Teste de Gaussianidade χ^2

Leonardo Tôrres

30 de agosto de 2005

Resumo

Breve revisão do teste χ^2 usado para determinar se a função densidade de probabilidade associada a variável observada de um instrumento é Gaussiana ou não, com um certo nível de significância (chance de cometer um equívoco).

1 O problema

Precisamos descobrir se os diversos valores observados em um instrumento de medição estão distribuídos segundo uma função de densidade de probabilidade Gaussiana, para o caso em que a variável desejada é mantida fixa. Isto é, desejamos saber se a função de densidade de probabilidade associada a x_{obs} pode ser escrita como

$$f(x_{\text{obs}}) = \frac{1}{s_{\text{obs}}\sqrt{2\pi}}\, e^{\frac{-(x_{\text{obs}}-\bar{x}_{\text{obs}})^2}{2s^2_{\text{obs}}}}, \tag{1}$$

sendo $\bar{x}_{\text{obs}}$ a média amostral dos valores observados e s_{obs} o desvio padrão amostral dos valores observados:

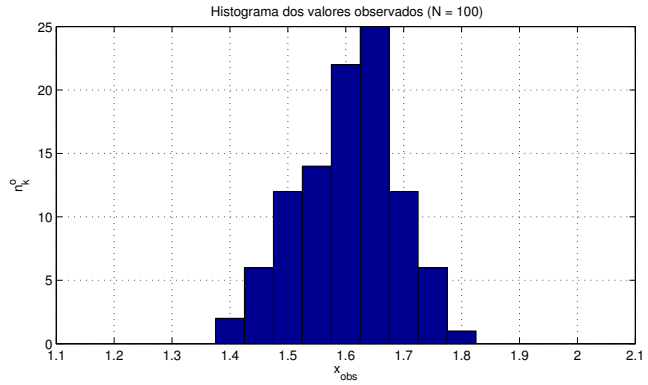
$$\begin{aligned} \bar{x}_{\text{obs}} &= \frac{1}{N}\sum_{k=1} N x_{\text{obs}}(k) \\ s_{\text{obs}} &= \frac{1}{N-1}\sum \left[x_{\text{obs}}(k) - \bar{x}_{\text{obs}}\right]^2 \end{aligned}$$

2 O Histograma

Suponha que foram observados $N = 100$ valores de x_{obs} para um dado instrumento, mantendo-se a variável desejada constante. Fazendo-se um histograma para visualizar a distribuição de valores, obtém-se a Fig.1a.

A partir dos valores de $\bar{x}_{\text{obs}}$ e s_{obs} calculados dos dados observados, e conhecendo-se a expressão (1), obtém-se a curva vermelha mostrada na Fig.1b, que corresponde ao número

(a)



(b)

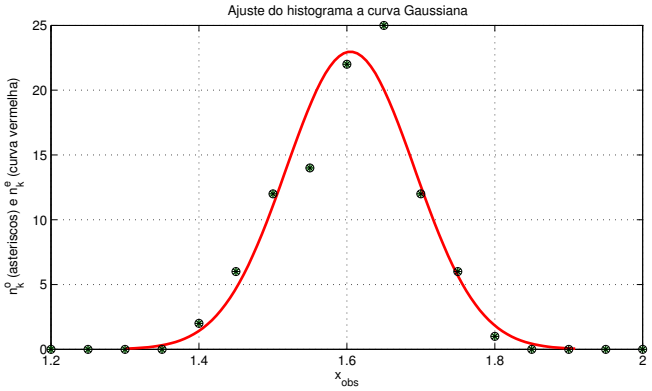


Figura 1: (a) Histograma para 100 valores de $x_{\rm obs}$. (b) Pontos do histograma sobre a curva Normal com parâmetros $\bar{x}_{\rm obs}$ e $s_{\rm obs}$.

de valores que *deveriam ter sido observados*, se a variável $x_{\rm obs}$ realmente fosse Gaussiana e o número de pontos $N \to \infty$:

$$n_{\rm e}^k = f(x_{\rm obs}^k) N \Delta x_{\rm obs}, \tag{2}$$

sendo $n_{\rm e}^k$ o número de amostras *esperadas* no intervalo k, $x_{\rm obs}^k$ o valor médio do intervalo k; N o número total de valores observados e $\Delta x_{\rm obs}$ o tamanho dos intervalos (classes) do histograma.

3 A variável Q^2

A fim de medir o quanto os pontos da curva vermelha (Fig.1b) se distanciam dos pontos obtidos para o histograma dos valores observados, podemos definir a variável Q^2, que

pode ser vista como um índice de qualidade de ajuste da curva:

$$Q^2 = \sum_{k=1}^{N_c} \frac{(n_{\mathrm{e}}^k - n_{\mathrm{o}}^k)^2}{n_{\mathrm{e}}^k} \tag{3}$$

sendo N_c o número de classes do histograma e n_{o}^k o número de valores observados na classe k do histograma traçado.

Como Q^2 é uma variável formada pela composição de valores aleatórios, é também uma variável aleatória.

Um fato interessante é que a função de densidade de probabilidade da nova variável Q^2 tende para uma fdp conhecida como função de densidade de probabilidade χ^2, mostrada na Fig.2, ou seja:

$$\lim_{N\to\infty} Q^2 = \chi^2.$$

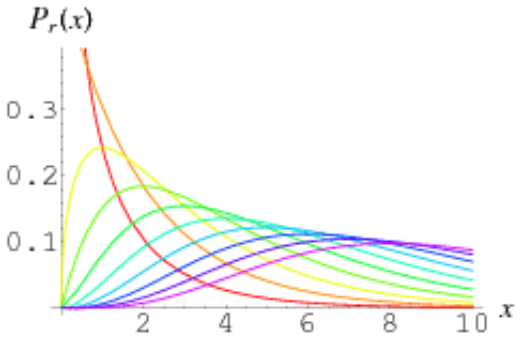


Figura 2: Funções de densidade de probabilidade χ^2 para diferentes graus de liberdade.

A fdp χ^2 é caracterizada pelo *número de graus de liberdade* g, definido como

$$g = N_c - p - 1;$$

sendo N_c o número de classes do histograma e p o número de parâmetros que precisaram ser estimados, pois não eram conhecidos a priori, para traçar a função de densidade de probabilidade Gaussiana. No presente caso, usamos a média e desvio-padrão amostrais para este fim, e portanto temos $p = 2$.

4 O Teste de Aderência χ^2

O teste de Gaussianidade χ^2, também chamado de teste de aderência χ^2 (qui-quadrado), fundamenta-se na minimização da chance de se cometer o seguinte equívoco, conhecido como *Erro do Tipo I* [Magalhães and de Lima, 2002]:

Rejeitar a hipótese de que os dados observados têm distribuição Gaussiana, para o caso em que os dados observados têm realmente distribuição Gaussiana.

Para isto, observamos que quanto maior for o valor de Q^2 em (3), maior será a discrepância entre o histograma obtido e a curva Gaussiana (ou curva Normal) correspondente, indicando que há maior chance de os dados não estarem distribuídos de forma Gaussiana.

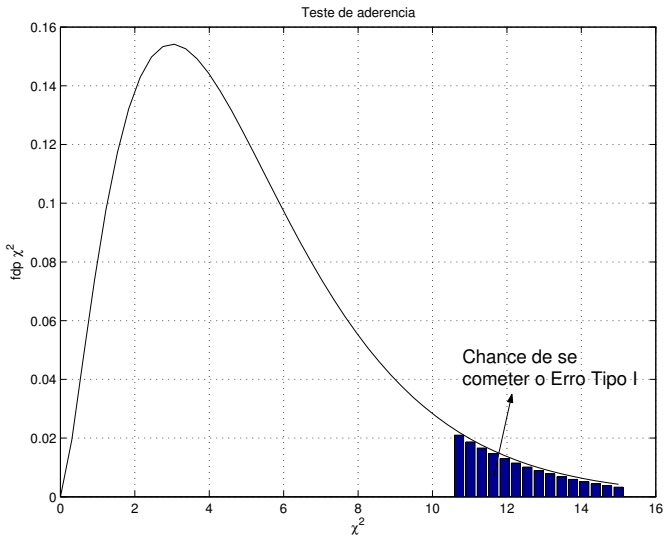


Figura 3: Exemplo de aplicação do teste de aderência χ^2 para um nível de siginificância $\alpha = 5\%$, ou seja, $q_c \approx 10{,}5$, para $g = 5$.

Para avaliarmos a chance de cometermos o equívoco supracitado, definimos um limite q_c para o valor de Q^2, tal que:

1. Se $Q^2 \geq q_c \Rightarrow$ rejeitamos a hipótese de que os dados observados têm distribuição Gaussiana.

2. Se $Q^2 < q_c \Rightarrow$ aceitamos a hipótese de que os dados observados têm distribuição Gaussiana.

Neste caso, quanto *maior* o valor de q_c, *menor* será a chance de cometermos o erro do Tipo I, ao rejeitarmos a hipótese de que a distribuição é Gaussiana. Por outro lado, mais fácil será aceitarmos que a distribuição é Gaussiana, mesmo que ela não seja (Erro Tipo II)!

É possível mostrar que, infelizmente, não conseguimos minimizar simultaneamente tanto a chance de cometermos o Erro Tipo I, quanto a chance de cometermos o Erro Tipo II. Escolhe-se, usualmente, a minimização do Erro Tipo I, como descrito neste documento.

Em função disto, dizemos que aceitamos ou rejeitamos a hipótese de que os dados têm distribuição Normal, a um *nível de significância* de α, onde α é a chance de cometermos o Erro Tipo I. Esta chance corresponde a área sob a curva da função de densidade de probabilidade χ^2, com g graus de liberdade, para o intervalo $\chi^2 > q_c$, conforme mostrado na Fig.3.

Os valores de q_c são obtidos de tabelas que contém valores para as probabilidades cumulativas χ^2, para diferentes graus de liberdade, em função do nível de significância α [Magalhães and de Lima, 2002].

Referências

[Magalhães and de Lima, 2002] Magalhães, M. N. and de Lima, A. C. P. (2002). *Noções de Probabilidade e Estatística*. EdUSP.