A gentle introduction to combinatorial stochastic processes VI

Enrico Scalas

Department of Mathematics, University of Sussex, UK

Stochastic Models for Complex Systems
10 June - 7 July, 2021

Outline

Infinite number of categories and innovations

Up to now, the case of n objects to be allocated into g categories was discussed. But, what happens if the number of categories is infinite? The situation with an infinite number of categories can naturally take into account "innovations". Consider for instance the case in which the n elements are agents who work in k firms. The random variable $\mathbf{Y} = \mathbf{n} = (n_1, \ldots, n_k)$ with $\sum_{i=1}^{k} n_i = n$ represents the distribution of agents into the k occupied firms. Assume that, at a certain point, an agent decides to leave his/her firm (say the firm labeled by j) and create a new one (a one-person firm). The new situation is represented by the vector $\mathbf{Y}' = \mathbf{n}' = (n_1, \ldots, n_j - 1, \ldots, n_k, n_{k+1} = 1)$. Now, there are $k + 1$ firms (occupied categories) with a new distribution of agents among them. Often, it is useful to study the size distribution of these categories. As usual, this is described by the random variable (note that $z_0 = \infty$) $\mathbf{Z} = \mathbf{z} = (z_1, \ldots, z_n)$, where z_i is the number of categories with i elements, but now, one has $\sum_{i=1}^{n} z_i = k$ and $\sum_{i=1}^{n} i z_i = n$ with k not fixed (and in general randomly varying!). It turns out that the methods developed in the previous chapters can be used to quantitatively discuss these problems and to derive useful distributions that can be compared with empirical data.

The Chinese restaurant process I

In order to visualize the situation described above, it is useful to refer to the so-called *Chinese restaurant process*. There is a Chinese restaurant; usually, it is described as having an infinite number of tables each one with infinite capacity. However, for our purposes, it is better to think of a Chinese restaurant in which each table can accommodate an arbitrary number of customers and each customer can be hosted at a new table if the customer wishes so. In principle, this process can be effectively realized in reality! (A table can be always expanded to host a new person and an additional table can be added to host a new customer). A customer comes and decides where to sit. Then, a second customer comes and decides where to sit. The second customer can sit at the same table of the previous one or choose a new table. After the accommodation of the second customer, the restaurant may have either $k = 2$ tables occupied by 1 person ($n_1 = 1, n_2 = 1$) or $k = 1$ table occupied by 2 persons $n_1 = 2$. In the former case, there are 2 clusters of size 1 ($z_1 = 2$), in the latter case, there is 1 cluster of size 2 ($z_1 = 0, z_2 = 1$). Again, when a third customer comes, he/she can join a table occupied by the previous two customers or a new table. After the third customer sits down the possibilities are as follows:

1. all the three customers sit at the same table, then, $k = 1$, $n_1 = 3$ and $z_1 = 0, z_2 = 0, z_3 = 1$;
2. two customers sit at the same table and another one occupies a different table; then $k = 2$ and one has either $n_1 = 2, n_2 = 1$ or $n_1 = 1, n_2 = 2$ and $z_1 = 1, z_2 = 1$;
3. the three customers sit at separate tables; in this case $k = 3$ and $n_1 = 1, n_2 = 1, n_3 = 1$ with $z_1 = 3$.

The Chinese restaurant process II

The process is then iterated for the desired number of times. For instance, the following configuration is compatible with the allocation of $n = 10$ agents in the Chinese restaurant process: $k = 4$ tables are occupied, $n_1 = 3$, $n_2 = 5$, $n_3 = 1$, $n_4 = 1$ and $z_1 = 2$, $z_2 = 0$, $z_3 = 1$, $z_4 = 0$, $z_5 = 1$, $z_6 = 0$, $z_7 = 0$, $z_8 = 0$, $z_9 = 0$, $z_{10} = 0$. Note that in the scheme proposed above, tables are numbered (labelled) according to the order of occupation. There is an alternative way of naming tables (categories). One can introduce an auxiliary urn with integer numbers from 1 to n and randomly extract a label without replacement every time, one of the n customers decides to choose a new table. For instance, after the arrival and accommodation of the third customer the situation is as follows. As three agents are expected to arrive, one has

1. all the three customers sit at the same table, then, $k = 1$, $n_j = 3$, where $j = 1$ or $j = 2$ or $j = 3$ and $z_1 = 0$, $z_2 = 0$, $z_3 = 1$; the number of possible situations is 3;
2. two customers sit at the same table and another one occupies a different table; then $k = 2$ and one has either $n_i = 2$, $n_j = 1$ or $n_j = 1$, $n_i = 2$ and $z_1 = 1$, $z_2 = 1$; i can take 3 values and j one of the two remaining values after the choice of i. Therefore, there are 6 possible situations;
3. the three customers sit at separate tables; in this case $k = 3$ and $n_i = 1$, $n_j = 1$, $n_l = 1$ with $z_1 = 3$. Also in this case there are 6 possible situations, corresponding to the permutations of the three labels.

After a little thought, one should convince him/herself that, if there are k occupied tables and n is the number of the labels, the number of situations corresponding to a case is given by $n(n-1)\cdots(n-k+1)$. Note that, in general, one can prepare g labels for the tables, with $g \leq n$ of which only n will be selected. The numbers of situations corresponding to each case will be given by $g(g-1)\cdots(g-k+1)$ if k tables are occupied as for the first occupied table there are g choices, for the second occupied table $g-1$ choices are left, until $g-k+1$ choices available for the k-th occupied table.

Relation with the Pólya process I

We showed that the Pólya distribution converges to the Dirichlet distribution for $n \to \infty$ and g and $\boldsymbol{\alpha}$ fixed. The question with which this Chapter began was: What happens, when the number of categories becomes infinite, that is when $g \to \infty$ with fixed n? Indeed, in this case the frequency description $\mathbf{n} = (n_1, \ldots, n_g)$ has infinite terms, and some caution is necessary. The vector $\mathbf{n}$ can be considered as the description of the size of categories (or clusters) of "colours" labelled from 1 to g. As discussed above, when $g \to \infty$, it is useful to introduce the description of clusters $\mathbf{z} = (z_1, \ldots, z_n)$, $\sum_i i z_i = n$, taking into account the number of clusters of size $i = 1, \ldots, n$ (and not their names). In order to understand what happens, it is useful to consider the case in which g is large but finite, and assume that there are $k \leq n$ distinct clusters (categories initially occupied). Relabel them, and set $\mathbf{n} = (n_1, \ldots, n_k, n_{k+1})$, where the $(k+1)$-th category collects all the $g - k$ empty categories at present. One can see that $n_1, \ldots, n_k$ are positive, whereas $n_{k+1} = 0$. In this setting, one has the following predictive probabilities

$$\mathbb{P}(X_{n+1} = j | X_1 = x_1, \ldots, X_n = x_n) = \mathbb{P}(X_{n+1} = j | n_j, n) = \begin{cases} \dfrac{\alpha_j + n_j}{\alpha + n} & \text{for } j \leq k \\ \dfrac{\alpha - \sum_{i=1}^{k} \alpha_i}{\alpha + n} & \text{for } j = k+1, \end{cases} \tag{1}$$

where $\alpha = \sum_{i=1}^{k+1} \alpha_i$, and $\alpha - \sum_{i=1}^{k} \alpha_i$ is the the total weight of the $g-k$ empty categories at present.

Relation with the Pólya process II

If the weight α_i of each category is finite, for large g both α and $\alpha - \sum_{i=1}^{k} \alpha_i$ diverge, so that $\mathbb{P}(X_{n+1} = j | n_j, n) \to 0$ for $j \leq k$, and $P(X_{n+1} = j | n_j, n) \to 1$ for $j = k + 1$. In other words, the next category observed will be a new one with probability one. The *a priori* weights dominate on the empirical observations, so that no correlations are present. New categories are always chosen and one expects to observe n clusters occupied by a single object, that is $\mathbf{z} = (z_1 = n, z_2 = 0, \ldots, z_n = 0)$ with probability one. A different situation is obtained if the total weight of the initial distribution converges to a constant value, $\lim_{g \to \infty} \sum_i^g \alpha_i = \alpha = \theta < \infty$. (Note that the previous case can be recovered setting $\theta = \infty$). Now the initial weight θ and the number of empirical observations are both finite, and an interesting behaviour is expected. In this limit, one can study the stochastic process characterized by the following predictive probability

$$\mathbb{P}(X_{n+1} = j | X_1 = x_1, \ldots, X_n = x_n) = \mathbb{P}(X_{n+1} = j | n_j, n) = \begin{cases} \dfrac{n_j}{\theta + n} & \text{for } j \leq k \\ \dfrac{\theta}{\theta + n} & \text{for } j = k + 1; \end{cases} \tag{2}$$

We are now going to study the process defined by equation (2).

Hoppe's urn

Consider n random variables $X_1, \ldots, X_n$, whose range is a set of labels $1, \ldots, n$. Assume that the first random variable is labelled 1 and that a new label is introduced sequentially whenever a new category appears. In this case, the label j denotes the j-th label that has been introduced. To be more specific, $\mathbf{Y}_m = (m_1, \ldots, m_k)$ is the current occupation vector, with $m_j = \#\{X_i = j, i = 1, \ldots, m\}$, and $k = \#\{m_j > 0, j = 1, \ldots, k\}$ is the number of present labels. The conditional predictive distribution of X_{m+1} is given by equation (2), leading to (for $m = 0, 1, \ldots, n-1$):

$$\mathbb{P}(X_{m+1} = j | m_j, m) = \begin{cases} \dfrac{m_j}{m+\theta} & j \leq k \\ \dfrac{\theta}{m+\theta}, & j = k+1 \end{cases} \tag{3}$$

with $\mathbb{P}(X_1 = 1) = 1$ by definition. This sampling process can be modelled by an urn process: Hoppe's urn, that can be traced back to A. De Moivre, according to Zabell. Initially, the urn contains a single white ball whose weight is θ. The following rules define the drawing process:

1. Whenever the white ball is drawn, it is painted with a colour not yet present in the urn; then a new white ball (of the same weight θ) is replaced into the urn, together with the ball just painted (of weight 1).
2. If a colored ball is drawn, it is replaced into the urn, together with a ball (of weight 1) of the same colour, as in the Pólya scheme.

Individual and frequency descriptions

The probability of an individual sequence is obtained from (3) and is given by:

$$\mathbb{P}(X_1 = 1, X_2 = x_2, \ldots, X_n = x_n) = \frac{\theta^k}{\theta^{[n]}} \prod_{i=1}^{k} (n_i - 1)!; \tag{4}$$

You can try and derive equation (4) as a useful exercise. Given an occupation vector $\mathbf{n} = (n_1, \ldots, n_k)$ there are many corresponding sequences with the same probability. For instance, if $n = 4$, the sequences $X_1 = 1, X_2 = 2, X_3 = 3, X_4 = 2$ and $X_1 = 1, X_2 = 2, X_3 = 2, X_4 = 3$ are equiprobable. However, the sequences are not exchangeable! For instance the probability of the sequence $X_1 = 2, X_2 = 1, X_3 = 3, X_4 = 2$ is 0 by definition and different from the probability of the other two sequences, even if in all three cases one has $n_1 = 1, n_2 = 2, n_3 = 1$. Therefore, one cannot use the usual multinomial factor to determine the probability of an occupation vector. In particular, all the sequences where 1 is not the first label are forbidden, as well as all the sequences where 3 appears before 2, 4 before 3 and 2, and so on. Taking this into account, if there are n_1 elements belonging to category 1, only $n_1 - 1$ can be freely chosen out of $n - 1$ elements. Then, only $n_2 - 1$ elements in category 2 are free to be chosen out of the $n - n_1 - 1$ remaining elements, and so on. These considerations leads to the following equation

$$\begin{aligned} \mathbb{P}(\mathbf{Y}_n = \mathbf{n}) &= \binom{n-1}{n_1-1}\binom{n-n_1-1}{n_2-1}\cdots\binom{n_{k-1}+n_k-1}{n_{k-1}-1}\frac{\theta^k}{\theta^{[n]}}\prod_{i=1}^{k}(n_i-1)! \\ &= \frac{n!}{n_k(n_k+n_{k-1})\ldots(n_k+n_{k-1}+\ldots+n_1)}\frac{\theta^k}{\theta^{[n]}} \end{aligned} \tag{5}$$

called by Donnelly “the size-biased permutation of the Ewens sampling formula”.

The Chinese restaurant revisited I

Consider n random variables $X_1^*, \ldots, X_n^*$ whose range is a set of $g > n$ labels $L = \{l_1, \ldots, l_g\}$. Now, when X_1^* is observed, a label is randomly chosen from an urn U (without replacement) and we label the category of X_1^* with, say l_{i_1}. As for X_2^*, if $X_2^* = X_1^*$, it will be labeled by l_{i_1} otherwise a second draw is made from U and a new label, say l_{i_2}, is assigned to X_2^*, and so on. Continuing with this procedure, all the coincident values of X_i^* have the same label, whereas different values correspond to different labels. Comparing the label urn process with the Hoppe urn scheme, one can see that while $X_1 = 1$ with probability 1, $X_1^* \in \{l_1, \ldots, l_g\}$ equiprobably; and when the second label appears in Hoppe's scheme, a new label will be drawn from U. The predictive probability for the label process is

$$\mathbb{P}(X_{m+1}^* = j | m_j, m) = \begin{cases} \dfrac{m_j}{m+\theta} & \text{for } m_j > 0 \\ \dfrac{1}{g-k(\mathbf{m})}\dfrac{\theta}{m+\theta} & \text{for } m_j = 0; \end{cases} \tag{6}$$

indeed, as before, new labels are selected when the white ball of weight θ is drawn, but, then, independently, the name of the category is randomly selected among all the remaining labels. The probability of a sequence can be directly obtained from (6). However, it is easier to note that every sequence $X_1 = 1, X_2 = x_2, \ldots, X_n = x_n$ corresponds to $g!/(g-k)!$ label sequences. Moreover, a little thought should convince the reader that the label process is exchangeable. Therefore, one gets

$$\mathbb{P}(X_1^* = x_1^*, \ldots, X_n^* = x_n^*) = \frac{\mathbb{P}(X_1 = 1, X_2 = x_2 \ldots, X_n = x_n)}{g(g-1)\ldots(g-k+1)}. \tag{7}$$

In other words, the non-exchangeable sequence $X_1 = 1, X_2 = x_2 \ldots, X_n = x_n$ is partitioned into $g!/(g-k)!$ exchangeable sequences $X_1^*, \ldots, X_n^*$.

A remark on exchangeability

Equation (4) only depends on $\mathbf{n}$, but this does not mean that $X_1 = 1, X_2 = x_2 \ldots, X_n = x_n$ is exchangeable. As discussed before, sequences such as $X_1 = 1, X_2 = 2, X_3 = 1$ and $X_1 = 1, X_2 = 1, X_3 = 2$ are both possible and equiprobable with occupation vector $(n_1 = 2, n_2 = 1)$, whereas the permuted sequence $X_1 = 2, X_2 = 1, X_3 = 1$ is forbidden by construction. This is also in agreement with the multiplicity factor in $\mathbb{P}(\mathbf{Y}_n = \mathbf{n})$ in equation (5). The fact is that (4) is symmetric with respect to to n_i so that $Y_1 = 1, Y_2 = 2, Y_3 = 1$ and $Y_1 = 1, Y_2 = 2, Y_3 = 2$ are equiprobable, but the number of allowed sequences belonging to $(n_1 = 2, n_2 = 1)$ is greater than the number of sequences corresponding to $(n_1 = 1, n_2 = 2)$.

The Chinese restaurant revisited II

Thanks to exchangeability, we can use the multinomial formula and, from equations (4) and (7), the distribution of the occupation vectors defined on the g states of the label urn turns out to be

$$\mathbb{P}(\mathbf{Y}_n^* = \mathbf{n}) = \frac{n!}{\prod_{i\in A} n_i!}\mathbb{P}(\mathbf{X}^{*(n)} = \mathbf{x}^{*(n)}) = \frac{(g-k)!}{g!}\frac{n!}{\prod_{i\in A} n_i}\frac{\theta^k}{\theta^{[n]}}, \tag{8}$$

where A is the set of k labels representing the occupied clusters.

The Ewens sampling formula I

Equation (8) can be written in a way that directly leads to the Ewens sampling formula. One can see that $\prod_{i=1}^{g} n_i! = \prod_{i=0}^{n} (i!)^{z_i}$. Similarly, in our case, one has that

$$\prod_{i \in A} n_i = \prod_{i=1}^{n} i^{z_i}, \tag{9}$$

where, as usual, z_i is the number of clusters/categories with i elements. Replacing equation (9) into equation (8) leads to

$$\mathbb{P}(\mathbf{Y}_n^* = \mathbf{n}) = \frac{(g-k)!}{g!} \frac{n!}{\theta^{[n]}} \prod_{i=1}^{n} \left(\frac{\theta}{i} \right)^{z_i}. \tag{10}$$

Note that, in this framework $z_0 = g - k$ gives the number of empty clusters/categories. Again, thanks to the exchangeability of occupation vectors, using the multinomial formula, from equation (10) one immediately gets the celebrated Ewens sampling formula for partitions (remember that there are k occupied categories and $z_0 = g - k$)

$$\mathbb{P}(\mathbf{Z}_n = \mathbf{z}_n) = \frac{g!}{(g-k)! z_1! \cdots z_n!} \mathbb{P}(\mathbf{Y}_n^* = \mathbf{n}) = \frac{n!}{\theta^{[n]}} \prod_{i=1}^{n} \left(\frac{\theta}{i} \right)^{z_i} \frac{1}{z_i!}. \tag{11}$$

The Ewens sampling formula II

As a side remark, note that combining the multinomial formula for partitions and (7) leads to

$$\mathbb{P}(\mathbf{Z}_n = \mathbf{z}_n) = \frac{n!}{\prod_{i=1}^{k} n_i!} \frac{1}{\prod_{i=1}^{n} z_i!} \mathbb{P}(\mathbf{X}_n = \mathbf{x}_n) = \frac{n!}{\prod_{i=1}^{n} (i!)^{z_i} z_i!} \mathbb{P}(\mathbf{X}_n = \mathbf{x}_n), \qquad (12)$$

so that, one can directly obtain the joint cluster distribution from the knowledge of $\mathbb{P}(X_1 = 1, X_2 = x_2, \ldots, X_n = x_n)$. In this framework, the Ewens sampling formula appears as the distribution of partition vectors for the exchangeable label process characterized by the predictive probability given by equation (6).

The Ewens sampling formula III

This formula was introduced by W. Ewens in population genetics. If a random sample of n gametes is taken from a population and classified according to the gene at a particular locus, the probability that there are z_1 alleles represented once in the sample, z_2 alleles represented twice, and so on is given by (11) under the following conditions:

1. the sample size n is small if compared to the size of the whole population;
2. the population is in statistical equilibrium under mutation and genetic drift and the role of selection at the locus under scrutiny can be neglected;
3. every mutant allele is an “innovation”.

The Ewens sampling formula IV

As mentioned by Tavaré and Ewens, equation (11) provides a sort of null hypothesis for a non-Darwinian theory of evolution (a theory where selection plays no role). The case $\theta = 0$ corresponds to the situation where all the objects occupy the initial category (all the genes are copies of the same allele, in the genetic intepretation). This can be immediately seen from the equations for the predictive probability: when $\theta = 0$ no innovation is possible. The opposite case was discussed above and appears in the limit $\theta \to \infty$. In this case all the objects occupy a new category. The case $\theta = 1$ is also remarkable as it corresponds to the distribution of integer partitions induced by uniformly distributed random permutations meaning that each permutation has a probability equal to $(n!)^{-1}$. In this case, equation (11) becomes

$$\mathbb{P}(\mathbf{Z}_n = \mathbf{z}_n) = \prod_{i=1}^{n} \left(\frac{1}{i}\right)^{z_i} \frac{1}{z_i!}, \tag{13}$$

and is quite old: it be traced back at least to Cauchy.

Limit of Pólya partitions I

Consider the symmetric Pólya distribution where α denotes the common weight of each category. We are interested in the limit in which this weight vanishes, $\alpha \to 0$, the number of categories diverges, $g \to \infty$, but the total weight $\theta = g\alpha$ remains constant. For small α the rising factorial $\alpha^{[i]}$ can be approximated as follows:

$$\alpha^{[i]} = \alpha(\alpha - 1) \cdots (\alpha + i - 1) \simeq \alpha(i - 1)!. \tag{14}$$

Limit of Pólya partitions II

As a consequence of equation (14), the symmetric Pólya sampling distribution can be approximated as follows

$$\mathbb{P}(\mathbf{n}) = \frac{n!}{(g\alpha)^{[n]}} \prod_{j=1}^{g} \frac{\alpha^{[n_j]}}{n_j!} = \frac{n!}{(g\alpha)^{[n]}} \prod_{i=1}^{n} \left(\frac{\alpha^{[i]}}{i!} \right)^{z_i} \simeq \frac{n!}{\theta^{[n]}} \prod_{i=1}^{n} \left(\frac{\alpha}{i} \right)^{z_i}, \tag{15}$$

where the following identities has been used

$$\prod_{j=1}^{g} n_j! = \prod_{i=1}^{n} (i!)^{z_i},$$

and

$$\prod_{j=1}^{g} \alpha^{[n_j]} = \prod_{i=1}^{n} (\alpha^{[i]})^{z_i}.$$

Limit of Pólya partitions III

Recall that $z_0 = g - k$ and that $\sum_{i=1}^{n} z_i = k$. Then, in the limit of small α and large g, for partitions, one gets

$$\mathbb{P}(\mathbf{z}) = \frac{g!}{(g-k)! \prod_{i=1}^{n} z_i!} \mathbb{P}(\mathbf{n}) \simeq$$

$$\frac{n!}{\theta^{[n]}} g^k \alpha^k \prod_{i=1}^{n} \left(\frac{1}{i}\right)^{z_i} \frac{1}{z_i!} = \frac{n!}{\theta^{[n]}} \prod_{i=1}^{n} \left(\frac{\theta}{i}\right)^{z_i} \frac{1}{z_i!}, \quad (16)$$

which coincides with the Ewens sampling formula.

Cluster number distribution I

The distribution of the number of clusters at step n, K_n, can be obtained from equation (8) by summing over all the occupation vectors with exactly k clusters. Let us denote this set by B. The distribution is given by

$$\mathbb{P}(K_n = k) = \sum_{\mathbf{n} \in B} \mathbb{P}(Y_n^* = \mathbf{n}) = S(n,k) \frac{\theta}{\theta[n]}, \tag{17}$$

where $S(n,k)$ are unsigned Stirling numbers of the first kind, obeying the following recurrence equation

$$S(n+1,k) = S(k,n-1) + nS(n,k).$$

Cluster number distribution II

The expected value and variance of the number of clusters at step n are given by

$$\mathbb{E}(K_n) = \sum_{j=0}^{n-1} \frac{\theta}{\theta + j},$$

$$\mathbb{V}(K_n) = \sum_{j=0}^{n-1} \frac{\theta j}{(\theta + j)^2}.$$

Expectation of cluster sizes

As for the expected cluster sizes, one gets

$$\mathbb{E}(Z_i) = \frac{\theta}{i} \frac{\theta^{[n-1]}/(n-i)!}{\theta^{[n]}/n!}.$$

The formula and derivation for the variance of Z_i is omitted and you are referred to the book and to the literature listed below.

Hoppe's vs Zipf's urn I

Consider the label process for Hoppe's urn. Using the indicator functions $\mathbb{I}_{n_j=0}$ which is 1 when $Y_j^* = n_j = 0$ and 0 for $Y_j^* = n_j > 0$ and its complement $\mathbb{I}_{n_j>0} = 1 - \mathbb{I}_{n_j=0}$, equation (6) can be written in a single line

$$\mathbb{P}(X_{n+1}^* = j | n_j, n) = \frac{n}{n+\theta}\frac{n_j}{n}\mathbb{I}_{n_j>0} + \frac{\theta}{n+\theta}\frac{1}{g-K_n}\mathbb{I}_{n_j=0}, \tag{18}$$

where all the parameters and variables have the same meaning as above. Once more, note that at the beginning of this process, the probability of selecting one of the g categories/sites is given by $1/g$ and this is automatically included in equations (6) and (18). In the right-hand side of equation (18), the first term describes herding as it gives the probability of joining an already occupied site, whereas the second term describes innovation because it is the probability of joining an empty site as a pioneer. We have previously seen ithat the label process is exchangeable. For example, the probability of the sequence $X_1^* = i, X_2^* = j, X_3^* = i$ (with $i \neq j$) is

$$\mathbb{P}(X_1^* = i, X_2^* = j, X_3^* = i) = \frac{1}{g(g-1)}\frac{\theta^2}{\theta(\theta+1)(\theta+2)}, \tag{19}$$

and it coincides with the probability of any permutation of the values, that is with the probability of the sequences $X_1^* = i, X_2^* = i, X_3^* = j$ and $X_1^* = j, X_2^* = i, X_3^* = i$.

Hoppe's vs Zipf's urn II

The label process for Zipf's urn can be defined in a similar way. Using the indicator function $\mathbb{I}_{n=0}$ and $\mathbb{I}_{n>0}$, one can write

$$\mathbb{P}(X^*_{n+1} = j | n_j, n) = \left[(1-u)\frac{n_j}{n}\mathbb{I}_{n_j>0} + u\frac{1}{g-K_n}\mathbb{I}_{n_j=0} \right] \mathbb{I}_{n>0} + \frac{1}{g}\mathbb{I}_{n=0}, \tag{20}$$

where $0 < u < 1$ is the probability of innovation which is now independent of n and its complement $1-u$ is the probability of herding. A direct application of equation (20) on the sequences $X^*_1 = i, X^*_2 = j, X^*_3 = i$ and $X^*_1 = i, X^*_2 = i, X^*_3 = j$ immediately shows that this process is no longer exchangeable. Indeed, one has

$$\mathbb{P}(X^*_1 = i, X^*_2 = j, X^*_3 = i) = \frac{1}{2}\frac{u(1-u)}{g(g-1)}, \tag{21}$$

whereas

$$\mathbb{P}(X^*_1 = i, X^*_2 = i, X^*_3 = j) = \frac{u(1-u)}{g(g-1)}. \tag{22}$$

Given that the process of individual sequences is not exchangeable, the methods we have applied so far cannot be used in order to derive the distribution of site and/or cluster sizes. In 1955, Simon suggested to study the average cluster dynamics as a way to circumvent this problem and obtain quantitative (albeit approximate) results. His method will be described in the next part using the language developed in the course.

Average cluster dynamics and Yule distribution I

Simon refers to the sequential construction of a text adding a new word at each step. This is equivalent to consider a new customer entering the Chinese restaurant and choosing an existing table (a word already written in the text) or a new table (a new word not yet used in the text). Here, instead of using site labels as in the previous section, it is useful to refer to time labels. Equation (20) modifies to

$$\mathbb{P}(X_{n+1} = j | n_j, n) = (1-u)\frac{n_j}{n}\mathbb{I}_{j \leq K_n} + u\mathbb{I}_{j=K_n+1}. \tag{23}$$

Simon assumes that you are writing a text using the following rules. You have already written n words. If $z_{i,n}$ is the number of words appearing i times in the text, the probability of using one of these words at the next step is proportional to $iz_{i,n}$; the probability of using a new word not yet present in the text is constant and equal to u with $0 < u < 1$. Note that the first word in the text is new with probability 1. Morever, if $u = 0$, no innovation is possible and $\mathbb{P}(Z_{n,n} = 1) = 1$, in other words, there is only a cluster containing n repetitions of the same word. On the contrary, if $u = 1$, every word included in the text is a new word not yet used; therefore, in this case, the text is made up of n distinct words meaning that there are n clusters each containing 1 element, so that one has $\mathbb{P}(Z_{1,n} = n) = 1$.

Average cluster dynamics and Yule distribution II

If $Y_{n+1} = i$ denotes the event that the $(n+1)$-th word is among the words that occurred i times and $Y_{n+1} = 0$ means that the $(n+1)$-th word is a new one, equation (23) leads to the following probabilities conditioned on the partition vector $\mathbf{Z}_n = \mathbf{z}_n$:

$$\begin{cases} \mathbb{P}(Y_{n+1} = i | \mathbf{Z}_n = \mathbf{z}_n) &= (1-u)\dfrac{i z_{i,n}}{n} \\ \mathbb{P}(Y_{n+1} = 0 | \mathbf{Z}_n = \mathbf{z}_n) &= u \qquad\qquad n > 0 \\ \mathbb{P}(Y_1 = 0) &= 1. \end{cases} \tag{24}$$

Average cluster dynamics and Yule distribution III

The evolution of the partition vector is the following. If $Y_{n+1} = i$, it means that a cluster of size i is destroyed and a cluster of size $i+1$ is created meaning that $z_{i,n+1} = z_{i,n} - 1$ and $z_{i+1,n+1} = z_{i+1,n} + 1$. If $Y_{n+1} = 0$, it means that a cluster of size 1 is created, that is $z_{1,n+1} = z_{1,n} + 1$. Consider now the random variable difference $Z_{i,n+1} - Z_{i,n}$; it can assume two values, one has $Z_{i,n+1} - Z_{i,n} = 1$ if a cluster of size $(i-1)$ is destroyed and a cluster of size i is created and $Z_{i,n+1} - Z_{i,n} = -1$ if a cluster of size i is destroyed and a cluster of size $i+1$ is created. As a consequence of equation (24), one has for $i = 2, \ldots, n$

$$\mathbb{P}(Z_{i,n+1} - Z_{i,n} = 1 | \mathbf{Z}_n = \mathbf{z}_n) = (1-u)\frac{(i-1)z_{i-1,n}}{n}, \tag{25}$$

and

$$\mathbb{P}(Z_{i,n+1} - Z_{i,n} = -1 | \mathbf{Z}_n = \mathbf{z}_n) = (1-u)\frac{iz_{i,n}}{n}. \tag{26}$$

Average cluster dynamics and Yule distribution IV

Therefore, considering that $\mathbb{E}(Z_{i,n}|\mathbf{Z}_n = \mathbf{z}_n) = z_{i,n}$ for any i, one finds that

$$\mathbb{E}(Z_{i,n+1}|\mathbf{Z}_n = \mathbf{z}_n) - z_{i,n} = (1-u)\left(\frac{(i-1)z_{i-1,n}}{n} - \frac{iz_{i,n}}{n}\right), \tag{27}$$

an equation valid for $i = 2, \ldots, n$. For $i = 1$, as a consequence of equation (24), one finds that

$$\mathbb{E}(Z_{1,n+1}|\mathbf{Z}_n = \mathbf{z}_n) - z_{1,n} = u - (1-u)\frac{z_{1,n}}{n}. \tag{28}$$

Further note that one has

$$\mathbb{E}(\mathbb{E}(Z_{i,n+1}|\mathbf{Z}_n = \mathbf{z}_n)|\mathbf{Z}_{n-1} = \mathbf{z}_{n-1}) = \mathbb{E}(Z_{i,n+1}|\mathbf{Z}_{n-1} = \mathbf{z}_{n-1}) \tag{29}$$

as $\mathbf{Z}_{n-1} = \mathbf{z}_{n-1} \subset \mathbf{Z}_n = \mathbf{z}_n$. Therefore, defining

$$\bar{z}_{i,n} = \mathbb{E}(Z_{i,n}) = \sum_{\mathbf{z}_{n-1}} \mathbb{E}(Z_{i,n}|\mathbf{Z}_{n-1} = \mathbf{z}_{n-1})\mathbb{P}(\mathbf{Z}_{n-1} = \mathbf{z}_{n-1}), \tag{30}$$

one can derive the following equations by taking the averages of (27) and (28)

$$\bar{z}_{i,n+1} - \bar{z}_{i,n} = (1-u)\left(\frac{(i-1)\bar{z}_{i-1,n}}{n} - \frac{i\bar{z}_{i,n}}{n}\right), \tag{31}$$

and

$$\bar{z}_{1,n+1} - \bar{z}_{1,n} = u - (1-u)\frac{\bar{z}_{1,n}}{n}. \tag{32}$$

Average cluster dynamics and Yule distribution V

In principle, the recurrence equations (31) and (32) can be directly solved. However, Simon suggests to look for solutions such that $\bar{z}_{i,n} \propto n$ corresponding to a steady growth of the clusters. With Simon's *Ansatz*, one has that

$$\frac{\bar{z}_{i,n+1}}{\bar{z}_{i,n}} = \frac{n+1}{n}, \tag{33}$$

equivalent to

$$\bar{z}_{i,n+1} - \bar{z}_{i,n} = \frac{\bar{z}_{i,n}}{n}, \tag{34}$$

for $i = 1, \ldots, n$. Replacing equation (34) for $i = 1$ into equation (32) gives

$$\bar{z}^*_{1,n} = \frac{nu}{2-u} = \frac{\rho}{1+\rho} nu, \tag{35}$$

where $\rho > 1$ is a parameter defined as

$$\rho = \frac{1}{1-u}. \tag{36}$$

If equation (34) is replaced into equation (31), the recurrence equation simplifies to

$$\bar{z}^*_{i,n} = \frac{(1-u)(i-1)}{1+(1-u)i} \bar{z}^*_{i-1,n} = \frac{i-1}{\rho+i} \bar{z}^*_{i-1,n}. \tag{37}$$

Average cluster dynamics and Yule distribution VI

The iteration of (37) leads to a closed-form solution for $\bar{z}^*_{i,n}$:

$$\bar{z}^*_{i,n} = \frac{i-1}{\rho+i}\frac{i-2}{\rho+i-1}\cdots\frac{1}{\rho+2}\bar{z}^*_{1,n} = \frac{\Gamma(i)\Gamma(\rho+2)}{\Gamma(\rho+i+1)}\bar{z}^*_{1,n} = (\rho+1)\frac{\Gamma(i)\Gamma(\rho+1)}{\Gamma(\rho+i+1)}\bar{z}^*_{1,n} = \rho B(i,\rho+1)nu, \tag{38}$$

where equation (35) and the definition of Euler's Beta function were used. Direct replacement shows that equation (38) is a solution of equation (31). Note that the ratio $\bar{z}_{i,n}/n$ on the left-hand side of equation (34) is not a frequency. The expected number of clusters (of words in the text) is given by

$$\mathbb{E}(K_n) = \sum_{i=1}^{n}\mathbb{E}(Z_{i,n}) = \sum_{i=1}^{n}\bar{z}^*_{i,n} = nu\sum_{i=1}^{n}\rho B(i,\rho+1), \tag{39}$$

and it is possible to prove that

$$\sum_{i=1}^{\infty}\rho B(i,\rho+1) = 1. \tag{40}$$

Therefore, from equation (39) one has that $\mathbb{E}(K_n) \approx nu$ for large n meaning that there is a constant flow of new words. In other words, Simon's *Ansatz* means that the relative frequency of clusters of size i (the number of words that appeared i times in the text) given by $\mathbb{E}(Z_{i,n})/\mathbb{E}(K_n) \approx \bar{z}^*_{i,n}/nu$ is invariant once steady growth is reached.

Average cluster dynamics and Yule distribution VII

It is necessary to convince oneself that equation (40) holds true. Using

$$f_i = \rho B(i, \rho + 1), \tag{41}$$

one has that $\forall i$, $f_i > 0$, moreover $\forall i$, $f_i < 1$. As for the limiting properties, one has that $\lim_{i\to\infty} f_i = 0$ and that $\lim_{i\to\infty} if_i = 0$. From

$$f_i = \frac{i-1}{\rho + i} f_{i-1}, \tag{42}$$

one can derive the recurrence relation

$$\left(1 + \frac{1}{\rho}\right) f_i = (i-1)\frac{1}{\rho}(f_{i-1} - f_i) \tag{43}$$

valid for any $i \geq 2$, whereas, for $i = 1$, one has

$$\left(1 + \frac{1}{\rho}\right) f_1 = 1. \tag{44}$$

Now define $S_n = \sum_{i=1}^{n} f_i$. Using the previous two equations (43) and (44), one arrives at

$$\left(1 + \frac{1}{\rho}\right) S_n = 1 + \frac{1}{\rho} S_n - \frac{n}{\rho} f_n. \tag{45}$$

Average cluster dynamics and Yule distribution VIII

In order to derive (45), one must write equation (43) for $i = 2, \ldots, n$ and add all the $(n-1)$ equations and finally add also equation (44). Solving equation (45) for S_n leads to

$$\sum_{i=1}^{n} f_i = 1 - \frac{n}{\rho} f_n, \tag{46}$$

so that one finds

$$\sum_{i=1}^{\infty} f_i = \lim_{n\to\infty} \left(1 - \frac{n}{\rho} f_n\right) = 1. \tag{47}$$

In summary, equation (41) defines a legitimate probability distribution on the integers $i \geq 1$ called *Yule distribution*.

Average cluster dynamics and Yule distribution IX

Let Y be a random variable distributed according the Yule distribution, then one has

$$\mathbb{P}(Y = i) = f_i = \rho B(i, \rho + 1); \tag{48}$$

in particular, from equation (38), one can see that the ratio $\bar{z}^*_{i,n}/nu$ coincides with the Yule distribution. For the sake of completeness, and without proof, note that, if Y follows the Yule distribution, its expected value is

$$\mathbb{E}(Y) = \sum_{i=1}^{\infty} i f_i = \frac{\rho}{\rho - 1}, \tag{49}$$

and its variance is

$$\mathbb{V}(Y) = \frac{\rho^2}{(\rho - 1)^2(\rho - 2)}. \tag{50}$$

It is now time to summarize the previous results. If at each step a unit is added to the system (a word is added to the text) according to equation (24), the expected number of clusters of size i (the expected number of words with i occurrences) converges to $\bar{z}^*_{i,n}/nu$. Note that $\bar{z}_{i,n} = 0$ for $i > n$ and the convergence holds true only for $n \gg i$. In other words, considering a fixed size i, if the population grows, it is possible to find a size $\bar{n}$ such that for $n > \bar{n}$, one has $\bar{z}_{i,n} \propto n$ reaching a steady growth. Eventually, one can see that for $i \gg \rho$

$$f_i = \rho B(i, \rho + 1) = \rho \frac{\Gamma(i)\Gamma(\rho + 1)}{\Gamma(i + \rho + 1)} = \rho \frac{\Gamma(\rho + 1)}{i(i + 1) \cdots (i + \rho)} \approx \frac{\rho \Gamma(\rho + 1)}{i^{\rho+1}}, \tag{51}$$

so that one gets a power-law tail of the distribution with $f_i \approx i^{-(\rho+1)}$.

A Monte Carlo simulation

This is the background for the exercise session. In order to simulate the dynamics of clusters, one can consider a system of n elements which can be found in $g = n + 1$ sites, assuming that the initial state is $\mathbf{Y}^* = (1, 1, \ldots, 1, 0)$. At each step a cluster is removed. If k is the number of active clusters, for each cluster, the probability of removal is simply given by $1/k$. If m is the size of the removed cluster, these items are returned to the system following Zipf's scheme. Accommodations are such that elements either join existing clusters (proportionally to their size) or move to free sites with innovation probability u. However, they cannot join the cluster which was just destroyed. The herding probabilities of joining already existing clusters sum up to $1 - u$.

Further reading


J. Bertoin, Exchangeable Coalescents, Lecture Notes 2010.


U. Garibaldi and E. Scalas, Finitary Probabilistic Methods in Econophysics, Cambridge University Press, 2010.


J. Pitman, Combinatorial Stochastic Processes, Lecture Notes 2006.