UDC 519.24

D. KOROLIOUK, V.S. KOROLIUK, N. ROSATO

EQUILIBRIUM PROCESSES IN BIOMEDICAL DATA ANALYSIS: THE WRIGHT–FISHER MODEL

Abstract. The biological process of cooperative interaction with equilibrium state will be described as a model of binary statistical experiments with Wright–Fisher normalization, which sets the concentration of a certain characteristic. Such mathematical model is supposed to have a property of persistent regression which means that all current elementary transitions depend on the mean concentration of the said characteristics in the previous state. Equilibrium state of the model is expressed in the terms of the regression function, given by a cubic parabola with three real roots. We construct stochastic approximation of the model by autoregressive process with normal disturbances. Such approach was developed for effective and calculable mathematical description of dynamic concentration for experiment planning, parameters evaluation and hypotheses verification of mechanism of action.

Keywords: binary statistical experiment, persistent regression, Wright–Fisher normalization, equilibrium state, normal autoregression.

INTRODUCTION

In biological processes with equilibrium, the dynamics of concentration, or frequencies, of a predefined characteristic, can be described by mathematical model of binary statistical experiments, based on statistical data of elementary hypotheses validation about the presence or absence of a predefined attribute A in the set of elements that make up a complex system.

It is assumed that:

1) all the elements that make up the system can gain or lose the said attribute A over time, that is the frequency of attribute A is dynamic variable;

2) the basic object of our study are statistical experiments, characterized by relative frequencies of presence or absence of the attribute A in a sample of fixed volume at each time instant;

3) it is assumed the dependency of (average) results of next experiment (at time instant $k+1$) on average result of the present experiment at time instant k. This relationship is called the feature of persistent regression and used as fundamental condition for the subsequent analysis of the model.

The method of constructing and exploring of the proposed mathematical model is based on analysis of the following basic properties of statistical experiments:

1. Persistent regression.
2. Equilibrium value and fluctuations, as well as their asymptotic behavior.
3. Approximation by normal process of autoregression.

Note that in view of assumption 2), the value of statistical experiment tends to probability of the presence or absence of attribute A, by $N \to \infty$, according to the law of large numbers. However, such probability, according to assumption 3), has a complex relationship and can be constructively expressed in terms of regression function.

As a result, all static and dynamic characteristics of the model can be expressed in terms of the set of regression functions, or their transformations.

EXAMPLE: A MODEL OF BIOLOGICAL MOLECULES INTERACTION

Equilibrium processes are common in chemical and biochemical systems and play important role in many mechanisms of interaction and self-regulation.

In studies of these processes, one should develop an adequate method of description, analysis and prediction of the behavior of such systems, taking into account the actions of a wide variety of external factors (for example, [1]).

As an example, we can refer to the enzyme Glutathione Transferase (GST) that binds the tripeptide glutathione (GSH) as a substrate for subsequent organic reactions. The GST is a very important enzyme in cellular biology, as a matter of fact it is involved in many detoxifying reactions and in mechanism of drug resistance in cancer cells [2]. This is a typical case where the stochastic nature of biological molecules interaction is interpreted as a model of statistical experiments with regression function, which sets the concentration $S_N^+(k)$ of biological molecules bound to a ligand (enzymes plus ligand), and of free biological molecules (enzyme without ligand) $S_N^-(k)$ at each stage k of experiment. In this case N records the total number of bounded and free molecules in biological experiments. The probabilities of transition to capture or release of the ligand depend on the mean concentration of free and bound ligands in the previous step. The biological process of capture and release of ligands by enzymes will be described by a model of statistical experiments with Wright–Fisher normalization. The choice of this model is caused by similarity of the process of biological molecules interaction and the process of arrangement of two attributes in the loci.

Such mathematical model describes the dynamic process of interaction between ligands and enzymes. Under this model, one can analize experimental data for estimation of parameters which define its dynamics, verify the model adequacy, exercise fitting of the model, develop the protocols of measurement procedures etc.

STATISTICAL MODEL FORMULATION

The starting point for the mathematical model construction is Wright–Fisher model in population genetics which is used by more than seventy years.

Basing on the definition formulated in [3], we restrict ourselves to the case of two genetic attributes A_1 and A_2. This narrows our analysis on a special case which, however, is very rich for numerous applications.

Consider statistical model with pairwise arrangement of such attributes at particular locus for N representatives of the population. There are three ways of attributes pairing: A_1A_1, A_1A_2 or A_2A_1 and A_2A_2.

In order to study the model dynamics, let us denote p the frequency of attribute A_1, and q the frequency of attribute A_2, observed in previous experiment.

For the next experiment, the attribute A_1 frequency we denote P_+, and the attribute A_2 frequency we denote P_-. As already noted, the probabilities $P_\pm$ depend on the average value of experiment in previous time instant.

We will consider the Wright–Fisher model as a series of statistical experiments (SE), defined by the amounts of sample values $\delta(k) = (\delta_r(k),\ 1 \le r \le N),\ k \ge 0$:

$$S_N(k) = \frac{1}{N}\sum_{r=1}^{N} \delta_r(k),\ k \ge 0, \tag{1}$$

in which the random variables $\delta_r(k)$, $1 \le r \le N$, $k \ge 0$, are equally distributed and independent for each fixed $k \ge 0$ which take binary values ± 1.

Thus, the values of SE, determined by (1), mean that the following equation takes place:

$$S_N(k) = S_N^+(k) - S_N^-(k). \tag{2}$$

Here $S_N^\pm(k)$ means the relative frequency, of two values $+1$ and -1 in the sample:

$$S_N^+(k) = \frac{1}{N}\sum_{r=1}^{N} I\{\delta_r(k) = +1\},\ S_N^-(k) = \frac{1}{N}\sum_{r=1}^{N} I\{\delta_r(k) = -1\},\ k \ge 0, \tag{3}$$

where $I(A)$ is indicator of event A:

$$I(A) = \begin{cases} 1 & \text{if event } A \text{ occurs;} \\ 0 & \text{if event } A \text{ does not occurs.} \end{cases}$$

The frequencies $S_N^{\pm}(k)$ are uniquely determined by the values $S_N(k)$ in the following way:

$$S_N^{+}(k)=[1+S_N(k)]/2,\ S_N^{-}(k)=[1-S_N(k)]/2$$

with obvious complete probability relation

$$S_N^{+}(k)+S_N^{-}(k)=1.$$

The sampling binary values are defined by conditional probabilities

$$P\{\delta_r(k+1)=\pm 1\,|\,S_N(k)=s\}=P_{\pm}(s),\ k\geq 0. \tag{4}$$

We shall consider the regression function $P_{\pm}(s)$ expressed in terms of Wright–Fisher proportions [3, Ch. 10, p. 412]:

$$\begin{gathered}P_{\pm}(s):=W_{\pm}(p,q)/W(p,q),\\ W_{+}(p,q):=p(W_1p+q),\ W_{-}(p,q):=q(W_2q+p),\\ W(p,q):=W_{+}(p,q)+W_{-}(p,q)=W_1p^2+2pq+W_2q^2,\ p+q=1,\end{gathered} \tag{5}$$

where the viability parameters W_1, W_2 are positive constants.

In this case, the values $p, q=1-p$, and s are connected as follows:

$$p:=(1+s)/2,\ q:=(1-s)/2,\ \ s=p-q,\ |s|\leq 1. \tag{6}$$

The initial suppositions (1)–(5) imply that the sequence of SE (1) has the property of persistent regression:

$$E[S_N(k+1)\,|\,S_N(k)=s]=C(s),\ |s|\leq 1,\ k\geq 0, \tag{7}$$

in which the regression function with Wright–Fisher normalization has the following form (see [3, Ch. 10]):

$$C(s)=[W_{+}(s)-W_{-}(s)]/W(s). \tag{8}$$

TRANSFORMATION OF REGRESSION FUNCTIONS

Asymptotic analysis (as $N\to\infty$) of statistic experiments (1)–(5) and their increments involves the use of the corresponding regression functions or their transformations.

Introducing new notations for reasons of symmetry, we can express the functions in (8) using viability parameters:

$$\begin{gathered}V_{+}:=1-W_1,\ V_{-}:=1-W_2;\ \ p_{+}:=p,\ \ p_{-}:=q=1-p,\\ W_{+}(p_{+})=p_{+}(1-p_{+}V_{+}),\ W_{-}(p_{-})=p_{-}(1-p_{-}V_{-}),\\ \bar{p}:=(p_{+},\,p_{-});\ s=p_{+}-p_{-},\ \ p_{\pm}=(1\pm s)/2,\\ W(\bar{p})=1-[V_{+}p_{+}^2+V_{-}p_{-}^2],\\ W(s)=1-[V_{+}(1+s)^2+V_{-}(1-s)^2]/4.\end{gathered} \tag{9}$$

Now consider the process of increments of statistical experiments (1)–(3):

$$\Delta S_N^{\pm}(k):=S_N^{\pm}(k+1)-S_N^{\pm}(k),\ \Delta S_N(k):=S_N(k+1)-S_N(k).$$

Introduce regression functions with Wright–Fisher normalization of increments $C_0(s)$ and $C_0^{\pm}(p)$:

$$\begin{gathered}E[\Delta S_N(k)\,|\,S_N(k)=s]=C_0(s)/W(s),\\ E[\Delta S_N^{\pm}(k)\,|\,S_N^{\pm}(k)=\bar{p}]=C_0^{\pm}(\bar{p})/W(\bar{p}).\end{gathered} \tag{10}$$

The corresponding regression functions are denoted as:

$$C(s) := E[S_N(k+1) \mid S_N(k) = s],\ C^{\pm}(\bar{p}) := E[S_N^{\pm}(k+1) \mid S_N^{\pm}(k) = p_{\pm}]. \quad (11)$$

So one has

$$C(s) = s + C_0(s)/W(s);\ C^{\pm}(\bar{p}) = p_{\pm} + C_0^{\pm}(\bar{p})/W(\bar{p}). \quad (12)$$

Hereinafter, the regression function with Wright–Fisher normalization will be used in the following modified form.

Proposition 1. The regression functions $C_0^{\pm}(\bar{p})$ and $C_0(s)$ of increments have the following form:

$$C_0(s) = -\frac{1}{4}V(1-s^2)(s-\rho),$$
$$C_0^{\pm}(\bar{p}) = \mp p_+ p_- (V_+ p_+ - V_- p_-) = -p_+ p_- V(p_{\pm} - \rho_{\pm}). \quad (13)$$

Here:

$$V = V_+ + V_-,\ \rho = \rho_+ - \rho_-,\ \rho_{\pm} := V_{\mp}/V,\ 0 < V_{\pm} < 1,\ \bar{\rho} = (\rho_+, \rho_-).$$

The Wright–Fisher normalizing functions in (9) have the following representation:

$$W(\bar{p}) = W(\rho_+, \rho_-) - V[p_{\pm} - \rho_{\pm}]^2,$$
$$W(s) = W(\rho) - \frac{1}{4}V(s-\rho)^2, \quad (14)$$
$$W(\bar{\rho}) = W(\rho_+, \rho_-) = [1 - V_+ V_-/V].$$

The regression functions (12), (13) can be calculated as follows:

$$C_0(s) = C_0^+(p) - C_0^-(p),\ C_0^{\pm}(\bar{p}) = W_{\pm}(\bar{p}) - p_{\pm} W(\bar{p}),\ p_{\pm} := (1 \pm s)/2. \quad (15)$$

As seen in (13), the regression function with Wright–Fisher normalization in the transformed form contains a cubic parabola with three real roots:

$$s_{\pm} = \pm 1,\ s_0 = \rho.$$

Hence there are equilibrium values $\rho_{\pm}$ and ρ:

$$C_0^{\pm}(\rho_{\pm}) = 0;\ C_0(\rho) = 0. \quad (16)$$

The defined in (10) regression functions of increments satisfy the balance condition

$$C_0^+(\bar{p}) + C_0^-(\bar{p}) = 0.$$

STEADY STATE REGIME

Define the fluctuations

$$\hat{S}_N(k) := S_N(k) - \rho. \quad (17)$$

In the sequel we shall study fluctuations (1)–(3) as primary goal.

The specifics of binary SE allows to calculate the conditional variance:

$$D[S_N(k+1) \mid S_N(k) = s] = B(s)/N,\ k \geq 0, \quad (18)$$

where

$$B(s) = 1 - C^2(s). \quad (19)$$

The convergence of conditional variance (18) to zero as $N \to \infty$ ensures the availability of steady state.

Theorem 1. If the initial convergence holds (with probability 1)

$$S_N(0) \xrightarrow{P1} \rho,\ N \to \infty, \quad (20)$$

then the following convergence (with probability 1) takes place:

$$S_N(k) \xrightarrow{P1} \rho, \ N \to \infty, \tag{21}$$

for each finite $k \geq 0$.

Proof of Theorem 1. Just as in [4], we introduce martingale as a sum of martingale differences:

$$\mu_N(n) := \sum_{k=0}^{n} [S_N(k+1) - C(S_N(k))]. \tag{22}$$

The quadratic characteristic of martingale (22), taking into account (18), is given as follows:

$$\langle \mu_N \rangle_n := \frac{1}{N} \sum_{k=0}^{n} B(S_N(k)). \tag{23}$$

The boundness of variances ensures convergence (with probability 1) of quadratic characteristic (23)

$$\langle \mu_N \rangle_n \xrightarrow{P1} 0, \ N \to \infty, \ n \geq 0. \tag{24}$$

Hence we have the convergence (with probability 1) of martingales (22) (see [5]):

$$\mu_N(n) \xrightarrow{P1} 0, \ N \to \infty, \tag{25}$$

for each finite $n \geq 0$.

In particular when $n = 0$, one has:

$$M_N(0) = S_N(1) - C(S_N(0)) = \hat{S}_N(1) - [\hat{S}_N(0) + C_0(S_N(0))/W(S_N(0))]. \tag{26}$$

By assumption (20) of Theorem 1 and properties (13), (14), the term in square bracket of (26) tends to zero as $n \to \infty$. Hence the convergence (with probability 1)

$$\hat{S}_N(1) \xrightarrow{P1} 0 \text{ as } n \to \infty$$

takes place.

By induction, we deduce that for every finite $k \geq 1$ the convergence (21) takes place. Theorem 1 is proved.

STOCHASTIC APPROXIMATION

The regression function $C_0(s)$ (see (13)) has a multiplier $(s - \rho)$, so there is the possibility of approximating the SE (1)–(8) by normal process of autoregression with discrete time $k \geq 0$, by $N \to \infty$. Another approach see in [6].

Theorem 2. Under the conditions of Theorem 1 there takes place the limit by distribution:

$$\sqrt{N}[S_N(k+1) - C(S_N(k))] \xrightarrow{d} \sigma\omega(k+1), \ N \to \infty, \tag{27}$$

for each finite $k \geq 0$.

Here σ^2 is quadratic variation which is expressed through the equilibrium value:

$$\sigma^2 = 1 - \rho^2 \tag{28}$$

and $\omega(k)$, $k \geq 1$, is a sequence of independent, normally distributed random variables with parameters $(0, 1)$:

$$E\omega(k) = 0, \ D\omega(k) = 1, \ k \geq 1.$$

Proposition 2 (Approximation). The limit relation (27) forms the basis to consider the normal process of autoregression $\tilde{S}_N(k)$:

$$\tilde{S}_N(k+1) = C(\tilde{S}_N(k)) + \frac{\sigma}{\sqrt{N}} \omega(k+1), \ k \geq 0, \tag{29}$$

as an approximation of the original $S_N(k)$ with the same nonlinear regression function

$$C(s) = s - \frac{1}{4} V (1 - s^2)(s - \rho) / W(s), \tag{30}$$

$$W(s) = W(\rho) - \frac{1}{4} V (s - \rho)^2, \tag{31}$$

$$W(\rho) = 1 - V_+ V_- / V, \ \rho = (V_- - V_+) / V.$$

Remark 1. The process of normal autoregression $\widetilde{S}_N(k)$, $k \geq 0$, of course, differs from the sequence of statistical experiments $S_N(k)$, $k \geq 0$.

In particular, the martingale differences $\sqrt{N}[S_N(k+1) - C(S_N(k))]$, generated by SE, are limited almost surely. However, the fluctuations of normal autoregression (29) are defined by normal random variables $\omega(k+1)/\sqrt{N}$, $k \geq 0$, and bounded only in probability. In accordance with Chebyshev inequality

$$P\left\{\frac{1}{\sqrt{N}} |\omega(k+1)| > C\right\} \leq \frac{\sigma^2}{C \cdot N} \to 0, \ N \to 0,$$

for every finite $C > 0$. So fluctuations in (29) are bounded with the probability, arbitrarily close to 1 for sufficiently large sample size N.

Remark 2. The normal process of autoregression (28)–(31) retains the property of persistent regression (7):

$$E[\widetilde{S}_N(k+1) | \widetilde{S}_N(k) = s] = C(s), \ |s| \leq 1, \tag{32}$$

with the same regression function (30) as for initial SE.

Proposition 3. In the neighborhood of equilibrium point ρ, the nonlinear regression function $C(s)$ can be approximated by linear regression function

$$\widetilde{S}_N(k+1) = C_\rho(\widetilde{S}_N(k)) + \frac{\sigma}{\sqrt{N}} \omega(k+1),$$

$$C_\rho(s) = bs + (1-b)\rho = \rho + b(s - \rho), \ b = 1 - \frac{1}{4} V \sigma^2 / W(\rho), \tag{33}$$

$$\sigma^2 = 1 - \rho^2, \ W(\rho) = 1 - V_+ V_- / V = 1 - V\sigma^2 / 4.$$

Proof of Theorem 2. Since the regression function (30) has in denominator the function $W(s)$, it greatly complicates the asymptotic analysis of SE. We need the following lemma.

Lemma 1. The function $W(p, q)$ (see (9)) admits uniform estimation

$$0 < 1 - \max(V_+, V_-) \leq W(p, q) \leq W(\rho)$$

for all values of the parameters V_+, V_- which satisfy the conditions:

$$0 < \max(V_+, V_-) < 1.$$

Introduce a martingale as sum of martingale-differences

$$\mu_N(n) := \sum_{k=0}^{n} \sqrt{N}[S_N(k+1) - C(S_N(k)], \ n \geq 0. \tag{34}$$

In order to transform the martingale (34), taking into account the regression function representation (30), let us introduce the normalized fluctuations

$$\zeta_N(k) := \sqrt{N}[S_N(k) - \rho], \ k \geq 0. \tag{35}$$

Lemma 2. The martingale (34) has the following asymptotic representation:

$$\mu_N(n) = \sum_{k=0}^{n} [\zeta_N(k+1) - b\zeta_N(k)] + \frac{1}{\sqrt{N}} \sum_{k=0}^{n} \zeta_N(k) R_N(\hat{S}_N(k)), \ n \geq 0, \tag{36}$$

where $\hat{S}_N(k)$ is defined in (17),

$$b = 1 - \frac{1}{4} V\sigma^2 / W(\rho),\ W(\rho) = 1 - V_+ V_- / V = 1 - V\sigma^2 / 4. \tag{37}$$

The residual function in (36) has the follwing form:

$$R_N(\hat{s}) = -\frac{V}{4}\left[2\rho W(\rho) + \hat{s}\left(1 - \frac{V}{2}\sigma^2\right)\right] / W(s)W(\rho),\ \sigma^2 = 1 - \rho^2. \tag{38}$$

The function $R_N(S_N(k))$, $k \geq 0$, is bounded in probability under the conditions of Lemma 1.

Now let us prove, in the same way as in [4], the normal approximation of martingale (34) or, equivalently, (36)–(38). First, the quadratic characteristics of martingale (34) are the following:

$$\langle \mu_N \rangle_n = \sum_{k=0}^{n} B(S_N(k)),\ B(s) = 1 - C^2(s). \tag{39}$$

Then, by Theorem 1, we find the limit (with probability 1):

$$\langle \mu_N \rangle_n \xrightarrow{P1} (n+1)\sigma^2,\ N \to \infty,\ \sigma^2 = 1 - \rho^2, \tag{40}$$

Further, according to central limit theorem, the main part of martingale (36) converges in distribution to the sum of normally distributed random variables:

$$\mu_N^0(n) := \sum_{k=0}^{n} [\zeta_N(k+1) - b\zeta_N(k)] \xrightarrow{d} \sum_{k=0}^{n} \omega_\sigma(k+1),\ N \to \infty. \tag{41}$$

At the same time the convergence (41) means that the random values $\{\omega_\sigma(k), 0 \leq k \leq N\}$ are asymptotically independent and have equal variances:

$$\omega_\sigma(k) = \sigma\omega(k),\ \sigma^2 = 1 - \rho^2. \tag{42}$$

The convergence of martingales (34) means that we have the convergence in probability (27). Theorem 2 is proved.

CONCLUSIONS

The models of SE given in Proposition 2 and Proposition 3, can be used in statistical analysis of real experiments.

For this purpose we intend to use the methods of mathematical statistics for parameters estimation of the normal process of autoregression, namely for estimation of guiding parameters $V \pm$, equilibrium value ρ and square variance σ^2.

The normal processes of autoregression (29) and (32) can be used to predict the behavior of real SE, or to detect deviations in the behavior of real statistical data, due to external conditions change.

REFERENCES

1. A stochastic model for the sigmoidal behaviour of cooperative biological systems / M. Abundo, L. Accardi, Finazzi Agro' A., G. Mei, N. Rosato // Biophysical Chemistry. — 1996. — **58**. — P. 313–323.
2. Glutathione transferases and development of new principles to overcome drug resistance / A. Sau, F.P. Tregno, F. Valentino, G. Federici, A.M. Caccuri // Archives of Biochemistry and Biophysics. — 2010. — **500**, N 2. — P. 116–122.
3. Ethier S.N., Kurtz T.G. Markov processes: Characterization and convergence. — New York: Wiley, 1986. — 640 p.
4. Korolyuk D. Recurrent statistical experiments with persistent linear regression // J. Math. Sci. — 2013. — **190**, N 4. — P. 600–607.
5. Borovskikh Yu.V., Korolyuk V.S. Martingale approximation. — Utrecht: VSP, 1997. — 345 p.
6. Korolyuk V.S., Koroliouk D. Diffusion approximation of stochastic Markov models with persistent regression // Ukr. Math. J. — 1995. — **47**, N 7. — P. 1065–1073.

Поступила 13.02.2014