НАУЧНОЕ ИЗДАНИЕ МГТУ ИМ. Н. Э. БАУМАНА

НАУКА и ОБРАЗОВАНИЕ

Эл № ФС77 - 48211. Государственная регистрация №0421200025. ISSN 1994-0408

электронный научно-технический журнал

Информационно-статистические методы анализа случайных процессов

77-48211/516648

01, январь 2013

Юдин С. В.

УДК 51-77:330.4:519.21

Тульский филиал Российского государственного торгово-экономического университета

svjudin@rambler.ru

Введение

На протяжении нескольких десятилетий в экономическую науку, в целом, и в эконометрику, в частности, широко внедряются методы исследований и математический аппарат, применяемые в физике. В частности, используется такое понятие как «энтропия».

Этому подходу посвящено много публикаций, среди которых следует отметить работы таких ученых, как В. Бурлачков [2], E.T. Jaynes [14], R. Kümmel [15], В.В. Глущенко [4], С.Д. Хайтун [9] и другие. Автор также применял подобный подход в исследованиях процессов контроля и управления в машиностроении [11].

Предложенный метод проводит аналогию между экономическими и термодинамическими процессами, что, с одной стороны, позволяет добиться новых интересных результатов, имеющих практическое значение, но, с другой стороны, не всегда позволяет получить интересующие практиков зависимости и оценить значимость моделей. Автор предлагает новый подход, базирующийся на методах теории информации, основанной К. Шенноном [10] и Н. Винером [3].

Постановка задачи

Как известно, практически все статистические методы исследования взаимосвязей предполагают, что модель является линейной либо может быть сведена к ней и все

переменные, входящие в нее, имеют нормальное распределение. Если же эти допущения отсутствуют, то практически невозможно оценить значимость коэффициентов модели и получить надежные доверительные интервалы.

Еще в 1975 году Н.С. Райбман и В.М. Чадеев [8] показали, что корреляционные модели не дают надежных результатов. Более того, остаточная дисперсия, являющейся основой для оценки значимости параметров влияния факторов на признак, никоим образом не может служить мерой взаимодействия. Также возникают проблемы с определением законов распределения случайных величин. Как правило, в этих задачах используют критерий Пирсона, хотя структура его такова, что он дает достоверные результаты только при решении вопроса о нормальности распределения.

Толчок к широкому применению методов теории информации автор получил при изучении работ Г.П. Башарина [1] и С. Кульбака [5].

В предисловии к книге Кульбака [5] академик А.Н.Колмогоров писал: «... аналитический аппарат теории информации был создан тогда, когда здание математической статистики было в своих основных, находящих наиболее широкое применение, частях уже построено и кодифицировано. Но новые мысли и аналитический аппарат теории информации должны, по-видимому, привести к заметной перестройке этого здания».

Можно отметить, что 90-м годам XX века сформировался новый подход в разных областях науки и техники, который можно назвать «энтропийным подходом». Теория информации позволяет унифицировать известные результаты теории статистических выводов, что наглядно показано в работе С. Кульбака [5].

Анализ литературы, посвященной информационной теории управления, приводит к выводу, что «... информационный подход дает единую точку зрения на все виды управления, независимо от его цели и типа управления системы» [7]. Теория информации, как и статистическая физика, благодаря своим методам и обобщениям позволяет исследовать объекты сложной природы на относительно простых и наглядных математических моделях.

Изначально энтропия рассматривалась как мера величины, характеризующей процессы перехода тепловой энергии в механическую. Связь между энтропией как мерой неопределенности и термодинамической энтропией достаточно долго оставалась неясной, но в последней четверти XX века она была установлена [6].

Рассмотрим некоторые соотношения из теории информации, которые мы будем использовать в дальнейшем.

Энтропия, как мера связи

Пусть некоторая система имеет дискретный набор состояний $x_1, x_2, \ldots, x_k$, которые она может принимать с вероятностями $p_1, p_2, \ldots, p_k$. Тогда, согласно К. Шеннону [10], мера неопределенности или энтропия системы имеет следующее числовое значение

$$H = h = -\sum_{i=1}^{k} p_i \ln p_i \quad , \quad (1)$$

Вероятности, используемые в выражении (1), как правило, определяются опытным путем, следовательно, при расчетах, вместо точных значений приходится использовать эмпирические оценки, рассчитанные через частоты наблюдений соответствующих состояний, т.е. в качестве оценки энтропии будет использоваться величина

$$H^* = -\sum_{i=1}^{k} p_i^* \ln p_i^* , \quad (2)$$

где $p_i^* = \frac{f_i}{n}$- частости; f_i – частоты наблюдений соответствующих состояний; n – количество наблюдений (объем выборки).

Как было показано Г.П. Башариным [1] при достаточно общих предположениях, в том числе и в предположении о стремлении количества состояний к бесконечности, статистическая оценка энтропии (2) имеет асимптотически нормальное распределение с параметрами

$$\begin{cases} \mathbf{M(H^*)} = h - \dfrac{k-1}{n} \\ \mathbf{D(H^*)} = \dfrac{a^2 - h^2}{n} \end{cases} \quad (3)$$

Здесь

$$a^2 = \sum_{i=1}^{k} p_i \ln^2 p_i \quad (4)$$

Параметры a^2 и h в дальнейшем будем называть энтропийными параметрами распределения.

Приведенные выше соотношения справедливы для дискретных случайных величин, в то время как на практике мы имеем дело с непрерывными распределениями, в связи с чем приходится проводить процедуру дискретизации.

Пусть *W(x)* – функция плотности вероятности некоторой случайной величины **X** с дисперсией σ^2.

Нормируем случайную величину **X** и введем новую случайную величину $\tilde{X} = X / \sigma$, имеющую функцию плотности *w(x)*. Разобьем область изменения случайной величины на интервалы шириной Δx. Пронумеруем эти интервалы от нижней до верхней границы числами натурального ряда от 1 до *k*. Пусть вероятности попадания в каждый интервал равны соответственно $p_i, i = 1...k$. Введем энтропийные параметры непрерывного распределения следующим образом:

$$\begin{cases} h = -\int_{-\infty}^{\infty} w(x) \ln w(x) dx \approx -\sum_{i=1}^{k} p_i \ln p_i + \ln \frac{\Delta x}{\sigma} \\ a^2 = \int_{-\infty}^{\infty} w(x) \ln^2 w(x) dx \approx \sum_{i=1}^{k} p_i \ln^2 p_i + 2h \ln \frac{\Delta x}{\sigma} - \ln^2 \frac{\Delta x}{\sigma} \end{cases} \quad (5)$$

Чем меньше ширина интервалов Δx, тем точнее равенства в формуле (5).

Эмпирическая энтропия непрерывных распределений, прошедших процедуру дискретизации, найденная по результатам опытов, имеет нормальное распределение с параметрами, вычисленными по формулам (3), (4) с учетом (5).

Помимо вышеуказанного свойства эмпирической энтропии, что позволяет ввести новый критерий идентификации вида закона распределения [11], методы теории информации дают возможность построения адекватных моделей нелинейных процессов.

Рассмотрим двумерную случайную величину **Z=(X,Y)**. Пусть **X** - входной параметр, а **Y** – выходной параметр. Дискретизируем области изменения одномерных случайных величин **X** и **Y**. Пусть p_{xi} (i=1…k_1) – вероятности попадания значений случайной величины **X** в соответствующие интервалы, а p_{yj} (j=1…k_2) – то же для **Y**. Обозначим через p_{ij} вероятность попадания случайной величины **Z** в соответствующую клетку.

Вычислим энтропии всех трех величин:

$$\mathbf{H(X)} = -\sum_{i=1}^{k_1} p_{xi} \ln p_{xi}; \ \mathbf{H(Y)} = -\sum_{j=1}^{k_2} p_{yj} \ln p_{yj}; \ \mathbf{H(Z)} = -\sum_{j=1}^{k_2} \sum_{i=1}^{k_1} p_{ij} \ln p_{ij} \quad (6)$$

Количество информации, передаваемое от входного параметра **X** выходному параметру **Y**, равно

$$\mathbf{I(X \to Y) = H(X) + H(Y) - H(Z)} \quad (7)$$

Можно показать, что в случае статистической независимости случайных величин **X** и **Y** энтропия двумерной величины **Z** равна сумме энтропий одномерных величин, а в случае детерминированной монотонной зависимости одномерных величин все три энтропии равны. На основании этого можно ввести параметр «коэффициент информационной связи»

$$q = \frac{\mathbf{I(X \to Y)}}{\mathbf{H(Y)}}, \tag{8}$$

который равен нулю при статистической независимости одномерных величин и равен единице при детерминированной монотонной связи.

В работах F. Attneave [12] и A. von Eye [13] показано, что эмпирическая информация $\mathbf{I}^*(\mathbf{X \to Y})$ с точностью до постоянного множителя имеет χ^2-распределение:

$$2n\mathbf{I}^* = \chi^2_m \tag{9}$$

Здесь $m = (k_1\text{-}1)(k_2\text{-}1)$ - количество степеней свободы; k_1, k_2 - количество интервалов разбиения входного и выходного параметров соответственно; n - объем выборки.

Информация, передаваемая от одного параметра к другому, считается значимой, если

$$2n\mathbf{I}^* \geq \chi^2_{m,\alpha}, \tag{10}$$

где $\chi^2_{m,\alpha}$ – α-квантиль χ^2_m-распределения; α – доверительная вероятность.

В этом случае считается также значимым и коэффициент информационной связи $\mathbf{q}^*$.

Пример построения информационной модели

По данным исследования зависимости производительности труда (Y) от заработной платы (X) (в процентах от базовой величины) получены следующие данные (табл. 1).

Таблица 1.

№	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	8n	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

Первый шаг – построение диаграммы рассеивания (рис. 1).

На рис. 1 по горизонтали отложены значения фактора X, а по вертикали – признака Y. Визуальный анализ дает основание утверждать, что нет аномальных наблюдений, все точки не сильно удалены от других.

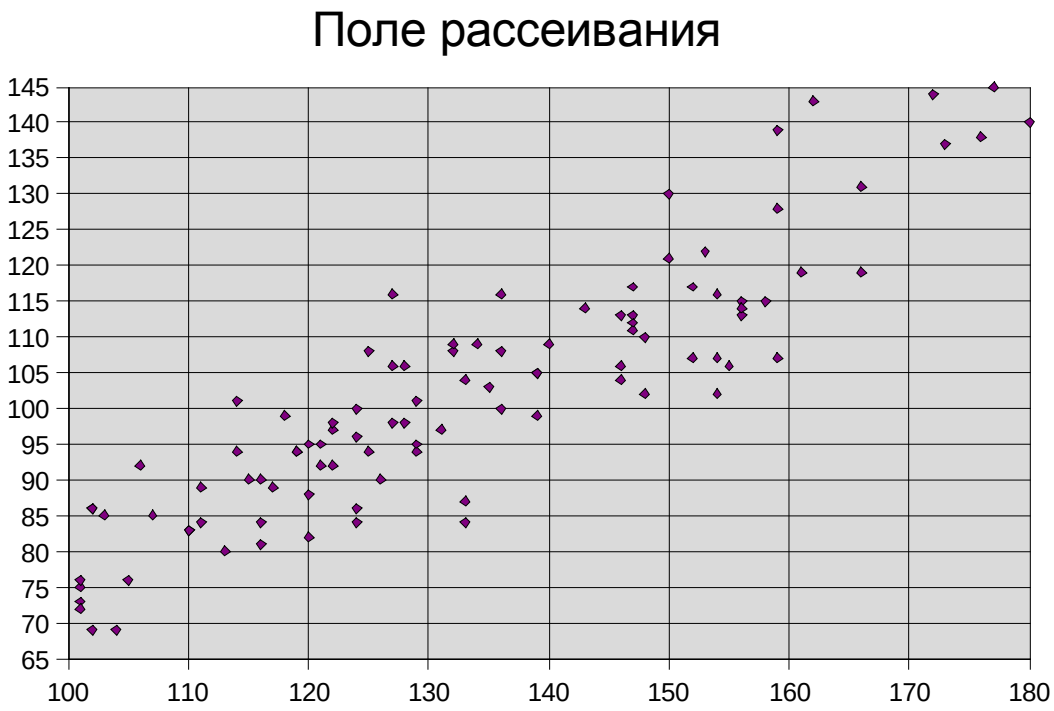

Рис. 1. Поле рассеивания экспериментальных наблюдений зависимости производительности труда (Y) от заработной платы (X) (в процентах от базовой величины).

Первый шаг. Строим двумерную гистограмму. Для этого разбиваем области изменения фактора X и признака Y на интервалы, ширина которых близка к среднему квадратическому отклонению.

В табл. 2 приведены основные статистические характеристики исследуемых случайных величин.

В табл. 3 приведены частоты попадания значений двумерной случайной величины в соответствующие интервалы.

Таблица 2

Основные статистические характеристики

	X	*Y*
Среднее	133,21	102,1
Стандартное отклонение	20,27154	17,34062
Дисперсия выборки	410,9353	300,697
Минимум	101	69
Максимум	180	145

Таблица 3.

Двумерная гистограмма.

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	

Второй шаг. Вычисляем энтропии H(X), H(Y), H(X,Y).

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

В результате расчетов получаем:

H(X)=1,300196; H(Y)=1,396325; H(X,Y)=2,134019.

Взаимная информация равна I(X→Y)=H(X)+H(Y)-H(X,Y)=0,562502.

Коэффициент информационной связи равен

q(X→Y)= I(X→Y)/H(Y)=0,402844.

Третий шаг. Оценка значимости найденной связи.

Значимость связи оценивается по критерию Пирсона χ^2.

Взаимная информация с точностью до множителя $2n$ имеет распределение Пирсона: $2nI = \chi^2_{(k_1-1)\cdot(k_2-1)}$, где k_1 – число интервалов разбиения области изменения случайной величины X, а k_2 – число интервалов разбиения области изменения случайной величины Y. В нашем случае k_1=4, k_2=5.

Итак, расчетное значение критерия Пирсона равно $\chi^2_{расч} = 2nI = 112{,}503$.

Табличное значение при числе степеней свободы m=(4-1)·(5-1)=12 и доверительной вероятности α=0,95 равно $\chi^2_{12;0,95} = 21{,}02606$.

Т.к. расчетное значение критерия Пирсона больше табличного, то связь между признаком Y и фактором X значима.

Заключение

Таким образом, в статье обоснована возможность использования статистики «энтропия» для решения разнообразных задач эконометрики.

В дальнейших публикациях автора предполагается рассмотреть информационный критерий идентификации вида закона распределения, проблемы построения и анализа информационных моделей, решение задач дискриминации и идентификации.

Список литературы

1. Башарин Г.П. О статистической оценке энтропии независимых случайных величин // Теория вероятностей и ее применения. 1956. Т. IV, № 3. С. 361-364

2. Бурлачков В. Экономическая наука и эконофизика: главные темы диалога // Вопросы экономики. 2007. № 12. С. 111-122.

3. Винер Н. Кибернетика, или управление и связь в животном и машине : пер. с англ. М.: Наука, 1983. 340 с.

4. Глущенко В.В. Использование спирали развития в виде n - кратно циклической би-спирали информации в борьбе с энтропией и хаосом в системных исследованиях экономики и управления // Управление экономическими системами. 2011. № 12. Режим доступа: http://uecs.ru/uecs-36-122011/item/852--n- (дата обращения 12.12.2012).

5. Кульбак С. Теория информации и статистика. М.: Наука, 1967. 408 с.

6. Мартин Н., Ингленд Дж. Математическая теория энтропии. М.: Мир, 1988. 350 с.

7. Петров Б.Н., Петров В.В., Уланов Г.М., Агеев В.М., Запорожец А.В., Усков А.В., Кочубиевский И.Д. Начала информационной теории управления // Техническая кибернетика. 1970. № 3. С. 10-14.

8. Райбман Н.С., Чадеев В.М. Построение моделей процессов производства. М.: Энергия, 1975. 376 с.

9. Хайтун С.Д. Социальная революция, энтропия и рынок // Общественные науки и современность. 2000. № 6. С. 94-109.

10. Шеннон К. Работы по теории информации и кибернетике : сб. статей : пер. с англ. М.: Иностранная литература, 1963. 829 с.

11. Юдин С.В. Информационный анализ // Известия Тульского государственного университета. Сер. Математика. Механика. Информатика. 1995. Т. 1, вып. 3. С. 136-145.

12. Attneave F. Information theory in der Psychologie. 2 Aufl. Bern, Stuttgart, Wienn. 1991.

13. Eye A. von. On the Equivalence of the Information-Theoretic Transmission-Measure to the Common 7c 52 0-Statistic // Biom. J. 1982. Vol. 24. P. 391-398.

14. Jaynes E.T. How Should We Use Entropy In Economics? St. John's College, Cambridge. Available at: http://bayes.wustl.edu/etj/articles/entropy.in.economics.pdf , accessed 12.12.2012.

15. Kümmel Reiner. The Second Law of Economics: Energy, Entropy, and the Origins of Wealth. New York: Springer, 2011. 293 p.

SCIENTIFIC PERIODICAL OF THE BAUMAN MSTU

SCIENCE and EDUCATION

EL № FS77 - 48211. №0421200025. ISSN 1994-0408

electronic scientific and technical journal

Informational and statistical methods of random process analysis

77-48211/516648

01, January 2013

Yudin S.V.

Tula branch of the Russian State Trade and Economic University

svjudin@rambler.ru

A new approach to analysis of random processes, based on applying information theory methods and new “distribution entropy” statistics, is proposed in this paper. Use of information-theoretical methods was justified when solving several statistical problems such as tests of statistical hypothesis, analysis of process state, and so forth. It was shown that entropy is a universal statistical estimate which allows to solve almost every simulation problems and analyze complex process, including identification of the distribution law and solving discrimination problems. The main advantage of this method is a possibility of considering non-linear models with non-Gaussian factors. An example of creation and analysis of informational model of labor productivity dependence on the salary is considered.

Publications with keywords: entropy, modelling, analyses, information theory, econometrics
Publications with words: entropy, modelling, analyses, information theory, econometrics

References

1. Basharin G.P. O statisticheskoi otsenke entropii nezavisimykh sluchainykh velichin [On the statistical estimation of the entropy of independent random variables]. *Teoriia veroiatnostei i ee primeneniia* [Probability theory and its applications], 1956, vol. 4, no. 3, pp. 361-364

2. Burlachkov V. Ekonomicheskaia nauka i ekonofizika: glavnye temy dialoga [Economics and Econophysics: The main topic of dialogue]. *Voprosy ekonomiki*, 2007, no. 12, pp. 111-122.

3. Wiener N. *Cybernetics: or, Control and Communication in the Animal and the Machine*. 2nd ed. New York-London, The M.I.T. Press and John Wiley & Sons, Inc., 1961. (Russ. ed.: Viner N. *Kibernetika, ili upravlenie i sviaz' v zhivotnom i mashine*. Moscow, Nauka, 1983. 340 p.).

4. Glushchenko V.V. Ispol'zovanie spirali razvitiia v vide n - kratno tsiklicheskoi bi-spirali informatsii v bor'be s entropiei i khaosom v sistemnykh issledovaniiakh ekonomiki i upravleniia [The use of a spiral of development in the form of the n-fold cyclic bi-spiral of information in the struggle against entropy and chaos in the system studies of economics and management]. *Upravlenie ekonomicheskimi sistemami* [Management of economic systems], 2011, no. 12. Available at: http://uecs.ru/uecs-36-122011/item/852--n- , accessed 12.12.2012.

5. Kul'bak S. *Teoriia informatsii i statistika* [Information theory and statistics]. Moscow, Nauka, 1967. 408 p.

6. Martin N.F.G., England J.W. *Mathematical theory of entropy. Encyclopedia of Mathematics and its Applications*, vol. 12. Addison-Wesley Publishing Co., Reading, Mass., 1981. 257 p. (Russ. ed.: Martin N., Inglend Dzh. *Matematicheskaia teoriia entropii*. Moscow, Mir, 1988. 350 p.).

7. Petrov B.N., Petrov V.V., Ulanov G.M., Ageev V.M., Zaporozhets A.V., Uskov A.V., Kochubievskii I.D. Nachala informatsionnoi teorii upravleniia [Principles of information management theory]. *Tekhnicheskaia kibernetika*, 1970, no. 3, pp. 10-14.

8. Raibman N.S., Chadeev V.M. *Postroenie modelei protsessov proizvodstva* [The construction of models of processes of production]. Moscow, Energiia, 1975. 376 p.

9. Khaitun S.D. Sotsial'naia revoliutsiia, entropiia i rynok [The social revolution, the entropy and the market]. *Obshchestvennye nauki i sovremennost'* [Social sciences and modernity], 2000, no. 6, pp. 94-109.

10. Shennon K. *Raboty po teorii informatsii i kibernetike: sb. statei* [Works on information theory and Cybernetics : collection of articles]. Moscow, Inostrannaia literatura, 1963. 829 p. (Russian translation).

11. Iudin S.V. Informatsionnyi analiz [Information analysis]. *Izvestiia Tul'skogo gosudarstvennogo universiteta. Ser. Matematika. Mekhanika. Informatika* [News of the Tula state University. Ser. Mathematics. Mechanics. Informatics], 1995, vol. 1, no. 3, pp. 136-145.

12. Attneave F. *Informationtheory in der Psychologie*. 2 Aufl. Bern, Stuttgart, Wienn. 1991. (in German)

13. Eye A. von. On the Equivalence of the Information-Theoretic Transmission-Measure to the Common 7c 52 0-Statistic. *Biom. J.*, 1982, vol. 24, pp. 391-398.

14. Jaynes E.T. *How Should We Use Entropy in Economics?* St. John's College, Cambridge. Available at: http://bayes.wustl.edu/etj/articles/entropy.in.economics.pdf , accessed 12.12.2012.

15. Kümmel Reiner. *The Second Law of Economics: Energy, Entropy, and the Origins of Wealth*. New York, Springer, 2011. 293 p.