ENTROPY AND ECONOMETRICS. SIMULATION

Sergey Vladimirovich Yudin,

Doctor of Engineering, professor,

professor of the department of the Finances and Information Technology of Management,

Plekhanov Russian University of Economics, Tula branch,

Russian Federation, Tula

email: svjudin@rambler.ru

ABSTRACT

The principles of the construction and the analysis of complex processes based on the informational-statistics methods and the entropy statistics is offered. The new principles are universal and they allow to take into account any nonlinear and non-Gaussian processes.

Modeling of economic and technical processes using statistical "entropy" increases the accuracy and reliability of the results, thus reducing the risks of decisions.

The article describes an example of design and analysis of process information model, it is shown that substantially reduces the complexity of calculations in comparison with regression models.

Key words*: econometrics, entropy, simulation*

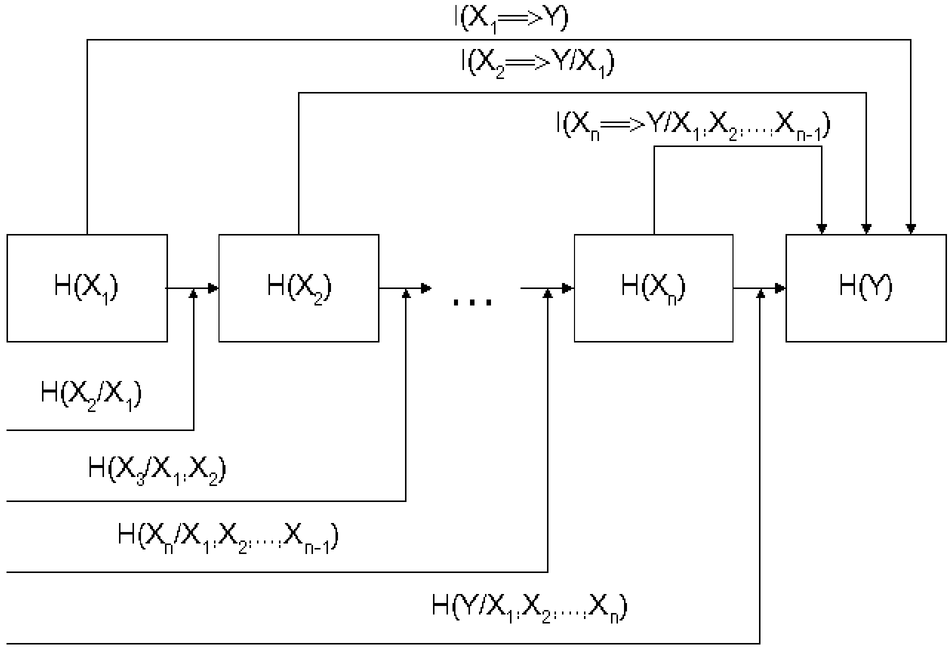
The informational method of simulation considered in this paper, is presented in the form of a solution of two tasks: 1) the analysis of hereditary effects in a line-up of relationships of cause and effect; 2) development of associations between factors and the indications simultaneously interacting with each other. The mathematical fundamentals of a solution of these tasks in machine industry have been developed for the analysis of engineering procedures in the eighties the XX-th centuries and generalised in the monography [4]. Essential principles of using of the entropy methods were based by works of such scientists as Basharin G.P. [1], Kullback S. [3]. The author applied statistics entropy at the decision of some tasks in mechanical engineering to the analysis of technological processes [5, 6].

I. *Simulation of hereditary effects*

The informational method of simulation of hereditary effects is grounded on introducing of a line-up of factors interacting sequentially against each other in the form of an information channel in which the information on the first factor arrives and sequentially will be transformed to the information on a total indication.

Let's consider a case when initially unique factor **X** sequentially will be transformed to **Y** (graf. 1).

Here $H(X_1), H(X_2), \ldots H(X_n)$ – an information quantity concluded in factor X after first, second, ... last operation; H (Y) – an information quantity concluded in indication Y; $I(X_k \Rightarrow Y/X_1, X_2, \ldots, X_{k-1})$ – an information quantity, transmitted to Y after working off *k* factors.



Graf. 1. The Information channel

The consecutive increment of the information is equal:

$$\begin{cases} I(X_1 \to Y) = H(Y) - H(Y/X_1) \\ I(X_2 \to Y/X) = H(Y/X_1) - H(Y/X_1X_2) \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ I(X_n \to Y/X_1 \ldots X_{n-1}) = H(Y/X_1 \ldots X_{n-1}) - H(X \ldots X_n), \end{cases} \quad (1)$$

Here H(Y) - an information quantity (entropy) about Y; $H(Y/X_1X_2\ldots)$ - the information quantity (entropy) received as a result of action on Y of various not considered factors. Since $H(Y|X_1X_2\ldots X_n) = H(Y, X_1X_2\ldots X_n) - H(X_1X_2\ldots X_n)$, then

$$\begin{cases} I(X_1 \to Y) = H(X_1) + H(Y) - H(X_1Y) \\ I(X_2 \to Y|X_1) = H(X_1X_2) - H(X_1X_2Y) - H(X_1) + H(X_1Y) \\ I(X_3 \to Y|X_1X_2) = H(X_1X_2X_3) - H(X_1X_2X_3Y) - H(X_1X_2) + H(X_1X_2Y) \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ I(X_n \to Y|X_1X_2..X_{n-1}) = H(X_1X_2..X_n) - \\ -H(X_1X_2..X_nY) - H(X_1X_2..X_{n-1}) + H(X_1X_2..X_{n-1}Y). \end{cases} \quad (2)$$

The level of influence of factor X on an indication Y at informational simulation can be evaluated by means of coefficient of informational connection q:

$$\begin{cases} q(X_1 \to Y) = I\left(X_1 \to Y\right) / H(Y)\right) \\ q(X_2 \to Y) = I\left(X_2 \to Y / X_1\right) / H(Y)\right) \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ q(X_n \to Y) = I\left(X_n \to Y / X_1 \ldots X_{n-1}\right) / H(Y)\right) \end{cases} \quad (3)$$

The coefficient of informational correlation is equal to unit if the information on an indication is completely defined by the information on factors; it is equal to zero if the indication does not depend on factors; generally the coefficient of informational correlation is concluded between zero and unit.

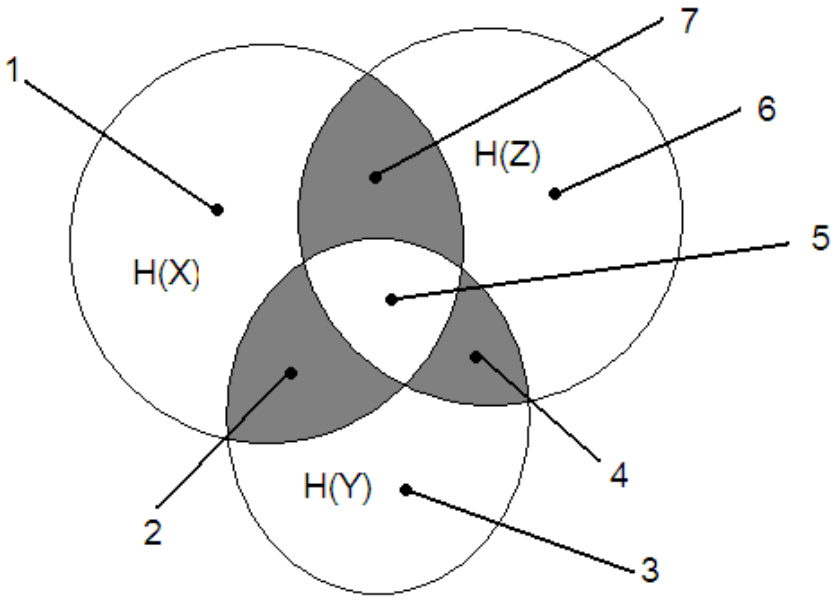
II. *Simulation of simultaneous processes*

Let the correlation between three factors-indications X, Y, Z is defined. On the basis of the chart of John Venn (graf. 2) is received:

$$\begin{cases} I(X \to Y) = H(X) + H(Y) - H(X,Y) \\ I(X \to Z) = H(X) + H(Z) - H(X,Z) \\ I(Y \to Z) = H(Y) + H(Z) - H(Y,Z) \\ I(XY \to Z) = H(X,Y) + H(Z) - H(X,Y,Z) \\ I(XZ \to Y) = H(XZ) + H(Y) - H(X,Y,Z) \\ I(YZ \to X) = H(YZ) + H(X) - H(X,Y,Z) \end{cases} \quad (4)$$

For the quantitative estimation of associations between parametres it is necessary to calculate coefficients of informational connection

$$\begin{cases} q(X \to Y) = I(X \to Y) / H(Y) \\ q(X \to Z) = I(X \to Z) / H(Z) \\ q(Y \to Z) = I(Y \to Z) / H(Z) \\ q(XY \to Z) = I(XY \to Z) / H(Z) \\ q(XZ \to Y) = I(XZ \to Y) / H(Y) \\ q(YZ \to X) = I(YZ \to X) / H(X) \end{cases} \quad (5)$$



Graf. 2. The chart of informational connection

Generalising the received outcomes on *n* parametres, association in between we will express the following formula:

$$I(X_1X_2,...,X_{n-1} \to X_n) = H(X_1X_2...X_{n-1}) + H(X_n) - H(X_1X_2...X_n). \quad (6)$$

III. *The Analysis of models*

At creation of models all theoretical values of entropies in the formulas reduced above are substituted by their estimations:

$$\hat{H}(X) = \sum_{i=1}^{k} \hat{p}_i \ln \hat{p}_i, \quad (7)$$

where $\hat{p}_i = f_i / n$ - empirical probability of hit of an aleatory variable X in a state number *i*; f_i - empirical frequency of hit of values X in this state; *n* - number of experiences.

It is displayed that the estimation of the information I (XY) →to within a constant factor has χ^2 allocation (see [2]):

$$2n\hat{I} = \chi^2_m \quad (8)$$

Here $m = (k_1 - 1)(k_2 - 1)$ - number of degree of freedoms; k_1, k_2 – an amount of intervals of a partition of input and output parametres accordingly.

The information transmitted from one parametre to another, is considered significant, if

$$2n\hat{I} \geq \chi^2_{m,\alpha} \quad (9)$$

Where $\chi^2_{m,\alpha}$ - α - a quantile χ^2 - allocations; α – a confidence level.

Allocation of the Pearson at *m*> 25 can be substituted Gaussian distribution with a variance

$\sigma^2=2m$ that gives the chance to define a confidence interval for the information:

$$\hat{I} - \frac{t_\alpha \sqrt{2m}}{2n} \le I \le \hat{I} + \frac{t_\alpha \sqrt{2m}}{2n} \qquad (10)$$

Value t_α - α-kvantil of a normal distribution. A confidence interval for coefficient of informational connection q is:

$$\hat{q} - \frac{t_\alpha \sqrt{2m}}{2nH(Y)} \le q \le \hat{q} + \frac{t_\alpha \sqrt{2m}}{2nH(Y)}. \qquad (11)$$

The minimum sample size is determined by means of necessary precision of ΔI value:

$$n_{\min} = \frac{t_\alpha}{\Delta I}\sqrt{\frac{m}{2}} \qquad (12)$$

In case of linear model the coefficient of correlation and coefficient of informational connection have a close connection among themselves, defined by statistical equality $q=r^2$ [1].

IV. *The Example of application of an informational model*

It is necessary to research association of labour productivity (Y) from a salary (X) (in percentage of basic value) (tab. 1).

The first step – creation of the chart of dispersion (graf. 3.).

Than we build the two-dimensional histogram. It is for this purpose defined the main statistical performances of researched aleatory variables (tab. 2) and we divide ranges of factors X and Y into the intervals which breadth is close to an average quadratic deviation.

Table 1

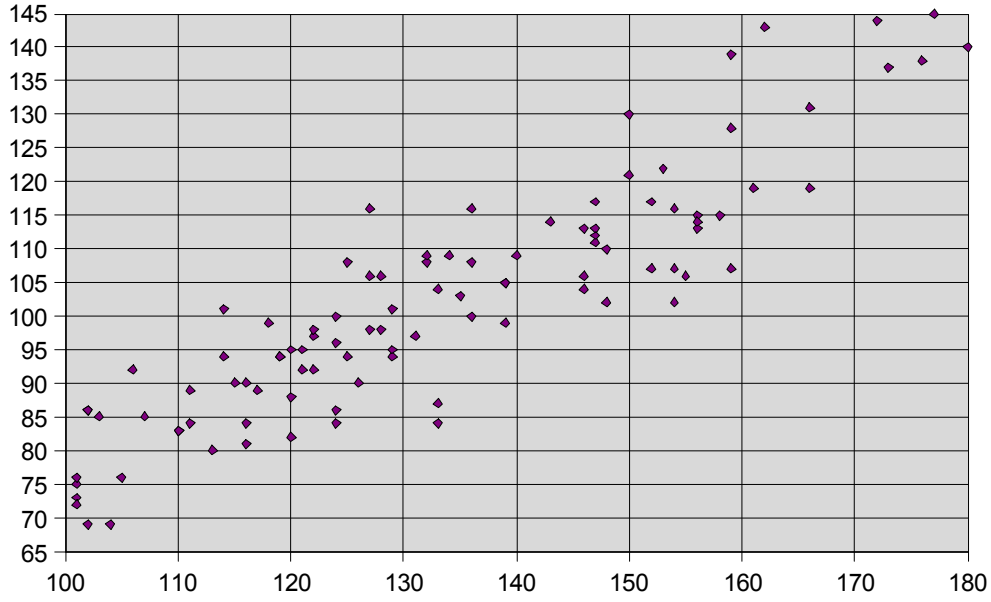
Input datas for model creation

№	X	Y	№	X	Y	№	X	Y	№	X	Y	№	X	Y
1	134	109	21	147	112	41	102	69	61	146	113	81	127	116
2	136	116	22	180	140	42	132	108	62	154	102	82	101	75
3	148	102	23	125	108	43	131	97	63	176	138	83	153	122
4	127	98	24	113	80	44	122	98	64	128	106	84	111	89
5	133	87	25	124	84	45	139	99	65	116	81	85	152	107
6	121	95	26	116	84	46	147	117	66	110	83	86	154	107
7	155	106	27	147	113	47	121	92	67	124	96	87	129	101
8	104	69	28	159	107	48	122	97	68	103	85	88	156	115
9	146	106	29	110	83	49	136	108	69	166	119	89	143	114
10	158	115	30	101	72	50	136	100	70	173	137	90	120	88
11	154	116	31	132	109	51	133	84	71	124	86	91	117	89

12	166	131	32	107	85	52	105	76	72	122	92	92	128	98
13	101	73	33	106	92	53	135	103	73	118	99	93	139	105
14	129	94	34	152	117	54	133	104	74	159	128	94	148	110
15	102	86	35	124	100	55	111	85	75	161	119	95	146	104
16	119	94	36	120	82	56	116	90	76	172	144	96	156	113
17	156	114	37	126	90	57	140	109	77	139	105	97	101	76
18	150	121	38	127	106	58	159	139	78	125	94	98	129	95
19	177	145	39	150	130	59	162	143	79	114	101	99	102	86
20	147	111	40	114	94	60	115	90	80	120	95	100	119	94

The third step is an evaluation of entropies H (X), H (Y), H (X, Y).

$$\begin{cases} H(X) = -\sum_{i=1}^{k_1} \frac{f_i(x)}{n} \cdot \ln\left(\frac{f_i(x)}{n}\right) = 1{,}300196 \left(f_i(x) \neq 0\right) \\ H(Y) = -\sum_{j=1}^{k_2} \frac{f_j(y)}{n} \cdot \ln\left(\frac{f_j(y)}{n}\right) = 1{,}396325 \left(f_j(y) \neq 0\right) \\ H(X,Y) = -\sum_{i=1}^{k_1} \sum_{j=1}^{k_2} \frac{f_{ij}}{n} \cdot \ln\left(\frac{f_{ij}}{n}\right) = 2{,}134019 \left(f_{ij} \neq 0\right) \end{cases}$$



Graf. 3. A field of dispersion of the experimental observations of association of labour productivity (Y) from a salary (X) (in percentage of basic value).

Table 2

The main statistical performances

	X	*Y*
Average	133,21	102,1
Standard deviation	20,27154	17,34062
Sampling variance	410,9353	300,697
Minimum	101	69
Maxima	180	145

In tab. 3 frequencies of hit of values of a two-dimensional aleatory variable in appropriate intervals are reduced.

Table 3

Two-dimensional bar graph

Y	X				f (y)
	100-120	120-140	140-160	160-180	
69-86	17	4			21
86-103	13	17	3		33
103-120		12	20	2	34
120-137			5	2	7
137-145				5	5
f (x)	30	33	28	9	

The mutual information is equal I(X→Y)=H(X)+H(Y)-H(X,Y)=0,562502 , and coefficient of informational connection q(X→Y)= I(X→Y)/H(Y)=0,402844.

The fourth step. An estimation of significance of the discovered connection by criterion of the Pearson (8). In our case k_1=4, k_2=5. Calculated value of Pearson criterion of the Pearson is equal to $\chi^2_{расч} = 2nI = 112{,}503$. Table value at number of degree of freedoms m = (4-1 (5-1) =12 ·and a fiducial probability α=0,95 is equal to $\chi^2_{12;0,95} = 21{,}02606$. Since a calculated value of Pearson criterion more than table connection between Y and X it is significant.

Thus, in paper the technique of simulation which is based on methods of the information theory is offered and justified. The example of creation of an informational model is reduced.

REFERENCES:

1. Basharin G.P. About a statistical estimation of entropy of independent random

variables//Probability theory and its applications.-1956, т. IV, № 3. - With. 361-364

2. Eye, A. von. On the Equivalence of the Information-Theoretic Transmission Measure to the Common χ2-Statistic. – “Biom. J.”, v. 22, 1925, p.p. 700-725.

3. Kullback S. Information Theory and Statistics. New York – John Willey & Songs, Inc. – 1965.

4. Yudin S.V., Yudin A.S. Informational-Statistics Methods of solution econometrical, sociological and psychometric problems. – Tula: Publishing house of the Tula state university, 2010. – 124 p.p.

5. Yudin S.V., Grigorovich V. G, Yudin A.S. Information-statistical methods of an estimation of quality of a flow of repousses in the conditions of acceptance testing//Blacksmith's forming production. Handling of metals by pressure.

6. Yudin S. The Informational Criterion of Identification of the Distribution // Modern European Researches. - 2015. - № 4. - C. 133-137