SENSIBILIDADE DA DISTRIBUIÇÃO A POSTERIORI EM RELAÇÃO AS DIFERENTES DISTRIBUIÇÕES A PRIORI PARA A PROBABILIDADE DE SUCESSO DE UMA DISTRIBUIÇÃO BINOMIAL CORRELACIONADA

Marcelo Hiroshi Tutia
Universidade Federal de São Carlos - UFSCar
Faculdade Estácio de Sá de Ourinhos - FAESO

Carlos Alberto Ribeiro Diniz
Universidade Federal de São Carlos - UFSCar

José Galvão Leite
Universidade Federal de São Carlos - UFSCar

Resumo. É comum o levantamento de dados onde a variância observada é maior que a variância esperada sob um modelo adotado. Observações com essa propriedade são denominadas superdispersas ou que apresentam superdispersão. Para modelar essas observações devemos considerar modelos que levem em consideração esse acréscimo de variabilidade. Neste trabalho apresentamos a distribuição binomial correlacionada, utilizada, entre outras, para modelar o fenômeno da superdispersão. Adotamos uma abordagem bayesiana e analisamos a sensibilidade da distribuição a posteriori em relação a diferentes distribuições a priori para a probabilidade de sucesso. Concluímos que nas situações onde a informação a priori não esteja disponibilizada, qualquer priori de referência apresentada neste trabalho pode ser usada sem prejuízo para os resultados a posteriori.
Palavras-chave: Inferência bayesiana, Distribuição binomial correlacionada.

Abstract. It is usual the occurrence of data where the observed variance is larger than the expected variance under a certain specific model. This phenomenon is denominated overdispersion. In modeling of these observations we should consider models that take in consideration this variability increment. In this work we presented the correlated binomial distribution proposed by Luceño (1995), used, among other, to model the phenomenon of the overdispersion. We adopted an Bayesian approach and we analyzed the sensibility of the posterior distribution, in relation to different prior distributions, for the probability of success p.
Keywords: Bayesian Inference, Correlated binomial distribution.

Recebido em Março 2004; revisão recebida em Julho 2004; aceito em Setembro 2004.
RETEC – Revista de Tecnologias, Ourinhos, v. 1, n. 1, p. 49-52, julho-dezembro. 2004

1. INTRODUÇÃO

Dados com grande variabilidade, não permitindo que os modelos usuais expliquem de maneira adequada toda a variabilidade observada, são denominados superdispersos ou que apresentam superdispersão. Nestes casos a variância observada é maior que a variância esperada sob o modelo adotado. A superdispersão não é incomum na prática. McCullagh e Nelder [6] afirmam que superdispersão é uma regra e a dispersão nominal a exceção. São muitas e diferentes as possíveis causas para a superdispersão. Algumas possibilidades são: variabilidade do material, correlação entre respostas individuais, amostragem por conglomerado, outliers e agrupamento de dados.

A literatura apresenta diferentes explicações para o processo de superdispersão, mas, em geral, é difícil concluir sobre a causa exata ou o processo que produz a superdispersão. Vários autores propuseram diferentes modelos para ajustar o fenômeno da superdispersão, destacando-se entre eles Kupper e Haseman [3], Williams [8], Luceño [4], Luceño e Ceballos [5] e Hinde e Demétrio [2].

O modelo apresentado por Luceño [4] estuda a variação extra-binomial considerando Y = W1 +...+ Wn, Bernoulli modificada *"0 e n"* Fu e Sproule [1], *Mbern(p)*, presente com probabilidade ρ. Desta forma, a função de distribuição de probabilidade de Y dado n, p, ρ pode ser escrita como

$$P(Y=y\mid n,p,\rho)=\binom{n}{y}p^{y}(1-p)^{n-y}(1-\rho)I_{A1}(y)+p^{\frac{y}{n}}(1-p)^{\frac{(n-y)}{n}}\rho I_{A2}(y),$$

onde A1 = {0, 1, ..., n}, A2 = {0, n} e y = 0, 1, ..., n.

As estruturas de média e variância são dadas por $E(Y) = np$ e $Var(Y) = p(1-p)\{n+\rho n(n-1)\}$.

Observamos que a variância apresenta um termo a mais do que a variância do modelo binomial ordinário. A variação extra-binomial tem origem no coeficiente de correlação ρ entre as variáveis de Bernoulli W_r, $r = 1, ..., n$. Para $\rho = 0$ o modelo $BC(n,p,\rho)$ equivale ao modelo $B(n,p)$.

2.1. Função de verossimilhança

Dada uma amostra aleatória $Y_1, Y_2, ..., Y_k$, da distribuição $BC(n,p,\rho)$, o logaritmo da função de verossimilhança para o modelo binomial correlacionado é dado pela equação (1).

$$l(p,\rho\mid n,y_1,y_2,\ldots,y_k)=\sum_{i=1}^{k}\left(\log\left\{\binom{n}{y_i}p^{y_i}(1-p)^{n-y_i}(1-\rho)I_{A1}(y_i)+p^{\frac{y_i}{n}}(1-p)^{\frac{(n-y_i)}{n}}\rho I_{A2}(y_i)\right\}\right) \quad (1)$$

onde

Wr, r = 1, ..., n, é uma variável de Bernoulli e ρ = Corr(Wr,Ws), $r \neq s$, $0 \leq \rho \leq 1$, é o coeficiente de correlação existente entre as variáveis de Bernoulli. O modelo proposto assume implicitamente que qualquer um dos n indivíduos incluídos na amostra pertencem a um exclusivo bloco equicorrelacionado.

Neste trabalho estudamos o modelo binomial correlacionado, BC(n,p,ρ), apresentado por Luceño [4] sob o enfoque Bayesiano e verificamos a sensibilidade a posteriori em relação a diferentes distribuições a priori utilizadas para a probabilidade de sucesso p.

2. MODELO BINOMIAL CORRELACIONADO: BC(N,P,ρ)

Luceño [4] apresenta um modelo binomial correlacionado, $BC(n,p,\rho)$, que pode ser utilizado para modelar o fenômeno da superdispersão. O modelo generaliza a distribuição binomial ordinária, $B(n,p)$. Seja $Y = W_1 +...+ W_n$, onde W_r, $r = 1, ..., n$, é uma variável de Bernoulli, com $E(W_r) = p$, $Var(W_r) = p(1-p)$ e $Corr(W_r,W_s) = \rho$, $0 \leq \rho \leq 1$. A distribuição de probabilidade de Y é obtida da combinação de duas variáveis aleatórias: uma variável binomial, $B(n,p)$, presente com probabilidade $(1-\rho)$ e uma variável de

$$\pi(p\mid n,\rho,\alpha,\beta,\mathbf{y})\propto\prod_{i=1}^{k}\left\{\binom{n}{y_i}p^{y_i+\frac{\alpha-1}{k}}(1-p)^{(n-y_i)+\frac{\beta-1}{k}}(1-\rho)I_{A1}(y_i)+p^{\frac{y_i}{n}+\frac{\alpha-1}{k}}(1-p)^{\frac{(n-y_i)}{n}+\frac{\beta-1}{k}}\rho I_{A2}(y_i)\right\} \quad (2)$$

3. INFERÊNCIA BAYESIANA

Uma análise Bayesiana do modelo binomial correlacionado é feita considerando n e ρ fixos e assumindo uma distribuição a priori $Beta(\alpha,\beta)$ para p, com hiperparâmetros α e β conhecidos, ou seja,

$$p \sim Beta(\alpha,\beta),\ \alpha,\beta \geq 0.$$

A notação $Beta(\alpha,\beta)$ será usada também para denotar distribuições a priori de referência (ver Smith, 1991), tais como $Beta(0,0)$, $Beta(0,1)$, $Beta(½, ½)$ e $Beta(1,1)$.

Combinando a função de verossimilhança, dada em (1), com a função densidade a priori para p encontramos a seguinte distribuição a posteriori para p dados n, ρ, α, β e $\mathbf{y}$, onde $\mathbf{y} = (y_1, ..., y_n)$ é o vetor de observações,

Não é difícil mostrar que se $\pi(p)$ é uma das quatro distribuições a priori de referência dadas anteriormente, a posteriori $\pi(p|n,\rho,\alpha,\beta,\mathbf{y})$ existe. A esperança e variância a posteriori de p são dadas, respectivamente, por

$$E(p\mid n,\rho,\alpha,\beta,\mathbf{y})=\int_0^1 p\pi(p\mid n,\rho,\alpha,\beta,\mathbf{y})dp; \quad (3)$$

$$Var(p \mid n,\rho,\alpha,\beta,\mathbf{y}) = E(p^2 \mid n,\rho,\alpha,\beta,\mathbf{y}) - (E(p \mid n,\rho,\alpha,\beta,\mathbf{y}))^2 \quad (4)$$

e a mediana m a posteriori de p é calculada da forma

$$\int_0^m \pi(p \mid n,\rho,\alpha,\beta,\mathbf{y})dp = \frac{1}{2}. \quad (5)$$

3.1. Sensibilidade da distribuição a posteriori

Nesta seção estudamos a sensibilidade da distribuição a posteriori para p, dada pela expressão (2), com relação à quatro distribuições a priori não informativas e cinco distribuições informativas. As características a posteriori são calculadas de forma exata através de (3), (4) e (5). Os pares de valores para os hiperparâmetros α e β da distribuição a priori Beta são $(\alpha, \beta) = (0, 0), (0, 1), (½, ½)$ *e* $(1, 1)$, determinando distribuições a priori não informativas ou prioris de referência, e os pares $(5, 15), (15, 5), (10, 10), (2, 100)$ e $(100, 1)$ determinando distribuições a priori informativas.

Na Tabela 1 são apresentadas as características a posteriori exatas de p considerando uma amostra aleatória y_i, $i = 1, ..., 30$, seguindo distribuição $BC(20; 0,5; 0,8)$.

Tabela 1 – Características a posteriori de p considerando diferentes distribuições a priori, considerando uma amostra aleatória seguindo distribuição BC(20; 0,5; 0,8).

Priori	$E(p\|n,\rho,\alpha,\beta,\mathbf{y})$	$Var(p\|n,\rho,\alpha,\beta,\mathbf{y})$	*Mediana*
Beta(0,0)	0,5000	0,0017	0,5000
Beta(0,1)	0,4966	0,0017	0,4965
Beta(½, ½)	0,5000	0,0017	0,5000
Beta(1,1)	0,5000	0,0017	0,5000
Beta(5,15)	0,4695	0,0015	0,4694
Beta(15,5)	0,5305	0,0015	0,5306
Beta(10,10)	0,5000	0,0015	0,5000
Beta(2,100)	0,3007	0,0009	0,3002
Beta(100,1)	0,7022	0,0009	0,7027

Para uma amostra aleatória seguindo $BC(20; 0,5; 0,8)$ as características a posteriori, considerando as quatro prioris não informativas, apresentam pequena variabilidade nos resultados em relação ao verdadeiro valor do parâmetro, indicando insensibilidade a posteriori em relação a estas prioris.

Apesar de as distribuições a priori *Beta(5, 15)* (esperança = 0,25 e variância = 0,0089) e *Beta(15,5)* (esperança = 0,75 e variância = 0,0089) serem informativas, as características a posteriori também apresentam pequena variabilidade em relação ao verdadeiro valor do parâmetro. Percebemos, nestes casos, uma predominância da verossimilhança.

A distribuição a priori *Beta(10,10)* (esperança = 0,50 e variância = 0,0119) possui melhor resultado a posteriori, pois apresenta demasiada informação a priori.

As informações fornecidas pelas distribuições a priori *Beta(2,100)* (esperança = 0,0196 e variância = 0,0002), *Beta(100,1)* (esperança = 0,9901 e variância = 0,0001) são extremamente informativas, com variâncias muito pequenas. Nestes casos as esperanças a posteriori estão distantes do verdadeiro valor de p e mesmo assim notamos a influência da verossimilhança nos resultados.

Na Tabela 2 são apresentadas as características a posteriori exatas de p considerando uma amostra aleatória de y_i, $i = 1, ..., 30$, seguindo distribuição $BC(20; 0,1; 0,8)$.

Tabela 2 – Características a posteriori de p considerando diferentes distribuições a priori, considerando uma amostra aleatória seguindo distribuição BC(20; 0,1; 0,8).

Priori	$E(p\|n,\rho,\alpha,\beta,\mathbf{y})$	$Var(p\|n,\rho,\alpha,\beta,\mathbf{y})$	*Mediana*
Beta(0,0)	0,1196	0,0006	0,1183
Beta(0,1)	0,1190	0,0005	0,1177
Beta(½, ½)	0,1216	0,0006	0,1203
Beta(1,1)	0,1236	0,0006	0,1223
Beta(5,15)	0,1321	0,0005	0,1309
Beta(15,5)	0,1792	0,0007	0,1782
Beta(10,10)	0,1557	0,0006	0,1547
Beta(2,100)	0,0847	0,0003	0,0837
Beta(100,1)	0,4139	0,0008	0,4137

Neste caso, as características a posteriori considerando distribuições a priori não informativas também apresentam resultados próximos do verdadeiro valor do parâmetro. Quando consideramos priori informativas novamente percebemos a predominância da verossimilhança nas características a posteriori.

Vários outros cenários foram executados, tais como $BC(20; 0,5; 0,5)$, $BC(20; 0,5; 0,2)$ e $BC(20; 0,9; 0,2)$. Todos os resultados encontrados foram similares aos apresentados.

4. CONCLUSÃO

Para distribuições a priori não informativas, ou prioris de referência, as características a posteriori para p, considerando diferentes amostras aleatórias de y_i, $i = 1, ..., 30$, seguindo distribuições $BC(n,p,\rho)$, apresentam pequena variabilidade em relação ao verdadeiro valor do parâmetro. Para algumas das distribuições a priori informativas as características a posteriori, considerando certas amostras aleatórias de y_i, apresentam uma pequena variabilidade em relação ao verdadeiro valor de p. Nos casos onde atribuímos distribuições a priori informativas notamos a predominância da verossimilhança nas características a posteriori, fazendo com que os resultados estejam próximos do verdadeiro valor do parâmetro. Em algumas situações os resultados estão distantes do verdadeiro valor de p, mas mesmo assim notamos a influência da verossimilhança. Assim, concluímos que nas situações onde a informação a priori não esteja disponibilizada, qualquer priori de referência apresentada neste trabalho pode ser usada sem prejuízo para os resultados a posteriori.

REFERÊNCIAS

[1] FU, J., SPROULE, R. A. A generalizations of the binomial distribution, *Communications in Statistics: Theory e Methods*, v. 24, n. 10, p. 2645-2658, 1995.

[2] HINDE, J., DEMÉTRIO, C. G. B. Overdispersion models and estimation, *Computational Statistics & Data Analysis*, v. 27, p. 151-170, 1998.
[3] KUPPER, L. L., HASEMAN, J. K. The use of a correlated binomial model for the analysis of certain toxicological experiments, *Biometrics*, v. 34, p. 67-76, 1978.
[4] LUCEÑO, A. A family of partially correlated Poisson models for overdispersion, *Computational Statistics & Data Analysis*, v. 20, p. 511-520, 1995.
[5] LUCEÑO, A., CEBALLOS, F. Describing extra-binomial variation with partially correlated models, *Communications in Statistics: Theory and Methods*, v. 24, n. 6, p. 1637-1653, 1995.
[6] MCCULLAGH, P., NELDER, J. A. *Generalized Linear Model*, 2 ed., London: Chapman & Hall, 1989.
[7] SMITH, P. J. Bayesian analyses for a multiple capture-recapture model, *Biometrika*, v. 78, n. 2, p. 399-407, 1991.
[8] WILLIAMS, D. A. Extra-binomial variation in logistic linear models, *Journal of the Royal Statistical Society Series C*, v. 31, n. 2, pl 144-148, 1982.