

AN APPROXIMATION LEADING TO THE BINOMIAL DISTRIBUTION IN THE STOCHASTIC TREATMENT OF A REVERSIBLE CHEMICAL REACTION

By

A. A. BABIKER* and E. K. ELSHEIKH**

*Department of Mathematics, Faculty of Science, University of Qatar, Doha, Qatar

**School of Mathematical Sciences, University of Khartoum, Sudan.

تقريب يؤدي إلى توزيع ذي الحدين عند المعالجة العشوائية للتفاعلات الكيميائية العكوسة

ع. ع. بابكر و أ. خ. الشيخ

ندرس في هذا البحث نموذجاً رياضياً عشوائياً للتفاعل الكيميائي العكوس $A+B \rightleftharpoons C$ ونبين أن توزيع الاحتمال الإتزاني لعدد جزيئات C يؤول إلى توزيع ذي الحدين وذلك عندما يؤول M العدد الكلي لجزيئات A إلى مالانهاية وتؤول النسبة K بين معدلي التفاعلين الأمامي والعكسي إلى الصفر بحيث يظل KM مقداراً منتهياً. كما نقدم مقارنة حسابية بين التوزيع الإتزاني الحقيقي وتوزيع ذي الحدين المُقَرَّب له وذلك لعدد من مجموعات قيم المقادير N, M, K حيث N يساوي العدد الكلي لجزيئات B وتحت نفس هذه الشروط نثبت أن توزيع الاحتمال العابر للمتغير العشوائي C(t) الممثل لعدد جزيئات C عند الزمن t يؤول إلى توزيع مجموع متغيرين عشوائيين مستقلين يتبع كل منهما توزيع ذي الحدين.

Key Words: Binomial distribution, Reversible Chemical reactions.

ABSTRACT

The exact equilibrium distribution of the number of molecules of C in the reversible chemical reaction $A+B \rightleftharpoons C$ is shown to converge to the binomial distribution when N is finite,

$M \rightarrow \infty, K \rightarrow 0$ in such a way that KM is finite, where M, N and K are respectively, the total number of molecules of A, the total number of molecules B and the ratio of the forward to the backward reaction rates, K_1/K_2 . The practical implication of this limiting condition have been investigated by numerically computing the two distributions for a wide range of values of the parameters. Under the same conditions, the transient distribution of C(t), the number of C molecules at time t, is shown to be approximately a convolution of two binomial distributions.

INTRODUCTION

The reversible chemical reaction



has received much attention (cf. eg. Darvey *et al.* (1966), Staff (1967), Oppenheim *et al.* (1969), Thakur *et al.* (1978), Formosinho and Miguel (1979). It is one of four basic reactions described by McQuarrie (1967) in his survey paper on the stochastic approach to chemical reactions. In this approach the quantities of the reactants present at any time are treated as random variables and the course of the reaction is determined by finding the distributions of these random variables. The equilibrium distributions are easy to obtain but are computationally demanding. On the other hand an exact solution of the differential difference equation governing the reaction seems to be out of reach.

Various approximations of the equilibrium solution have been considered. A normal approximation derived by Dunstan and Reynolds (1981) is reported to give accurate results in a range of values of interest when compared with the exact solution. Hall (1983) derived explicit conditions on the parameters under which the normal approximation is valid. Hall also considered a number of limiting situations that lead to the Bessel and Poisson distributions. Reference is made there to other efforts towards approximations. The approximations of Dunstan and Reynolds and Hall cover all the four basic reactions:



Here we derive a binomial approximation of the exact equilibrium distribution confining ourselves only to the reaction (1.1). The exact equilibrium distribution and its binomial approximation are compared for a wide range of parameter values using a FORTRAN program written for that purpose. The program works out the complete distribution for the two

cases. To facilitate comparison the output consists only of the means, variances and selected standardized tail quantiles. A summary of the results of the numerical comparison is presented. We also consider the transient distribution under the same limiting conditions and show that it can be approximately by a sum of two independent binomial variables.

ASSUMPTIONS AND BASIC RESULTS

Let

$$\begin{aligned} A(t) &= \text{number of molecules of type A at time } t, \\ B(t) &= \text{number of molecules of type B at time } t, \\ C(t) &= \text{number of molecules of type C at time } t, \end{aligned} \quad (2.1)$$

Then we have the following relations

$$\begin{aligned} A(t) + C(t) &= M, \\ B(t) + C(t) &= N, \end{aligned} \quad (2.2)$$

where M and N are the total number of molecules of type A and type B respectively. Thus, we have three random variables satisfying two relations. To describe the reaction we need to find the distribution of only one of them. We choose C(t). It is worth mentioning here that the parametrization of the problem adopted here is different from the one usually followed but is suitable for our purpose. For the usual parametrization see Dunstan and Reynolds (1981) or Hall (1983).

The usual assumptions for modelling the reaction as a stochastic process are as follows:

In any small interval of time (t, t+h)

- (i) the probability of an AB association is $K_1 A(t)B(t)h + o(h)$,
- (ii) the probability of a dissociation of a molecule of C is $K_2 C(t)h + o(h)$, and
- (iii) the probability of the occurrence of more than one event is $o(h)$

where

$$\lim_{h \rightarrow 0} \frac{o(h)}{h} = 0.$$

If we let

$$P_r(t) = P(C(t) = r), \quad (2.3)$$

then using the standard procedure of representing $P_r(t+h)$ as a difference equation in $P_r(t)$, it is easy to show that $P_r(t)$ satisfies the following differential difference equation

$$\frac{dP_r(t)}{dt} = K_1(M-r+1)(N-r+1)P_{r-1}(t) + K_2(r+1)P_{r+1}(t) - (K_1(M-r)(N-r) + K_2r)P_r(t), \quad (2.4)$$

where r can take any integer value provided $P_r(t)$ is defined as 0 for $r < 0$ and $r > \min.(N, M)$.

The equilibrium solution is obtained by putting the derivative equal to zero in (2.4) and solving for $P_r = P_r(\infty)$. It is easy to see that it satisfies

$$K(M-r)(N-r)P_r = (r+1)P_{r+1}, \quad (2.5)$$

where

$$K = K_1/K_2. \quad (2.6)$$

This leads to

$$P_r = \frac{K^r M! N!}{r!(M-r)!(N-r)!} P_0 \quad (2.7)$$

P_0 can be obtained since $\sum_r P_r = 1$

The probability generating function of $P_r(t)$, defined by

$$G(x, t) = \sum x^r P_r(t), \quad (2.8)$$

can be obtained by multiplying both sides of (2.4) by x^r , summing over all r and simplifying the resulting expression. We get

$$\frac{\partial G}{\partial t} = K_1 MN(x-1)G - (x-1)[(K_1 M + K_1 N - K_1)x + K_2] \frac{\partial G}{\partial x} + K_1(x-1)x^2 \frac{\partial^2 G}{\partial x^2} \quad (2.9)$$

Attempts at solving this second order partial differential equation have not met with success. We note that if $C(0) = r_0$, the equation has to be solved under the initial condition

$$G(x, 0) = x^{r_0} \quad (2.10)$$

It is useful to have approximation formulae for the mean μ and the variance, σ^2 , of the equilibrium distribution (2.7). In what follows we denote the third central moment of the equilibrium distribution by m_3 . Summing both sides of (2.5) over r and putting $C(\infty) = C$, we get:

$$KE[(M-C)(N-C)] = E(C); \quad (2.11)$$

where E denotes the expected value. Simplifying we get:

$$K(M-\mu)(N-\mu) - \mu + K\sigma^2 = 0. \quad (2.12)$$

Multiplying both sides of (2.5) by r and summing over all r, gives:

$$KE[C(M-C)(N-C)] = E[C(C-1)]. \quad (2.13)$$

This leads to:

$$K\mu(M-\mu)(N-\mu) + KE(C^3) - K\mu^3 - K(M+N)\sigma^2 - \mu^2 - \sigma^2 + \mu = 0 \quad (2.14)$$

Now

$$m_3 = E[(C-\mu)^3] = E(C^3) - 3\mu\sigma^2 - \mu^3. \quad (2.15)$$

Substituting in (2.14) gives

$$\mu[K(M-\mu)(N-\mu) - \mu + K\sigma^2] + Km_3 + 2K\mu\sigma^2 - K(M+N)\sigma^2 - \sigma^2 + \mu = 0. \quad (2.16)$$

Using (2.12) we get

$$\sigma^2 = \frac{\mu + Km_3}{K(M-\mu) + K(N-\mu) + 1}. \quad (2.17)$$

Under the assumption that the equilibrium distribution is symmetric we get the following approximate formula for σ^2

$$\sigma^2 = \frac{\mu}{K(M-\mu) + K(N-\mu) + 1}. \quad (2.18)$$

The contribution of $K\sigma^2$, with σ^2 as given above, to equation (2.12) is bound to be negligible for small K. This is confirmed by

numerical computations. We can thus approximate μ by the following equation derived from (2.12) by putting $K\sigma^2 = 0$;

$$K(M - \mu)(N - \mu) - \mu = 0. \quad (2.19)$$

Note that from (2.18)

$$\frac{1}{\sigma^2} = [K(M - \mu) + K(N - \mu) + 1]/\mu$$

$$= \frac{K(M - \mu)(N - \mu)}{\mu} \left[\frac{1}{M - \mu} + \frac{1}{N - \mu} + \frac{1}{K(M - \mu)(N - \mu)} \right];$$

and using (2.19) we get:

$$\frac{1}{\sigma^2} = \frac{1}{M - \mu} + \frac{1}{N - \mu} + \frac{1}{\mu}. \quad (2.20)$$

This is the approximate formula for the variance given by Dunstan and Reynolds (1981), where μ in their case is the mode rather than the mean.

THE BINOMIAL APPROXIMATION

In this section we derive a limiting condition under which the equilibrium distribution (2.7) converges to the binomial distribution. We begin by a heuristic argument. Assume without loss of generality that $M \geq n$. Each molecule of B can be in one of two states, either free or bound to a molecule of A. C is then the number of bound molecules of B. One way in which the binomial distribution can arise is to set the conditions for the molecules of B to act independently in an identical manner. C will then have the distribution of the number of successes out of N independent trials, i.e. will be binomially distributed. We should note that the molecules of B are physically indistinguishable. So it seems that this solves the problem of identity. One reason for the molecules not to act independently is their competition over the molecules of A. If we supply an abundance of the latter, each molecule of B should be free or bound independently of the other molecules. But there is another situation in which the competition will heat, namely when association rate constant K_1 is large. Hence to achieve independence we should make M large and K_1 small.

More formally, we now show that when $M \rightarrow \infty$, $K \rightarrow 0$ such that $KM \rightarrow \lambda$, a finite value, then the distribution of C converge to the binomial distribution provided N is kept finite. Referring to (2.7) we note that when $M \geq N$, then

$$\frac{M!}{(M - r)!} \sim M^r \quad (3.1)$$

since r is at most equal to N. Thus, under the limiting condition considered, (2.7) becomes

$$P_r = (KM)^r \binom{N}{r} P_0 \quad (3.2)$$

i.e.
$$P_r = \lambda^r \binom{N}{r} P_0, \quad r = 0, 1, \dots, N \quad (3.3)$$

Thus
$$1 = \sum_{r=0}^N P_r = (1 + \lambda)^N P_0 \quad (3.4)$$

Therefore
$$P_0 = \left(\frac{1}{1 + \lambda}\right)^N \quad (3.5)$$

It follows that

$$P_r = \binom{N}{r} \lambda^r \left(\frac{1}{1 + \lambda}\right)^N \quad (3.6)$$

i.e.
$$P_r = \binom{N}{r} \left(\frac{\lambda}{1 + \lambda}\right)^r \left(\frac{1}{1 + \lambda}\right)^{N-r}, \quad r = 0, 1, \dots, N \quad (3.7)$$

i.e.
$$C \sim B(N, \frac{\lambda}{\lambda + 1}) \quad (3.8)$$

where $C = C(\infty)$ and $\lambda = KM$. (3.9)

We observe that relation (3.1) holds irrespective of the value of K. Thus, it seems that the condition for the convergence to the binomial distribution is simply $M \rightarrow \infty$. We should note, however, that if the condition $K \rightarrow 0$ is not imposed, the limiting distribution will be a degenerate binomial concentrated at N.

THE NUMERICAL COMPARISON

Having shown that the exact equilibrium distribution (2.7) converges to the binomial distribution under the condition N finite, $\rightarrow \infty$, $K \rightarrow 0$ such that $KM = \lambda$ is finite, we now investigate the practical implications of this condition by numerically computing the two distributions for a range of values of the parameters.

The computer program written to compare the two distributions uses the recurrence relation (2.5) valid for the exact distribution and the corresponding relation for the binomial distribution obtained by replacing $M - r$ by M in (2.5). The recurrence relation for the exact distribution is started at m, rather than zero, where m is the nearest integer to the approximate mean given and especially (2.19). This is in anticipation that probabilities for values far away from the mean and especially at $r=0$ will be practically zero when dealing with a large number of molecules. The probabilities are then computed (in terms of P_m) for values up to seven standard deviations away from the mean, the standard deviation being approximated by relation (2.18). Use is then made of the fact that the sum of probabilities should be 1. As expected, truncation beyond seven standard deviations was found not to result in any improvement of accuracy. The same procedure is carried out to compute the binomial probabilities.

The range of parameter values considered, with $R = \frac{M}{N}$, is as follows:

$$\lambda = 10^k, \quad \text{for } k = -4, -3, \dots, 4$$

$$R = 10^k, \quad \text{for } k = 1, 2, \dots, 5$$

$$N = 10, 10^2, 10^3, 10^5, 10^6, 10^7$$

The program output consists of the means, variances and selected standardized tail quantities for the two distributions for each combination of the parameter values. The agreement is assessed in terms of the percentage differences for the means and variances and in terms of decimal places for the standardized quantiles.

The results obtained can be summarized as follows:

For $\lambda \leq 0.001$ there is agreement up to around 0.1% in the

means and 0.2% in the variances even for $R=1$. The agreement in tail quantiles is up to two decimal places and improves as N increases. The two distributions are highly skewed to the right for small N (e.g. $N=100$ and 1000), but the skewness tends to decrease as N increases. However, some slight skewness still remains even at $N=10^6$. Increasing R to 10 results in 10 fold accuracy of the binomial as an approximation to the exact distribution.

In summary as long as $R \geq 10$, there is close agreement between the two distributions in location, scale and shape at the tails for all N . For $\lambda \leq 0.01$ and $R=1$, the difference may reach around 1% for the means and 2% for the variances. The quantiles agree to at least one decimal place. Thus there is fairly good agreement in this case even when $R=1$. Increasing N for fixed λ improves the agreement in shape but not in scale and location. When $\lambda \leq 0.1$ and $R=1$, The percentage difference in means may exceed 8% and that in variances may exceed 16%. For example when $\lambda=0.1$, $R=1$ and $N=10^6$, the mean of the exact distribution is 8392.03 while the mean of the binomial distribution is 9090.91. Here we have a difference of 8.33% between the means. For $\lambda \geq 1$ and $R=1$ or 10 there are large differences between the two distributions in location and scale. However, when $R=100$, the maximum percentage difference between the mean is around 0.25%, while the maximum percentage difference between variances is less than 1%. See for example Table 1. The percentage difference for $R=1000$ are one tenth of those for $R=100$.

Table 1
Characteristics of the Equilibrium Distribution and its Binomial Approximation
 $\lambda=1, R=100, N=1000$

	Exact	Binomial
Mean	498.75	500.00
Variance	249.37	250.00
$Z_{.005}$	-2.607	-2.607
$Z_{.01}$	-2.358	-2.358
$Z_{.025}$	-1.992	-1.992
$Z_{.05}$	-1.677	-1.677
$Z_{.1}$	-1.313	-1.313
$Z_{.90}$	1.134	1.324
$Z_{.95}$	1.677	1.677
$Z_{.975}$	1.992	1.992
$Z_{.99}$	2.357	2.358
$Z_{.995}$	2.607	2.607

It may be said then that irrespective of the values of λ and N , there will be very good agreement between the two distributions in location and scale when $R=100$. When $R=1000$ there is hardly any difference between the two distributions in all respects. The agreement in shape is found to be invariably good and improves as N increases even though the two distributions may be quite different regarding location and scale. Since the binomial distribution is well approximated by the normal

distribution for large N , the same claim may be made for the equilibrium distribution. However, the normal approximation of the binomial gets worse the smaller the value of its mean. Hence, when the mean is small we would expect the binomial distribution to provide an approximation of to the exact distribution that is not shared by the normal distribution. For example, when $\lambda=0.0001$, $M=100N$ and $N=10^6$, the mean and variance of the binomial distribution are both 10, a fairly small value and as we see from Table 2, the two distributions have the same shape which is quite different from the shape of the normal distribution.

Table 2
Characteristics of the Exact Equilibrium Distribution and its Binomial Approximation
 $\lambda=0.0001, M=100N, N=10^6$

	Exact	Binomial
Mean	10.00	10.00
Variance	10.00	10.00
$Z_{.005}$	-2.437	-2.437
$Z_{.01}$	-2.228	-2.228
$Z_{.025}$	-1.969	-1.969
$Z_{.05}$	-1.724	-1.724
$Z_{.1}$	-1.416	-1.416
$Z_{.90}$	1.481	1.481
$Z_{.95}$	1.886	1.886
$Z_{.975}$	2.264	2.264
$Z_{.99}$	2.721	2.721
$Z_{.995}$	3.031	3.031

APPROXIMATING THE TRANSIENT SOLUTION

Assume as before that N is finite $M \rightarrow \infty$ and $K_1 \rightarrow 0$ in such a way that $K_1 M \rightarrow K_2 \lambda$ a finite value. Invoking these limiting conditions on the infinitesimal assumptions it can be easily shown that the partial differential equation defining the probability generating function of $C(t)$ takes the limiting form

$$\frac{\partial G}{\partial t} = K_2 \lambda N(x-1)G - K_2(x-1)(\lambda x + 1) \frac{\partial G}{\partial x}; \quad (5.1)$$

with initial condition

$$G(x,0) = x^{C(0)}$$

Since we have the binomial distribution in mind, we try a solution for (5.1) of the form:

$$G(x,t) = [1 - p(t) + xp(t)]^N$$

substituting this in (5.1) we obtain;

$$\frac{dp}{dt} = K_2 \lambda - K_2(1 + \lambda)p$$

which gives

$$p(t) = \lambda / (\lambda + 1) + \alpha \exp(-K_2(1 + \lambda)t) \quad (5.2)$$

The initial condition is then equivalent to

$$[1 - p(0) + xp(0)]^N = x^{C(0)}$$

This restricts the value of $C(0)$ to either 0 or N . In the first case ($C(0)=0$ and $p(0)=0$), from which (5.2) becomes

$$p(t) = [\lambda / (\lambda + 1)] [1 - \exp(-K_2(1 + \lambda)t)],$$

and $C(t) \sim B(N, p(t))$. (5.3)

This means that each molecule of B will be bound to a molecule of A at time t with probability $p(t)$ given by (5.3) independently of other molecules.

In the second case $C(0)=N$, $p(0)=1$ and $C(t) \sim B(N, q(t))$ where

$$q(t) = (1 + \lambda)^{-1} [\lambda + \exp(-K_2(1 + \lambda)t)]. \quad (5.4)$$

Thus if a molecule of B was initially bound to a molecule of A , it will be bound at time t with probability $q(t)$ independently of other molecules.

Under the general initial condition $G(x, 0) = x^{C(0)}$, equation (5.1) can be solved by the method of characteristics (Zauderer 1989). The characteristic equations are

$$\frac{dx}{ds} = K_2(x - 1)(\lambda x + 1); \quad (5.4a)$$

$$\frac{dt}{ds} = 1; \quad (5.4b)$$

$$\frac{dG}{ds} = K_2 \lambda N(x - 1)G; \quad (5.4c)$$

with initial values: $s=0$; $x=\tau$; $t=0$; $G=\tau^{C(0)}$.

These equations give $t = s$ and

$$(x - 1)/(\lambda x + 1) = [(\tau - 1)/(\lambda \tau + 1)] \exp(K_2(1 + \lambda)t). \quad (5.5)$$

A few steps of simple algebra starting from (5.5) give

$$\tau = [1 - q(t) + xq(t)] / [1 - p(t) + xp(t)]; \quad (5.6)$$

where $p(t)$ and $q(t)$ are given by (5.3) and (5.4), and

$$x = 1 + \left[\frac{(\lambda + 1)(\tau - 1)}{\lambda \tau + 1} \exp(K_2(1 + \lambda)t) \right] / \left[1 - \frac{\lambda(\tau - 1)}{\lambda \tau + 1} \exp(K_2(1 + \lambda)t) \right]. \quad (5.7)$$

Substituting (5.7) in (5.4c) and integrating the resulting equation with initial value $G(0) = \tau^{C(0)}$ we get

$$G(x, t) = [1 - p(t) + xp(t)]^N \tau^{C(0)} \\ = [1 - p(t) + xp(t)]^N [1 - q(t) + xq(t)] / [1 - p(t) + xp(t)]^{C(0)}$$

(by 5.6) and thus $G(x, t) = [1 - q(t) + xq(t)]^{C(0)} [1 - p(t) + xp(t)]^{N - C(0)}$, which is the probability generating function of the convolution of two binomial distributions. i.e. $C(t)$ has the same distribution as the sum of two independent binomial random variables, one with parameters $(C(0), q(t))$ and the other with parameters $(N - C(0), p(t))$.

The interpretation of this is that the molecules of B will act independently. Each of the $C(0)$ molecules that were initially bound, will be bound at time t with probability $q(t)$ independently of all other molecules. Hence, the number out of these that will be bound at time t will follow the binomial distribution with parameters $C(0)$ and $q(t)$. Similarly, the number bound at time t out of the $N - C(0)$ initially free molecules will follow the binomial distribution with parameters $N - C(0)$ and $p(t)$. The number of molecules of C at time t is the sum of these two random variables and therefore has the probability generating function shown above.

If independence of molecules is assumed, the reaction can be studied by considering the behaviour of a single molecule of B , for example, which can be in any of two states: free (state 0) and bound (state 1). If we let $X(t)$ denote the state of the molecule at time t , $X(t)$ is then a two-state Markov process with matrix of transition rates given by

$$Q = \begin{matrix} & \begin{matrix} 0 & 1 \end{matrix} \\ \begin{matrix} 0 \\ 1 \end{matrix} & \begin{bmatrix} -K_2 \lambda & K_2 \lambda \\ K_2 & -K_2 \end{bmatrix} \end{matrix}$$

From this simple Markov process it is easy to obtain $p(t)$ and $q(t)$ as given by (5.3) and (5.4) and this completely solves the problem for the whole reaction.

ACKNOWLEDGEMENT

The authors wish to thank the referee for his helpful comments and suggestions.

REFERENCES

- Darvey, I.G., B.W. Ninham and P.J. Staff, 1966. Stochastic model for second-order chemical reaction kinetics: The equilibrium state. *J. Chem. Phys.* 45: 2145-2155.
- Darvey, I.G. and P.J. Staff, 1967. The application of the theory of Markov processes to the reversible one substrate - one intermediate - one product enzymic reaction. *J. Theoret. Biol.* 14: 157-172.
- Dunstan, F.D.J. and J.F. Reynolds, 1981. Normal approximation for distributions arising in the stochastic approach to chemical reactions kinetics. *J. Appl. Prob.* 18: 263-267.
- Formosinho, S.J. and M.DAG. M. Miguel, 1979. Markov chains for plotting the course of chemical reactions. *J. Chem. Education* 56: 582-585.
- Hall, P. 1983. On the roles of the Bessel and Poisson distributions in chemical kinetics. *J. Appl. Prob.* 20: 585-599.
- McQuarrie, D.A. 1967. Stochastic approach to chemical kinetics. *J. Appl. Prob.* 4: 413-478.
- Oppenheim, I., K.E. Shuler and G.H. Weiss, 1969. Stochastic and deterministic formulation of chemical rate equations. *J. Chem. Phys.* 50: 460-466.
- Staff, P.J. 1967. Approximation method for equilibrium distributions in second-order chemical reaction. *J. Chem. Phys.* 46: 2209-2212.
- Thakur, A.K., A. Rescigno and C. De Lisi, 1978. Stochastic theory of second-order chemical reactions. *J. Phys. Chem.* 82: 552-558.
- Zauderer, E. 1989. Partial differential equations of applied mathematics. John Wiley and Sons, New York.