A Reprint from

INFORMATION THEORY

THIRD LONDON SYMPOSIUM

Papers read at a Symposium on 'Information Theory'
held at the Royal Institution, London,
September 12th to 16th 1955

Edited by
COLIN CHERRY

Published by
BUTTERWORTHS SCIENTIFIC PUBLICATIONS
88 KINGSWAY, LONDON, W.C.2

3

ON SOME MEASURES OF INFORMATION USED IN STATISTICS

M. P. SCHÜTZENBERGER
Paris

As is well known, the concepts of information with which statistics and communication theory deal are different, both in their formal expression and in their content. In statistics, when the problem is to estimate the value of an unknown parameter θ through the observation of the state ξ of a physical system, the *a priori* probability $P(\xi \mid \theta)$ depending upon θ, one is led to introduce the expression

$$F = \sum \left(\frac{\partial}{\partial \theta} P(x \mid \theta) \right)^2 \frac{1}{P(x \mid \theta)}$$

with the summation running over all the possible states x of ξ. A whole family of theorems[2, 4, 6] relates F, under appropriate regularity conditions, to a lower bound of the variance of the difference $\theta - \hat{\theta}$ between the true value of θ and its estimated value $\hat{\theta}$.

On the other hand, in communication theory, one is accustomed to evaluate the amount of information on ξ itself by:

$$H = -\sum P(\xi) \log P(\xi)$$

It is remarkable that so much interest has been devoted to this last quantity rather than to the older expression F, which was defined by Sir RONALD FISHER[3] as early as in 1921 and which has been very cursorily dealt with by communication specialists.

Moreover, F and H are not the only measures of the information relative to 'something' contained in an experiment involving *a priori* probability. In the second main problem of statistics—that of deciding on the basis of observation of ξ which of the hypotheses $\theta = \theta_0$ or $\theta = \theta_1$ is true—the following expression

$$W_i = \sum P(x \mid \theta_i) \log \frac{P(x \mid \theta_i)}{P(x \mid \theta_j)} \qquad (i = 0, 1, j = 0, 1, j \neq i)$$

enters in a natural way. WALD[9] has shown that whatever be the procedure used (sequential or not) for the test, the expectation of the number of independent trials needed to reach a given level of security could not be smaller than K/W, where K depends on the probability of error which defines the level of security. In consequence, W could be termed the measure of the information relative to the dilemma $\theta = \theta_0$ or θ_1 afforded by ξ. Indeed, quite close connexions do exist between F, H and W.

After BARTLETT[1], let us consider the modified form

$$H^{\star} = -\Sigma P(x \mid \theta) \log P(x \mid \theta + \varepsilon)$$

and suppose that $\log P(x \mid \theta + \varepsilon)$ may be developed into a series in ascending powers of ε. Then, after some simplification

$$H^{\star} = H + \varepsilon^2 F + \text{terms of higher order in } \varepsilon$$

On the same line of reasoning, if $\theta_0 = \theta_2 + \varepsilon$ and $\theta_1 = \theta_2 - \varepsilon$, where ε is infinitely small, it may be shown that the random variate

$$z(x) = \log \frac{P(x \mid \theta_0)}{P(x \mid \theta_1)}$$

(the expectation of which is W_0, when $\theta = \theta_0$) is distributed with mean $2\varepsilon^2 F$ and variance $4\varepsilon^2 F$, to terms of higher order in ε. More general relations between F and W have been recently studied by KULLBACK[5].

The aim of this present communication (also of reference 8) is to show that these analogies are deeply rooted in the very nature of what we are ready to call a 'measure of information'. As a matter of fact, the leading principle of the axiomatization we shall attempt is more or less a sophistication of WOODWARD's[10] approach to the same problem: that when performing the complete determination of ξ, one may stop at an intermediate level and obtain the total information by adding together: (a) a term corresponding to the information up to this point; (b) a term corresponding to the information from this point on, weighted with the adequate conditional probabilities.

We shall however restrict the postulation of this 'Huygens principle' to those intermediate observations only, which exclude definitely some contingencies, instead of requiring it for all of them, as is the case with Woodward's axiomatic approach which turns out to be unnecessarily exacting. With this weaker form, a purely algebraic treatment is possible giving, besides the 'conventional' H, the expressions F and W, as special cases of the complete solution, which can be explicitly given under some regularity conditions.

For the sake of simplicity, the argument will be split into two parts: the first (Condition I and Theorem I) could be extended to cases other than information, and entails the abstract equivalent of a principle of separation of variates. The second (Condition II and Theorem II) determines the specific character of the information *i.e.* introduces the 'log P' function.

The regularity Conditions III and IV could be presumably weakened by introducing another postulate which is satisfied by H, F and W: the condition that the information be non-negative. Further research would be needed along this line, which we shall only mention here together with the not too difficult possibility of extending Theorem I to non-finite cases, under proper restrictions. A more general theory (using *modular* instead of *distributive* lattices and *idempotent linear operator* instead of *equivalence relations*) may be developed and gives an axiomatic definition of the variance (as corresponding to H and *not* to F) and of the so-called 'chi-square' measure of discrepancy (as corresponding to W) [see reference 11].

First of all, let us make it clear that we are not looking for a measure of the information provided by one given result of the observations, but for a

measure of what amount may be obtained on the average with the help of a given observational set-up. For representing the general situation, we shall consider a physical system whose state ξ is still unknown. For the sake of simplicity it will be supposed, throughout the paper, that ξ can take on only one of a finite set E of values $x, y, z \ldots$. In practice, even if ξ were a continuous variate, this quantification could always be assumed, since any real measurement may be done with a finite precision only.

The *a priori* probabilities with which ξ may be any of the states $x, y \ldots$ will be written $P(x), P(y) \ldots$ and we shall suppose that they are functions of some unknown parameter(s) symbolized by θ. With respect to the physical system, an observer Ω_i is characterized by the degree of accuracy with which he is able to recognize ξ. For instance, if ξ is a numerical variate with possible values 0, 1, 2 or 3, an observer Ω_0 may be unable to know more about ξ than to ascertain whether it is zero or not; another one, Ω_2, whether it is odd or even *etc.* Accordingly, to each observed Ω_i corresponds an equivalence relation ρ_i between the possible states of ξ; that is to say, a partition of E into disjoint subsets $(X), (Y) \ldots (Z)$, Ω_i being unable to 'separate' two states when they pertain to the same 'class of equivalence' of ρ_i.

Between equivalence relations on E exists the usual partial ordering relation $\rho' < \rho$ (ρ' is finer than ρ: every class of ρ' is contained in a class of ρ), which means that the observer Ω' is able to perform every distinction between states which Ω can do. Further, if $\rho' < \rho$, and if X is a class of ρ, we shall denote by $\rho'[X]$ (the '*restriction* of ρ' to X') the equivalence relation induced by ρ' on the subset X of E. If, for some subset X, $\rho'[X] = \rho''[X]$ we shall write: $\rho' \equiv \rho[X]$ ('ρ and ρ' are identical on X').

With these notions at hand we may now compare observers, or rather pairs of observers.

Definition: The two pairs of equivalence relations (ρ_i, ρ_j) and (ρ_k, ρ_e), where $\rho_i < \rho_j$ and $\rho_k < \rho_e$, will be said to be in the relation $\sim$ if, and only if, there exists a partition of E into disjoint subsets E' and E'' such that

$$\rho_i \equiv \rho_k[E']; \quad \rho_j \equiv \rho_e[E']; \quad \rho_i \equiv \rho_j[E'']; \quad \rho_k \equiv \rho_e[E'']$$

For instance, let $E = \{a, b, c, d, e, f, g\}$ and:

$$\rho_i = (abc)(d)(e)(fg); \quad \rho_j = (abc)(d)(efg); \quad \rho_k = (ab)(cd)(e)(fg);$$
$$\rho_e = (ab)(cd)(efg)$$

One sees that $(\rho_i, \rho_j) \sim (\rho_k, \rho_e)$, by taking $E' = (efg)$ and $E'' = (abcd)$, for then:

$$\rho_i[E'] = \rho_k[E'] = (e)(fg); \qquad \rho_j[E'] = \rho_e[E'] = (efg)$$
$$\rho_i = [E''] = \rho_j[E''] = (abc)(d); \quad \rho_k[E''] = \rho_e[E''] = (ab)(cd)$$

It is readily demonstrated that $\sim$ is again an equivalence relation on the set of all ordered pairs of equivalence relations. For that it is enough to remark that from ρ_i and ρ_j (with $\rho_i < \rho_j$) E'' is unequivocally determined as the union of those classes which are in the same time classes of ρ_i and of ρ_j. [Parenthetically, let us observe that if, in addition, $\rho_j < \rho_e$ one has too, $\rho_i < \rho_k$ and $(\rho_i, \rho_k) \sim (\rho_j, \rho_e)$ so that the $\sim$ relation looks quite like an abstract version of the equality relation between two fractions.]

Consider now, for a given finite E, the set R of all the equivalence relations on E, and a function $f(\ )$ of R in to some additive group **a**.

Definition: The function $f(\ )$ will be called a valuation on R if for every quadruple of relations $(\rho_i, \rho_j) \sim (\rho_k, \rho_e)$ entails $f(\rho_i) - f(\rho_j) = f(\rho_k) - f(\rho_e)$.

Theorem I: Any valuation $f(\ )$ of the set of all equivalence relations on the finite E may be written in the form $f(\rho) = \Sigma g(X)$ where the summation extends to all classes X of ρ.

Proof: Consider for any equivalence relation into k classes $\rho = (X)(Y) \ldots (T)$ the three other relations: ρ_X, the finest among the equivalence relations admitting the class (X), $\rho_{\bar{X}}$ the finest among all equivalence relations admitting the class (Y), $(Z) \ldots (T)$, and ρ_0 the finest among all the equivalence relations; one has

$$(\rho_0, \rho_X) \sim (\rho_{\bar{X}}, \rho) \textit{ i.e. if } f(\ ) \text{ is a valuation: } f(\rho) = f(\rho_X) + f(\rho_{\bar{X}}) - f(\rho_0)$$

Thus if the theorem is proved for all the equivalence relations with no more than $k - 1$ classes, it is proved for the relations with k classes, since

$$f(\rho) = \{\sum_{x \epsilon X} g(x)\} + \{g(x) + \sum_{y \epsilon E - X} g(g)\} - \{\sum_{x \epsilon E} g(x)\}$$

Now, choose for any elements of E, an arbitrary value of $g(x)$ with the sole condition that $\sum_{x \epsilon E} g(x) = f(\rho_0)$. For any X we define $g(X) = f(\rho_X) - \sum_{x \epsilon X} f(x) - f(\rho_0)$ which achieves the proof.

Remark: Suppose now that we discover that ρ_0 was *not* really the finest equivalence relation, but that there exists a still finer one, ρ'_0 differing from it only by the fact that in ρ'_0 the class (X) is split into (X') and (X''). Every equivalence relation ρ'_i, less fine than ρ_0 either is $> \rho_0$ or is of the form (X'), (X''), (U), $(V) \ldots (W)$—*i.e.* is identical on $E - X$ to some $\rho_i > \rho_0$.

Thus in this case $(\rho'_0, \rho'_i) \sim (\rho_0, \rho_i)$

and
$$f(\rho'_i) = f(\rho_i) + f(\rho'_0) - f(\rho_0)$$

and $f(\rho')$ will be of the form needed if, having chosen arbitrarily $g(X')$, we take $g(X'') = g(X) - g(X') + f(\rho'_0) - f(\rho_0)$.

Suppose now that $\rho' < \rho$ entails $f(\rho) \leqslant f(\rho') \leqslant f < \infty$ and that we find some finite ρ_0 such that $f - f(\rho_0) \leqslant \varepsilon$ and for all classes Y of $\rho_0 = 0 < |g(Y)| < \eta$. Then the same condition may be made to hold for ρ'_0; for instance by taking

$$0 < g(X'') = g(X') = \tfrac{1}{2}\{f(\rho'_0) - f(\rho_0) + g(X)\} < \varepsilon + \eta/2$$

This remark would lead to the possibility of extending Theorem I to general E, under a proper definition of what is meant by an equivalence relation with infinitely many classes and corresponding restrictions on f.

Reverting to our main purpose—which is to measure information—we shall postulate that:

Condition I: The measure of the information $H(\rho_i)$ attached to the observer is a valuation on the set (lattice) R of all equivalence relation on E.

What this means exactly seems a fair enough requirement. If Ω_i differs from Ω_j by the same ability in finer distinctions as Ω_k does from Ω_e, we ask that the differences $H(\rho_i) - H(\rho_j)$ and $H(\rho_k) - H(\rho_e)$ be equal.

If one prefers, this condition may be interpreted in an equivalent way by assimilating H to a cost of equipment, and requiring that the expenses, involved in the addition of a special gadget, intended to perform some finer analysis, be independent from the total cost of the other parts of the observational machinery.

From *Theorem I*, we know only that if the number of states is finite, this implies that $H(\rho)$ be a sum of terms depending only on the classes of ρ. Of course, so broad a definition is not enough to determine H in a really interesting way, and we shall further postulate:

Condition II: If $\rho < \rho'$ in such a way that for a class U of ρ', $\rho \equiv \rho'[E-U]$ (*i.e.* if Ω differs from Ω' only by a further splitting of one of the classes), then

$$H(\rho) = H(\rho') + P(U)H(\rho'')$$

where $H(\rho'')$ is the measure of information attached to $\rho'' = \rho[U]$ for an observer knowing that the state ξ pertains to the subset U. To this condition we add:

Condition III: The probabilities $P(X) = x \neq 0$ are elements of some topological communitative ring* **a** and $H(\rho)$ a continuous functional in the $x, y \ldots$.

Condition IV: **a** is such, that if for all $a, b \in \mathbf{a}$, $a + b = 1$, $ab \neq 0$, h_i is a continuous functional satisfying

(1) $h_1(a + b) = h_1(a) + h_1(b)$ then $h_1(x) = \Delta(x)$

(2) $h_2(ab) = h_2(a) + h_2(b)$ then $h_2(x) = \Delta \log x$

where Δ is a semi-linear functional and log is some fixed function within the ring of functions of **a**.

Theorem II: Under IV, the necessary and sufficient condition for $H(\rho)$ to satisfy I, II, III, is that it has the form:

$$H(\rho) = \sum_x x \, \Delta \log x$$

where the summation runs over classes X of ρ, and Δ is any continuous semi-linear functional.

Proof: The sufficiency is a matter of straightforward verification. As to the necessity, Condition I implies that:

$$H(\rho) = \Sigma g'(x) = \Sigma \, x \, . \, g(x) \text{ for some } g(X) = g'(x)/x.$$

Consider an equivalence relation with four classes: $\rho = (X)(Y)(Z)(T)$. Condition II with $U = Y + Z + T$ implies:

$$\begin{aligned} H(\rho) &= xg(x) + yg(y) + zg(z) + tg(t) \\ &= xg(x) + (y + z + t)g(y + g + t) \\ &\quad + (y + z + t)\left\{\frac{y}{y + z + t}g\left(\frac{y}{y + z + t}\right)\right. \\ &\quad \left. + \frac{z}{y + z + t}g\left(\frac{z}{y + z + t}\right) + \frac{t}{y + z + t}g\left(\frac{t}{y + z + t}\right)\right\} \end{aligned}$$

* *i.e.* very roughly speaking, a set in which abstract operations $+$ and $\times$ are defined, so as to satisfy the usual conditions except that it may happen that $ab = 0$ even when $a \neq 0$ and $b \neq 0$.

Let us transform this by writing: $1 - x = y + z + t$; $y' = y(1 - x)^{-1}$; $z' = z(1 - x)^{-1}$; $t' = t(1 - x)^{-1}$ and $k(a; b) = g(ab) - g(a) - g(b)$. Then after regrouping the terms we get

$$yk(y'; 1 - x) + zk(z'; 1 - x) + tk(t'; 1 - x) = 0$$

in particular, if t were zero, x and y keeping the same values, one would have

$$yk(y'; 1 - x) + (z + t)k(z' + t'; 1 - x) = 0$$

Then, by subtraction of the last two equations, one obtains:

$$(z' + t')k(z' + t'; 1 - x) = z'k(z'; 1 - x) + t'k(t'; 1 - x)$$

Hence after IV (1), since $ak(a; b)$ is additive in its first argument: $ak(a; b) = \Delta(a)$ where Δ may depend on b, but not on a. But $k(a, b)$ is symmetrical in a and b; this, in turn, implies that $k(a; b) = 1/ab\ \Delta'(ab)$ and the above equation gives $\Delta'(y'(1 - x)) + \Delta'((z' + t')(1 - x)) = \Delta'(y) + \Delta'(z + t) = 0$ where y and $z + t$ are restricted only by $y + z + t \leqslant 1$. Since Δ' is additive, this means that $\Delta'(u) = 0$ for all $u(0 < u < 1)$ so that one has $h(a; b) = g(ab) - g(a) - g(b) = 0$ for all $a, b \in \mathbf{a}$; $a + b < 1$; $ab \neq 0$. Now, by Condition IV (2) this gives $g(a) = \Delta \log a$ which achieves the proof of the theorem.

Remark (1): The two successive steps *I* and *II* involved in the above axiomatic could be replaced by the single postulate *I'* (under Conditions III and IV and finite *E*).

I': For all $\rho = (X)(Y)(Z)$; $\rho_1' = (X)(Y+Z)$; $\rho_1'' = \rho\,[E - X] = (Y)(Z)$ $\rho_2' = (X + Y)(Z)$; $\rho_2'' = (X)(Y)$ one has:

$$H(\rho_1') + P(X)H(\rho_1'') = H(\rho_2') + P(Z)H(\rho_2'')$$

Indeed, it can be shown by recurrence that *I'* implies *I*. This formulation which does not require the concept of valuation may be interpreted as a principle of 'virtual decomposition of the observations into successive dichotomies', since it requires that the information attached to the distinction of ξ between X, Y, Z may be computed on the sole basis of the information attached to the dichotomies ρ' and ρ''.

Remark (2): When one confines oneself to information depending only on the numerical values of the $P(X)$, as is the case in communication theory, the Condition II may be replaced by additivity for the composition of independent variates (Woodward). Then, if a valuation is continuous and depends only on the $P(X)$, a necessary and sufficient condition for it to be additive *for the composition of independent variates is that it should have the form*: $\Sigma x \log x$.

Proof: Let η and ζ be two independent variates, taking respectively the states $Y_1Y_2Y_3$ and $Z_1Z_2Z_3$. Let $\xi = \eta \times \zeta$ be their abstract product, taking the 3×3 states X_{ij} $(i, j = 1, 2, 3)$. Let η^x (respectively $\zeta^\star$) be the variates obtained from η (respectively ζ) by confounding the states numbered 2 and 3. Since H is assumed to be a valuation, the information $H(\eta \times \zeta)$ on ξ is a sum: $H(\eta \times \zeta) = \sum_{ij} x_{ij}g(x_{ij})$ which by hypothesis is equal to

$$H(y) + H(\zeta) = \sum_i y_i g(y_i) + \sum_j z_j g(z_j)$$

One has

$$\begin{aligned} D &= H(\eta \times \zeta) - H(\eta \times \zeta^\star) - H(\eta^\star \times \zeta) + H(\eta^\star \times \zeta^\star) \\ &= (H(\eta) + H(\zeta)) - (H(\eta) + H(\zeta^\star)) - (H(\eta^\star) + H(\zeta)) \\ &\quad + (H(\eta^\star) + H(\zeta^\star)) \\ &= 0 \text{ identically} \end{aligned}$$

and, on the other hand, supposing that $z_2 = z_3 = z$ and writing $g(ab) - g(2ab) = k(a, b)$:

$$D = y_2 k(y_2, z) + y_3 k(y_3, z) - (y_2 + y_3) k(y_2 + y_3;\ z) = 0$$

which implies, since continuity is postulated, that $k(y_i, z)$ be independent from y_i.

But $k(y_i, z)$, again, is symmetrical in y_i and z, so that it is a constant K, and we obtain finally, letting $u = y_i$, z

$$g(u) - g(2u) = K \quad \text{for all} \quad 0 \leqslant u \leqslant 1$$

The theorem follows, since this is Schröder's equation, which is known[7] to have as its only solutions $g(u) = K \log u$. Observe that the proof would have failed if we had not assumed that g and k are numerical functions, for we could not have proved that $k(a, b)$ is a constant. Indeed, in the more general case, not only information, but the results of applying any linear operation to them (*i.e.* expressions of the form $\Sigma\, \Delta_1 P(X) \Delta_2 \log P(X)$) satisfy the requirements of additivity for the composition of independent variates.

REFERENCES

[1] BARTLETT, M. S. In *Proc. London Symp. Information Theory*, p. 81, London; Ministry of Supply, 1950

[2] DARMOIS, G. *Rev. Inst. Int. Stat.* (1945) 132

[3] FISHER, R. A. *Phil. Trans. R. Soc.*, A. 22 (1921) 309

[4] FRECHET, M. *Rev. Inst. Int. Stat.*, 3/4 (1942) 182

[5] KULLBACK, S. *Ann. Math. Stat.*, 2 (1954) 745

[6] RAO, G. Q. *Bull. Calcutta Math. Soc.*, 37 (1945) 81

[7] SCHRÖDER, K. *Math. Ann.*, 3 (1871) 296

[8] SCHÜTZENBERGER, M. P. *Pub. Inst. Stat. Univ. Paris*, 3 (1953) 27

[9] WALD, A. *Sequential Analysis*, N.Y.; J. Wiley and Son, 1947

[10] WOODWARD, P. M. and DAVIES, I. L. *Proc. Institute of Electrical Engineers*, III, 99 (1952) 37

[11] For further details, see the forthcoming book *Decision et Informations* by B. Mandelbrot and M. P. Schützenberger.

DISCUSSION

J. C. R. LICKLIDER: As I understand it, Shannon's measure H, would be 'just another measure' if it did not lead to the Channel-Capacity Theorem. The fact that H leads to that remarkable insight gives H a definite status. In problems concerning coding of information for efficient transmission through restricted channels H is the natural measure.

If it is true that measures are noteworthy insofar as they lead to new orderings and relations of facts, then the question arises: do the other measures you have discussed lead to discoveries comparable with the Channel-Capacity Theorem?

DISCUSSION

R. SYSKI: The author has formulated his results in terms of lattice theory, and postulated that the amount of information $H(\rho)$ is the valuation on the modular lattice R of all equivalence relations on the set E.

I think that a formulation in terms of measure theory could be possible. In fact, some particular cases to which the author refers admit such a formulation. Shannon's entropy is defined as an integral of a certain measurable function, taken with respect to the probability measure, over the measure space with appropriate Borel field. Similarly, the concept of sufficient statistic was discussed by Halmos with the help of the Radon-Nikodym theorem (*Ann. Math. Stat.*, 20 (1949) 225.) Since the measure on sets is also the valuation on lattices (Boolean algebra), the author's approach and the measure theory approach are closely related. I should like to ask, therefore, what are the advantages in using lattice theory here?

Secondly, Dr. McMillan recently used metric informational lattices (*Bull. Amer. Math. Soc.*, 60 (1954) 558). Is his treatment related to that of the author's as far as the selective information is concerned?

M. P. SCHÜTZENBERGER in reply: I would answer Dr. Licklider: partly, yes. The Frechet-Darmois-Cramer- Rao and the Wald-Wolfowitz theorems are the counterpart of the Channel-Capacity Theorem; the first applies when the signal is of a continuous nature and the loss function is quadratic and provides a basis for Tuller's inequality; the second when a fixed signal has to be detected with as few elementary observations as possible. Both give an upper limit to the efficiency of a given transmission set-up. This may be shown under proper restrictions to be attained asymptotically when a long enough delay is allowed.

In reply to Mr. Syski, measure theory could be used as well. The point here is that the lattice under consideration is not Boolean nor modular so that the present approach (where the aim is quite different from McMillan's) seems to me to be more direct.