KYBERNETIKA — VOLUME 23 (1987), NUMBER 5

THE MUTUAL INFORMATION. ESTIMATION IN THE SAMPLING WITHOUT REPLACEMENT*

MARIA ANGELES GIL, RIGOBERTO PEREZ, PEDRO GIL

In previous papers, the "mutual information of order β concerning two random variables" was defined from the concept of conditional entropy of order β (Z. Daróczy, 1970). The aim of the present paper is to approach the value of the mutual information of order $\beta = 2$ in a large population on the basis of a sample drawn at random and without replacement from it. This purpose is achieved by obtaining an unbiased estimator of that value and estimating its mean square error. In addition, a contrast between samplings with and without replacement shows that the second one entails an improvement in the estimation precision with respect to the first one. Finally, we discuss the suitability of adopting the measure of order $\beta = 2$ against Shannon's amount of information.

1. INTRODUCTION

Consider an experiment involving the observation of two random variables, X and Y, corresponding to measurable characteristics associated with each random choice from a certain finite population. The randomness of each variable includes uncertainty which usually decreases by revealing the value of the other variable.

In order to evaluate how much information is conveyed about one of the variables by the other one, a usual procedure is to measure it as a reduction in uncertainty. When the joint probability distribution of X and Y is known, the uncertainty about the identity of the value of the variable X, and the uncertainty about the identity of the value of X when the value of Y is revealed, can be quantified by means of probabilistic uncertainty measures. Let $H(X)$ and $H(X \mid Y)$ denote such uncertainties. Then, the "information conveyed about X by Y" may be evaluated by the expression $I(X \mid Y) = H(X) - H(X \mid Y)$, that is, by means of the "mean decrease in uncertainty about the first variable by the revealment of the value of the second one".

Particularly, when $H(X)$, $H(X, Y)$ and $H(X \mid Y)$ represent respectively the entropy

* This work was supported in part by the Comisión Asesora de Investigación Científica y Técnica (CAICYT) under the contract P B85-0401.

of order β of X, the joint entropy of order β of X and Y (cf. Havrda-Charvát [21], Daróczy [13]), and the conditional entropy of order β of X with respect to Y (cf. Daróczy [13]), then $I(X \mid Y) = H(X) - H(X \mid Y) = H(X) + H(Y) - H(X, Y)$ that is, the information conveyed about X by Y is a symmetric measure with respect to X and Y (in other words, it coincides with the information conveyed about Y by X). From now on, this symmetric measure will be called mutual information concerning X and Y, and will be denoted by $I(X, Y)$.

The mutual information, $I(X, Y)$, may be applied and interpreted in different fields. The following are some of the most interesting applications:

i) Consider the problem of measuring the mutual information between the presence of certain species and an ecological parameter or factor, or the mutual information concerning the presence of two species. When the entropy associated with the presence of the species is quantified through the entropy of order β, then $I(X, Y)$ could be used for quantifying the preceding mutual information (see [11] and [12] for the Shannon measures).

ii) Consider the problem of measuring the "ecological diversity" of a finite population under a classification process X dividing it into M classes or species, when the population is subjected to an additional separate classification process Y that divides it into M' classes. If the diversity under a classification within each class determined by the other one is measured by means of the entropy of order β ([5], [6], [7], [30], [31]), then $I(X, Y)$ quantifies the mean decrease in diversity (or, the mean increase in concentration) under the X-classification caused by the adoption of the additional classification process Y, and conversely, the mean decrease in diversity under the Y-classification caused by the adoption of the additional classification process X.

iii) Consider a discrete constant channel with input alphabet X, characterized by M symbols, and output alphabet Y, characterized by M' symbols. Then, $I(X, Y)$ could be used for quantifying the information processed by the discrete constant channel (cf. [13]).

iv) Consider a probabilistic questionnaire having a set of M questions and a set of M' answers. Then, $I(X, Y)$ could be used for quantifying the information processed and transmitted by the questionnaire ([3], [15], [29]).

2. PRELIMINARY CONCEPTS

Statistical Inference deals with the drawing of conclusions about the variables in a population on the basis of a sample from it. In this sense, this paper is devoted to estimate the mutual information concerning two variables in a finite population from the knowledge of their joint probability distribution in a sample drawn at random and without replacement from the population.

In order to achieve this purpose we are going to consider the quadratic mutual information, or mutual information of order $\beta = 2$, which will allow us to construct an unbiased estimator.

Consider a finite population with N members, and let X and Y be two random variables in the population such that the random vector (X, Y) takes on the values (x_i, y_j) with joint probabilities p_{ij} $(i = 1, \ldots, M,\ j = 1, \ldots, M')$, respectively. Let $p_{i.} = \sum_{j=1}^{M'} p_{ij}$ $(i = 1, \ldots, M)$ and $p_{.j} = \sum_{i=1}^{M} p_{ij}$ $(j = 1, \ldots, M')$ the marginal probability distribution of X and Y in the population.

According to definition and results stated by Havrda and Charvát [21] and Daróczy [13], the mutual information concerning the variables X and Y in the population can be quantified as follows:

Definition 2.1. The value $I^2(X, Y)$ defined by

$$I^2(X, Y) = H^2(X) + H^2(Y) - H^2(X, Y) =$$
$$= 2(1 - \sum_{i=1}^{M} p_{i.}^2 - \sum_{j=1}^{M'} p_{.j}^2 + \sum_{i=1}^{M} \sum_{j=1}^{M'} p_{ij}^2)$$

is called *quadratic population mutual information* concerning the variables X and Y.

It should be remarked that the quadratic entropy H^2 has the qualitative signification and essential properties in Shannon's entropy (limit entropy of order β as $\beta \to 1$) for quantifying the probabilistic uncertainty.

In previous papers ([36], [37]) the average conditional quadratic entropies were defined by the values

$$\tilde{H}^2(X \mid Y) = \sum_{j=1}^{M'} p_{.j} H^2(X \mid y_j) = 2 \sum_{j=1}^{M'} p_{.j}[1 - \sum_{i=1}^{M} (p_{ij}/p_{.j})^2]$$
$$\tilde{H}^2(Y \mid X) = \sum_{i=1}^{M} p_{i.} H^2(Y \mid x_i) = 2 \sum_{i=1}^{M} p_{i.}[1 - \sum_{j=1}^{M'} (p_{ij}/p_{i.})^2]$$

Applications of this concept can be viewed in [15], [36] and [37].

On the other hand, Definition 2.1 admits the equivalent expression $I^2(X, Y) = H^2(X) - H^2(X \mid Y) = H^2(Y) - H^2(Y \mid X)$, when the conditional quadratic entropies $H^2(X \mid Y)$ and $H^2(Y \mid X)$ are intended as in [13] (that is,

$$H^2(X \mid Y) = \sum_{j=1}^{M'} p_{.j}^2 H^2(X \mid y_j) = 2 \sum_{j=1}^{M'} p_{.j}^2[1 - \sum_{i=1}^{M} (p_{ij}/p_{.j})^2]$$
$$H^2(Y \mid X) = \sum_{i=1}^{M} p_{i.}^2 H^2(Y \mid x_i) = 2 \sum_{i=1}^{M} p_{i.}^2[1 - \sum_{j=1}^{M'} (p_{ij}/p_{.i})^2])$$

We have taken into account the conditional entropies in the sense of Daróczy since they are well adapted to the problem in this paper. More precisely, the following are the basic arguments justifying this consideration:

– the strong non-additivity of the quadratic entropy H^2 entails (cf. [24]) that the joint entropy (uncertainty) is

$$H^2(X, Y) = H^2(Y) + \sum_{j=1}^{M'} p_{.j}^2 H^2(X \mid y_j) = H^2(X) + \sum_{i=1}^{M} p_{i.}^2 H^2(Y \mid x_i)$$

and hence, $\sum_{j=1}^{M'} p_{.j}^2 H^2(X \mid y_j)$ and $\sum_{i=1}^{M} p_{i.}^2 H^2(Y \mid x_i)$ could be respectively interpreted as the entropy (uncertainty) remaining in X given the value of Y, and the entropy (uncertainty) remaining in Y when the value of X is known.

- the measure I^2 in Definition 2.1 is symmetric, that is, the information contained in X about Y and that contained in Y about X coincide, so that it represents a "mutual" information.
- in [6] the use of the conditional quadratic entropy $H^2(X \mid Y)$ is justified by its application to evaluate the conditional diversity of a classification process X under the classification process Y. In addition, in [13] and [15] the adoption of the measure in Definition 2.1 has been warranted in terms of its adaptability to the problems of information transmission in discrete constant channels and pseudoquestionnaires.
- the measures $H^2(X \mid Y)$ and $H^2(Y \mid X)$ are easier to estimate unbiasedly than $\tilde{H}^2(X \mid Y)$ and $\tilde{H}^2(Y \mid X)$, as it can be later corroborated.

We now state some immediate properties of the quadratic mutual information in Definition 2.1.

Whatever the random variables X and Y may be, we have:

Theorem 2.1. (Nonnegativity.)

$$I^2(X, Y) \geqq 0 .$$

Theorem 2.2. (Symmetry.)

$$I^2(X, Y) = I^2(Y, X) .$$

Theorem 2.3. (Maximum Information.)

$$I^2(X, Y) \leqq I^2(X, X) = H^2(X) .$$

Theorem 2.4. If X and Y are independent variables, then

$$I^2(X, Y) = H^2(X) \,.\, H^2(Y)/2 .$$

It is worth emphasizing that the quadratic mutual information does not vanish for independent variables. This circumstance must be hoped because of the nonadditivity of the quadratic entropy from which the mutual information is defined (since $H^2(X, Y) = H^2(X) + H^2(Y) - H^2(X) \,.\, H^2(Y)/2$ for X and Y independent variables). Thus, although Shannon's entropy is additive for independent variables the entropies suggested in [13] and [21] do not satisfy such a property and, consequently, they are often called "nonadditive measures of order β".

Theorem 2.5. $I^2(X, Y)$ is a concave function with respect to the vectors $(p_{1.}, \ldots, p_{M.})$ and $(p_{.1}, \ldots, p_{.M'})$.

Remark 2.1. It should be pointed out that when the mutual information of order β concerning two random variables X and Y is defined by $I^\beta(X, Y) = H^\beta(X) + H^\beta(Y) - H^\beta(X, Y)$, then the Shannon mutual information becomes the limit of $I^\beta(X, Y)$ as $\beta \to 1$.

3. AN UNBIASED ESTIMATOR OF THE POPULATION MUTUAL INFORMATION IN THE SAMPLING WITHOUT REPLACEMENT

Following ideas in the estimation of certain population parameters, such as the population mean or variance, in order to estimate the quadratic population mutual information we first introduce the analogue estimator of that concept and we then construct an unbiased estimator from the analogue one. In this way, let $(\mathscr{x}, \mathscr{y})$ denote the random vector taking on the values (x_i, y_j) with joint relative frequencies $f_{ij}(\mathscr{x}, \mathscr{y})$ $(i = 1, \ldots, M,\ j = 1, \ldots, M')$, respectively, in a generic sample of size n drawn at random from the population. Let $f_{i.} = \sum_{j=1}^{M'} f_{ij}$ $(i = 1, \ldots, M)$ and $f_{.j} = \sum_{i=1}^{M} f_{ij}$ $(j = 1, \ldots, M')$ be the marginal relative frequencies associated with X and Y, respectively, in the sample. Then, the mutual information concerning X and Y in the sample can be quantified as follows:

Definition 3.1. The value $I_n^2(\mathscr{x}, \mathscr{y})$ defined by

$$I_n^2(\mathscr{x}, \mathscr{y}) = H_n^2(\mathscr{x}) + H_n^2(\mathscr{y}) - H_n^2(\mathscr{x}, \mathscr{y}) =$$
$$= 2(1 - \sum_{i=1}^{M} [f_{i.}(\mathscr{x})]^2 - \sum_{j=1}^{M'} [f_{.j}(\mathscr{y})]^2 + \sum_{i=1}^{M} \sum_{j=1}^{M'} [f_{ij}(\mathscr{x}, \mathscr{y})]^2)$$

is called *quadratic sample mutual information* concerning the variables X and Y.

In order to analyze the suitability of the analogue estimate of $I^2(X, Y)$, $I_n^2(\mathscr{x}, \mathscr{y})$, we first wish to point out that the random vectors $(nf_{11}, \ldots, nf_{MM'})$, $(nf_{1.}, \ldots, nf_{M.})$ and $(nf_{.1}, \ldots, nf_{.M'})$ have multivariate hypergeometric distributions with parameters $(N, D_{11} = Np_{11}, \ldots, D_{MM'} = Np_{MM'}, n)$, $(N, D_{1.} = Np_{1.}, \ldots, D_{M.} = Np_{M.}, n)$ and $(N, D_{.1} = Np_{.1}, \ldots, D_{.M'} = Np_{.M'}, n)$, respectively. Then, the expected value of the analogue estimator over all samples $(\mathscr{x}, \mathscr{y})$ of size n in a random sampling without replacement is given by

$$\mathsf{E}(I_n^2) = 2[1 - \sum_{i=1}^{M} \mathsf{E}(f_{i.}^2) - \sum_{j=1}^{M'} \mathsf{E}(f_{.j}^2) + \sum_{i=1}^{M} \sum_{j=1}^{M'} \mathsf{E}(f_{ij}^2)] =$$
$$= (n - 1)\, N\, I^2(X, Y)/n(N - 1)$$

Consequently, the analogue estimator I_n^2 is consistent in the Cochran sense (that is, the estimate becomes equal to the population mutual information when $n = N$). In addition, it allows us to construct an unbiased estimator of $I^2(X, Y)$, since whatever the sample size n may be we have

Theorem 3.1. Let (X, Y) be a random vector in a finite population of N members taking on the values (x_i, y_j) $(i = 1, \ldots, M, j = 1, \ldots, M')$. In the random sampling without replacement from this population, the estimator $(\hat{I_n^2})^c$ allocating to each sample $(\mathfrak{x}, \mathfrak{y})$ of n members the value $(\hat{I_n^2})^c(\mathfrak{x}, \mathfrak{y}) = n(N - 1) I_n^2(\mathfrak{x}, \mathfrak{y})/(n - 1) N$ is an unbiased estimator of the quadratic population mutual information concerning X and Y.

Theorem 3.1 suggests the introduction of the following concept:

Definition 3.2. The estimator $(\hat{I_n^2})^c$ allocating to each sample $(\mathfrak{x}, \mathfrak{y})$ of size n without replacement the value $n(N - 1) I_n^2(\mathfrak{x}, \mathfrak{y})/(n - 1) N$ is called the *quadratic sample mutual quasi-information concerning X and Y corriged by finite population.*

4. EXACT PRECISION OF THE UNBIASED ESTIMATOR

For the sake of evaluating the precision of the quadratic sample mutual quasi-information corriged by finite population in estimating the quadratic population mutual information, we now measure the mean square error of that unbiased estimator.

Theorem 4.1. Let (X, Y) be a random vector in a finite population with N members taking on the values (x_i, y_j) $(i = 1, \ldots, M, j = 1, \ldots, M')$. If $(\hat{I_n^2})^c$ is the quadratic mutual quasi-information corriged by finite population for random samples of size n drawn without replacement from the population, then its variance (or mean squared error)

$$V((\hat{I_n^2})^c) = (N - n)\{[(6 - 4n) N + 6(n - 1)] N[I^2(X, Y)]^2 -$$
$$- 12(n - 2) N(N - 1) I^3(X, Y) + 4[(4n - 7) N - (n + 1)] (N - 1) I^2(X, Y) +$$
$$+ 32(n - 2) N(N - 1) \sum_{i=1}^{M} \sum_{j=1}^{M'} p_{ij}(p_{i.} - p_{ij})(p_{.j} - p_{ij})\} :$$
$$: 1/(n(n - 1) N(N - 2)(N - 3))$$

being $I^3(X, Y) = H^3(X) + H^3(Y) - H^3(X, Y) = \frac{4}{3}(1 - \sum_{i=1}^{M} p_{i.}^3 - \sum_{j=1}^{M'} p_{.j}^3 + \sum_{i=1}^{M} \sum_{j=1}^{M'} p_{ij}^3)$ the mutual information of order $\beta = 3$ concerning X and Y.

Proof. Indeed, the random vectors $(nf_{11}, \ldots, nf_{MM'})$, $(nf_{1.}, \ldots, nf_{M.})$ and $(nf_{.1}, \ldots, nf_{.M'})$ have the multivariate hypergeometric distribution mentioned in

Section 3, and hence

$$V((\hat{I_n^2})^c) = (N - n)\{16(N - n - 1) - 32(N - n - 1)\sum_{i=1}^{M}\sum_{j=1}^{M'} p_{ij}^2 +$$

$$+ 8N[(6 - 4n)N + 6(n - 1)][(\sum_{i=1}^{M} p_{i.}^2)(\sum_{j=1}^{M'} p_{.j}^2) -$$

$$- (\sum_{i=1}^{M} p_{i.}^2)(\sum_{i=1}^{M}\sum_{j=1}^{M'} p_{ij}^2) - (\sum_{j=1}^{M'} p_{.j}^2)(\sum_{i=1}^{M}\sum_{j=1}^{M'} p_{ij}^2)] +$$

$$+ 32(n - 2)N(N - 1)\sum_{i=1}^{M}\sum_{j=1}^{M'} p_{ij}(p_{i.}p_{.j} - p_{i.}p_{ij} - p_{.j}p_{ij}) -$$

$$- 16(N - n - 1)(N - 1)\sum_{i=1}^{M}\sum_{j=1}^{M'} p_{ij}^2\}/n(n - 1)N(N - 2)(N - 3)$$

which accounts for proving the theorem. □

Remark 4.1. the term $\sum_{i=1}^{M}\sum_{j=1}^{M'} p_{ij}(p_{i.} - p_{ij})(p_{.j} - p_{ij})$ in $V((\hat{I_n^2})^c)$ cannot only be expressed by means of mutual information measures of order β, but it may be further expressed in terms of measures concerning another related concept in the Information Theory and Statistics: the inaccuracy. Thus, the preceding term equals $I^2(X, Y)[1 - H_2(P; Q)/2]/2$, where $H_2(P; Q)$ is the inaccuracy of order $\beta = 2$ of P with respect to Q, where P and Q denote the probability distribution $\{p_{ij}\}$ and $\{(p_{i.} - p_{ij})(p_{.j} - p_{ij})/I^2(X, Y)\}$, respectively (cf. [31]).

Remark 4.2. Theorem 4.1 implies that the mean square error equals zero as $n = N$ (i.e., when the considered sample drawn without replacement is the whole population). In addition, zero is the limit of the mean square error as $n \to N$.

Remark 4.2 allows us to conclude that for a large sample the mean square error of the unbiased estimator $(\hat{I_n^2})^c$ is small. In addition, we can verify that the greater the size of the sample, the lower its mean square error is, since

Theorem 4.2. Whatever the sample size $n > 3$ may be, we have

$$V((\hat{I_n^2})^c) - V((\hat{I_{n-1}^2})^c) = -(N - 1)\{[(n - 4)N + n](N - 2)V((\hat{I_{N-1}^2})^c) +$$

$$+ 4(N - n)V((\hat{I_2^2})^c)/(N - 2)\}/n(n - 1)(n - 2)(N - 3)$$

5. COMPARING THE PRECISION OF ESTIMATIONS IN SAMPLINGS WITH AND WITHOUT REPLACEMENT

In this section we are going to corroborate that the estimation of the population mutual information in a random sampling without replacement improves its estimation (more precisely, its precision) in the sampling with replacement. It should be remarked that the sampling with replacement may be regarded as a limit situation of the sampling without replacement as the size population N tends to ∞ (since the multinomial distribution with parameter $n, p_1, \ldots, p_k$ is the limit of the multivariate hypergeometric distribution with parameters $N, D_1 = Np_1, \ldots, D_k = Np_k$ and n as $N \to \infty$) On the basis of this remark, we can state

Theorem 5.1. In estimating $I^2(X, Y)$ the precision of $(\hat{I}_n^2)^c$ is greater than the precision of $\hat{I}_n^2 = \lim_{N\to\infty} (\hat{I}_n^2)^c$.

Proof. Indeed, if ΔV denotes the variation $V(\hat{I}_n^2) - V((\hat{I}_n^2)^c)$, we have

$$\Delta V = 2(N - n)\{(n + 1)(N - 1)V(\hat{I}_2^2) + (N - n - 1)[I^2(X, Y)]^2\} :$$
$$: n(n - 1)N(N - 2)(N - 3) + \{(N - 2)(N - 3) - (N - 1)(N - n)\} \times$$
$$\times V(\hat{I}_n^2)/(N - 2)(N - 3)$$

and then, ΔV takes on a nonnegative value, whatever the sample size n may be. □

Remark 5.1. ΔV converges to zero as $N \to \infty$, and this result is coherent with comments at the beginning of this section.

6. ESTIMATED PRECISION OF THE UNBIASED ESTIMATOR

As the mean square error of $(\hat{I}_n^2)^c$ involves population probabilities of the variable values, this error will not be known in practice. However, this error can be estimated from the considered sample. We are now going to construct an unbiased estimator of $V((\hat{I}_n^2)^c)$ following arguments in Section 3.

Theorem 6.1. Let (X, Y) be a random vector in a finite population of N members taking on the values (x_i, y_j) $(i = 1, \ldots, M, j = 1, \ldots, M')$. If $v((\hat{I}_n^2)^c)$ is the estimator allocating to each sample $(\mathscr{x}, \mathscr{y})$ of n members in a random sampling without replacement the value given by

$$v((\hat{I}_n^2)^c) = n(N - 1)(N - n)\{[(6 - 4n)N + 6(n - 1)]\, n[I_n^2(\mathscr{x}, \mathscr{y})]^2 -$$
$$- 12n(n - 1)(N - 2)I_n^3(\mathscr{x}, \mathscr{y}) + 4(n - 1)[(4n - 1)N - 7n - 1]\, I_n^2(\mathscr{x}, \mathscr{y}) +$$
$$+ 32n(n - 1)(N - 2)\sum_{i=1}^{M}\sum_{j=1}^{M'} f_{ij}(\mathscr{x}, \mathscr{y})\,[f_{i.}(\mathscr{x}) - f_{ij}(\mathscr{x}, \mathscr{y})] \times$$
$$\times [f_{.j}(\mathscr{y}) - f_{ij}(\mathscr{x}, \mathscr{y})]\}/(n - 1)^2(n - 2)(n - 3)N^3$$

being $I_n^3(\mathscr{x}, \mathscr{y}) = H_n^3(\mathscr{x}) + H_n^3(\mathscr{y}) - H_n^3(\mathscr{x}, \mathscr{y}) = \frac{4}{3}\{1 - \sum_{i=1}^{M} [f_{i.}(\mathscr{x})]^3 - \sum_{j=1}^{M'} [f_{.j}(\mathscr{y})]^3 + \sum_{i=1}^{M}\sum_{j=1}^{M'} [f_{ij}(\mathscr{x}, \mathscr{y})]^3\}$ the mutual information of order $\beta = 3$ concerning X and Y in the sample $(\mathscr{x}, \mathscr{y})$, then, $v((\hat{I}_n^2)^c)$ is an unbiased estimator of $V((\hat{I}_n^2)^c)$. □

Remark 6.1. Theorem 6.1 connects n with an unbiased estimator of the precision of $(\hat{I}_n^2)^c$. With such a connection one could readily estimate the suitable size for estimating the quadratic population mutual information (and consequently the mutual information between the presence of species and ecological factors, the mean decrease in diversity, the information processed by a discrete constant channel, and so on), by means of the quadratic sample mutual quasi-information corriged by finite population, with a desired degree of precision. The estimation of this suitable size could be accomplished either by using a previous sampling from the population (in order to approximate $I_n^2(x, y)$, $I_n^3(x, y)$, $f_{ij}(x, y)$, $f_{i.}(x)$ and $f_{.j}(y)$), or by using a sequential sampling.

7. ADVANTAGES OF THE QUADRATIC MUTUAL INFORMATION AGAINST THE SHANNON MUTUAL INFORMATION

Consider a finite population with N members, and let X and Y be two random variables in the population such that the random vector (X, Y) takes on the values (x_i, y_j) with joint probabilities p_{ij} $(i = 1, \ldots, M, j = 1, \ldots, M')$, respectively. Let $p_{i.}$ and $p_{.j}$ the marginal probabilities of the values x_i and y_j in the population.

Following Shannon, the mutual information concerning the variables X and Y in the population (limit of the population mutual information of order β as $\beta \to 1$) is quantified by means of

Definition 7.1. The value $I^1(X, Y)$ defined by

$$I^1(X, Y) = H^1(X) + H^1(Y) - H^1(X, Y) =$$
$$= -\sum_{i=1}^{M} p_{i.} \log_2 p_{i.} - \sum_{j=1}^{M'} p_{.j} \log_2 p_{.j} + \sum_{i=1}^{M} \sum_{j=1}^{M'} p_{ij} \log_2 p_{ij}$$

is called *Shannon's population mutual information* concerning the variables X and Y.

The analogue estimator of $I^1(X, Y)$ for a random sample (x, y) drawn from the population and characterized by a random vector taking on the values (x_i, y_j) with joint relative frequencies $f_{ij}(x, y)$ $(i = 1, \ldots, M, j = 1, \ldots, M')$, respectively, is defined as follows:

Definition 2.2. The value $I_n^1(x, y)$ defined by

$$I_n^1(x, y) = H_n^1(x) + H_n^1(y) - H_n^1(x, y) = -\sum_{i=1}^{M} f_{i.}(x) \log_2 f_{i.}(x) -$$
$$- \sum_{j=1}^{M'} f_{.j}(y) \log_2 f_{.j}(y) + \sum_{i=1}^{M} \sum_{j=1}^{M'} f_{ij}(x, y) \log_2 f_{ij}(x, y)$$

is called *Shannon's sample mutual information* concerning the variables X and Y.

When we examine the expected value of I_n^1 over all samples of size n in a random sampling without replacement, we obtain

$$\mathsf{E}(I_n^1) = \sum_{i=1}^{M} \sum_{j=1}^{M'} \mathsf{E}(f_{ij} \log_2 f_{ij}) - \sum_{i=1}^{M} \mathsf{E}(f_{i.} \log_2 f_{i.}) - \sum_{j=1}^{M'} \mathsf{E}(f_{.j} \log_2 f_{.j})$$

Nevertheless, an exact relation, irrespective of the variables X and Y, cannot be established between $\mathsf{E}(f_{ij} \log_2 f_{ij})$ and $\mathsf{E}(f_{ij}) \log_2 \mathsf{E}(f_{ij}) = p_{ij} \log_2 p_{ij}$, and so on.

The preceding argument leads to the following conclusion: An unbiased estimator of Shannon's population mutual information irrespective of the concerning variables X and Y cannot be immediately defined from Shannon's sample mutual information.

In the same way, if I^β and I_n^β denote respectively the population and sample mutual information of order β (where $\beta \neq 1, 2$) which can be defined in a similar form following Daróczy [13], to state exact relations between $\mathsf{E}(I_n^\beta)$ and I^β is either impossible or more complicated.

8. CONCLUDING REMARKS

The results we have just expounded could be used for estimating the information conveyed by a random sample about its corresponding population with respect to a random variable.

On the other hand, the study in this paper was developed in a previous paper [19] for the random sampling with replacement and it might be accomplished for the stratified random sampling, which would provide greater precisions. In the same way, the estimation of the population mutual information could be examined for the case when the adopted sampling is not random.

It is worth remarking that in [27] we have recently analyzed the problem of estimating the uncertainty associated with a random variable in a finite population in both, the samplings with and without replacement. This analysis leads to a conclusion similar to that in the present paper: the quadratic entropy, or entropy of order $\beta = 2$, is the best adapted for the random sampling and, consequently, it is more suitable than Shannon's entropy in such a situation.

Another interesting study leading also to similar conclusions would be determined by the estimation of the unquietness associated with a random variable in a finite population, which could directly be applied to the estimation of the income inequality in a large population (cf. [8], [10], [14], [16], [17], [18], [26], [33], [35] and [38]. This study has been introduced in [28].

Finally, it should be emphasized that the research about the unbiased estimation of the mutual information could be complemented by examining the asymptotic distribution of the estimators, following ideas in [23] and [39], and using results in [7].

APPENDIX

MOMENTS OF THE MULTIVARIATE HYPERGEOMETRIC DISTRIBUTION

Let $(n_1, \ldots, n_M)$ be a random vector with multivariate hypergeometric distribution, where N, $D_1 = Np_1, \ldots, D_M = Np_M$ and n are the corresponding parameters. Then,

$$\mathsf{E}(n_i) = np_i$$

$$\mathsf{E}(n_i^2) = (N-n)\, np_i[N(n-1)\, p_i/(N-n) + 1]/(N-1)$$

$$\mathsf{E}(n_i^3) = (N-n)\, np_i[N^2(n-1)(n-2)\, p_i^2/(N-n) + 3N(n-1)\, p_i + N - 2n] : \\ :(N-1)(N-2)$$

$$\mathsf{E}(n_i^4) = (N-n)\, np_i[N^3(n-1)(n-2)(n-3)\, p_i^3/(N-n) + 6(n-1)(n-2) \times \\ \times N^2 p_i^2 + N(n-1)(7N - 11n + 1)\, p_i + N(N+1) - 6n(N-n)] : \\ :(N-1)(N-2)(N-3)$$

$$\mathsf{E}(n_i n_j) = Nn(n-1)\, p_i p_j/(N-1)\,, \quad i \neq j$$

$$\mathsf{E}(n_i^2 n_j^2) = n(n-1)\, N^2(N-n)\, p_i p_j[N(n-2)(n-3)\, p_i p_j/(N-n) + \\ + (n-2)(p_i + p_j) + (N - n - 1)/N]/(N-1)(N-2)(N-3)\,, \\ i \neq j$$

$$\mathsf{E}(n_i^3 n_j) = (N-n)\, N^2 n(n-1)\, p_i p_j[N(n-2)(n-3)\, p_i^2/(N-n) + \\ + 3(n-2)\, p_i + (N - 2n + 1)/N]/(N-1)(N-2)(N-3)\,, \\ i \neq j$$

$$\mathsf{E}(n_i^2 n_j n_k) = (N-n)\, N^2 n(n-1)(n-2)\, p_i p_j p_k[N(n-3)\, p_i/(N-n) + 1] : \\ :(N-1)(N-2)(N-3) \quad j \neq i\,,\ k \neq i, j$$

$$\mathsf{E}(n_i n_j n_k n_l) = N^3 n(n-1)(n-2)(n-3)\, p_i p_j p_k p_l/(N-1)(N-2)(N-3), \\ j \neq i;\ k \neq i, j;\quad l \neq i, j, k$$

ACKNOWLEDGEMENT

The authors thank the helpful suggestions of the referee in connection with this paper.

(Received April 4, 1986.)

REFERENCES

[1] J. Aczél and Z. Daróczy: On Measures of Information and Their Characterizations. Academic Press, New York 1975.

[2] J. Aczél and Pl. Kannapan: A mixed theory of information, III, inset entropies of degree β. Inform. and Control *39* (1978), 315—322.

[3] N. L. Aggarwal, Y. Cesari and C. F. Picard: Propriétés de branchement liées aux questionnaires de Campbell et à l'information de Renyi C. R. Acad. Sci. Paris Sér. A *275* (1972), 437—440.

[4] F. Azorin: Curso de Muestreo y Aplicaciones. Ed. Aguilar, Madrid 1972.
[5] F. Azorin: Diversity and uncertainty. Proc FISAL-83, Palma de Mallorca 1983, pp. 37—44.
[6] T. N. Bhargava and V. R. R. Uppuluri: On an axiomatic derivation of Gini diversity with applications. Metron *XXXIII* (1975), No. 1—2, 1—13.
[7] T. N. Bhargava and V. R. R. Uppuluri: Sampling distribution of Gini's index of diversity. Appl. Math. Comput. *3*, (1977), 1—24.
[8] F. Bourguignon: Decomposable income inequality measures. Econometrica *47* (1979), 901—920.
[9] W. G. Cochran: Sampling Techniques. John Wiley, New York 1967.
[10] F. A. Cowell: On the structure of additive inequality measures Rev. Econom. Stud. *47* (1980), 521—531.
[11] Ph. Daget et M. Godron: Analyse Fréquentielle de l'écologie. Masson, Paris 1982.
[12] Ph. Daget, M. Godron et J. L. Guillerm: Profils écologiques et information mutuelle entre especes et factuers écologiques, Grundfragen iun Methoden in the Pflanzensoziologie. Verlag Dr. W. Junk N. V., Den Haag 1972, pp. 121—149.
[13] Z. Daróczy: Generalized information functions. Inform. and Control *16* (1970) 36—51.
[14] W. Eichhorn and W. Gehrig: Measurement of inequality in economics. In: Modern Applied Mathematics Optimization and Operations Research (B. Korte ed.), North Holland, Amsterdam 1982.
[15] H. Emptoz: Informations Utiles et Pseudoquestionnaires. Thése, Université Claude-Bernard, Lyon 1976.
[16] M. A. Gil: Estudio de una medida para la Incertidumbre correspondiente a las Utilidades. Trabajos Estadíst. Investigación Oper. *32* (1981), 3, 45—66.
[17] M. A. Gil: Criterion of maximizing the expected quietness (invariant by homotheties in relation to the utilities). R. A. I. R. O. — Rech. Opér. *16* (1982), 319—331.
[18] M. A. Gil: Mixed criterion of expected utility and quietness (invariant by homotheties with respect to the utilities). Statistica *XLII* (1982), 1, 21—37.
[19] M. A. Gil, R. Perez and I. Martinez: The mutual information. Estimation in the sampling with replacement. R. A. I. R. O. — Rech. Opér. *20* (1986), 3, 257—268.
[20] P. Gil: Teoría Matemática de la Información. Ed. ICE, Madrid 1981.
[21] J. Havrda and F. Charvát: Quantification method of classification processes. Kybernetika *3* (1967), 30—35.
[22] M. Kendall and A. Stuart: The Advanced Theory of Statistics. Charles Griffin Co. Ltd., London 1977—83.
[23] Z. Lomnicki and S. Zaremba: The asymptotic distributions of estimators of the amount of transmitted information. Inform. and Control *2* (1959), 260—284.
[24] A. M. Mathai and P. N. Rathie: Basic Concepts in Information Theory and Statistics. Wiley Eastern Limited, New Delhi 1975.
[25] O. Onicescu: Energie informationnelle. C. R. Acad. Sci. Paris, Sér. A *263* (1966), 841—842.
[26] R. Pérez: Estimación de la Incertidumbre, la Incertidumbre útil y la Inquietud en Poblaciones Finitas. Rev. Real Acad. Ciencias de Madrid *LXXIX* (1985), 4, 651—654.
[27] R. Pérez, M. A. Gil and P. Gil: Estimating the uncertainty associated with a variable in a finite population. Kybernetes *15* (1986), 251—256.
[28] R. Pérez, C. Caso and M. A. Gil: Unbiased estimation of income inequality. Statist. Hefte *27* (1986), 227—237.
[29] C. F. Picard: Graphes et Questionnaires, Tome 2. Gauthier-Villars, Paris 1972.
[30] E. C. Pielou: An Introduction to Mathematical Diversity. Wiley Interscience, New York 1969.
[31] P. R. Rathie and Pl. Kannapan: An inaccuracy function of the type β. Ann. Inst. Statist. Math. *25* (1973), 205—214.

[32] C. E. Shannon: A mathematical theory of communication Bell System Tech. J. *27* (1949), 379—423.
[33] A. F. Shorrocks: The class of additively decomposable inequality measures. Econometrica *48* (1980), 613—625.
[34] E. H. Simpson: Measurement of diversity. Nature *163* (1949), 688.
[35] H. Theil: Economics and Information Theory. North Holland, Amsterdam 1967.
[36] I. Vajda: Bounds of the minimal error probability of checking a finite or countable number of hypotheses. Problems Inform. Transmission *4* (1968), 1, 9—19.
[37] I. Vajda: A contribution to the informational Analysis of Pattern. Methodologies of Pattern Recognition, Academic Press, New York 1969, pp. 509—519.
[38] D. Zagier: On the decomposability of the Gini coefficient and other indices of inequality. Discussion Paper No. 108, Projectgruppe Theoretische Modelle, Universität Bonn, 1983.
[39] J. Zvárová: On asymptotic behaviour of sample estimator of Renyi's information of order α. In: Trans. Sixth Prague Conf. on Inform. Theory, Statist. Dec. Functions, Random Processes, Academia, Prague 1971, pp. 919—924.

Dr. María Angeles Gil, Dr. Rigoberto Pérez, Prof. Dr. Pedro Gil, Department of Mathematics, University of Oviedo, 33071 Oviedo. Spain.