Análise de Dados Categóricos

Tabelas $r \times c$ e Teste de Mantel-Haenszel

Enrico A. Colosimo

Departamento de Estatística
Universidade Federal de Minas Gerais

Tabelas $r \times c$

- Tabelas 2×2 são estendidas naturalmente para de maiores dimensões, chamadas $r \times c$.
- As estatísticas qui-quadrado e razão de verossimilhanças somam sobre as $r \times c$ caselas e, sob H_0, têm uma distribuição qui-quadrado com $(r-1) \times (c-1)$ graus de liberdade.
- Os resíduos padronizados podem ser utilizados para identificar as caselas que estão gerando a dependência, quando rejeitamos H_0.

Tabela r x c (r-linhas e c-colunas)- Desenho Multinomial (*n* fixo)

A tabela de contingência r x c é representada por

		Y				
		1	2	.	.	c
	1	π_{11}	π_{12}	.	.	π_{1c}
	2	.	.	.	.	.
X	.	.	.	.	.	.
	r	π_{r1}	π_{r2}	.	.	π_{rc}

em que, $r \geq 2$ e $c \geq 2$. Em um desenho multinomial:

$H_0 : \pi_{ij} = \pi_{i+}\pi_{+j}$

$i = 1, ..., r$ e $j = 1, ...c$

Teste Qui-Quadrado e Razão de Verossimilhança

- $X^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(n_{ij} - \widehat{E}_{ij})^2}{\widehat{E}_{ij}}$

- $G^2 = 2 \sum_{i=1}^{r} \sum_{j=1}^{c} n_{ij} log \frac{n_{ij}}{\widehat{E}_{ij}}$

em que,

$\widehat{E}_{ij} = \frac{n_{i+} n_{+j}}{n}$

$GL = (r - 1)(c - 1)$

Análise de resíduos

Os resíduos proporcionam uma avaliação das caselas da tabela que tem o maior impacto no valor das estatísticas de teste.

- Resíduos

  $$e_{ij} = (n_{ij} - \hat{E}_{ij})/\sqrt{\hat{E}_{ij}}$$

  $$X^2 = \sum_{i=1}^{r} \sum_{i=1}^{c} e_{ij}^2$$

  Problema: $\hat{Var}(e_{ij}) < 1$.

- Resíduos padronizados

  $$d_{ij} = \frac{e_{ij}}{\sqrt{[(1 - \frac{n_{i+}}{n})(1 - \frac{n_{+j}}{n})]}}$$

Os resíduos padronizados que estão fora do intervalo (-2,2) são considerados influentes no resultado do teste de Qui-Quadrado

Exemplo: Nature/97

scientific correspondence

bladder preparations were similar in *R. sylvatica* and the closely related, but freeze-intolerant, common frog (*R. temporaria*) of Europe and the leopard frog (*R. pipiens*) of North America. We also noted bladder permeability to glucose in the taxonomically distant bufonid, *Bufo marinus*, and the neotenic urodele, *Necturus maculosus*.

The taxonomic diversity of species exhibiting glucose permeability of the bladder indicates that this organ is fundamental for energy balance in amphibians whose carnivorous diet contains little carbohydrate[10]. The urinary bladder has long been used in studies of solute and water permeability, and may prove to be an ideal model for investigating transepithelial glucose flux.

Jon P. Costanzo, Phyllis A. Callahan
Richard E. Lee Jr, Michael F. Wright
Department of Zoology, Miami University, Oxford, Ohio 45056, USA
e-mail: costanjp@muohio.edu

1. Bentley, P. J. *Science* **152,** 619–623 (1966).
2. Boutilier, R. G., Stiffler, D. F. & Toews, D. P. in *Environmental Physiology of the Amphibians* (eds Feder, M. E. & Burggren, W. W.) 81–124 (Univ. Chicago Press, Chicago, Illinois, 1992).
3. Shoemaker, V. H. & Nagy, K. A. *Annu. Rev. Physiol.* **39** 449–471 (1977).
4. Costanzo, J. P., Lee, R. E. Jr, DeVries, A. L., Wang, T. & Layne, J. R. Jr *FASEB J.* **9,** 351–357 (1995).
5. Storey, K. B. & Storey, J. M. *Annu. Rev. Ecol. Syst.* **27,** 365–386 (1996).
6. Storey, K. B. & Storey, J. M. *J. Comp. Physiol.* **155,** 29–36 (1984).
7. Costanzo, J. P., Lee, R. E. Jr & Lortz, P. H. *J. Exp. Biol.* **181,** 245–255 (1993).
8. Russell, E. L. & Storey, K. B. *Cryo-Lett.* **16,** 263–266 (1995).
9. Layne, J. R. Jr, Lee, R. E. Jr & Cutwa, M. *J. Herpetol.* **30,** 85–87 (1996).
10. Pough, F. H. in *Behavioral Energetics: the Cost of Survival in Vertebrates* (eds Aspey, W. P. & Lustick, S. I.) 141–188 (Ohio State Univ. Press, Columbus, Ohio, 1983).

Parental age gap skews child sex ratio

The proportion of male to female births increases during and shortly after periods of war[1,2]. We show that the age difference between parents (age of husband − age of wife) predicts the sex of the first child. We also find that in England and Wales, the mean spouse age difference increased during and immediately after the two World Wars and was strongly correlated with the sex ratio during the period 1911–52.

We obtained the age and sex of children from 301 families who attended secondary schools that recruited from a wide range of socioeconomic groups. The mean age difference D_a (age of husband − age of wife) was 2.48 years ± 0.23 (s.e.m.) and there were 301 first-born and 260 second-born children. Among first-borns there was an excess of daughters from couples with low D_a and an excess of sons from those with high D_a ($D_a = -9$ to -1 years: 14 sons and 29 daughters; $D_a = 0$ to 5 years: 117 sons and 84 daughters; $D_a = 5$ to 15 years: 37 sons and 20 daughters; $\chi^2 = 11.86$, $P = 0.0027$). Among second-borns there was the opposite but non-significant tendency ($D_a = -9$ to -1 years: 22 sons and 11 daughters; $D_a = 0$ to 4 years: 93 sons and 89 daughters; $D_a = 5$ to 17 years: 20 sons and 25 daughters; $\chi^2 = 3.93$, $P = 0.14$).

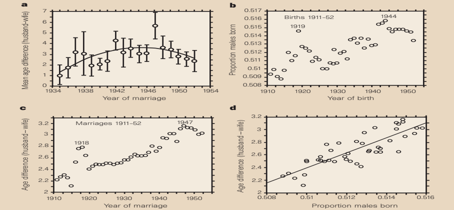


Figure 1 Parental age differences and sex-ratio statistics, 1911–52. **a,** The relationship between the mean (± s.e.m) of the difference in age between husbands and wives (D_a) and year of marriage (1935–52) in the Woolton area of Liverpool. There is a significant curvilinear relationship with a peak value of D_a in 1947 (second order polynomial, $y = -42.15 + 2.024x - 0.022x^2$, $F = 5.88$, $P = 0.013$, $n = 469$ marriages). **b,** Sex ratios of births registered in England and Wales from 1911–52; and **c,** D_a for marriages in the same period. **d,** Linear regression of sex ratio of births in England and Wales against D_a, 1911–52 ($r^2 = 0.68$, $F = 86.46$, $P = 0.0001$).

The age of parents at the birth of the child has a weak effect on the child's sex[3]. However, multiple regression analyses with sex of child as the dependent variable and D_a and age of mother or father at birth as independent variables showed that D_a remained significantly associated with sex of child (D_a/age of mother — D_a: standardized partial regression coefficient $b_1 = -0.14$, $t = 2.35$, $P = 0.02$; age of mother: $b_2 = 0.13$, $t = 0.22$, $P = 0.83$; D_a/age of father — D_a: $b_1 = 0.14$, $t = 2.34$, $P = 0.02$; age of father: $b_2 = 0.13$, $t = 0.21$, $P = 0.83$).

Local and national patterns of D_a during the period 1911–52 (ref. 4) are shown in Fig. 1a, c. If couples do not delay the birth of their first child, D_a and sex ratio should be correlated and changes in the sex ratio should be preceded by changes in D_a. This is seen in 1914–18 but not during the Second World War (Fig. 1b, c). Registration of second and subsequent births will weaken the relationship between D_a and sex ratio so that an exact correlation is unlikely. Nevertheless a regression of sex ratio on D_a shows that the latter explains 68% of the variance of the former (Fig. 1d). Age of woman at marriage was negatively related to the sex ratio ($b = -0.003$, $r^2 = 0.23$, $F = 12.19$, $P = 0.001$). However a multiple regression analysis with sex ratio as the dependent variable and D_a and bride's age as independent variables left D_a as the only significant correlate of sex ratio (D_a: $b_1 = 0.78$, $t = 8.26$, $P = 0.0001$; age of bride: $b_2 = -0.14$, $t = 1.51$, $P = 0.14$).

Rank in many animals is related to the sex of their offspring[5]. In humans, the elite often form partnerships with high D_a[6] and have more sons than daughters[7]. It may be that during wartime women prefer to marry older men with high resources and this leads to an increase in D_a. We do not know how the sex of first-borns is adjusted in relation to D_a. Women could influence the motility of sperm bearing either X or Y chromosomes or they may invest differentially in males and females *in utero* leading to higher miscarriage rates of one or the other sex.

J. T. Manning, R. H. Anderton, M. Shutt
Population Biology Research Group,
School of Biological Sciences,
University of Liverpool, Liverpool L69 3BX, UK
e-mail: jtmann@liv.ac.uk

1. Martin, W. J. *Lancet* **1,** 807 (1943).
2. MacMahon, B. & Pugh, T. F. *J. Hum. Genet.* **6,** 284–292 (1954).
3. Bromwich, P. *Prog. Obstet. Gynaecol.* **7,** 217–231 (1989).
4. *The Registrar General's Statistical Review of England and Wales, Part II Civil* (HMSO, London, 1921–1952).
5. Clutton-Brock, T. H. & Iason, G. R. *Q. Rev. Biol.* **61,** 339–374 (1986).
6. Kenrick, D. T. & Keefe, R. C. *Behav. Brain Sci.* **15,** 75–133 (1992).
7. Mueller, U. *Nature* **363,** 490 (1993).

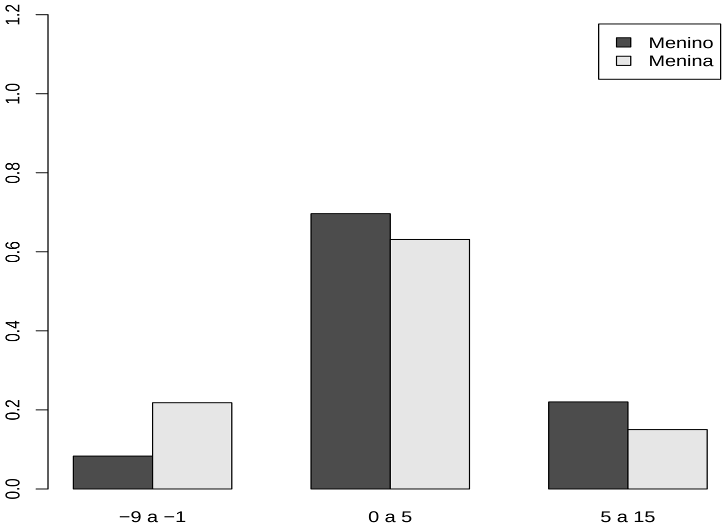
Exemplo: Tabelas $r \times c$

Exemplo: Diferença de idade entre os pais e sexo do primeiro filho na Inglaterra e País de Gales (Nature, set/97).

Sexo Recém-nascido	Dif Idade: Pai - Mãe -9 a -1	0 a 5	5 a 15	Total
Menino	14	117	37	168
Menina	29	84	20	133
Total	43	201	57	301

$\widehat{E}_{ij} > 5, i = 1, 2;\ j = 1, 2, 3.$
$X^2 = 11,81, gl = 2, \text{valor-p} = 0,002725$
$G^2 = 11,86, gl = 2, \text{valor-p} = 0,0027$
MC: valor-p=0,0026

Descrevendo os Dados



Exemplo: Diferença de idade entre os pais e sexo do récem-nascido (Nature, set/97).

- Resíduos

Sexo	Dif Idade: Pai - Mãe		
Recém-nascido	-9 a -1	0 a 5	5 a 15
Menino	-2,04	0,45	0,92
Menina	2,29	-0,51	-1,03

- Os resíduos indicam que existem mais meninas em casais com diferença de idade entre -9 e -1 anos do que a hipótese de independência prediz.

Análise de resíduos padronizados: Tabelas $r \times c$

Exemplo: Diferença de idade entre os pais e sexo do récem-nascido (Nature, set/97).

- Resíduos Padronizados

Sexo	Dif Idade: Pai - Mãe		
Recém-nascido	-9 a -1	0 a 5	5 a 15
Menino	-3,32	1,19	1,54
Menina	3,32	-1,19	-1,54

- Conclusão similares ao anterior.

Dividindo em sub-tabelas

Os testes avaliam a dependência entre as variáveis, mas não permitem uma análise localizada. Para isso, a tabela pode ser particionada em sub-tabelas. Os principais objetivos são:

- Dividir a tabela em sub-tabelas menores.
- A decisão sobre quais colunas combinar deve ser feita em conjunto com o pesquisador.
- Identificar categorias que estão causando a dependência.
- Identificar categorias similares que podem, se for de interesse, serem combinadas, reduzindo a dimensão da tabela.

Particionando em sub-tabelas independentes.
Regras para Dividir em Sub-tabelas

- Os graus de liberdade das sub-tabelas devem somar à da tabela original;
- O valor de cada casela deve aparecer uma única vez em uma das sub-tabelas.
- Cada total marginal da tabela original deve ser o total marginal em uma única sub-tabela.

O valor das estatísticas qui-quadrado somam ao da tabela original.

Exemplo: Tabelas $r \times c$

Exemplo: Diferença de idade entre os pais e sexo do primeiro filho (Nature, set/97).

- Tabela 1 ($X_1^2 = 11$)

Sexo	Dif Idade: Pai - Mãe		
Recém-nascido	-9 a -1	0 a 15	Total
Menino	14	154	168
Menina	29	104	133
Total	43	258	301

- Tabela 2 ($X_2^2 = 0,81$)

Sexo	Dif Idade: Pai - Mãe		
Recém-nascido	0 a 5	5 a 15	Total
Menino	117	37	154
Menina	84	20	104
Total	201	57	301

Tabelas $r \times c$: RC (IC 95%)

- Razão de Chances (IC 95%)
  - $\widehat{RC}_{12} = 2,9\ (1,4; 5,9)$, a chance de ocorrer uma menina para pais com diferença de idade entre -9 e -1 anos é cerca de 3 vezes a chance daqueles entre 0 e 5 anos.
  - $\widehat{RC}_{13} = 3,8\ (1,6; 9,0)$,a chance de ocorrer uma menina para pais com diferença de idade entre -9 e -1 anos é cerca de 3,8 vezes a chance daqueles entre 5 e 15 anos.

Teste de Tendência Linear: Tabelas $2 \times c(> 2)$

- Teste Cochran-Armitage - (Agresti, ICDA, p.34-39; Giolo, p.53).
- Níveis ordenados da variável coluna.
- Teste para detectar aumento linear nos níveis da variável linha.
- Necessário atribuir escores aos níveis da variável coluna. Em variáveis categorizadas, os escores surgem naturalmente. Em outros casos, usamos usualmente, 1,2,3,...
- Por exemplo: (1) no exemplo anterior podemos usar os pontos médios de classe: -5; 2,5 e 10; (2) para gravidade: baixa (1), média (2) e alta (3).

Exemplo:Variáveis X vs Y1 e X vs Y2 .

X	Y1			Y2		
	1	2	3	1	2	3
1	10	20	40	10	40	20
2	90	80	60	90	60	80

- As duas tabelas são equivalentes somente trocando as colunas 2 e 3 da primeira na segunda tabela.
- Isto significa que o teste qui-quadrado para a associação entre X e Y é EXATAMENTE o mesmo nas duas tabelas apresentando o valor 26,09 (valor-p$< 0,01$).
- No entanto, na primeira tabela existe uma tendência linear para os níveis de X (10, 20 e 40%) enquanto que na segunda este efeito não está presente.

Teste de Tendência Linear: Tabelas $2 \times c(> 2)$

- Definir escores para as colunas ordenadas: $s_1, \ldots, s_c$.
- Notação para a tabela ordenada com $c = 3$.

X	Y			
	1	2	3	Total
1	n_{11}	n_{12}	n_{13}	n_{1+}
2	n_{21}	n_{22}	n_{23}	n_{2+}
Total	n_{+1}	n_{+2}	n_{+3}	n

- Escore médio estimado para cada linha de X:

$$em_i = \sum_{j=1}^{3} \frac{s_j n_{ij}}{n_{i+}} = (s_1 n_{i1} + s_2 n_{i2} + s_3 n_{i3})/n_{i+}$$

o numerador soma o escore em Y para todos os indivíduos na i-ésima linha.

Teste de Tendência Linear: Tabelas $2 \times c (> 2)$

- Para obtermos o teste, devemos encontrar a média (μ) e a variância (σ^2) da estatística em_i, sob a hipótese nula de não haver tendência linear.
- $$\mu = \sum_{j=1}^{3} \frac{s_j n_{+j}}{n}$$
- $$\widehat{\sigma}^2 = \frac{(n - n_{1+})}{(n-1)n_{1+}} \sum_{j=1}^{3} (s_j - \mu)^2 \frac{n_{+j}}{n}$$
- Estatística teste é dada por $(em_1 - \mu)^2/\sigma^2$, que tem sob a hipótese nula, uma distribuição aproximada qui-quadrado com 1 grau de liberdade.

Teste de Tendência Linear: Tabelas $2 \times c > 2$: Exemplos

Exemplo:Variáveis X vs Y1 e X vs Y2 .

X	Y1			Y2		
	1	2	3	1	2	3
1	10	20	40	10	40	20
2	90	80	60	90	60	80

- O teste para tendência linear na tabela apresenta um resultado igual a 25,07, com um valor-p $< 0,01$. No caso da segunda tabela, o valor é 2,79, com valor-p$= 0,095$.

Teste de Tendência Linear: Exemplo Nature

- No caso do exemplo da Nature obtemos um valor de 9,38 para a estatística teste, resultando em um valor-p = $0,0022$.
- Este teste confirma a tendência linear da proporção de meninos como primeiro filho com o aumento da diferença de idade entre o pai e a mãe.
- Observe que o teste qui-quadrado de independência tem dois gl, com valor igual a 11,81.
- Isto significa que

$$X^2 = 11,81 = X_L^2 + X_Q^2 = 9,38 + 2,43$$

O que fazer se queremos controlar por outras variáveis?

Tratamento Geral

- Estratificação
  - Controlar por uma ou mais variáveis.
  - Teste de Mantel-Haenszel.
  - Limitação: tamanho de amostra.
- Modelos Estatísticos.
  - Modelo multivariado log-linear.
  - Regressão de Poisson (resposta contagem).
  - Regressão Logística (resposta categórica).

- Testar associação entre duas variáveis, controlando por uma terceira (ou por mais de uma variável).
- Exemplo: controlar por idade em três faixas etárias.
- A terceira variável define os estratos. O teste de Mantel-Haenszel combina as tabelas em um único teste e em uma única estimativa para a RC.

Exemplo: Campanha Publicitária para um determinado produto em duas cidades (A e B). Preferência de 2000 consumidores pelo produto X após a campanha publicitária.

- Cidade A

	Preferência		
	Sim	Não	Total
Semana 1	60	140	200
Semana 2	320	480	800

- X-squared = 6,79, df = 1, p-value = 0,0092
- $\hat{RC}$ = 0,64 (0,46; 0,90): a chance de venda na semana 1 é 0,64 vezes a chance da semana 2.

- Cidade B

	Preferência		
	Sim	Não	Total
Semana 1	640	160	800
Semana 2	180	20	200

- X-squared = 10,84, df = 1, p-value = $< 0{,}001$
- $\hat{RC}$ = 0,44 (0,27; 0,73)

Exemplo: Campanha Publicitária em duas cidades: Combinando Cidades A e B.

- Cidades A + B (Associação Marginal)

	Preferência		
	Sim	Não	Total
Semana 1	700	300	1000
Semana 2	500	500	1000

  - X-squared = 83,33, df = 1, p-value = $< 0,001$
  - $\hat{RC}$ = 2,33 (1,94; 2,80)

O que está acontecendo?
Tratamento Geral

- Condições das Cidades:
  - Cidade A obteve menos vendas e foi mais amostrada na semana 2.
  - Cidade B obteve mais vendas e foi mais amostrada na semana 1.
- Realidade: existe um aumento das vendas.
- Razão do Problema: a variável Z (cidade) está relacionada tanto com X (vendas) quanto com Y (semana) (variável de confusão).
- Solução: Testar a associação de X e Y controlando por Z.
- Teste de Independência Condicional: X indep. de Y, dado Z.
- Teste de Mantel-Haenszel

Teste de Mantel-Haenszel

Notação: k-ésima tabela; $k = 1, \ldots, I$

.

	Y		
X	1	2	Total
1	n_{11k}	n_{12k}	n_{1+k}
2	n_{21k}	n_{22k}	n_{2+k}
Total	n_{+1k}	n_{+2k}	n_k

Teste de Mantel-Haenszel (1958)

A estatística de MH para l tabelas é dada por:

$$MH = \frac{(|\sum_{k=1}^{l}(n_{11k} - \widehat{E}(n_{11k}))| - 0,5)^2}{\sum_{k=1}^{l}\widehat{Var}(n_{11k})}$$

em que

$$\widehat{E}(n_{11k}) = \frac{n_{1+k}n_{+1k}}{n_k}$$

e

$$\widehat{Var}(n_{11k}) = \frac{n_{1+k}n_{2+k}n_{+1k}n_{+2k}}{n_k^2(n_k - 1)}$$

sob H_0, MH tem uma distribuição qui-quadrado com 1 gl.

Obs. Esta é a versão do MH com correção de continuidade. Basta retirarmos o termo $-0,5$ do numerador para termos a versão padrão.

$$\widehat{RC}_{MH} = \frac{\sum_{k=1}^{I} \frac{n_{11k} n_{22k}}{n_k}}{\sum_{k=1}^{I} \frac{n_{12k} n_{21k}}{n_k}}$$

- $\widehat{RC}_{MH}$ é chamado de razão de chances combinado para a associação entre X e Y ou simplesmente de razão de chances de Mantel-Haenszel.
- Da mesma forma anterior obtemos uma $\widehat{Var}(log(\widehat{RC}_{MH})$ e o intervalo de 95% de confiança para RC é dado por

$$\exp(\log(\widehat{RC}_{MH}) \pm 1,96\sqrt{\widehat{Var}(log(\widehat{RC}_{MH}))}$$

Resultos do Exemplo: Campanha Publicitária em duas cidades (Paradoxo de Simpson).
Teste de Mantel-Haenszel

- MH = 16,17, df = 1, p-value = 5,798e-05.
- $\hat{RC}$ = 0,57 (0,43; 0,74): a chance de venda na semana 1 é 0,57 vezes a chance da semana 2.

Exemplo: Teste de Mantel-Haenszel

Giolo, p. 88

Exemplo: Ensaio clínico para comparar duas drogas para o tratamento de infecções respiratórias em dois diferentes centros.

		Resposta		
Centro	Tratamento	Favorável	Não favorável	Total
1	Novo	29	16	45
1	Padrão	14	31	45
Total		43	47	90
2	Novo	37	8	45
2	Padrão	24	21	45
Total		61	29	90

Exemplo: Ensaio clínico para comparar duas drogas para o tratamento de infecções respiratórias em dois diferentes centros.

		Resposta			
Centro	Tratamento	Favorável	Não favorável	Total	$\widehat{RC}$
1	Novo	29 (**64%**)	16	45	4,0
1	Padrão	14 (**31%**)	31	45	
Total		43	47	90	
2	Novo	37 (**82%**)	8	45	4,0
2	Padrão	24 (**53%**)	21	45	
Total		61	29	90	

A estimativa da RC para a tabela combinada é $3,76$.

Exemplo: Ensaio clínico para comparar duas drogas para o tratamento de infecções respiratórias em dois diferentes centros.

- MH = 18,41, GL = 1, p-value = 1,78e-05.
- Razão de Chances

$$\widehat{RC} = 4,0 \quad (IC, 95\%; 2,1; 7,7)$$

- MH é um teste de independência condicional

$$H_0 : \pi_{ij(k)} = \pi_{+j(k)}\pi_{i+(k)}.$$

- Podemos somar sobre centros e obter uma tabela de duas entradas?
- Sim, se houver independência marginal.

- O teste de MH é adequado para situações em que queremos verificar a associação entre duas variáveis binárias controlando pelas demais.
- Este teste é chamado de independência condicional.
- O teste é inapropriado quando a associação varia muito entre as tabelas parciais.
- As variáveis a serem controladas têm que ser categóricas ou categorizadas.
- O teste fica muito limitado na presença de muitas tabelas ou tabelas com pequeno tamanho amostral.