

MATD49-Estatística não paramétrica

Testes de Qui-quadrado

Testes de Qui-quadrado I

Testes de qui-quadrado são desenhados a partir de tabelas de contingência. Trataremos dos seguintes testes:

- **Homogeneidade:** testamos se uma determinada variável se distribui da mesma forma em várias populações de interesse;
- **Independência:** neste teste, temos interesse em verificar se as variáveis com distribuições nas marginais são independentes ou não.
- **Aderência:** neste tipo de teste, testamos se uma determinada amostra é proveniente de uma suposta distribuição de probabilidade;
 - Ao longo do curso, veremos outros testes não paramétricos para a aderência: Kolmogorov-Smirnov, Liliefors, Shapiro-Wilk, Anderson-Darling.
- Nesta aula trabalharemos com tabelas de contingência $r \times c$ de forma geral. Posteriormente, discutiremos o caso particular 2×2 juntamente com o teste exato de Fisher e o teste de Mantel-Haenszel.

Teste de Homogeneidade I

- Tabelas de contingência são construídas a partir da contagem de indivíduos em c classes de uma variável aleatória categórica X de interesse, ao longo de amostras coletadas a partir de r populações diferentes, as quais temos interesse em comparar.

População	Categorias de X					Total
	1	2	3	...	c	
1	O_{11}	O_{12}	O_{13}	...	O_{1c}	n_1
\vdots						
r	O_{r1}	O_{r2}	O_{r3}	...	O_{rc}	n_r
Total	C_1	C_2	C_3	...	C_c	N

- Assumiremos amostra aleatórias em cada população e independência entre as populações;
- Cada observação coletada deve estar categorizada em uma (e só uma) das categorias de X .

Teste de Homogeneidade II

Hipóteses

H_0 : Todas as probabilidades numa mesma colunas são iguais ($p_{1j} = \dots = p_{rj} = p_j$, para todo j).

H_a : Pelo menos uma das probabilidades é diferente de p_{ij} é diferente de p_j , para algum j

Estatística de Teste I

Estatística de Teste:

Testamos se as frequências observadas diferem muito das frequências esperadas da seguinte forma:

$$Q = \sum_{i=1}^r \sum_{j=1}^c \frac{(O_{ij} - E_{ij})^2}{E_{ij}},$$

em que:

r, c = Número de populações e categorias;

O_{ij} = Frequência observada na categoria j da população i ;

E_{ij} = Frequência esperada na categoria j da população i , i.e. $E_{ij} = \frac{n_i C_j}{N}$.

Estatística de Teste II

- Quanto maior o valor de Q maior será a probabilidade de as frequências observadas estarem divergindo das frequências esperadas;
- Sob H_0 , Q tem distribuição aproximada de Qui-Quadrado com $\nu = (r - 1) \times (c - 1)$ graus de liberdade;

Decisão do Teste:

- Se $Q > \chi^2_{\nu; 1-\alpha}$, rejeita-se H_0 para o nível de significância α .
 - Pelo p-valor, temos que para rejeitar a hipótese nula $P(\chi^2_{\nu} > Q | H_0) = p < \alpha$.
-
- Esta aproximação é razoável quando os valores de E_{ij} não são pequenos, caso contrário há um aumento excessivo nos níveis do erro tipo I (perda de poder);
 - Dependendo do quão pequenos são os valores de E_{ij} , o teste pode ser pouco discriminatório (frequências esperadas entre 0.5 e 4) ou muito conservador (< 0.5);

Estatística de Teste III

- Veja [Cochran, 1952], [Cochran, 1954], [Koehler, 1986] e [Koehler, 1998] para uma discussão mais aprofundada sobre a inadequação da aproximação de qui-quadrado para caselas com poucas observações;
- Sugere-se considerar não mais do que 20% das caselas com menos do que 5 observações e um tamanho amostral total (N) mínimo de 25. Veja [Agresti, 2007] e [Conover, 1996];
- Caso tenha-se poucas observações ou frequências esperadas baixas em muitas caselas, recomenda-se agrupar as categorias ou recorrer a um outro teste.

Exemplo [Bussab and Morettin, 2017] I

Em uma empresa com 15.800 empregados, temos os funcionários distribuídos de acordo com a tabela abaixo.

	<25 anos	25-40 anos	>40 anos	Total
Homens	2000	4500	800	8300
Mulheres	800	2500	4200	7500
Total	2800 (17.7%)	7000 (44.3%)	6000 (37.97%)	15800

Tabela: Valores observados

	<25 anos	25-40 anos	>40 anos	Total
Homens	1471	3677	3152	8300
Mulheres	1329	3323	2848	7500
Total	2800	7000	6000	15800

Tabela: Valores esperados

Exemplo [Bussab and Morettin, 2017] II

O valor da estatística é

$$Q = \frac{(2000 - 1471)^2}{1471} + \cdots + \frac{(4200 - 2848)^2}{2848} = 3185$$

que é muito maior do que $\chi^2_{2,\alpha=0.1\%} = 13.82$. Logo, rejeitamos a hipótese de homogeneidade das distribuição de idade ao longo das populações de homem e mulher.

Teste Qui-quadrado de Independência

Temos agora interesse em avaliar a relação de dependência entre duas variáveis categóricas coletadas a partir de uma amostra aleatória.

Relembre que, se X e Y são v.a. independentes assumindo valores nos conjuntos enumeráveis $\mathcal{X} = \{x_1, x_2, \dots\}$ e $\mathcal{Y} = \{y_1, y_2, \dots\}$ respectivamente, então

$$P(X = x, Y = y) = P(X = x).P(Y = y), \quad (1)$$

para todo $x \in \mathcal{X}$ e todo $y \in \mathcal{Y}$.

- Ou seja, a probabilidade conjunta é dada pelo produto das marginais sob a hipótese de independência entre X e Y ;
- Utilizamos então (1) nas tabelas de contingência para definir os valores esperados sob H_0 : X e Y são independentes.

Tabelas Observada e Esperada

Categorias de Y	Categorias de X				Total
	1	2	...	c	
1	O_{11}	O_{12}	...	O_{1c}	R_1
\vdots	\vdots	\vdots	\ddots	\vdots	\vdots
r	O_{r1}	O_{r2}	...	O_{rc}	R_r
Total	C_1	C_2	...	C_c	N

Tabela: Valores observados

Categorias de Y	Categorias de X				Total
	1	2	...	c	
1	E_{11}	E_{12}	...	E_{1c}	R_1
\vdots	\vdots	\vdots	\ddots	\vdots	\vdots
r	E_{r1}	E_{r2}	...	E_{rc}	R_r
Total	C_1	C_2	...	C_c	N

Tabela: Valores esperados

Estatística de Teste I

Testamos se as freqüências observadas diferem muito das freqüências esperadas da seguinte forma:

$$Q = \sum_{i=1}^r \sum_{j=1}^c \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim \chi^2_{\nu}, \text{ sob } H_0,$$

em que:

$$\nu = (r - 1) \times (c - 1);$$

c, r : Número de categorias de X e Y , respectivamente;

O_{ij} : Freqüência observada na casela i, j ;

E_{ij} : Freqüência esperada na casela i, j , calculado a partir das probabilidades marginais:

$$E_{ij} = R_i C_j / N.$$

Estatística de Teste II

- Sob H_0 , Q tem distribuição aproximada de Qui-Quadrado com $\nu = (r - 1) \times (c - 1)$ graus de liberdade;
- Quanto maior o valor de Q maior será a probabilidade de as frequências observadas estarem divergindo das frequências esperadas;

Decisão do Teste:

- Se $Q > \chi^2_{\nu; 1-\alpha}$, rejeita-se H_0 para o nível de significância α ;
- Pelo p-valor, temos que para rejeitar a hipótese nula $P(\chi^2_{\nu} > Q | H_0) = p < \alpha$.

Exemplo [Bussab and Morettin, 2017] I

Verificar se a criação de cooperativas está associada com um fator regional.

	Consumidor	Produtor	Escola	Outros	Total
São Paulo	214	237	78	119	648
Paraná	51	102	126	22	301
R.G. Sul	111	304	139	48	602
Total	376	643	343	189	1551

Tabela: Dados observados. Fonte: Sinopse Estatística do Brasil – IBGE, 1977.

	Consumidor	Produtor	Escola	Outros	Total
São Paulo	157	269	143	79	648
Paraná	73	125	67	37	301
R.G. Sul	146	250	133	73	602
Total	376	643	343	189	1551

Tabela: Tabela esperada supondo independência entre região e tipo de cooperativa.

Exemplo [Bussab and Morettin, 2017] II

Estatística do teste:

$$Q = \frac{(214 - 157)^2}{157} + \dots + \frac{(48 - 73)^2}{73} = 173.38$$

que é muito maior do que $\chi^2_{6, \alpha=0.1\%} = 22.46$.

Logo, há evidências de dependência entre tipo de cooperativa e região.

Teste de Aderência

Hipóteses do Teste:

Hipótese Nula (H_0):

Não existe diferença entre as frequências obtidas na amostra e as frequências esperadas, dado que a população tem distribuição de probabilidade conhecida.

Hipótese Alternativa (H_1):

Existe diferença entre as frequências obtidas na amostra e as frequências esperadas, dado que a população tem distribuição de probabilidade conhecida.

Pressupostos I

Variável	Categorias				
	1	2	3	...	k
Frequência observada	O_1	O_2	O_3	...	O_k
Frequência esperada	E_1	E_2	E_3	...	E_k

$E_i = n \times p_i$, sendo p_i dado pelo modelo de probabilidade em H_0

Estatística de Teste I

Estatística de Teste:

Testamos se as frequências observadas diferem muito das frequências esperadas da seguinte forma:

$$Q = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i},$$

em que:

k = Número de categorias;

O_i = Frequência observada na categoria i ;

E_i = Frequência esperada na categoria i .

Estatística de Teste II

- Quanto maior o valor de Q maior será a probabilidade de as frequências observadas estarem divergindo das frequências esperadas;
- Q tem distribuição Qui-Quadrado com ν graus de liberdade, em que:
 - $\nu = k - 1$ se as frequências esperadas puderem ser calculadas sem precisar estimar os parâmetros da distribuição;
 - $\nu = k - m - 1$ se as frequências esperadas só puderem ser calculadas após a estimação dos m parâmetros da distribuição.

Decisão do Teste:

- Se $Q > \chi^2_{\nu; \alpha}$, rejeita-se H_0 para o nível de significância α .
- Pelo p-valor, temos que para rejeitar a hipótese nula $P(\chi^2_{\nu} > Q | H_0) = p\text{-valor} < \alpha$.

Comentários

Alguns Comentários:

- Muito utilizado na verificação de adequação de ajuste em variáveis nominais (*goodness of fit*);
- Se os dados são originalmente em escala intervalar ou de razão, o uso do teste χ^2 de aderência pode desperdiçar informações;
- Para variáveis ordinais não é sensível ao efeito de ordem. Quando a hipótese levar em conta a ordem, o teste qui-quadrado deixa de ser a melhor opção.

Acknowledgements

O Exemplo a seguir foi extraído do material didático do prof. Anderson Ara, com sua autorização. Agradecemos ao professor pela disponibilização de seu material didático.

Exemplo I

Deseja-se verificar, a um nível de 5% de significância, se o comportamento aleatório expresso pelo número de termostatos vendidos por uma indústria mensalmente segue uma distribuição normal com média 250 e desvio padrão 50. Para isso, a amostra de 65 observações relativas a demanda mensal de termostatos é dada abaixo.

158 222 248 216 226 239 206 178 169 177 290 245 318 158 274 255 191 244 149 195 247 233 156 203
182 272 256 184 299 246 184 224 316 285 251 248 222 135 337 235 225 282 302 391 261 315 231 231
263 177 270 169 172 243 278 283 200 324 266 214 240 239 224 262 200

- Número de intervalos (Regra de Sturges)

$$k = 1 + 3,31 \log_{10} n \Rightarrow k = 1 + 3,31 \log_{10} 65 \approx 7$$

- Amplitude da cada intervalo:

$$h = \frac{A}{k} = \frac{x_{(n)} - x_{(1)}}{k} = \frac{391 - 135}{7} \approx 36.57$$

Exemplo II

- Construindo a tabela de frequências

INTERVALO	AMPLITUDE
1	$\min(x) \vdash \min(x)+h$
2	$\min(x)+h \vdash \min(x)+2h$
3	$\min(x)+2h \vdash \min(x)+3h$
4	$\min(x)+3h \vdash \min(x)+4h$
5	$\min(x)+4h \vdash \min(x)+5h$
6	$\min(x)+5h \vdash \min(x)+6h$
7	$\geq \min(x)+6h$

Exemplo III

- Construindo a tabela de frequências

INTERVALO	AMPLITUDE
1	135.00 \vdash 171.57
2	171.57 \vdash 208.14
3	208.14 \vdash 244.71
4	244.71 \vdash 281.28
5	281.28 \vdash 317.85
6	317.85 \vdash 354.42
7	≥ 354.42

Exemplo IV

- Calculando p_i

$$\begin{aligned} p_1 &= P(135 < X < 171.57) \\ &= P\left(\frac{135 - \mu_0}{\sigma_0} < \frac{X - \mu_0}{\sigma_0} < \frac{171.57 - \mu_0}{\sigma_0}\right) \\ &= P\left(\frac{135 - 250}{50} < Z < \frac{171.57 - 250}{50}\right) \\ &\approx 0.05 \end{aligned}$$

Exemplo V

- Construindo a tabela de frequências com valores esperados

I_i	AMPLITUDE	O_i	p_i	$E_i = n \times p_i$
1	135.00 ┤ 171.57	7	0.05	3.25
2	171.57 ┤ 208.14	13	0.14	9.10
3	208.14 ┤ 244.71	17	0.26	16.90
4	244.71 ┤ 281.28	16	0.28	18.20
5	281.28 ┤ 317.85	8	0.18	11.70
6	317.85 ┤ 354.42	3	0.07	4.55
7	≥ 354.42	1	0.02	1.30
SOMA		65	1	65

Exemplo VI

- Construindo a tabela de frequências com valores esperados

$$Q = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i} = \frac{(7 - 3.25)^2}{3.25} + \dots + \frac{(1 - 1.3)^2}{1.3} \approx 8.03$$

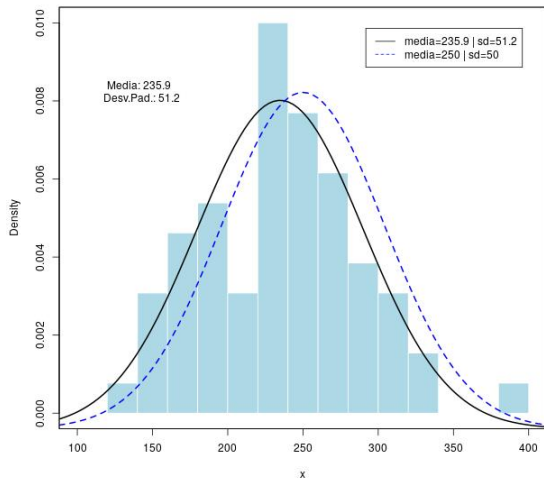
No R:

```
qchisq(0.95,df=6) = 12.59159
```

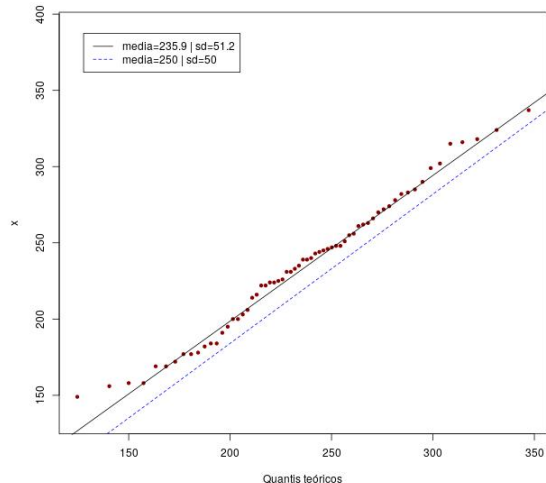
Como $Q < \chi^2_{6;5\%}$, ou seja, $8,03 < 12.59$, não rejeitamos H_0 à 5% de significância. A demanda mensal de termostatos possui, estatisticamente, uma distribuição normal com média 250 e desvio padrão 50.

QQplot do número de termostatos

Distribuição do número de termostatos



QQ-plot para número de termostatos



Aspecto Computacional I

No R:

```
chisq.test()
```

No SAS: FREQ procedure





```
proc freq data=dataset;  
tables var1*var2 /chisq;  
run;
```

Leitura complementar I




Veja:

- [Conover, 1996], capítulo 4;
- [Agresti, 2007], capítulos 2 e 3 e seção 9.8.4;
- [Bussab and Morettin, 2017], capítulo 14.

Referências I

-  Agresti, A. (2007).
An introduction to categorical data analysis.
Wiley-Interscience.
-  Bussab, W. O. and Morettin, P. A. (2017).
Estatística Basica.
Saraivauni, 9 edition.
-  Cochran, W. G. (1952).
The χ^2 test of goodness of fit.
The Annals of Mathematical Statistics, 23:315–345.
-  Cochran, W. G. (1954).
Some methods for strengthening the common χ^2 tests.
Biometrics, 10:417.

Referências II

-  Conover, W. J. (1996).
Practical nonparametric statistics.
John Wiley and sons, 3 ed. edition.
-  Koehler, K. J. (1986).
Goodness-of-fit tests for log-linear models in sparse contingency tables.
Journal of the American Statistical Association, 81:483–493.
-  Koehler, K. J. (1998).
Chi-square tests.
Encyclopedia of Biostatistics, pages 608–622.