RT-MAE 2001-29

UMA NOVA VERSÃO DO MODELO DE URNA DE PÓLYA-EGGENBERGER

by

L.G. Esteves, S. Wechsler and P. Iglesias

Palavras-Chave: Pólya's urn scheme, urn models, exchangeable processes, De Finetti's Representation Theorem, mixtures.
Classificação AMS: 60G09, 60A99, 62F15, 62A15.
(AMS Classification)

- Dezembro de 2001 -

Uma nova versão do modelo de urna de Pólya-Eggenberger

L.G. Esteves, S. Wechsler, P. Iglesias

Abstract

In this work we present a new version of Pólya's urn scheme via the introduction of a probability distribution for the initial composition, that is, for the numbers of black and white balls. We also determine when an exchangeable process taking values in $\{0, 1\}^{\infty}$ may be well aproximated by a suitable Pólya's process with unknown initial configuration.

Keywords: Pólya's urn scheme, urn models, exchangeable processes, De Finetti's Representation Theorem, mixtures.

1 Introdução

O estudo de modelos de urnas, cujos primórdios remontam ao século XVII (segundo Johnson & Kotz, 1977, p. 22, "a primeira referência a modelos de urnas em problemas de probabilidade aparece nos trabalhos de Huygen (1629–1695)"), tem sido, desde então, alvo de trabalhos de muitos autores devido a sua importância, que, resumidamente, reside em pelo menos dois aspectos: primeiro, a possibilidade de derivação de muitos resultados da Teoria de Probabilidade (ao menos no que diz respeito à probabilidade discreta) via modelos de urnas, o que os tornam um vigoroso instrumento didático (Johnson & Kotz, 1977, por exemplo, obtêm as distribuições de probabilidade discretas comumente utilizadas em métodos estatísticos a partir de um esquema de urnas mais geral); segundo, a possibilidade de modelagem de diversos fenômenos da natureza bem como de problemas reais em várias áreas do conhecimento através de esquemas de urnas (Heitele, 1975, afirma que é possível associar modelos de urnas a grande parte dos experimentos que envolvem incerteza). Finalmente, Pólya (1954) ratifica a importância de modelos de urnas: "Qualquer problema de probabilidade parece comparável a um adequado problema de urnas contendo bolas e qualquer fenômeno aleatório parece similar, em certos aspectos essenciais, a sucessivas retiradas de bolas de um sistema de urnas combinadas convenientemente".

Johnson & Kotz (1977) distinguem os principais modelos de urnas em duas categorias: (i) os modelos de urnas com reposição de bolas, dentre os quais podemos destacar os modelos estocásticos de Pólya-Eggenberger para fenômenos envolvendo algumas formas de “contágio” e de Ehrenfest para transferência de calor entre dois corpos isolados e (ii) os modelos de urnas para problemas de ocupação (sem reposição de bolas) tais como os modelos de Bose-Einstein e de Maxwell-Boltzmann. Neste trabalho, no entanto, vamos nos ater apenas a um particular modelo da categoria (i) citada acima: o modelo de Pólya-Enggenberger e suas variações. Uma descrição bastante detalhada sobre o estudo de modelos de urnas em geral é encontrado em Johnson & Kotz (1977).

No que segue, recordamos o modelo de urna de Pólya-Eggenberger bem como apresentamos uma variação deste modelo que acreditamos não ter sido ainda contemplada na literatura.

2 O modelo de Pólya-Eggenberger e suas variações

Dentre os diversos modelos de urnas desenvolvidos para representar formas de contágio, o modelo de Pólya-Eggenberger ocupa, indubitavelmente, uma posição central, não só pelo seu cunho pioneiro nesta área de estudo, mas também pela generalidade e riqueza de propriedades que encerra. Na seqüência, faremos uma breve descrição deste modelo e de algumas de suas propriedades.

Podemos descrever o modelo de urna de Pólya-Eggenberger de modo bastante simples. Imaginemos uma urna contendo, inicialmente, a bolas brancas e b bolas pretas. Pólya e Eggenberger (1923) consideram o seguinte procedimento de retiradas sucessivas de bolas da urna: retira-se uma bola da urna de maneira equiprovável e, em seguida, retorna-se esta bola à urna juntamente com c bolas desta mesma cor (o caso particular $c = 1$ foi estudado anteriormente por Markov (1906)). Notemos que, segundo esta construção, a retirada de uma bola de uma determinada cor, digamos branca, em uma certa etapa do processo aumenta a probabilidade de nova ocorrência deste evento (retirada de uma bola branca) na etapa seguinte (Feller, 1957, dá o nome de “aftereffect” a este fenômeno em que a ocorrência de um evento aumenta (ou diminui) a probabilidade de nova ocorrência deste evento). Esta característica do processo, em concordância com seu conceito de “influence globale”, teria levado Pólya a considerá-lo um protótipo bastante razoável para descrever específicas formas de “contágio”. A seguir, apresentamos uma formalização do modelo de Pólya-Eggenberger.

Seja $\{X_n\}_{n\geq 1}$ um processo estocástico assumindo valores em $\{0,1\}^\infty$ com medida $\mathbb{P}$ tal que, $\forall n \in \mathbb{N}$, $\forall (x_1,\ldots,x_n) \in \{0,1\}^n$,

$$\mathbb{P}(X_1 = x_1,\ldots,X_n = x_n) = \frac{a(a+c)\cdots(a+(t-1)c)b(b+c)\cdots(b+(n-t-1)c)}{(a+b)(a+b+c)\cdots(a+b+(n-1)c)}$$

ou

$$P(X_1 = x_1,\ldots,X_n = x_n) = \frac{\Gamma\left(\frac{a}{c}+t_n\right)\Gamma\left(\frac{b}{c}+n-t_n\right)\Gamma\left(\frac{a+b}{c}\right)}{\Gamma\left(\frac{a}{c}\right)\Gamma\left(\frac{b}{c}\right)\Gamma\left(\frac{a+b}{c}+n\right)}, \qquad (2.1)$$

onde $t_n = \sum_{i=1}^{n} x_i$, $a, b, c \in \mathbb{N}$. O processo $\{X_n\}_{n \geq 1}$ descreve a evolução do modelo de urna de Pólya-Eggenberger com configuração inicial (a, b), isto é, com a bolas brancas e b bolas pretas inicialmente na urna, e c bolas acrescidas em cada etapa. X_n corresponde então à variável indicadora de retirada de uma bola branca na n-ésima etapa do processo, $n \in \mathbb{N}$. De (2.1), é fácil ver que o processo de Pólya-Eggenberger é permutável. Deste modo, o processo de Pólya-Eggenberger satisfaz as condições do Teorema da Representação de De Finetti e, portanto, qualquer marginal n-dimensional deste processo pode ser escrita como uma mistura de n variáveis aleatórias de Bernoulli independentes e identicamente distribuídas (neste caso, a medida de De Finetti, ou misturadora, possui função densidade de probabilidade Beta com parâmetros $\frac{a}{c}$ e $\frac{b}{c}$). A partir do processo de Pólya-Eggenberger, são derivadas as distribuições de Pólya-Eggenberger e Pólya inversa, bem como suas versões multivariadas para as situações em que existem mais de duas cores para as bolas da urna. Um estudo bastante detalhado dessas distribuições e outras caracterizações do modelo de Pólya-Eggenberger são encontrados em Johnson & Kotz (1977).

A partir do trabalho de Pólya e Eggenberger (1923), modificações (ou generalizações) em várias direções para o modelo de Pólya-Eggenberger têm sido desenvolvidas, desde a consideração de mais de duas cores para as bolas na urna até a adoção de diferentes mecanismos de reposição de bolas à urna. Dentre estas alterações destacamos: (1) reposição não apenas de c bolas da mesma cor da bola retirada, mas também de d bolas da outra cor; (2) reposição de bolas das duas cores, como em (1), dependendo da etapa do processo, isto é, $c = c_n$ e $d = d_n$, ou da cor da bola retirada, $c = c_n(x_n)$ e $d = d_n(x_n)$ e (3) reposição de números C e D aleatórios de bolas à urna. Em qualquer uma das três situações acima, o processo obtido não é, em geral, permutável. No que segue, apresentamos agora uma nova versão para o modelo de urna de Pólya-Eggenberger.

Consideremos uma urna contendo bolas brancas e pretas. Suponhamos, no entanto, que sejam desconhecidos os números iniciais A e B de bolas brancas e pretas, respectivamente. Como o par (A, B) é desconhecido, é natural, do ponto de vista subjetivista, que atribuamos uma medida de probabilidade sobre $\mathbb{N}^2$ que expresse nossa incerteza a respeito da composição inicial da urna. Assim, a configuração inicial da urna para o modelo que propomos é um vetor aleatório em $(\mathbb{N}^2, \mathcal{P}(\mathbb{N}^2))$. Quanto à evolução do processo, admitimos que, em cada etapa, seja escolhida uma bola da urna uniformemente e que, em seguida, esta seja reposta à urna juntamente com outra bola da mesma cor (notemos que este é o procedimento de retiradas sucessivas do modelo de urna de Pólya-Eggenberger com $c = 1$). Ao processo descrito acima, damos o nome de Modelo de Urna de Pólya-Eggenberger com composição (ou configuração) inicial aleatória. Neste caso, se $\{((a_k, b_k), p_k)\}_{k \in \mathbb{N}}$, $(a_k, b_k) \in \mathbb{N}^2$, $p_k \geq 0$, $k \in \mathbb{N}$, com $\sum_{k \in \mathbb{N}} p_k = 1$, exprime nossa incerteza a respeito da composição inicial da urna, temos, utilizando a notação do começo desta seção, que $\forall n \in \mathbb{N}$, $\forall (x_1, \ldots, x_n) \in \{0,1\}^n$,

$$\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n) = \sum_{k=1}^{\infty} p_k \cdot \text{Pólya}\left((x_1, \ldots, x_n)|(a_k, b_k)\right),$$

ou

$$\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n) = \int_{\mathbb{N}^2} \text{Pólya}\left((x_1, \ldots, x_n)|(a, b)\right) \mathrm{d}\mu(a, b), \tag{2.2}$$

onde $\mu : \mathcal{P}(\mathbb{N}^2) \to [0,1]$ é tal que $\mu(A) = \sum_{k:(a_k,b_k)\in A} p_k$, $A \in \mathcal{P}(\mathbb{N}^2)$, e

$$\text{Pólya}\left((x_1, \ldots, x_n)|(a, b)\right) = \frac{\Gamma\left(a + \sum_{i=1}^{n} x_i\right) \Gamma\left(b + n - \sum_{i=1}^{n} x_i\right) \Gamma(a+b)}{\Gamma(a)\Gamma(b)\Gamma(a+b+n)}.$$

A expressão (2.2) acima é facilmente obtida condicionando-se a evolução do processo à composição inicial da urna, isto é,

$$P(X_1 = x_1, \ldots, X_n = x_n) = \sum_{k\in\mathbb{N}} P((A, B) = (a_k, b_k)) \prod_{j=1}^{n} P(X_j = x_j \mid x_1, \ldots, x_{j-1}, (a_k, b_k)).$$

Notemos que em (2.2) poderíamos ter formalmente considerado medidas misturadoras μ mais gerais sobre $(\mathbb{R}_+^2, \mathcal{B}(\mathbb{R}_+^2))$. No entanto, como tal medida deve refletir nossa incerteza sobre a composição inicial da urna, (A, B), sabidamente assumindo valores em $\mathbb{N}^2$, não contemplamos estas situações neste trabalho. No que segue, apresentamos algumas propriedades (sem demostração) do Modelo de Urna de Pólya-Eggenberger com configuração inicial aleatória.

Proposição 2.1 *Qualquer que seja a medida misturadora μ, o modelo de Urna de Pólya-Eggenberger com composição inicial aleatória é permutável, isto é, $\forall n \in \mathbb{N}$, $\forall (x_1, \ldots, x_n) \in \{0,1\}^n$ e $\forall \pi : \{1, \ldots, n\} \to \{1, \ldots, n\}$ bijetora (permutação dos índices $1, \ldots, n$)*

$$P(x_{\pi(1)}, \ldots, x_{\pi(n)}) = P(x_1, \ldots, x_n).$$

A Proposição 2.1 revela que o Modelo de Urna de Pólya-Eggenberger com composição inical aleatória é permutável, ao contrário das outras variações do Modelo de Pólya-Eggenberger mencionadas anteriormente, que não gozam de tal propriedade. Também por isso, vemos que o modelo que propomos não corresponde a generalização alguma do modelo de urna de Pólya-Eggenberger, dentre as mencionadas. Ademais, o modelo de urna de Pólya-Eggenberger com configuração inicial aleatória, sendo um processo permutável em $\{0,1\}^\infty$, satisfaz as condições do Teorema da Representação de De Finetti (1937), de modo que qualquer marginal n-dimensional deste processo equivale a uma mistura de n variáveis de Bernoulli independentes e identicamente distribuídas. A seguir, caracterizamos então a medida de De Finetti deste processo.

Proposição 2.2 *Seja* $\{X_n\}_{n\geq 1}$ *um modelo de urna de Pólya-Eggenberger com configuração inicial aleatória dada por* $\{((a_k, b_k); p_k)\}_{k\in\mathbb{N}}$, $(a_k, b_k) \in \mathbb{N}^2$, $p_k \geq 0$, $k \in \mathbb{N}$, *com* $\sum_{k\in\mathbb{N}} p_k = 1$. *Então a medida de De Finetti deste processo possui função densidade de probabilidade dada por*

$$g(\theta) = \sum_{k\in\mathbb{N}} p_k \mathrm{Beta}\,(\theta; (a_k, b_k)), \tag{2.3}$$

onde $\mathrm{Beta}\,(\theta; (a,b)) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}\mathbb{I}_{[0,1]}(\theta)$.

A Proposição 2.2 mostra que o modelo de urna de Pólya-Eggenberger com configuração inicial dada por $\{((a_k, b_k); p_k)\}_{k\in\mathbb{N}}$ tem medida de De Finetti possuindo função densidade de probabilidade igual a uma mistura enumerável de funções densidade Betas, a saber, Betas com parâmetros a_k e b_k, $k \in \mathbb{N}$, ponderadas pelos respectivos pesos p_k.

Na próxima seção, apresentamos o principal resultado deste trabalho, que afirma que, sob certas condições, um processo permutável em $\{0,1\}^\infty$ pode ser "bem aproximado" por um modelo de urna de Pólya-Eggenberger com configuração inicial aleatória.

3 Resultado principal

Estabelecemos, a partir de agora, condições para que a medida de probabilidade de um processo permutável a valores em $\{0,1\}^\infty$ possa ser bem aproximada, em distância de variação total, pela medida de um conveniente modelo de urna de Pólya-Eggenberger com composição inicial aleatória. Antes, porém, relembremos o Teorema de Stone-Weierstrass, em sua versão para o caso particular da reta real devida a Bernstein, que terá papel fundamental nos resultados subseqüentes.

Teorema 3.1 (S. Bernstein) *Seja* $f : [0,1] \to \mathbb{R}$ *contínua. Então, a seqüência de polinômios de Bernstein*

$$P_n(x) = \sum_{k=0}^{n} \binom{n}{k} f\left(\frac{k}{n}\right) x^k(1-x)^{n-k}\ ,\ n = 1, 2, \ldots$$

converge uniformemente a f *quando* $n \to \infty$.

Marsden (1974) argumenta que o conhecimento de Teoria da Probabilidade possibilitou a Bernstein melhor compreender e demonstrar este teorema. De fato, várias passagens da prova acima correspondem a resultados bem conhecidos em probabilidade. Além disso, a consideração de um cenário envolvendo um jogo com premiação fornece uma explicação intuitiva, ainda que sob uma perspectiva frequentista, da validade do teorema 3.1. Com efeito, considerando-se n lançamentos independentes de uma moeda, nos quais a probabilidade de resultar cara é x, e admitindo-se que $f(\frac{k}{n})$ unidades monetárias são pagas a um jogador quando k caras são obtidas nestes n lançamentos da moeda, $P_n(x)$ é o ganho esperado deste jogador. Para valores grandes de n, a fração de lançamentos que resultam em cara aproxima-se de x e, consequentemente, o ganho médio de tal jogador aproxima-se de $f(x)$, isto é, $P_n(x)$ deve aproximar $f(x)$.

Na seqüência, exibimos as condições para que a solução do problema dos momentos de Hausdorff (Shohat e Tamarkin (1943)) seja absolutamente contínua possuindo função densidade limitada.

Teorema 3.2 *Sejam $\{\alpha_n\}_{n\geq 1}$, $\alpha_n \in [0,1]$, uma sequência de números e $\psi(t)$ uma função de distribuição com $\psi(0) = 0$ e $\psi(1) = 1$ tais que*

$$\alpha_n = \int_0^1 t^n \mathrm{d}\psi(t), \quad \forall\, n \in \mathbb{N}, \tag{3.1}$$

Então, $\psi(t) = \int_0^t \varphi(\mu)\mathrm{d}\mu$, com $\varphi \geq 0$ limitada, se, e somente se, existe $M > 0$ tal que, $\forall n \in \mathbb{N}$, tem-se

$$\max_{v\in\{0,\ldots,n\}} \left\{ \binom{n}{v} \int_0^1 t^v (1-t)^{n-v} \mathrm{d}\psi(t) \right\} \leq \frac{M}{n+1}.$$

As demonstrações dos teoremas 3.1 e 3.2 são encontradas, respectivamente, em Marsden (1974) e Shohat e Tamarkin (1943).Enunciamos a seguir o principal resultado deste capítulo.

Teorema 3.3 *Seja $\{X_n\}_{n\geq 1}$ um processo permutável com valores em $\{0,1\}^\infty$. Suponha que exista $M > 0$ tal que, $\forall n \in \mathbb{N}$, $\max_{v\in\{0,\ldots,n\}} P\left(\sum_{i=1}^n X_i = v\right) \leq \frac{M}{n+1}$. Então, $\forall \varepsilon > 0$, $\exists m_0 = m_0(\varepsilon) \in \mathbb{N}$ e $\{p_0, \ldots, p_{m_0}\}$ com $p_k \geq 0$, $k = 0, \ldots, m_0$, e $\sum_{k=0}^{m_0} p_k = 1$, tais que $\forall n \in \mathbb{N}$ e $\forall (x_1, \ldots, x_n) \in \{0,1\}^n$,*

$$\left| P(x_1, \ldots, x_n) - \sum_{k=0}^{m_0} p_k \cdot \text{Pólya}((x_1, \ldots, x_n)|(k+1, m_0 - k + 1)) \right| < \varepsilon, \tag{3.2}$$

onde P denota a medida do processo $\{X_n\}_{n\geq 1}$.

Prova Como $\{X_n\}_{n\geq 1}$ é permutável assumindo valores em $\{0,1\}^\infty$, segue, pelo Teorema da Representação de De Finetti, que, $\forall n \in \mathbb{N}$, $\forall (x_1, \ldots, x_n) \in \{0,1\}^n$,

$$P(x_1, \ldots, x_n) = \int_0^1 \theta^{\sum_{i=1}^n x_i} (1-\theta)^{n - \sum_{i=1}^n x_i} \mathrm{d}\psi(\theta),$$

onde $\psi : \mathcal{B}([0,1]) \to [0,1]$ é a medida de De Finetti. Em particular, para o vetor $1_n = (1, \ldots, 1)_{1\mathrm{x}n}$, temos

$$\underbrace{P(1_n)}_{\alpha_n} = \int_0^1 \theta^n \mathrm{d}\psi(\theta), \quad \forall n \in \mathbb{N}.$$

Como por hipótese existe $M > 0$ tal que

$$\max_{v\in\{0,\ldots,n\}} P\left(\sum_{i=1}^{n} X_i = v\right) = \max_{v\in\{0,\ldots,n\}} \binom{n}{v} \int_0^1 \theta^v (1-\theta)^{n-v} \mathrm{d}\psi \leq \frac{M}{n+1}, \forall n \in \mathbb{N},$$

segue, pelo Teorema 3.2, que ψ é absolutamente contínua e que $\exists\ \varphi : [0,1] \to \mathbb{R}_+$, limitada, tal que

$$\psi(t) = \varphi(t)\mathrm{d}t\ .$$

Neste caso, existe $\{\varphi_n\}_{n\geq 1}$, $\varphi_n \geq 0$, simples, $\forall n \in \mathbb{N}$, tal que $\varphi_n \xrightarrow{u} \varphi$. Assim, $\forall \varepsilon > 0$, $\exists\ \varphi_{n_0} \geq 0$, simples, tal que

$$\int_0^1 |\varphi - \varphi_{n_0}| \mathrm{d}\lambda_1 \ < \ \frac{\varepsilon}{4}\ . \tag{3.3}$$

Além disso, existem uma "step function" $S : [0,1] \to \mathbb{R}_+$ tal que

$$\int_0^1 |\varphi_{n_0} - S| \mathrm{d}\lambda_1 \ < \ \frac{\varepsilon}{4}\ , \tag{3.4}$$

uma função contínua limitada $C : [0,1] \to \mathbb{R}_+$ tal que

$$\int_0^1 |S - C| \mathrm{d}\lambda_1 \ < \ \frac{\varepsilon}{4} \tag{3.5}$$

e uma sequência de polinômios de Bernstein $\{B_m\}_{m\geq 1}$ tal que $B_m \xrightarrow{u} C$. Avaliemos, então,

$$\left| P(x_1,\ldots,x_n) - \sum_{k=0}^{m} p_k^{(m)} \cdot \text{Pólya}\left((x_1,\ldots,x_n)|(k+1, m-k+1)\right) \right|,$$

com $p_k^{(m)} = \dfrac{C(\frac{k}{m})}{\sum_{j=0}^{m} C(\frac{j}{m})}$, $k = 0,\ldots,m$. Temos

$$\left| P(x_1,\ldots,x_n) - \sum_{k=0}^{m} p_k^{(m)} \cdot \text{Pólya}\left((x_1,\ldots,x_n)|(k+1, m-k+1)\right) \right| =$$

$$= \left| \int_0^1 \theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i}\varphi(\theta)\mathrm{d}\theta \right. -$$

$$\left. - \sum_{k=0}^{m} p_k^{(m)} \cdot \int_0^1 \theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i}\,\mathrm{Beta}\,(\theta;(k+1,m-k+1))\mathrm{d}\theta \right| ,$$

onde a última igualdade segue do Teorema da Representação de De Finetti e (??). Pelo Teorema de Tonelli, segue que

$$= \left| P(x_1,\ldots,x_n) - \sum_{k=0}^{m} p_k^{(m)} \cdot \text{Pólya}\,((x_1,\ldots,x_n);(k+1,m-k+1)) \right| =$$

$$= \left| \int_0^1 \theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i}\varphi(\theta)\mathrm{d}\theta \right. -$$

$$\left. - \int_0^1 \theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i}\left(\sum_{k=0}^{m} p_k^{(m)} \cdot \mathrm{Beta}\,(\theta;(k+1,m-k+1))\right)\mathrm{d}\theta \right| =$$

$$= \left| \int_0^1 \theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i}\left[\varphi(\theta) - \sum_{k=0}^{m} \frac{C(\frac{k}{m})}{\sum_{j=0}^m C(\frac{j}{m})}\frac{\Gamma(m+2)}{\Gamma(k+1)\Gamma(m-k+1)}\theta^k(1-\theta)^{m-k}\right]\mathrm{d}\theta \right|$$

$$= \left| \int_0^1 \theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i}\left[\varphi(\theta) - \frac{1}{\sum_{j=0}^m \frac{C(\frac{j}{m})}{m+1}} \sum_{k=0}^{m} \binom{m}{k} C\left(\frac{k}{m}\right)\theta^k(1-\theta)^{m-k}\right]\mathrm{d}\theta \right| =$$

$$= \left| \int_0^1 \theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i}\left[\varphi(\theta) - \frac{P_m(\theta)}{b_m}\right]\mathrm{d}\theta \right| ,$$

onde $P_m(\theta) = \sum_{k=0}^m \binom{m}{k} C\left(\frac{k}{m}\right)\theta^k(1-\theta)^{m-k}$ e $b_m = \dfrac{\sum_{j=0}^m C(\frac{j}{m})}{m+1}$. Por fim,

$$\left| P(x_1,\ldots,x_n) - \sum_{k=0}^{m} p_k^{(m)} \cdot \text{Pólya}\,((x_1,\ldots,x_n)|(k+1,m-k+1)) \right| =$$

$$= \left| \int_0^1 \theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i}\left[\varphi(\theta) - \frac{P_m(\theta)}{b_m}\right]\mathrm{d}\theta \right| \leq$$

$$\leq \int_0^1 \theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i}\left|\varphi(\theta) - \frac{P_m(\theta)}{b_m}\right|\mathrm{d}\theta \,.$$

Mas, como $b_m \to 1$, quando $m \to \infty$, temos que

$$\frac{P_m(\theta)}{b_m} \xrightarrow{u} C(\theta), \quad \text{quando } m \to \infty$$

e, portanto, $\forall \varepsilon > 0$, $\exists m_0 = m_0(\varepsilon) \in \mathbb{N}$ tal que se $m \geq m_0$, então

$$\int_0^1 \left| \frac{P_m(\theta)}{b_m} - C(\theta) \right| d\lambda_1 \ < \ \frac{\varepsilon}{4} \tag{3.6}$$

Assim, temos, finalmente, que

$$\left| P(x_1, \ldots, x_n) - \sum_{k=0}^{m_0} p_k^{(m_0)} \cdot \text{Pólya}\left((x_1, \ldots, x_n)|(k+1, m-k+1)\right) \right| \ \leq$$

$$\leq \ \int_0^1 \theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i} \left| \varphi(\theta) - \frac{P_m(\theta)}{b_m} \right| d\theta \ \leq$$

$$\leq \ \int_0^1 \left| \varphi(\theta) - \frac{P_m(\theta)}{b_m} \right| d\theta \ \leq$$

$$\leq \ \int_0^1 |\varphi - \varphi_{n_0}| \, d\lambda_1 \ + \ \int_0^1 |\varphi_{n_0} - S| \, d\lambda_1 \ + \ \int_0^1 |S - C| \, d\lambda_1 \ + \ \int_0^1 \left| C - \frac{P_{m_0}}{b_{m_0}} \right| d\lambda_1 \ \Rightarrow$$

$$\Rightarrow \ \left| P(x_1, \ldots, x_n) - \sum_{k=0}^{m_0} p_k^{(m_0)} \cdot \text{Pólya}\left((x_1, \ldots, x_n)|(k+1, m-k+1)\right) \right| \ < \ \varepsilon \ ,$$

onde a última desigualdade segue de (3.3), (3.4), (3.5) e (3.6). Portanto, $\forall \varepsilon > 0$, $\exists m_0 = m_0(\varepsilon) \in \mathbb{N}$ e $\{p_0, \ldots, p_{m_0}\}$ com $p_k \geq 0$, $k = 0, \ldots, m_0$, e $\sum_{k=0}^{m_0} p_k = 1$, tais que $\forall n \in \mathbb{N}$ e $\forall (x_1, \ldots, x_n) \in \{0,1\}^n$,

$$\left| P(x_1, \ldots, x_n) - \sum_{k=0}^{m_0} p_k \cdot \text{Pólya}\left((x_1, \ldots, x_n)|(k+1, m_0-k+1)\right) \right| < \varepsilon \ ,$$

concluindo a demonstração. ■

O Teorema 3.3 estabelece que qualquer processo permutável com valores em $\{0,1\}^\infty$ satisfazendo a condição, $\forall n \in \mathbb{N}$, $\max_{v \in \{0,\ldots,n\}} P\left(\sum_{i=1}^n X_i = v\right) \leq \frac{M}{n+1}$, para algum $M > 0$, pode ser bem aproximado por um adequado modelo de urna de Pólya-Eggenberger com configuração inicial aleatória. Mais precisamente, a probabilidade de qualquer evento relativo a uma marginal n-dimensional deste processo, $n \in \mathbb{N}$, pode ser avaliada, com erro máximo $\varepsilon > 0$ fixado, pela probabilidade do correspondente evento num modelo de urna de Pólya-Eggenberger com composição inicial aleatória contendo, inicialmente, k (k desconhecido) bolas brancas e m_0+2-k bolas pretas com probabilidade $p_{k-1}^{(m_0)}$ especificada no teorema, $k = 1, \ldots, m_0 + 1$.

A condição citada acima indica que tal aproximação é possível sempre que a distribuição de $\sum_{i=1}^{n} X_i$ não apresente, em seu suporte, pontos com probabilidade de ordem superior a $\frac{1}{n}$, isto é, sempre que a sequência $\{(n+1)P\left(\sum_{i=1}^{n} X_i = k\right)\}_{n\geq 1}$ seja convergente (finita), $\forall k \in \mathbb{N}$. Esta condição é suficiente para que a medida de De Finetti do processo $\{X_n\}_{n\geq 1}$ sob consideração seja absolutamente contínua possuindo função densidade de probabilidade limitada (na verdade, é também condição necessária). Deste modo, o resultado do Teorema 3.3 não se aplica, por exemplo, a um processo estocástico permutável em $\{0,1\}^{\infty}$ cuja medida de De Finetti possui função densidade de probabilidade Beta (1/2,1), que não é limitada (com efeito, temos, neste caso, $\lim_{n\to\infty}(n+1)P\left(\sum_{i=1}^{n} X_i = 0\right) = \infty$).

Ainda nas condições do Teorema 3.3, com a hipótese adicional da medida de De Finetti do processo possuir função densidade de probabilidade não apenas limitada, mas também contínua em $[0,1]$, obtemos uma majoração bem mais "fina" em (3.2). Neste caso, temos, $\forall n \in \mathbb{N}$ e $\forall (x_1, \ldots, x_n) \in \{0,1\}^n$,

$$\left| P(x_1, \ldots, x_n) - \sum_{k=0}^{m_0} p_k^{(m_0)} \cdot \text{Pólya}\left((x_1, \ldots, x_n)|(k+1, m-k+1)\right) \right| <$$

$$\leq \frac{\varepsilon\, \Gamma\left(\sum_{i=1}^{n} x_i + 1\right) \Gamma\left(n - \sum_{i=1}^{n} x_i + 1\right)}{\Gamma(n+2)} \leq \frac{\varepsilon}{n+1} << \varepsilon \quad \text{e},$$

além disso, a distribuição do número de "sucessos" ($\{X_i = 1\}$) nas n primeiras etapas do processo, $\sum_{i=1}^{n} X_i$, $n \geq 1$, é bem aproximada por uma mistura de distribuições de Pólya-Eggenberger, isto é,

$$\left| P\left(\sum_{i=1}^{n} X_i = t\right) - \sum_{k=0}^{m_0} p_k \cdot P - E(t|(k+1, m_0 - k + 1)) \right| < \varepsilon,$$

$\forall n \in \mathbb{N}$, $\forall t = 0, 1, \ldots, n$, $P - E(\ \cdot\ |(k+1, m_0 - k + 1))$ denotando a distribuição de Pólya-Eggenberger para o número de retiradas de bola branca nas n primeiras etapas do processo de urna de Pólya-Eggenberger com configuração inicial $(k+1, m_0 - k + 1)$:

$$P - E(t|(k+1, m_0 - k + 1)) = \binom{n}{t} \text{Pólya}\left((x_1, \ldots, x_n)|(k+1, m_0 - k + 1)\right),$$

com $(x_1, \ldots, x_n) \in \{0,1\}^n$ tal que $\sum_{i=1}^{n} x_i = t$.

Devemos salientar que o resultado do Teorema 3.3 é essencialmente qualitativo, no sentido em que apenas aponta a existência de um modelo de urna de Pólya-Eggenberger com configuração inicial aleatória conveniente para "aproximar" o processo estocástico permutável original. Assim, neste trabalho, não contemplamos, por exemplo, a questão da possível otimalidade do modelo de urna de Pólya-Eggenberger com configuração inicial aleatória considerado no Teorema 3.3 (Diaconis e Ylvisaker (1985) sugerem uma modificação nos coeficientes do Polinômio de Bernstein de modo a obter uma majoração mais "fina" no Teorema 3.1).

Destacamos ainda que o Teorema 3.3, embora estabeleça condições somente sobre as quantidades observáveis $\{X_n\}_{n\geq 1}$ para que (3.2) se verifique, não corresponde a uma representação preditivista exata para o modelo de urna de Pólya-Eggenberger com configuração inicial aleatória com $c = 1$. Afinal, o resultado em (3.2) fornece apenas aproximações para as distribuições marginais do processo $\{X_n\}_{n\geq 1}$, não garantindo a existência de uma medida $\mu^* : \mathcal{P}(\mathbb{N}^2) \to [0,1]$ tal que, $\forall n \in \mathbb{N}$ e $\forall (x_1, \ldots, x_n) \in \{0,1\}^n$,

$$P(x_1, \ldots, x_n) \;=\; \int_{\mathbb{N}^2} \text{Pólya}((x_1, \ldots, x_n)|(a,b))\mathrm{d}\mu^*(a,b)\ .$$

De fato, como a sequência de medidas em $(\mathbb{N}^2, \mathcal{P}(\mathbb{N}^2))$ $\{\{((k+1, m-k+1), p_k^m)\}_{k=0}^m\}_{m\geq 1}$ não é "tight" (pois, seus respectivos suportes são $\{(a,b) \in \mathbb{N}^2 : a+b = m+2\}$), não podemos assegurar a existência de tal medida μ^*. Ainda assim, a hipótese do teorema 3.3 é totalmente preditivista, uma vez que leva em conta somente condições sobre as quantidades observáveis $X_1, X_2, \ldots$.

Devemos salientar ainda que o resultado do Teorema 3.3 possui forte apelo intuitivo, uma vez que envolve a incerteza que possuímos a respeito da composição inicial de uma urna contendo bolas de duas cores; os modelos de urnas são objetos de concepção bem mais simples, se não mais natural, que limites de freqüências relativas ou probabilidades de cara em um lançamento de uma moeda (usuais interpretações para a quantidade aleatória θ que figura no Teorema da Representação de De Finetti), além de serem passíveis de experimentação e de possuirem grande flexibilidade para modelagem probabilística (Freudenthal (1960) afirma que "Lançamentos de moedas ou dados ou jogos de cartas não são suficientemente flexíveis. O instrumento de aleatoriedade mais geral é a urna preenchida com bolas de diferentes cores").

4 Conclusões

Neste capítulo, apresentamos uma nova versão para o modelo de Pólya-Eggenberger, introduzindo uma distribuição de probabilidade para a composição inicial da urna. Este novo processo, diferentemente de outras variações do modelo de Pólya-Eggenberger, é permutável, com medida de De Finetti possuindo função densidade de probabilidade igual a uma mistura de funções densidade Betas. Exibimos, também, em nosso principal resultado, uma condição suficiente e preditivista para que um processo permutável a valores em $\{0,1\}^\infty$ possa ser bem aproximado por um conveniente modelo de urna de

Pólya-Eggenberger com configuração inicial aleatória. Além disso, mostramos que o fortalecimento desta condição conduz a uma aproximação bem melhor, de ordem $\frac{1}{n}$. Devemos destacar que o resultado do teorema 3.3 possui forte apelo intuitivo, uma vez que envolve a incerteza que possuímos a respeito da composição inicial de uma urna contendo bolas de duas cores, ao invés de objetos de concepção mais sofisticada como limites de frequencias relativas ou probabilidades de cara em um lançamento de uma moeda.

Algumas questões, no entanto, não foram contempladas neste trabalho e, indubitavelmente, constituem objeto de futura investigação. Dentre outras, destacamos a obtenção de condições necessárias para a validade do teorema 3.3 (do qual foi caracterizada apenas uma condição suficiente), a possível derivação de uma representação preditivista exata para o modelo de urna de Pólya-Eggenberger, o que corresponde à especificação de condições adicionais sobre um processo permutável a valores em $\{0,1\}^\infty$ de modo a restringir a sua medida de De Finetti à classe das medidas em $([0,1],\mathcal{B}([0,1]))$ com função densidade igual a mistura de Betas (na verdade, desconhecemos a existência de uma representação preditivista neste caso), a introdução de uma medida de probabilidade conjunta para a configuração inicial da urna (A,B) e para o número de bolas repostas à urna em cada etapa (C) e, também, a determinação de condições para que as distribuições preditivas de um processo estocástico em $\{0,1\}^\infty$ possam ser bem aproximadas por respectivas distribuições preditivas de um processo de urna de Pólya-Eggenberger com configuração inicial aleatória (Dalal e Hall (1983) demonstram resultados desta natureza para famílias de distribuições conjugadas em Inferência Bayesiana).

Referências

[1] Dalal, S.R.; Hall, W.J. (1983). *Approximating Priors by Mixtures of Natural Conjugate Priors.* Journal of Royal Statistical Society, 45(B), 2, 278-286.

[2] de Finetti, B. (1937). *Le prévision: ses lois logiques, ses sources subjectives.* Ann. Inst. H. Poincaré. Prob. Statist., 7, 1–68.

[3] Diaconis, P.; Ylvisaker, D. (1985). *Quantifying prior opinion.* Em Bayesian Statistics, 2, Elsevier Science Publishers B. V. and Valencia University Press.

[4] Feller, W. (1957). *An Introduction to Probability Theory and its Applications.* Vol. I. Wiley, New York.

[5] Freudenthal, H. (1960). *Models in applied probability.* Synthèse, 12, 202–212.

[6] Heitele, D. (1975). *An epistemological view on fundamental stochastic ideas.* Educational Studies in Mathematics, Vol. 6, 187–205. Reidel, Dordrecht, Netherlands.

[7] Johnson, N.L.; Kotz, S. (1977). *Urn Models and Their Applications.* Wiley, New York.

[8] Markov, A.A. (1906). *Extension of the law of large numbers to dependent variables.* Izv. Fiz-Mat. Obschch. Kazan Univ., Ser. 2, 15(4), 135–256. (in Russian)

[9] Marsden, J.E. (1974). *Elementary Classical Analysis.* W. H. Freeman and Company, San Francisco.

[10] Pólya, G.; Eggenberger, F. (1923). *Uber die Statistik verketteter Vorgange.* Z. Angew. Math. Mech., 3, 279–289.

[11] Pólya, G. (1954). *Patterns of Plausible Inference.* Princeton University Press, Princeton, N.J..

[12] Shohat, J.A.; Tamarkin, J.D. (1943). *The problem of moments.* American Mathematical Society, New York.

[13] Stieltjes, T.J. (1894 - 1895). *Recherches sur les fractions continues.* Annales de la Faculté des Sciences de Toulouse, (1) 8 (1894), T 1–122, (1) 9 (1895), A 5–47.

ÚLTIMOS RELATÓRIOS TÉCNICOS PUBLICADOS

2001-01 - KOTTAS, A., BRANCO, M.D., GELFAND, A.E., A Nonparametric bayesian modeling approach for cytogenetic dosimetry. 2001. 19p. (RT-MAE-2001-01)

2001-02 - AOKI, R., BOLFARINE, H., SINGER, J.M., Null Intercept measurement error regression models. 2001. 18p.(RT-MAE-2001-02)

2001-03 - AOKI, R., BOLFARINE, H., SINGER, J.M., Asymptotic Efficiency of null intercept measurement error regression models. 2001. 11p. (RT-MAE-2001-03)

2001-04 - ALMEIDA, S.S., LIMA, C.R.O.P., SANDOVAL, M.C., Linear Calibration in functional models without the normality assumption. 2001. 18p. (RT-MAE-2001-04)

2001-05 - GARCIA-ALFARO, K.E., BOLFARINE, H. Comparative calibration with subgroups. 2001. 18p. (RT-MAE-2001-05)

2001-06 - DUNLOP, F.M., FERRARI, P.A., FONTES, L.R.G. A dynamic one-dimensional interface interacting with a wall. 2001. 21p. (RT-MAE-2001-06)

2001-07 - FONTES, L.R., ISOPI, M., NEWMAN, C.M., STEIN, D.L. 1*D* Aging. 2001. 10p. (RT-MAE-2001-07)

2001-08 - BARROS, S.R.M., FERRARI, P.A., GARCIA, N.L., MARTÍNEZ, S. Asymptotic behavior of a stationary silo with absorbing walls. 2001. 20p. (RT-MAE-2001-08)

2001-09 - GONZALEZ-LOPEZ, V.A., TANAKA, N.Y. Characterization of copula and its relationships with TP_2 (RR_2) association. 2001. 24p. (RT-MAE-2001-09)

2001-10 - ORTEGA, E.M.M., BOLFARINE, H., PAULA, G.A. Influence Diagnostics in Generalized Log-Gamma Regression Models. 2001. 24p. (RT-MAE-2001-10)

2001-11 - M.D, BRANCO, BOLFARINE, H., IGLESIAS, P., ARELLANO-VALLE, R.B. Bayesian and classical solutions for binomial cytogenetic dosimetry problem. 2001. 16p. (RT-MAE-2001-11)

2001-12 - GONÇALEZ-LOPES, V.A., TANAKA, N.I. Dependence structures and a posteriori distributions. 2001. 41p. (RT-MAE-2001-12)

2001-13 - MUTAFCHIEV, L., KOLEV, N. The Number of Empty Cells in an Allocation Scheme Generated by a Zero-Inflated Distribution: Exact Results and Poisson Convergence. 2001. 11p. (RT-MAE-2001-13)

2001-14 - CORDEIRO, G.M., BOTTER, D.A., BARROSO, L.P., FERRARI, S.L.P. Three Corrected Score Tests for Generalized Linear Models with Dispersion Covariates. 2001. 19p. (RT-MAE-2001-14)

2001-15 - TAVARES, H.R., ANDRADE, D.F. Item Response Theory for Longitudinal Data: item Parameter Estimation. 2001. 13p. (RT-MAE-2001-15)

2001-16 - WECHSLER, S., ESTEVES, L.G., SIMONIS, A., PEIXOTO, C.M. Indifference, Neutrality and Informativeness: Generalizing the Three Prisoners Paradox. 2001. 12p. (RT-MAE-2001-16)

2001-17 - FONTES, L.R., SCHONMANN, R.H., SIDORAVICIUS, V. Stretched Exponential Fixation in Stochastic Ising Models at Zero Temperature. 2001. 25p. (RT-MAE-2001-17)

2001-18 - BUENO, V.C. Minimal Standby Redundancy Allocation in a K-Out-Of-N:F System of Dependent Components. 2001. 13p. (RT-MAE-2001-18)

2001-19 - AUBIN, E.C.Q., CORDEIRO, G.M. Bartlett adjustments for two-parameter exponential family models. 2001. 27p. (RT-MAE-2001-19)

2001-20 - HOKAMA, J., MORETTIN, P.A., BOLFARINE, H., GALEA, M. Consistent Estimation in Functional Linear Relationships with Replications. 2001. 25p. (RT-MAE-2001-20)

2001-21 - AOKI, R., SINGER, J.M., BOLFARINE, H. Local Influence in Measurement Error Regression Models with Null Intercept. 2001. 20p. (RT-MAE-2001-21)

2001-22 - ZANDONADE, E., MORETTIN, P. Wavelets in State Space Models. 2001. 31p. (RT-MAE-2001-22)

2001-23 - BUENO, V.C. A Stochastic Process Approach to Multivariante Increasing Failure Rate Distribution of Order2. 2001. 14p. (RT-MAE-2001-23)

2001-24 - BUENO, V.C. Conditional Reliability Importance of Components. 2001. 13p. (RT-MAE-2001-24)

2001-25 - BUENO, V.C. Analysing the Barbow and Proschan reliablity importance under dependence conditions. 2001. 11p. (RT-MAE-2001-25)

2001-26 - SALINAS-TORRES, V.H., PEREIRA, C.A.B., TIWARI, R.C. Bayesian Nonparametric Estimation in a Series System or a Competing-Risks Model. 2001. 14p. (RT-MAE-2001-26)

2001-27 - FONTES, L.R.G., MEDEIROS, D.P., VACHKOVSKAIA, M. Time fluctuations of the random average process with parabolic initial conditions. 2001. 25p. (RT-MAE-2001-27)

2001-28 - ESTEVES, L.G., IGLESIAS, P., WECHSLER, S. On characterizations of Uniform distributions. 2001. 15p. (RT-MAE-2001-28)

The complete list of "Relatórios do Departamento de Estatística", IME-USP, will be sent upon request.

Departamento de Estatística
IME-USP
Caixa Postal 66.281
05315-970 - São Paulo, Brasil