Iniciação à Estatística

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

Departamento de Estatística
Universidade Federal de Viçosa
Campus UFV - Viçosa

Sumário

Um teste qui-quadrado, também escrito como teste χ^2, é qualquer teste de hipótese estatística em que a distribuição amostral da estatística de teste é uma distribuição qui-quadrada quando a hipótese nula é verdadeira. O teste qui-quadrado é utilizado para determinar se existe uma diferença significativa entre as frequências esperadas e as frequências observadas em uma ou mais categorias.

Existem três tipos:

- **Teste de aderência:** testa a hipótese da amostra ser proveniente de uma distribuição de probabilidade definida em H_0. Com essa distribuição definida em H_0 são obtidos as frequências esperadas (E);

Existem três tipos:

- **Teste de aderência:** testa a hipótese da amostra ser proveniente de uma distribuição de probabilidade definida em H_0. Com essa distribuição definida em H_0 são obtidos as frequências esperadas (E);
- **Teste de homogeneidade:** testa a hipótese H_0 de duas ou mais amostras serem provenientes de uma mesma distribuição de probabilidades. Os valores esperados são obtidos pelo produto da linha marginal e tamanho das amostras;

Existem três tipos:

- **Teste de aderência:** testa a hipótese da amostra ser proveniente de uma distribuição de probabilidade definida em H_0. Com essa distribuição definida em H_0 são obtidos as frequências esperadas (E);
- **Teste de homogeneidade:** testa a hipótese H_0 de duas ou mais amostras serem provenientes de uma mesma distribuição de probabilidades. Os valores esperados são obtidos pelo produto da linha marginal e tamanho das amostras;
- **Teste de independência:** testa a hipótese H_0 de que a distribuição conjunta é o produto das distribuições marginais, o que só ocorre quando existe independência entre as variáveis aleatórias. No caso de duas variáveis aleatórias organizadas numa tabela de dupla entrada, os valores esperados são obtidos como produto dos valores marginais.

Nos testes chi-quadrado o que muda é só a hipótese envolvida no cálculo dos valores esperados. Para os três tipos de hipótese, a estatística do teste é:

$$\chi^2_{cal} = \sum_{i=1}^{k} \frac{(f_{oi} - f_{ei})^2}{f_{ei}}$$

em que f_{oi} e f_{ei} são, respectivamente, as frequências observadas e esperadas. Sendo que sob H_0 a variável aleatória $\chi^2_{cal} \sim \chi^2_{\nu}$ em que ν são os graus de liberdade.

Teste de Aderência

Temos uma população P e queremos verificar se ela segue uma distribuição especificada P_0, isto é, queremos testar a hipótese $H_0 : P = P_0$. O procedimento consiste em considerar classes, segundo as quais a variável X, característica da população, pode ser classificada. A variável X pode ser qualitativa ou quantitativa.

Supondo H_0 verdadeira,

$$\chi^2 = \sum_{i=1}^{k} \frac{(f_{oi} - f_{ei})^2}{f_{ei}} \sim \chi^2_q,$$

em que $q = k - 1$ representa o número de graus de liberdade.
Observação: Este resultado é válido para n grande e para $f_{ei} \geq 5, i = 1, 2, \cdots, k$.

Regra de decisão: Pode ser baseada no nível descritivo ou valor-P, neste caso

$$valor - p = P(\chi^2_q \geq \chi^2_{cal}),$$

em que χ^2_{cal} é o valor calculado, a partir dos dados, usando a expressão apresentada para χ^2. Se para α fixado, $p - valor < \alpha$ rejeitamos H_0.

Exemplo 14.1 - Morettin and Bussab (2009)

Um dado é lançado 300 vezes, com os resultados dados na próxima Tabela. Por enquanto, considere somente a linha correspondente às frequências observadas. Com os resultados observados, queremos saber se o dado é "honesto", isto é, se a probabilidade de ocorrência de qualquer face é $1/6$. Ou seja, queremos testar a hipótese

$$H_0 : p_1 = p_2 = \cdots = p_6 = \frac{1}{6},$$

em que $p_i = P(\text{face } i), i = 1, 2, \cdots, 6$. Isso equivale a dizer que P_0 segue uma distribuição uniforme discreta.

Tabela: Resultados do lançamento de um dado 300 vezes. (Morettin and Bussab (2009))

Ocorrência (i)	1	2	3	4	5	6	Total
Freq. Observada	43	49	56	45	66	41	300
Freq. Esperada	50	50	50	50	50	50	300

$$H_0 : \text{O dado é honesto}$$
$$H_1 : \text{Não } H_0$$

$$\chi^2_{cal} = \frac{(43-50)^2}{50} + \cdots + \frac{(41-50)^2}{50} = 8.96$$

```
> (x2_t <- qchisq(0.05,df,lower.tail = FALSE))

[1] 11.0705

> (pvalor <- pchisq(x2_c,df,lower.tail = FALSE))

[1] 0.1106703
```

Conclusão: Não rejeitamos H_0 ao nível de 5% de significância.

```
> #Frequencia de acidentes nos dias da semana
> #hipótese H_0 é de as frequências são dadas por uma
> #distribuição uniforme discreta com n=5, ou seja,
> #p_i=1/5 para todo i={seg,ter,qua,qui,sex}
> #Bussab & Morettin-Estatística Básica-6 edição, pg-404
> Oi <- c(seg=32,ter=40,qua=20,qui=25,sex=33)
> Ei <- sum(Oi)*1/length(Oi)  #esperados sob H_0
> (X2 <- sum((Oi-Ei)^2/Ei))   #estatística do teste

[1] 7.933333

> nu <- length(Oi)-1          #graus de liberdade
> (pchisq(X2, df=nu, lower.tail=FALSE))#valor-p do teste

[1] 0.09405103
```

Teste de Homogeneidade

Considere o seguinte exemplo 14.2 do livro Morettin and Bussab (2009). Uma prova básica de Estatística foi aplicada a 100 alunos de Ciências Humanas e a 100 alunos de Ciências Biológicas. As notas são classificadas segundo os graus A, B, C, D e E (onde D significa que o aluno não recebe créditos e E indica que o aluno foi reprovado). Os resultados estão na Tabela abaixo:

Tabela: Frequências observadas

Aluno de	Grau					Total
	A	B	C	D	E	
C. Humanas	15	20	30	20	15	100
C. Biológicas	8	23	18	34	17	100
Total	23	43	48	54	32	200

Queremos testar se as distribuições das notas, para as diversas classes, são as mesmas para os dois grupos de alunos. Esse teste pode ser estendido para o caso de três ou mais populações. Novamente, para efetuar o teste, consideramos amostras das duas populações, P_1 e P_2, e classificamos os seus elementos de acordo com certo número de categorias para as duas variáveis características de P_1 e P_2.

Considerando P_1 como a população de alunos de Ciências Humanas e P_2 a dos alunos de Ciências Biológicas, nosso objetivo é testar a hipótese $H_0 : P_1 = P_2$, usando os resultados amostrais da Tabela anterior. Para isso, precisamos encontrar os valores esperados f_e, para aplicar a fórmula do χ^2.

A frequência esperada de cada entrada da tabela é obtido fazendo

$$f_{eij} = \frac{T_i * T_j}{T_G}, \ i = 1, \cdots, I, \ j = 1, \cdots, J.$$

em que T_i representa o total da linha i, T_j representa o total da coluna j e T_G representa o total geral. Logo, temos:

Tabela: Frequências esperadas

Aluno de	Grau					Total
	A	B	C	D	E	
C. Humanas	11,5	21,5	24	27	16	100
C. Biológicas	11,5	21,5	24	27	16	100
Total	23	43	48	54	32	200

```
> obs <- c( 15 ,  20 , 30, 20, 15,8,23,18,34,17)
> esp <- c(11.5, 21.5, 24, 27, 16,11.5, 21.5, 24, 27, 16)
> nlinhas <- 2
> ncolunas <- 5
> df <- (nlinhas-1)*(ncolunas-1)
> (x2_c <- sum(((obs-esp)^2)/esp))

[1] 9.094367

> alpha <- 0.05
> (x2_t <- qchisq(0.05,df,lower.tail = FALSE))

[1] 9.487729

> (pvalor <- pchisq(x2_c,df,lower.tail = FALSE))

[1] 0.05878356
```

De outra forma:

```
> c_hum<-c(15,20,30,20,15)
> c_bio<-c( 8,23,18,34,17)
> tab14_2<-rbind(c_hum,c_bio)
> test_tab14_2=chisq.test(tab14_2)
> test_tab14_2

Pearson's Chi-squared test

data:  tab14_2
X-squared = 9.0944, df = 4, p-value = 0.05878

> tab14_8 <-  rbind(test_tab14_2$expected,
+                 total=apply(test_tab14_2$expected,2,sum))
```

Exemplo 14.7 - Pag 408 - Morettin and Bussab (2009)

```
> tab14_10<-rbind(P_1T=c(29,60,9,2),P_2C=c(37,44,13,6))
> chisq.test(tab14_10)

Pearson's Chi-squared test

data:  tab14_10
X-squared = 6.1585, df = 3, p-value = 0.1041
```

Obs: O aviso "Chi-squared approximation may be incorrect" aparece por conta de termos uma das caselas menor que 5.

Teste de Independência

O teste de independência supõe a existência de duas v.a.'s X e Y, e os valores de amostras delas são classificados segundo categorias, obtendo-se uma tabela de dupla entrada. Queremos testar a hipótese que X e Y são independentes.

Exemplo 14.3 - Morettin and Bussab (2009)

Uma companhia de seguros analisou a frequência com que 2.000 segurados (1.000 homens e 1.000 mulheres) usaram hospitais. Os resultados estão na Tabela abaixo. A hipótese a testar é que o uso de hospital independe do sexo do segurado.

```
> M<-data.frame(uso_hospital=c("usaram_hospital",
+                              "nao_usaram_hospital")
+               ,homens=c(100,900),mulheres=c(150,850),
+               row.names = 1)
> M

                    homens mulheres
usaram_hospital        100      150
nao_usaram_hospital    900      850
```

Exemplo

Uma companhia de seguros analisou a frequência com que 2.000 segurados (1.000 homens e 1.000 mulheres) usaram hospitais. Os resultados estão na Tabela abaixo. A hipótese a testar é que o uso de hospital independe do sexo do segurado.

```
> Xsq <- chisq.test(M, correct = FALSE)
> Xsq

Pearson's Chi-squared test

data:  M
X-squared = 11.429, df = 1, p-value = 0.0007232
```

```
> Xsq$observed   # observed counts (same as M)

                   homens mulheres
usaram_hospital        100      150
nao_usaram_hospital    900      850

> Xsq$expected   # expected counts under the null

                   homens mulheres
usaram_hospital        125      125
nao_usaram_hospital    875      875

> Xquad <- (((100-125)^2)/125)+(((150-125)^2)/125)+
(((900-875)^2)/875)+(((850-875)^2)/875)
> Xquad

[1] 11.42857
```

```
> Xsq <- chisq.test(M, correct = TRUE)
> Xsq

Pearson's Chi-squared test with Yates' continuity correction

data:  M
X-squared = 10.976, df = 1, p-value = 0.000923
```

Referências

P. Morettin and W. Bussab. *Estatística básica*. Editora Saraiva, São Paulo, 6 edition, 2009.

W. Zeviane. Rídiculas - dicas curtas sobre o r, 2011. URL https://ridiculas.wordpress.com/2011/07/04/os-testes-chi-quadrado/.