International Journal of Theoretical Physics, Vol. 37, No. 12, 1998

Measures of Information and Error Laws

B. H. Lavenda[1]

Received February 13, 1998

The logarithm of joint error densities for the most common means are shown to be proportional to the difference of two weighted means which discriminate between a complete, nonuniform probability distribution and the uniform distribution. The difference in the weighted means is related to a new Shannon-type inequality for the discrimination between two probability distributions. Measures of the distance between the two distributions are determined, and a new statistic, comparable to χ^2, is derived from a first-order approximatiom of the directed divergence. Comparison is made between the error laws and the method of maximum likelihood.

1. INTRODUCTION AND SUMMARY

Statisticians have introduced many *ad hoc* procedures to discriminate between a 'null' hypothesis, which assigns a set of probabilities, $p = p_1, p_2, \ldots, p_N$, to a set of events, and the observed frequencies, $q = q_1, q_2, \ldots, q_N$. If the former set of numbers is sufficiently 'close' to the latter set, then we are satisfied that theory fits experiment. but how close is 'close'? At the beginning of this century, Pearson introduced a χ^2 test to deal with this problem analytically. The χ^2 test measures the mean-squared error committed when the probabilities p are taken to be the true values.

Another way to discriminate between two distributions is to construct the log-likelihood ratio for discriminating in favor p against q. This utilizes the Shannon inequality, based on the concavity of the Shannon entropy, which Kullback has referred to as the 'directed' divergence (Kullback, 1959). A more symmetrical form, called the symmetric divergence, puts the two distributions on the same footing by adding two Shannon inequalities, where the second inequality results from interchanging p and q. However, in all these attempts to discriminate or measure the 'distance' between the two distribu-

[1] Università di Camerino, I-63032 Camerino, (MC) Italy.

0020-7748/98/1200-311 9$15.00/0

tions no attempt has been made to relate the criteria with error laws, other than the normal one, that lead to a whole host means as the most probable values of the quantity measured (Keynes, 1921). Furthermore, the normal or χ^2 distribution arises from the assumption that the two distributions p and q are close in the sense that higher than quadratic terms in their difference can be neglected.

In this paper, we relate the logarithm of the joint error laws to a difference in the weighted means. At least as far as the common means are concerned (i.e., the arithmetic, geometric, and harmonic means), the difference in the two weighted means always results in a discrimination between a nonuniform and a uniform distribution. Well-known inequalities are shown to be responsible for the fact that the joint error law is a *bona fide* probability density. We then give a number of different proofs showing that the difference in the weighted means is always of the same sign. This cannot be derived from the well-known theorems of comparability of weighted means (Hardy *et al.*, 1952) because the functions to be compared are the same, while their arguments are different. The difference in the weighted means will be shown to be related to a new discrimination inequality through the mean-value theorem. Finally, we will derive a new statistic, akin to the χ^2 statistics of Pearson and Neyman, from a first-order approximation of the symmetric divergence of the difference in the powers of the two distributions.

2. CRITERIA OF 'SPREADING'

If sets of N nonnegative numbers $a_1, a_2, \ldots, a_N$ (or b) are the realizations of random events with the same probability density f, then thermostatistics tells us that the error law leading to the arithmetic mean $\overline{a}$ of the measurements as the most probable value of the quantity measured is given by (Lavenda, 1991)

$$\log f_N(a; \overline{a}) = N\overline{a} \log \overline{a} - \sum_{i=1}^{N} a_i \log a_i \leq 0 \tag{1}$$

The inequality is due to the fact that Shannon entropy

$$H_1(a) := -\sum_{i=1}^{N} a_i \log a_i \tag{2}$$

tends to increase as the set of variables become less spread out. This is none other than a manifestation of concavity (Marshall and Olkin, 1979), $H_1(\overline{a}) \geq H_1(a)$. The same property of concavity is responsible for Shannon's inequality, where if p and q are two complete distributions, then

$$-\sum_{i=1}^{N} p_i \log\left(\frac{p_i}{q_i}\right) \leq 0 \tag{3}$$

We may say that p majorizes q: $p \succ q$. Interchanging p and q gives a second inequality which when added to (3) gives

$$\mathcal{J}(p, q) = \sum_{i=1}^{N} (p_i - q_i) \log\left(\frac{p_i}{q_i}\right) \leq 0 \tag{4}$$

This is known as the 'symmetric divergence' for the discrimination between the two probability distributions p and q (Kullback, 1959). It is well known that (4) does not satisfy the triangle inequality (Kullback, 1959, p. 6), and therefore the symmetric divergence cannot be considered a distance in a finite-dimensional space equipped with a metric.

In fact, one would like to generalize (4) to

$$\mathcal{J}(p, q) = \sum_{i=1}^{N} (p_i - q_i)[h(q) - h(p)] \geq 0 \tag{5}$$

where h is continuous and strictly monotonic. The goals of this paper are (i) to show that monotonic power laws of the form $t^{\alpha-1}$, where $\alpha \in (0, 1)$ or $\alpha \in (1, 2]$, also satisfy (5), and (ii) to relate the discrimination with the error laws of probability theory that select out the familiar mean values as the most probable ones, by expressing the logarithm of the joint error laws as the difference of weighted means for the discrimination between a nonuniform probability distribution and the uniform one.

From what has been said so far, one might be tempted to generalize (1) in the following manner. Consider the difference of the weighted means as determining the logarithm of the joint error law:

$$\log f_N(a) = \chi^{-1}\left(\sum_{i=1}^{N} p_i \chi(a_i)\right) - \psi^{-1}\left(\sum_{i=1}^{N} p_i \psi(a_i)\right) \tag{6}$$

where χ and ψ are strictly monotonic and continuous. Denoting by

$$\chi\{\psi^{-1}(x)\} := \chi \circ \psi^{-1} = \Phi(x) \tag{7}$$

the composition function, we can write (6) as

$$\log f_N(a) = \chi^{-1}\left(\sum_{i=1}^{N} p_i \chi(a_i)\right) - \chi^{-1}\left[\Phi\left(\sum_{i=1}^{N} p_i \psi(a_i)\right)\right]$$

If we set $x = \psi(a)$ and $a = \psi^{-1}(x)$ (Hardy *et al.*, 1952, p. 70), then by virtue of the mean-value theorem, the difference in the weighted means can cast in the form (Cargo and Shisha, 1970)

$$\log f_N(x) = (\chi^{-1})'(\check{x})\left\{\sum_{i=1}^{N} p_i\Phi(x_i) - \Phi\left(\sum_{i=1}^{N} p_i x_i\right)\right\} \tag{8}$$

where $\check{x}$ is some value in the domain of x. If χ is increasing, then a necessary and sufficient condition that (8) be negative is that Φ should be concave (Hardy *et al.*, 1952, p. 75).

However, in order to compare the weighted means in (6), the composition function (7) cannot be linear, for otherwise the difference in the weighted means vanishes. If, however, we are discriminating between two sets of probabilities p and q with weighted means of the same order, then their difference will not vanish. The discrimination is of the type (5), with power laws replacing logarithms. The relationship between the difference of the weighted means in (6) to the directed divergence (5) is given by the mean-value theorem. Hence, we will have a direct chain of relations leading from the joint error laws, for the common means to be the most probable values, in terms of the difference of weighted means, comparing a probability distribution p or q with the uniform distribution $1/N$ to the directed divergence involving powers instead of logarithms. In the same way that the first-order approximation of the logarithm of a ratio, by the mean of its upper and lower bounds, leads to the Pearson χ^2 static, the first-order approximation of the difference in power laws, by the mean of its upper and lower bounds, gives a new statistic which reduces to the χ^2 statistic in a given limit.

3. ERROR LAWS AND WEIGHTED MEANS

In this section we show that the logarithm of the joint error laws leading to the most familiar means as the most probable value of the quantity measured are proportional to the negative of the difference of the weighted means:

$$\left(\sum_{i=1}^{N} p_i^{\alpha}\right)^{1/(\alpha-1)} - \left(\sum_{i=1}^{N} p_i q_i^{\alpha-1}\right)^{1/(\alpha-1)} \tag{9}$$

where p and q are two complete probability distributions, and the characteristic exponent is either restricted to the semiclosed interval $\alpha \in (1, 2]$, or the open interval, $\alpha \in (0, 1)$.

3.1. Arithmetic Mean

In order that the arithmetic mean $\bar{a}$ be the most probable value of the quantities a_i measured,

$$\sum_{i=1}^{N} (a_i - \bar{a}) = 0 \tag{10}$$

it must coincide with the stationary condition for the extremum of the probability density of a_i

$$\frac{d}{d\bar{a}} \log f(a_i; \bar{a}) = 0 \tag{11}$$

Since the quantities a_i are independent and have a common distribution, whose density is f, (10) will be equivalent to (11) provided that they are proportional to one another (Keynes, 1921)

$$\frac{d}{d\bar{a}} \log f(a_i; \bar{a}) = \varphi''(\bar{a})(\bar{a} - a_i)$$

where, for convenience, the function of proportionality $\varphi''(\bar{a})$ is taken to be the second derivative of some function of $\bar{a}$, independent of the measurements a_i. For if it did depend upon a measurement a_i, the proportionality would be violated for the density $f(a_j; \bar{a})$ since it does not depend on a_i.

An integration by parts leads to the error law:

$$\log f(a_i; \bar{a}) = \varphi'(\bar{a})(\bar{a} - a_i) - \varphi(\bar{a}) + \psi(a_i) \leq 0 \tag{12}$$

where $\psi(a_i)$ is a function of integration that does not depend upon $\bar{a}$. Ideally, the error law should be a function of the error only. This, more special assumption was invoked by Gauss to obtain the normal law of error. For if we set $\varphi(\bar{a}) = -\bar{a}^2/2$ and $\psi(a_i) = -a_i^2/2$, (12) becomes

$$\log f(a_i; \bar{a}) = -\tfrac{1}{2}(a_i - \bar{a})^2 \tag{13}$$

which is none other than Gauss' law of error showing that negative and positive errors of the same absolute amounts are equally likely.

Since the errors committed are independent of one another, the N-joint error law is the product of the individual densities. In terms of its logarithm, we have

$$\log f_N(\bar{a}) := \sum_{i=1}^{N} \log f(a_i; \bar{a}) = \frac{1}{2}\left(\bar{a}^2 N - \sum_{i=1}^{N} a_i^2\right) \tag{14}$$

which must be less than zero if f_N is to be a *bona fide* probability density. It is precisely the Cauchy inequality (Hardy *et al.*, 1952, Theorem 7, p. 16)

$$\sum_{i=1}^{N} a_i^2 \sum_{i=1}^{N} b_i^2 > \left(\sum_{i=1}^{N} a_i b_i\right)^2$$

with $b = 1$, which guarantees (14) is negative.

Introducing the probabilities $p_i = a_i/\Sigma_{i=1}^{N} a_i$ into the joint error law (14) leads to

$$\log f_N(p, q) = \frac{1}{2}(\bar{a}N)^2\left(\sum_{i=1}^{N} p_i q_i - \sum_{i=1}^{N} p_i^2\right) \tag{15}$$

In order that (15) be equivalent to (14), we have to set $q = 1/N$. It will then be appreciated that the logarithm of the joint error law, leading to the arithmetic mean as the most probability value of the quantity measured, is proportional to the negative of the difference of the weighted means (9) for $\alpha = 2$.

It is quite remarkable—and it will reappear time and time again—that simple inequalities are behind the error laws. Moreover, in all the error laws to be derived, we will be always be comparing a nonuniform distribution of probabilities, either p or q, with the uniform distribution, $1/N$.

For classical thermodynamic systems $\varphi(\bar{a}) = -\bar{a}(\log \bar{a} - 1)$, and ψ is the same function of a_i that φ is of $\bar{a}$ (Lavenda, 1991). Otherwise, there would be a different thermodynamics for the microcanonical and canonical ensembles. With these choices, the error law (12) becomes the discrete Poisson distribution,

$$f(a_i; \bar{a}) = \frac{\bar{a}^{a_i}}{a_i!}\exp(-\bar{a}) \tag{16}$$

provided the a_i are large enough to validate the use of Stirling's approximation. Taking the product of the different error laws (16) gives the N-joint error law for the Poisson distribution as

$$f_N(\bar{a}) = \exp\left\{-\sum_{i=1}^{N}\left(a_i \log\left(\frac{a_i}{\bar{a}}\right) - a_i + \bar{a}\right)\right\} \tag{17}$$

Each term in the sum is nonnegative by virtue of the inequality $x \log(x/y) \geq x - y$. Now setting $a_i = p_i N$ so that $\bar{a} = 1$ gives

$$f_N(p) = \exp\left(-N\sum_{i=1}^{N} p_i \log p_i\right) \tag{18}$$

which can be taken as Boltzmann's principle relating the probability of committing N errors to the Shannon entropy (2). In terms of our original quantities a_i the condition that (18) be a *bona fide* probability distribution is

$$H_1(\bar{a}) \geq \frac{1}{N}\sum_{i=1}^{N} H_1(a_i)$$

The Shannon entropy manifests a tendency to increase as the measurements a_i become less spread out.

3.2. Geometric Mean

The error law leading to the geometric mean $\tilde{p}$ as the most probable value of the quantity measured equates each term in the sum

$$\sum_{i=1}^{N} \left(p_i \log p_i - \frac{1}{N} \log \tilde{p} \right) = 0$$

with the corresponding term in $\Sigma_{i=1}^{N} d \log f(p_i; \tilde{p})/d\tilde{p} = 0$. The probability densities $f(p_i; \tilde{p})$ for each of the probabilities p_i satisfies

$$\frac{d}{d\tilde{p}} \log f(p_i; \tilde{p}) = \varphi'(\tilde{p}) \log \left(\frac{p_i^{p_i}}{\tilde{p}^{1/N}} \right)$$

where the function of proportionality φ' can only depend upon $\tilde{p}$ since the p_i are independent. An integration by parts then gives

$$\log f(p_i; \tilde{p}) = \varphi(\tilde{p}) \left(p_i \log p_i - \frac{1}{N} \log \tilde{p} \right) + \frac{1}{N} \int \varphi(\tilde{p})\, d \log \tilde{p} + \psi(p_i)$$

where again the function of integration ψ can only depend on p_i and must be independent of $\tilde{p}$ (Keynes, 1921).

This error law is not symmetrical: positive and negative errors of the same magnitude are not equally as likely. The simplest law of error leading to the geometric mean is obtained by putting $\varphi(\tilde{p}) = -\tilde{p}N$ (Keynes, 1921), and for the sake of symmetry we set $\psi(p_i) = \prod_{i=1}^{N} q_i^{p_i} = \tilde{q}$, where $q_i \neq p_i$ and $\Sigma_{i=1}^{N} q_i = 1$. The resulting law of error is

$$f(p_i; \tilde{p}) = \left(\frac{\tilde{p}}{p_i^{p_i N}} \right)^{\tilde{p}} \exp\{-(\tilde{p} - \tilde{q})\} \tag{19}$$

Since the measurements p_i are independent, the joint probability density is the product of the individuals densities,

$$f_N(\tilde{p}) = \prod_{i=1}^{N} f(p_i; \tilde{p}) = \exp\{-N(\tilde{p} - \tilde{q})\} \tag{20}$$

It remains to be shown that $\tilde{p} \geq \tilde{q}$. Shannon's inequality (3), being a consequence of the concavity of the Shannon entropy (2), can be written as $\log \prod_{i=1}^{N} (q_i/p_i)^{p_i} \leq 0$. This means that $\prod_{i=1}^{N} (q_i/p_i)^{p_i} \leq 1$, and it follows at once that $\tilde{q} \leq \tilde{p}$. If it happens that $q_i = 1/N$, Shannon's inequality (3) requires $\tilde{p} > 1/N$, or equivalently,

$$H_0 := \log N \geq H_1(p) := -\log \tilde{p}$$

In other words, the Hartley entropy H_0 cannot be inferior to the Shannon entropy H_1.

Moreover, the concavity of $\tilde{q}$ implies the concavity of $\log \tilde{q}$, while the convexity of $\log \tilde{p}$ implies the convexity of $\tilde{p}$ (Marshall and Olkin, 1979, B.7.b p. 451). The difference in the exponent of the joint error law for the geometric mean (20) is a limiting case of the difference of the weighted means (9) as $\alpha \uparrow 1$. The characteristic exponent α now lies in the open interval (0, 1).

3.3. Harmonic Mean

We now consider the diametrically opposite limit as $\alpha \to 0$, and determine the law of error leading to the harmonic mean $\hat{a}$,

$$\sum_{i=1}^{N} \left(\frac{Np_i}{a_i} - \frac{1}{\hat{a}} \right) = 0 \tag{21}$$

as the most probable value of the quantity measured. Deviations from (21) must be equivalent to deviations from $\Sigma_{i=1}^{N}\, d \log f(a_i; \hat{a})/d\hat{a} = 0$, so that the usual procedure gives the error law:

$$\log f(a_i; \hat{a}) = \varphi(\hat{a}) \left(\frac{Np_i}{a_i} - \frac{1}{\hat{a}} \right) - \int \frac{\varphi(\hat{a})}{\hat{a}^2}\, d\hat{a} + \psi(a_i) \tag{22}$$

The simplest form of the error law is obtained by setting $\varphi = -\hat{a}^2$ and $\psi = -a_i/Np_i$ (Keynes, 1921). The error law (22) then reduces to

$$\log f(a_i; \hat{a}) = -\left(\frac{Np_i}{a_i} \right) (\hat{a} - a_i)^2 \tag{23}$$

showing again that errors of the same magnitude and different sign are not equally as likely.

Since the a_i are the outcomes of independent events with the same distribution, the logarithm of the joint error law is obtained by summing (23) over all N. If we want to use the fact that the harmonic mean is less than the arithmetic mean unless all the a_i are equal, we now have to set $p = 1/N$. We then obtain

$$\log f_N(\hat{a}) = \sum_{i=1}^{N} \log f(a_i; \hat{a}) = -N(\overline{a} - \hat{a}) < 0 \tag{24}$$

asserting that the arithmetic mean is always greater than the geometric mean unless all the a_i are equal.

Introducing the probabilities $q_i = a_i/\Sigma_{i=1}^{N}\, a_i$ into the joint error law (24) gives

$$\log f_N(\hat{q}) = -\bar{a}N^2\left\{\frac{1}{N} - \left(\frac{1}{N}\sum_{i=1}^{N}\frac{1}{q_i}\right)^{-1}\right\}$$

The term within the curly brackets is precisely the difference of the weighted means (9), with the p distribution as uniform, in the limit as $\alpha \downarrow 0$. In this limit, the second weighted mean in (9) coincides with the harmonic mean, $\tilde{q}$.

Thus, we have proved the following.

Theorem 1. The joint error laws leading to the common means as the most probable values of the quantity measured are proportional to the negative of the difference in the weighted means (9). The arithmetic mean arises in the limit $\alpha = 2$, while in the limits $\alpha \uparrow 1$ and $\alpha \downarrow 0$, the geometric mean and the harmonic means are most probable values of the quantity measured, respectively.

We now prove that the difference of the weighted means (9) is always positive for $p \neq q$. We then relate it to a new class of directed divergences.

4. PROOF THROUGH MAJORIZATION

A nonuniform distribution p is said to majorize the uniform distribution $1/N$: $p \succ 1/N$ (Marshall and Olkin, 1979). Since $\Sigma_{i=1}^{N} p_i^2$ is Schur-convex, it follows that $\Sigma_{i=1}^{N} p_i^2 \geq 1/N$, showing that (15) is negative when $q = 1/N$.

First consider the weighted mean

$$\mathcal{M}_{\alpha-1}(a) = \left(\sum_{i=1}^{N} p_i^{\alpha}\right)^{1/(\alpha-1)} \tag{25}$$

For $\alpha \in (1, 2]$, p^{α} is convex, and $\mathcal{M}_{(\alpha-1)}$ is increasing. Hence, (25) is Schur-convex, while for $\alpha \in (0, 1)$, p^{α} is concave, and $\mathcal{M}_{-(1-\alpha)}$ is decreasing. Consequently, (25) is Schur-convex on both the intervals $\alpha \in (1, 2]$ and $\alpha \in (0, 1)$, and $p \succ 1/N$ implies $(\Sigma_{i=1}^{N} p_i^{\alpha})^{1/(\alpha-1)} \geq 1/N$.

Next consider the weighted mean

$$\mathcal{M}_{\alpha-1}(q) = \left(\sum_{i=1}^{N} p_i q_i^{\alpha-1}\right)^{1/(\alpha-1)} \tag{26}$$

for $q \neq p$. For fixed p, $q^{\alpha-1}$ is concave in the interval $\alpha \in (1, 2]$, and $\mathcal{M}_{\alpha-1}(q)$ is increasing. This implies that (26) is Schur-concave on this interval. Alternatively, for $\alpha \in (0, 1)$, $q^{\alpha-1}$ is convex and $\mathcal{M}_{-(1-\alpha)}(q)$ is decreasing. Again the weighted mean (26) is Schur-concave. Majorization $p \succ 1/N$ now implies $(\Sigma_{i=1}^{N} p_i q_i^{\alpha-1})^{1/(\alpha-1)} \leq 1/N$. Hence, the difference in the weighted means (9) is positive semidefinite.

5. DERIVATION OF A NEW DISCRIMINATION INEQUALITY

The difference of the weighted means (9) for $\alpha \in (0, 1)$ can be looked upon as the error committed in choosing the probabilities q when the true probabilities are p. In fact, the inequality of their logarithms,

$$-\frac{1}{1-\alpha}\log\left(\sum_{i=1}^{N} p_i^{\alpha}\right) \geq -\frac{1}{1-\alpha}\log\left(\sum_{i=1}^{N} p_i q_i^{-(1-\alpha)}\right)$$

is just Shannon's inequality (3) in the limit as $\alpha \uparrow 1$.

Employing the mean-value theorem, we may write the difference (9) as (Cargo and Shisha, 1970)

$$\mathcal{M}_{(\alpha-1)}(p) - \mathcal{M}_{(\alpha-1)}(q) = (h^{-1})'(r)\{\mathcal{M}_{\alpha-1}^{\alpha-1}(p) - \mathcal{M}_{\alpha-1}^{\alpha-1}(q)\} \tag{27}$$

where $(h^{-1})'(r) = r^{(2-\alpha)/(\alpha-1)}/(\alpha - 1)$ for some $r \in [q, p]$. Since $(h^{-1})'(r) \lesseqgtr 0$ as $\alpha \lesseqgtr 1$, the condition that the difference of the weight means be greater than zero implies

$$\mathcal{C}_{\alpha<1} = \sum_{i=1}^{N} p_i\{q_i^{\alpha-1} - p_i^{\alpha-1}\} \geq 0 \tag{28}$$

for $\alpha < 1$, while

$$\mathcal{C}_{\alpha>1} = \sum_{i=1}^{N} p_i\{p_i^{\alpha-1} - q_i^{\alpha-1}\} \geq 0 \tag{29}$$

for $\alpha \in (1, 2]$. These inequalities have the form of a Shannon-type inequality:

$$\sum_{i=1}^{N} p_i\{h(q_i) - h(p_i)\}\begin{cases}\geq 0 & \text{if } \alpha \in (0, 1)\\ \leq 0 & \text{if } \alpha \in (1, 2]\end{cases} \tag{30}$$

expressing the error committed when the estimates q are used instead of the true probabilities p (Good, 1952).

Aczél and Daróczy (1975, p. 114) ask what type of functions h satisfy the first inequality in (30). Certainly $h(x) = -\log x$ satisfies it, because the first inequality in (30) is a direct consequence of the concavity of $-x \log x$. Aczél and Daróczy (1975, Theorem, 4.3.8, p. 116) make even the following stronger statement:

> If and only if h satisfies $\Sigma_{i=1}^{N} p_i h(p_i) \leq \Sigma_{i=1}^{N} p_i h(q_i)$ for a fixed $N > 2$ does it have to be of the form $h(p) = c \log p + b$ for all $p \in (0, 1)$ and $c \leq 0$, b constants.

It is to the "if and only if" assertion that we take exception. Their derivation makes use of two distinct points $p_1 + p_2 = q_1 + q_2 = r$, where $r \in (0, 1)$, and, for simplicity of notation, they write $p_1 = p$ and $q_1 = q$. Assuming the first inequality holds in (30), they obtain an inequality which is symmetric

in p and q. Interchanging p and q gives a second inequality. At this point they assume $q > p$ and get the upper and lower bounds on the ratio $[h(q) - h(p)]/(q - p)$, namely

$$\frac{r-q}{q}\frac{h(r-p)-h(r-q)}{(r-p)-(r-q)} \geq \frac{h(q)-h(p)}{q-p} \geq \frac{r-p}{p}\frac{h(r-p)-h(r-q)}{(r-p)-(r-q)}$$

Letting $q \to p$, the bounds tend to

$$\frac{r-p}{p}h'(r-p)$$

But this they claim must be equal to $h'(p)$, the derivative of h at p. Hence, $ph'(p) = (r - p)h'(r - p)$. Since the only restriction on r is that it lie in the open interval $(p, 1)$, they conclude that $ph'(p) = \text{const} = \gamma < 0$. From this they deduce that $h(p) = -\gamma \log p$.

Aczél and Daróczy establish the properties that h is monotonic and nonincreasing. Hence, its *logarithm* is also monotonic and nonincreasing. Replacing h by $\log h$, the above condition gives $ph'(p)/h(p) = \alpha - 1$, which upon integration gives $h(p) = cp^{\alpha-1}$, where the constant of integration $c > 0$. Thus, the first inequality in (30) is satisfied by any completely monotone, nonincreasing, Schur-convex function, and in particular by $h(p) = cp^{\alpha-1}$ for $\alpha \in (0, 1)$.

Prior to a general proof, let us consider the limiting situations. Multiplying (28) by a positive number $(1 - \alpha)^{-1} > 0$ does not change the inequality, so that

$$\frac{1}{1-\alpha}\sum_{i=1}^{N} p_i(q_i^{\alpha-1} - p_i^{\alpha-1}) \geq 0 \tag{31}$$

In the limit as $\alpha \uparrow 1$, (31) becomes Shannon's inequality (3), while in the opposite limit as $\alpha \downarrow 0$, the inequality reduces to $\Sigma_{i=1}^{N} p_i/q_i - N \geq 0$. Inserting for the probabilities $q_i = a_i/\Sigma_{i=1}^{N} a_i$ converts this into the arithmetic-geometric mean inequality $\bar{a} = \Sigma_{i=1}^{N} a_i/N > \hat{a} = 1/\Sigma_{i=1}^{N} (p_i/a_i)$ unless all the a_i are equal.

As we have mentioned, Aczél and Daróczy (1975, p. 117) deduce the properties that h must be strictly monotonic and nonincreasing. If and only if h is Schur-convex will (Marshall and Olkin, 1979, p. 447)

$$\sum_{i=1}^{N} (p_i - q_i)[h'(p_i) - h'(q_i)] \geq 0 \tag{32}$$

hold. There are only two classes of such functions (Hardy *et al.*, 1952, p. 65): $t^{\alpha-1}$, for $\alpha < 1$, and $-\log t$. Both are strictly monotonic and nonincreasing. Therefore, Kullback's directed divergence can be extended to a second class of completely monotone functions $t^{\alpha-1}$, so that (28) and (29) can be considered

as directed divergences when $\alpha \in (0, 1)$ or $\alpha \in (1, 2]$, respectively. We now proceed to give a first general proof.

Consider the Hölder inequality (Hardy *et al.*, 1952, Theorem 10, p. 21)

$$\sum_{i=1}^{N} p_i^{\alpha} q_i^{1-\alpha} \leq \left(\sum_{i=1}^{N} p_i \right)^{\alpha} \left(\sum_{i=1}^{N} q_i \right)^{1-\alpha} = 1$$

for $0 < \alpha < 1$, while for $\alpha > 1$ the inequality is reversed. This can be written as (Aczél and Daróczy, 1975, p. 208)

$$\sum_{i=1}^{N} q_i \left(\frac{p_i}{q_i} \right)^{\alpha} \leq 1, \qquad \alpha \in (0, 1). \tag{33}$$

Kullback (1959, p. 40) defined (33) as the moment generating function,

$$\mathcal{I}(\alpha) = \sum_{i=1}^{N} q_i e^{-\alpha \theta_i} \tag{34}$$

where $\theta_i = \log(q_i/p_i)$ is the conjugate of α in the sense of a Legendre transform. The logarithm of the moment generating function is strictly convex in α, and its Legendre transform $\alpha\theta - \log \mathcal{I}$ varies continuously and monotonically from 0 to the 'minimum discrimination information' $\Sigma_{i=1}^{N} q_i \log(q_i/p_i)$ as α varies from 0 to 1.

The discrimination between the probabilities p_i and q_i will be informative when they have diametrically opposite properties. For the sake of concreteness, let us consider the probabilities p_i and q_i as nondecreasing and nonincreasing, respectively, in $i = 1, \ldots, N$. The two distributions are *oppositely ordered* so that (Hardy *et al.*, 1952, p. 43)

$$\left(\frac{1}{p_i} - \frac{1}{p_j} \right)(q_i - q_j) \geq 0$$

and consequently Tchebychef's inequality for the weighted mean, $\mathcal{U}(a) = \Sigma_{i=1}^{N} p_i a_i$, will read (Hardy *et al.*, 1952, Theorem 43, p. 43)

$$\begin{aligned} &\mathcal{U}[(p/q)^{\alpha-1}] - \mathcal{U}(p^{\alpha-1})\mathcal{U}(1/q^{\alpha-1}) \\ &\quad = \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} p_i p_j \left(\frac{1}{p_i^{1-\alpha}} - \frac{1}{p_j^{1-\alpha}} \right)(q_i^{1-\alpha} - q_j^{1-\alpha}) \geq 0 \end{aligned}$$

Moreover, it follows from Hölder's inequality (33) that $\mathcal{U}(p^{\alpha-1})\mathcal{U}(1/q^{\alpha-1}) \leq 1$, or, equivalently,

$$\sum_{i=1}^{N} p_i^{\alpha} \leq 1 \Big/ \sum_{i=1}^{N} p_i / q_i^{\alpha-1} \tag{35}$$

The right-hand side of (35) is the harmonic mean of $q^{\alpha-1}$, and in virtue of the arithmetic-harmonic mean inequality

$$1 \Big/ \sum_{i=1}^{N} p_i / q_i^{\alpha-1} \leq \sum_{i=1}^{N} p_i q_i^{\alpha-1}$$

there results inequality (28). The reverse inequality (29) can be derived from the reverse of Hölder's inequality, $\Sigma_{i=1}^{N} p_i^{\alpha} q_i^{1-\alpha} \geq 1$, for $\alpha > 1$.

A second proof of inequality (28) can be given in terms of majorization (Marshall *et al.*, 1967). Suppose, as before, that p_i/q_i is increasing in $i = 1, \ldots, N$. Then the ratio of the weighted means,

$$\left(\frac{\Sigma_{i=1}^{N} p_i q_i^{\alpha-1}}{\Sigma_{i=1}^{N} p_i^{\alpha}} \right)^{1/(\alpha-1)}$$

is decreasing in $\alpha \neq 1$. Thus, $p \succ q$, implying that the partial sums of

$$\sum_{i=1}^{k} \frac{p_i^{\alpha}}{\Sigma_{i=1}^{N} p_i^{\alpha}} \geq \sum_{i=1}^{k} \frac{p_i q_i^{\alpha-1}}{\Sigma_{i=1}^{k} p_i q_i^{\alpha-1}} \tag{36}$$

for $k = 1, \ldots, N-1$. The inequality is a consequence of the ordering condition that p_i/q_i should be increasing in $i = 1, \ldots, N$, namely

$$\sum_{i=1}^{k} p_i^{\alpha} \sum_{j=1}^{N} p_j q_j^{\alpha-1} - \sum_{i=1}^{k} p_i q_i^{\alpha-1} \sum_{j=1}^{N} p_j^{\alpha}$$

$$= \sum_{i=1}^{k} \sum_{j=k+1}^{N} p_i^{\alpha} p_j^{\alpha} \left[\left(\frac{p_j}{q_j} \right)^{1-\alpha} - \left(\frac{p_i}{q_i} \right)^{\alpha-1} \right] \geq 0$$

which is a sufficient condition for majorization (Marshall *et al.*, 1967).

It will suffice to consider the $k = 1$ term in (36). Upon rearrangement we have

$$\frac{\Sigma_{i=1}^{N} p_i^{\alpha}}{\Sigma_{i=1}^{N} p_i q_i^{\alpha-1}} \leq \left(\frac{p_i}{q_i} \right)^{\alpha-1}$$

Multiplying both sides by p_i and summing over all N gives the following result.

Theorem 2. Inequality (28), resulting from Hölder's inequality (33), is the directed divergence which measures the difficulty in discriminating between the distributions p and q for values of $\alpha \in (0, 1)$. Inequality (29) can likewise be derived from the reverse of the Hölder inequality for $\alpha > 1$.

6. DISTANCE AND DISCRIMINATION

When attempting to discriminate between alternative probability distributions, it is desirable to have a measure of their 'distance,' or how 'close' they are to one another. Any candidate for a distance, $\rho(p, q)$, between p and q, must possess the following properties:

1. $\rho(p, q) > 0$, for $p \neq q$, and $\rho(p, q) = 0$ for $p = q$.
2. $\rho(p, q) = \rho(q, p)$.
3. $\rho(p, q) \geq \rho(p, r) + \rho(r, q)$.

The triangle inequality (3) is none other than the property of subadditivity (Rudin, 1973). For positively homogeneous functions of degree 1, subadditivity coincides with convexity. this is the reason for using convexity and subadditivity interchangeably when discussing distances and norms. As we have mentioned, the Kullback divergence does not satisfy the triangle inequality, (3), and hence cannot be considered a distance (Kullback, 1959, p. 6).

Euclidean space uses the Minkowski inequality (Hardy *et al.*, 1952, Theorem 25, p. 31)

$$\left(\sum_{i=1}^{N}(a_i + b_i)^s\right)^{1/s} \leq \left(\sum_{i=1}^{N} a_i^s\right)^{1/s} + \left(\sum_{i=1}^{N} a_i^s\right)^{1/s}$$

to define the distance

$$\|x - y\| := \left(\sum_{i=1}^{N} |x_i - y_i|^2\right)^{1/2}$$

on $\mathbf{R}^N$ for $1 \leq s < \infty$. For $s < 1$ or $s < 0$ Minkowski's inequality is reversed, so that it cannot be associated with a triangle inequality. However, there does exist a companion to Minkowski's inequality (Hardy *et al.*, 1952, Theorem 27, p. 32)

$$\sum_{i=1}^{N} p_i(a_i + b_i)^t < \sum_{i=1}^{N} p_i a^t + \sum_{i=1}^{N} p_i b_i^t \tag{37}$$

$0 < t < 1$, which can formally be associated with the property of subadditivity (Rudin, 1973, p. 35).

We first shall consider the case $t = \alpha - 1$, where $\alpha \in (1, 2)$. On the open interval $(0, 1)$ we set $a = q$ and $b = p - q$ in (37) to obtain

$$\sum_{i=1}^{N} p_i(p_i - q_i)^{\alpha-1} > \sum_{i=1}^{N} p_i[q_i^{\alpha-1} - p_i^{\alpha-1}]$$

The inequality

$$\sum_{i=1}^{N} p_i(p_i - q_i)^{\alpha-1} \leq \sum_{i=1}^{N} p_i\left|p_i - q_i\right|^{\alpha-1} \leq \sum_{i=1}^{N} \left|p_i - q_i\right|^{\alpha-1}$$

leads to the definition

$$\rho_{\alpha>1}(p, q) := \sum_{i=1}^{N} \left|p_i - q_i\right|^{\alpha-1} \tag{38}$$

of a distance. Although (38) has the properties of being positive semidefinite, symmetric, and subadditive, it is not a norm, since it is not positively homogeneous of degree 1. In order to satisfy this condition, a weighted mean is required. but such a weighted mean, with $\alpha \in (0, 1)$, would not satisfy a triangle type of inequality, and hence would only be a partial norm. Moreover, the weighted mean would turn out to be concave function and consequently there would be no convex open sets, so that 0 would be the only continuous linear mapping of the space into any locally convex space (Rudin, 1973, p. 35). We now relate (38) to the logarithm of the joint error law.

Appealing to the mean-value theorem (27), the difference of the weighted means (9) is bounded from above by

$$\mathcal{M}_{\alpha-1}(p) - \mathcal{M}_{\alpha-1}(q) \leq \sup\left|(h^{-1})'\right| \cdot \rho_{\alpha>1}(p, q)$$

Since the logarithm of the joint error law is proportional to the negative difference of the weighted means (9), namely

$$\log f_N(p, q) = -\kappa\{\mathcal{M}_{\alpha-1}(p) - \mathcal{M}_{\alpha-1}(q)\} \tag{39}$$

where $\kappa > 0$ is a constant of proportionality, we have

$$\left|\log f_N(p, q)\right| \leq \kappa_1 \rho_{\alpha>1}(p, q) \tag{40}$$

where the constant $\kappa_1 = \sup\left|(h^{-1})'\right|\kappa$.

Next consider values of $\alpha \in (0, 1)$. The distance (38) cannot be used since it is infinite when $p = q$. However, if we set $t = 1 - \alpha$, $a = 1/p$, and $b = 1/q - 1/p > 0$ in (37), we get

$$\sum_{i=1}^{N} p_i(1/q_i - 1/p_i)^{1-\alpha} = \sum_{i=1}^{N} p_i^{\alpha}\left(\frac{p_i - q_i}{q_i}\right)^{1-\alpha} \geq \sum_{i=1}^{N} p_i/q_i^{1-\alpha} - \sum_{i=1}^{N} p_i^{\alpha}$$

Since

$$\sum_{i=1}^{N} p_i(1/q_i - 1/p_i)^{1-\alpha} \leq \sum_{i=1}^{N} p_i\left|1/q_i - 1/p_i\right|^{1-\alpha} \leq \sum_{i=1}^{N} \left|1/q_i - 1/p_i\right|^{1-\alpha}$$

the distance from q to p can be defined as

$$\rho_{\alpha<1}(q, p) := \sum_{i=1}^{N} \left|1/q_i - 1/p_i\right|^{1-\alpha} \tag{41}$$

The distance (41) on the open interval $\alpha \in (0, 1)$ has all the same properties that (38) has on the interval $\alpha \in (1, 2)$. In particular, it vanishes when $p = q$, which is the reason why (38) cannot be used for values of $\alpha \in (0, 1)$.

Unlike (40), there is no direct relationship between (41) and the logarithm of the joint error density. However, if we use the Minkowski inequality[2]

$$\left(\sum_{i=1}^{N} p_i(a_i + b_i)^r\right)^{1/r} > \left(\sum_{i=1}^{N} p_i a_i^r\right)^{1/r} + \left(\sum_{i=1}^{N} p_i b_i^r\right)^{1/r} \tag{42}$$

for $r < 0$, we obtain an upper bound on the logarithm of the error law. Specifically, setting $r = -(1 - \alpha)$, $a = q$, and $b = p - q$, we find that Minkowski's inequality is converted into

$$\left(\sum_{i=1}^{N} p_i^{\alpha}\right)^{-1/(1-\alpha)} - \left(\sum_{i=1}^{N} p_i/q_i^{(1-\alpha)}\right)^{-1/(1-\alpha)} > \left(\sum_{i=1}^{N} p_i/(p_i - q_i)^{(1-\alpha)}\right)^{-1/(1-\alpha)}$$

Consequently, the logarithm of the joint error law is bounded by

$$\log f_N(p, q) < -\kappa \cdot \left(\sum_{i=1}^{N} p_i/(p_i - q_i)^{(1-\alpha)}\right)^{-1/(1-\alpha)}; \qquad 0 < \alpha < 1 \tag{43}$$

[2] This inequality can be derived in an analogous way to the usual Minkowski inequality (Hardy *et al.*, 1952, Theorem 24, p. 30), with the exception that the direction of Hölder's inequality is the reverse of the usual one for a negative exponent. In the simplest case we have

$$\sum_{i=1}^{N} p_i(a_i + b_i)^r = \sum_{i=1}^{N} p_i a_i(a_i + b_i)^{r-1} + \sum_{i=1}^{N} p_i b_i(a_i + b_i)^{r-1}$$

Using Hölder's inequality in the form

$$\sum_{i=1}^{N} p_i a_i(a_i + b_i)^{r-1} = \sum_{i=1}^{N} (p_i^{1/r} a_i)(p_i^{1/r}(a_i + b_i))^{r-1} > \left(\sum_{i=1}^{N} (p_i^{1/r} a_i)^r\right)^{1/r} \left(\sum_{i=1}^{N} [p_i^{1/r}(a_i + b_i)]^r\right)^{1/r'}$$

where $r' = r/(r - 1)$, gives

$$\sum_{i=1}^{N} p_i(a_i + b_i)^r > \left[\left(\sum_{i=1}^{N} p_i a_i^r\right)^{1/r} + \left(\sum_{i=1}^{N} p_i b_i^r\right)^{1/r}\right] \left(\sum_{i=1}^{N} p_i(a_i + b_i)^r\right)^{1/r'}$$

and from which Minkowski's inequality (42) follows directly. However, it cannot be used to define a metric because it is the reverse of the triangle inequality. In other words, inequality (42) is a statement of *superadditivity* . Note also that if $r' < 1$ then $r < 0$, and vice versa.

where κ is a positive constant. Inequality (43) guarantees maximum probability for $p = q$. Moreover, due to the conjugate nature of the exponents r and r' in the proof of the Minkowski inequality (42) (cf. footnote 2), by setting $r = \alpha - 1$ for values of $\alpha \in (1, 2)$, we get an analogous upper bound

$$\log f_N(p, q) < -\kappa \cdot \left(\sum_{i=1}^{N} p_i (p_i - q_i)^{\alpha - 1} \right)^{1/(\alpha - 1)} ; \qquad 1 < \alpha < 2 \quad (44)$$

on the logarithm of the joint error probability density.

It is well known that Shannon's inequality (3) can be related to the χ^2 statistic (Kullback, 1959, p. 114). That is, if p is not too distant from q, $\log(p/q)$ can be estimated as the mean of its upper, $(p - q)/q$, and lower, $(p - q)/p$, bounds. Thus, Kullback's symmetric divergence becomes

$$\begin{aligned} \mathscr{E} &= \sum_{i=1}^{N} (p_i - q_i) \log \left(\frac{p_i}{q_i} \right) \\ &\approx \frac{1}{2} \sum_{i=1}^{N} \frac{(p_i - q_i)^2}{p_i} + \frac{1}{2} \sum_{i=1}^{N} \frac{(p_i - q_i)^2}{q_i} : \\ &= \frac{1}{2} (\chi^2 + \chi'^2) \end{aligned}$$

where the first sum is Pearson's χ^2, and the second sum is Neyman's χ'^2 statistic. The closer q is to p, the better the approximation.

In contrast to the χ^2 statistic, the directed divergences (28) and (29) implicate a different type of statistic. For $\alpha < 1$, the symmetric divergence for power laws is

$$\mathscr{T}_{\alpha<1} = \sum_{i=1}^{N} (p_i - q_i)\{q_i^{\alpha - 1} - p_i^{\alpha - 1}\} \geq 0 \quad (45)$$

The difference in the curly brackets in (45) is bounded below and above by [Hardy *et al.*, 1952, Theorem 41, inequality (2.15.1), p. 39]

$$(1 - \alpha) q^{\alpha - 2} (p - q) > q^{\alpha - 1} - p^{\alpha - 1} > (1 - \alpha) p^{\alpha - 2} (p - q)$$

We may therefore use, as a first approximation to $(q^{\alpha - 1} - p^{\alpha - 1})$, the mean of its upper and lower bounds. With this approximation, the symmetric divergence (45) becomes

$$\mathscr{T}_{\alpha<1} \approx \frac{(1 - \alpha)}{2} \sum_{i=1}^{N} \left\{ p_i^{\alpha} \left(\frac{p_i - q_i}{p_i} \right)^2 + q_i^{\alpha} \left(\frac{p_i - q_i}{q_i} \right)^2 \right\} \geq 0 \quad (46)$$

We may say that the variable order α necessitates the weights p^α and q^α on the χ^2 statistics. In the limit as $\alpha \uparrow 1$, the first sum in the curly brackets of (46) reduces to the Pearson χ^2 statistic, while the second sum becomes the Neyman χ'^2 statistic.

For $\alpha > 1$, the difference in the curly brackets of

$$\mathcal{T}_{\alpha>1} = \sum_{i=1}^{N} (q_i - p_i)\{q_i^{\alpha-1} - p_i^{\alpha-1}\} \geq 0 \tag{47}$$

is bounded above and below by [Hardy *et al.*, 1952, Theorem 41, inequality (2.15.2), p. 39]

$$(\alpha - 1)q^{\alpha-2}(q - p) < q^{\alpha-1} - p^{\alpha-1} < (\alpha - 1)p^{\alpha-2}(q - p)$$

If, as a first approximation, we again use the mean of the upper and lower bounds to estimate the difference in the curly brackets of (47), we obtain

$$\mathcal{T}_{\alpha>1} \approx \frac{(\alpha - 1)}{2} \sum_{i=1}^{N} \left\{ p_i^\alpha \left(\frac{p_i - q_i}{p_i} \right)^2 + q_i^\alpha \left(\frac{p_i - q_i}{q_i} \right)^2 \right\} \geq 0 \tag{48}$$

Like the χ^2 statistic, both (46) and (48) measure the departure of the expected from observed frequencies. The χ^2 statistic is an approximation to the logarithm of the likelihood ratio when the distance between the observed from the expected values is small. The likelihood functions are products of the individual probabilities, which must be assumed from the beginning. The assumption of normality is based on a mixture of mathematical convenience and large sample theory, which rely on the law of large numbers and the central limit theorem (Bickel and Doksum, 1977). The choice of the distribution not only determines the statistic, but also, in certain cases, its critical value. In place of the log-likelihood function, we have the log-joint error law. The statistic is determined by the assumption that it is the most probable value of the quantity measured. The sample distribution is a consequence of this choice. And varying the order α gives a continuous range of means which are the modes of the distributions.

ACKNOWLEDGMENTS

The author is grateful to Professors S. Frigo and L. Mangiarotti for helpful discussions. This work was sponsored in part by MURST (60% and 40%) and CNR.

REFERENCES

Aczél, J., and Daróczy, Z. (1975). *On Measures of Information and Their Characterizations* (Academic Press, New York, 1975).

Bickel, P. J., and Doksum, K. A. (1977). *Mathematical Statistics* (Prentice-Hall, Englewood Cliffs, N.J.), p. 225.

Cargo, G. T., and Shisha, O. (1970). In *Inequalities II,* O. Shisha, ed. (Academic Press, New York), pp. 163–178.

Good, I. J. (1952). *J. Roy. Stat. Soc. B* **14**, 107–114.

Hardy, G. H., Littlewood, J. E., and Pólya, G. (1952). *Inequalities* , 2nd ed., Cambridge University Press, Cambridge, §3.4.

Keynes, J. M. (1921). *A Treatise on Probability* (St. Martin's Press, New York), pp. 218–228.

Kullback, S. (1959). *Information Theory and Statistics* (Wiley, New York), Ch. 1.

Lavenda, B. H. (1991). *Statistical Physics: A Probabilistic Approach* (Wiley-Interscience, New York), Ch. 1.

Marshall, A. W., Olkin, I., and Proschan, F. (1967). In *Inequalities,* O. Shisha, ed. (Academic Press, New York), pp. 177–190.

Marshall, A. W., and Olkin, I. (1979). *Inequalities: Theory of Majorization and its Applications* (Academic Press, New York), p. 71.

Rudin, W. (1973). *Functional Analysis* (McGraw-Hill, New York), p. 24.