**Academic Advances of the CTO, Vol. 1, Issue 1, 2017**
**Original Research Article**

# A Generalized Multinomial Distribution from Dependent Categorical Random Variables

Rachel Traylor*, Ph.D.

**Abstract**
Categorical random variables are a common staple in machine learning methods and other applications across disciplines. Many times, correlation within categorical predictors exists, and has been noted to have an effect on various algorithm effectiveness, such as feature ranking and random forests. We present a mathematical construction of a sequence of identically distributed but dependent categorical random variables, and give a generalized multinomial distribution to model the probability of counts of such variables.

**Keywords**
categorical variables — correlation — multinomial distribution — probability theory

## Contents



## Introduction

Bernoulli random variables are invaluable in statistical analysis of phenomena having binary outcomes, however, many other variables cannot be modeled by only two categories. Many topics in statistics and

*Office of the CTO, Dell EMC

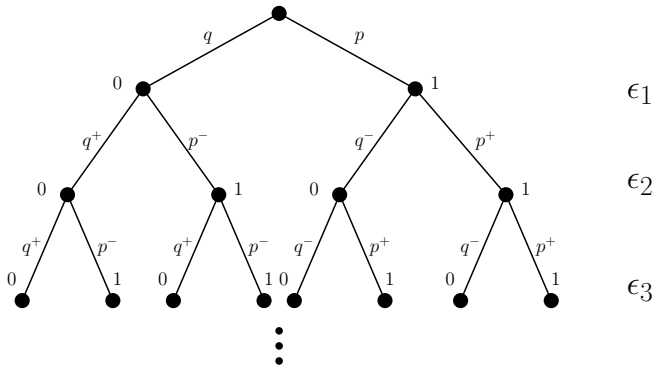


**Figure 1.** First-Dependence Tree Structure

machine learning rely on categorical random variables, such as random forests and various clustering algorithms. [6, 7]. Many datasets exhibit correlation or dependency among predictors as well as within predictors, which can impact the model used. [6, 9]. This can result in unreliable feature ranking [9], and inaccurate random forests [6].

Some attempts to remedy these effects involve Bayesian modeling [2] and various computational and simulation methods [8]. In particular, simulation of correlated categorical variables has been discussed in the literature for some time. [1, 3, 5]. Little research has been done to create mathematical framework of correlated or dependent categorical variables and the resulting distributions of functions of such variables.

Korzeniowski [4] studied dependent Bernoulli variables, formalizing the notion of identically distributed but dependent Bernoulli variables and deriving the distribution of the sum of such dependent variables, yielding a Generalized Binomial Distribution.

In this paper, we generalize the work of Korzeniowski [4] and formalize the notion of a sequence of identically distributed but dependent categorical random variables. We then derive a Generalized Multinomial Distribution for such variables and provide some properties of said distribution. We also give an algorithm to generate a sequence of correlated categorical random variables.

## 1. Background

Korzeniowski defined the notion of dependence in a way we will refer to here as *dependence of the first kind* (FK dependence). Suppose $(\varepsilon_1,...,\varepsilon_N)$ is a sequence of Bernoulli random variables, and $P(\varepsilon_1 = 1) = p$. Then, for $\varepsilon_i, i \geq 2$, we weight the probability of each binary outcome toward the outcome of $\varepsilon_1$, adjusting the probabilities of the remaining outcomes accordingly.

Formally, let $0 \leq \delta \leq 1$, and $q = 1 - p$. Then define the following quantities

$$\begin{aligned} p^+ &:= P(\varepsilon_i = 1|\varepsilon_1 = 1) = p + \delta q \qquad & p^- &:= P(\varepsilon_i = 0|\varepsilon_1 = 1) = q - \delta q \\ q^+ &:= P(\varepsilon_i = 1|\varepsilon_1 = 0) = p - \delta p \qquad & q^- &:= P(\varepsilon_i = 0|\varepsilon_1 = 0) = q + \delta p \end{aligned} \tag{1}$$

Given the outcome $i$ of $\varepsilon_1$, the probability of outcome $i$ occurring in the subsequent Bernoulli variables $\varepsilon_2, \varepsilon_3, ...\varepsilon_n$ is $p^+, i = 1$ or $q^+, i = 0$. The probability of the opposite outcome is then decreased to $q^-$ and $p^-$, respectively.

Figure 1 illustrates the possible outcomes of a sequence of such dependent Bernoulli variables. Korzeniowski showed that, despite this conditional dependency, $P(\varepsilon_i = 1) = p \; \forall \; i$. That is, the sequence of Bernoulli variables is identically distributed, with correlation shown to be

$$Cor(\varepsilon_i, \varepsilon_j) = \begin{cases} \delta, & i = 1 \\ \delta^2, & i \neq j, i, j \geq 2 \end{cases}$$

These identically distributed but correlated Bernoulli random variables yield a Generalized Binomial distribution with a similar form to the standard binomial distribution. In our generalization, we use the same form of FK dependence, but for categorical random variables. We will construct a sequence of identically distributed but dependent categorical variables from which we will build a generalized multinomial distribution. When the number of categories $K = 2$, the distribution reverts back to the generalized binomial distribution of Korzeniowski [4]. When the sequence is fully independent, the distribution reverts back to the independent categorical model and the standard multinomial distribution, and when the sequence is independent and $K = 2$, we recover the standard binomial distribution. Thus, this new distribution represents a much larger generalization than prior models.

## 2. Construction of Dependent Categorical Variables

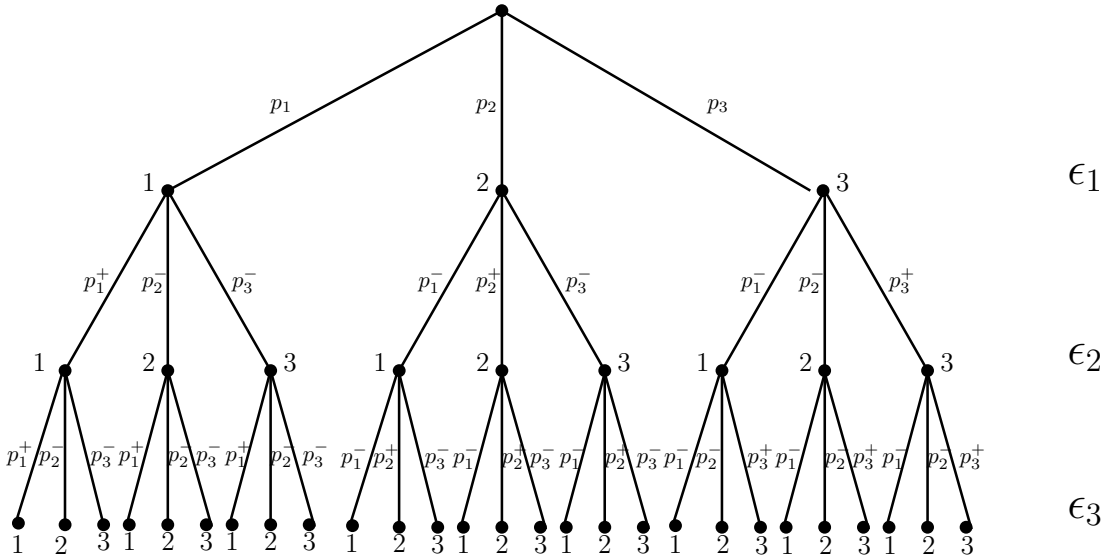


**Figure 2.** Probability distribution at $N = 3$ for $K = 3$

Suppose each categorical variable has $K$ possible categories, so the sample space $S = \{1,...,K\}$[1] The construction of the correlated categorical variables is based on a probability mass distribution over $K-$adic partitions of $[0,1]$. We will follow graph terminology in our construction, as this lends a visual representation of the construction. We begin with a parent node and build a $K$-nary tree, where the end nodes are labeled by the repeating sequence $(1,...,K)$. Thus, after $N$ steps, the graph has $K^N$ nodes, with each node labeled $1,...,K$ repeating in the natural order and assigned injectively to the intervals $\left(0,\frac{1}{K^N}\right],\left(\frac{1}{K^n},\frac{2}{K^N}\right],...,\left(\frac{K^N-1}{K^n},1\right]$. Define

$$\begin{aligned} \varepsilon_N &= i \qquad \text{on } \left(\tfrac{Kj}{K^N},\tfrac{Kj+i}{K^N}\right] \\ \varepsilon_N &= K \qquad \text{on } \left(\tfrac{K(j+1)-1}{K^N},\tfrac{K(j+1)}{K^N}\right] \end{aligned} \tag{2}$$

where $i = 1,...,K-1$, and $j = 0,1,...,K^{N-1}-1$. An alternate expression for (2) is

$$\begin{aligned} &\varepsilon_N = i \text{ on } \left(\tfrac{l-1}{K^N},\tfrac{l}{K^N}\right], \quad i \equiv l \mod K, \quad i = 1,...,K-1 \\ &\varepsilon_N = K \text{ on } \left(\tfrac{l-1}{K^N},\tfrac{l}{K^N}\right], \quad 0 \equiv l \mod K \end{aligned} \tag{3}$$

[1]These integers should not be taken as ordered or sequential, but rather as character titles of categories.

To each of the branches of the tree, and by transitivity to each of the $K^N$ partitions of $[0,1]$, we assign a probability mass to each node such that the total mass is 1 at each level of the tree in a similar manner to [4].

Let $0 < p_i < 1$, $i = 1,...,K$ such that $\sum_{i=1}^{K} p_i = 1$, and let $0 \leq \delta \leq 1$ be the *dependency coefficient*. Define $p_i^+ := p_i + \delta \sum_{l \neq i} p_l$, $p_i^- = p_i - \delta p_i$
for $i = 1,...,K$. These probabilities satisfy two important criteria:

**Lemma 1.**

- $\sum_{i=1}^{K} p_i = p_i^+ + \sum_{l \neq i} p_l^- = 1$
- $p_i p_i^+ + p_i^- \sum_{l \neq i} p_l = p_i$

*Proof.* The first is obvious from the definitions of $p_i^{+/-}$ above. The second statement follows clearly from algebraic manipulation of the definitions: $p_i p_i^+ + p_i^- \sum_{l \neq i} p_l = p_i^2 + p_i(1 - p_i) = p_i$ □

We now give the construction in steps down each level of the $K$-nary tree.

LEVEL 1:
*Parent node* has mass 1, with mass split $1 \cdot \prod_{i=1}^{K} p_i^{[\varepsilon_1 = i]}$, where $[\cdot]$ is an Iverson bracket. This level corresponds to a sequence $\varepsilon$ of dependent categorical variables of length 1.

| $\varepsilon_1$/Branch | Path | Mass at Node | Interval |
|---|---|---|---|
| 1 | *parent* $\rightarrow$ 1 | $p_1$ | $(0, 1/K]$ |
| 2 | *parent* $\rightarrow$ 2 | $p_2$ | $(1/K, 2/K]$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| i | *parent* $\rightarrow$ i | $p_i$ | $((i-1)/K, i/K]$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| K | *parent* $\rightarrow$ K | Mass $p_K$ | $((K-1)/K, 1]$ |

**Table 1.** Probability mass distribution at Level 1

LEVEL 2:
Level 2 has $K$ nodes, with $K$ branches stemming from each node. This corresponds to a sequence of length 2: $\varepsilon = (\varepsilon_1, \varepsilon_2)$. Denote $i.1$ as node $i$ from level 1. For $i = 1,...,K$,

**Node i.1** has mass $p_i$, with mass split $p_i \left(p_i^+\right)^{[\varepsilon_2 = i]} \prod_{j=1, j \neq i}^{K} \left(p_j^-\right)^{[\varepsilon_2 = j]}$

| $\varepsilon_2$/Branch | Path | Mass at Node | Interval |
|---|---|---|---|
| 1 | $i.1 \rightarrow 1$ | $p_i p_1^-$ | $\left(\frac{(i-1)K}{K^2}, \frac{(i-1)K+1}{K^2}\right]$ |
| 2 | $i.1 \rightarrow 2$ | $p_i p_2^-$ | $\left(\frac{(i-1)K+1}{K^2}, \frac{(i-1)K+2}{K^2}\right]$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| i | $i.1 \rightarrow i$ | $p_i p_i^+$ | $\left(\frac{(i-1)K+(i-1)}{K^2}, \frac{(i-1)K+i}{K^2}\right]$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| K | $i.1 \rightarrow K$ | $p_i p_K^-$ | $\left(\frac{iK-1}{K^2}, \frac{iK}{K^2}\right]$ |

**Table 2.** Probability mass distribution at Level II, Node i

In this fashion, we distribute the total mass 1 across level 2. In general, at any level $r$, there are $K$ streams of probability flow at Level $r$. For $\varepsilon_1 = i$, the probability flow is given by

$$p_i \prod_{j=2}^{r} \left[ (p_i^+)^{[\varepsilon_j = i]} \prod_{l \neq i} (p_l^-)^{[\varepsilon_j = l]} \right], i = 1, ..., K \tag{4}$$

We use this flow to assign mass to the $K^r$ intervals of $[0,1]$ at level $r$ in the same way as above. We may also verify via algebraic manipulation that

$$p_i = p_i \left( p_i^+ + \sum_{l \neq i} p_i^- \right)^{r-1} = p_i \sum_{\varepsilon_2, ..., \varepsilon_r} \prod_{j=2}^{r} \left[ (p_i^+)^{[\varepsilon_j = i]} \prod_{l \neq i} (p_l^-)^{[\varepsilon_j = l]} \right] \tag{5}$$

where the first equality is due to the first statement of Lemma 1, and the second is due to (4).

## 2.1 Example construction

For an illustration, refer to Figure 2. In this example, we construct a sequence of length $N = 3$ of categorical variables with $K = 3$ categories. At each level $r$, there are $3^r$ nodes corresponding to $3^r$ partitions of the interval $[0,1]$. Note that each time the node splits into 3 children, the sum of the split probabilities is 1. Despite the outcome of the previous random variable, the next one always has three possibilities. The sample space of categorical variable sequences of length 3 has $3^3 = 27$ possibilities. Some example sequences are $(1,3,2)$ with probability $p_1 p_3^- p_2^-$, $(2,1,2)$ with probability $p_2 p_1^- p_2^+$, and $(3,1,1)$ with probability $p_3 p_1^- p_1^-$. These probabilities can be determined by tracing down the tree in Figure 2.

## 2.2 Properties

### 2.2.1 Identically Distributed but Dependent

We now show the most important property of this class of sequences– that they remain identically distributed despite losing independence.

**Lemma 2.** $P(\varepsilon_r = i) = p_i; i = 1, ..., K, r \in \mathbb{N}$.

*Proof.* The proof proceeds via induction. $r = 1$ is clear by definition, so we illustrate an additional base case. For $r = 2$, from Lemma 1, and for $i = 1, ..., K$,

$$P(\varepsilon_2 = i) = P\left[ \bigcup_{j=1}^{K^2 - 1} \left( \frac{Kj}{K^2}, \frac{Kj+1}{K^2} \right] \right] = p_i p_i^+ + p_i^- \sum_{j \neq i} p_j = p_i$$

Then at level $r$, in keeping with the alternative expression (3), we express $\varepsilon_r$ in terms of the specific nodes at level $r$:

$$\varepsilon_r = \sum_{l=1}^{K^r} \varepsilon_r^l \mathbb{1}_{\left( \frac{l-1}{K^r}, \frac{l}{K^r} \right]}, \text{ where } \varepsilon_r^l = \begin{cases} l \mod K, & 0 \not\equiv l \mod K \\ K, & 0 \equiv l \mod K \end{cases} \tag{6}$$

Let $m_r^l$ be the probability mass for $\varepsilon_r^l$. Then, for $l = 1, ..., K^r$ and $i = 1, ..., K-1$,

$$m_r^l = \begin{cases} P(\varepsilon_r^l = i), & i \equiv l \mod K \\ P(\varepsilon_r = K), & 0 \equiv l \mod K \end{cases} \tag{7}$$

As the inductive hypothesis, assume that for $i = 1,...,K-1$,

$$p_i = P(\varepsilon_{r-1} = i) = P\left(\bigcup_{\substack{l=1 \\ i \equiv l \mod K}}^{K^{r-1}} \{\varepsilon_{r-1}^l = i\}\right) = \sum_{\substack{l=1 \\ i \equiv l \mod K}}^{K^{r-1}} P\left(\varepsilon_{r-1}^l = i\right) = \sum_{\substack{l=1 \\ i \equiv l \mod K}}^{K^{r-1}} m_{r-1}^l$$

Each mass $m_{r-1}^l$ is split into $K$ pieces in the following way

$$m_{r-1}^l = \begin{cases} m_{r-1}^l \left(p_1^+ + \sum_{j=2}^K p_i^-\right), & l = 1,...,K \\ m_{r-1}^l \left(p_2^+ + \sum_{j\neq 2} p_i^-\right), & l = K+1,...,2K \\ m_{r-1}^l \left(p_3^+ + \sum_{j\neq 3} p_i^-\right), & l = 2K+1,...,3K \\ \vdots & \\ m_{r-1}^l \left(p_K^+ + \sum_{j\neq K} p_i^-\right), & l = K^{r-1}-K+1,...,K^{r-1} \end{cases} \tag{8}$$

which may be written as

$$m_{r-1}^l = \begin{cases} P(\varepsilon_r = 1) + P(\varepsilon_r^l \neq 1) = m_r^l + P(\varepsilon_r^l \neq 1), & l = 1,...,K \\ P(\varepsilon_r = 1) + P(\varepsilon_r^l \neq 1) = m_r^l + P(\varepsilon_r^l \neq 1), & l = K+1,...,2K \\ P(\varepsilon_r = 1) + P(\varepsilon_r^l \neq 1) = m_r^l + P(\varepsilon_r^l \neq 1), & l = 2K+1,...,3K \\ \vdots & \\ P(\varepsilon_r = 1) + P(\varepsilon_r^l \neq 1) = m_r^l + P(\varepsilon_r^l \neq 1), & l = K^{r-1}-K+1,...,K^{r-1} \end{cases} \tag{9}$$

Then

$$P(\varepsilon_r = 1) = \sum_{\substack{\xi=1 \\ 1 \equiv \xi \mod K}}^{K^r} m_r^\xi = \sum_{l=1}^K m_{r-1}^l p_1^+ + \sum_{l=K+1}^{K^{r-1}} m_{r-1}^l p_1^- \tag{10}$$

When $1 \equiv l \mod K$,

$$p_1^+ \left(\sum_{\xi=0}^{K-1} m_{r-1}^{l+\xi}\right) = m_{r-1}^l p_1^+ + m_{r-1}^l \left(\sum_{j=2}^K p_j^-\right) = m_{r-1}^l \tag{11}$$

Equation 11 holds because $m_{r-1}^{l+\xi} = \frac{p_{l+\xi}^-}{p_1^+}$, $\xi = 0,...,K-1$, and by of Lemma 1. Thus,

$$P(\varepsilon_r = 1) = \sum_{\substack{\xi=1 \\ 1 \equiv \xi \mod K}}^{K^r} m_r^\xi = \sum_{l=1}^K m_{r-1}^l p_1^+ + \sum_{l=K+1}^{K^{r-1}} m_{r-1}^l p_1^- = \sum_{\substack{l=1 \\ 1 \equiv l \mod K}}^{K} m_{r-1}^l + \sum_{l=K+1}^{K^{r-1}} m_{r-1}^l = \sum_{l=1}^{K^{r-1}} m_{r-1}^l = p_1 \tag{12}$$

A similar procedure for $i = 2,...,K$ follows and the proof is complete. □

#### 2.2.2 Pairwise Cross-Covariance Matrix

We now give the pairwise cross-covariance matrix for dependent categorical random variables.

**Theorem 1** (Cross-Covariance of Dependent Categorical Random Variables)**.** *Denote $\Lambda^{\iota,\tau}$ be the $K \times K$ cross-covariance matrix of $\varepsilon_\iota$ and $\varepsilon_\tau$, $\iota, \tau = 1,...,n$, defined as $\Lambda^{\iota,\tau} = E[(\varepsilon_\iota - E[\varepsilon_\iota])(\varepsilon_\tau - E[\varepsilon_\tau])]$. Then the entries of the matrix are given by $\Lambda^{1,\tau}_{ij} = \begin{cases} \delta p_i(1-p_i), & i = j \\ -\delta p_i p_j, & i \neq j \end{cases}$, $\tau \geq 2$, and $\Lambda^{\iota,\tau}_{ij} = \begin{cases} \delta^2 p_i(1-p_i), & i = j \\ -\delta^2 p_i p_j, & i \neq j \end{cases}$, $\tau > \iota$, $\iota \neq 1$.*

*Proof.* The $ij$th entry of $\Lambda^{\iota,\tau}$ is given by

$$\mathrm{Cov}([\varepsilon_\iota = i],[\varepsilon_\tau = j]) = \mathbb{E}[[\varepsilon_\iota = i][\varepsilon_\tau = j]] - \mathbb{E}[[\varepsilon_\iota = i]]\mathbb{E}[[\varepsilon_\tau = j]] = P(\varepsilon_\iota = i, \varepsilon_\tau = j) - P(\varepsilon_\iota = i)P(\varepsilon_\tau = j) \quad (13)$$

Let $\iota = 1$. For $j = i$, and $\tau = 2$,

$$\mathrm{Cov}([\varepsilon_1 = i],[\varepsilon_2 = i]) = P(\varepsilon_1 = i, \varepsilon_2 = i) - p_i^2 = p_i p_i^+ - p_i^2 = \delta p_i(1-p_i) \quad (14)$$

For $j \neq i$ and $\tau = 2$,

$$\mathrm{Cov}([\varepsilon_1 = i],[\varepsilon_2 = j]) = P(\varepsilon_1 = i, \varepsilon_2 = j) - p_i p_j = p_i p_j^- - p_i p_j = -\delta p_i p_j \quad (15)$$

For $\tau \neq 2$, it suffices to show that $P(\varepsilon_1 = i, \varepsilon_\tau = j) = \begin{cases} p_i p_i^+, & j = i \\ p_i p_j^-, & j \neq i \end{cases}$.

Starting from $\varepsilon_1$, the tree is split into $K$ large branches governed by the results of $\varepsilon_1$. Then at level $r$, there are $K^r$ nodes indexed by $l$. Each of the K large branches contains $K^{r-1}$ of these nodes. That is,

$$\begin{aligned} &\varepsilon_1 = 1 \text{ branch contains nodes } l = 1,...,K^{r-1} \\ &\varepsilon_1 = 2 \text{ branch contains nodes } l = K^{r-1}+1,...,2K^{r-1} \\ &\qquad\qquad\qquad\qquad \vdots \\ &\varepsilon_1 = K \text{ branch contains nodes } l = K^r - K + 1,...,K^r \end{aligned}$$

Then $P(\varepsilon_1 = 1, \varepsilon_r^l = i) = m_r^l$; $\quad i \equiv l \mod K, l = 1,...,K^{r-1}$. We have already shown the base case for $r = 2$. So, as an inductive hypothesis, assume

$$P(\varepsilon_1 = 1, \varepsilon_{r-1} = i) = \begin{cases} p_1 p_1^+, & i = 1 \\ p_1 p_i^-, & i \neq 1 \end{cases}$$

Then we have that for $i = 1$,

$$p_1 p_1^+ = P(\varepsilon_1 = 1, \varepsilon_{r-1} = 1) = \sum_{\substack{l=1 \\ 1 \equiv l \bmod K}}^{K^{r-2}} m_{r-1}^l$$

and for $i \neq 1$,

$$p_1 p_i^- = P(\varepsilon_1 = 1, \varepsilon_{r-1} = i) = \sum_{\substack{l=1 \\ i \equiv l \bmod K}}^{K^{r-2}} m_{r-1}^l$$

Moving one step down the tree (still noting that $\varepsilon_1 = 1$), we have seen that the each mass $m_{r-1}^l$ splits as

$$m_{r-1}^l = p_1^+ m_{r-1}^l + m_{r-1}^l \sum_{j=2}^{K} p_j^- = P(\varepsilon_1 = 1, \varepsilon_r = 1) + m_{r-1}^l \sum_{j=2}^{K} p_j^-$$

Therefore,

$$\begin{aligned}
P(\varepsilon_1 = 1, \varepsilon_r = 1) &= \sum_{\substack{l=1 \\ 1 \equiv l \mod K}}^{K^{r-1}} m_r^l \\
&= \sum_{\substack{l=1 \\ 1 \equiv l \mod K}}^{K^{r-2}} m_{r-1}^l p_1^+ + \sum_{j=2}^{K} \sum_{\substack{l=1 \\ 1 \not\equiv l \mod K}}^{K^{r-2}} m_{r-1}^l p_j^- \\
&= p_1 p_1^+ p_1^+ + p_1 p_1^+ \sum_{j=2}^{K} p_j^- \\
&= p_1 p_1^+
\end{aligned}$$

A similar strategy shows that $P(\varepsilon_1 = 1, \varepsilon_r = i) = p_1 p_i^-$ and that $P(\varepsilon_1 = i, \varepsilon_r = j) = \begin{cases} p_i p_i^+, & j = i \\ p_i p_j^-, & j \neq i \end{cases}$.

Next, we will compute the $ij$th entry of $\Lambda^{\iota,\tau}$, $\iota > 1, \tau > \iota$. For $\iota = 2, \tau = 3$,

$$P(\varepsilon_2 = i, \varepsilon_3 = j) = \begin{cases} p_i \left(p_i^+\right)^2 + (p_i^-)^2 \sum_{j \neq i} p_j, & i = j \\ p_i p_i^+ p_j^- + p_j p_i^- p_j^+ + \sum_{\substack{l \neq i \\ l \neq j}} p_l p_i^- p_j^-, & i \neq j \end{cases} \tag{16}$$

This can be seen by tracing the tree construction from level 2 to level 3, with an example given in Figure 2. Next, we will show that (16) holds for $\tau > 3$.

First, note that $P(\varepsilon_2 = 1, \varepsilon_\tau^l = 1) = m_\tau^l$ for $1 \equiv l \mod K$, and $l = (\xi - 1)L^{\tau-1} + j$, where $\xi = 1, ..., K$, $j = 1, ..., K^{\tau-2}$. We have shown the base case where $\tau = 3$. For the inductive hypothesis, assume that

$$P(\varepsilon_2 = 1, \varepsilon_{\tau-1} = 1) = p_1 \left(p_1^+\right)^2 + (p_1^-)^2 \sum_{j=2}^{K} p_j^-$$

Then we have that $\left(p_1^-\right)^2 \sum_{j=2}^{K} p_j^- = \sum_{1 \equiv l \mod K} m_{\tau-1}^l$ for the $l$ defined above. Then, at each $m_{\tau-1}^l$, we have the following mass splits:

$$\begin{aligned}
m_{\tau-1}^l &= m_{\tau-1}^l p_1^+ + m_{\tau-1}^l \sum_{j=2}^{K} p_j^-, && \xi = 1 \\
m_{\tau-1}^l &= m_{\tau-1}^l p_i^+ + m_{\tau-1}^l \sum_{j \neq i} m_{\tau-1}^l p_j^-, && \xi = i, \quad i = 2, ..., K
\end{aligned} \tag{17}$$

Then

$$P(\varepsilon_2 = 1, \varepsilon_\tau = 1) = \sum_{\xi=1}^{K} \sum_{\substack{l,\xi \\ 1 \equiv l \mod K}} m_\tau^l \tag{18}$$

Using the same tactic as (11), we see that using (17), adding the components, and combining the terms correctly, the proof is complete for any $\tau$. The proof for $P(\varepsilon_2 = i, \varepsilon_\tau = j)$ for any $i, j$ follows similarly, and the proof for $P(\varepsilon_\iota = i, \varepsilon_\tau = j)$ follows from reducing to the above proven claims.

□

In the next section, we exploit the desirable identical distribution of the categorical sequence in order to provide a generalized multinomial distribution for the counts in each category.

## 3. Generalized Multinomial Distribution

From the construction in Section 2, we derive a generalized multinomial distribution in which all categorical variables are identically distributed but no longer independent.

**Theorem 2** (Generalized Multinomial Distribution). *Let $\varepsilon_1,...,\varepsilon_N$ be categorical random variables with categories $1,...,K$, constructed as in Section 2. Let $X_i = \sum_{j=1}^{N}[\varepsilon_j = i], i = 1,...,K$, where $[\cdot]$ is the Iverson bracket. Denote $\mathbf{X} = (X_1,...,X_K)$, with observed values $\mathbf{x} = (x_1,...,x_K)$. Then*

$$P(\mathbf{X}=\mathbf{x}) = \sum_{i=1}^{K} p_i \frac{(N-1)!}{(x_i-1)!\prod_{j\neq i} x_j!} \left(p_i^+\right)^{x_i-1} \prod_{j\neq i} \left(p_j^-\right)^{x_j}$$

*Proof.* For brevity, let $\varepsilon_{(-1)} = (\varepsilon_2,...,\varepsilon_n)$ denote the sequence of $n$ categorical random variables with the first variable removed. Conditioning on $\varepsilon_1$,

$$\begin{aligned} P(\mathbf{X}=\mathbf{x}) &= \sum_{i=1}^{K} P(\mathbf{X}=\mathbf{x}|\varepsilon_1 = i) \\ &= \sum_{i=1}^{K} p_i \sum_{\substack{\varepsilon_{-(1)} \\ \mathbf{X}=\mathbf{x}}} \prod_{j=1}^{N} \left(p_i^+\right)^{[\varepsilon_j=i]} \prod_{\substack{k=1 \\ k\neq i}}^{K} \prod_{l=1}^{N} \left(p_l^-\right)^{[\varepsilon_l=k]} \\ &= \sum_{i=1}^{K} p_i \left( \sum_{\substack{\varepsilon_2,...,\varepsilon_N \\ \mathbf{X}=\mathbf{x}}} \left(p_i^+\right)^{\sum_{j=1}^{N}[\varepsilon_j=i]} \prod_{l\neq i} \left(p_l^-\right)^{\sum_{j=1}^{N}[\varepsilon_j=l]} \right) \end{aligned}$$

Now, when $\varepsilon_1 = 1$, there are $\binom{N-1}{x_1-1}$ combinations of the remaining $N-1$ categorical variables $\{\varepsilon_i\}_{i=2}^{N}$ to reside in category 1, $\binom{N-1-x_1-1}{x_2}$ ways the remaining $N-1-x_1-1$ categorical variables can reside in category 2, and so forth. Finally, there are $\binom{N-1-x_1-1-\sum_{i=2}^{K-1} x_i}{x_K}$ ways the final unallocated $\{\varepsilon_i\}$ can be in category $K$. Thus,

$$\begin{aligned} &P(\mathbf{X}=\mathbf{x}|\varepsilon_1 = 1) \\ &\quad = p_1 \binom{N-1}{x_1-1} \binom{N-1-x_1-1}{x_2} \cdots \binom{N-1-x_1-1-\sum_{j=2}^{K-1} x_j}{x_K} \left(p_1^+\right)^{x_1-1} \prod_{j=2}^{K} \left(p_j^-\right)^{x_j} \\ &\quad = p_1 \frac{(N-1)!}{(x_1-1)!\prod_{j=2}^{K} x_j!} \left(p_1^+\right)^{x_1-1} \prod_{j=2}^{K} \left(p_j^-\right)^{x_j} \end{aligned} \tag{19}$$

Similarly, for $i = 2,...,K$

$$\begin{aligned} &P(\mathbf{X}=\mathbf{x}|\varepsilon_1 = i) \\ &\quad = p_i \binom{N-1}{x_1} \cdots \binom{N-1-\sum_{j=1}^{i-1}}{x_i} \cdots \binom{N-1-x_i-1-\sum_{j\neq i, j=1}^{K-1} x_j}{x_K} \left(p_1^+\right)^{x_1-1} \prod_{j=2}^{K} \left(p_j^-\right)^{x_j} \\ &\quad = p_i \frac{(N-1)!}{(x_i-1)!\prod_{j\neq i} x_j!} \left(p_i^+\right)^{x_i-1} \prod_{j\neq i} \left(p_j^-\right)^{x_j} \end{aligned} \tag{20}$$

Summing completes the proof. □

## 4. Properties

This section details some useful properties of the Generalized Multinomial Distribution and the dependent categorical random variables.

### 4.1 Marginal Distributions

**Theorem 3** (Univariate Marginal Distribution)**.** *The univariate marginal distribution of the Generalized Multinomial Distribution is the Generalized Binomial Distribution. That is,*

$$P(X_i = x_i) = q\binom{N-1}{x_i}\left(p_i^-\right)^{x_i}\left(q^-\right)^{N-1-x_i} + p_i\binom{N-1}{x_i-1}\left(p_i^+\right)^{x_i-1}\left(q^+\right)^{N-1-(x_i-1)} \tag{21}$$

*where* $q = \sum_{j\neq i} p_j$, $q^+ = q + \delta p_i$, *and* $q^- = q - \delta q$

*Proof.* First, we claim the following: $q^+ = q + \delta p_i = p_l^+ + \sum_{\substack{j\neq l\\ j\neq i}} p_j^-$, $l = 2,...,K$. This may be justified via a simple manipulation of definitions:

$$p_l^+ + \sum_{j\neq l} p_j^- = p_l + \delta\sum_{j\neq l}(p_j - \delta p_j) + \sum_{\substack{j\neq l\\ j\neq i}}(p_j - \delta p_j) = p_l + \sum_{\substack{j\neq l\\ j\neq i}} p_j + \delta p_i = q + \delta p_i$$

Similarly, $q^- = q - \delta q$. Thus, we may collapse the number of categories to 2: Category $i$, and everything else. Now, notice that for $l \neq i$, $\left(p_l^+\right)^{x_l-1}\prod_{\substack{j\neq l\\ j\leq i}}\left(p_j^-\right)^{x_j} = \left(q^+\right)^{N-x_i-1}$ for $l = 1,...,K$ and $l \neq i$. Fix $k \neq i$. Then

$$\begin{aligned} p_k\frac{(N-1)!}{(x_k-1)!\prod_{j\neq k}x_j!}\left(p_k\right)^{x_k-1}\prod_{j\neq k}\left(p_j^-\right)^{x_j} &= p_k\left(p_i^-\right)^{x_i}\left(q^+\right)^{N-x_i-1}\frac{(N-1)!}{x_i!(x_k-1)!\prod_{\substack{j\neq k\\ j\neq i}}x_j!} \\ &= p_k\binom{N-1}{x_i}\left(p_i^-\right)^{x_i}\left(q^+\right)^{N-x_i-1} \end{aligned} \tag{22}$$

Then

$$\begin{aligned} \sum_{i=1}^{K} p_i\frac{(N-1)!}{(x_i-1)!\prod_{j\neq i}x_j!}\left(p_i^+\right)^{x_i-1}\prod_{j\neq i}\left(p_j^-\right)^{x_j} &= p_i\frac{(N-1)!}{(x_i-1)!\prod_{j\neq i}x_j!}\left(p_i^+\right)^{x_i-1}\left(q^-\right)^{N-1-(x_i-1)} \\ &\quad + \sum_{k\neq i} p_k\frac{(N-1)!}{(x_k-1)!\prod_{j\neq k}x_j!}\left(p_k^+\right)^{x_k-1}\prod_{j\neq k}\left(p_j^-\right)^{x_j} \\ &= p_i\binom{N-1}{x_i-1}\left(p_i^+\right)^{x_i-1}\left(q^-\right)^{N-1-(x_i-1)} \\ &\quad + \sum_{k\neq i} p_k\binom{N-1}{x_i}\left(p_i^-\right)^{x_i}\left(q^+\right)^{N-x_i-1} \\ &= p_i\binom{N-1}{x_i-1}\left(p_i^+\right)^{x_i-1}\left(q^-\right)^{N-1-(x_i-1)} \\ &\quad + q\binom{N-1}{x_i}\left(p_i^-\right)^{x_i}\left(q^+\right)^{N-x_i-1} \end{aligned}$$

□

The above theorem shows another way the generalized multinomial distribution is an extension of the generalized binomial distribution.

### 4.2 Moment Generating Function

**Theorem 4** (Moment Generating Function). *The moment generating function of the generalized multinomial distribution with K categories is given by*

$$M_{\mathbf{X}}(\mathbf{t}) = \sum_{i=1}^{K} p_i e^{t_i} \left( p_i^+ e^{t_i} + \sum_{j \neq i} p_j^- e^{t_j} \right)^{n-1} \tag{23}$$

*where* $\mathbf{X} = (X_1, ..., X_K), \mathbf{t} = (t_1, ..., t_K)$.

*Proof.* By definition, $M_{\mathbf{X}}(\mathbf{t}) = \mathbb{E}\left[e^{\mathbf{t}^T X}\right] = \sum_{\mathbf{X}} e^{\mathbf{t}^T \mathbf{x}} P(\mathbf{X} = \mathbf{x}) = \sum_{\mathbf{X}} e^{\mathbf{t}^T \mathbf{x}} \sum_{i=1}^{K} p_i \frac{(N-1)!}{(x_i-1)! \prod_{j \neq i} x_j!} \left(p_i^+\right)^{x_i - 1} \prod_{j \neq i} \left(p_j^-\right)^{x_j}$. Let $S_m = \sum_{i=1}^{m} x_i$, and Expanding the above,

$$\begin{aligned}
\mathbb{E}\left[e^{\mathbf{t}^T X}\right] = &\sum_{x_1=1}^{N} \sum_{x_2=0}^{N-x_1} \cdots \sum_{x_{k-1}=0}^{N-S_{K-2}} e^{\mathbf{t}^T \mathbf{x}} p_1 \frac{(N-1)!}{(x_1-1)! \prod_{j=2}^{k-1} x_j! (N - \sum_{j=1}^{K-1} x_j)!} \left(p_1^+\right)^{x_1 - 1} \left[ \prod_{j=2}^{K-1} \left(p_j^-\right)^{x_j} \right] \left(p_K^-\right)^{N - S_{K-1}} \\
&+ \sum_{i=2}^{K} \left[ \sum_{x_i=1}^{N} \sum_{x_1=0}^{N-S_1} \cdots \sum_{x_{i-1}=0}^{N-x_i-S_{i-2}} \sum_{x_{i+1}=0}^{N-S_i} \cdots \sum_{x_{K-1}=0}^{N-S_{K-2}} e^{\mathbf{t}^T \mathbf{x}} p_i \frac{(N-1)!}{(x_i-1)! \prod_{j=1, j \neq i}^{K-1} x_j! (N - S_{K-1})!} \right. \\
&\left. \times \left(p_i^+\right)^{x_i - 1} \left[ \prod_{\substack{j=1 \\ j \neq i}}^{K-1} \left(p_j^-\right)^{x_j} \right] \left(p_K^-\right)^{N - S_{K-1}} \right]
\end{aligned}$$

Taking the first term, denoted $T_1$, let $y = x_1 - 1$. Then

$$T_1 = p_1 e^{t_1} \sum_{y=0}^{N-1} \sum_{x_2=0}^{N-1-y} \cdots \sum_{x_{K-1}=0}^{N-1-y-\sum_{j=2}^{K-2} x_j} e^{t_1 y + t_2 x_2 + ... + t_K x_K} \left(p_1^+\right)^y \left[ \prod_{j=2}^{K-1} \left(p_j^-\right)^{x_j} \right] \left(p_K^-\right)^{N-1-y-\sum_{j=2}^{K-1} x_j}$$

The summation of the above is simply the moment generating function of a standard multinomial distribution with probabilities $\mathbf{p} = (p_1^+, p_2^-, ..., p_K^-)$. Thus,

$$T_1 = p_1 e^{t_1} \left( p_1^+ e^{t_1} + \sum_{j=2}^{K} p_j^- e^{t_j} \right)^{N-1}$$

A similar procedure follows with the remaining terms, and summing finishes the proof.

□

### 4.3 Moments of the Generalized Multinomial Distribution

Using the moment generating function in the standard way, the mean vector $\mu$ and the covariance matrix $\Sigma$ may be derived.

**Expected Value** The expected value of $\mathbf{X}$ is given by $\mu = n\mathbf{p}$ where $\mathbf{p} = (p_1, ..., p_K)$

**Covariance Matrix** The entries of the covariance matrix are given by

$$\Sigma_{ij} = \begin{cases} p_i(1-p_i)(n + \delta(n-1) + \delta^2(n-1)(n-2)), & i = j \\ p_i p_j (\delta(1-\delta)(n-2)(n-1) - n), & i \neq j \end{cases}$$

Note that if $\delta = 0$, the generalized multinomial distribution reduces to the standard multinomial distribution and $\Sigma$ becomes the familiar multinomial covariance matrix. The entries of the corresponding correlation matrix are given by

$$\rho(X_i, X_j) = -\sqrt{\frac{p_i p_j}{(1-p_i)(1-p_j)}} \left( \frac{n - \delta(n-1)(n-2)}{n + \delta(n-1) + \delta^2(n-1)(n-2)} \right)$$

If $\delta = 1$, the variance of $X_i$ tends to $\infty$ with $n$. This is intuitive, as $\delta = 1$ implies perfect dependence of $\varepsilon_2,...,\varepsilon_n$ on the outcome of $\varepsilon_1$. Thus, $X_i$ will either be 0 or $n$, and this spread increases to $\infty$ with $n$.

## 5. Generating a Sequence of Correlated Categorical Random Variables

For brevity, we will take the acronym DCRV for a **D**ependent **C**ategorical **R**andom **V**ariable. A DCRV sequence $\varepsilon = (\varepsilon_1,...,\varepsilon_n)$ is in itself a random variable, and thus has a probability distribution. In order to provide an algorithm to generate such a sequence, we first derive this probability distribution.

### 5.1 Probability Distribution of a DCRV Sequence

The probability distribution of the DCRV sequence $\varepsilon$ of length $n$ is given formally in the following theorem. The proof follows in a straightforward fashion from the construction in Section 2 and is therefore omitted.

**Theorem 5** (Distribution of a DCRV Sequence)**.** *Let* $(\Sigma,\mathcal{F},\mathbb{P}) = ([0,1],\mathcal{B},\mu)$. *Let* $\varepsilon_i : [0,1] \to \{1,...,K\}$, $i = 1,...,n$, $n \in \mathbb{N}$ *be DCRVs as defined in* (2). *Let* $\varepsilon = (\varepsilon_1,...,\varepsilon_n)$ *denote the DCRV sequence with observed values* $e = (e_1,...,e_n)$. *Then* $\mu$ *has the density*

$$f(x) = \sum_{i=1}^{K^n} K^n m^i \mathbb{1}_{((i-1)/K^n, i/K^n]}(x) \tag{24}$$

*and*

$$P(\varepsilon = e) = \int_{\left(\frac{i-1}{K^n}, \frac{i}{K^n}\right]} f(x)dx = m^i \tag{25}$$

*where* $m^i$ *is the mass allocated to the interval* $\left(\frac{i-1}{K^n}, \frac{i}{K^n}\right]$ *by* (4) *and* $i$ *as the lexicographic order of* $e$ *in the sample space* $\{1,...,K\}^n$ *given by the relation* $\frac{i}{K^n} = \sum_{j=1}^{K^n} \frac{\varepsilon_j - 1}{K^j}$.

### 5.2 Algorithm

We take a common notion of using a uniform random variable in order to generate the desired random variable $\varepsilon$. For $\varepsilon$ with distribution $F(x) = \int_0^x f(y)dy$, $f(x)$ as in (24), it is clear that $F$ is invertible with inverse $F^{-1}$. Thus, $F^{-1}(U)$ for $U \sim \text{Uniform}[0,1]$ has the same distribution as $\varepsilon$. Therefore, sampling $u$ from $U$ is equivalent to the sample $e = F^{-1}(u)$ from $\varepsilon$.

In Section 2, we associated $\varepsilon_n$ to the intervals given in (2)

$$\begin{aligned} &\varepsilon_N = i \text{ on } \left(\tfrac{l-1}{K^N}, \tfrac{l}{K^N}\right], \quad i \equiv l \mod K, \quad i = 1,...,K-1 \\ &\varepsilon_N = K \text{ on } \left(\tfrac{l-1}{K^N}, \tfrac{l}{K^N}\right], \quad 0 \equiv l \mod K \end{aligned} \tag{26}$$

From the construction in Section 2, each sequence has a 1-1 correspondence with the interval $\left[\frac{i-1}{K^n}, \frac{i}{K^n}\right)$ for a unique $i = 1,...,K^n$. The probability of such a sequence can be found using Theorem 5:

$$P(\varepsilon = e) = F\left(\left[\frac{i-1}{K^n}, \frac{i}{K^n}\right)\right) = m^i = l\left([s_{i-1}, s_i)\right)$$

where $l$ is the length of the above interval, and $s_i = \sum_{j=1}^{i} m^j$. Therefore, we have now partitioned the interval $[0,1)$ according to the distribution of $\varepsilon$ bijectively to the $K-$nary partition of $[0,1)$ corresponding to the particular sequence. Thus, sampling $u \in [0,1)$ from a uniform distribution and finding the interval $[s_{i-1}, s_i)$ and corresponding $i$ will yield the unique DCRV sequence.

**Algorithm Strategy:** Given $u \in [0,1)$ and $n \in \mathbb{N}$, find the unique interval $[s_{i-1}, s_i)$, $i = 1,...,K^n$ that contains $u$ by "moving down the tree" and narrowing down the "search interval" until level $n$ is reached.

We provide an explicit example prior to the pseudocode to illustrate the strategy.

### 5.2.1 Example DCRV Sequence Generation

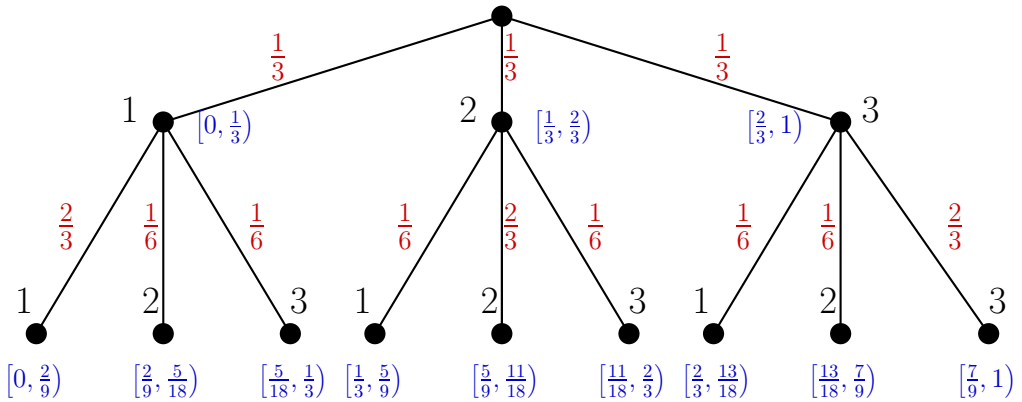


**Figure 3.** Probabilistic partitions of [0,1) for a DCRV sequence of length 2 with $K = 3$.

Suppose $K = 3$, $p_1 = p_2 = p_3 = 1/3$, and $\delta = 1/2$, and suppose we wish to generate a DCRV sequence of length 2. Figure 3 gives the probability flow and the corresponding probability intervals $[s_{i-1}, s_i)$ that partition $[0,1)$ according to Theorem 5. Now, suppose $u = \frac{3}{4}$. We now illustrate the process of moving down the above tree to generate the sequence.

1. The first level of the probability tree partitions the interval $[0,1)$ into three intervals given in Figure 3. $u = \frac{3}{4}$ lies in the third interval, which corresponds to $\varepsilon_1 = 3$. Thus, the first entry of $e$ is given by $e_1 = 3$.

2. Next, the search interval is reduced to the interval from the first step $[2/3,1)$. We then generate the partitions of $[2/3,1)$ by cumulatively adding $p_3 p_i^-, i = 1,2,3$ to the left endpoint $2/3$. Thus, the next partition points are

   - $2/3 + (1/3)(1/6) = 13/18$,
   - $2/3 + (1/3)(1/6) + (1/3)(1/6) = 7/9$, and
   - $2/3 + (1/3)(1/6) + (1/3)(1/6) + (1/3)(2/3) = 1$.

   Yielding the subintervals of $[2/3,1)$:

   - $[2/3, \quad 13/18)$,
   - $[13/18, \quad 7/9)$, and
   - $[7/9, \quad 1)$.

   We now find the interval from above that contains $u$ is the second: $[13,18, \quad 7/9)$. Thus, $\varepsilon_2 = 2$.

Since we only sought a sequence of length 2, the algorithm is finished, and we have generated the sequence $e = (3,2)$. If a longer sequence is desired, we repeat step 2 until we are at level $n$.

In general, the algorithm is given below. The **IntervalSearch**$(x, p = (p_1, p_2, ..., p_m))$ procedure finds the interval $i$ built from the partitions given in the vector $p$ containing $x$ via binary search. Let **USample**$(n)$ be the procedure that samples $n$ instances of a uniformly distributed random variable. Also, let $p_i' = (p_1^-, ..., p_{i-1}^-, p_i^+, p_{i+1}^-, ..., p_K)$ be the "altered" probability vector for the categorical variables $2, ..., n$ given $\varepsilon_1 = i$.

**Algorithm 1** DCRV Sequence Generation

**procedure** DCRVSEQUENCE(n, $\delta$, $p = (p_1, ..., p_K)$)
  $u \leftarrow$ **USample**$(1)$
  $sequence \leftarrow vector(n)$
  $partitions \leftarrow$ **cumsum**$(p_1, ..., p_k)$
  $s \leftarrow$ **IntervalSearch**$(u, partitions)$
  $sequence[1] \leftarrow s + 1$
  $p_{new} \leftarrow p'_{s+1}$
  $p_{prev} \leftarrow p_{s+1}$
  **for** $i = 2, i \leq n, i+ = 1$ **do**
    $endPoint_l \leftarrow partitions[s]$
    $endPoint_r \leftarrow partitions[s + 1]$
    $partitions \leftarrow endPoint_l +$ **cumsum**$(p_{prev} \cdot p_{new})$
    $s \leftarrow$ **IntervalSearch**$(u, partitions)$
    $sequence[i] \leftarrow s + 1$
    $l \leftarrow length(sequence)$
    $p_{prev} \leftarrow p_{prev} \cdot p'_{sequence[l]}$
  **return** sequence

## 6. Conclusion

Categorical variables play a large role in many statistical and practical applications across disciplines. Moreover, correlations among categorical variables are common and found in many scenarios, which can cause problems with conventional assumptions. Different approaches have been taken to mitigate these effects, because a mathematical framework to define a measure of dependency in a sequence of categorical variables was not available. This paper formalized the notion of dependent categorical variables under a first-dependence scheme and proved that such a sequence is identically distributed but now dependent. With an identically distributed but dependent sequence, a generalized multinomial distribution was derived in Section 3 and important properties of this distribution were provided. An efficient algorithm to generate a sequence of dependent categorical random variables was given.

## Acknowledgments

The author extends thanks to Jason Hathcock for his suggestions and review.

## References

[1] BISWAS, A. Generating correlated ordinal categorical random samples. *Statistics and Probability Letters* (2004), 25–35.

[2] HIGGS, M. D., AND HOETING, J. A. A clipped latent variable model for spatially correlated ordered categorical data. *Computational Statistics and Data Analysis* (2010), 1999–2011.

[3] IBRAHIM, N., AND SULIADI, S. Generating correlated discrete ordinal data using r and sas iml. *Computer Methods and Programs in Biomedicine* (2011), 122–132.

[4] KORZENIOWSKI, A. On correlated random graphs. *Journal of Probability and Statistical Science* (2013), 43–58.

[5] LEE, A. Some simple methods for generating correlated categorical variates. *Computational Statistics and Data Analysis* (1997), 133–148.

[6] NICODEMUS, K. K., AND MALLEY, J. D. Predictor correlation impacts machine learning algorithms: implications for genomic studies. *Bioinformatics* (2009), 1884–1890.

[7] NISTOR GROZAVU, L. L., AND BENNANI, Y. Autonomous clustering characterization for categorical data. *Machine Learning and Applications (ICMLA), 2010 Ninth International Conference on* (2010).

[8] S.J. TANNENBAUM, N.H.G. HOLFORD, H. L. E. A. Simulation of correlated continuous and categorical variables using a single multivariate distribution. *Journal of Pharmacokinetics and Pharmacodynamics* (2006), 773–794.

[9] TOLOSI, L., AND LENGAURER, T. Classification with correlated features: unreliability of feature ranking and solutions. *Bioinformatics* (2011), 1986–1994.