_{k+1})]^2$. This difference sequence must converge to zero, since the sequence $I(Y_k)$ is decreasing and bounded. Therefore $E[(\rho_k(Z_{\tau/2}) + \rho_k(Z_{\tau/2}'))/\sqrt{2} - \rho_{k+1}((Z_{\tau/2} + Z_{\tau/2}')/\sqrt{2})]^2$ also converges to zero, since the density for $Y_k = S_{2^k} + Z_\tau$ is bounded below by a multiple of the normal $(0, \tau/2)$ density. But for an arbitrary function v with finite $Ev^2(Z)$, the mean-squared error of estimation of $v(Z) + v(Z')$ using any function of the sum $Z + Z'$ is shown to exceed a multiple of the mean-squared error using the best linear estimate. Consequently, Brown established that

$$\lim_{k\to\infty} E\big(\rho_k(Z_{\tau/2}) - \rho_\phi(Z_{\tau/2})\big)^2 = 0 \tag{4.2}$$

and furthermore, since the score is the derivative of the logarithm of the density, the sequence of densities g_k for $Y_k = S_{2^k} + Z_\tau$ satisfies

$$g_k(y) \to \phi(y) \tag{4.3}$$

uniformly on compact subsets. Here ϕ is the normal $(0, \sigma^2 + \tau)$ density. Brown used (4.3) and let $\tau \to 0$ to prove convergence in distribution. We use (4.2) and (4.3) plus a uniform integrability argument to identify the limit of the Fisher informations.

PROOF OF LEMMA 2. We show that $\lim I(S_n') = 1/\sigma^2$. Let $g_k = g_{k,\tau}$ be the density for $Y_k = S_{2^k} + Z_\tau$. Note that the Fisher information satisfies $I(Y_k) = E\rho_k^2(Y_k) = \int \rho_k^2 g_k/\phi \, d\Phi$ where Φ is the normal $(0, \sigma^2 + \tau)$ distribution. From (4.2) the score ρ_k converges to ρ_ϕ in normal $(0, \tau/2)$ probability and hence in Φ probability. From (4.3) and Scheffe's lemma $\int |g_k - \phi| \to 0$ and hence $g_k/\phi \to 1$ in Φ probability. The product of sequences convergent in probability is also convergent in probability. Thus $\rho_k^2 g_k/\phi$ converges to ρ_ϕ in Φ probability. Consequently,

$$\lim_{k\to\infty} \int \rho_k^2 \frac{g_k}{\phi} \, d\Phi = \int \rho^2 \phi \, d\Phi \tag{4.4}$$

provided the integrand $\rho_k^2 g_k/\phi$ is uniformly Φ-integrable. By Lemma 3 the integrand is less than a multiple of $g_{k,2\tau}/\phi$. From inequality (2.2) $H(S_{2^k} + Z_{2\tau}) \geq H(Z_{2\tau})$, we find that the relative entropy $\int (g_{k,2\tau}/\phi)\log(g_{k,2\tau}/\phi) \, d\Phi$ is bounded. But bounded relative entropy implies uniform integrability of the density ratio by a standard argument (Billingsley, 1979, page 188). Hence Equation (3.4) is valid, which means $\lim I(Y_k) = 1/(\sigma^2 + \tau)$. By change of variables $\lim I(S_{2^k}') = 1/\sigma^2$ where $S_n' = \sqrt{t} S_n + \sqrt{1-t} Z$. By subadditivity, the entire sequence has the same limit as the subsequence, hence $\lim I(S_n') = 1/\sigma^2$. □

In the above proof we used the following simple result.

LEMMA 3. *Let g_τ be the density for $Y = X + Z_\tau$ where X is an arbitrary random variable and Z_τ is an independent normal $(0, \tau)$ variable. Then $(g_\tau'(y))^2 \leq c_\tau g_{2\tau}(y) g_\tau(y)$ where $c_\tau = 4\sqrt{2}\, e^{-1}/\tau$.*

PROOF. Let ϕ_τ denote the normal density function for Z_τ. The normal density has a bounded derivative $\phi_\tau'(z) = -z\phi_\tau(z)/\tau$ and hence the density $g_\tau(y) = E\phi_\tau(y - X)$ has a bounded derivative $g_\tau'(y) = E\phi_\tau'(y - X)$. (The implicit exchange of limit and expectation is valid by application of the mean value and bounded convergence theorems.) Thus $(g_\tau'(y))^2 = (E(y - X)\phi_\tau^{1/2}(y - X)\phi_\tau^{1/2}(y - X)/\tau)^2$ and by the Cauchy–Schwartz inequality this does not exceed $E(y - X)^2\phi_\tau(y - X) g_\tau(y)/\tau^2$ which is less than $c_\tau E(\phi_{2\tau}(y - X) g_\tau(y) = c_\tau g_{2\tau}(y) g_\tau(y)$. This is the desired result. □

5. Examples. For the relative entropy to be finite for some n it is sufficient but not necessary for the density to be bounded for some n. By modifying an example in Kolmogorov and Gnedenko (1954, page 223), we find unbounded densities for which the relative entropy either becomes finite or remains infinite.

Let $f_r(x)$, $r > 0$ be the family of probability density functions proportional to $|x|^{-1}/(\log|x|^{-1})^{1+r}$ on $|x| \leq e^{-1}$. This density has entropy $H(f_r) = -\infty$ for $r \leq 1$ and $H(f_r) = -(r^2(r-1))^{-1} - \log r/2$ for $r > 1$. The n-fold convolution $f_r^{(n)}$ exceeds a multiple of $f_{nr}(x)$ and is less than a multiple of $f_{nr}(x/n)$ for x in a neighborhood of zero. Consequently, the entropy $H(f_r^{(n)})$ is infinite $(-\infty)$ for $n \leq 1/r$ and finite for $n > 1/r$. Therefore, the standardized n-fold convolution converges to the normal density in the relative entropy sense, although the density remains unbounded for all n.

For an example where convergence in the relative entropy sense fails, let $f_r(x)$, $r > 0$ be proportional to $|x|^{-1}/\log|x|^{-1}(\log\log|x|^{-1})^{1+r}$ on $|x| \leq e^{-e}$. This density is more sharply peaked and has entropy $H(f_r) = -\infty$ for all $r > 0$. Also, the n-fold convolution $f_r^{(n)}(x)$ exceeds a multiple of $f_{nr}(x)$ for x in a neighborhood of zero. Therefore, the entropy $H(f_r^{(n)})$ is infinite for all n and the density does not converge in the relative entropy sense. A consequence is that the normal density provides an inefficient description of samples from $f_r^{(n)}$. Indeed, the Shannon redundancy is infinite.

Acknowledgment. Professors Tom Cover and Imre Csiszár are acknowledged for their helpful suggestions. Cover showed that Shannon's entropy power inequality implies the monotonicity of the entropy and he posed the problem of identifying the limit.

REFERENCES

BARRON, A. R. (1984). Monotonic central limit theorem for densities. Technical Report 50, Dept. Statistics, Stanford Univ.

BILLINGSLEY, P. (1979). *Probability and Measure*. Wiley, New York.

BLACHMAN, N. M. (1965). The convolution inequality for entropy powers. *IEEE Trans. Inform. Theory* **IT-11** 267–271.

BROWN, L. D. (1982). A proof of the central limit theorem motivated by the Cramér–Rao inequality. In *Statistics and Probability: Essays in Honor of C. R. Rao* (G. Kallianpur, P. R. Krishnaiah, and J. K. Ghosh, eds.) North-Holland, Amsterdam.

CHERNOFF, H. (1956). Large sample theory—parametric case. *Ann. Math. Statist.* **27** 1–22.

CSISZÁR, I. (1967). Information-type measures of difference of probability distributions and indirect observations. *Studia Sci. Math. Hungar.* **2** 299–318.

CSISZÁR, I. (1975). I-divergence geometry of probability distributions and minimization problems. *Ann. Probab.* **3** 146–158.

GALLAGER, R. G. (1968). *Information Theory and Reliable Communications*. Wiley, New York.

KOLMOGOROV, A. N. and GNEDENKO, B. V. (1954). *Limit Distributions for Sums of Independent Random Variables*. Translated by K. L. Chung, Addison-Wesley, Reading, Mass.

KULLBACK, S. (1967). A lower bound for discrimination in terms of variation. *IEEE Trans. Inform. Theory* **IT-13** 126–127.

LINNIK, YU. V. (1959). An information-theoretic proof of the central limit theorem with the Lindeberg condition. *Theory Probab. Appl.* **4** 288–299.

PINSKER, M. S. (1964). Information and Information Stability of Random Variables. *Translated by A. Feinstein, Holden-Day, San Francisco.*

PROHOROV, YU. V. (1952). On a local limit theorem for densities. *Dokl. Akad. Nauk. SSSR* **83** 797–800.

RÉNYI, A. (1970). *Probability Theory*. North-Holland, Amsterdam.

SHANNON, C. (1948). A mathematical theory of communication. *Bell System Tech. J.* **27** 379–423, 623–656.

STAM, A. J. (1959). Some inequalities satisfied by the quantities of information of Fisher and Shannon. *Inform. and Control.* **2** 101–112.

DEPARTMENT OF STATISTICS
THE UNIVERSITY OF ILLINOIS
URBANA, ILLINOIS 61801