Is the entropy a good measure of correlation?

Anita Dobek, Krzysztof Moliński,
Ewa Skotarczak

Poznań Univeristy of Life Sciences
Wojska Polskiego 28, 60-637 Poznań

Będlewo, 2016

Introduction

In the life sciences there are many traits which can be observed only in a categorical scale but are determined by many factors including genetic and environmental components, for example fertility, calving difficulty, resistance to diseases or resistance of pathogenic bacteria to different antibiotics.

It is natural to suppose that the categorical phenotype of those traits is determined by a continuous, unobservable variable, often called liability.

Introduction

For example, when we observe only two categories (success or failure), the relation between the categorical and the continuous variables is as follows: we can notice the success when the values of the liability reached sufficient value in the unobservable scale, in the opposite case we expect the failure.

Similarly, for more categories, we observe one from several states of the categorical trait as the consequence of fact that the underlying liability exceeds the corresponding, unobservable threshold.

Idea

Let us suppose we observe two threshold traits X and Y which are possibly correlated. This correlation referring to corresponding for X and Y liabilities cannot be measured by Pearson' correlation coefficient because the values of X and Y are not observable in the continuous scale.

So, we need to use a measure of correlation for the categorical values of X and Y, for example the entropy.

Idea

The question is:

Is it possible to estimate the correlation between the threshold traits on the basis of information which can be collected from the categorical observations?

Entropy

According to Shannon's fundamental paper "*A Mathematical Theory of Communication*"(1948), we define the entropy of a discrete variable X with the probability mass function $p(x)$ as

$$H(X) = E_X[I(x)] = -\sum_x p(x) log_b(p(x)),$$

where $I(x) = -log_p(p(x))$ is the information context of X, b is the base of logarithm used.

The unit of entropy is *shannon* or *bit* when $b = 2$, *nat* for $b = e$ and *hartley* for $b = 10$.

Conditional entropy

The conditional entropy of two variables X and Y taking values x and y respectively is defined as:

$$H(X|Y) = E_Y[H(X, y)] = -\sum_y p(y) \sum_x p(x|y) log_b p(x|y).$$

The common entropy of two variables X and Y taking values x and y respectively is given by:

$$H(X, Y) = H(X) + H(Y|X) = H(Y) + H(X|Y).$$

Properties of entropy

1. $H(X) = 0$ if and only of when there exist one event x with $p(x) = 1$.
2. The value of entropy reaches the maximum when all events x have the same probability.
3. For two independent variables X and Y
$$H(X, Y) = H(X) + H(Y)$$

Mutual information

Mutual information is a measure of information about variable X with the observation given for variable Y:

$$I(X,Y) = H(X) + H(Y) - H(X,Y) = H(Y) - H(Y|X) = H(X) - H(X|Y)$$

Mutual information is zero for independent variables, so the following coefficient can be used as a measure of correlation:

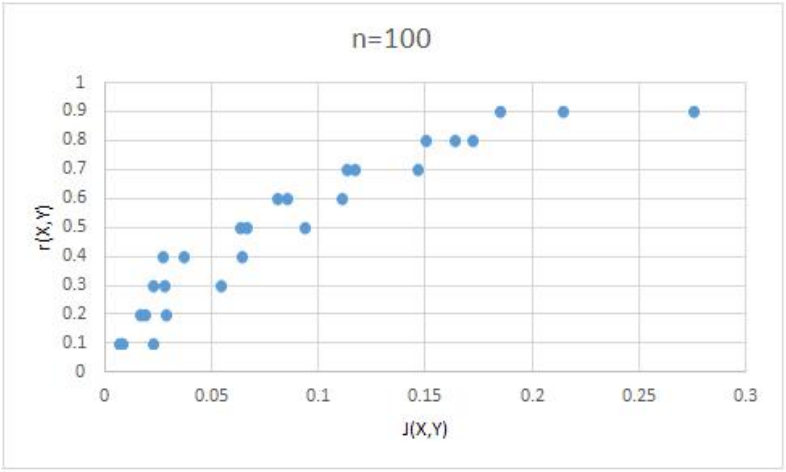
$$J(X,Y) = \frac{I(X,Y)}{H(X,Y)} \in [0,1]$$

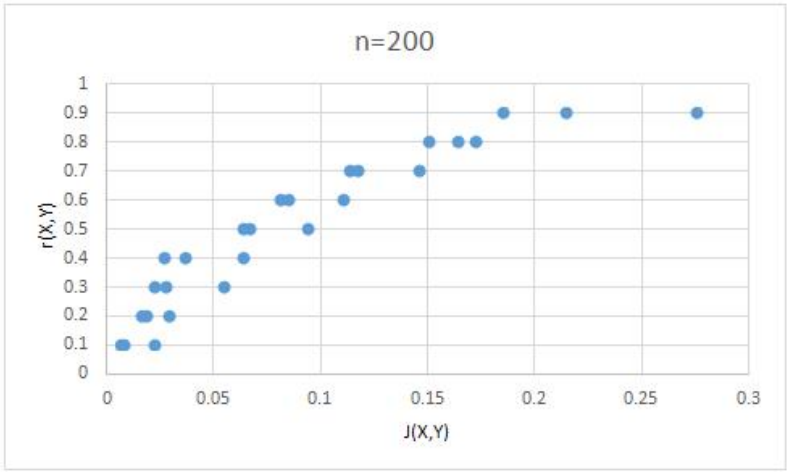
Data simulation

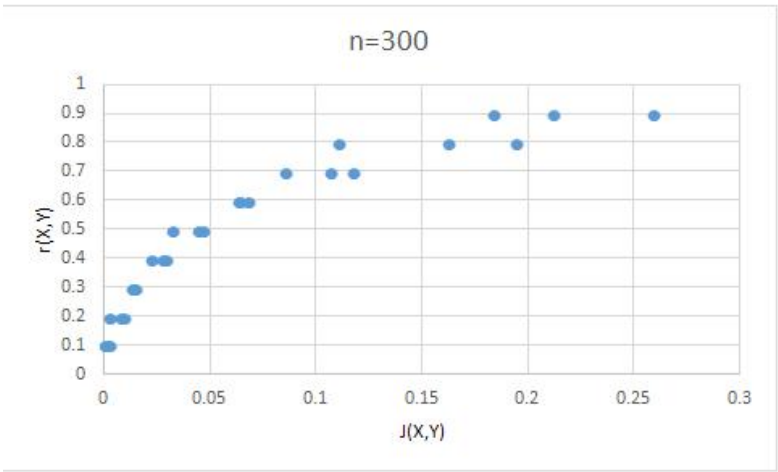
1. The continuous variable X length $n = 100$, $n = 200$ and $n = 300$ was simulated from two variants of the normal distribution: $N(10, 2^2)$ and $N(50, 5^2)$.
2. The values of X were transformed to obtain Y variable which was correlated with X according with assumed Pearson' correlation coefficient r. Nine values of r were checked: from $r = 0.1$ to $r = 0.9$ with step 0.1.
3. In each case the values of X were divided into two categories (i.e. success or failure) while the values of Y were categorized into two, three or four classes.
4. The categorized data were organized in $2x2$, $2x3$ or $2x4$ tables. For each table the information J(X,Y) was calculated.

Results for data generated from $N(10, 2^2)$

The dimensions of data tables are treated as the replications

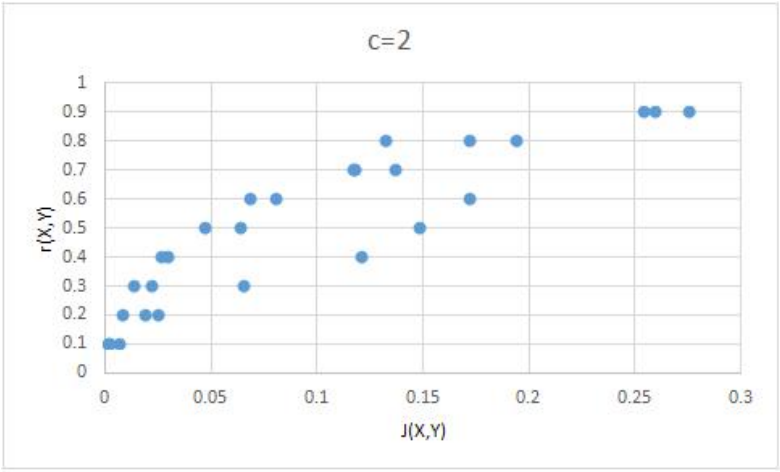


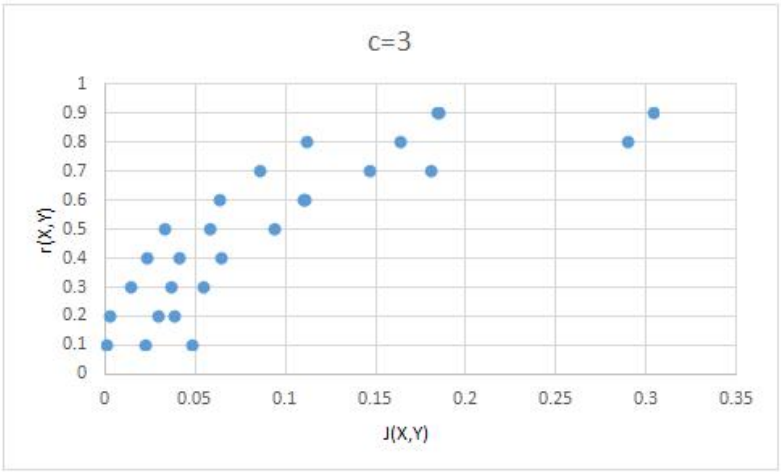


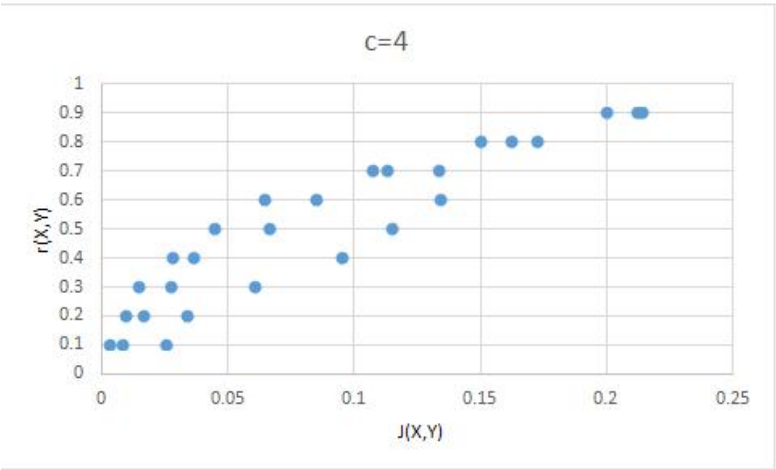


Results for data generated from $N(10, 2^2)$

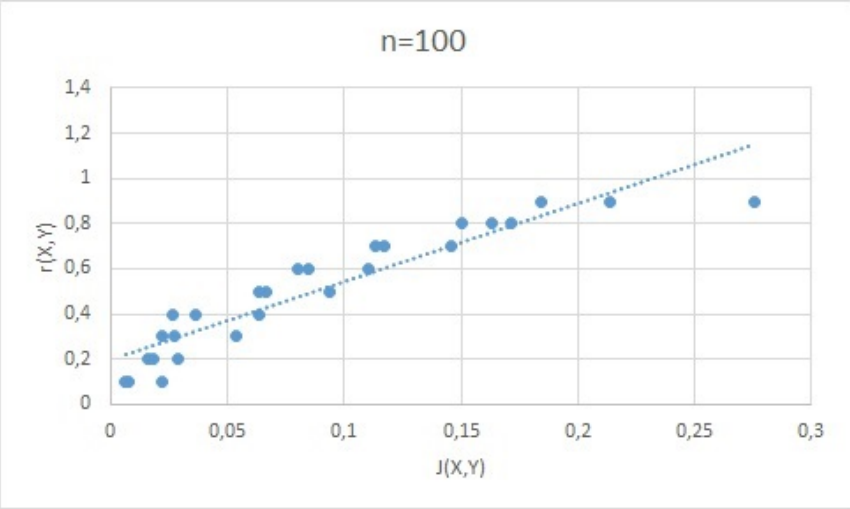
The length of X variable is treated as the replication

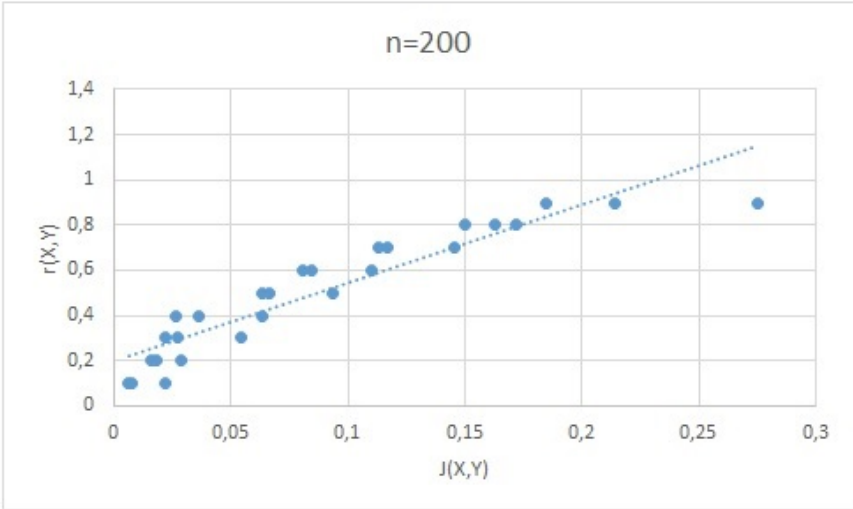


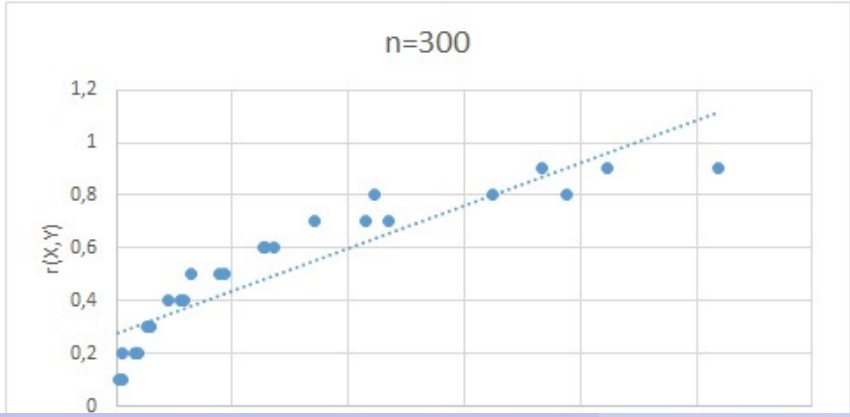




Regression


n=100
1,4
1,2
1
0,8
0,6
0,4
0,2
0
r(X,Y)
0
0,05
0,1
0,15
0,2
0,25
0,3
J(X,Y)


n=200
1,4
1,2
1
0,8
0,6
0,4
0,2
0
r(X,Y)
0
0,05
0,1
0,15
0,2
0,25
0,3
J(X,Y)


n=300
1,2
1
0,8
0,6
0,4
0,2
0
r(X,Y)

Regresion

Because in all cases considered, the values of $J(X, Y)$ were small (less than 0.3), $ln(J(X, Y))$ were used in the regression and in a consequence also $ln(r(X, Y))$ instead of $J(X, Y)$ and $r(X, Y)$ (only positive values of r were considered).

Linear regression

$$-ln(r(X, Y)) = -B_1 ln(J(X, Y)) + B_0$$

was estimated.

Regression

	N=300	N=200	N=100
Data 1 N(10, 2^2)			
B_0	-0.61	-0.94	-0.48
B_1	0.43	0.64	0.54
Data 2 N(10, 2^2)			
B_0	-0.92	-0.90	-0.90
B_1	0.54	0.56	0.65
Data 3 N(50, 5^2)			
B_0	-0.72	-0.90	-0.56
B_1	0.48	0.55	0.40
Data 4 N(50, 5^2)			
B_0	-0.43	-0.78	-0.44
B_1	0.38	0.55	0.35

Regression

	N=300	N=200	N=100
Data 1 N(10, 2^2)			
B_0	-0.61	-0.94	-0.48
B_1	0.43	0.64	0.54
Data 2 N(10, 2^2)			
B_0	-0.92	-0.90	-0.90
B_1	0.54	0.56	0.65
Data 3 N(50, 5^2)			
B_0	-0.72	-0.90	-0.56
B_1	0.48	0.55	0.40
Data 4 N(50, 5^2)			
B_0	-0.43	-0.78	-0.44
B_1	0.38	0.55	0.35

Suggestions

1. The analysis of all checked cases enabled to observe that the value of regression coefficient is near to 0.5 (with minimum 0.3, maximum 0.64 and mean 0.495) and the intercept is near to -0.7 (with minimum -0.99, maximum -0.21 and mean -0.688).
2. On the basis of the regression equation the following relation between $r(X, Y)$ and $J(X, Y)$ can be proposed:

$$r(X, Y) = exp^{|B_0|} J(X, Y)^{B_1}$$

3. Used the averaged values of regression coefficients we obtain that

$$r(X, Y) = 2\sqrt{J(X, Y)}$$

Problems

1. It is possible to find in the analytical way a relationship between $r(X, Y)$ and $J(X, Y)$ which could confirm (or deny) the relation presented above?
2. Which other continuous distribution could be reasonable to use for X variable?
3. What would be more valuable from the practical point of view: to increase the length of X or to increase the number of categories for X and Y (empty categories problem)?

Bibliography

1. Jakulin A., 2005, Machine Learning Based on Attribute Interactions. PhD thesis.
2. Shannon C.E., 1948, A mathematical theory of communication, *The Bell System Technical Journal*, Vol. 27, pp. 379-423, 623-656.