78. Information and Statistics. I

By Yukiyosi KAWADA
Department of Mathematics, Faculty of Science,
University of Tokyo

(Communicated by Shokichi IYANAGA, M. J. A., Sept. 14, 1987)

This short note is a summary of an axiomatic consideration of an information and its applications to statistics.[1)]

I. Axioms of an information **1.** Under an information we usually understand the *Kullback-Leibler information* (1951):

$$(1) \qquad I_{KL}(\boldsymbol{p},\boldsymbol{q})=\sum_{k=1}^{m} p_k \log (p_k/q_k)$$

where $\boldsymbol{p}=(p_1,\cdots,p_m)$, $\boldsymbol{q}=(q_1,\cdots,q_m)$ are two finite probability distributions. There are, however, similar known functions $I(\boldsymbol{p},\boldsymbol{q})$ which may be also called informations. For example,

$$(2) \qquad I_P(\boldsymbol{p},\boldsymbol{q})=\left(\sum_{k=1}^{m} p_k^2 q_k^{-1}\right)-1 \quad (\textit{Pearson's information}, 1900)$$

which can be expressed as $(\sum_{k=1}^m (n_k-nq_k)^2/nq_k)/n$ when $\boldsymbol{p}=(n_1/n,\cdots,n_m/n)$ $(n=n_1+\cdots+n_m)$, and

$$(3) \qquad I_K(\boldsymbol{p},\boldsymbol{q})=2\left(1-\sum_{k=1}^{m} p_k^{1/2}q_k^{1/2}\right) \quad (\textit{Kakutani's information}, [5], 1948).$$

These are included as special cases of the family

$$(4) \qquad I^{\lambda}(\boldsymbol{p},\boldsymbol{q})=\frac{1}{\lambda}\left\{\left(\sum_{k=1}^{m} p_k^{1+\lambda}q_k^{-\lambda}\right)-1\right\}, \quad -\frac{1}{2}\leqq\lambda<\infty, \quad \lambda\neq 0,$$

namely, $I_P=I^1$, $I_K=I^{-1/2}$, and we define $I^0=I_{KL}$.

We can easily see that

$$I^{\lambda}(\boldsymbol{p},\boldsymbol{q})\leqq I^{\mu}(\boldsymbol{p},\boldsymbol{q}) \qquad \text{for } \lambda<\mu$$

and

$$\lim_{n\to\infty} I^{\lambda_n}(\boldsymbol{p},\boldsymbol{q})=I^{\lambda_0}(\boldsymbol{p},\boldsymbol{q}) \qquad \text{for } \lim_{n\to\infty}\lambda_n=\lambda_0.$$

We call I^0 the *parabolic* information, I^{λ} $(\lambda>0)$ a *hyperbolic* information and $I^{-\mu}(1/2\geqq\mu>0)$ an *elliptic* information.

Remark. (i) The function $I^{-\mu}(\boldsymbol{p},\boldsymbol{q})$ for $1/2\geqq\mu>0$ was introduced by several authors [4], [7] and the general case was also considered in [9].

(ii) In the definition (4) we can extend the value λ for $\lambda<-1/2$ formally. Then we have

$$I^{-\lambda}(\boldsymbol{p},\boldsymbol{q})=-\frac{\lambda-1}{\lambda}I^{\lambda-1}(\boldsymbol{q},\boldsymbol{p}), \quad \lambda>1$$

$$I^{-1}(\boldsymbol{p},\boldsymbol{q})=0$$

$$I^{-\mu}(\boldsymbol{p},\boldsymbol{q})=\frac{1-\mu}{\mu}I^{\mu-1}(\boldsymbol{q},\boldsymbol{p}), \quad 1/2<\mu<1.$$

1) The details will be published in the Proceedings of the Institute of Statistical Mathematics (Tōkei Sūri) in Japanese.

Theorem 1. *The function* $I^\lambda(\boldsymbol{p},\boldsymbol{q})=I^\lambda(p_1,\cdots,p_m;q_1,\cdots,q_m)$ $(-1/2\leqq\lambda<\infty)$ $(p_k\geqq 0,\ q_k\geqq 0,\ p_1+\cdots+p_m=q_1+\cdots+q_m=1,\ m=1,2,\cdots)$ *satisfies the following conditions*:

(I) *Reducibility. If* $p_m=q_m=0$, *then*

$$I^\lambda(p_1,\cdots,p_m;q_1,\cdots,q_m)=I^\lambda(p_1,\cdots,p_{m-1};q_1,\cdots,q_{m-1}).$$

(II) *Symmetry.*

$$I^\lambda(p_1,\cdots,p_m;q_1,\cdots,q_m)=I^\lambda(p_{i_1},\cdots,p_{i_m};q_{i_1},\cdots,q_{i_m})$$

for any substitution $(i_1,\cdots,i_m)$ *of* $(1,\cdots,m)$.

(III) *Non-negativity.*

$$I^\lambda(\boldsymbol{p},\boldsymbol{q})\geqq 0$$

for any $\boldsymbol{p},\boldsymbol{q}$, *and the equality holds if and only if* $\boldsymbol{p}=\boldsymbol{q}$.

(IV) *Convexity and* (V) *Invariance.*

$$I^\lambda(p_1+p_2,p_3,\cdots,p_m;q_1+q_2,q_3,\cdots,q_m)$$
$$\leqq I^\lambda(p_1,p_2,\cdots,p_m;q_1,q_2,\cdots,q_m)$$

holds in general and the equality holds if and only if $q_1/p_1=q_2/p_2$.

(VI) *Additivity and pseudo-additivity. Let* $\boldsymbol{p}\otimes\boldsymbol{p}'$, $\boldsymbol{q}\otimes\boldsymbol{q}'$ *be the direct product distributions. If* $\lambda=0$ *the additivity holds*:

$$I^0(\boldsymbol{p}\otimes\boldsymbol{p}',\ \boldsymbol{q}\otimes\boldsymbol{q}')=I^0(\boldsymbol{p},\boldsymbol{q})+I^0(\boldsymbol{p}',\boldsymbol{q}').$$

In general the pseudo-additivity holds:

$$I^\lambda(\boldsymbol{p}\otimes\boldsymbol{p}',\ \boldsymbol{q}\otimes\boldsymbol{q}')=I^\lambda(\boldsymbol{p},\boldsymbol{q})+I^\lambda(\boldsymbol{p}',\boldsymbol{q}')+\lambda I^\lambda(\boldsymbol{p},\boldsymbol{q})\cdot I^\lambda(\boldsymbol{p}',\boldsymbol{q}').$$

(VII) *Continuity. If* $\lim_{n\to\infty}\boldsymbol{p}_n=\boldsymbol{p}_0$, *and* $\lim_{n\to\infty}\boldsymbol{q}_n=\boldsymbol{q}_0$, *then*

$$\lim_{n\to\infty} I^\lambda(\boldsymbol{p}_n,\boldsymbol{q}_n)=I^\lambda(\boldsymbol{p}_0,\boldsymbol{q}_0).$$

(VIII) *Relativity. Let* $\boldsymbol{p}^*=(p_{kj})$, $\boldsymbol{q}^*=(q_{kj})$ $(k=1,\cdots,m;\ j=1,\cdots,r_k)$ *be probability distributions. Put* $p_k=\sum_{j=1}^{r_k}p_{kj}$, $q_k=\sum_{j=1}^{r_k}q_{kj}$ *and* $\boldsymbol{p}=(p_k)$, $\boldsymbol{q}=(q_k)$. *Define the conditional probability*: $\boldsymbol{p}^{(k)}=(p_{kj}/p_k)$, $\boldsymbol{q}^{(k)}=(q_{kj}/q_k)$ $(j=1,\cdots,r_k)$ *for* $k=1,\cdots,m$. *Then*

$$I^\lambda(\boldsymbol{p}^*,\boldsymbol{q}^*)=I^\lambda(\boldsymbol{p},\boldsymbol{q})+\sum_{k=1}^{m}p_k^{1+\lambda}q_k^{-\lambda}I^\lambda(\boldsymbol{p}^{(k)},\ \boldsymbol{q}^{(k)}).$$

For $\lambda=0$ *these properties are proved in Kullback* [8].

Remark. We see easily that (VIII) implies (VI), and (III) and (VIII) imply (IV) and (V).

Theorem 2. *Let us fix a constant* λ $(-1/2\leqq\lambda<\infty)$ *and assume that a function* I $(p_1,\cdots,p_m;q_1,\cdots,q_m)$ *satisfies* (I) *reducibility,* (II) *symmetry,* (III) *non-negativity,* (VII) *continuity, and* (VIII) *relativity. Then*

$$I(\boldsymbol{p},\boldsymbol{q})=cI^\lambda(\boldsymbol{p},\boldsymbol{q})$$

for some constant $c>0$.

2. Now we shall consider about "information" which we define by the following system of axioms.

Definition 1. Let $\boldsymbol{p}=(p_1,\cdots,p_m)$, $\boldsymbol{q}=(q_1,\cdots,q_m)$ $(\sum_{k=1}^m p_k=\sum_{k=1}^m q_k=1)$ be finite probability distributions $(m=1,2,\cdots)$. A function $I(\boldsymbol{p},\boldsymbol{q})=I(p_1,\cdots,p_m;q_1,\cdots,q_m)$ is called an *information* if the function I satisfies the axioms: (I) *reducibility,* (II) *symmetry,* (III) *non-negativity,* (IV) *convexity, and* (V) *invariance.*

The functions $I^\lambda(\boldsymbol{p},\boldsymbol{q})$ $(-1/2\leqq\lambda<\infty)$ are informations in the above sense. We can also define an information in the form $I(\boldsymbol{p},\boldsymbol{q})=F(I_1(\boldsymbol{p},\boldsymbol{q}),\cdots,I_r(\boldsymbol{p},\boldsymbol{q}))$ by using a suitable function $F(x_1,\cdots,x_r)$ from known informations $I_1,\cdots,I_r$. For example, $I=aI_1+bI_2$, $I=aI_1^2+bI_2^2$, $a>0$, $b>0$, etc. In particular,

$$(5)_1 \qquad \tilde{I}^{\lambda}(\boldsymbol{p},\boldsymbol{q})=\frac{1}{\lambda}\log(1+\lambda I^{\lambda}(\boldsymbol{p},\boldsymbol{q})),\quad \lambda>0,$$

$$(5)_2 \qquad \tilde{I}^{-\mu}(\boldsymbol{p},\boldsymbol{q})=\frac{-1}{\mu}\log(1-\mu I^{-\mu}(\boldsymbol{p},\boldsymbol{q})),\quad 0<\mu<\frac{1}{2}$$

are also informations which satisfy the additivity.

Remark. (i) $\tilde{I}^{-\mu}$ was introduced by Kudō [7] (1953) and $\tilde{I}^{\lambda}$ and $\tilde{I}^{-\mu}$ were also considered by Rényi [10] (1961).

(ii) The functions $d(\boldsymbol{p},\boldsymbol{q})=(\sum_{k=1}^m|p_k-q_k|)/\sqrt{2}$ and $D(\boldsymbol{p},\boldsymbol{q})=(\sum_{k=1}^m(p_k-q_k)^2)^{1/2}$ are not informations in the above sense.

3. **Definition 2.** A continuous information is called *fundamental* if $I(\boldsymbol{p},\boldsymbol{q})$ can be expressed as

$$(6) \qquad I(\boldsymbol{p},\boldsymbol{q})=L(p_1,q_1)+\cdots+L(p_m,q_m)$$

by a continuous function $L(x,y)$ defined for $0\leqq x\leqq 1$ and $0\leqq y\leqq 1$.

For example, $I_{KL}, I_P, I_K, I^{\lambda}$ are fundamental, but $\tilde{I}^{\lambda}, \tilde{I}^{-\mu}$ are not fundamental.

Theorem 3. *In order that a function $I(\boldsymbol{p},\boldsymbol{q})$ defined by (6) is an information it is necessary and sufficient that*

(i) $L(0,0)=0$, $L(1,1)=0$,

(ii) *if* $p_1/q_1=p_2/q_2=(p_1+p_2)/(q_1+q_2)$ $(p_1+p_2\leqq 1,\ q_1+q_2\leqq 1)$, *then*
$$L(p_1,q_1)+L(p_2,q_2)=L(p_1+p_2,\ q_1+q_2),$$

(iii) *if* $p_1+p_2\leqq 1$, $q_1+q_2\leqq 1$, *then*
$$L(p_1,q_1)+L(p_2,q_2)\geqq L(p_1+p_2,\ q_1+q_2)$$
and the equality holds if and only if $p_1/q_1=p_2/q_2$.

Theorem 4. *A fundamental information $I(\boldsymbol{p},\boldsymbol{q})$ can be expressed in the form*

$$(7) \qquad I(\boldsymbol{p},\boldsymbol{q})=\sum_{k=1}^m p_kK(q_k/p_k)$$

by a non-negative strictly convex function $K(x)$ with $K(1)=0$, and conversely the function defined by (7) is a fundamental information. If we assume furthermore the differentiability of $K(x)$, such function $K(x)$ is uniquely determined by I.

Examples.

$$I^0(\boldsymbol{p},\boldsymbol{q})=\sum_{k=1}^m p_kK^0(q_k/p_k),\qquad K^0(x)=-\log x+(x-1)\geqq 0.$$

$$I^{\lambda}(\boldsymbol{p},\boldsymbol{q})=\sum_{k=1}^m p_kK^{\lambda}(q_k/p_k),\quad \lambda\neq 0,\quad K^{\lambda}(x)=(x^{-\lambda}-1)/\lambda+(x-1)\geqq 0.$$

Theorem 5. *Let $I(\boldsymbol{p},\boldsymbol{q})$ be a differentiable fundamental information (i.e. $L(x,y)$ in (6) be three times differentiable in x and y).*

(i) *If $I(\boldsymbol{p},\boldsymbol{q})$ satisfies the additivity:*
$$I(\boldsymbol{p}\otimes\boldsymbol{p}',\ \boldsymbol{q}\otimes\boldsymbol{q}')=I(\boldsymbol{p},\boldsymbol{q})+I(\boldsymbol{p}',\boldsymbol{q}'),$$

then we have

$$I(\boldsymbol{p},\boldsymbol{q})=c_1 I^0(\boldsymbol{p},\boldsymbol{q})+c_2 I^0(\boldsymbol{q},\boldsymbol{p}),\quad c_1\geqq 0,\quad c_2\geqq 0.$$

(ii) *If* $I(\boldsymbol{p},\boldsymbol{q})$ *satisfies the relation*

$$I(\boldsymbol{p}\otimes\boldsymbol{p}',\ \boldsymbol{q}\otimes\boldsymbol{q}')=I(\boldsymbol{p},\boldsymbol{q})+I(\boldsymbol{p}',\boldsymbol{q}')+I(\boldsymbol{p},\boldsymbol{q})\cdot I(\boldsymbol{p}',\boldsymbol{q}')$$

then we have

$$I(\boldsymbol{p},\boldsymbol{q})=\lambda I^{\lambda}(\boldsymbol{p},\boldsymbol{q}),\quad or\quad =\lambda I^{\lambda}(\boldsymbol{q},\boldsymbol{p})$$

by some $\lambda>0$.

(iii) *If* $I(\boldsymbol{p},\boldsymbol{q})$ *satisfies the relation*

$$I(\boldsymbol{p}\otimes\boldsymbol{p}',\ \boldsymbol{q}\otimes\boldsymbol{q}')=I(\boldsymbol{p},\boldsymbol{q})+I(\boldsymbol{p}',\boldsymbol{q}')-I(\boldsymbol{p},\boldsymbol{q})\cdot I(\boldsymbol{p}',\boldsymbol{q}')$$

then we have

$$I(\boldsymbol{p},\boldsymbol{q})=\mu I^{-\mu}(\boldsymbol{p},\boldsymbol{q}),\quad or\quad =\mu I^{-\mu}(\boldsymbol{q},\boldsymbol{p})$$

by some μ $(1/2\geqq\mu>0)$.

Remark. Rényi [10] characterized I^0, $\tilde{I}^{\lambda}$ and $\tilde{I}^{-\mu}$ by different axioms.

Let I be a differentiable fundamental information (7). Let $\boldsymbol{p}=(p_k)$, $\boldsymbol{q}=(q_k)$ and $\boldsymbol{q}^0=(q_k^0)$ be probability distributions and put $p_k=q_k^0+u_k$, $q_k=q_k^0+v_k$ $(k=1,\cdots,m)$.

If $|u_k|<\varepsilon$, $|v_k|<\varepsilon$ $(k=1,\cdots,m)$, then we have

$$I(\boldsymbol{p},\boldsymbol{q})=\frac{\alpha}{2}\sum_{k=1}^{m}\frac{1}{q_k^0}(u_k-v_k)^2+R,\qquad R=O(\varepsilon^3), \tag{8}$$

where $\alpha=(d^2K/dx^2)(1)\geqq 0$. We call α the *invariant* of I.

The invariant α of I^{λ} is given by $\alpha(I^{\lambda})=1+\lambda$ $(-1/2\leqq\lambda<\infty)$.

References

[1] H. Akaike: Information theory and an extension of the maximum likelihood principle. 2nd Internat. Symp. on Information Theory, Akademiai Kiado, Budapest, pp. 267–281 (1973).

[2] ——: A new look at the statistical model identification. IFEE Trans. Automatic Control, AC-19, pp. 716–723 (1974).

[3] S. M. Ali and S. D. Silvey: A general class of divergence of one distribution from another. J. Roy. Statist. Soc., B **28**, 131–142 (1966).

[4] I. Csizár: Information measures; a critical survey. Trans. 7-th Prague Conference on Information Theory, Statistical Decision Functions, Random Processes and of the 1974 European Meeting of Statisticians, vol. B, pp. 73–86 (1978).

[5] S. Kakutani: On equivalence of infinite product measures. Ann. of Math., **49**, 214–224 (1948).

[6] H. Kudō: A theorem of Kakutani on infinite product measures. Nat. Sci. Rep. Ochanomizu Univ., **3**, 10–22 (1952).

[7] ——: Theory of time series and informations, and their applications, Chap. 2. Statistical Experiments and Their Informations. Nippon Kagaku-Gizyutu Renmei, pp. 104–124 (1953) (in Japanese).

[8] S. Kullback: Information Theory and Statistics, Wiley (1959).

[9] P. N. Rathie and P. L. Kannappan: A directed-divergence function of type β*. Information and Control, **20**, 38–45 (1972).

[10] A. Rényi: On measures of entropy and information. Proc. Fourth Berkeley Symposium Math. Statist. and Probability, **1**, 547–561 (1961).