Distribuição Multinomial: Propriedades, Verossimilhança e Exemplos

André F. B. Menezes

Department of Statistics, Universidade Estadual de Maringá
Maringá, PR, Brazil

1 Introdução

A presente nota de aula foi elaborada tendo como principais fontes de referencias os seguintes livros e artigos: Agresti (1990), Davis and Jones (1992), Johnson et al. (1996) e Agresti (2007).

Considere uma série de n ensaios independentes, em que cada um dos quais apenas um dos k eventos mutuamente exclusivos $E_1, E_2, \ldots, E_k$ pode ser observado, e na qual a probabilidade de ocorrência do evento E_k é igual a p_k (com $p_1+p_2+\ldots+p_k = 1$). Sejam $N_1, N_2, \ldots, N_k$ variáveis aleatórias denotando o número de ocorrências dos eventos $E_1, E_2, \ldots, E_k$, respectivamente, nos n ensaios, com $\sum\limits_{i=1}^{k} N_i = n$. Então a distribuição conjunta de $N_1, N_2, \ldots, N_k$ é dada por

$$\Pr\left[N_1 = n_1, N_2 = n_2, \ldots, N_k = n_k\right] = \frac{n!}{n_1! \cdots n_k!}\, p_1^{n_1} \cdots p_k^{n_k} = n! \prod_{i=1}^{k} \left(\frac{p_i^{n_i}}{n_i!}\right), \qquad (1)$$

em que $n_1, \ldots, n_k$ são inteiros não negativos satisfazendo $\sum\limits_{i=1}^{k} n_i = n$.

A função massa de probabilidade conjunta apresentada em (1) é denominada na literatura de distribuição Multinomial, com parâmetros $(n; p_1, p_2, \ldots, p_k)$. Os parâmetros p_i, $i = 1, \ldots, k$, são chamados de probabilidades das células e n é chamado de índice.

A distribuição Binomial é um caso espacial quando existem $k = 2$ categorias/eventos. Além disso, pode-se notar que a distribuição marginal dos N_i, $i = 1, \ldots, k$, é Binomial com parâmetros (n, p_i). Portanto, a distribuição Multinomial pode ser considerada uma generalização multivariada da distribuição Binomial.

Pode-se mostrar para a distribuição Multinomial que

$$\begin{aligned} \mathbb{E}\left(N_i\right) &= n\, p_i, & \mathrm{Var}\left(N_i\right) = n\, p_i\, (1 - p_i), \\ \mathrm{Cov}\left(N_i, N_j\right) &= -n\, p_i\, p_j, & \mathrm{Corr}\left(N_i, N_j\right) = -\sqrt{\frac{p_i\, p_j}{(1-p_i)\,(1-p_j)}}, \quad i \neq j. \end{aligned}$$

A matriz de variância-covariância da distribuição Multinomial pode ser escrita como

$$V = n\left(\mathrm{diag}(\boldsymbol{p}) - \boldsymbol{p}\,\boldsymbol{p}^{\top}\right), \qquad (2)$$

em que $\boldsymbol{p} = (p_1, \ldots, p_k)$ e $\top$ denota o operador transposto.

Johnson et al. (1996) apontam que a distribuição Multinomial pode surgir em diversos contextos. Por exemplo, se $X_1, X_2, \ldots, X_k$ são variáveis aleatórias com distribuição de Poisson com média $\lambda_1, \lambda_2, \ldots, \lambda_k$, respectivamente, então a distribuição condicional conjunta de $X_1, X_2, \ldots, X_k$ dado $\sum_{i=1}^{k} X_i = n$ é Multinomial com parâmetros $(n; \pi_1, \pi_2, \ldots, \pi_k)$ em que $\pi_i = \dfrac{\lambda_i}{\lambda_1 + \lambda_2 + \cdots + \lambda_k}$, $i = 1, \ldots, k$.

2 Inferência

2.1 Abordagem Clássica

Observando uma amostra aleatória $\boldsymbol{n} = (n_1, \ldots, n_k)$ da distribuição Multinomial com $\sum_{i=1}^{n} n_i = n$, então os estimadores de máxima verossimilhança, $\widehat{\boldsymbol{p}} = (\widehat{p}_1, \widehat{p}_2, \ldots, \widehat{p}_k)$ de $\boldsymbol{p} = (p_1, p_2, \ldots, p_k)$, são obtidos maximizando a função de log-verossimilhança, cuja expressão é dada por

$$\ell = \ell(\boldsymbol{p} \mid \boldsymbol{x}) = \log n! - \sum_{i=1}^{k} \log n_i! + \sum_{i=1}^{k} n_i \log p_i. \tag{3}$$

A restrição $\sum_{i=1}^{k} p_i = 1$ pode ser incorporada utilizando os multiplicadores de Lagrange (ver, por exemplo, Bishop et al., 1975, p. 446). Alternativamente, a redundância nos parâmetros pode ser eliminada tratando ℓ como uma função de $p_1, p_2, \ldots, p_{k-1}$ como em Agresti (1990) e Davis and Jones (1992).

Diferenciando (3) com respeito aos parâmetros têm-se

$$\frac{\partial \ell}{\partial p_i} = \frac{n_i}{p_i} - \frac{n_k}{1 - \sum_{i=1}^{k-1} p_i} = \frac{n_i}{p_i} - \frac{n_k}{p_k}. \tag{4}$$

Igualando $\dfrac{\partial \ell}{\partial p_i}$ a zero obtemos

$$n_i = \frac{n_k}{p_k} p_i. \tag{5}$$

Somando ambos os lados de (5) em $i = 1, \ldots, k$, obtemos que $\widehat{p}_k = \dfrac{n_k}{n}$. De (5) temos então que

$$\widehat{p}_i = \frac{n_i}{n_k} \widehat{p}_k = \frac{n_i}{n_k} \frac{n_k}{n} = \frac{n_i}{n}, \tag{6}$$

para $i = 1, \ldots, k - 1$.

Como muito bem aponta Agresti (1990) no ano de 1900 o eminente estatístico britânico Karl Pearson introduziu um teste de hipótese que foi um dos primeiro métodos da inferência estatística. O teste de Pearson avalia se os parâmetros da distribuição são estatisticamente iguais a certos valores especificados. Como motivação Pearson buscava analisar se todos os resultados possíveis numa roda de roleta Monte Carlo eram igualmente prováveis.

Considere a hipótese nula $\mathcal{H}_0 : p_j = p_{j0},\ j = 1, \ldots, k$ em que $\sum\limits_{j=1}^{k} p_{j0} = 1$. Quando $\mathcal{H}_0$ é verdadeira, os valores esperados de $\{n_j\}$, denominadas frequências esperadas, são $\mu_j = n\, p_{j0}$, $j = 1, \ldots, k$. Pearson propôs a seguinte estatística de teste

$$\chi^2 = \sum_{j=1}^{k} \frac{(n_j - \mu_j)^2}{\mu_j}. \tag{7}$$

Para grandes amostras, χ^2 tem aproximadamente distribuição qui-quadrado com $k-1$ graus de liberdade. Assim, o valor-p pode ser aproximado por $\Pr(\chi^2_{k-1} > \chi^2_0)$, em que χ^2_{k-1} denota uma variável aleatória com distribuição qui-quadrado com $k-1$ graus de liberdade.

Um teste alternativo para os parâmetros da distribuição Multinomial é o teste da razão de verossimilhanças. Sob H_0 a verossimilhança é maximizada quando $\widehat{p}_j = p_{j0}$. No caso geral, ela é maximizada quando $\widehat{p}_j = n_j/n$. A razão de verossimilhanças é igual a

$$\Lambda = \frac{\prod\limits_{j=1}^{k} (p_{j0})^{n_j}}{\prod\limits_{i=1}^{k} (n_j/n)^{n_j}}. \tag{8}$$

Assim, a estatística da razão de verossimilhança, denotada por S_{LR}, é

$$S_{LR} = -2\, \log \Lambda = 2 \sum_{j=1}^{k} n_j \log\left(\frac{n_j}{n\, p_{j0}}\right). \tag{9}$$

A estatística (9) é denominada de *likelihood-ratio chi-squared statistic*. Grandes valores de S_{LR}, maior a evidência contra a hipótese nula.

2.2 Abordagem Bayesiana

Alternativamente, a inferência sob os parâmetros da distribuição Multinomial pode ser realizada sob o enfoque Bayesiano. Logo, devemos assumir uma distribuição a priori para os parâmetros $\boldsymbol{p}$. A distribuição a priori conjugada para $\boldsymbol{p}$ é a versão multivariada da distribuição Beta denominada de distribuição Dirichlet, cuja densidade é dada por

$$\pi(\boldsymbol{p}) = \frac{\Gamma\left(\sum\limits_{i=1}^{k} \alpha_i\right)}{\prod\limits_{i=1}^{k} \Gamma(\alpha_i)} \prod_{i=1}^{k} p_i^{\alpha_i - 1}, \qquad \alpha_i > 0, \tag{10}$$

em que $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_k)$, $p_i > 0$ com $\sum_{i=1}^{k} p_i = 1$.

O hiperparâmetro α_i pode ser interpretado como uma contagem virtual para o valor i, antes de observar o n_i. Grandes valores para α_i corresponde a um forte conhecimento a priori sobre a distribuição, ao passo que pequenos valores de α_i corresponde a ignorância. A distribuição Dirichlet tem média e variância expressas, respectivamente, por

$$\mathbb{E}\left(p_i\right) = \frac{\alpha_i}{\alpha_0} \qquad \text{e} \qquad \mathrm{Var}\left(p_i\right) = \frac{\alpha_i\left(\alpha_0 - \alpha_i\right)}{\alpha_0^2\left(\alpha_0 + 1\right)}, \tag{11}$$

em que $\alpha_0 = \sum_{i=1}^{k} \alpha_i$.

Dessa forma, a distribuição a posteriori é proporcional a

$$\pi(\boldsymbol{p} \mid \boldsymbol{n}) \propto \pi(\boldsymbol{p})\, L(\boldsymbol{p} \mid \boldsymbol{n}) \tag{12}$$

em que $\pi(\boldsymbol{p})$ foi definido em (10) e $L(\boldsymbol{p} \mid \boldsymbol{n})$ é a função de verossimilhança da Multinomial dada em (1). Substituindo as expressões pode-se mostrar que

$$\pi(\boldsymbol{p} \mid \boldsymbol{n}) = \frac{\Gamma\left(\sum_{i=1}^{k} n_i + \alpha_i\right)}{\prod_{i=1}^{k} \Gamma(n_i + \alpha_i)} \prod_{i=1}^{k} p_i^{n_i + \alpha_i - 1} \sim \mathcal{D}(n_i + \alpha_i), \tag{13}$$

isto é, a densidade a posteriori tem distribuição Dirichlet com parâmetros $n_i + \alpha_i$.

Portanto toda inferência será realizada considerando a densidade (13). Por exemplo, uma estimativa Bayesiana para os parâmetros $\boldsymbol{p}$ pode ser a esperança a posteriori, dada por

$$\mathbb{E}\left(p_i \mid n_1, \ldots, n_k\right) = \frac{n_i + \alpha_i}{n + \alpha_0}. \tag{14}$$

3 Exemplos

3.1 Dados simulados

Utilizando o software R simulamos 1000 amostras pseudos aleatórias da distribuição multinomial supondo três categorias (X_1, X_2 e X_3) com parâmetros $n = 100$ e $\boldsymbol{p} = (0.5, 0.3, 0.2)$. Abaixo exibimos a média teórica ($\boldsymbol{\mu}$), média empírica ($\widehat{\boldsymbol{\mu}}$), a matriz de correlação teórica ($\boldsymbol{\rho}$) e empírica ($\widehat{\boldsymbol{\rho}}$).

Como esperado verificamos que as estimativas empíricas são bastante próximas da teórica. Além disso, pela matriz de correlação teórica e empírica observamos que as variáveis estão negativamente correlacionadas. Isso faz sentido pois, existem $n = 100$ itens e se a frequência for maior em X_1 então haverá baixa frequência nas demais categorias.

$$\boldsymbol{\mu} = \begin{bmatrix} 50 & 30 & 20 \end{bmatrix}, \qquad \boldsymbol{\rho} = \begin{bmatrix} 1.00000 & -0.65465 & -0.50000 \\ -0.65465 & 1.00000 & -0.32733 \\ -0.50000 & -0.32733 & 1.00000 \end{bmatrix},$$

$$\widehat{\boldsymbol{\mu}} = \begin{bmatrix} 50.1040 & 29.9120 & 19.9840 \end{bmatrix}, \qquad \widehat{\boldsymbol{\rho}} = \begin{bmatrix} 1.00000 & -0.65453 & -0.46755 \\ -0.65453 & 1.00000 & -0.36228 \\ -0.46755 & -0.36228 & 1.00000 \end{bmatrix}.$$

Podemos visualizar a distribuição Multinomial plotando frequência das duas categorias X_1 e X_2, sendo que a terceira X_3 é obtida pelo complementar, isto é, 100 menos a soma das frequências observadas de X_1 e X_2. Uma vez que a distribuição Multinomial é discreta em que seus valores observados são contagens existe uma considerável sobreposição (*overploting*) dos pontos. Para resolver isso utilizamos uma estimativa da densidade Kernel. Assim, áreas com alta densidade correspondem a áreas onde há muitos pontos de sobreposição.

Na Figura 1 exibimos o gráfico da distribuição Multinomial considerando dois cenários, à esquerda para $\boldsymbol{p} = (0.5, 0.3, 0.2)$ e à direita para $\boldsymbol{p} = (0.7, 0.2, 0.1)$. Observamos que a maior parte dos pontos se encontram próximo a média populacional (50, 20) e (70, 20), respectivamente.

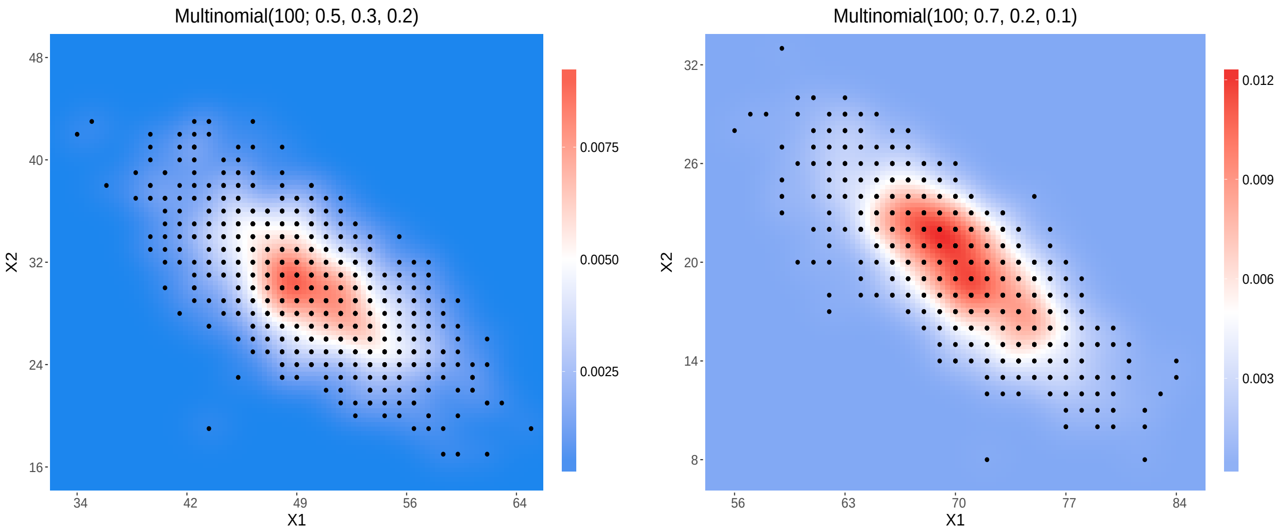

Figure 1: Diagrama de dispersão com densidade Kernel entre as categorias X_1 e X_2 da distribuição Multinomial.

3.2 Votação eleitoral

Para um exemplo simples da distribuição Multinomial consideramos as respostas de uma questão de uma pesquisa amostral com três repostas possíveis (Gelman et al., 2004). Um total de $n = 1447$ adultos foram entrevistados para indicar a sua preferência, sendo que $n_1 = 727$

apoiaram George Bush, $n_2 = 583$ apoiaram Michael Dukakis e $n_3 = 137$ apoiaram outros candidatos ou não expressaram opinião. Assumimos que as contagens n_1, n_2 e n_3 são observações da distribuição Multinomial com tamanho amostral n e respectivas probabilidades p_1, p_2 e p_3. Logo, a função de verossimilhança pode ser escrita na forma

$$L(\boldsymbol{p} \mid \boldsymbol{n}) \propto p_1^{n_1}\, p_2^{n_2}\, p_3^{n_3}. \tag{15}$$

O objetivo desta análise é comparar a proporção de votos populacional entre Bush e Dukakis. Assim a estatística mais usual para realizar a comparação é $\theta = p_1 - p_2$, isto é, a diferença da proporção populacional entre os dois principais candidatos. Podemos realizar um teste de hipótese para verificar se a proporção de votos entre Bush e Dukakis é igual ou de maneira análoga construir um intervalo de confiança/credibilidade para θ.

Inferências a respeito de θ sera realizada sob as abordagens Clássica e Bayesiana. Sob o enfoque clássico sabemos que as estimativas de máxima verossimilhança $\widehat{p}_1, \widehat{p}_2$ e $\widehat{p}_3$ de p_1, p_2 e p_3, respectivamente são dadas por (6). Os intervalos assintóticos para os parâmetros e para θ foram obtidos utilizando o erro padrão assintótico dado pela inversa da matriz de informação de Fisher observada.

Sob a abordagem Bayesiana atribuímos distribuição a priori não informativa Dirichlet, isto é, $\pi(\boldsymbol{p}) \sim \mathcal{D}(1, 1, 1)$. Portanto como discutido em (13) temos que $\pi(\boldsymbol{p} \mid \boldsymbol{n}) \sim \mathcal{D}(y_1{+}1, y_2{+}1, y_3{+}1)$, e então podemos simular valores aleatórios dessa distribuição e obter resumos a posterioris dos parâmetros de interesse. Ressalta-se que valores pseudoaleatórios da distribuição Dirichlet foram gerados baseado no fato de que se W_1, W_2 e W_3 tem distribuição Gamma$(\alpha_i, 1), i = 1, 2, 3$ e $T = W_1{+}W_2{+}W_3$, então a distribuição das proporções $(W_1/T, W_2/T, W_3/T)$ tem distribuição $\mathcal{D}(\alpha_1, \alpha_2, \alpha_3)$.

A *procedure* **NLMIXED** do software SAS foi utilizada para os cálculos computacionais da inferência clássica. Enquanto que sob o enfoque Bayesiano uma função foi implementada em R para simular valores aleatórias da distribuição Dirichlet.

Na Tabela 1 apresenta-se as estimativas dos parâmetros sob o enfoque Clássico e Bayesiano. Os parâmetros p_1, p_2 e p_3 fornecem estimativas para a proporção populacional de cada candidato. Por exemplo, esperasse que George Bush obtenha aproximadamente 50% dos votos, analisando os intervalos verifica-se que a proporção de votos de Bush na população varia entre aproximadamente 47% e 53%, portanto Bush seria o vencedor das eleições. Além disso, observando as estimativas para o parâmetro θ conclui-se que existe diferença entre a proporção de votos de Bush e Duakakis, em outras palavras os candidatos não estão tecnicamente empatados, isto é, Bush tem maior apoio do que Dukakis na população da pesquisa.

Table 1: Resumos inferências sob o enfoque Clássico e Bayesiano.

Parâmetro	EMV (E.P.)	Média (D.P.)	I.C. 95%	H.P.D. 95%
p_1	0.5024 (0.0131)	0.5020 (0.0130)	(0.4767, 0.5282)	(0.4763, 0.5264)
p_2	0.4029 (0.0129)	0.4029 (0.0127)	(0.3776, 0.4282)	(0.3795, 0.4305)
p_3	0.0947 (0.0079)	0.0951 (0.0077)	(0.0796, 0.1098)	(0.0801, 0.1100)
θ	0.0995 (0.0248)	0.0991 (0.0245)	(0.0508, 0.1483)	(0.0497, 0.1456)

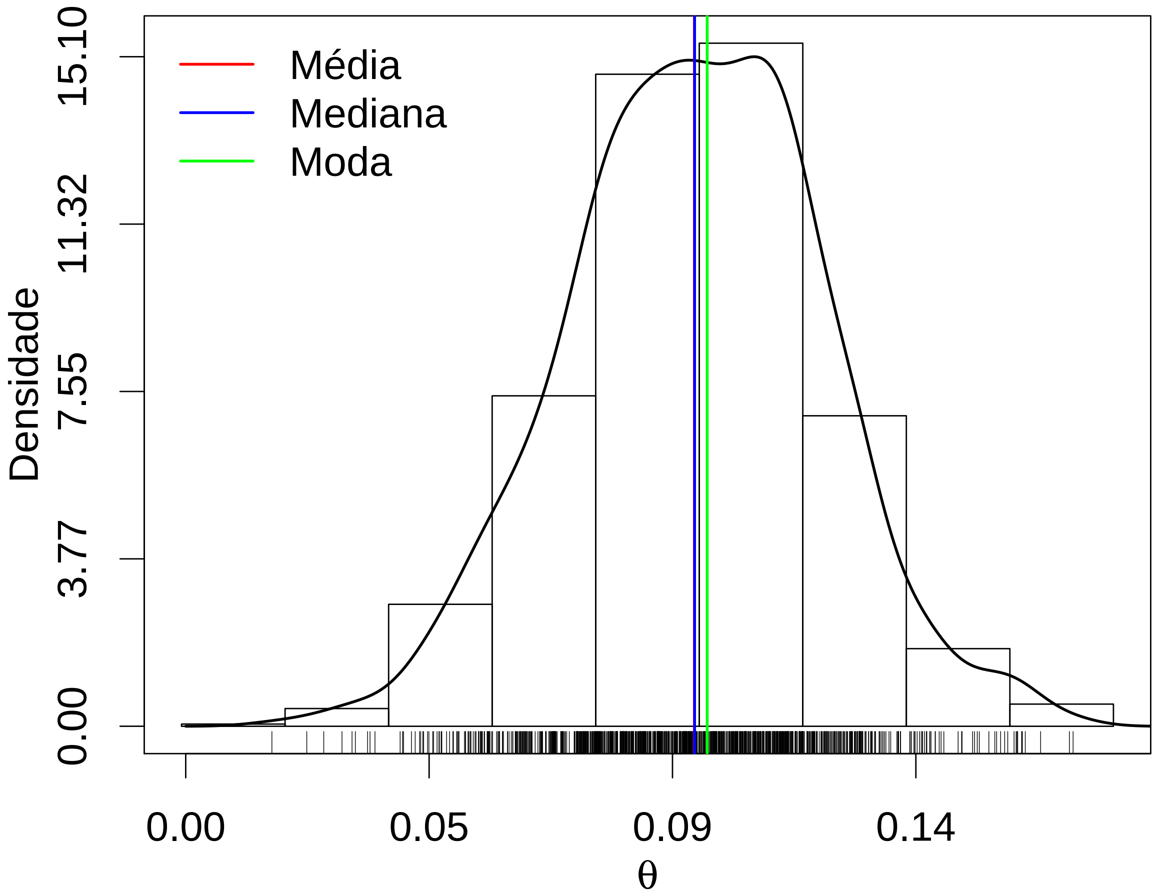

Figure 2: Densidade a posteriori de $\theta = p_1 - p_2$.

Curiosamente, os resultados da eleição de 1988 evidenciam que a inferência realizada acertou no resultado da eleição, sendo que Bush alcançou 53.37% e Dukakis 45.65% dos votos.

References

Agresti, A., 1990. Categorical Data Analysis. Wiley, New York.

Agresti, A., 2007. An Introduction to Categorical Data Analysis. John Wiley & Sons, Inc., New Jersey.

Bishop, Y. M. M., Fienberg, S. E., Holland, P. W., 1975. Discrete Multivariate Analysis: Theory and Practice. Cambridge, MA: The MIT Press.

Davis, C. S., Jones, M. P., 1992. Maximum likelihood estimation: for the Multinomial distribution. Teaching Statistics 14 (3), 9–11.

Gelman, A., Carlin, J. B., Stern, H. S., Rubin, D. B., 2004. Bayesian Data Analysis. Chapman & Hall/CRC, New York.

Johnson, N. L., Kotz, S., Balakrishnan, N., 1996. Discrete multivariate distributions. Wiley, New York.