Bul. Stiinţ. Univ. Baia Mare, Ser. B,
Matematică-Informatică, Vol. XIV(1998), Nr. 2, 161-172

THE INFORMATIONAL STUDY OF A STATISTICAL COLLECTIVITY

Cristina Ioana Fătu – Mihoc, Ion Mihoc

Abstract. In this paper we consider an informational characterization of the homogeneity of a statistical population relative to a common property of hers elements.

Keywords: statistical collectivity, homogeneity, directed divergence.

AMS classification: 94A17

1. Introduction

Let $\mathbf{C}$ be a population (collectivity) and X a common property of hers elements. We want to study this collectivity relative to this common property (characteristic) when $\mathbf{S}_X = (X_1, X_2, ..., X_N)$ is a $N-$dimensional sample vector which corresponds to a sample $\mathbf{S}$, N is the size of this sample.

Formally, the process of sampling may be defined as a procedure for selecting a number of elements (constituting a sample) from a collectivity in such away that any particular property of the sample will corresponds as closely as possible to the true value of the property in the collectivity.

Because X is a random variable any statistical investigation of hers deals with a collection of the results of observations $(x_1, x_2, ..., x_N)$ representing the values of the random variable X when the size of the sample is N. More precisely, $(x_1, x_2, ..., x_N)$ is the value of the random vector $\mathbf{S}_X = (X_1, X_2, ..., X_N)$ whose components are independent identically distributed according to the random variable X.

Also, if we consider a random experiment $\mathcal{A}$, connected with this random variable X and if we make N independent repetitions of $\mathcal{A}$, (that is, if we take a sample of the size N from the population $\mathbf{C}$) then we shall obtain a sequence of N observed values of the random variable X. A such experiment (this sample of size N) must to be „representative" of the all population, that is, this sample must to satisfy the following conditions:

– the size of the sample N must to be large;

– the collectivity (population) $\mathbf{C}$ must to be homogeneous relative to the common property X;

– the sample must to be a random sample;

– the probability of being chosen is the same for all elements of the population **C**.

Therefore, a representative sample must to have and the property of the homogenous.

Contributions to measures of the homogeneity of a statistical population have been made by Kapur (1986), Kullbak and Leibler (1951), Rényi (1961), Theil (1986) and Văduva (1987).

In this paper, we consider an informational characterization of the homogeneity of a statistical population relative to a common property of hers elements.

2. Generalized probability distributions

Let (Ω, K, P) be a probability space, that is, Ω an arbitrary nonempty set, called the set of elementary events; K a $\sigma-$algebra of subsets of Ω containing Ω itself, the elements of K being called events; and P a probability measure, that is, a nonnegative and additive set function, defined on K, for which $P(\Omega) = 1$.

Let

$$\Delta_N^* = \left\{ P = (p_1, p_2, ..., p_N);\ p_i > 0,\ i = \overline{1,N};\ \sum_{i=1}^{N} p_i = 1 \right\} \tag{2.1}$$

be the set of all N-component complete probability distributions with positive elements associated with a discrete finite random variable X.

Rényi [5] introduced a generalization of the notion of a random variable.

Definition 2.1. An incomplete random variable random variable X is a function $X = X(\omega)$ measurable with respect to the measure on K and defined on a subset Ω_1 of Ω, where $\Omega_1 \subset \Omega$ and $P(\Omega_1) > 0$.

The only difference between an ordinary random variable (X is an ordinary or complete random variable if $P(\Omega_1) = 1$) and an incomplete random variable is thus that the latter is not necessarily defined for even $\omega \in \Omega$. Therefore, an incomplete random variable can be interpreted as a quantity describing the result of an experiment $\mathcal{A}$ depending on chance which is not always observable, any with the probability $0 < P(\Omega_1) < 1$.

Definition 2.2. If $0 < P(\Omega_1) \leq 1$, then the random variable X, defined on Ω_1, is a generalized random variable. The distribution of a generalized random variable X will be called a generalized probability distribution.

We denote by

$$w(P) = \sum_{i=1}^{N} p_i \tag{2.2}$$

the weight of the distribution P.

Using the above definitions it follows that:

– if $w(P) = 1$,then P is an ordinary (complete) probability distribution;

– if $0 < w(P) < 1$,then P is an incomplete probability distribution;

– if $0 < w(P) \leq 1$, then P is a generalized probability distribution.

Also, we denote by

$$\Delta_N = \left\{ P = (p_1, p_2, ..., p_N);\ p_i > 0,\ i = \overline{1, N};\ 0 < w(P) \leq 1 \right\} \tag{2.3}$$

the set of all finite discrete generalized probability distributions.

3. Informational measures of the homogeneity of a statistical population

Let $P = (p_1, p_2, ..., p_N)$, $P \in \Delta_N^*$ be the probability distribution associated with a complete system of events $\mathbf{S}_X = \{A_1, A_2, ..., A_N\}$, where

$$p_i = P(X = x_i) = P(A_i) > 0,\ i = \overline{1, N} \tag{3.1}$$

and X a discrete finite random variable which represents a common property for the all elements of the population C.

Definition 3.1. [5] The amount of uncertainty of the distribution P, that is, the amount of uncertainty concerning the outcome of an experiment $\mathcal{A}$, the possible results of which have the probabilities $p_1, p_2, ..., p_N$, is called Shannon's entropy of the distribution P and is usually measured by the quantity

$$H^*(P) = H^*(X) = H^*(\mathcal{A}) = -\sum_{i=1}^{N} p_i \log_2 p_i. \tag{3.2}$$

In connection with the notion of uncertainty we also have to mention the concept of information. Such, $H^*(P)$ is the information contained in the values of X or the amount of information contained by the random variable X generated by a random experiment $\mathcal{A}$ and we may write $H^*(X) = H^*(\mathcal{A})$ instead of $H^*(P)$.

This measure $H^*(P)$ is additive in the following sense:

$$H^*(P * Q) = H^*(P) + H^*(Q), \tag{3.3}$$

where

$$P * Q = \{p_i q_j \mid p_i \in P, q_j\} \in Q \in \Delta_{N^2}^* \tag{3.4}$$

is the direct product of the probability distributions $P, Q \in \Delta_N^*$.

There are also and the other measures of information which are often considered. These measures have been characterized by different sets of postulates by choosing proper algebraic properties satisfied by them. One of the algebraic properties shared by all these measures is the property of additivity.

Definition 3.2. Let $P, Q \in \Delta_N^*$ two ordinary probability distributions, that is, $w(P) = w(Q) = 1$.Then, the directed divergence [4] or the gain of information [5] is defined as

$$D^*(P;Q) = I^*(P \mid\mid Q) = \sum_{i=1}^{N} p_i \log_2 \frac{p_i}{q_i}. \tag{3.5}$$

Notice that the quantity (3.5) is defined for two finite discrete probability distribution $P, Q \in \Delta_N^*$ only if $q_i > 0$ for, $i = \overline{1, N}$, and if there is given a one-to-one correspondence between the elements of the distribution P and Q, which must therefore consist of an equal number of terms.

The gain of information is one of the most important notions in the information theory; it may even be considered as the fundamental one, from which all others can be derived. The quantity $I^*(P \mid\mid Q) = D^*(P;Q)$ give us the gain of information resulting from the replacement of the probability distribution P by the probability distribution Q.

This measure is additive in the following sense

$$D^*(P;Q) = D^*(P_1;Q_1) + D^*(P_2;Q_2), \tag{3.6}$$

where

$$P_1, P_2, Q_1, Q_2 \in \Delta_N^*, \tag{3.7}$$

$$P = P_1 * P_2, Q = Q_1 * Q_2, P, Q \in \Delta_{N^2}^* \tag{3.8}$$

and the correspondence between the elements of P and Q is that induced by the correspondence between the elements of P_1 and Q_1,and those of P_2 and Q_2.

Definition 3.3. [1] Let $P \in \Delta_N$ be a generalized probability distribution, that is, $0 < w(P) \leq 1$. The generalized mean value of the generalized probability distribution P is a positive quantity $M_g[P]_f$ defined by

$$M_g[P]_f = g^{-1}\left(\frac{\sum_{i=1}^{N} f(p_i)g(p_i)}{\sum_{i=1}^{N} f(p_i)}\right), \tag{3.9}$$

where the weight function f and the representation function g must to satisfy the following conditions:

c_1) f is positive and bounded function on the interval $(0, 1]$;

c_2) g is a strictly monotonic and continuous function on the interval $(0, 1)$.

Definition 3.4.[2] The quantity

$$H_g(P)_f = -\log_2 M_g[P]_f \tag{3.10}$$

represents the generalized information of the Daróczy sense if the generalized mean value $M_g[P]_f$ satisfies and the following conditions:

c_3) $M_g[P]_f$ depends only on the probabilities $p_i, i = \overline{1, N}$;

c_4) if $P = (p_1, p_2, ..., p_N), Q = (q_1, q_2, ..., q_N)$, $P, Q \in \Delta_N$,

then

$$M_g[P * Q]_f = M_g[P]_f \cdot M_g[Q]_f, \tag{3.11}$$

where the distribution

$$P * Q = (p_1q_1, ..., p_1q_N, ..., p_Nq_1, ..., p_Nq_N) \in \Delta_{N^2} \tag{3.12}$$

is the direct product of the distributions P and Q.

Now, if we consider two generalized probability distributions $P, Q \in \Delta_N$,that is, $0 < w(p) \leq 1, 0 < w(Q) \leq 1$,then according to the Definitions 3.3 and 3.4, we obtain

$$D(P;Q) = \log_2 M_g[P \mid Q]_f, \tag{3.13}$$

$$M_g[P \mid Q]_f = g^{-1}\left(\frac{\sum_{i=1}^{N} f(p_i)g\left(\frac{p_i}{q_i}\right)}{\sum_{i=1}^{N} f(p_i)}\right), \tag{3.14}$$

where

d_1) g is continuous and strict monotonic for positive values;

d_2) f is bounded and positive in $(0,1]$,
d_3) if $P = P_1 * P_2$ and $Q = Q_1 * Q_2$ and the elements of P_1 and Q_1 correspond to that of P_2 and Q_2 respectively, then

$$M_g[P_1 * P_2 \mid Q_1 * Q_2]_f = M_g[P_1 \mid Q_1]_f \cdot M_g[P_2 \mid Q_2]_f. \tag{3.15}$$

Now we present a measure of the homogeneity using the notion of the directed divergence.

Definition 3.5. [7] Let $P = (p_1, p_2, ..., p_N)$ be a complete distribution associated with a discrete finite random variable X which represents a common property for the all elements of a population $\mathbf{C}$. Then, the informational measure of the statistical homogeneity, relative to the complete probability distribution P, is defined by

$$I^*(P) = \frac{D^*(P;Q)}{H^*(P)}, \tag{3.16}$$

where

$$D^*(P;U) = H^*(U) - H^*(P), \tag{3.16a}$$

$$H^*(P) = -\sum_{i=1}^{N} p_i \log_2 p_i, \; H^*(U) = \log_2 N \tag{3.16b}$$

$$U = \left(\frac{1}{N}, \frac{1}{N}, \cdots, \frac{1}{N}\right), P = (p_1, p_2, ..., p_N), \; U, P \in \Delta_N^*. \tag{3.16c}$$

Theorem 3.1. The measure $I^*(P)$ has the following properties:
1^0 $I^*(U) = 0$;
2^0 $I^*(V) = 1$,where $V = (0, 0, ..., 1, 0, ..., 0)$, $V \in \Delta_N^*$;
3^0 $0 \leq I^*(P) \leq 1$.

Proof. Indeed, from the relation of the definition (3.16), we obtain

$$I^*(U) = \frac{H^*(U) - H^*(U)}{H^*(U)} = 0,$$

$$I^*(V) = \frac{D^*(V;U)}{H^*(V)} = \frac{\log_2 N}{\log_2 N} = 1,$$

and, from here, it follows and the property 3^0.

Corollary 3.1. If the complete directed divergence $D^*(P;U)$ can be write in the form

$$D^*(P;U) = H^*(U) - H^*(P) = \log_2 N - H^*(P), \tag{3.17}$$

then it induces a measure of the statistical homogeneity collectivity relative to a common property X, where X is a discrete finite random variable and $H^*(X)$ represents a measure of the amount of information contained by this random variable X.

Remark 3.1. If in the relations (3.13) and (3.14) we consider that the functions f and g have the forms

$$f(x) = x,\ g(x) = x^{\alpha-1}, \tag{3.18}$$

where $\alpha > 0, \alpha \neq 1$, then we obtain a new quantity

$$D_\alpha(P;Q) = \frac{1}{\alpha - 1} \log_2 \left(\frac{\sum\limits_{i=1}^{N} p_i^\alpha q_i^{1-\alpha}}{w(P)} \right), 0 < w(P) \leq 1. \tag{3.19}$$

Theorem 3.2. The quantity $D_\alpha(P;Q)$ represents the measure of the generalized directed divergence of order α of Rényi.

Proof. Let P be a generalized probability distribution, that is, $0 < w(P) \leq 1$, associated with the generalized discrete finite random variable X. Rényi [5] has defined the measure of information of order α, associated with the generalized probability distribution P, in the following form

$$H_\alpha(P) = \frac{1}{1 - \alpha} \log_2 \left(\frac{\sum\limits_{i=1}^{N} p_i^\alpha}{w(P)} \right), \alpha > 0, \alpha \neq 1. \tag{3.20}$$

Using the relation

$$\log_2 A = \log_2 \cdot \log_e A = k \cdot \log_e A, \tag{3.21}$$

we find

$$H_\alpha(P) = \frac{k}{1-\alpha}\log_e\left(\frac{\sum\limits_{i=1}^{N} p_i^\alpha}{w(P)}\right) = k\cdot\widetilde{H}_\alpha(P),\ k=\log_2 e \tag{3.22}$$

where

$$\overline{H}_\alpha(P) = \frac{1}{1-\alpha}\log_e\left(\frac{\sum\limits_{i=1}^{N} p_i^\alpha}{w(P)}\right). \tag{3.23}$$

Now, it is easy to see that to prove the Theorem 3.2 is equivalent with to prove the following relation

$$D_\alpha(P;U) = H_\alpha(U) - H_\alpha(P), \tag{3.24}$$

respectively, to prove that

$$D_\alpha(P;U) = k.\left[\overline{H}_\alpha(U) - \overline{H}_\alpha(P)\right]. \tag{3.25}$$

Indeed, we have

$$\begin{aligned}
\widetilde{H}_\alpha(U) - \widetilde{H}_\alpha(P) &= \frac{1}{1-\alpha}\log_e\left(\frac{\sum\limits_{i=1}^{N}\left(\frac{1}{N}\right)^\alpha}{\sum\limits_{i=1}^{N}\frac{1}{N}}\right) - \frac{1}{1-\alpha}\log_e\left(\frac{\sum\limits_{i=1}^{N} p_i^\alpha}{w(P)}\right)\\
&= \frac{1}{1-\alpha}\log_e N^{1-\alpha} - \frac{1}{1-\alpha}\log_e\left(\frac{\sum\limits_{i=1}^{N} p_i^\alpha}{w(P)}\right)\\
&= \frac{1}{\alpha-1}\log_e\left(\frac{\sum\limits_{i=1}^{N} p_i^\alpha N^{\alpha-1}}{w(P)}\right).
\end{aligned} \tag{3.26}$$

Therefore, we have the equality

$$k\left[\widetilde{H}_\alpha(U)-\widetilde{H}_\alpha(P)\right]=\frac{k}{\alpha-1}\log_e\left(\frac{\sum_{i=1}^{N}p_i^\alpha N^{\alpha-1}}{w(P)}\right),\alpha>0,\alpha\neq 1, \qquad (3.27)$$

respectively, the equality

$$D_\alpha(P;U)=H_\alpha(U)-H_\alpha(P), \qquad (3.28)$$

and hence it follows that indeed the quantity (3.19) represents just the generalized directed divergence of order α of Rényi.

Corollary 3.2. The measure of information of order α of Rényi, associated to the uniform probability distribution $U=\left(\frac{1}{N},\frac{1}{N},...,\frac{1}{N}\right)$, has the following value

$$H_\alpha(U)=\log_2 N. \qquad (3.29)$$

Theorem 3.3. The informational measure of order α of Rényi, for the statistical homogeneity of the collectivity, relative to the generalized random variable X, has the following form

$$I_\alpha(P)=\log_2 .H_\alpha^{-1}(P)-1,\ \alpha>0,\alpha\neq 1. \qquad (3.30)$$

Proof. If we have in view the Definition 3.4, we obtain

$$\begin{aligned} I_\alpha(P) &= \frac{D_\alpha(P;U)}{H_\alpha(P)}=\frac{H_\alpha(U)-H_\alpha(P)}{H_\alpha(P)}= \\ &= \frac{\log_2 w(P)-\log_2\left(\sum_{i=1}^{N}p_i^\alpha\right)+(1-\alpha)\log_2 N}{\log_2\left(\sum_{i=1}^{N}p_i^\alpha\right)-\log_2 w(P)}= \\ &= \log_2 N.H_\alpha^{-1}(P)-1,\alpha>0,\alpha\neq 1. \end{aligned}$$

Corollary 3.3. The measure $I_\alpha(P)$ has the following properties:

$$I_\alpha(U)=0 - \text{ the total homogeneity;} \qquad (3.31)$$

$$I_\alpha(V) = 1 \; - \text{ the total non-homogeneity,} \tag{3.32}$$

$$0 \leq I_\alpha(P) \leq 1. \tag{3.33}$$

Theorem 3.4. The measure $H_\alpha(P)$ satisfies the following relation

$$\lim_{\alpha \to 1} H_\alpha(P) = H_1(P), \tag{3.34}$$

where

$$H_\alpha(P) = \frac{1}{1-\alpha} \log_2 \left(\frac{\sum_{i=1}^{N} p_i^\alpha}{w(P)} \right) = k.\widetilde{H}_\alpha(P), \alpha > 0, \alpha \neq 1, \tag{3.35}$$

$$\widetilde{H}_\alpha(P) = \frac{1}{1-\alpha} \log_e \left(\frac{\sum_{i=1}^{N} p_i^\alpha}{w(P)} \right), \alpha > 0, \alpha \neq 1, \tag{3.36}$$

$$H_1(P) = k.\widetilde{H}_1(P), \tag{3.37}$$

$$\widetilde{H}_1(P) = -\frac{1}{w(P)} \sum_{i=1}^{N} p_i \log_e p_i, \tag{3.38}$$

$$k = \log_2 e. \tag{3.39}$$

Proof. Indeed, we have

$$\begin{aligned} \lim_{\alpha \to 1} \widetilde{H}_\alpha(P) &= \lim_{\alpha \to 1} \frac{1}{1-\alpha} \log_e \left(\frac{\sum_{i=1}^{N} p_i^\alpha}{w(P)} \right) = \left(\frac{0}{0} \right) = \\ &= -\frac{1}{w(P)} \sum_{i=1}^{N} p_i \log_e p_i = \widetilde{H}_1(P). \end{aligned}$$

Theorem 3.5. For the measure $D_\alpha(P;Q)$ we have

$$\lim_{\alpha\to 1} D_\alpha(P;Q) = D(P;Q), \tag{3.40}$$

where

$$D_\alpha(P;Q) = \frac{1}{\alpha-1}\log_2\left(\frac{\sum_{i=1}^{N} p_i^\alpha q_i^{1-\alpha}}{w(P)}\right), \alpha>0, \alpha\neq 1, \tag{3.41}$$

$$D(P;Q) = \frac{1}{w(P)}\sum_{i=1}^{N} p_i \log_2 \frac{p_i}{q_i}. \tag{3.42}$$

Proof. We have

$$D_\alpha(P;Q) = k.\widetilde{D}_\alpha(P;Q), \tag{3.43}$$

$$D(P;Q) = k.\widetilde{D}(P;Q), \tag{3.44}$$

where

$$\overline{D}_\alpha(P;Q) = \frac{1}{\alpha-1}\log_e\left(\frac{\sum_{i=1}^{N} p_i^\alpha q_i^{1-\alpha}}{w(P)}\right), \alpha>0, \alpha\neq 1, \tag{3.45}$$

$$\overline{D}(P;Q) = \frac{1}{w(P)}\sum_{i=1}^{N} p_i \log_e \frac{p_i}{q_i}, \tag{3.46}$$

$$k = \log_2 e. \tag{3.47}$$

From (3.45), when take the limit, we obtain

$$\lim_{\alpha\to 1} \widetilde{D}_\alpha(P;Q) = \widetilde{D}(P;Q), \tag{3.48}$$

and, hence, we find just the equality (3.40).

Corollary 3.4. We have

$$\lim_{\alpha\to 1} I_\alpha(P) = I_1(P), \tag{3.49}$$

where

$$I_\alpha(P) = \frac{D_\alpha(P;U)}{H_\alpha(P)}, \tag{3.50}$$

$$I(P) = \frac{D(P;U)}{H(P)}. \tag{3.51}$$

References

[1] Bechenbach, E.F., A class of mean value functions, Amer. Math. Monthly, 57(1950), pp.1-6.

[2] Daóczy,Z., Einige Ungleichungen über die mit Gewichtsfunktionen gebildeten Mittelwerte Monatah. Math. 68(1964), pp.102-112.

[3] Kapur,J.N., Entropic measure of economic inequality, Indian J.Pure Appl.Math., 17(3),(1986), pp. 273-285.

[4] Kullbach,S., Leibler,R.A., On information and sufficiency, Ann. Math. Stat. 22 (1951), pp.79-86.

[5] Rényi,A., On measure of entropy and information, Proc. 4-rth Berkeley Symp. Prob. Stat. 1(1961),pp.547-561.

[6] Shannon,C.E., The mathematical theory of communications, Bell System Tech. J., 27(1948), pp.379-423.

[7] Theil, H., Economics and information, North Holland, Amsterdam 1967.

[8] Văduva, I., Măsuri ale omogenităţii unei colectivităţi statistice, Studii şi cercetări de calcul economic şi cibernetică economică, 2(1987), pp.75-84.

Studiul informaţional al unei colectivităţi statistice

Rezumat

In lucrare se consideră o caracterizare informaţională a unei colectivităţi statistice relativ la o proprietate comună a elementelor ei.

Faculty of Mathematics and Informatics
„Babeş-Bolyai" University
Str. M. Kogălniceanu nr.1
3400 Cluj-Napoca
Romania

Received 12.07.1998