

Inferência Estatística II

Prof. Fernando de Souza Bastos
fernando.bastos@ufv.br

Departamento de Estatística
Programa de Pós-Graduação em Estatística Aplicada e Biometria
Universidade Federal de Viçosa
Campus UFV - Viçosa



Sumário

- 1 Introdução
- 2 Resultado Importante
- 3 Teorema
- 4 Teste Qui-Quadrado
- 5 Exemplo 1
- 6 Exemplo 2
- 7 Distribuição F
- 8 Teste F
 - Exemplo 1

Teste Qui-Quadrado

Testes qui-quadrado foram originalmente proposto por Karl Pearson em 1900, este teste representou um dos primeiros métodos de inferência estatística. Considere a variável aleatória X_i distribuída como $N(\mu_i, \sigma_i^2)$, em que $i = 1, 2, \dots, n$, e as variáveis X_1, X_2, \dots, X_n mutuamente independentes. A função de densidade conjunta dessas variáveis é dada por:

$$f(x_1, x_2, \dots, x_n) = \frac{1}{\sigma_1 \dots \sigma_n (2\pi)^{n/2}} e^{\left(-\frac{1}{2} \sum_{i=1}^n \left(\frac{x_i - \mu_i}{\sigma_i} \right)^2 \right)}, -\infty < x_i < \infty$$

A variável aleatória definida pelo expoente (exceto o coeficiente $-1/2$) é a soma $\sum_{i=1}^n \left(\frac{x_i - \mu_i}{\sigma_i} \right)^2$, essa variável aleatória segue uma distribuição Qui-Quadrado com n graus de liberdade, denotada como $\chi^2(n)$.

Prove que $Z^2 \sim \chi^2(1)$

Considere $V = Z^2$. Quero mostrar que a densidade de V é a densidade de uma variável qui-quadrado com 1 grau de liberdade, dada por:

$$g(v) = \frac{1}{2^{\frac{1}{2}} \Gamma\left(\frac{1}{2}\right)} v^{\frac{1}{2}-1} e^{-\frac{v}{2}}, \quad v > 0$$

Prove que $Z^2 \sim \chi^2(1)$

Considere $V = Z^2$. Quero mostrar que a densidade de V é a densidade de uma variável qui-quadrado com 1 grau de liberdade, dada por:

$$g(v) = \frac{1}{2^{\frac{1}{2}} \Gamma\left(\frac{1}{2}\right)} v^{\frac{1}{2}-1} e^{-\frac{v}{2}}, \quad v > 0$$

Para mostrar tal resultado, notem que

$$\begin{aligned} G(v) &= P(V \leq v) = P(Z^2 \leq v) \\ &= P(-\sqrt{v} \leq Z \leq \sqrt{v}) \\ &= 2P(0 \leq Z \leq \sqrt{v}), \quad Z \text{ é simétrica em relação a origem!} \\ &= 2 \int_0^{\sqrt{v}} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz \end{aligned}$$

Considere $z = \sqrt{y} \Rightarrow dz = \frac{y^{-\frac{1}{2}}}{2} dy = \frac{1}{2\sqrt{y}} dy$, nesse caso, $z = 0 \Rightarrow y = 0$, $z^2 \Rightarrow y = v$, logo,

Considere $z = \sqrt{y} \Rightarrow dz = \frac{y^{-\frac{1}{2}}}{2} dy = \frac{1}{2\sqrt{y}} dy$, nesse caso, $z = 0 \Rightarrow y = 0$, $z^2 \Rightarrow y = v$, logo,

$$\begin{aligned} G(v) &= 2 \int_0^{\sqrt{v}} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz \\ &= 2 \int_0^v \frac{1}{\sqrt{2\pi}} e^{-\frac{y}{2}} \frac{1}{2\sqrt{y}} dy \\ &= \int_0^v \frac{1}{\sqrt{\pi}\sqrt{2}} y^{\frac{1}{2}-1} e^{-\frac{y}{2}} dy, \quad v > 0 \end{aligned}$$

Considere $z = \sqrt{y} \Rightarrow dz = \frac{y^{-\frac{1}{2}}}{2} dy = \frac{1}{2\sqrt{y}} dy$, nesse caso, $z = 0 \Rightarrow y = 0$, $z^2 \Rightarrow y = v$, logo,

$$\begin{aligned} G(v) &= 2 \int_0^{\sqrt{v}} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz \\ &= 2 \int_0^v \frac{1}{\sqrt{2\pi}} e^{-\frac{y}{2}} \frac{1}{2\sqrt{y}} dy \\ &= \int_0^v \frac{1}{\sqrt{\pi}\sqrt{2}} y^{\frac{1}{2}-1} e^{-\frac{y}{2}} dy, \quad v > 0 \end{aligned}$$

Segue que,

$$G'(v) = g(v) = \frac{1}{\sqrt{\pi}\sqrt{2}} v^{\frac{1}{2}-1} e^{-\frac{v}{2}}, \quad v > 0$$

Usando o fato de $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$, temos que,

$$g(v) = \frac{1}{\Gamma\left(\frac{1}{2}\right) \sqrt{2}} v^{\frac{1}{2}-1} e^{-\frac{v}{2}}, \quad v > 0 \text{ cqd} \blacksquare$$

Teorema 1

Sejam Z_1, Z_2, \dots, Z_ν variáveis aleatórias independentes com distribuição normal padrão, onde ν é um inteiro positivo. Então, a variável aleatória $\chi^2(\nu) = \sum_{i=1}^{\nu} Z_i^2$ segue uma distribuição qui-quadrado com ν graus de liberdade.

Demonstração:

1. $Z_i \sim N(0, 1)$, então $Z_i^2 \sim \chi^2(1)$, como acabamos de demonstrar.

Demonstração:

1. $Z_i \sim N(0, 1)$, então $Z_i^2 \sim \chi^2(1)$, como acabamos de demonstrar.
2. A FGM de uma variável qui-quadrado com n graus de liberdade é dada por:

$$M_Y(t) = (1 - 2t)^{-n/2} \quad \text{para } t < \frac{1}{2}.$$

Demonstração:

1. $Z_i \sim N(0, 1)$, então $Z_i^2 \sim \chi^2(1)$, como acabamos de demonstrar.
2. A FGM de uma variável qui-quadrado com n graus de liberdade é dada por:

$$M_Y(t) = (1 - 2t)^{-n/2} \quad \text{para } t < \frac{1}{2}.$$

3. A FGM de Y é o produto das FGMs de V_1, V_2, \dots, V_ν porque elas são independentes:

$$M_Y(t) = M_{V_1}(t) \cdot M_{V_2}(t) \cdot \dots \cdot M_{V_\nu}(t).$$

Como $V_i \sim \chi^2(1)$, temos que $M_{V_i}(t) = (1 - 2t)^{-1/2}$.

4. Então:

$$M_Y(t) = [(1 - 2t)^{-1/2}]^\nu = (1 - 2t)^{-\nu/2}.$$

4. Então:

$$M_Y(t) = [(1 - 2t)^{-1/2}]^\nu = (1 - 2t)^{-\nu/2}.$$

5. Portanto:

$$Y \sim \chi^2(\nu).$$

Antes de entender o teste qui-quadrado de aderência, entenda a observação seguinte:

Antes de entender o teste qui-quadrado de aderência, entenda a observação seguinte:

Observação:

Devido aos resultados anteriores, se X_1 segue uma distribuição binomial $b(n, p_1)$, considere a variável aleatória:

$$Y = \frac{X_1 - np_1}{\sqrt{np_1(1 - p_1)}}$$

Antes de entender o teste qui-quadrado de aderência, entenda a observação seguinte:

Observação:

Devido aos resultados anteriores, se X_1 segue uma distribuição binomial $b(n, p_1)$, considere a variável aleatória:

$$Y = \frac{X_1 - np_1}{\sqrt{np_1(1 - p_1)}}$$

Observação:

Conforme n se aproxima do infinito, a variável Y tem uma distribuição aproximada $N(0, 1)$. Além disso, a distribuição de Y^2 é aproximadamente $\chi^2(1)$.

Observação:

Seja $X_2 = n - X_1$ e $p_2 = 1 - p_1$. Defina $Q_1 = Y^2$, então

$$Q_1 = \frac{(X_1 - np_1)^2}{np_1(1 - p_1)} = \frac{(X_1 - np_1)^2}{np_1} + \frac{(X_2 - np_2)^2}{np_2}$$

Isso ocorre porque $(X_1 - np_1)^2 = (n - X_2 - n + np_2)^2 = (X_2 - np_2)^2$.

Observação:

Seja $X_2 = n - X_1$ e $p_2 = 1 - p_1$. Defina $Q_1 = Y^2$, então

$$Q_1 = \frac{(X_1 - np_1)^2}{np_1(1 - p_1)} = \frac{(X_1 - np_1)^2}{np_1} + \frac{(X_2 - np_2)^2}{np_2}$$

Isso ocorre porque $(X_1 - np_1)^2 = (n - X_2 - n + np_2)^2 = (X_2 - np_2)^2$.

Observação:

Notem que se $(X_1 - np_1)^2 = (X_2 - np_2)^2$ e $p_2 = 1 - p_1$, então

$$\frac{(X_1 - np_1)^2}{np_1} + \frac{(X_2 - np_2)^2}{np_2} = \frac{np_2(X_1 - np_1)^2 + np_1(X_1 - np_1)^2}{n^2 p_1 p_2}$$

Usa-se o fato de $p_1 + p_2 = 1$.

Observação:

Esse resultado pode ser generalizado da seguinte maneira. Suponha que X_1, X_2, \dots, X_{k-1} tenham uma distribuição multinomial com os parâmetros n e p_1, p_2, \dots, p_{k-1} . Defina $X_k = n - (X_1 + X_2 + \dots + X_{k-1})$ e $p_k = 1 - (p_1 + p_2 + \dots + p_{k-1})$. Defina Q_{k-1} por:

$$Q_{k-1} = \sum_{i=1}^k \frac{(X_i - np_i)^2}{np_i}$$

À medida que n se aproxima do infinito, Q_{k-1} possui uma distribuição aproximada $\chi^2(k-1)$. Para usar essa aproximação, é importante garantir que n seja grande o suficiente para que cada np_i , onde $i = 1, 2, \dots, k$, seja pelo menos igual a 5.

Teste Qui-Quadrado

Considere um espaço amostral A de um experimento aleatório que seja a união de k conjuntos mutuamente disjuntos A_1, A_2, \dots, A_k . Suponha que $P(A_i) = p_i$, $i = 1, 2, \dots, k$, e $p_k = 1 - p_1 - p_2 - \dots - p_{k-1}$. Em que p_i é a probabilidade de que o resultado do experimento aleatório seja um elemento do conjunto A_i .

Teste Qui-Quadrado

Considere um espaço amostral A de um experimento aleatório que seja a união de k conjuntos mutuamente disjuntos A_1, A_2, \dots, A_k . Suponha que $P(A_i) = p_i$, $i = 1, 2, \dots, k$, e $p_k = 1 - p_1 - p_2 - \dots - p_{k-1}$. Em que p_i é a probabilidade de que o resultado do experimento aleatório seja um elemento do conjunto A_i .

Repita um experimento aleatório n vezes de forma independente, e considere X_i como o número de vezes que o resultado do experimento é um elemento do conjunto A_i . Ou seja, $X_1, X_2, \dots, X_k = n - X_1 - X_2 - \dots - X_{k-1}$ representam as frequências com as quais o resultado é, respectivamente, um elemento de A_1, A_2, \dots, A_k . Assuma que a função de massa de probabilidade conjunta de X_1, X_2, \dots, X_{k-1} é a função de massa de probabilidade multinomial com os parâmetros $n, p_1, p_2, \dots, p_{k-1}$.

Considere a hipótese simples:

$$H_0 : p_1 = p_{10}, p_2 = p_{20}, \dots, p_{k-1} = p_{k-1,0} \\ (p_k = p_{k0} = 1 - p_{10} - p_{20} - \dots - p_{k-1,0})$$

Em que $p_{10}, p_{20}, \dots, p_{k-1,0}$ são valores especificados. Deseja-se testar H_0 em relação a todas as alternativas.

Considere a hipótese simples:

$$H_0 : p_1 = p_{10}, p_2 = p_{20}, \dots, p_{k-1} = p_{k-1,0} \\ (p_k = p_{k0} = 1 - p_{10} - p_{20} - \dots - p_{k-1,0})$$

Em que $p_{10}, p_{20}, \dots, p_{k-1,0}$ são valores especificados. Deseja-se testar H_0 em relação a todas as alternativas.

Se a hipótese H_0 for verdadeira, a variável aleatória

$$Q_{k-1} = \sum_{i=1}^k \frac{(X_i - np_{i0})^2}{np_{i0}}$$

tem uma distribuição aproximada qui-quadrado com $k - 1$ graus de liberdade. Pois, se H_0 é verdadeira, np_{i0} é o valor esperado de X_i . Os valores observados de Q_{k-1} não devem ser muito grandes se H_0 for verdadeira.

Podemos, portanto, considerar o teste com nível de significância α que rejeita H_0 quando $Q_{k-1} \geq c$. Usando o software R, calculamos o valor crítico c por meio da função `qchisq(1 - α , k - 1)`. Se a hipótese H_0 for rejeitada quando o valor observado de Q_{k-1} for pelo menos igual a c , o teste de H_0 terá um nível de significância aproximadamente igual a α . Além disso, se q for o valor realizado da estatística do teste Q_{k-1} , o nível de significância observado do teste pode ser calculado em R pela função `1-pchisq(q, k - 1)`. Isso é frequentemente chamado de teste de bondade de ajuste ou de teste qui-quadrado para aderência.

Exemplo 1

Escolher um dos seis primeiros números inteiros positivos por meio de um experimento aleatório (por exemplo, lançamento de um dado). Seja $A_i = \{x : x = i\}$, onde $i = 1, 2, \dots, 6$.

$$H_0 : P(A_i) = p_{i0} = \frac{1}{6}, \quad i = 1, 2, \dots, 6,$$

$\alpha = 5\%$. Para realizar o teste, o experimento aleatório é repetido sob as mesmas condições, 60 vezes de forma independente. Ou seja, $k = 6$ e $n \cdot p_{i0} = 60 \cdot \frac{1}{6} = 10$, $i = 1, 2, \dots, 6$. Seja X_i a frequência com que o experimento aleatório termina com o resultado em A_i , $i = 1, 2, \dots, 6$. Defina Q_5 como:

$$Q_5 = \sum_{i=1}^6 \frac{(X_i - 10)^2}{10}$$

Uma vez que existem $6 - 1 = 5$ graus de liberdade, o valor crítico para um teste com nível $\alpha = 0.05$ é $qchisq(0.95, 5) = 11.0705$. Suponha agora que as frequências experimentais de A_1, A_2, \dots, A_6 sejam, respectivamente, 13, 19, 11, 8, 5 e 4. O valor observado de Q_5 é calculado como:

$$\begin{aligned} Q_{5,cal} &= \frac{(13 - 10)^2}{10} + \frac{(19 - 10)^2}{10} + \frac{(11 - 10)^2}{10} \\ &\quad + \frac{(8 - 10)^2}{10} + \frac{(5 - 10)^2}{10} + \frac{(4 - 10)^2}{10} \\ &= 15.6 \end{aligned}$$

Como $15.6 > 11.0705$, a hipótese $P(A_i) = \frac{1}{6}$, $i = 1, 2, \dots, 6$, é rejeitada a um nível de significância (aproximado) de 5%.

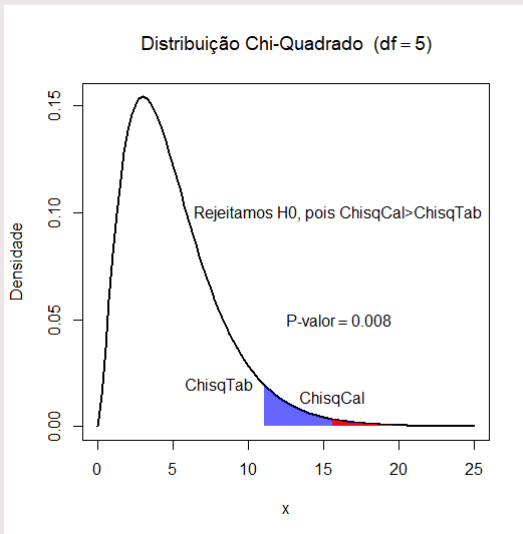


Figura: Gráfico com Resultados do Teste Qui-Quadrado

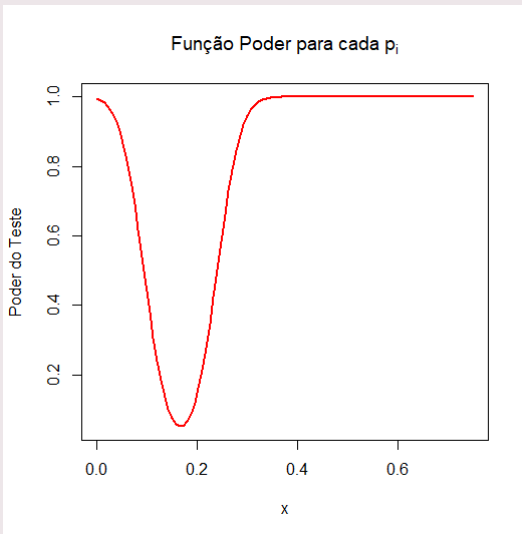


Figura: Poder do Teste Qui-Quadrado (Exemplo 1)

Exemplo 2

Neste exemplo, um ponto deve ser selecionado a partir do intervalo unitário $\{x : 0 < x < 1\}$ por meio de um processo aleatório. Definimos os conjuntos $A_1 = \{x : 0 < x \leq \frac{1}{4}\}$, $A_2 = \{x : \frac{1}{4} < x \leq \frac{1}{2}\}$, $A_3 = \{x : \frac{1}{2} < x \leq \frac{3}{4}\}$ e $A_4 = \{x : \frac{3}{4} < x < 1\}$. As probabilidades p_i , onde $i = 1, 2, 3, 4$, atribuídas a esses conjuntos sob a hipótese são determinadas pela função de densidade de probabilidade $2x$, em que $0 < x < 1$, e zero caso contrário. Essas probabilidades são:

$$\begin{aligned} p_{10} &= \int_0^{\frac{1}{4}} 2x \, dx = \frac{1}{16}, & p_{20} &= \int_{\frac{1}{4}}^{\frac{1}{2}} 2x \, dx = \frac{3}{16} \\ p_{30} &= \int_{\frac{1}{2}}^{\frac{3}{4}} 2x \, dx = \frac{5}{16}, & p_{40} &= \int_{\frac{3}{4}}^1 2x \, dx = \frac{7}{16} \end{aligned}$$

Portanto, a hipótese a ser testada é que p_1, p_2, p_3 e $p_4 = 1 - p_1 - p_2 - p_3$ possuem os valores anteriores em uma distribuição multinomial com $k = 4$. Considere $\alpha = 0.025$, repetindo o experimento aleatório $n = 80$ vezes de forma independente sob as mesmas condições, temos $n \cdot p_{i0}$, $i = 1, 2, 3, 4$ são, respectivamente, 5, 15, 25 e 35.

Portanto, a hipótese a ser testada é que p_1, p_2, p_3 e $p_4 = 1 - p_1 - p_2 - p_3$ possuem os valores anteriores em uma distribuição multinomial com $k = 4$. Considere $\alpha = 0.025$, repetindo o experimento aleatório $n = 80$ vezes de forma independente sob as mesmas condições, temos $n \cdot p_{i0}$, $i = 1, 2, 3, 4$ são, respectivamente, 5, 15, 25 e 35.

Suponha que as frequências observadas de A_1, A_2, A_3 e A_4 sejam 6, 18, 20 e 36, respectivamente. Assim,

$$Q_{3,cal} = \frac{(6 - 5)^2}{5} + \frac{(18 - 15)^2}{15} + \frac{(20 - 25)^2}{25} + \frac{(36 - 35)^2}{35} = \frac{64}{35} \approx 1.83.$$

Ou seja, falhamos em rejeitar H_0 , pois $\chi_{tab}^2 = qchisq(0.975, 3) = 9.34$.

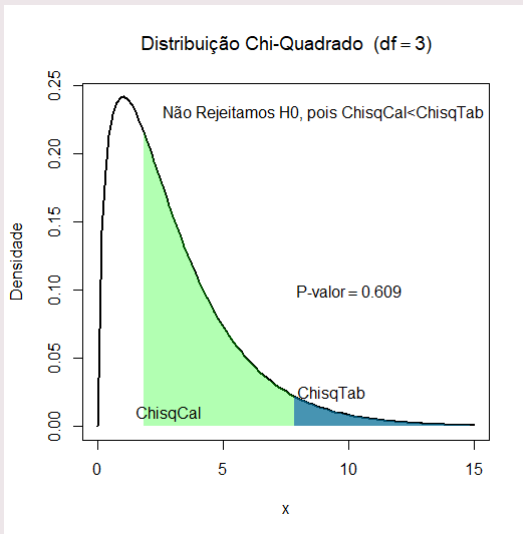


Figura: Gráfico com Resultados do Teste Qui-Quadrado

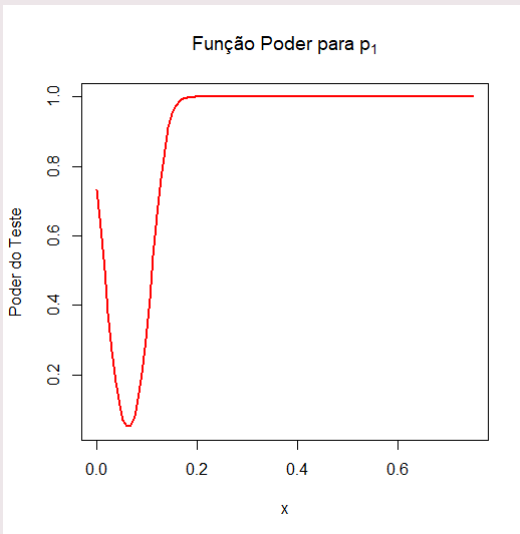


Figura: Poder do Teste Qui-Quadrado (Exemplo 2)

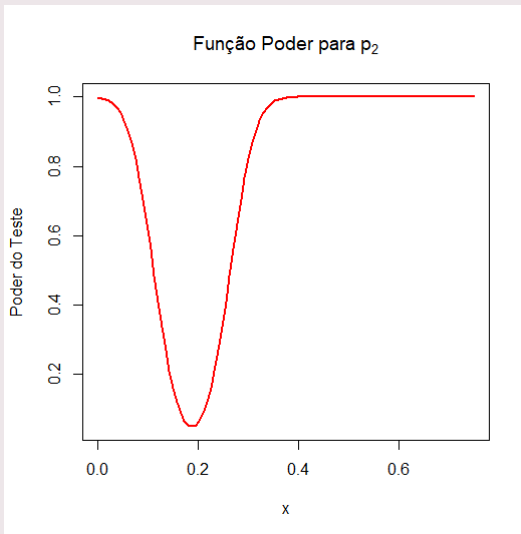


Figura: Poder do Teste Qui-Quadrado (Exemplo 2)

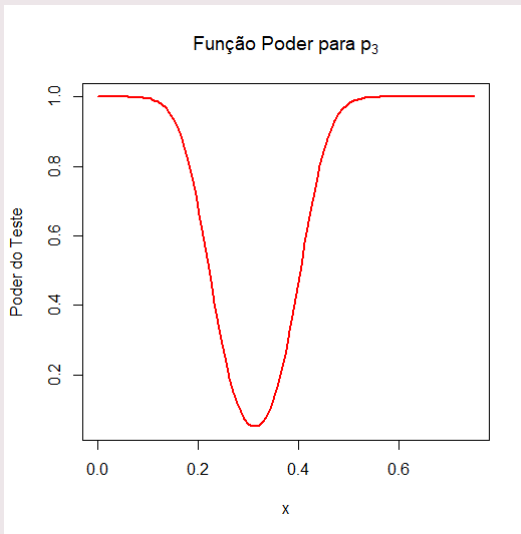


Figura: Poder do Teste Qui-Quadrado (Exemplo 2)

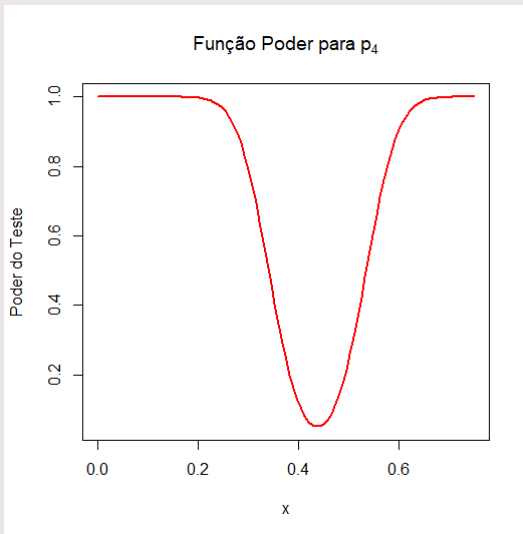


Figura: Poder do Teste Qui-Quadrado (Exemplo 2)

Para

- Exercícios da seção 4.7: 1, 4 e 9.

Distribuição F

Definição 1

Sejam Y e W variáveis aleatórias independentes, em que Y segue uma distribuição qui-quadrado com m graus de liberdade e W segue uma distribuição qui-quadrado com n graus de liberdade, m e n inteiros positivos dados. Defina uma nova variável aleatória X da seguinte forma:

$$X = \frac{Y/m}{W/n} = \frac{nY}{mW}.$$

Então, a distribuição de X é chamada de distribuição F com m e n graus de liberdade. A função de densidade de probabilidade da distribuição F é apresentada no próximo slide.

Seja X uma variável aleatória com distribuição F com m e n graus de liberdade. Então, sua função de densidade de probabilidade $f(x)$ é a seguinte, para $x > 0$:

$$f(x) = \frac{\Gamma\left[\frac{1}{2}(m+n)\right] m^{m/2} n^{n/2}}{\Gamma\left[\frac{1}{2}m\right] \Gamma\left[\frac{1}{2}n\right]} \frac{x^{m/2} - 1}{(mx + n)^{(m+n)/2}},$$

e $f(x) = 0$ para $x \leq 0$.

Suponha que as variáveis aleatórias X_1, \dots, X_m formem uma amostra aleatória de m observações de uma distribuição normal para a qual tanto a média μ_1 quanto a variância σ_1^2 são desconhecidas. Suponha também que as variáveis aleatórias Y_1, \dots, Y_n formem uma amostra aleatória independente de n observações de outra distribuição normal para a qual tanto a média μ_2 quanto a variância σ_2^2 são desconhecidas. Por fim, suponha que as seguintes hipóteses devam ser testadas a um nível de significância especificado α_0 ($0 < \alpha_0 < 1$):

$$H_0 : \sigma_1^2 \leq \sigma_2^2, \quad H_1 : \sigma_1^2 > \sigma_2^2.$$

Definimos S_X^2 e S_Y^2 como as variâncias amostrais de X e Y , respectivamente. Então, $\frac{(m-1)S_X^2}{\sigma_X^2} \sim \chi^2(m-1)$ e $\frac{(n-1)S_Y^2}{\sigma_Y^2} \sim \chi^2(n-1)$. Faz sentido intuitivo rejeitar H_0 se a razão desses dois estimadores for grande. Ou seja, definimos

$$V = \frac{S_X^2}{S_Y^2} \underset{\text{sob } H_0}{=} \frac{\frac{(m-1)S_X^2}{(m-1)\sigma_X^2}}{\frac{(n-1)S_Y^2}{(n-1)\sigma_Y^2}} \sim F(m-1, n-1),$$

e rejeitamos H_0 se $V \geq c$, em que c é escolhido para que o teste tenha um nível de significância desejado.

Exemplo 1

Suponha que seis observações X_1, \dots, X_6 sejam selecionadas aleatoriamente de uma distribuição normal para a qual tanto a média μ_1 quanto a variância σ_1^2 são desconhecidas, e é encontrado que $S_X^2 = 30$. Suponha também que 21 observações, Y_1, \dots, Y_{21} , sejam selecionadas aleatoriamente de outra distribuição normal para a qual tanto a média μ_2 quanto a variância σ_2^2 são desconhecidas, e é encontrado que $S_Y^2 = 40$. Realize um teste F das hipóteses

$$H_0 : \sigma_1^2 \leq \sigma_2^2, \quad H_1 : \sigma_1^2 > \sigma_2^2.$$

Neste exemplo, temos $m = 6$ e $n = 21$. Portanto, quando H_0 é verdadeira, a estatística V , seguirá a distribuição F com 5 e 20 graus de liberdade. Segue que o valor de V para as amostras dadas é $V = \frac{30}{40} = 0,75$.

O valor-p do teste, pode ser calculado no R por:

$$\begin{aligned} P_{\text{valor}} &= P(F > F_{\text{cal}}) \\ &= \text{pf}(F_{\text{cal}}, nx - 1, ny - 1, \text{lower.tail} = \text{FALSE}) = 0.596 \end{aligned}$$

A hipótese H_0 de que $\sigma_1^2 \leq \sigma_2^2$ seria, portanto, não rejeitada no nível de significância $\alpha_0 = 0.05$.

É possível mostrar que o poder deste teste pode ser escrito como:

$$\gamma_d(\sigma_1^2, \sigma_2^2) = 1 - F_{m-1, n-1}(C * \frac{\sigma_1^2}{\sigma_2^2}), C = F_{tab}$$

Suponha, assim, que seja importante rejeitar H_0 se σ_1^2 for três vezes maior que σ_2^2 . Nesse caso, gostaríamos que a função poder fosse alta quando $\sigma_1^2 = 3\sigma_2^2$. Nesse caso, $1 - pf(qf(0.95, 5, 20) * (1/3), 5, 20) = 1 - F_{5,20}(2.71 \times 1/3) \approx 0.5$. Logo, se σ_1^2 for três vezes maior que σ_2^2 , o teste com nível de 0.05 tem cerca de 50% de chance de rejeitar H_0 .

Referências I



Hogg, RV, J McKean e AT Craig (2019). *Introduction to Mathematical Statistics*.