A General Class of Coefficients of Divergence of One Distribution from Another

By S. M. Ali and S. D. Silvey

University of Manchester

[Received July 1965. Revised September 1965]

Summary

Let P_1 and P_2 be two probability measures on the same space and let ϕ be the generalized Radon–Nikodym derivative of P_2 with respect to P_1. If C is a continuous convex function of a real variable such that the P_1-expectation (generalized as in Section 3) of $C(\phi)$ provides a reasonable coefficient of the P_1-dispersion of ϕ, then this expectation has basic properties which it is natural to demand of a coefficient of divergence of P_2 from P_1. A general class of coefficients of divergence is generated in this way and it is shown that various available measures of divergence, distance, discriminatory information, etc., are members of this class.

1. Introduction

Several coefficients have been suggested in statistical literature to reflect the fact that some probability distributions are "closer together" than others and consequently that it may be "easier to distinguish" between the distributions of one pair than between those of another. Such coefficients have been variously called measures of distance between two distributions (see Adhikari and Joshi, 1956, for references), measures of separation (Rao, 1952), measures of discriminatory information (Chernoff, 1952; Kullback, 1959) and measures of variation-distance (Kolmogorov, 1963). While these coefficients have not all been introduced for exactly the same purpose, as the names given them imply, they have the common property of increasing as the two distributions involved "move apart". Following Kullback (1959), we shall call a coefficient with this property a coefficient of divergence of one distribution from another. It is the object of the present paper to discuss a general method of generating coefficients of divergence and to show that various available coefficients of divergence belong to a general class defined by this method.

The idea underlying the method can be illustrated by considering two probability distributions P_1 and P_2 on the real line, absolutely continuous with respect to Lebesgue measure and with respect to each other, and by visualizing what happens to the ratio $\phi(x) = p_2(x)/p_1(x)$ of their densities as P_1 and P_2 "move apart". If P_1 and P_2 are the same then $\phi(x) \equiv 1$. As P_1 and P_2 move apart ϕ takes larger values on a set of decreasing P_1-probability (incidentally of increasing P_2-probability) and smaller values on a set of increasing P_1-probability (and decreasing P_2-probability). Bearing in mind that $E_1(\phi) = 1$ for all such P_1 and P_2, where E_1 denotes expectation relative to P_1, this picture suggests that the P_1-dispersion of ϕ increases as P_1 and P_2 move apart. Hence we might anticipate that a coefficient of the P_1-dispersion of ϕ will have the basic properties required of a coefficient of divergence of P_2 from P_1. Now the expectation of a convex function of a real random variable measures its dispersion to a greater or lesser extent depending on the nature of this function and so we are led to consider the P_1-expectations of convex functions of ϕ as possible coefficients of divergence.

We need not limit ourselves to distributions on the real line. From now on, P_1 and P_2 will denote probability measures on the same sample space $(\mathscr{X}, \mathscr{F})$; ϕ will be the generalized Radon–Nikodym derivative of P_2 with respect to P_1, generalized in the sense that it is allowed to take the value $+\infty$ on a P_1-null set if P_2 has a component singular with respect to P_1; and we shall consider as potential coefficients of divergence, functionals which can be expressed in the form $E^*\{C(\phi)\}$ where C is a real continuous convex function on the non-negative real numbers, and E^* is a generalized expectation to be defined subsequently.

2. Basic Properties of a Coefficient of Divergence

In this Section we shall list four properties that it seems reasonable to demand of a real coefficient $d(P_1, P_2)$ if this coefficient is to reflect the facts that some distributions may be closer together than others and that it may be more difficult to distinguish between the distributions of one pair than between those of another.

First property. The coefficient $d(P_1, P_2)$ should be defined for all pairs of measures P_1 and P_2 on the same sample space.

Second property. Suppose that $y = t(x)$ is a measurable transformation from $(\mathscr{X}, \mathscr{F})$ onto a measure space $(\mathscr{Y}, \mathscr{G})$. Then we should have

$$d(P_1, P_2) \geqslant d(P_1 t^{-1}, P_2 t^{-1}).$$

Here $P_i t^{-1}$ denotes the induced measure on $\mathscr{Y}$ corresponding to P_i.

This property, which has already been laid down by various authors, for example Adhikari and Joshi (1956) and Kullback (1959), is possibly not self-evident in terms of the formal statement given. It is best thought of in terms of distinguishability of one distribution from another. We should not be able to increase our ability to distinguish between two different distributions by "grouping observations together" —mathematically by limiting consideration to a coarser σ-field than that on which the two distributions are defined. To take an extreme case, suppose that X_1 and X_2 are disjoint measurable sets of $\mathscr{X}$ such that $P_1(X_1) > 0$ and $P_2(X_1) = 0$ whereas $P_1(X_2) = 0$ and $P_2(X_2) > 0$. Given that a sample point is in $(X_1 \cup X_2)$, the additional information about which of the sets X_1 and X_2 it belongs to is of great help in enabling us to distinguish between P_1 and P_2. We cannot suppress the latter information without making it more difficult to distinguish P_2 from P_1. The second property above states this idea in a more general way.

Another particular case may be of value in the justification of this second demand. Suppose that $\{x_n; n = 1, 2, \ldots\}$ is a stochastic process and that P_1 and P_2 are two possible distributions for this process. It is plausible that the more observations made on the process, the greater is our ability to distinguish the true distribution. In other words, if $P_i^{(n)}$ denotes the marginal distribution of $x_1, x_2, \ldots, x_n$ corresponding to P_i $(i = 1, 2)$, then we ought to have

$$d(P_1^{(m)}, P_2^{(m)}) \leqslant d(P_1^{(n)}, P_2^{(n)}) \quad \text{for} \quad m < n.$$

While some writers, for example Adhikari and Joshi (1956), have singled this out as a separate demand it is really a particular case of the second property with t defined by

$$t(x_1, x_2, \ldots, x_n) = (x_1, x_2, \ldots, x_m).$$

Third property. $d(P_1, P_2)$ should take its minimum value when $P_1 = P_2$ and its maximum value when $P_1 \perp P_2$.

This requires no explanation. It is an immediate consequence of the notion that $d(P_1, P_2)$ should increase as P_1 and P_2 move apart. The next requirement also springs directly from this notion. While it is, in a sense, a very particular property, it is the one which tells us most directly that the coefficient $d(P_1, P_2)$ is measuring what we want to measure.

Fourth property. Let θ be a real parameter and let $\{P_\theta;\ \theta \in (a, b)\}$ be a family of equivalent (mutually absolutely continuous) distributions on the real line such that the family of densities $p_\theta(x)$ with respect to a fixed measure μ has monotone likelihood ratio in x (see Lehmann, 1959, p. 68). Then if $a < \theta_1 < \theta_2 < \theta_3 < b$, we should have

$$d(P_{\theta_1}, P_{\theta_2}) \leqslant d(P_{\theta_1}, P_{\theta_3}).$$

This is a situation where, by any reasonable standard, P_{θ_3} is "further away" from P_{θ_1} than is P_{θ_2}. So the demand we have made is very natural.

These four properties which we have stated are not intended to comprise an exhaustive list. There may well be other properties that would be required of a coefficient of divergence designed for some specific purpose. However, they do provide a framework in which we can study the class of coefficients which we shall now consider.

3. The Relevance of Expectations of Convex Functions of ϕ

It is well known that ϕ is sufficient for (P_1, P_2). Therefore it is reasonable to restrict attention to functions of ϕ when we are considering the divergence of P_2 from P_1, at least if we are thinking of divergence in terms of our ability to distinguish P_2 from P_1 as a result of an "observation". If we wish to consider *numerical coefficients* of divergence then the natural place to look is among the expectations of functions of ϕ. However, we first have to decide on the distribution relative to which those expectations should be taken. A fairly general choice for this distribution is a mixture $Q = \lambda P_1 + (1-\lambda) P_2$ of P_1 and P_2 with $0 \leqslant \lambda \leqslant 1$.

Now there is no real loss of generality in restricting attention to expectations relative to P_1 since

$$E_Q\{h(\phi)\} = \lambda E_1\{h(\phi)\} + (1-\lambda) E_2\{h(\phi)\}$$

$$= E_1\{\lambda h(\phi) + (1-\lambda)\phi h(\phi)\}.$$

Given that, for these or some other reasons, we are going to consider only coefficients which can be expressed in the form $E_1\{g(\phi)\}$ we shall show in Theorem 1 that g must be convex if the corresponding coefficient is to have the second property above. Before doing so, however, we shall deal with one difficulty which we have ignored up to this point. If we allow ϕ to take the value ∞ on a P_1-null set, how are we going to interpret $E_1\{g(\phi)\}$?

In general, by the Lebesgue decomposition theorem, P_2 can be expressed as a sum $Q + S$ of two measures with Q absolutely continuous with respect to P_1 and S singular with respect to P_1. The Radon–Nikodym derivative $\phi = dQ/dP_1$ is defined a.e. P_1. Moreover, there exists a P_1-null set N such that, for every measurable set X,

$$S(X) = S(X \cap N).$$

We shall define ϕ to be $+\infty$ on N and define a "generalized expectation" E^* for functions $g(\phi)$ by

$$E^*\{g(\phi)\} = \int_{\phi<\infty} g(\phi)\,dP_1 + P_2(N) \lim_{\phi\to\infty} \frac{g(\phi)}{\phi}, \tag{3.1}$$

provided that the right-hand side is meaningful, i.e. that $\lim_{\phi\to\infty}\{g(\phi)/\phi\}$ exists and that the stated expression does not take the indeterminate form $\infty-\infty$.

(The motivation behind this somewhat peculiar-looking definition is as follows, this having to be read in inverted commas.

$$E_1\{g(\phi)\} = \int_{\bar{N}} g(\phi)\,dP_1 + \int_N g(\phi)\,dP_1$$

and

$$\int_N g(\phi)\,dP_1 = \int_N \frac{g(\phi)}{\phi}\,dP_2.$$

Since $\phi=\infty$ on N, if we interpret $g(\phi)/\phi$ on N as $\lim_{\phi\to\infty} g(\phi)/\phi$, we are led to the stated definition of E^*.)

We shall adopt the convention that if P_2 is absolutely continuous with respect to P_1, then whatever $\lim_{\phi\to\infty} g(\phi)/\phi$ may be, $P_2(N)\lim\{g(\phi)/\phi\}=0$. With this convention, E^* becomes the ordinary expectation operator E_1 when $P_2 \ll P_1$. Furthermore, as a first indication that this definition of E^* does not lead to inconsistency, we note that $E^*(\phi)=1$ always. Subsequent results will confirm this consistency.

We turn now to a result which confirms the relevance, to the problem in hand, of the generalized expectations of *convex* functions of ϕ, a relevance anticipated to a certain extent by the heuristic argument in Section 1.

Theorem 1. In order that a coefficient of the form $E^*\{g(\phi)\}$ should possess the second property for all P_1 and P_2 on the same sample space, it is necessary that g be a convex function.

Proof. Consider the sample space $\mathscr{X}$ consisting of three elements x_1, x_2 and x_3, the distribution P_1 which assigns probabilities $1-2p$, p and p to these points respectively and the distribution P_2 which assigns probabilities q_1, q_2 and q_3 to them. Let $\mathscr{Y}$ have two elements y_1 and y_2 and let t be the transformation

$$t(x_1)=y_1, \quad t(x_2)=t(x_3)=y_2.$$

If $0<p<\frac{1}{2}$ then we are in the absolutely continuous case and the second property implies

$$(1-2p)\,g\left(\frac{q_1}{1-2p}\right)+pg\left(\frac{q_2}{p}\right)+pg\left(\frac{q_3}{p}\right) \geqslant (1-2p)\,g\left(\frac{q_1}{1-2p}\right)+2pg\left(\frac{q_2+q_3}{2p}\right),$$

that is,

$$g\left(\frac{q_2}{p}\right)+g\left(\frac{q_3}{p}\right) \geqslant 2g\left\{\frac{1}{2}\left(\frac{q_2}{p}+\frac{q_3}{p}\right)\right\},$$

this being true for all p, q_2 and q_3 which satisfy $0<p<\frac{1}{2}$, $0\leqslant q_2, q_3, q_2+q_3 \leqslant 1$.

Now given any two non-negative numbers v_1 and v_2, it is clear that these can always be expressed in the form $v_1=q_2/p$, $v_2=q_3/p$, with p, q_2 and q_3 satisfying the above conditions. Hence the second property implies that, for any non-negative numbers v_1 and v_2, we have

$$g(v_1)+g(v_2) \geqslant 2g\{\tfrac{1}{2}(v_1+v_2)\},$$

that is, that g is convex on $[0,\infty)$.

We are thus led to consider, as potential coefficients of divergence, functionals of pairs of measures on the same sample space which can be expressed in the form $E^*\{C(\phi)\}$, where C is a real continuous convex function on $(0,\infty)$. More generally, if f is a non-decreasing function on $(-\infty,\infty)$ we might consider functionals of the form $f[E^*\{C(\phi)\}]$ since if $E^*\{C(\phi)\}$ has the basic property of increasing as P_1 and P_2 move apart, so also has $f[E^*\{C(\phi)\}]$. We shall now show that coefficients of this form have the properties demanded in Section 2.

4. Properties of Coefficients of the Form $E^*C(\phi)$

4.1. *First Property*

The first property that we demanded of a coefficient of divergence was that it should be defined for all pairs of measures on the same sample space, and we now prove that this is so for *any* coefficient of the form $E^*\{C(\phi)\}$, where C is a continuous convex function on $(0,\infty)$. If $\lim_{\phi\to\infty} C(\phi) = +\infty$ (this limit is either finite or $+\infty$), we define $C(0)$ to be $+\infty$ and allow the value of $+\infty$ for $E^*\{C(\phi)\}$. Then we have to show, according to (3.1), that $\lim_{\phi\to\infty}\{C(\phi)/\phi\}$ exists and that the right-hand side of (3.1) cannot in this case take an indeterminate value. Now since $\{C(\phi)-C(a)\}/(\phi-a)$ is a non-decreasing function of ϕ for $\phi>a$, a being a fixed positive number, it follows that $\lim_{\phi\to\infty}\{C(\phi)/\phi\}$ exists either as a finite number or as $+\infty$.

Moreover, since $C(0)>-\infty$ and since

$$C(\phi) \geqslant \frac{\phi-a}{b-a}\{C(b)-C(a)\}+C(a) \quad \text{for} \quad \phi>b>a,$$

and $\int_{\phi<\infty} \phi \, dP_1 \geqslant 0$, then $\int_{\phi<\infty} C(\phi)\, dP_1$ is either finite or $+\infty$, and it follows that $E^*\{C(\phi)\}$ is defined either as a finite number or $+\infty$, whatever P_1 and P_2 may be.

4.2. *Second Property*

The second property (Section 2) is easily established for coefficients of the form $E^*\{C(\phi)\}$ in the case where P_2 is absolutely continuous with respect to P_1. For in this case, as we have mentioned above, the generalized expectation E^* reduces to the ordinary expectation E_1. Furthermore, $P_2 t^{-1}$ is absolutely continuous with respect to $P_1 t^{-1}$ and it is easily verified that if $\phi(x) = dP_2(x)/dP_1(x)$ and

$$\phi(y) = \frac{dP_2 t^{-1}(y)}{dP_1 t^{-1}(y)},$$

then

$$\phi\{t(x)\} = \phi(y) = E_1\{\phi(x)\,|\,t\}.$$

Also

$$E_1[C\{\phi(x)\}] = E_1(E_1[C\{\phi(x)\}\,|\,t])$$

$$\geqslant E_1[C\{E_1(\phi(x)\,|\,t)\}]$$

by Jensen's inequality,

$$= E_1[C\{\phi(y)\}],$$

giving the required inequality. Note that there is equality here, for all C, if and only if t is sufficient for (P_1, P_2).

In the general case where P_2 may not be absolutely continuous with respct to P_1, we cannot appeal to conditional generalized expectations which have not been defined. However, the result does extend in the following way.

As before, P_2 decomposes into Q_x+S_x with Q_x absolutely continuous, and S_x singular, with respect to P_1. Let N_x be a P_1-null set such that

$$S_x(X) = S_x(X \cap N_x)$$

for every measurable set X, so that $\phi(x) = +\infty$ on N_x and

$$\int_{\phi(x)<\infty} \phi(x)\, dP_1(x) = 1 - S_x(N_x) = 1 - P_2(N_x).$$

Now let t be a measurable transformation from $\mathscr{X}$ onto a measurable space $\mathscr{Y}$, and denote, as usual, induced measures on $\mathscr{Y}$ by $P_1 t^{-1}$, etc.

Since $Q_x \ll P_1$, $Q_x t^{-1} \ll P_1 t^{-1}$. Let $\phi_1(y) = d(Q_x t^{-1})/d(P_1 t^{-1})$.

However, $S_x t^{-1}$ is not necessarily singular with respect to $P_1 t^{-1}$ and in general $S_x t^{-1}$ decomposes into

$$S_x t^{-1} = Q_y + S_y,$$

where $Q_y \ll P_1 t^{-1}$ and $S_y \perp P_1 t^{-1}$. Let $\phi_2(y) = dQ_y/dP_1 t^{-1}$ and let N_y be a $P_1 t^{-1}$-null set such that $S_y(Y) = S_y(Y \cap N_y)$ for every measurable set Y.

We now have a decomposition of $P_2 t^{-1}$,

$$P_2 t^{-1} = (Q_x t^{-1} + Q_y) + S_y$$

in which $Q_x t^{-1} + Q_y \ll P_1 t^{-1}$ and $S_y \perp P_1 t^{-1}$, so that

$$\phi(y) = \begin{cases} \phi_1(y) + \phi_2(y) & \text{a.e. } P_1 t^{-1} \\ \infty & \text{on } N_y. \end{cases}$$

It follows that

$$E^*[C\{\phi(y)\}] = \int_{\phi(y)<\infty} C(\phi_1+\phi_2)\, dP_1 t^{-1} + S_y(N_y) \lim_{\phi\to\infty} \frac{C(\phi)}{\phi}.$$

Now it is easily proved from the convexity of C that

$$C(\phi_1+\phi_2) \leqslant C(\phi_1) + \phi_2 \lim_{\phi\to\infty} \{C(\phi)/\phi\}.$$

Hence

$$\begin{aligned} E^*[C\{\phi(y)\}] &\leqslant \int_{\phi(y)<\infty} C(\phi_1)\, dP_1 t^{-1} + \left\{ \int_{\phi(y)<\infty} \phi_2(y)\, dP_1 t^{-1} + S_y(N_y) \right\} \lim_{\phi\to\infty} \frac{C(\phi)}{\phi} \\ &= \int_{\phi(y)<\infty} C(\phi_1)\, dP_1 t^{-1} + \left\{ \int_{\phi(y)<\infty} dQ_y + S_y(N_y) \right\} \lim_{\phi\to\infty} \frac{C(\phi)}{\phi} \\ &= \int_{\phi(y)<\infty} C(\phi_1)\, dP_1 t^{-1} + P_2(N_x) \lim_{\phi\to\infty} \frac{C(\phi)}{\phi}. \end{aligned}$$

Finally, by applying the argument given at the beginning of this Section to the measure Q_x (which is absolutely continuous with respect to P_1) we have

$$\int_{\phi(y)<\infty} C(\phi_1)\, dP_1\, t^{-1} \leqslant \int_{\phi(x)<\infty} C\{\phi(x)\} dP_1,$$

and this yields the result that

$$E^*[C\{\phi(y)\}] \leqslant E^*[C\{(x)\}].$$

This establishes in general that coefficients of the form $E^*\{C(\phi)\}$ enjoy the second property.

4.3. *Third Property*

When $P_1 = P_2$, $E^*\{C(\phi)\} = C(1)$, since then $\phi \equiv 1$. Now whether or not $P_1 = P_2$, if we consider the trivial mapping t, which maps $\mathscr{X}$ onto a single point, so that $P_1 t^{-1}$ and $P_2 t^{-1}$ coincide, we have from the result proved in 4.2 the fact that

$$E^*\{C(\phi)\} \geqslant C(1).$$

Hence $E^*\{C(\phi)\}$ takes its minimum value when $P_1 = P_2$.

If $P_1 \perp P_2$, then $\phi = 0$ a.e. P_1 and there is a P_1-null set N_x such that

$$P_2(X) = P_2(X \cap N_x) \quad \text{and} \quad P_2(N_x) = 1,$$

so that

$$E^*\{C(\phi)\} = C(0) + \lim_{\phi \to \infty} \{C(\phi)/\phi\}.$$

Now for $\phi \geqslant 0$, $C(\phi) \leqslant C(0) + \phi \lim_{\phi \to \infty} \{C(\phi)/\phi\}$, and from this it follows easily from the definition that in general

$$E^*\{C(\phi)\} \leqslant C(0) + \lim_{\phi \to \infty} \{C(\phi)/\phi\}.$$

Thus we have demonstrated that coefficients of the form $E^*\{C(\phi)\}$ have the third property.

4.4. *Fourth Property*

The fourth property is concerned with a one-parameter family of equivalent distributions on the line whose densities $p_\theta(x)$, $\theta \in \Omega$, relative to a fixed measure μ, have monotone likelihood-ratio in x. To show that the type of coefficient which we are discussing has the fourth property we have to show that, if $\theta_1 < \theta_2 < \theta_3$, then

$$E^*\{C(\phi_2)\} \leqslant E^*\{C(\phi_3)\},$$

where $\phi_i(x) = p_{\theta_1}(x)/p_{\theta_l}(x)$, $i = 2, 3$, and E^* denotes generalized expectation relative to the distribution corresponding to θ_1. Because the distributions concerned are equivalent (we have made this restriction only to avoid somewhat irrelevant detail), E^* reduces in this case to the ordinary expectation operator E_1.

Suppose that the family of distributions has monotone non-decreasing likelihood-ratio. Then $\phi_2(x)$ and $\phi_3(x)$ are non-decreasing functions of x and so also is the ratio $\phi_3(x)/\phi_2(x) = p_{\theta_3}(x)/p_{\theta_2}(x)$. The fact that the ratio is non-decreasing implies that either

(i) $\phi_3(x) < \phi_2(x)$, for all x,

(ii) $\phi_3(x) > \phi_2(x)$, for all x, or

(iii) there exists a number a such that $\phi_3(x) \leqslant \phi_2(x)$ for $x < a$ and $\phi_3(x) \geqslant \phi_2(x)$ for $x > a$.

Then, since $E_1(\phi_3) = E_1(\phi_2) = 1$, neither (i) nor (ii) is possible and therefore (iii) holds. Now it follows from the monotonicity of ϕ_2 and ϕ_3 that if $b < \phi_2(a)$,

$$\{x\colon \phi_2(x) \leqslant b\} \subset \{x\colon \phi_3(x) \leqslant b\},$$

while if $b > \phi_2(a)$,

$$\{x\colon \phi_2(x) \leqslant b\} \supset \{x\colon \phi_3(x) \leqslant b\}.$$

Hence, if F_2 and F_3 are the distribution functions of ϕ_2 and ϕ_3 respectively (the distributions of these random variables being determined by P_{θ_1}), these distribution functions "cross once" in the sense that

$$F_2(\phi) \leqslant F_3(\phi) \quad \text{for} \quad \phi < \phi_2(a)$$

and

$$F_2(\phi) \geqslant F_3(\phi) \quad \text{for} \quad \phi > \phi_2(a).$$

Since $E_1(\phi_2) = E_1(\phi_3)$, this means that, according to any reasonable interpretation of dispersion, the P_{θ_1}-dispersion of ϕ_2 is no more than that of ϕ_3, so that one would anticipate that the expected value of any continuous convex function of ϕ_2 would be no more than the corresponding expected value for ϕ_3. This is in fact true and the essence of a proof is contained in two results proved by Ali and Silvey (1965). The first of these results (Lemma 3) shows that the "crossing-once" of the distribution functions and the fact that $E_1(\phi_2) = E_1(\phi_3)$ imply that, for every k,

$$E_1(|\phi_2 - k|) \leqslant E_1(|\phi_3 - k|),$$

and the second (Theorem 1 of the paper referred to) shows that this in turn implies that, for any continuous convex function C on $(0, \infty)$,

$$E_1\{C(\phi_2)\} \leqslant E_1\{C(\phi_3)\}.$$

The argument for the case of monotone non-increasing likelihood-ratio is obviously similar.

4.5. *A Summarizing Theorem*

We have now shown that any coefficient of the form $E^*\{C(\phi)\}$, where C is a continuous convex function on the positive real numbers, has the four properties demanded in Section 2. It is obvious that if f is an increasing real function of a real variable then a coefficient of the form $f[E^*\{C(\phi)\}]$ also will have these properties. So we can state the following theorem.

Theorem 2. Let C be a continuous convex function on $(0, \infty)$ and let f be an increasing real-valued function of a real variable. Then if P_1 and P_2 are two probability distributions on the same sample space and $\phi(P_1, P_2)$ is the generalized Radon–Nikodym derivative of P_2 with respect to P_1, the coefficient

$$d(P_1, P_2) = f[E^*\{C(\phi)\}]$$

is a possible measure of divergence of P_2 from P_1 in the sense that it enjoys the four properties demanded in Section 2.

It remains to remark that some coefficients of the form indicated will be trivial and so not very useful in practice: for instance, this will be the case if C is a linear function. But it is clear that the class of coefficients of this form is quite large and contains many non-trivial coefficients also. We shall now see that it contains several more or less well-known coefficients.

5. VARIOUS SUGGESTED MEASURES OF DIVERGENCE

Adhikari and Joshi (1956) have discussed, in considerable detail, the properties of various measures of divergence having different interpretations depending on their purpose. We shall now show that most of these can be expressed in the form $f[E^*\{C(\phi)\}]$. In what follows p_1 and p_2 will be the density functions of P_1 and P_2 with respect to a measure λ such that $P_i \ll \lambda$, $i = 1, 2$; such a λ always exists.

(i) $$E^*\{(\phi-1)\log\phi\}.$$

This is Jeffreys' measure of divergence

$$J(1,2) = \int (p_2 - p_1) \log\frac{p_2}{p_1}\, d\lambda.$$

(ii) $$E^*(-\log\phi) \quad \text{and} \quad E^*(\phi\log\phi).$$

These are Kullback's and Leibler's measures of discriminatory information $I(1,2)$ and $I(2,1)$ respectively:

$$I(1,2) = \int p_1 \log\frac{p_1}{p_2}\, d\lambda,$$

$$I(2,1) = \int p_2 \log\frac{p_2}{p_1}\, d\lambda.$$

(iii) $$\tfrac{1}{2}E^*(\sqrt{\phi}-1)^2.$$

This is Kolmogorov's measure of distance, namely,

$$\tfrac{1}{2}\int (\sqrt{p_2} - \sqrt{p_1})^2\, d\lambda.$$

(iv) $$\{E^*(\sqrt{\phi}-1)^2\}^{\frac{1}{2}}.$$

This is Matusita's measure of distance.

(v) $$\tfrac{1}{2}E^*|\phi-1|.$$

This is Kolmogorov's measure of "variation distance",

$$\tfrac{1}{2}\int |f_2 - f_1|\, d\lambda.$$

(vi) $$E^*(-\phi^{1-t}), 0 < t < 1.$$

This is the basis of Chernoff's measure of discriminatory information,

$$-\log \inf_{0<t<1} \int p_1^t p_2^{1-t} d\lambda.$$

It is easily deduced from Theorem 2 that the latter coefficient has all four properties demanded in Section 3. However, it does not appear that this coefficient can be expressed in the form $f[E^*\{C(\phi)\}]$, with f increasing and C convex.

(vii) We conclude this Section by a discussion of two coefficients of divergence arising from the classification problem.

Given a measurement x on an individual which may have come from one of two populations G_1 and G_2 with density functions $p_1(x)$ and $p_2(x)$ respectively, it is required to assign this individual to either G_1 or G_2. A solution of this classification problem consists in partitioning the sample space $\mathscr{X}$ into complementary regions R_1 and R_2 and allocating the individual to G_i if $x \in R_i$. The region R_1 is usually chosen to minimize, in some way, the probability of misclassification. If the distributions of G_1 and G_2 are "far apart", then the probability of misclassification should be small. So one might expect to be able to derive coefficients of divergence from probabilities of misclassification.

We consider first the case where the individual concerned is drawn at random from a mixed population in which G_1 and G_2 are mixed in *known* proportions π_1 and π_2. Then the probability α of misclassification is given by

$$\alpha = \pi_2 \int_{R_1} p_2\,dx + \pi_1 \int_{R_2} p_1\,dx.$$

If R_1 is chosen to minimize α, we have, as is well known,

$$R_1 = \{x: p_2/p_1 < \pi_1/\pi_2\},$$

and consequently

$$R_2 = \{x: p_2/p_1 \geqslant \pi_1/\pi_2\}.$$

In this case it is easy to show that

$$\begin{aligned} 1-\alpha &= \pi_1 + \pi_2 \int_{\phi > \pi_1/\pi_2} \left(\phi - \frac{\pi_1}{\pi_2}\right) dx \\ &= \tfrac{1}{2} + \tfrac{1}{2}\pi_2 E^* |\phi - \pi_1/\pi_2|. \end{aligned}$$

Now one might expect the coefficient $1-2\alpha$ to reflect the distance apart of the distributions of G_1 and G_2 and this is confirmed by the fact that

$$1-2\alpha = \pi_2 E^* |\phi - \pi_1/\pi_2|,$$

which demonstrates that this coefficient has properties 1–4.

Rao (1952) has discussed this problem in the case where π_1 and π_2 are *unknown.* If we demand that the probability of misclassification should be minimized subject to the condition that it be the same for members of G_1 and G_2, then this common value α^* can be expressed in the form

$$\alpha^* = \int_{R_1^*} p_2\,dx,$$

where $R_1^* = \{x: \phi < b\}$, b being chosen to ensure that

$$\int_{R_2^*} p_1\,dx = \int_{R_1^*} p_2\,dx.$$

Rao has proposed $1-\alpha^*$ as a measure of separation of G_2 from G_1.

There is no doubt that this coefficient does measure the divergence of G_2 from G_1. However, it does not appear to be expressible in the form which we are discussing. Of course, it would be over-optimistic to expect that *every* reasonable measure of divergence can be expressed in this form. After all, there are measures of the dispersion of a random variable, such as the quartile deviation, which are not the expectations of convex functions of it. So even if the problem of measuring divergence were equivalent to that of measuring the P_1-dispersion of ϕ (and this we do not claim), we still could not expect all coefficients to be of the form $f[E^*\{C(\phi)\}]$.

6. The Multivariate Normal Distribution

The normal distribution plays such an important part in multivariate theory generally, and in discrimination theory in particular, that it seems appropriate to make one or two remarks in this context about the class of coefficients of divergence which we are considering.

First, we consider the problem of divergence between two n-dimensional normal distributions with different mean vectors but the same variance matrix. Let these be $N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$, $i = 1, 2$. Mahalanobis's generalized distance is α^2, where

$$\alpha^2 = (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)' \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1).$$

α is a metric and a generally accepted measure of distance between the two distributions.

Now every coefficient in the class we are considering is an increasing function of α. This is easily demonstrated by considering the transformation

$$y = (\mathbf{x} - \boldsymbol{\mu}_1)' \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)/\alpha$$

and so reducing the problem to that of the divergence of a $N(\alpha, 1)$ distribution from a $N(0, 1)$. The family $\{N(\alpha, 1): \alpha \geqslant 0\}$ of distributions of y has monotonic increasing likelihood-ratio in y and it follows from Theorem 2 that if f is increasing and C convex then $f[E^*\{C(\phi)\}]$ is an increasing function of α.

Next, we consider two n-dimensional normal distributions with the same mean but different variance matrices. Suppose that these are $N(\boldsymbol{\mu}, \boldsymbol{\Sigma}_i)$, $i = 1, 2$. By making a suitable standard linear transformation, i.e. essentially by renaming the vectors involved, we can reduce the problem of the divergence between these two to that of the problem of the divergence of $N(\mathbf{0}, \boldsymbol{\Lambda})$ from $N(\mathbf{0}, \mathbf{I})$. Here $\mathbf{I}$ is the unit matrix of order n and $\boldsymbol{\Lambda}$ is diag $\{\lambda_1, \lambda_2, \ldots, \lambda_n\}$, the λ's being the roots of $|\boldsymbol{\Sigma}_2 - \lambda \boldsymbol{\Sigma}_1| = 0$.

Now the closer to unity are the λ's the nearer is $N(\mathbf{0}, \boldsymbol{\Lambda})$ to $N(\mathbf{0}, \mathbf{I})$, according to any reasonable interpretation of the proximity of two distributions. So in view of the several results which we have obtained we might anticipate that if $\phi(\mathbf{x})$ is the ratio $p(\mathbf{x}, \boldsymbol{\Lambda})/p(\mathbf{x}, \mathbf{I})$ of the $N(\mathbf{0}, \boldsymbol{\Lambda})$ and $N(\mathbf{0}, \mathbf{I})$ densities and E is the expectation operator relative to $N(\mathbf{0}, \mathbf{I})$, then $E\{C(\phi)\}$ will be an increasing function of each $|1 - \lambda_i|$, when C is convex. This is indeed the case and it can be proved as follows.

We note that

$$\phi(\mathbf{x}) = \prod_{i=1}^{n} \psi(x_i, \lambda_i)$$

where $\psi(\cdot, \lambda)$ is the ratio of the densities of a $N(0, \lambda)$ and a $N(0, 1)$ distribution. Now $\psi(y, \lambda)$ is a non-decreasing (non-increasing) function of y for $\lambda \geqslant 1$ $(\lambda \leqslant 1)$. It follows by the kind of argument used in Section 4.4 that $E\{C(\psi)\}$ is a non-decreasing function of $|1 - \lambda|$ and that in particular $E|\psi - k|$ has this property for every positive k.

From this we can deduce, by using another result of Ali and Silvey (1965, Lemma 4 and subsequently) that $E\{C(\phi)\}$ is a non-decreasing function of each $|1-\lambda_i|$, as anticipated.

These particular results on the normal distribution add a little more weight to the already strong case for the thesis that the problem of measuring the divergence of P_2 from P_1 is closely related to that of measuring the P_1-dispersion of ϕ, a thesis which provides a unifying principle underlying many of the numerous coefficients of divergence suggested in the literature.

Note added in proof. It has been brought to the notice of the authors that the result established in Section 4.2 has been proved (Theorem 1, Cor. 3) in the following paper: CSISZAR, I. (1963). "Eine informationstheoretische Ungleichung und ihre Anwendung auf den Beweis der Ergodizität von Markoffschen Ketten", *Publ. Math. Inst. Hung. Acad. Sc.*, **3**, 85–107.

ACKNOWLEDGEMENT

The authors are indebted to a referee for helpful comments on an earlier version of this paper.

REFERENCES

ADHIKARI, B. P. and JOSHI, D. D. (1956), "Distance, discrimination et résumé exhaustif", *Publ. Inst. Statist. Univ. Paris*, **5**, 57–74.

ALI, S. M. and SILVEY, S. D. (1965), "Association between random variables and the dispersion of a Radon–Nikodym derivative", *J. R. statist. Soc.* B, **27**, 100–107.

CHERNOFF, H. (1952), "A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations", *Ann. math. Statist.*, **23**, 493–507.

KOLMOGOROV, A. N. (1963), "On the approximation of distributions of sums of independent summands by infinitely divisible distributions", *Sankhyā*, **25**, 159–174.

KULLBACK, S. (1959), *Information Theory and Statistics.* New York: Wiley.

LEHMANN, E. L. (1959), *Testing Statistical Hypotheses.* New York: Wiley.

RAO, C. R. (1952), *Advanced Statistical Methods in Biometric Research.* New York: Wiley.