REVSTAT – Statistical Journal
Volume 6, Number 3, November 2008, 199–230

SPECIAL CHARACTERIZATIONS OF STANDARD DISCRETE MODELS

Authors: Carlos Alberto de Bragança Pereira
– Department of Statistics, University of São Paulo, Brazil
cpereira@ime.usp.br

Julio Michael Stern
– Department of Applied Mathematics, University of São Paulo, Brazil
jstern@ime.usp.br

Received: June 2007 Revised: April 2008 Accepted: April 2008

Abstract:

- This article presents important properties of standard discrete distributions and its conjugate densities. The Bernoulli and Poisson processes are described as generators of such discrete models. A characterization of distributions by mixtures is also introduced.

  This article adopts a novel singular notation and representation. Singular representations are unusual in statistical texts. Nevertheless, the singular notation makes it simpler to extend and generalize theoretical results and greatly facilitates numerical and computational implementation.

Key-Words:

- *Characterization of discrete distribution; conjugate densities; singular representations.*

AMS Subject Classification:

- 62-02, 60E05, 62F15.

1. INTRODUCTION AND NOTATION

This article presents important properties of the distributions used for categorical data analysis. Regardless of the population size being known or unknown, or the specific observational stopping rule, the Bernoulli Processes generates the sampling distributions considered. On the other hand, the Gamma distribution generates the prior and posterior distributions obtained: Gamma, Gamma-Poisson, Dirichlet, and Dirichlet-Multinomial. The Poisson Processes as generator of sampling distributions is also considered.

The development of the theory in this article is self contained, seeking a unified treatment of a large variety of problems, including finite and infinite populations, contingency tables of arbitrary dimension, deficiently categorized data, logistic regressions, etc. These models also present a way of introducing non parametric solutions.

This article adopts a singular notation and representation, first used in Pereira and Stern (2005). Singular representations are unusual in statistical texts. Nevertheless, the singular notation makes it simpler to extend and generalize theoretical results and greatly facilitates numerical and computational implementation.

The generation form of the discrete sampling distributions presented in Section 2 is, in fact, a characterization method of such distributions. If one recalls that all the distribution classes being mixed are complete classes and are Blackwell sufficient for the Bernoulli processes, the mixing distributions are unique. This characterization method is completely described in Basu and Pereira (1983).

Section 9 describes the Reny–Aczel characterization of the Poisson distribution. Although it could be thought as a de Finetti type characterization this characterization is based on alternative requirements. While de Finetti characterization is based on a permutable infinite 0-1 process, Reny–Aczek characterization is based on a homogeneous Markov process in a finite interval, generating finite discrete Markov Chains. Using Reny–Aczel characterization, together with Theorem 3.1, one can obtain a characterization of Multinomial distributions.

Section 7 describes the Dirichlet of Second Kind. In this section we also show how to use a multivariate normal approximation to the logarithm of a random vector distributed as Dirichlet of Second Kind, and a log-normal approximation to a Gamma distribution, see Aitchison and Shen (1980). In many examples of the authors' consulting practice these approximations proved to be a powerful modeling tool, leading to efficient computational procedures.

Let us first define some matrix notation. The operator $f:s:t$, to be read *from f to t* with *step s*, indicates the vector $\left[f, f+s, f+2s, ..., t\right]$ or the corresponding index domain. $f:t$ is a short hand for $f:1:t$. Usually we write a matrix, A, with subscript row index and superscript column index. Hence, A_i^j is the element in the i-th row and j-th column of matrix A. Index vectors can be used to build a matrix by extracting from a larger matrix a given sub-set of rows and columns. For example, $A_{1:m/2}^{n/2:n}$ is the northeast block, i.e. the block with the first rows and last columns, from A. Alternatively, we may write a matrix with row and column indices in parenthesis. Hence, we may write the northeast block as $A(1:m/2, n/2:n)$. The next example shows a more general case of this notation:

$$A = \begin{bmatrix} 11 & 12 & 13 \\ 21 & 22 & 23 \\ 31 & 32 & 33 \end{bmatrix} , \qquad r = \begin{bmatrix} 1 & 3 \end{bmatrix} , \qquad s = \begin{bmatrix} 3 & 1 & 2 \end{bmatrix} ,$$

$$A_r^s = A(r,s) = \begin{bmatrix} 13 & 11 & 12 \\ 33 & 31 & 32 \end{bmatrix} .$$

$V > 0$ is a positive definite matrix. The Diagonal operator, diag, if applied to a square matrix, extracts the main diagonal as a vector, and if applied to a vector, produces the corresponding diagonal matrix:

$$\operatorname{diag}(A) = \begin{bmatrix} A_1^1 \\ A_2^2 \\ \vdots \\ A_n^n \end{bmatrix} , \qquad \operatorname{diag}(a) = \begin{bmatrix} a_1 & 0 & ... & 0 \\ 0 & a_2 & ... & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & ... & a_n \end{bmatrix} .$$

A list of matrices can be indexed with left subscript or superscript indices. In case of block matrices, these left indices indicate the row and column block position, like in the following example:

$$A = \begin{bmatrix} {}_1^1A & {}_1^2A & ... & {}_1^sA \\ {}_2^1A & {}_2^2A & ... & {}_2^sA \\ \vdots & \vdots & \ddots & \vdots \\ {}_r^1A & {}_r^2A & ... & {}_r^sA \end{bmatrix} .$$

Hence, ${}_r^sA_i^j$ is the element in the i-th row and j-th column of the block situated at the r-th block of rows and s-th block of columns of matrix A. Alternatively, we may write block indices in braces, that is, we may write ${}_r^sA_i^j$ as $A\{r,s\}(i,j)$.

The Vec operator stacks the columns of the argument matrix in a single vector. The Kronecker product, also known as direct or tensor product, is defined as follows:

$$\operatorname{Vec}(U^{1\,:\,n}) = \begin{bmatrix} u^1 \\ u^2 \\ \vdots \\ u^n \end{bmatrix} , \qquad A \otimes B = \begin{bmatrix} A_1^1 B & A_1^2 B & ... & A_1^n B \\ A_2^1 B & A_2^2 B & ... & A_2^n B \\ \vdots & \vdots & \ddots & \vdots \\ A_m^1 B & A_m^2 B & ... & A_m^n B \end{bmatrix} .$$

We now introduce some concepts and notations related to the permutation and partition of indices. Let $1\!:\!m$ be an index domain or, in this article context, a classification index. Let $p = \sigma(1\!:\!m)$ be a permutation of these indices. The corresponding (Row) Permutation Matrix is

$$P = I_p = \begin{bmatrix} I_{p(1)} \\ \vdots \\ I_{p(m)} \end{bmatrix} , \qquad \text{hence} , \qquad P \begin{bmatrix} 1 \\ \vdots \\ m \end{bmatrix} = \begin{bmatrix} p(1) \\ \vdots \\ p(m) \end{bmatrix} .$$

A permutation vector, p, and a termination vector, t, define a partition of the m original classes in s super-classes:

$$\begin{bmatrix} p(1) \\ \vdots \\ p\big(t(1)\big) \end{bmatrix} , \qquad \begin{bmatrix} p\big(t(1)+1\big) \\ \vdots \\ p\big(t(2)\big) \end{bmatrix} , \qquad \dots , \qquad \begin{bmatrix} p\big(t(s-1)+1\big) \\ \vdots \\ p\big(t(s)\big) \end{bmatrix} ,$$

$$\text{where} \qquad t(0) = 0 < t(1) < ... < t(s-1) < t(s) = m \ .$$

We define the corresponding permutation and partition matrices, P and T, as

$$P = I_{p(1\,:\,m)} = \begin{bmatrix} {}_1P \\ {}_2P \\ \vdots \\ {}_sP \end{bmatrix} , \qquad {}_rP = I_{p\big(t(r-1)+1\,:\,t(r)\big)} ,$$

$$T_r = \mathbf{1}'({}_rP) \qquad \text{and} \qquad T = \begin{bmatrix} T_1 \\ \vdots \\ T_s \end{bmatrix} .$$

These matrices facilitate writing functions of a given partition, like

- The class indices in the super-class r

$${}_rP\,(1\!:\!m) \ = \ {}_rP \begin{bmatrix} 1 \\ \vdots \\ m \end{bmatrix} = \begin{bmatrix} p\big(t(r-1)+1\big) \\ \vdots \\ p\big(t(r)\big) \end{bmatrix} ;$$

- The number of classes in the super class r

$$T_r\,\mathbf{1} \ = \ t(r) - t(r-1) \ ;$$

- A sub-matrix with the row indices in super-class r

$$_{r}P\,A \;=\; \begin{bmatrix} A_{p\left(t(r-1)+1\right)} \\ \vdots \\ A_{p\left(t(r)\right)} \end{bmatrix} \;;$$

- The summation of the rows of a submatrix with row indices in super-class r

$$T_r\,A \;=\; \mathbf{1}'({}_{r}P\,A)\,;$$

- The rows of a matrix, added over each super-class

$$T\,A \;=\; \begin{bmatrix} T_1\,A \\ \vdots \\ T_s\,A \end{bmatrix} \;.$$

Note that a matrix T represents a partition of m-classes into s-super-classes if T has dimension $s\!\times\!m$, $T_h^j \in \{0,1\}$ and T has orthogonal rows. The element T_h^j indicates if the class $j \in 1\!:\!m$ is in super-class $h \in 1\!:\!s$.

We introduce the following notation for observation matrices, and respective summation vectors:

$$U = \left[u^1, u^2, ...\right] \,, \qquad U^{1\,:\,n} = \left[u^1, u^2, ..., u^n\right] \,, \qquad x^n = U^{1\,:\,n}\,\mathbf{1} = \sum\nolimits_{j=1}^{n} u^j \;.$$

The tilde accent indicates some form of normalization like, for example, $\widetilde{x} = (1/\mathbf{1}'x)\,x$.

Lemma 1.1. *If $u^1, ..., u^n$ are i.i.d. random vectors,*

$$x = U^{1\,:\,n}\,\mathbf{1} \;\Longrightarrow\; \mathrm{E}(x) = n\,\mathrm{E}(u^1) \;\;\textit{and}\;\; \mathrm{Cov}(x) = n\,\mathrm{Cov}(u^1) \;.$$

Proof: The first result is trivial. For the second result, we only have to remember the transformation properties for the expectation and covariance operators by a linear operation on their argument,

$$\mathrm{E}(AY + b) = A\,\mathrm{E}(Y) + b \;, \qquad \mathrm{Cov}(AY + b) = A\,\mathrm{Cov}(Y)\,A' \;,$$

and write

$$\begin{aligned} \mathrm{Cov}(x) &= \mathrm{Cov}\big(U^{1\,:\,n}\,\mathbf{1}\big) \\ &= \mathrm{Cov}\Big(\big(\mathbf{1}'\otimes I\big)\;\mathrm{Vec}\big(U^{1\,:\,n}\big)\Big) \;=\; \big(\mathbf{1}'\otimes I\big)\,\big(I\otimes \mathrm{Cov}(u^1)\big)\,\big(\mathbf{1}\otimes I\big) \\ &= \big(\mathbf{1}'\otimes \mathrm{Cov}(u^1)\big)\,\big(\mathbf{1}\otimes I\big) \;=\; n\,\mathrm{Cov}(u^1) \;. \end{aligned}$$

□

2. THE BERNOULLI PROCESS

Let us consider a sequence of random vectors $u^1, u^2, ...$ where, $\forall u^i$ can assume only two values

$$I^1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \quad \text{or} \quad I^2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix} , \qquad \text{where} \;\; I = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} ,$$

representing success or failure. That is, u^i can assume the value of any column of the identity matrix, I. We say that u^i is of class k, $c(u^i) = k$, iff $u^i = I^k$, $k \in [1, 2]$.

Also assume that (in your opinion), this sequence is exchangeable, that is, if $p = \big[p(1), p(2), ..., p(n)\big]$ is a permutation of $[1, 2, ..., n]$, then, $\forall\, n, p$,

$$\Pr\big(u^1, ..., u^n\big) \; = \; \Pr\big(u^{p(1)}, ..., u^{p(n)}\big) \; .$$

Just from this exchangeability constraint, that can be interpreted as saying that the index labels are non informative, de Finetti Theorem establishes the existence of an unknown vector

$$\theta \in \Theta \;\; = \;\; \left\{ \mathbf{0} \le \theta = \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix} \le \mathbf{1} \;\middle|\; \mathbf{1}'\theta = 1 \right\}$$

such that, conditionally on θ, $u^1, u^2, ...$ are mutually independent, and the conditional probability of $\Pr(u^i = I^k \mid \theta)$ is θ_k, i.e.

$$\big(u^1 \amalg u^2 \amalg ...\big) \,\big|\, \theta \quad \text{or} \quad \coprod_{i=1}^{\infty} u_i \,|\, \theta \; , \qquad \text{and} \qquad \Pr\big(u^i = I^k \,|\, \theta\big) \; = \; \theta_k \; .$$

Vector θ is characterized as the limit of proportions

$$\theta = \lim_{n\to\infty} \frac{1}{n}\, x^n \; , \qquad x^n \; = \; U^{1\,:\,n}\,\mathbf{1} = \sum\nolimits_{j=1}^{n} u^j \; .$$

Conditionally on θ, the sequence $u^1, u^2, ...$ receives the name of Bernoulli process. As we shall see, many well known discrete distributions can be obtained from transformations of this process.

The expectation and covariance (conditionally on θ) of any vector in the sequence are:

- $\mathrm{E}(u^i) = \theta\,$;
- $\mathrm{Cov}(u^i) = \mathrm{E}\big(u^i \otimes (u^i)'\big) - \mathrm{E}(u^i) \otimes \mathrm{E}\big((u^i)'\big) = \mathrm{diag}(\theta) - \theta \otimes \theta'\,.$

When the summation domain $1\!:\!n$ is understood, we may use the relaxed notation x instead of x^n. We also define the Delta operator, or "pointwise power product" between two vectors of same dimension: Given θ, and x, $n \times 1$,

$$\theta \bigtriangleup x \equiv \prod_{i=1}^{n} (\theta_i)^{x_i} .$$

A stopping rule, δ, establishes, for every $n = 1, 2, ...$, a decision of observing (or not) u^{n+1}, after the observations $u^1, ..., u^n$.

For a good understanding of this text, it is necessary to have a clear interpretation of conditional expressions like $x^n | n$ or $x_2^n | x_1^n$. In both cases we are referring to a unknown vector, x^n, but with a different partial information. In the first case, we know n, and therefore we know the sum of components, $x_1^n + x_2^n = n$; however, we know neither component x_1^n nor x_2^n. In the second case we only know the first component, of x^n, x_1^n, and do not know the second component, x_2^n, obviously we also do not know the sum, $n = x_1^n + x_2^n$. Just pay attention: We list what we know to the right of the bar and, (unless we have some additional information) everything that can not be deduced from this list is unknown.

The first distribution we are going to discuss is the Binomial. Let $\delta(n)$ be the stopping rule where n is the pre-established number of observations. The (conditional) probability of the observation sequence $U^{1:n}$ is

$$\Pr\big(U^{1:n} \,|\, \theta\big) = \theta \bigtriangleup x^n .$$

The summation vector, x^n, has Binomial distribution with parameters n and θ, and we write $x^n \,|\, [n, \theta] \sim \mathrm{Bi}(n, \theta)$. When n (or $\delta(n)$) is implicit in the context we may write $x \,|\, \theta$ instead of $x^n \,|\, [n, \theta]$. The Binomial distribution has the following expression:

$$\Pr\big(x^n \,|\, n, \theta\big) = \binom{n}{x^n} \big(\theta \bigtriangleup x^n\big)$$

where

$$\binom{n}{x} \equiv \frac{\Gamma(n+1)}{\Gamma(x_1 + 1)\,\Gamma(x_2 + 1)} = \frac{n!}{x_1!\; x_2!} \qquad \text{and} \qquad n = \mathbf{1}' x .$$

It is not hard to check that expectation vector and the covariance matrix of $x^n \,|\, [n, \theta]$ have the following expressions:

$$\mathrm{E}(x^n) = n\theta \qquad \text{and} \qquad \mathrm{Cov}(x^n) = n\,(\theta \bigtriangleup \mathbf{1}) \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} .$$

The second distribution we discuss is the Negative Binomial. Let $\delta(x_1^n)$ be the rule establishing to stop at observation u^n when obtaining a pre-established number of x_1^n successes. The random variable x_2^n, the number of failures he have when we obtain the required x_1^n successes, is called a Negative Binomial with parameters x_1^n and θ. It is not hard to prove that the Negative Binomial distribution $x_2^n \,|\, [x_1^n, \theta] \sim \mathrm{NB}(x_1^n, \theta)$, has expression, $\forall\, x_2^n \in \mathbb{N}$,

$$\Pr\bigl(x^n \,|\, x_1^n, \theta\bigr) \;=\; \frac{x_1^n}{n} \binom{n}{x^n} \bigl(\theta \bigtriangleup x^n\bigr) \;=\; \theta_1 \Pr\Bigl((x^n - I^1) \,|\, (n-1),\, \theta\Bigr) \;.$$

Note that, from the definition of this distribution, x_1^n is a positive integer number. Nevertheless, we can extend the definition above for any real positive value a, and still obtain a probability function. For this, we use

$$\sum_{j=0}^{\infty} \frac{\Gamma(a+j)}{\Gamma(a)\, j!} \,(1-\pi)^j \;=\; \pi^{-a} \;, \qquad \forall\, a \in [0, \infty[\quad \text{and} \quad \pi \in \,]0,1[\;.$$

It is not hard to check the last equation, as well as the following expressions for the expectation and variance of x_2^n:

$$\mathrm{E}\bigl(x_2^n \,|\, x_1^n, \theta\bigr) = \frac{x_1^n\, \theta_2}{\theta_1} \qquad \text{and} \qquad \mathrm{Var}\bigl(x_2^n \,|\, x_1^n, \theta\bigr) = \frac{x_1^n\, \theta_2}{(\theta_1)^2} \;.$$

In the special case of $\delta(x_1^n = 1)$, the Negative Binomial distribution is also known as the Geometric distribution with parameter θ. If a random variables are independent and identically distributed (i.i.d.) as a geometric distribution with parameter θ, then the sum of these variables has Negative Binomial distribution with parameters a and θ.

The third distribution studied in this article is the Hypergeometric. Going back to the original sequence, $u^1, u^2, \ldots$, assume that a first observer knows the first N observations, while a second observer knows only a subsequence of $n < N$ of these observations. Since the original sequence, $u^1, u^2, \ldots$, is exchangeable, we can assume, without loss of generality, that the subsequence known to the second observer is the subsequence of the first n observations, $u^1, \ldots, u^n$. Using de Finetti theorem, we have that x^n and $x^N - x^n = U^{n+1\,:\,N}\mathbf{1}$ are conditionally independent, given θ. That is, $x^n \perp\!\!\!\perp (x^N - x^n) \,|\, \theta$. Moreover, we can write

$$x^n \,|\, [n, \theta] \sim \mathrm{Bi}(n, \theta) \;, \qquad x^N \,|\, [N, \theta] \sim \mathrm{Bi}(N, \theta) \qquad \text{and}$$

$$(x^N - x^n) \,\big|\, \bigl[(N-n), \theta\bigr] \;\sim\; \mathrm{Bi}(N-n, \theta) \;.$$

Our goal is to find the distribution function of $x^n \,|\, x^N$. Note that x^N is sufficient for $U^{1\,:\,N}$ given θ, and x^n is sufficient for $U^{1\,:\,n}$. Moreover $x^n \,|\, [n, x^N]$ has the same distribution of $x^n \,|\, [n, x^N, \theta]$. Using the basic rules of probability calculus and the properties above, we have that

$$\begin{aligned}\Pr\big(x^n\,|\,n, x^N, \theta\big) &= \frac{\Pr\big(x^n, x^N|\,n, N, \theta\big)}{\Pr\big(x^N|\,n, N, \theta\big)} \\ &= \frac{\Pr\big(x^n, (x^N - x^n)\,|\,n, N, \theta\big)}{\Pr\big(x^N|\,n, N, \theta\big)} \\ &= \frac{\Pr\big(x^n\,|\,n, N, \theta\big)\ \Pr\big(x^N - x^n\,|\,n, N, \theta\big)}{\Pr\big(x^N|\,n, N, \theta\big)} \ .\end{aligned}$$

Hence, $x^n\,|\,[n, x^N]$ has distribution function

$$\Pr\big(x^n\,|\,n, x^N\big) = \frac{\begin{pmatrix} n \\ x^n \end{pmatrix} \begin{pmatrix} N - n \\ x^N - x^n \end{pmatrix}}{\begin{pmatrix} N \\ x^N \end{pmatrix}}$$

$$\text{where} \quad \mathbf{0} \le x^n \le x^N \le N\mathbf{1} \ , \quad \mathbf{1}'x^n = n \ , \quad \mathbf{1}'x^N = N \ .$$

This is the vector representation of the Hypergeometric probability distribution:

$$x^n\,|\,[n, x^N] \sim \mathrm{Hy}(n, N, x^N) \ .$$

It is not hard to check the following expressions for the expectation and (conditional) covariance of $x^n\,|\,[n, N, x^N]$, and covariance of u^i and u^j, $i, j \le n$:

$$\mathrm{E}(x^n) = \frac{n}{N}\,x^N \quad \text{and} \quad \mathrm{Cov}(x^n) = \frac{n(N-n)}{(N-1)}\,\big(x^N \triangle \mathbf{1}\big) \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} ,$$

$$\mathrm{Cov}(u^i, u^j\,|x^N) = \frac{1}{(N-1)\,N^2}\,\big(x^N \triangle \mathbf{1}\big) \begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix} \ .$$

We finish this section presenting the derivation of the Beta-Binomial distribution. Let us assume that the first observer observed x_2^n failures, until observing a pre-established number of x_1^n successes. A second observer makes more observations, observing x_2^N failures until completing the pre-established number of x_1^N successes, $x_1^n < x_1^N$.

Since x_1^n and x_1^N are pre-established, we can write

$$x_2^N\,|\,\theta \sim \mathrm{NB}(x_1^N, \theta) \ , \qquad x_2^n\,|\,\theta \sim \mathrm{NB}(x_1^n, \theta) \ ,$$
$$(x_2^N - x_2^n)\,|\,\theta \sim \mathrm{NB}(x_1^N - x_1^n, \theta) \quad \text{and} \quad x_2^n \amalg (x_2^N - x_2^n)\,|\,\theta \ .$$

As before, our goal is to describe the distribution of $x_2^n\,|\,[x_1^n, x^N]$. If one notices that $[x_1^n, x^N]$ is sufficient for $[x^n, (x^N - x^n)]$, with respect to θ, the problem becomes similar to the Hypergeometric case, and one can obtain

$$\Pr\bigl(x_2^n \,|\, x_1^n, x^N\bigr) \;=\; \frac{x_2^N!\;\Gamma(x_1^N)}{\Gamma(x_2^N + x_1^N)}\;\frac{\Gamma(x_2^n + x_1^n)}{x_2^n!\;\Gamma(x_1^n)}\;\frac{\Gamma\bigl(x_2^N - x_2^n + x_1^N - x_1^n\bigr)}{(x_2^N - x_2^n)!\;\Gamma(x_1^N - x_1^n)}\;, \qquad x_2^n \in \bigl\{0, 1, ..., x_2^N\bigr\}\;.$$

This is the distribution function of a random variable called Beta Binomial with parameters x_1^n and x^N:

$$x_2^n \,|\, (x_1^n, x^N) \;\sim\; \mathrm{BB}(x_1^n, x^N)\;.$$

The properties of this distribution will be studied in the general case of the Dirichlet-Multinomial, in the following sections.

Generalized categories for $k > 2$ can be represented by the orthonormal base $I^1, I^2, ...I^k$, i.e., the columns of the k-dimensional identity matrix. The Multinomial and Hypergeometric multivariate distributions, presented in the next sections, are distributions derived of this basic generalization.

3. MULTINOMIAL DISTRIBUTION

Let u^i, $i = 1, 2, ...$, be random vectors with possible results in the set of columns of the m-dimensional identity matrix, I^k, $k \in 1\!:\!m$. We say that u^i is of class k, $c(u^i) = k$, iff $u^i = I^k$.

Let $\theta \in [0,1]^m$ be the vector of probabilities for an observation of class k in a m-variate Bernoulli process, i.e.,

$$\Pr\bigl(u^i = I^k \,|\, \theta\bigr) = \theta_k\;, \qquad \mathbf{0} \le \theta \le \mathbf{1}\,, \quad \mathbf{1}'\theta = 1\;.$$

Like in the last section, let U

$$U = [u^1, u^2, ...] \qquad \text{and} \qquad x^n = U^{1\,:\,n}\,\mathbf{1}\;.$$

Definition 3.1. If the knowledge of θ makes the vectors u^i independent, then the (conditional) distribution of x^n given θ is the Multinomial distribution of order m with parameters n and θ, given by

$$\Pr\bigl(x^n \,|\, n, \theta\bigr) \;=\; \binom{n}{x^n}\;(\theta \bigtriangleup x^n)$$

where

$$\binom{n}{x} \equiv \frac{\Gamma(n+1)}{\Gamma(x_1+1)\,\cdots\,\Gamma(x_m+1)} \;=\; \frac{n!}{x_1!\,\cdots\,x_m!} \qquad \text{and} \qquad n = \mathbf{1}'x\;.$$

We represent the m-Multinomial distribution writing

$$x^n \,|\, [n,\theta] \;\sim\; \mathrm{Mn}_m(n,\theta) \ .$$

When $m=2$, we have the binomial case.

Let us now examine some properties of the Multinomial distribution.

Lemma 3.1. *If $x|\theta \sim \mathrm{Mn}_m(n,\theta)$ then the (conditional) expectation and covariance of x are*

$$\mathrm{E}(x) = n\,\theta \qquad \textit{and} \qquad \mathrm{Cov}(x) = n\Big(\mathrm{diag}(\theta) - \theta \otimes \theta'\Big) \ .$$

Proof: Analogous to the binomial case. □

The next result presents a characterization of the Multinomial in terms of the Poisson distribution.

Lemma 3.2. *Reproductive property of the Poisson distribution.*

$$x_i \sim \mathrm{Ps}(\lambda_i) \implies \mathbf{1}'x \,|\, \lambda \;\sim\; \mathrm{Ps}(\mathbf{1}'\lambda) \ .$$

That is, the sum of (independent) Poisson variates is also Poisson.

Theorem 3.1. *Characterization of the Multinomial by the Poisson.*
Let $x = [x_1, ..., x_m]'$ be a vector with independent Poisson distributed components with parameters in the known vector $\lambda = [\lambda_1, ..., \lambda_m]' > 0$. Let n be a positive integer. Then, given λ,

$$x \,|\, [n = \mathbf{1}'x, \lambda] \;\sim\; \mathrm{Mn}_m(n,\theta) \qquad \textit{where} \;\; \theta = \frac{1}{\mathbf{1}'\lambda}\,\lambda \ .$$

Proof: The joint distribution of x, given λ is

$$\Pr(x|\lambda) \;=\; \prod_{k=1}^{m} \frac{e^{-\lambda_k}\lambda_i^{x_k}}{x_k!} \ .$$

Using the Poisson reproductive property,

$$\Pr\big(x \,|\, \mathbf{1}'x = n, \lambda\big) \;=\; \frac{\Pr\big(\mathbf{1}'x = n \wedge x \,|\, \lambda\big)}{\Pr\big(\mathbf{1}'x = n \,|\, \lambda\big)} \;=\; \delta(n = \mathbf{1}'x)\,\frac{\Pr(x \,|\, \lambda)}{\Pr\big(\mathbf{1}'x = n \,|\, \lambda\big)} \ . \qquad \square$$

The following results state important properties of the Multinomial distribution. The proof of these properties is simple, using the characterization of the Multinomial by the Poisson, and the Poisson reproductive property.

Theorem 3.2. *Multinomial Class Partition.*

Let $1\!:\!m$ be the index domain for the classes of a order m Multinomial distribution. Let T be a partition matrix breaking the m-classes into s-super-classes. Let $x \sim \mathrm{Mn}_m(n,\theta)$, then $y = Tx \sim \mathrm{Mn}_s(n, T\theta)$.

Theorem 3.3. *Multinomial Conditioning on the Partial Sum.*

If $x \sim \mathrm{Mn}_m(n,\theta)$, then the distribution of part of the vector x conditioned on its sum has Multinomial distribution, having as parameter the corresponding part of the original (normalized) parameters. In more detail, conditioning on the t first components, we have:

$$x_{1\,:\,t}\,|\,(\mathbf{1}'x_{1\,:\,t} = j) \;\sim\; \mathrm{Mn}_t\biggl(j\,,\, \frac{1}{\mathbf{1}'\theta_{1\,:\,t}}\,\theta_{1\,:\,t}\biggr) \qquad \textit{where}\quad 0 \le j \le n\ .$$

Theorem 3.4. *Multinomial-Binomial Decomposition.*

Using the last two theorems (3.2 and 3.3), if $x \sim \mathrm{Mn}_m(n,\theta)$,

$$\begin{aligned}\Pr(x\,|\,n,\theta) \;=\; \sum_{j=0}^{n} &\Pr\biggl(x_{1\,:\,t}\,|\,j\,,\, \frac{1}{\mathbf{1}'\theta_{1\,:\,t}}\,\theta_{1\,:\,t}\biggr)\\ &\cdot\;\Pr\biggl(x_{t+1\,:\,m}\,|\,(n-j)\,,\, \frac{1}{\mathbf{1}'\theta_{t+1\,:\,m}}\,\theta_{t+1\,:\,m}\biggr)\\ &\cdot\;\Pr\biggl(\begin{bmatrix} j \\ (n-j)\end{bmatrix}\bigg|\,n\,,\, \begin{bmatrix}\mathbf{1}'\theta_{1\,:\,t}\\ \mathbf{1}'\theta_{t+1\,:\,m}\end{bmatrix}\biggr)\ .\end{aligned}$$

Analogously, we could write the Multinomial-Trinomial decomposition for a three-partition of the class indices in three super-classes. More generally, we could also write the m-nomial-s-nomial decomposition for the partition of the m class indices into s super-classes.

4. MULTIVARIATE HYPERGEOMETRIC DISTRIBUTION

In the second section we have shown how an Hypergeometric variate can be generated from a Bernoulli process. The natural generalization of this result is obtained considering a Multinomial process. As in the last section, we say that u^i is of class k, $c(u^i) = k$, iff $u^i = I^k$.

We take a sample of size n from a finite population of size $N\,(> n)$, that is partitioned into m classes. The population frequencies (number of elements in each category) are represented by $[\psi_1, ..., \psi_m]$, hence $N = \mathbf{1}'\psi$. Based on the sample, we want to make an inference on ψ. x_k is the sample frequency of class k.

One way of describing this problem is to consider an urn with N balls of m different colors, indexed by $1, ..., m$. ψ_k is the number of balls of color k. Assume that the N balls are separated into two smaller boxes, so that box 1 has n balls and box 2 has the remaining $N-n$ balls. The statistician can observe the composition of box 1, represented by vector x of sample frequencies. The quantity of interest for the statistician is the vector $\psi - x$ representing the composition of box 2.

As in the bivariate case, we assume that $U^{1:N}$ is a finite sub-sequence in an exchangeable process and, therefore, any sub-sequence extracted from $U^{1:N}$ has the same distribution of $U^{1:n}$. Hence, $x = U^{1:n}\mathbf{1}$ has the same distribution of the frequency vector for a sample of size n.

As in the bivariate case, our objective is to find the distribution of $x|\psi$. Again, using de Finetti theorem, there is a vector $\mathbf{0} \leq \theta \leq \mathbf{1}$, $\mathbf{1}'\theta = 1$, such that $\coprod_{j=0}^{N} u^j \,|\, \theta$ and $\Pr\bigl(c(u^j) = k\bigr) = \theta_k$.

Theorem 4.1. *As in the Multinomial case, the following results follow:*

- $\psi|\theta \sim \mathrm{Mn}_m(N, \theta)$;
- $x|\theta \sim \mathrm{Mn}_m(n, \theta)$;
- $(\psi - x)\,|\,\theta \sim \mathrm{Mn}_m\bigl((N-n), \theta\bigr)$;
- $(\psi - x) \amalg x|\theta$.

Using the results of the last section and following the same steps as in the Hy_2 case in the first section, we obtain the following expression for m-variate Hypergeometric distribution, $x^n \,|\, [n, N, \psi] \sim \mathrm{Hy}_m(n, N, \psi)$:

$$\Pr\bigl(x^n | n, \psi\bigr) = \frac{\begin{pmatrix} n \\ x^n \end{pmatrix} \begin{pmatrix} N-n \\ \psi - x^n \end{pmatrix}}{\begin{pmatrix} N \\ \psi \end{pmatrix}}$$

$$\text{where} \quad \mathbf{0} \leq x^n \leq \psi \leq N\mathbf{1}, \quad \mathbf{1}'x^n = n, \quad \mathbf{1}'\psi = N .$$

This is the vector representation of the Hypergeometric probability distribution:

$$x^n \,|\, [n, x^N] \sim \mathrm{Hy}(n, N, x^N) .$$

Alternatively, we can write the more usual formula,

$$\Pr(x \,|\, \psi) = \frac{\begin{pmatrix} \psi_1 \\ x_1 \end{pmatrix} \begin{pmatrix} \psi_2 \\ x_2 \end{pmatrix} \cdots \begin{pmatrix} \psi_m \\ x_m \end{pmatrix}}{\begin{pmatrix} N \\ n \end{pmatrix}} .$$

Theorem 4.2. *The expectation and covariance of a random vector with Hypergeometric distribution, $x \sim \mathrm{Hy}_m(n, N, \psi)$, are:*

$$\mathrm{E}(x) = n\widetilde{\psi}\,, \qquad \mathrm{Cov}(x) = n\,\frac{N-n}{N-1}\left(\mathrm{diag}(\widetilde{\psi}) - \widetilde{\psi} \otimes \widetilde{\psi}'\right) \qquad \textit{where} \;\; \widetilde{\psi} = \frac{1}{N}\,\psi\ .$$

Proof: Use that

$$\begin{aligned}
\mathrm{Cov}(x^n) &= n\,\mathrm{Cov}(u^1) + n(n-1)\,\mathrm{Cov}(u^1, u^2)\ ,\\
\mathrm{Cov}(u^1) &= \mathrm{E}\big(u^1 \otimes (u^1)'\big) - \mathrm{E}(u^1) \otimes \mathrm{E}(u^1)' \;=\; \mathrm{diag}(\widetilde{\psi}) - \widetilde{\psi} \otimes \widetilde{\psi}'\\
\mathrm{Cov}(u^1, u^2) &= \mathrm{E}\big(u^1 \otimes (u^2)'\big) - \mathrm{E}(u^1) \otimes \mathrm{E}(u^2)'\ .
\end{aligned}$$

The second term of the last two equations are equal, and the first term of the last equation is

$$\mathrm{E}(u^1_i\, u^2_j) = \begin{cases} \dfrac{\psi_i}{N}\,\dfrac{\psi_i - 1}{N-1} & \text{if} \quad i = j\,,\\[2ex] \dfrac{\psi_i}{N}\,\dfrac{\psi_j}{N-1} & \text{if} \quad i \neq j\ . \end{cases}$$

Algebraic manipulation yields the result. □

Note that, as in the order 2 case, the diagonal elements of $\mathrm{Cov}(u^1)$ are positive, while the diagonal elements of $\mathrm{Cov}(u^1, u^2)$ are negative. In the off diagonal elements, the signs are reversed.

5. DIRICHLET DISTRIBUTION

In the second section we presented the multinomial distribution, $\mathrm{Mn}_m(n, \theta)$. In this section we present the Dirichlet distribution for the parameter θ. Let us first recall the univariate Poisson and Gamma distributions.

A random variable has Gamma distribution, $x \,|\, [a, b] \sim G(a, b)$, $a, b > 0$, if its distribution is continuous with density

$$f(x|a, b) \;=\; \frac{b^a}{\Gamma(a)}\,x^{a-1}\exp(-bx)\ , \qquad x > 0\ .$$

The expectation and variance of this variate are

$$E(x) = \frac{a}{b} \qquad \text{and} \qquad \mathrm{Var}(x) = \frac{a}{b^2}\ .$$

Lemma 5.1. *Reproductive property for the Gamma distribution.*
If n independent random variables $x_i \,|\, a_i, b \sim G(a_i, b)$, then

$$\mathbf{1}'x \;\sim\; G(\mathbf{1}'a, b)\ .$$

Lemma 5.2. *The Gamma distribution is conjugate to the Poisson distribution.*

Proof: If $y|\lambda \sim \mathrm{Ps}(\lambda)$ and λ has prior $\lambda|a,b \sim G(a,b)$, then

$$\begin{aligned} f(\lambda|y,a,b) \;&\propto\; L(\lambda|y)\, f(\lambda) \;= \\ &=\; \exp(-\lambda)\,\frac{\lambda^y}{y!}\;\frac{b^a}{\Gamma(a)}\,\lambda^{a-1}\exp(-b\lambda) \;\propto\; \lambda^{y+a-1}\exp\bigl(-(b+1)\,\lambda\bigr)\;. \quad \square \end{aligned}$$

That is, the posterior distribution of λ is Gamma with parameters $[a+y,\, b+1]$.

Definition 5.1. Dirichlet distribution.
A random vector

$$y \in \mathcal{S}_{m-1} \;\equiv\; \Bigl\{ y \in \mathbb{R}^m \mid\; \mathbf{0} \le y \le \mathbf{1} \;\wedge\; \mathbf{1}'y = 1 \Bigr\}$$

has Dirichlet distribution of order m with positive $a \in \mathbb{R}^m$ if its density is

$$\Pr(y|a) \;=\; \frac{y \bigtriangleup (a-\mathbf{1})}{B(a)}\;.$$

Note that $\mathcal{S}_{m-1}$, the $m-1$ dimensional Simplex, is the region of $\mathbb{R}^m$ subject to the "constraint", $\mathbf{1}'y = 1$. Hence, a point in the Simplex has only $m-1$ "degrees of freedom". In this sense we say that the Dirichlet distribution has a "singular" representation. It is possible to give a non-singular representation to the distribution $[y_1, ..., y_{m-1}]'$, known as the Multivariate Beta distribution, but at the cost of obtaining a convoluted algebraic formulation that also loses the natural geometric interpretation of the singular form.

The normalization factor for the Dirichlet distribution is

$$B(a) \;\equiv\; \int_{y\in\mathcal{S}_{m-1}} \Bigl(y \bigtriangleup (a-1) \Bigr)\, dy\;.$$

Lemma 5.3. *Beta function.*
The normalization factor for the Dirichlet distribution defined above is the Beta function, defined as

$$B(a) \;=\; \frac{\prod_{k=1}^m \Gamma(a_k)}{\Gamma(\mathbf{1}'a)}\;.$$

The proof is given at the end of this section.

Theorem 5.1. *Dirichlet as Conjugate of the Multinomial.*
If $\theta \sim \mathrm{Di}_m(a)$ *and* $x|\theta \sim \mathrm{Mn}_m(n,\theta)$ *then*

$$\theta|x \sim \mathrm{Di}_m(a+x) \ .$$

Proof: We only have to remember that the Multinomial likelihood is proportional to $\theta \triangle x$, and that a Dirichlet prior is proportional to $\theta \triangle (a-\mathbf{1})$. Hence, the posterior is proportional to $\theta \triangle (x+a-1)$. At the other hand, $B(a+x)$ is the normalization factor, i.e., equal to the integral on θ of $\theta \triangle (x+a-1)$, and so we have a Dirichlet density function, as defined above. □

Theorem 5.2. *Dirichlet Moments.*
If $\theta \sim \mathrm{Di}_m(a)$ *and* $p \in \mathbb{N}^m$, *then*

$$\mathrm{E}(\theta \triangle p) = \frac{B(a+p)}{B(a)} \ .$$

Proof:

$$\begin{aligned}\int_\Theta (\theta \triangle p)\, f(\theta|a)\, d\theta &= \frac{1}{B(a)} \int_\Theta (\theta \triangle p) \Big(\theta \triangle (a-1)\Big)\, d\theta \\ &= \frac{1}{B(a)} \int_\Theta \Big(\theta \triangle (a+p-1)\Big)\, d\theta = \frac{B(a+p)}{B(a)} \ . \qquad \square\end{aligned}$$

Choosing the exponents, p, appropriately, we have

Corollary 5.1. *If* $\theta \sim \mathrm{Di}_m(a)$, *then*

$$\begin{aligned}\mathrm{E}(\theta) &= \widetilde{a} \equiv \frac{1}{\mathbf{1}'a}\, a \ , \\ \mathrm{Cov}(\theta) &= \frac{1}{\mathbf{1}'a+1}\Big(\mathrm{diag}(\widetilde{a}) - \widetilde{a} \otimes \widetilde{a}'\Big) \ .\end{aligned}$$

Theorem 5.3. *Characterization of the Dirichlet by the Gamma.*
Let the components of the random vector $x \in \mathbb{R}^m$ *be independent variables with distribution* $G(a_k, b)$. *Then, the normalized vector*

$$y = \frac{1}{\mathbf{1}'x}\, x \sim \mathrm{Di}_m(a) \ , \qquad \mathbf{1}'x \sim \mathrm{Ga}(\mathbf{1}'a) \qquad \textit{and} \qquad y \perp\!\!\!\perp \mathbf{1}'x \ .$$

Proof: Consider the normalization

$$y = \frac{1}{t}\, x \ , \qquad t = \mathbf{1}'x \ , \qquad x = t\, y \ ,$$

as a transformation of variables. Note that one of the new variables, say $y_m \equiv t(1 - y_1 \cdots - y_{m-1})$, becomes redundant.

The Jacobian matrix of this transformation is

$$J = \frac{\partial\,(x_1, x_2, ..., x_{m-1}, x_m)}{\partial\,(y_1, y_2, ..., y_{m-1}, t)} = \begin{bmatrix} t & 0 & \cdots & 0 & y_1 \\ 0 & t & \cdots & 0 & y_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & t & y_{m-1} \\ -t & -t & \cdots & -t & 1 - y_1 \cdots - y_{m-1} \end{bmatrix} .$$

By elementary operations that add all rows to the last one, we obtain the LU factorization of the Jacobian matrix, $J = LU$, where

$$L = \begin{bmatrix} 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \\ -1 & -1 & \cdots & -1 & 1 \end{bmatrix} \quad \text{and} \quad U = \begin{bmatrix} t & 0 & \cdots & 0 & y_1 \\ 0 & t & \cdots & 0 & y_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & t & y_{m-1} \\ 0 & 0 & \cdots & 0 & 1 \end{bmatrix} .$$

A triangular matrix determinant is equal to the product of the elements in its main diagonal, hence $|J| = |L|\,|U| = 1\,t^{m-1}$.

At the other hand, the joint distribution of x is

$$f(x) = \prod_{k=1}^{m} \mathrm{Ga}(x_k\,|\,a_k, b) = \prod_{k=1}^{m} \frac{b^{a_k}}{\Gamma(a_k)}\, e^{-bx_k} (x_k)^{a_k - 1}$$

and the joint distribution in the new system of coordinates is

$$\begin{aligned} g([y,t]) &= |J|\, f\big(x^{-1}([y,t])\big) \\ &= t^{m-1} \prod_{k=1}^{m} \frac{b^{a_k}}{\Gamma(a_k)}\, e^{-bx_k} (x_k)^{a_k-1} = t^{m-1} \prod_{k=1}^{m} \frac{b^{a_k}}{\Gamma(a_k)}\, e^{-bty_k} (ty_k)^{a_k-1} \\ &= \left(\prod_{k=1}^{m} \frac{(y_k)^{a_k-1}}{\Gamma(a_k)} \right) b^{\mathbf{1}'a}\, e^{-bt}\, t^{\mathbf{1}'a-m}\, t^{m-1} = \left(\prod_{k=1}^{m} \frac{(y_k)^{a_k-1}}{\Gamma(a_k)} \right) b^{\mathbf{1}'a}\, e^{-bt}\, t^{\mathbf{1}'a-1} . \end{aligned}$$

Hence, the marginal distribution of $y = [y_1, ..., y_k]'$ is

$$\begin{aligned} g(y) &= \int_{t=0}^{\infty} g([y,t])\, dt \\ &= \left(\prod_{k=1}^{m} \frac{(y_k)^{a_k-1}}{\Gamma(a_k)} \right) \int_{t=0}^{\infty} b^{\mathbf{1}'a}\, e^{-bt}\, t^{\mathbf{1}'a-1}\, dt \\ &= \left(\prod_{k=1}^{m} \frac{(y_k)^{a_k-1}}{\Gamma(a_k)} \right) \Gamma(\mathbf{1}'a) = \frac{y \bigtriangleup (a-1)}{B(a)} . \end{aligned}$$

In the last passage, we have replaced the integral by the normalization factor of a Gamma density, $\mathrm{Ga}(\mathbf{1}'a, b)$. Hence, we obtain a density proportional to $y \bigtriangleup (a-1)$, i.e., a Dirichlet. □

In the last passage we also obtain the Dirichlet normalization factor, proving the Beta function lemma.

Lemma 5.4. *Bipartition of Indices for the Dirichlet.*

Let $1\!:\!t$, $t\!+\!1\!:\!m$ *be a bipartition of the class index domain,* $1\!:\!m$, *of an order* m *Dirichlet, in two super-classes. Let* $y \sim \mathrm{Di}_m(a)$, *and*

$$z^1 = \frac{1}{\mathbf{1}'y_{1\,:\,t}}\, y_{1\,:\,t} \ , \qquad z^2 = \frac{1}{\mathbf{1}'y_{t+1\,:\,m}}\, y_{t+1\,:\,m} \ , \qquad w = \begin{bmatrix} \mathbf{1}'y_{1\,:\,t} \\ \mathbf{1}'y_{t+1\,:\,m} \end{bmatrix} .$$

We then have $z^1 \amalg z^2 \amalg w$ *and*

$$z^1 \sim \mathrm{Di}_t(a_{1\,:\,t}) \ , \qquad z^2 \sim \mathrm{Di}_{m-t}(a_{t+1\,:\,m}) \qquad \text{and} \qquad w \sim \mathrm{Di}_2\left(\begin{bmatrix} \mathbf{1}'a_{1\,:\,t} \\ \mathbf{1}'a_{t+1\,:\,m} \end{bmatrix}\right) .$$

Proof: From the Dirichlet characterization by the Gamma we can imagine that the vector y is built by normalizing of a vector x, as follows:

$$y = \frac{1}{\mathbf{1}'x}\, x \ , \qquad x_k \sim \mathrm{Ga}(a_k, b) \ , \qquad \coprod_{k=1}^{m} x_k \ .$$

Considering separately each one of the super-classes, we build the vectors z^1 and z^2 that are distributed as

$$\begin{aligned} z^1 &= \frac{1}{\mathbf{1}'y_{1\,:\,t}}\, y_{1\,:\,t} = \frac{1}{\mathbf{1}'x_{1\,:\,t}}\, x_{1\,:\,t} \sim \mathrm{Di}_t(a_{1\,:\,t}) \ , \\ z^2 &= \frac{1}{\mathbf{1}'y_{t+1\,:\,m}}\, y_{t+1\,:\,m} = \frac{1}{\mathbf{1}'x_{t+1\,:\,m}}\, x_{t+1\,:\,m} \sim \mathrm{Di}_{m-t}(a_{t+1\,:\,m}) \ . \end{aligned}$$

$z^1 \amalg z^2$, that are in turn independent of the partial sums

$$\mathbf{1}'x_{1\,:\,t} \sim \mathrm{Ga}(\mathbf{1}'a_{1\,:\,t}, b) \qquad \text{and} \qquad \mathbf{1}'x_{t+1\,:\,m} \sim \mathrm{Ga}(\mathbf{1}'a_{t+1\,:\,m}, b) \ .$$

Using again the theorem characterizing the Dirichlet by the Gamma distribution for these two Gamma variates, we obtain the result. □

We can generalize this result for any partition of the set of classes, as follows. If $y \sim \mathrm{Di}_m(a)$ and T is a s-partition of the m classes, the intra and extra super-class distributions are independent Dirichlets, as follows:

$$\begin{aligned} z^r &= \frac{1}{T_r y}\, {}_rPy \sim \mathrm{Di}_{T_r 1}({}_rPa) \ , \\ w &= Ty \sim \mathrm{Di}_s(Ta) \ . \end{aligned}$$

6. DIRICHLET-MULTINOMIAL

We say that a random vector $x \in \mathbb{N}^n \mid \mathbf{1}'x = n$ has Dirichlet-Multinomial (DM) distribution with parameters n and $a \in \mathbb{R}^m$, iff

$$\Pr(x|n,a) \;=\; \frac{B(a+x)}{B(a)} \binom{n}{x} \;=\; \frac{B(a+x)}{B(a)\,B(x)} \, \frac{1}{x \triangle \mathbf{1}} \;.$$

Theorem 6.1. *Characterization of the DM as a Dirichlet mixture of Multinomials.*

If $\theta \sim \mathrm{Di}_m(a)$ *and* $x|\theta \sim \mathrm{Mn}(n,\theta)$ *then* $x \mid [n,a] \sim \mathrm{DM}_m(n,a)$.

Proof: The joint distribution of θ, x is proportional to $\theta \triangle (a+x-1)$, which integrated on θ is $B(a+x)$. Hence, multiplying by the joint distribution constants, we have the marginal for x, Q.E.D. Therefore, we have also proved that the function DM is normalized, that is

$$\begin{aligned} \Pr(x) \;&=\; \int_{\theta \in \mathcal{S}_{m-1}} \binom{n}{x} (\theta \triangle x) \, \frac{1}{B(a)} \, \theta \triangle (a-\mathbf{1}) \; d\theta \\ &=\; \frac{1}{B(a)} \binom{n}{x} \int_{\theta \in \mathcal{S}_{m-1}} \Big(\theta \triangle (x+a-\mathbf{1}) \Big) \, d\theta \;=\; \frac{B(x+a)}{B(a)} \binom{n}{x} \;. \qquad \square \end{aligned}$$

Theorem 6.2. *Characterization of the DM by m Negative Binomials.*

Let $a \in \mathbb{N}^m_+$, and $x \in \mathbb{N}_m$, be a vector whose components are independent random variables, $a_k \sim \mathrm{NB}(a_k, \theta)$. Then

$$x \mid [\mathbf{1}'x = n, a] \;\sim\; \mathrm{DM}_m(n,a) \;.$$

Proof:

$$\begin{aligned} \Pr(x|\theta,a) \;&=\; \prod_{k=1}^{m} \binom{a_k + x_k - 1}{x_k} \, \theta^{a_k} (1-\theta)^{x_k} \;, \\ \Pr(\mathbf{1}'x|\theta,a) \;&=\; \binom{\mathbf{1}'a + \mathbf{1}'x - 1}{\mathbf{1}'x} \, \theta^{\mathbf{1}'a} (1-\theta)^{\mathbf{1}'a} \;. \end{aligned}$$

Then,

$$\Pr\big(x \mid \mathbf{1}'x = n,\, \theta,\, a\big) \;=\; \frac{\Pr(x|a,\theta)}{\Pr(\mathbf{1}'x = n \mid \theta)} \;=\; \frac{\prod_{k=1}^{m} \binom{a_k + x_k - 1}{x_k}}{\binom{\mathbf{1}'a + \mathbf{1}'x - 1}{\mathbf{1}'x}} \;.$$

Hence,

$$\begin{aligned}\Pr\big(x \,|\, \mathbf{1}'x = n,\, \theta,\, a\big) \;&=\; \Pr\big(x \,|\, \mathbf{1}'x = n,\, a\big) \\ &=\; \prod_{k=1}^{m} \frac{\Gamma(a_k + x_k)}{x!\; \Gamma(a_k)} \Big/ \frac{\Gamma(\mathbf{1}'a + n)}{\Gamma(\mathbf{1}'a)\; n!} \;=\; \frac{B(a+x)}{B(a)} \binom{n}{x} \,. \qquad \square\end{aligned}$$

Theorem 6.3. *The DM as Pseudo-Conjugate for the Hypergeometric.*
If $x \sim \mathrm{Hy}_m(n, N, \psi)$ *and* $\psi \sim \mathrm{DM}_m(N, a)$ *then* $(\psi - x) \,|\, x \sim \mathrm{DM}_m(N - n, a)$.

Proof: Using the properties of the Hypergeometric already presented, we have the independence relation, $(\psi - x)\,\mathrm{II}\, x|\theta$. We can therefore use the Multinomial sample $x|\theta$ for updating the prior and obtain the posterior

$$\theta | x \;\sim\; \mathrm{Di}_m(a + x) \ .$$

Hence, the distribution of the non sampled pat of the population, $\psi - x$, given the sample x, is a mixture of $(\psi - x)\theta$ by the posterior for θ. By the characterization of the DM as a mixture of Multinomials by a Dirichlet, the theorem follows, i.e.,

$$\left.\begin{array}{r} (\psi - x) \,|\, [\theta, x] \;\sim\; (\psi - x) \,|\, \theta \;\sim\; \mathrm{Mn}_m(N - n, \theta) \\ \theta | x \;\sim\; \mathrm{Di}_m(a + x) \end{array}\right\} \Longrightarrow$$

$$\Longrightarrow \;\; (\psi - x) \,|\, x \;\sim\; \mathrm{Di}_m(N - n, a + x) \ . \quad \square$$

Lemma 6.1. *DM Expectation and Covariance.*
If $x \sim \mathrm{DM}_m(n, a)$ *then*

$$\begin{aligned}\mathrm{E}(x) \;&=\; n\,\widetilde{a} \;\equiv\; \frac{1}{\mathbf{1}'a}\, a \ , \\ \mathrm{Cov}(x) \;&=\; \frac{n(n + \mathbf{1}'a)}{\mathbf{1}'a + 1} \big(\mathrm{diag}(\widetilde{a}) - \widetilde{a} \otimes \widetilde{a}'\big) \ .\end{aligned}$$

Proof:

$$\mathrm{E}(x) \;=\; \mathrm{E}_\theta\big(\mathrm{E}_x(x|\theta)\big) \;=\; \mathrm{E}_\theta(n\theta) \;=\; n\widetilde{a} \ ;$$

$$\begin{aligned}\mathrm{E}(x \otimes x') \;&=\; \mathrm{E}_\theta\big(\mathrm{E}_x(x \otimes x' \,|\, \theta)\big) \\ &=\; \mathrm{E}_\theta\Big(\mathrm{E}(x|\theta) \otimes \mathrm{E}(x|\theta)' + \mathrm{Cov}(x|\theta)\Big) \\ &=\; \mathrm{E}_\theta\Big(n\big(\mathrm{diag}(\theta) - \theta \otimes \theta'\big) + n^2\, \theta \otimes \theta'\Big) \\ &=\; n\, \mathrm{E}_\theta\big(\mathrm{diag}(\theta)\big) + n\,(n - 1)\, \mathrm{E}_\theta(\theta \otimes \theta') \\ &=\; n\; \mathrm{diag}(\widetilde{a}) + n\,(n - 1)\,\Big(\mathrm{E}(\theta) \otimes \mathrm{E}(\theta)' + \mathrm{Cov}(\theta)\Big) \\ &=\; n\; \mathrm{diag}(\widetilde{a}) + n\,(n - 1)\left(\widetilde{a} \otimes \widetilde{a}' + \frac{1}{\mathbf{1}'a + 1}\,\big(\mathrm{diag}(\widetilde{a}) - \widetilde{a} \otimes \widetilde{a}'\big)\right) \\ &=\; n\; \mathrm{diag}(\widetilde{a}) + n(n - 1)\left(\frac{1}{\mathbf{1}'a + 1}\; \mathrm{diag}(\widetilde{a}) + \frac{\mathbf{1}'a}{\mathbf{1}'a + 1}\; \widetilde{a} \otimes \widetilde{a}'\right) \ ;\end{aligned}$$

$$\begin{aligned}
\mathrm{Cov}(x) &= \mathrm{E}(x \otimes x') - \mathrm{E}(x) \otimes \mathrm{E}(x)' \;=\; \mathrm{E}(x \otimes x') - n^2 \widetilde{a} \otimes \widetilde{a}' \\
&= \left(n + \frac{n(n-1)}{\mathbf{1}'a + 1} \right) \mathrm{diag}(\widetilde{a}) \;+\; \left(n\,(n-1)\, \frac{\mathbf{1}'a}{\mathbf{1}'a + 1} - n^2 \right) \widetilde{a} \otimes \widetilde{a}' \\
&= \frac{n\,(n + \mathbf{1}'a)}{\mathbf{1}'a + 1} \left(\mathrm{diag}(\widetilde{a}) - \widetilde{a} \otimes \widetilde{a}' \right) \; . \qquad \square
\end{aligned}$$

Theorem 6.4. *DM Class Bipartition.*

Let $1\!:\!t$, $t\!+\!1\!:\!m$ a bipartition of the index domain for the classes of an order m DM, $1\!:\!m$, in two super-classes. Then, the following conditions (i) to (iii) are equivalent to condition (iv):

(**i**) $x_{1:t} \,\mathrm{II}\, x_{t+1:m} \,|\, n_1 \,=\, \mathbf{1}' x_{1:t}$;

(**ii**-1) $x_{1:t} \,|\, n_1 = \mathbf{1}' x_{1;t} \;\sim\; \mathrm{DM}_t(n_1, a_{1:t})$;

(**ii**-2) $x_{t+1:m} \,|\, n_2 = \mathbf{1}' x_{t+1:m} \;\sim\; \mathrm{DM}_{m-t}(n_2, a_{t+1:m})$;

(**iii**) $\begin{bmatrix} n_1 \\ n_2 \end{bmatrix} \sim \mathrm{DM}_2\!\left(n, \begin{bmatrix} \mathbf{1}' a_{1:t} \\ \mathbf{1}' a_{t+1:m} \end{bmatrix} \right)$;

(**iv**) $x \;\sim\; \mathrm{DM}_m(n, a)$.

Proof: We only have to show that the joint distribution can be factored in this form. By the DM characterization as a mixture, we can write it as Dirichlet mixture of Multinomials. By the bipartition theorems, we can factor both, the Multinomials and the Dirichlet, so the theorem follows. □

7. DIRICHLET OF THE SECOND KIND

Consider $y \sim \mathrm{Di}_{m+1}(a)$. The vector $z = (1/y_{m+1}) y_{1\,:\,m}$ has Dirichlet of the Second Kind (D2K) distribution.

Theorem 7.1. *Characterization of D2K by the Gamma distribution.*

Using the characterization of the Dirichlet by the Gamma, we can write the D2K variate as a function of $m+1$ independent Gamma variates,

$$z_{1\,:\,m} \sim (1/x_{m+1}) x_{1\,:\,m} \qquad \textit{where} \quad x_k \sim Ga(a_k, b) \; .$$

Similar to what we did for the Dirichlet (of the first kind), we can write the D2K distribution and its moments as:

$$\begin{aligned}
f(z|a) &= \frac{z \bigtriangleup (a_{1\,:\,m} - 1)}{(1 + \mathbf{1}'z)^{\mathbf{1}'a}\; B(a)} \; , \\
E(z) &= e \;=\; (1/a_{m+1})\, a_{1\,:\,m} \; , \\
\mathrm{Cov}(z) &= \frac{1}{a_{m+1} - 2} \left(\mathrm{diag}(e) + e \otimes e' \right) \; .
\end{aligned}$$

The logarithm of a Gamma variate is well approximated by a Normal variate, see Aitchison and Shen (1980). This approximation is the key to several efficient computational procedures, and motivates the computation of the first two moments of the log-D2K distribution. For that, we use the Digamma, $\psi(\ )$, and Trigamma function, $\psi'(\ )$, defined as:

$$\psi(a) = \frac{d}{da} \ln \Gamma(a) = \frac{\Gamma'(a)}{\Gamma(a)} \ , \qquad \psi'(a) = \frac{d}{da} \psi(a) \ .$$

Lemma 7.1. *The expectation and covariance of a log-D2K variate are:*

$$\begin{aligned} E\big(\log(z)\big) &= \psi(a_{1:m}) - \psi(a_{m+1})\mathbf{1} \ , \\ \operatorname{Cov}\big(\log(z)\big) &= \operatorname{diag}\big(\psi'(a_{1:m}) + \psi'(a_{m+1})\big)\mathbf{1} \otimes \mathbf{1}' \ . \end{aligned}$$

Proof: Consider a Gamma variate, $x \sim G(a,1)$:

$$1 = \int_0^\infty f(x)\,dx = \int_0^\infty \frac{1}{\Gamma(a)}\, x^{a-1} \exp(-x)\ dx \ .$$

Taking the derivative with respect to parameter a, we have

$$0 = \int_0^\infty \ln(x)\, x^{a-1}\, \frac{\exp(-x)}{\Gamma(a)}\ dx - \frac{\Gamma'(a)}{\Gamma^2(a)}\,\Gamma(a) = E\big(\ln(x)\big) - \psi(a) \ .$$

Taking the derivative with respect to parameter a a second time,

$$\begin{aligned} \psi'(a) &= \frac{d}{da} E\big(\ln(x)\big) = \frac{d}{da} \int_0^\infty \frac{\ln(x)}{\Gamma(a)}\, x^{a-1} \exp(-x)\ dx \\ &= \int_0^\infty \ln(x)^2\, x^{a-1}\, \frac{\exp(-x)}{\Gamma(a)}\ dx - \frac{\Gamma'(a)}{\Gamma(a)}\, E\big(\ln(x)\big) \\ &= E\big(\ln(x)^2\big) - E\big(\ln(x)\big)^2 = \operatorname{Var}\big(\ln(x)\big) \ . \end{aligned}$$

The lemma follows from the D2K characterization by the Gamma. □

8. EXAMPLES

Example 8.1. Let A, B be two attributes, each one of them present or absent in the elements of a population. Then each element of this population can be classified in exactly one of $2^2 = 4$ categories:

A	B	k	I^k
present	present	1	$[1,0,0,0]'$
present	absent	2	$[0,1,0,0]'$
absent	present	3	$[0,0,1,0]'$
absent	absent	4	$[0,0,0,1]'$

According to the notation above, we can write $x\,|\,n,\theta \sim \text{Mn}_4(n,\theta)$.

If $\theta = [0.35,\ 0.20,\ 0.30,\ 0.15]$ and $n = 10$, then

$$\Pr\big(x^{10}|n,\theta\big) \;=\; \binom{10}{x^{10}}\,(\theta \bigtriangleup x^{10})\ .$$

Hence, in order to compute the probability of $x = [1,2,3,4]'$ given θ, we use the expression above, obtaining

$$\Pr\left(\begin{bmatrix}1\\2\\3\\4\end{bmatrix}\;\middle|\;\begin{bmatrix}0.35\\0.20\\0.30\\0.15\end{bmatrix}\right) \;=\; 0.000888\ .$$

Example 8.2. If $X\,|\theta \sim \text{Mn}_3(10,\theta)$, $\theta = [0.20,\ 0.30,\ 0.15]$, one can conclude, using the result above, that

$$\text{E}(X) \;=\; (2,3,1.5)\ ,$$

while the covariance matrix is

$$\Sigma \;=\; \begin{bmatrix} 1.6 & -0.6 & -0.3\\ -0.6 & 2.1 & -0.45\\ -0.3 & -0.45 & 1.28\end{bmatrix}\ .$$

Example 8.3. Assume that $X|\theta \sim \text{Mn}_3(10,\theta)$, with $\theta = [0.20,\ 0.30,\ 0.15]$, as in Example 2. Let us take $A_0 = \{0,1\}$, $A_1 = \{2,3\}$. Then,

$$\sum_{A_1} X_i|\theta \;=\; X_2 + X_3\,|\,\theta \;\sim\; \text{Mn}_1(10,\,\theta_2+\theta_3)\ ,$$

or

$$X_2 + X_3\,|\,\theta \;\sim\; \text{Mn}_1(10,\,0.45)\ .$$

Analogously,

$$X_0 + X_1\,|\,\theta \;\sim\; \text{Mn}_1(10,\,0.55)\ ,$$
$$X_1 + X_3\,|\,\theta \;\sim\; \text{Mn}_1(10,\,0.35)\ ,$$
$$X_2\,|\,\theta \;\sim\; \text{Mn}_1(10,\,0.30)\ .$$

Note that, in general, if $X|\theta \sim \text{Mn}_k(n,\theta)$, then $X_i\,|\,\theta \sim \text{Mn}_1(n,\theta_i)$, for $i = 1,...,k$.

Example 8.4. 3×3 Contingency Tables.
Assume that $X\,|\,\theta \sim \text{Mn}_8(n,\theta)$, as in a 3×3 Contingency Tables:

x_{11}	x_{12}	x_{13}	$x_{1\bullet}$
x_{21}	x_{22}	x_{23}	$x_{2\bullet}$
x_{31}	x_{32}	x_{33}	$x_{3\bullet}$
$x_{\bullet 1}$	$x_{\bullet 2}$	$x_{\bullet 3}$	n

Applying Theorem 3.2 we get

$$(X_{1\bullet}, X_{2\bullet}) \,|\, \theta \;\sim\; \mathrm{Mn}_2(n, \theta')\ , \qquad \theta' = (\theta_{1\bullet}, \theta_{2\bullet}), \quad \theta'_0 = \theta_3\ .$$

This result tell us that

$$(X_{i1}, X_{i2}, X_{i3}) \,|\, \theta \;\sim\; \mathrm{Mn}_3(n, \theta'_i)\ ,$$

with

$$\theta'_i = (\theta_{i1}, \theta_{i2}, \theta_{i3}), \quad \theta'_{0i} = 1 - \theta_{i\bullet}\ , \qquad i = 1, 2, 3\ .$$

We can now apply Theorem 3.3 to obtain the probability distribution of each row of the contingency table, conditioned on its sum, or conditioned on the sum of the other rows. We have

$$(X_{i1}, X_{i2}) \,|\, x_{i\bullet}\ , \qquad \theta \sim \mathrm{Mn}_2(x_{i\bullet}, \theta'_i)$$

with

$$\theta'_i = \frac{(\theta_{il}, \theta_{i2})}{\theta_{i\bullet}}\ , \qquad \theta'_{0i} = \frac{\theta_{i3}}{\theta_{i\bullet}}\ .$$

The next result expresses the distribution of $X \,|\, \theta$ in term of the conditional distributions, of each row of the table, in its sum, and in term of the distribution of these sums.

Proposition 8.1. *If* $X \,|\, \theta \sim \mathrm{Mn}_{r^2-1}(n, \theta)$, *as in an* $r \times r$, *contingency table, then* $P(X \,|\, \theta)$ *can be written as*

$$P(X \,|\, \theta) \;=\; \left[\prod_{i=1}^{r} P\big(X_{i1}, ..., X_{i,r-1} \,|\, x_{i\bullet}\,,\, \theta\big)\right] P\big(X_{1\bullet}, ..., X_{r-1\bullet} \,|\, \theta\big)\ .$$

Proof: We have:

$$\begin{aligned} P(X \,|\, \theta) \;&=\; n! \prod_{i=1}^{r} \frac{\theta_i^{x_i}}{x_i!} \;=\; n!\ \frac{\theta_{11}^{x_{11}} \cdots \theta_{rr}^{x_{rr}}}{x_{11}! \cdots x_{rr}!} \\ &= \left[\prod_{i=1}^{r} \frac{x_{i\bullet}!}{x_{i1}! \cdots x_{ir}!} \left(\frac{\theta_{i1}}{\theta_{i\bullet}}\right)^{x_{i1}} \!\cdots \left(\frac{\theta_{ir}}{\theta_{i\bullet}}\right)^{x_{ir}}\right] \frac{n!}{x_{i\bullet}! \cdots x_{r\bullet}!}\ \theta_{1\bullet}^{x_{1\bullet}} \cdots \theta_{r\bullet}^{x_{r\bullet}}\ . \end{aligned}$$

From Theorems 3.2 and 3.3, as in the last example, we recognize each of the first r factors above as the probabilities of each row in the table, conditioned on its sum, and recognize the last factor as the joint probability distribution of sum of these r rows. □

Corollary 8.1. *If* $X|\theta \sim \mathrm{Mn}_{r^2-1}(n,\theta)$*, as in Theorems 3.2 and 3.3, then*

$$P\big(X\,|\,x_{1\bullet},...,x_{r-1\bullet}\,,\,\theta\big) \;=\; \prod_{i=1}^{r} P\big(X_{i1},...,X_{i,r-1}\,|\,x_{i\bullet}\,,\,\theta\big)$$

and, knowing θ, $x_{1\bullet},...,x_{r-1\bullet}$,

$$(X_{11},...,X_{1,r-1})\ \mathrm{II}\ ...\ \mathrm{II}\ (X_{r1},...,X_{r,r-1})\ .$$

Proof: Since

$$P(X|\theta) \;=\; P\big(X\,|\,x_{1\bullet},...,x_{r-1\bullet}\,,\,\theta\big)\,P\big(X_{1\bullet},X_{2\bullet},...,X_{r-1\bullet}\,|\,\theta\big)\ ,$$

from Theorems 3.2 and 3.3 we get the proposed equality. □

The following result will be used next to express Theorem 3.4 as a canonical representation for $P(X|\theta)$.

Proposition 8.2. *If* $X|\theta \sim \mathrm{Mn}_{r^2-1}(n,\theta)$*, as in Proposition, then a transformation*

$$T:\ \big(\theta_{11},...,\theta_{1r},...,\theta_{r1},...,\theta_{r,r-1}\big) \;\to\; \big(\lambda_{11},...,\lambda_{1,r-1},...,\lambda_{r1},...,\lambda_{r,r-1},\eta_1,...,\eta_{r-1}\big)$$

given by

$$\begin{array}{lcl} \lambda_{11} = \dfrac{\theta_{11}}{\theta_{1\bullet}}\ , & \cdots\ , & \lambda_{1,r-1} = \dfrac{\theta_{1,r-1}}{\theta_{1\bullet}} \\ \vdots & & \\ \lambda_{r1} = \dfrac{\theta_{r1}}{\theta_{r\bullet}}\ , & \cdots\ , & \lambda_{r,r-1} = \dfrac{\theta_{r,r-1}}{\theta_{r\bullet}} \end{array}$$

$$\eta_1 = \theta_{1\bullet}\,, \quad \eta_2 = \theta_{2\bullet}\,, \quad \ldots, \quad \eta_{r-1} = \theta_{(r-1)\bullet}$$

is a onto transformation defined in $\big\{0<\theta_{11}+\cdots+\theta_{r,r-1}<1\,;\ \ 0<\theta_{ij}<1\big\}$ *over the unitary cube of dimension* r^2-1*. Moreover, the Jacobian of this transformation, t, is*

$$J \;=\; \eta^{r-1}\,\eta_1^{r-1}\cdots\,\eta_{r-1}^{r-1}\,\big(1-\eta_1-\cdots-\eta_{r-1}\big)^{r-1}\ .$$

The proof is not hard to check.

Example 8.5. Let us examine the case of a 2×2 contingency table:

$$\begin{array}{|c|c|}\hline x_{11} & x_{12} \\ \hline x_{21} & x_{22} \\ \hline \end{array}_{\;n} \qquad\qquad \begin{array}{|c|c|}\hline \theta_{11} & \theta_{12} \\ \hline \theta_{21} & \theta_{22} \\ \hline \end{array}_{\;1}$$

In order to obtain the canonical representation of $P(X|\theta)$ we use the transformation T in the case $r = 2$:

$$\lambda_{11} = \frac{\theta_{11}}{\theta_{11}+\theta_{12}} \ ,$$

$$\lambda_{21} = \frac{\theta_{11}}{\theta_{21}+\theta_{22}} \ ,$$

$$\eta_1 = \theta_{11}+\theta_{12} \ ,$$

hence,

$$P(X|\theta) = \binom{x_{1\bullet}}{x_{11}} \lambda_{11}^{x_{11}}(1-\lambda_{11})^{x_{12}} \binom{x_{2\bullet}}{x_{21}} \lambda_{21}^{x_{21}}(1-\lambda_{21})^{x_{22}} \binom{n}{x_{1\bullet}} \eta_1^{x_{1\bullet}}(1-\eta_1)^{x_{2\bullet}} \ ,$$

$$0<\theta_{11}<1, \quad 0<\theta_{21}<1, \quad 0<\eta_1<1 \ .$$

9. FUNCTIONAL CHARACTERIZATIONS

The objective of this section is to derive the general form of a homogeneous Markov random process. Theorem 9.1, by Reny and Aczel, states that such a process is described by a mixture of Poisson distributions. Our presentation follows Aczél (1966, Sec. 2.1 and 2.3) and Jánossy, Rényi and Aczél (1950). It follows from the characterization of the Multinomial by the Poisson distribution given in Theorem 3.1, that Reny–Aczel characterization of a homogeneous and local time point process is analogous to de Finetti characterization of an infinite exchangeable 0-1 process as a mixture of Bernoulli distributions, see for example Feller (V. 2, Ch. VII, Sec. 4).

Cauchy's Functional Equations

Cauchy's additive functional equation has the form

$$f(x+y) = f(x)+f(y) \ .$$

The following argument from Cauchy (1821) shows that a continuous solution of this functional equation must have the form

$$f(x) = cx \ .$$

Repeating the sum of the same argument, x, n times, we must have $f(nx) = nf(x)$. If $x = (m/n)t$, then $nx = mt$ and

$$nf(x) = f(nx) = f(mt) = mf(t) \ ,$$

hence

$$f\left(\frac{m}{n}\,t\right) \;=\; \frac{m}{n}\,f(t)\ ,$$

taking $c = f(1)$, and $x = m/n$, it follows that $f(x) = cx$, over the rationals, $x \in \mathbb{Q}$. From the continuity condition for $f(x)$, the last result must also be valid over the reals, $x \in \mathbb{R}$. Q.E.D.

Cauchy's multiplicative functional equation has the form

$$f(x+y) \;=\; f(x)\,f(y)\ , \qquad \forall\, x, y > 0\,,\quad f(x) \geq 0\ .$$

The trivial solution of this equation is $f(x) \equiv 0$. Assuming $f(x) > 0$, we take the logarithm, reducing the multiplicative equation to the additive equation,

$$\ln f(x_y) \;=\; \ln f(x) + \ln f(y)\ ,$$

hence

$$\ln f(x) = cx\,, \quad \text{or} \quad f(x) = \exp(cx)\ .$$

Homogeneous Discrete Markov Processes

We seek the general form of a homogeneous discrete Markov process. Let $w_k(t)$, for $t \geq 0$, be the probability of occurrence of exactly k events. Let us also assume the following hypotheses:

Time Locality: If $t_1 \leq t_2 \leq t_3 \leq t_4$ then, the number of events in $[t_1, t_2[$ is independent of the number of events in $[t_3, t_4[$.

Time Homogeneity: The distribution for the number of events occurring in $[t_1, t_2[$ depends only on the interval length, $t = t_2 - t_1$.

From time locality and homogeneity, we can decompose the occurrence of no (zero) events in $[0, t+u[$ as ,

$$w_0(t+u) \;=\; w_0(t)\,w_0(u)\ .$$

Hence, $w_0(t)$ must obey Cauchy's functional equation, and

$$w_0(t) \;=\; \exp(ct) \;=\; \exp(-\lambda t)\ .$$

Since $w_0(t)$ is a probability distribution, $w_0(t) \leq 1$, and $\lambda > 0$.

Hence, $v(t) = 1 - w_0(t) = 1 - \exp(-\lambda t)$, the probability of one or more events occurring before $t > 0$, must be the familiar exponential distribution.

For $k \geq 1$ occurrences before $t + u$, the general decomposition relation is

$$w_n(t+u) \;=\; \sum_{k=0}^{n} w_k(t)\,w_{n-k}(u)\ .$$

Theorem 9.1 (Reny–Aczel). *The general (non trivial) solution of this this system of functional equations has the form:*

$$w_k(t) \;=\; e^{-\lambda t} \sum_{\langle r,k\rangle} \prod_{j=1}^{k} \frac{(c_j\,t)^{r_j}}{r_j!} \;, \qquad \lambda = \sum_{j=1}^{\infty} c_j \;.$$

where the index set $\langle r, k, n\rangle$ *is defined as*

$$\langle r,k,n\rangle \;=\; \Big\{ r_1, r_2, ..., r_k \;\Big|\; r_1 + 2r_2 + \cdots + k\, r_k = n \Big\} \;.$$

and $\langle r, k\rangle$ *is a shorthand for* $\langle r, k, k\rangle$.

Proof: By induction: The theorem is true for $k = 0$. Let us assume, as induction hypothesis, that it is true to $k < n$. The last equation in the recursive system is

$$\begin{aligned} w_n(t+u) \;&=\; \sum_{k=0}^{n} w_k(t)\, w_{n-k}(u) \;= \\ &=\; w_n(t)\, e^{-\lambda u} + w_n(u)\, e^{-\lambda t} + e^{-\lambda(t+u)} \sum_{k=1}^{n-1} \sum_{\langle r,k\rangle} \sum_{\langle s,n-k\rangle} \prod_{i=1}^{k} \frac{(c_i\,t)^{r_i}}{r_i!} \prod_{j=1}^{k} \frac{(c_j\,u)^{s_j}}{s_j!} \;. \end{aligned}$$

Defining

$$f_n(t) \;=\; e^{\lambda t}\, w_n(t) - \sum_{\langle r,n-1,n\rangle} \prod_{j=1}^{n-1} \frac{(c_j\,t)^{r_j}}{r_j!} \;,$$

the recursive equation takes the form

$$f_n(t+u) \;=\; f_n(t) + f_n(u) \;,$$

and can be solved as a general Cauchy's equation, that is,

$$f_n(t) \;=\; c_n\, t \;.$$

From the last equation and the definition of $f_n(t)$, we get the expression of $w_n(t)$ as in Theorem 9.1. The constant λ is chosen so that the distribution is normalized. □

The general solution given by Theorem 9.1 represents a composition (mixture) of Poisson processes, where an event in the j-th process in the composition corresponds to the simultaneous occurrence of j single events in the original homogeneous Markov process. If we impose the following rarity condition, the general solution is reduced to a mixture of ordinary Poisson processes.

Rarity Condition: The probability that an event occurs in a short time at least once is approximately equal to the probability that it occurs exactly once, that is, the probability of simultaneous occurrences is zero.

10. FINAL REMARKS

This work is in memory of Professor D. Basu who was the supervisor of the first author PhD dissertation, the starting point for the research in Bayesian analysis of categorical data presented here. A long list of papers follows Basu and Pereira (1982). We have chosen a few that we recommend for additional reading: Albert (1985), Gunel (1984), Irony, Pereira and Tiwari (2000), Paulino and Pereira (1992, 1995) and Walker (1996). To make the analysis more realistic, extensions and mixtures of Dirichlet also were considered. For instance see Albert and Gupta (1983), Carlson (1977), Dickey (1983), Dickey, Jiang and Kadane (1987), and Jiang, Kadane and Dickey (1992).

Usually the more complex distributions are used to realistic represent situations for which the strong properties of Dirichlet seems to be not realistic. For instance, in a 2×2 contingency table, the first line to be conditional independent of the second line given the marginal seems to be unrealistic in some situations. Mixtures of Dirichlet in some cases take care of the situation as shown by Albert and Gupta (1983).

The properties presented here are also important in non-parametric Bayesian statistics in order to understand the Dirichlet process for the competitive risk survival problem. See for instance Salinas-Torres, Pereira and Tiwari (1997, 2002). In order to be historically correct we cannot forget the important book of Wilks, published in 1962, where one can find the definition of Dirichlet distribution.

This article adopts a singular notation and representation, first used in Pereira and Stern (2005). Singular representations are unusual in statistical texts. Nevertheless, the singular notation makes it simpler to extend and generalize theoretical results and greatly facilitates numerical and computational implementation.

We end this article presenting the Reny–Aczel characterization of the Poisson mixture. This result can be interpreted as an alternative to de Finetti characterization theorem introduced in Finetti (1937). Using the characterization of binomial distributions by Poisson processes conditional arguments, as given by Theorem 3.1, and Blackwell (minimal) sufficiency properties discussed in Basu and Pereira (1983), Section 9 leads in fact to a De Finetti characterization for Binomial distributions. Also, if one recall the indifference principle (Mendel, 1989) the finite version of Finetti argument can simply be obtained. See also Irony and Pereira (1994) for the motivation of these arguments. The consideration of Section 9 could be viewed as a very simple formulation of the binomial distribution finite characterization.

ACKNOWLEDGMENTS

The authors had support from IME-USP, FAPESP and CNPq.

REFERENCES

[1] Aczél, J. (1966). *Lectures on Functional Equations and their Applications*, New York, Academic Press.

[2] Albert, J.H. (1985). *Bayesian estimation methods for incomplete two-way contingency tables using prior belief of association.* In "Bayesian Statistics" (J.M. Bernardo, ... and A.F.M. Smith, Eds.), Amsterdam, North Holland, 2:589–602.

[3] Albert, J.H. and Gupta, A.K. (1983). Bayesian estimation methods for 2×2 contingency tables using mixtures of Dirichlet distributions, *JASA*, **78**, 831–841.

[4] Aitchison, J. (2003). *The Statistical Analysis for Compositional Data*, (2nd Ed.), Blackburn Press, Caldwell.

[5] Aitchison, J. and Shen, S.M. (1980). Logistic-normal distributions: some properties and uses, *Biometrika*, **67**, 261–272.

[6] Basu, D. and Pereira, C.A.B. (1982). On the Bayesian analysis of categorical data: the problem of nonresponse, *JSPI*, **6**, 345–362.

[7] Basu, D. and Pereira, C.A.B. (1983). A note on Blackwell sufficiency and a Shibinsky characterization of distributions, *Sankhya A*, **45**(1), 99–104.

[8] de Finetti, B. (1947). La prévision: des lois logiques, ses sourses subjectives, *Annalles de l'Institut Henri Poincaré*, **7**, 1–68. English translation: *Foresight: Its logical laws, its subjective sources.* In "Studies in Subjective Probability" (Kiburg and Smoker, Eds.) (1963), pp. 93–158, Wiley, New York.

[9] Dickey, J.M. (1983). Multiple hypergeometric functions: probabilistic interpretations and statistical uses, *JASA*, **78**, 628–637.

[10] Dickey, J.M.; Jiang, T.J. and Kadane, J.B. (1987). Bayesian methods for categorical data, *JASA*, **82**, 773–781.

[11] Feller, W. (1957). *An Introduction to Probability Theory and Its Applications* (2nd Ed.), Vol. 1, Wiley, New York.

[12] Feller, W. (1966). *An Introduction to Probability Theory and Its Applications* (2nd Ed.), Vol. 2, Wiley, New York.

[13] Gunel, E. (1984). A Bayesian analysis of the multinomial model for a dichotomous response with non-respondents, *Communications in Statistics – Theory and Methods*, **13**, 737–751.

[14] Irony, T.Z. and Pereira, C.A.B. (1994). Motivation for the use of discrete distributions in quality assurance, *Test*, **3**(2), 181–193.

[15] Irony, T.Z.; Pereira, C.A.B. and Tiwari, R.C. (2000). Analysis of opinion swing: comparison of two correlated proportions, *The American Statistician*, **54**, 57–62.

[16] Jánossy, L.; Rényi, A. and Aczél, J. (1950). On composed Poisson distributions, *Acta Math. Hungarica*, **1**, 209–224.

[17] Jiang, T.J.; Kadane, J.B. and Dickey, J.M. (1992). Computation of Carlson's multiple hypergeometric function R for Bayesian applications, *Journal of Computational and Graphical Statistics*, **1**, 231–251.

[18] Kadane, J.B. (1985). Is victimization chronic? A Bayesian analysis of multinomial missing data, *Journal of Econometrics*, **29**, 47–67.

[19] Little, R.J.A. and Rubin, D.B. (1987). *Statistical Analysis with Missing Data*, Wiley, New York.

[20] Martin, J.J. (1975). *Bayesian decision and problems and Markov Chains.*

[21] Mendel, M.B. (1989). *Development of Bayesian parametric theory with application in control*, PhD Thesis, MIT, Cambridge: MA.

[22] Paulino, C.D.M. and Pereira, C.A.B. (1992). Bayesian analysis of categorical data informatively censored, *Communications in Statistics – Theory and Methods*, **21**, 2689–2705.

[23] Paulino, C.D.M. and Pereira, C.A.B. (1995). Bayesian methods for categorical data under informative general censoring, *Biometrika*, **82**(2), 439–446.

[24] Pereira, C.A.B. and Stern, J.M. (2005). Inferência indutiva com dados discretos: uma visão genuinamente Bayesiana, *COMCA-2005*, Universidad de Antofagasta, Chile.

[25] Salinas-Torres, V.H.S.; Pereira, C.A.B. and Tiwari, R.C. (1997). Convergence of Dirichlet measures arising in context of Bayesian analysis of competing risks models, *J. Multivariate Analysis*, **62**(1), 24–35.

[26] Salinas-Torres, V.H.S.; Pereira, C.A.B. and Tiwari, R.C. (2002). Bayesian nonparametric estimation in a series system or a competing-risks model, *J. of Nonparametric Statistics*, **14**(4), 449–458.

[27] Smith, P.J. and Gunel, E. (1984). Practical Bayesian approaches to the analysis of 2×2 contingency table with incompletely categorized Data, *Communication of Statistics – Theory and Methods*, **13**, 1941–63.

[28] Stern, J.M. (2007). *Cognitive constructivism and the epistemic significance of sharp statistical hypotheses*, 2007 Summer Program, Institute of Mathematics and Statistics, University of São Paulo, Brazil.

[29] Tian, G.L.; Ng, K.W. and Geng, Z. (2003). Bayesian computation for contingency tables with incomplete cells-counts, *Statistica Sinica*, **13**, 189-206.

[30] Walker, S. (1986). A Bayesian maximum posteriori algorithm for categorical data under informative general censoring, *The Statistician*, **45**, 293–298.

[31] Wilks, S.S. (1962). *Mathematical Statistics*, Wiley, New York.