OS TESTES NÃO-PARAMÉTRICOS G E QUI-QUADRADO E SUAS APROXIMAÇÕES

Marcos A. Antonello Scremin

Curso de Pós-graduação em Métodos Quantitativos. Departamento de Estatística. Centro de Ciências Naturais e Exatas. UFSM. Santa Maria, RS.

Anaelena B. de Moraes Ethur e Maria Emília Camargo

Departamento de Estatística. Centro de Ciências Naturais e Exatas. UFSM. Santa Maria, RS.

RESUMO

Neste trabalho é apresentada a procedência e a aproximação de um teste não-paramétrico alternativo ao teste do Qui-quadrado (X^2), conhecido na literatura como teste G, mas pouco divulgado. Foi verificada a origem desta estatística através de deduções e também a aproximação à estatística do Qui-quadrado, utilizada nos testes de hipóteses.

SUMMARY

SCREMIN, M. A. A.; ETHUR, A. B. de M. and CAMARGO, M. E., The no-parametries tests g and chi-square and their aproximations. Ciência e Natura, 13:55-59, 1991.

In this paper we show from where became and the aproach of a alternative to the few published no-parametric test, knowed chis-quare in the literature as G test. By deduction, we research the origin of this statistics and the aproach of the statistics of chi-square that is used at the hypothesis tests.

1 - ESTATÍSTICA G

A Estatística G é definida como sendo duas vezes o logarítmo natural da razão entre a probabilidade da amostra com todos os parâmetros estimados a partir dos dados e a probabilidade da amostra, admitindo-se que a hipótese nula é verdadeira. Esta definição origina-se do teste de razão de verossimilhança, como segue:

Considerando-se k categorias com probabilidades de ocorrência $p_1, p_2, \ldots, p_k$. com $\Sigma p_i = 1$ para i = 1, 2, ..., k. Seja $x_1, x_2, \ldots, x_k$, as frequências observadas naquelas k categorias em n tentativas, com $n = \Sigma x_i$, para i = 1, 2, ..., k. Então a função de verossimilhança é:

$$L(p) = (p_1)^{x_1} . (p_2)^{x_2} . (p_3)^{x_3} \ldots (p_k)^{x_k} \qquad (1)$$

Para maximizar L(p), por método de cálculo, deve-se considerar que existe k-1 parâmetros independentes, pois $\Sigma p = 1$.

Assim é conveniente escolher p_k como parâmetro a ser expresso em termos dos parâmetros remanescentes.

Aplicando logarítmos naturais a ambos os membros da Eq. (1), obtém-se:

$$\ln L(p) = x_1 . \ln(p_1) + x_2 . \ln(p_2) + \ldots + x_k . \ln(p_k) \qquad (2)$$

Derivando-se a Eq. (2) em relação a p , obtém-se:

$$\frac{\ln L(p)}{p_i} = \frac{x_1}{p_1} + \frac{x_2}{p_2} + \frac{x_3}{p_3} + \ldots + \frac{x_k}{p_k} = \frac{x_i}{p_i}$$

Como só existem k - 1 parâmetros independentes, deve-se eliminar o termo $\frac{x_k}{p_k}$. Então:

$$\frac{\ln L(p)}{p_i} = \frac{x_i}{p_i} - \frac{x_k}{p_k} \qquad (3)$$

Para maximizar é necessário que $\frac{\ln L(p)}{p_i}$ seja igual a zero, ou seja, todas as k - 1 derivadas parciais desapareçam. Então:

$$\frac{x_i}{p_i} - \frac{x_k}{p_k} = \Phi \text{ , para } i = 1, 2, \ldots, k-1$$

Logo:

$$p_i = \frac{p_k}{x_k} \cdot x_i \text{ , para } i = 1, 2, \ldots, k \qquad (4)$$

Aplicando-se o somatório a ambos os membros da Eq. (4) e considerando que $\Sigma\, p = 1$, tem-se:

$$1 = \Sigma \frac{p_k}{x_k} \cdot x_i \longleftrightarrow 1 = \frac{p_k}{x_k} \Sigma\, x_i$$

Como $\Sigma\, x_i = n \longleftrightarrow 1 = \frac{p_k}{x_k} \cdot n$, então:

$$\frac{p_k}{x_k} = \frac{1}{n} \qquad (5)$$

Substituindo-se a Eq. (5) na Eq. (4), obtém-se:

$$p_i = \frac{1}{n} \cdot x_i \longleftrightarrow p_i = \frac{x_i}{n} \text{ , } i = 1, 2, \ldots, k \qquad (6)$$

Considerando o teste da razão de verossimilhança para testar a hipótese:

$$H_0 : p_i = p_{i0},\ i = 1, 2, \ldots k$$

Como não existe nenhum parâmetro não específico remanescente quando H_0 é verdadeiro, segue-se que a razão de verossimilhança é dada por:

$$\lambda = \frac{L(\hat{p})}{L(p)} = \frac{(\hat{p}_1)^{x_1} \cdot (\hat{p}_2)^{x_2} \cdot \ldots \cdot (\hat{p}_k)^{x_k}}{(p_1)^{x_1} \cdot (p_2)^{x_2} \cdot \ldots \cdot (p_k)^{x_k}} \qquad (7)$$

pode-se observar que a Eq. (7) pode ser expressa como sendo a razão de duas distribuições multinomiais, então:

$$\lambda = \frac{\frac{n!}{x_1!\, x_2! \ldots x_k!} (\hat{p}_1)^{x_1} \cdot (\hat{p}_2)^{x_2} \cdot \ldots \cdot (\hat{p}_k)^{x_k}}{\frac{n!}{x_1!\, x_2! \ldots x_k!} (p_1)^{x_1} \cdot (p_2)^{x_2} \cdot \ldots \cdot (p_k)^{x_k}}$$

De (8), tem-se:

$$\lambda = \prod_{i=1}^{k} \left[\frac{\hat{p}_i}{p_i} \right]^{x_i} \qquad (9)$$

Se as proporções p e $\hat{p}$ forem iguais, então a razão será igual a 1. Quanto maior a diferença entre p e $\hat{p}$, com maior intensidade λ será diferente de 1. Isto indica que a razão dessas probabilidades pode ser usada como estatística para medir a conformidade entre as freqüências observadas e esperadas. Um teste baseado em tal razão é denominado de teste de razão de probabilidade.

Sendo a freqüência esperada $e_i = n.\hat{p}_i$ e a freqüência observada $x_i = n.p_i$, então:

$$\hat{p}_i = \frac{e_i}{n} \qquad e \qquad p_i = \frac{x_i}{n}$$

substituindo estes resultados na eq. (9), tem-se:

$$\lambda = \prod_{i=1}^{k} \left[\frac{e_i}{x_i} \right]^{x_i} \tag{10}$$

Como a distribuição teórica dessa razão é complexa e pouco conhecida, considera-se o seguinte teorema "Sob certas condições de regularidade, a variável aleatória $-2.\ell n\ \lambda$, onde λ é dada por (7), tem uma distribuição que se aproxima da variável X^2, quando n é finito, com graus de liberdade iguais ao número de parâmetros que são determinados pela hipótese H_0"[8] , ou seja, os graus de liberdade para o teste são iguais ao do teste de qui-quadrado.

$$-2\ \ell n\ \lambda = -2. \sum_{i=1}^{k} x_i\ \ell n \left[\frac{e_i}{x_i} \right] = 2. \sum_{i=1}^{k} x_i\ \ell n \left[\frac{e_i}{x_i} \right]^{-1}$$

Logo:

$$G = 2. \sum_{i=1}^{k} x_i\ \ell n \left[\frac{x_i}{e_i} \right] \tag{11}$$

Esta estatística é pouco conhecida e utilizada devido a ser mais trabalhoso aplicar um teste utilizando logarítmos, mas com o uso do computador torna-se mais fácil sua aplicação.

Em geral, G será numericamente similar a X^2, sendo que o emprego do teste G requer nenhuma sofisticação matemática especial e tem algumas vantagens consideráveis em projetos mais complexos sobre os testes convencionais que utilizam a estatística X^2, pois permite certos tipos de análises mais detalhadas que são possíveis usando-se a estatística de qui-quadrado [13].

Mesmo com o uso de logaritmos para o teste G, este muitas vezes requer muito menos esforço computacional do que é requerido para o teste de X^2, isto é verdadeiro para alguns testes [13].

Portanto, pode-se empregar o teste G, sempre que puder ser utilizado um computador, pois os resultados numéricos entre G e X^2 podem diferir em alguns casos, sugerindo uma análise mais cuidadosa dos dados coletados [13].

2- ORIGEM DA ESTATÍSTICA QUI-QUADRADO A PARTIR DA ESTATÍSTICA G

Em 1960, Karl Pearson propôs o seguinte teste estatístico, o qual é funçáo dos quadrados dos desvios das freqüências observadas, com relação aos valores esperados respectivos.

$$X^2 = \sum_{i=1}^{k} \frac{(x_i - e_i)^2}{e_i} \tag{12}$$

Pode-se demonstrar a origem da estatística X^2 a partir da estatística G, da seguinte maneira:

Considerando-se a Estatísitca G, a Eq. (11), e fazendo-se $y_i = x_i - e_i$, a diferença entre as freqüências observadas e esperadas na i-ésima cela, tem-se:

$$G = 2 \Sigma (e_i + y_i) \ln \frac{e_i + y_i}{e_i} = 2 \Sigma (e_i + y_i) \ln \left[1 + \frac{y_i}{e_i} \right]$$

Expandindo $\ln \left[1 + \frac{y_i}{e_i} \right]$ em Série de Taylor em torno de $x = \Phi$ obtém-se:

$$G = 2 \Sigma (y_i + e_i) \cdot \left[\frac{y_i}{e_i} - \frac{1}{2} \left[\frac{y_i}{e_i} \right]^2 + \frac{1}{3} \left[\frac{y_i}{e_i} \right]^3 - \ldots \right]$$

Efetuando, resulta:

$$G = \Sigma \frac{y_i^2}{e_i} - \frac{1}{3} \Sigma \frac{y_i^3}{e_i} + \ldots\ldots\ldots \qquad (13)$$

A variável y é uma variável binomial com média $\mu_i = n.p_i = e_i$ e variância $\sigma_i^2 = np_i (1-p)$; conseqüentemente, a variável $\frac{y_i}{e_i}$ pode ser expressa na forma:

$$\frac{y_i}{e_i} = \frac{x_i - e_i}{e_i} = \frac{x_i - \mu_i}{\sigma_i} \sqrt{(1 - p_i)/np_i} \qquad (14)$$

Do Teorema "Se X é normalmente distribuída com média e variância σ^2 e uma amostra aleatória de tamanho n é retirada, então a média amostral $\bar{X}$ será normalmente distribuída com média μ e variâcia $\frac{\sigma^2}{n}$ "[8], a variável $(x_i - \mu_i)/\sigma_i$ tem distribuição aproximada à de uma variável normal padrão quando $n \longrightarrow \infty$. Sendo assim o fator da raiz quadrada da Eq. (14) se aproxima de zero na mesma taxa que $1/\sqrt{n}$. Então, para n muito grande y_i/e_i será muito pequeno e da ordem de $1/\sqrt{n}$. Como consequência os termos sucessivos na expansão anterior serão da ordem de $1/\sqrt{n}$ vezes o termos precedente.

O resultado aproximado de $-2 \ln \lambda$ para amostras grandes é dado pelo 1º termos da direita da Eq. (13). Então:

$$-2 \ln \lambda \tilde{=} \Sigma \frac{y_i^2}{e_i} = \sum_{i=1}^{k} \frac{(x_i - e_i)^2}{e_i} \qquad (15)$$

Como $-2 \ln \lambda$ possui distribuição X^2 aproximada, então o termo da direita da Eq. (15) também possui distribuição X^2.

3- CONCLUSÕES:

Através da literatura citada chegou-se as seguintes conclusões:

- Em projetos mais complexos a aplicação da estatística G tem algumas vantagens em relação a aplicação da estatística X^2, pois possibilita uma análise mais detalhada dos dados;
- A estatística G é pouco utilizada devido, principalmente, a falta de informações sobre o assunto causada pela escassa li

teratura e também pela maior dificuldade de cálculos pelo emprego de logarítmos, o que no momento não é mais problema com o uso do computador.

4- REFERÊNCIAS BIBLIOGRÁFICAS

01. AYRES, M. & AYRES JUNIOR, M Aplicações Estatísticas em Basic. São Paulo, Mc Graw-Hill, 1987.
02. BISHOP, Y. M. et alii. Discrete Multivariate Analysis: Theory and Practice. The mit Press, 1975.
03. CHRISTMANN, Raul U. Estatística Aplicada. 2ª ed. São Paulo, Edgard Blücher Ltda, 1978.
04. CONOVER, W. J. Practical Nonparametric Statistic. New York, John Wiley & Sons, 1971.
05. COSTA NETO, Pedro L. de O. Estatística. São Paulo, Edgard Blücher Ltda, 1977.
06. FONSECA, Jairo S. da, MARTINS, Gilberto de A. Curso de Estatística. 3ª ed. São Paulo, Atlas, 1982.
07. GOODMAN, Richard. Estatística. São Paulo, Livraria Editora Pioneira, 1965.
08. HOEL, Paul G. Estatística Matemática. 4ª ed. Rio de Janeiro, Guanabara Dois S.A., 1980.
09. MENDENHALL, William. Probabilidade e Estatística. Rio de Janeiro, Campus, 1985, v. 2.
10. RODRIGUES, Aroldo. A Pesquisa Experimental em Psicologia e Educação. 2ª ed. Rio de Janeiro, Vozes, 1976.
11. SCREMIN, Marcos A. A. Comparação Entre os Testes G e Qui-Quadrado Através de um Programa Computacional. Monografia do Curso de Pós-Graduação em Métodos Quantitativos, Universidade Federal de Santa Maria, Santa Maria-RS, 1989.
12. SIEGEL, Sidney. Estatística Não-paramétrica. São Paulo, McGraw-Hill do Brasil, Ltda, 1975.
13. SOKAL, R. and ROHLF, F. J. BIOMETRY. San Francisco, Freeman and Company, 1969.
14. STEVENSON, Willian J. Estatística Aplicada à Administração. São Paulo, Harper & Row do Brasil Ltda, 1981.
15. YAMANE, Taro. Estatística. 3ª ed. México, Harla, 1974.

Recebido em outubro, 1990; aceito em abril, 1991.