Time series: entropy and informational energy

George Daniel Mateescu[1]

Abstract. In the present work, we propose to investigate the possibility of using, in the analysis of data series, some notions similar to the entropy and the informational energy of a random distribution. Using these tools it is possible to appreciate some characteristics of the data series, exemplified in the article by the linear model.

Key words: entropy, frequency, histogram, informational energy

JEL Codes: C22, C40

Context. We remember that a probability distribution is given

$$(p_i), i = 1..m$$

then the Shannon entropy is:

$$\sum_{i=1..m} -p_i log(p_i)$$

The main property ([1]) is the maximum value, which is reached in the case of equal probabilities,

$$p_i = \frac{1}{m}, i = 1..m, \sum_{i=1..m} -\frac{1}{m} log\left(\frac{1}{m}\right) = \log(m)$$

In the case of a deterministic distribution, with a single value $p = 1$, the entropy is 0.

Starting from this definition, we consider a series of data:

$$(y_i)_{i=0..n}$$

whose values are in the interval $(a, b]$.

[1] Institute of Economic Forecasting, Romanian Academy, e-mail: dan@mateescu.ro

We consider a division of it

$$a = a_0 < a_1 < \cdots < a_m = b$$

as well as the associated frequencies:

$$z_i = card\{y_k | y_k \in (a_{i-1}, a_i], k \in [0, \ldots n]\}, i = 1..m$$

and

$$Z = \sum_{i=1..m} z_i$$

We put by definition

$$p_i = \frac{z_i}{Z}, i = 1..m$$

Therefore, we can associate entropy with the data series ([1] Claude Shannon, *A Mathematical Theory of Communication*):

$$\sum_{i=1..m} -p_i log(p_i)$$

in which we define $p_i log(p_i) = 0$ if $p_i = 0$ (because $\lim_{x \to 0} x log(x) = 0$).

Time series

For such data series, the order of values is significant, the data are ordered. Consequently, it will be necessary to use some derived indicators. Such indicators are anyway used to eliminate, for example, systematic errors.

We will exemplify by two ways of exploring the data order, respectively by arithmetic growth and by proportional growth, that is:

$$u_i = y_i - y_{i-1}, i = 1..n$$

and

$$v_i = \frac{y_i}{y_{i-1}}, i = 1..n, y_{i-1} \neq 0$$

The periodic case. If the data series $(y_i)_{i=0..n}$ is periodic, then the histogram relative to an equidistant division has values proportional to those relative to a period. Consequently, the entropy of the entire data series is equal to the entropy relative to a period. Likewise, the data series derived by arithmetic growth or by proportional growth are periodic.

Indeed, if the data series is periodic, of period t, that is:

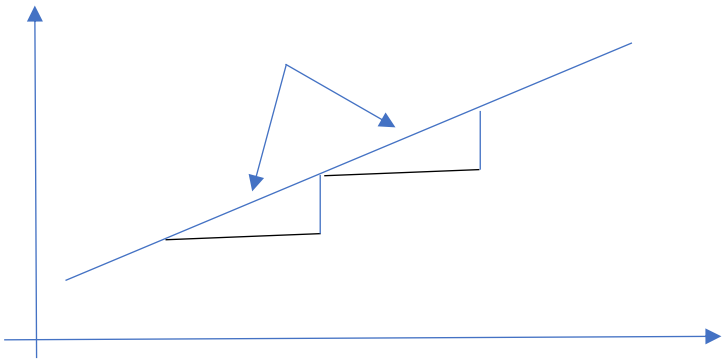
$$y_i = y_{i-t}$$

then arithmetic and proportional growth are also periodic:

$$y_i - y_{i-1} = y_{i-t} - y_{i-1-t}$$

$$\frac{y_i}{y_{i-1}} = \frac{y_{i-t}}{y_{i-1-t}}$$

The linear case. If the data series $(y_i)_{i=0..n}$ is linear, then the histogram relative to an equidistant division with *m* intervals has constant values



consequently the probabilities $(p_i)_{i=1..m}$ are equal, and the entropy is maximum. Instead, the data series derived by arithmetic growth has constant values, and the histogram has only one non-zero value, so the entropy is 0.

Indeed, if the data series is of the form:

$$y_i = \alpha x_i + \beta$$

and

$$x_i - x_{i-1} = h(constant)$$

then

$$y_i - y_{i-1} = \alpha h$$

The exponential case. If the data series $(y_i)_{i=0..n}$ follows an exponential curve, for example having the points (values) on a shape curve

$$y_i = \alpha e^{tx_i}, x_i - x_{i-1} = h$$

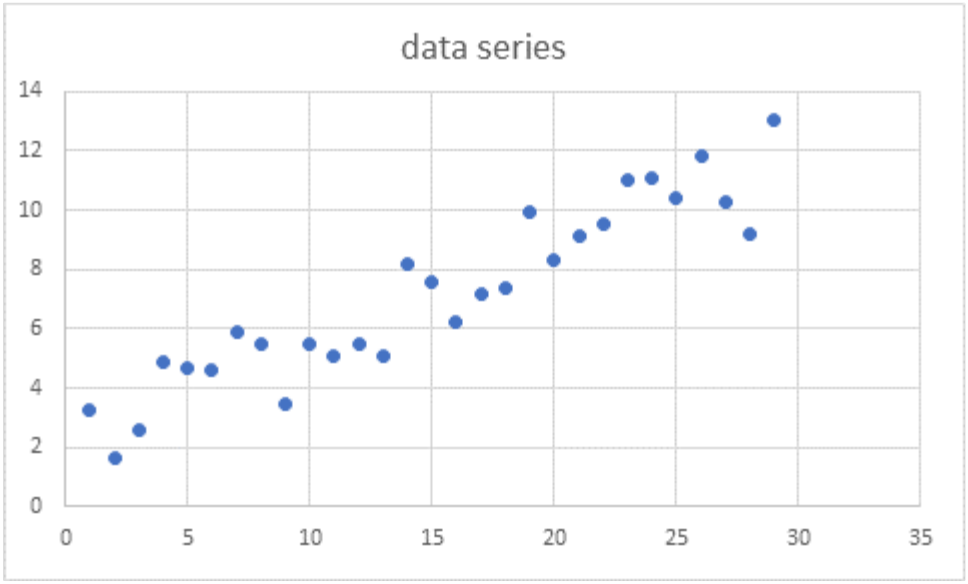
then

$$\frac{y_i}{y_{i-1}} = \frac{\alpha e^{tx_i}}{\alpha e^{tx_{i-1}}} = e^{th}$$

That is, the data series derived by proportional growth is constant, having 0 entropy.

Example, the linear model

3.286752
1.640412
2.593846
4.914973
4.704256
4.619854
5.921308
5.479213
3.448796
5.506525
5.100402
5.478127
5.05552
8.155143
7.595632
6.20528
7.204173
7.399293
9.944258
8.30864
9.142633
9.560191
11.02534
11.07022
10.39321
11.83122
10.29774
9.199645
13.05555



The minimum value is 1.640411 and the maximum is 13.055548 and accordingly, we consider an

equidistant division through 9 intervals, which corresponds to the nodes:

1.640312
2.90876
4.177109
5.445457
6.713806
7.982154
9.250503
10.51885
11.7872
13.05555

and the frequencies

2
2
5
5
3
4
4
2
2

For entropy calculation we will have:

probability	logarithm
0.068966	-0.08009
0.068966	-0.08009
0.172414	-0.13163
0.172414	-0.13163
0.103448	-0.10193
0.137931	-0.11867
0.137931	-0.11867
0.068966	-0.08009
0.068966	-0.08009

Finally, the entropy is 0.922888, close to the maximum entropy which is log(9)=0.954243,

corresponding to the linear model.

Informational energy.

The notion was introduced by O. Onicescu in 1966 and is defined by

$$E(p) = \sum_{i=1..m} p_i^2$$

where $(p_i), i = 1..m$ is a probability distribution

Some useful proprieties are ([2]):

$$\frac{1}{m} \leq E(p) \leq 1$$

and: entropy and informational energy are inversely proportional

In the case of the previous example, the informational energy is 0.127229, close to the minimum value which is $\frac{1}{9} = 0.111111$.

Both criteria, entropy and informational energy, show that the data series in the example corresponds to a linear model.

Bibliography

[1] Claude Shannon, *A Mathematical Theory of Communication*, Bell System Technical Journal 1948

[2] O.Onicescu, V.Stefanescu, *Elemente de statistica informationala cu aplicatii*, Editura Tehnica 1979