Testes de hipótese para tabelas de contingência: parte 2 (testes de aderência e medidas de associação/dependência)

Prof. Caio Azevedo

Exemplo 6: distribuição espacial de árvores

- Os dados a seguir (extraídos de Andrade e Ogliari (2010)) se referem ao número de árvores por quadrante da espécie *Guapira opposita*, obtidos de um estudo realizado com o objetivo de verificar a distribuição espacial dessa espécie num local de restinga.
- Foram considerados um total de 94 quadrantes e contou-se o número de quadrantes com zero árvores, uma árvore, duas árvores, assim por diante.
- Na última categoria foram contabilizados todos os quadrantes que apresentarem pelo menos nove árvores.

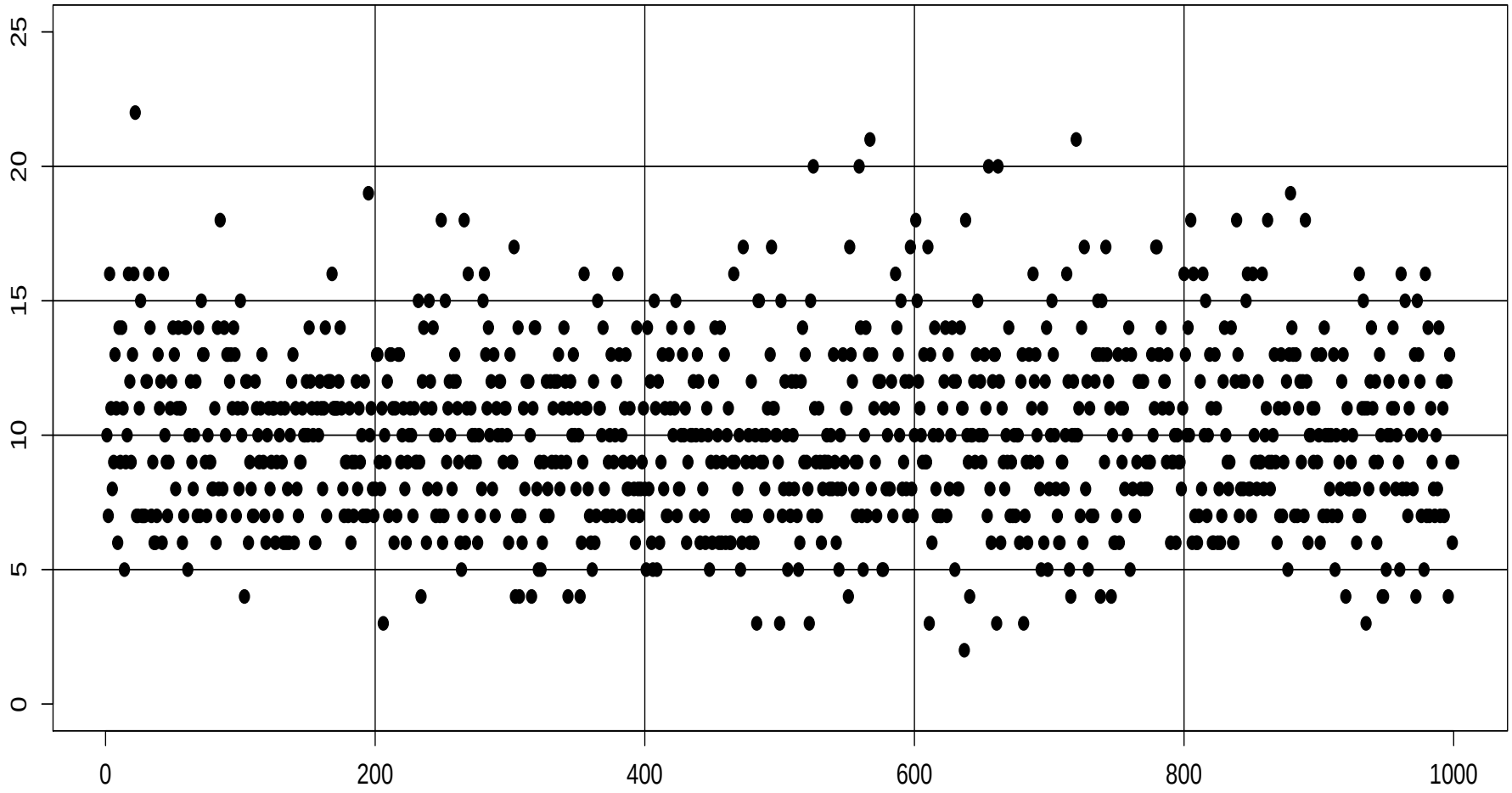
Exemplo 6 (cont.)

- As hipóteses de interesse são:
  - H_0 : A espécie se distribui aleatoriamente na região (a probabilidade de uma árvore ocorrer em qualquer ponto da região é a mesma e independe de qualquer outra árvore).
  - H_1 : A espécie não se distribui aleatoriamente.
- Equivalentemente:
  - H_0 : A distribuição de Poisson (discutível) é apropriada para modelar o comportamento (aleatório) da dispersão espacial.
  - H_1 : A distribuição de Poisson não é apropriada para modelar o comportamento (aleatório) da dispersão espacial.

Exemplo 6 (cont.)

- Lembrando que a estatística para testar a aderência (adequabilidade) é $Q_H = \sum_{i=1}^{m} \frac{(N_i - E_i)^2}{E_i}$.
- Temos que: $E_i = P(X_i = i), i = 1, 2, .., 9,$ $X_i \sim \text{Poisson}(\widetilde{\lambda}), \widetilde{\lambda} = \frac{1}{94} \sum_{i=1}^{n} x_i y_i$, x_i : número de árvores por quadrante, y_i : número de quadrantes com x_i árvores.
- Para calcular $\widetilde{\lambda}$ consideramos uma média ponderada de sorte que, na última categoria $x_i = 9$.

Ilustração da estrutura dos dados



Dados e análise

x_i	y_i	Prob. de Poisson	Num. esperado de quadrantes
0	6	0,0566	5,3172
1	18	0,1625	15,2729
2	23	0,2333	21,9345
3	19	0,2234	21,0011
4	11	0,1604	15,0806
5	6	0,0922	8,6633
6	5	0,0441	4,1473
7	4	0,0181	1,7018
8	1	0,0065	0,6110
9	1	0,0021*	0,1950

(* Calculada para $x_i = 9$). Nesse caso, $q_H = 9,59$ e $p - valor = P(Q \geq 9,59|H_0) = 0,4772, Q \sim \chi^2_{10}$. Assim, não rejeitamos a hipótese de distribuição espacial aleatória.

Voltemos ao Exemplo 3: estudo sobre a inclinação (identificação) partidária estadunidense

- Tabela de contingência (2×2) com os resultados da pesquisa.

		Inclinação partidária		
		Democrata	Republicano	Total
Gênero	Feminino	762	468	1230
	Masculino	484	477	961
Total	-	1246	945	2191

- Pergunta: as proporções de pessoas para cada inclinação partidária é a mesma entre os gêneros?

Produto de binomiais (condicionalmente) independentes

- A tabela anterior é uma realização (amostra) possível, oriunda da seguinte estrutura:

		Inclinação partidária		
		Democrata	Republicano	Total
Gênero	Feminino	$N_{11}(\theta_{11})$	$N_{12}(\theta_{12})$	$n_{1.} = 1230$
	Masculino	$N_{21}(\theta_{21})$	$N_{22}(\theta_{22})$	$n_{2.} = 961$
Total	-	$N_{.1}$	$N_{.2}$	$n_{..} = 2191$

Exemplo 3 (cont.)

- Já vimos que, nesse caso, as hipóteses de homogeneidade e independência são equivalentes.
- Há outras formas de se quantificar (testar) a dependência.
- Chances: $\lambda_1 = \frac{\theta_{11}}{1-\theta_{11}}$ e $\lambda_2 = \frac{\theta_{21}}{1-\theta_{21}}$.
- λ_1 quantifica o quão mais ($\lambda > 1$) ou menos ($\lambda < 1$) provável é um eleitor do gênero feminino ter uma inclinação "democrata" em relação à ter uma inclinação "republicana".
- Analogamente, para λ_2 (gênero masculino). Note que $\lambda_i \in (0, \infty), i = 1, 2$.

Exemplo 3 (cont.)

- Razão de chances:

$$\pi = \frac{\lambda_1}{\lambda_2} = \frac{\frac{\theta_{11}}{1-\theta_{11}}}{\frac{\theta_{21}}{1-\theta_{21}}}, \pi \in (0, \infty).$$

- Quantifica o quão maior ($\pi > 1$) ou menor ($\pi < 1$) é a chance de um eleitor do gênero feminino ter uma inclinação "democrata" em relação à ter uma inclinação "republicana", comparado com a equivalente chance para o gênero masculino.
- Podemos provar que $\theta_{11} = \theta_{21}$ (independência) $\leftrightarrow \pi = 1$ (exercício).

Exemplo 3 (cont.)

- Podemos, então, verificar (e quantificar) a existência de dependência testando as hipóteses $H_0 : \pi = 1$ vs $H_1 : \pi \neq 1$.
- Equivalentemente, podemos testar $H_0 : \eta = \ln \pi = 0$ vs $H_1 : \eta = \ln \pi \neq 0$.
- Temos que o estimador de máxima verossimilhança de η é dado por

$$\widehat{\eta} = \ln \widehat{\pi} = \ln \left(\frac{\frac{\widehat{\theta}_{11}}{1-\widehat{\theta}_{11}}}{\frac{\widehat{\theta}_{21}}{1-\widehat{\theta}_{21}}} \right) = \ln \left(\frac{N_{11} N_{22}}{N_{12} N_{21}} \right) = \ln N_{11} + \ln N_{22} - \ln N_{12} - \ln N_{21},$$

em que $\widehat{\theta}_{i1} = \frac{N_{i1}}{n_{i.}}$, $i = 1, 2$, devido à propriedade da invariância dos estimadores de MV.

Exemplo 3 (cont.)

- A distribuição assintótica de $\widehat{\eta}$ se aproxima mais de uma distribuição normal do que a distribuição assintótica de $\widehat{\pi}$, para um mesmo conjunto de dados.
- Isso ocorre, essencialmente, porque $\widehat{\eta} \in (-\infty, \infty)$ enquanto que $\widehat{\pi} \in (0, \infty)$. Além disso, a distribuição de $\widehat{\eta}$ é menos assimétrica do que a distribuição de $\widehat{\pi}$.
- Para $n_{i.}, i = 1, 2$ suficientemente grandes, temos que $\eta \approx N(\eta, \sigma^2_\eta)$, em que $\sigma^2_\eta = \frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}$ (é a estimativa de máxima verossimilhança da variância assintótica de η).

Exemplo 3 (cont.) Metodologias assintóticas

- Portanto, um $IC(\eta, \gamma) = [\widehat{\eta} - z_{\frac{1-\gamma}{2}}\sigma_\eta; \widehat{\eta} + z_{\frac{1-\gamma}{2}}\sigma_\eta]$, em que $P(Z \geq z_{\frac{1-\gamma}{2}}) = \frac{1-\gamma}{2}$ e $\sigma_\eta = \sqrt{\sigma^2_\eta}$.
- Um teste para testar $H_0 : \eta = \eta_0$ vs $H_1 : \eta \neq \eta_0$ é, rejeitar H_0 se $p - valor \leq \alpha$, em que $p - valor = 2P(Z \geq |z_t||H_0)$, em que z_t é o valor calculado da estatística

$$Z_t = \frac{\widehat{\eta} - \eta_0}{\sigma_\eta}$$

e $Z \sim N(0, 1)$.

- Também podemos obter uma aproximação numérica da distribuição de $\widehat{\eta}$ por reamostragem.

Exemplo 3 (cont.)

- Voltando ao exemplo, temos: $\widetilde{\eta} = \ln\left(\frac{n_{11}}{n_{12}} / \frac{n_{21}}{n_{22}}\right) = \ln(n_{11}) + \ln(n_{22}) - \ln(n_{12}) - ln(n_{21}) = 0,473$ e $\sigma_\eta = 0,087$.
- Também ,$IC(\eta, 0,95) = [0,302; 0,644]$ e p-valor $< 0,0001$ (associado ao teste de nulidade de η, como visto anteriormente).
- Além disso, $IC(\pi, 0,95) = [e^{0,302}; e^{0,644}] = [1,353; 1,904]$.
- Logo, como esperado, rejeitamos a hipótese de independência entre gênero e inclinação partidária.
- A função "oddsratio" do pacote "vcd" estima a razão de chances, o erro-padrão assintótico e executa o teste apresentado anteriormente.

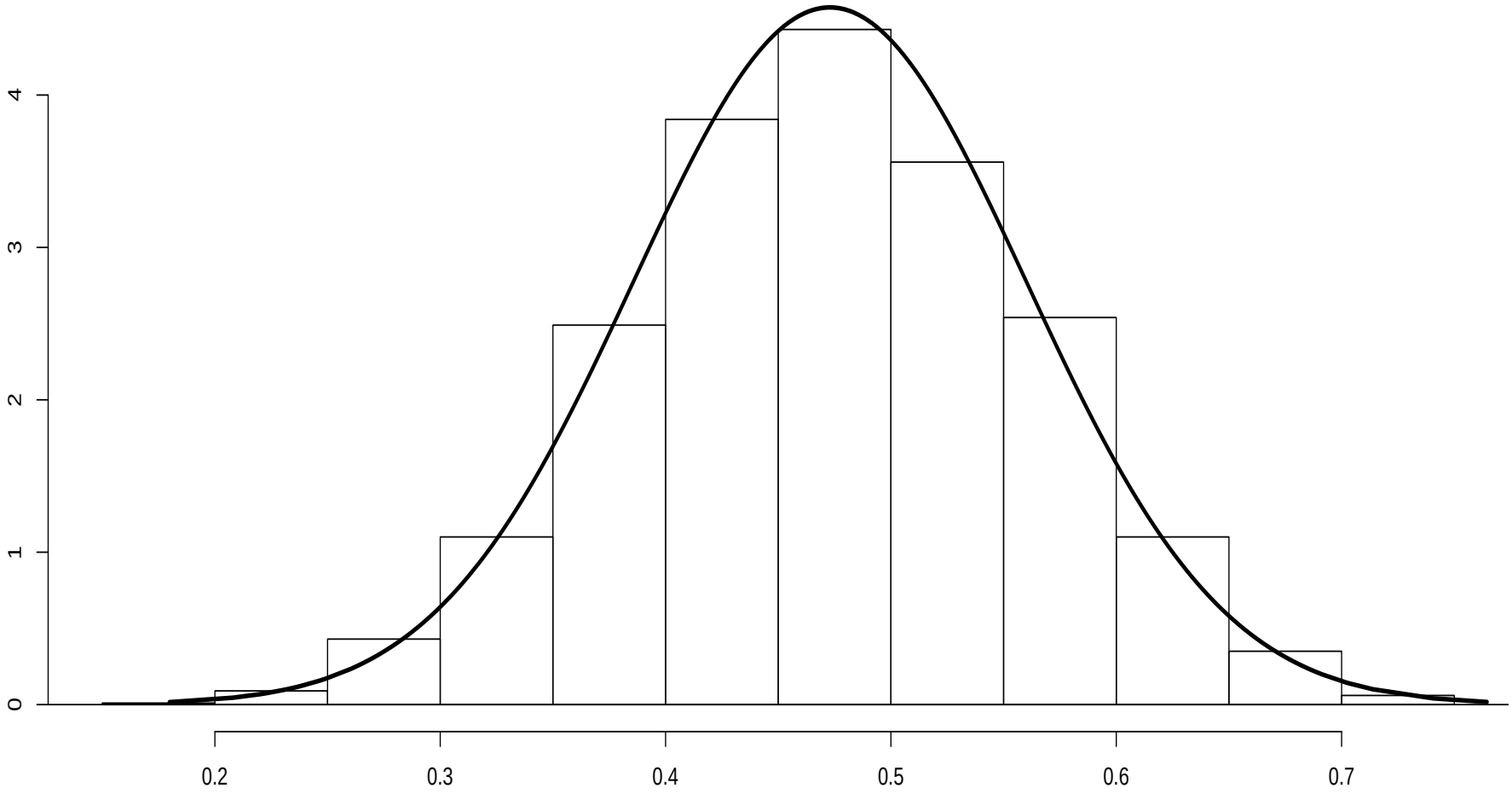
Um procedimento para se obter uma aproximação numérica da distribuição exata de $\widehat{\eta}$

- Estime os parâmetros associados ao modelo suposto gerador da tabela de contingência utilizando o método de MV.
- Para b=1,...,B execute os seguintes passos
  1. Gere uma tabela de contingência sob o modelo em questão, utilizando as estimativas calculadas anteriormente.
  2. Obtenha a estimativa de MV η.
- Ao final teremos uma amostra aleatória da distribuição exata de $\widehat{\eta}$ (ou seja, uma aproximação numérica).

Cont.

- Com essa amostra podemos construir um histograma, intervalos de confiança e estimar o poder do teste anteriormente apresentado (para isso temos que calcular a estatística do teste Z_t além da estimativa de η).
- Se quisermos obter uma aproximação da distribuição exata da estatística do teste sob H_0 e calcular o respectivo p-valor, devemos, além de calcular a estatística Z_t no passo 2, estimar os parâmetros e gerar a tabela de contingência, sob H_0 (no passo 1).

Histograma da distribuição exata obtida via simulação



Resultados numéricos

- $\sigma_{\boldsymbol{\eta}} = 0,085$, $IC(\eta, 0,95) = [0,302; 0,632]$.
- p-valor $< 0,0001$.
- Neste caso, a aproximação assintótica mostrou-se bastante apropriada.

Comentários

- Os resultados podem ser estendidos para tabelas $(2 \times s)$ e $(r \times s)$.
- No primeiro caso, teremos $(s-1)$ razões de chances.
- No segundo caso, teremos $\begin{pmatrix} r \\ 2 \end{pmatrix} \times (s-1)$ razões de chances.
- As definições anteriores permanecem, essencialmente, as mesmas.
- Chance: $\lambda_{ij} = \frac{\theta_{ij}}{1-\theta_{ij}}$.
- Razão de chances $\pi_{ilj} = \lambda_{ij}/\lambda_{lj}$.
- Pesquisar!

Tabela de contingência $r \times s$: produto de multinomiais independentes

		Variável 1 (resposta)					
		C_{11}	C_{12}	...	$C_{1(s-1)}$	C_{1s}	Total
Variável 2	C_{21}	$N_{11}(\theta_{11})$	$N_{12}(\theta_{12})$	...	$N_{1(s-1)}(\theta_{1(s-1)})$	$N_{1s}(\theta_{1s})$	$n_{1.}$
(explicativa)	C_{22}	$N_{21}(\theta_{21})$	$N_{22}(\theta_{22})$	...	$N_{1(s-1)}(\theta_{2(s-1)})$	$N_{2s}(\theta_{2s})$	$n_{2.}$
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	
	C_{2r}	$N_{r1}(\theta_{r1})$	$N_{r2}(\theta_{r2})$	...	$N_{r(s-1)}(\theta_{r(s-1)})$	$N_{rs}(\theta_{rs})$	$n_{r.}$
Total	-	$N_{.1}$	$N_{.2}$	...	$N_{.(s-1)}$	$N_{.s}$	$n_{..}$

Outras medidas de associação

- Existem famílias de medidas de associação para tabelas de contingência ($r \times s$) (multinomiais e produtos de multinomiais).
- Em geral, elas são baseadas na estatística de Pearson (qui-quadrado): $Q_H = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(N_{ij} - E_{ij})^2}{E_{ij}}$.
- A idéia é construir estatísticas com suporte limitado (intervalo (0,a), a $>$0), de tal forma que quanto maior/menor seu valor, maior/menor o grau de dependência.
- A fórmula geral é $M = Q_H/T$, em que T é algum limitante superior para Q_H. Assim, quanto mais próximo de zero for o valor de M menor será a magnitude da associação e quanto mais próximo de T, maior será a magnitude dessa associação.

Outras medidas de associação (cont.)

- Lembrando:
  - Q_H : estatística qui-quadrado.
  - $n_{..}$: número total de observações.
  - r : número total de linhas.
  - s : número total de colunas
- Coeficiente Phi: $\Phi = \sqrt{\frac{Q_H}{n_{..}}}$.
- Coeficiente de Cramer V: $V = \sqrt{\frac{\Phi^2}{min(r,s)}}$.
- Coeficiente de contingência de Pearson: $C = \sqrt{\frac{Q_H}{Q_H+n_{..}}}$.
- Coeficiente T de Tschuprow: $\sqrt{\frac{\Phi^2}{(r-1)(s-1)}}$.
- Os limites superiores para esses coeficientes podem depender dos valores de s, r e $n_{..}$ (não, necessariamente, são iguais à 1).

Comentários

- As medidas anteriores são apropriadas quando ambas as variáveis são nominais (ou quando pelo menos uma é nominal), embora possam ser utilizadas quando ambas forem ordinais se o interesse é medir associação.
- O coeficiente Φ não é muito apropriado para tabelas maiores do que 2×2. As outras não tem limitações quanto à isso.
- Quase sempre é difícil avaliar a magnitude de tais medidas considerando apenas seu valor numérico.
- O mais apropriado é comparar o valor obtido pela tabela observada com os valores oriundos obtidas de tabelas geradas sob H_0.

Um procedimento de quantificação (numérica) da magnitude dos coeficientes

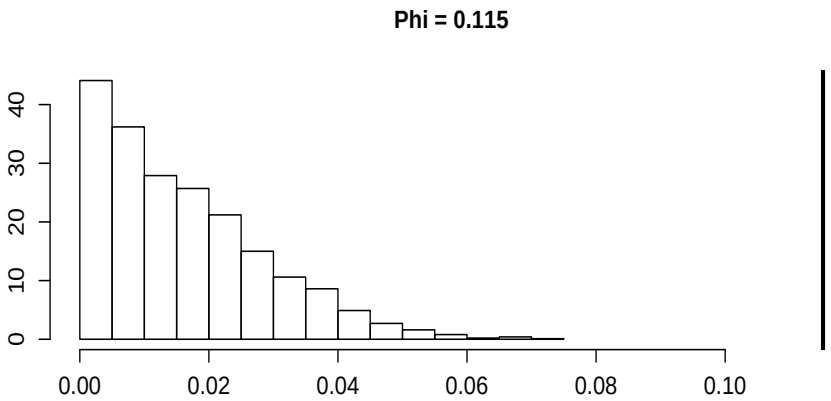
- Calcule os coeficientes de associação com base na tabela observada.
- Estime os parâmetros associados ao modelo suposto gerador da tabela de contingência (sob H_0, independência) utilizando o método de MV (por exemplo).
- Para b=1,...,B execute os seguintes passos
  1. Gere uma tabela de contingência sob o modelo em questão, utilizando as estimativas calculadas anteriormente.
  2. Calcule os coeficientes de associação com base na tabela simulada.

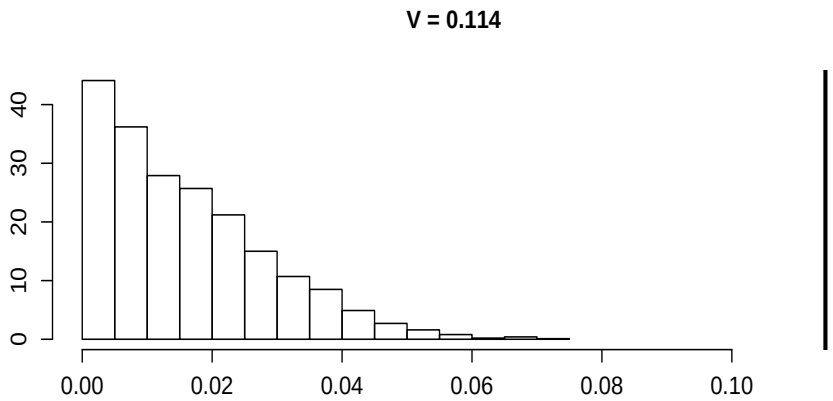
Um procedimento de quantificação (numérica) da magnitude dos coeficientes (cont.)

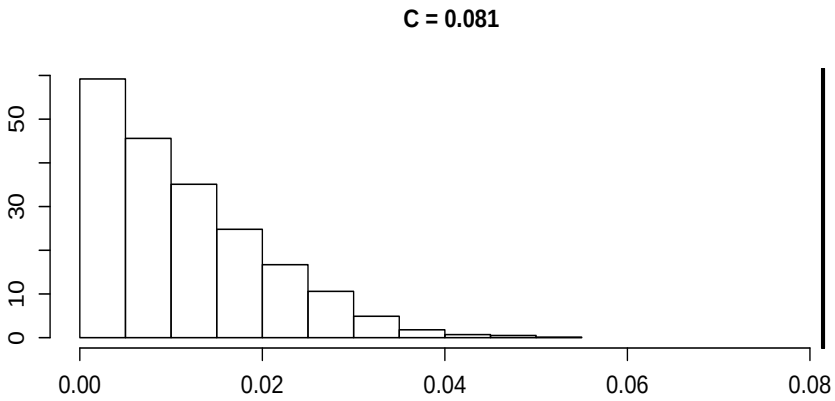
- Ao final teremos uma amostra aleatória da distribuição exata dos coeficientes.
- Assim, quanto maior for a proporção de valores simulados menores que a estimativa calculada através da tabela observada, maior será a magnitude do coeficiente e, consequentemente, maior será a magnitude da associcação.
- Pode-se calcular p-valores para hipóteses de interesse.

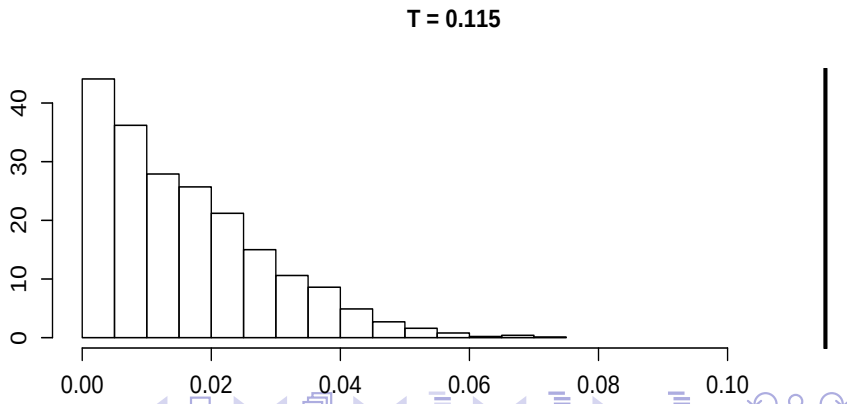
Histograma das distribuições exatas dos coeficientes (sob H_0) obtidas via simulação

(exemplo da inclinação partidária)

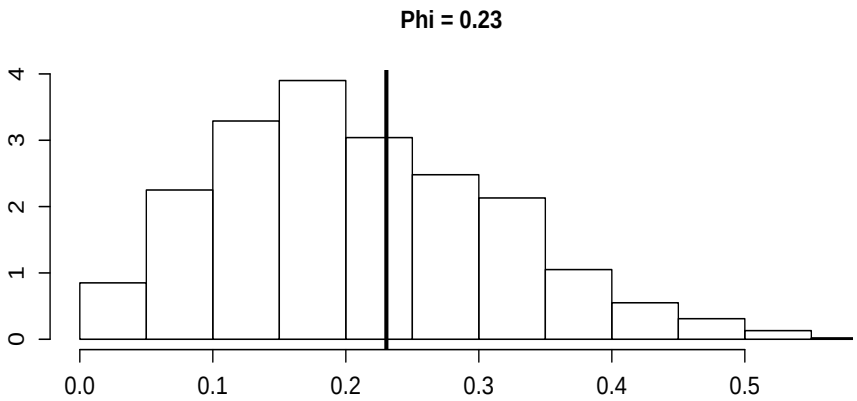


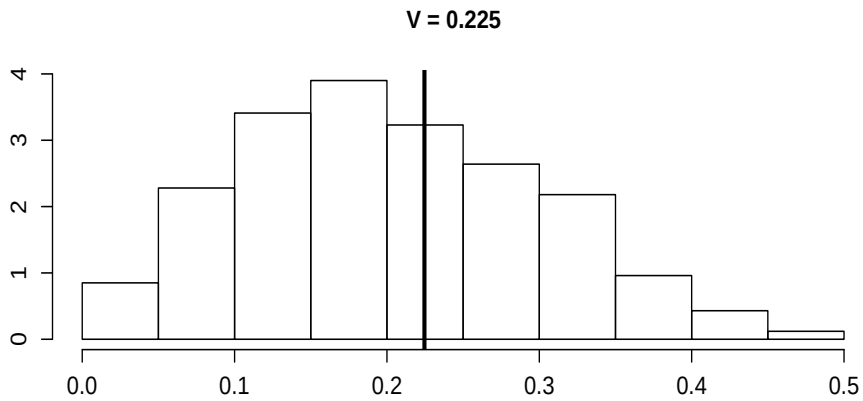


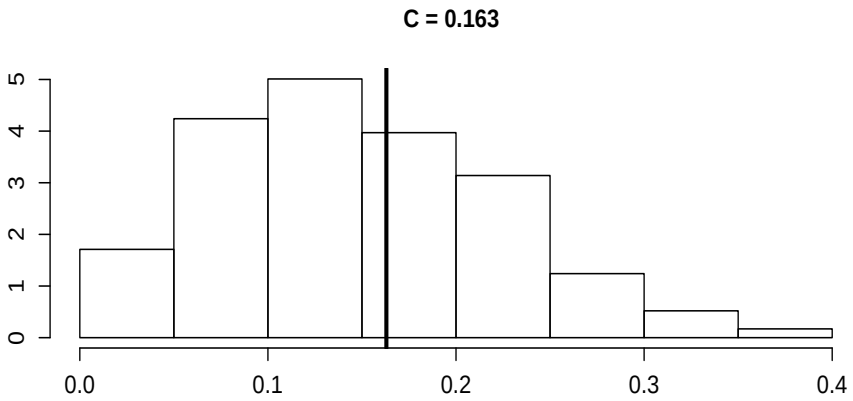


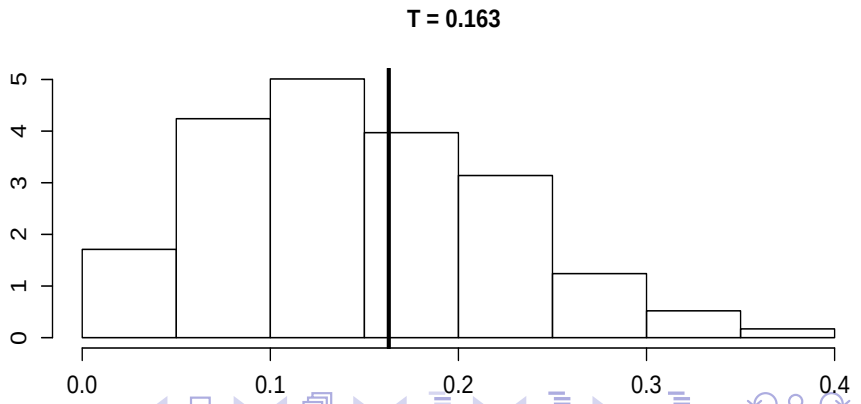


Histogramas das referidas distribuições (exemplo do estudo do estado civil com grau de instrução) (a independência não foi rejeitada)









Voltando ao Exemplo 1: comparação de métodos de detecção de cárie

		Risco de cárie segundo o método convencional			
		Baixo	Médio	Alto	Total
Risco de cárie segundo o método simplificado	Baixo	11	5	0	16
	Médio	14	34	7	55
	Alto	2	13	11	26
Total	-	27	52	18	97

Queremos verificar o grau de concordância (plena) entre os métodos.

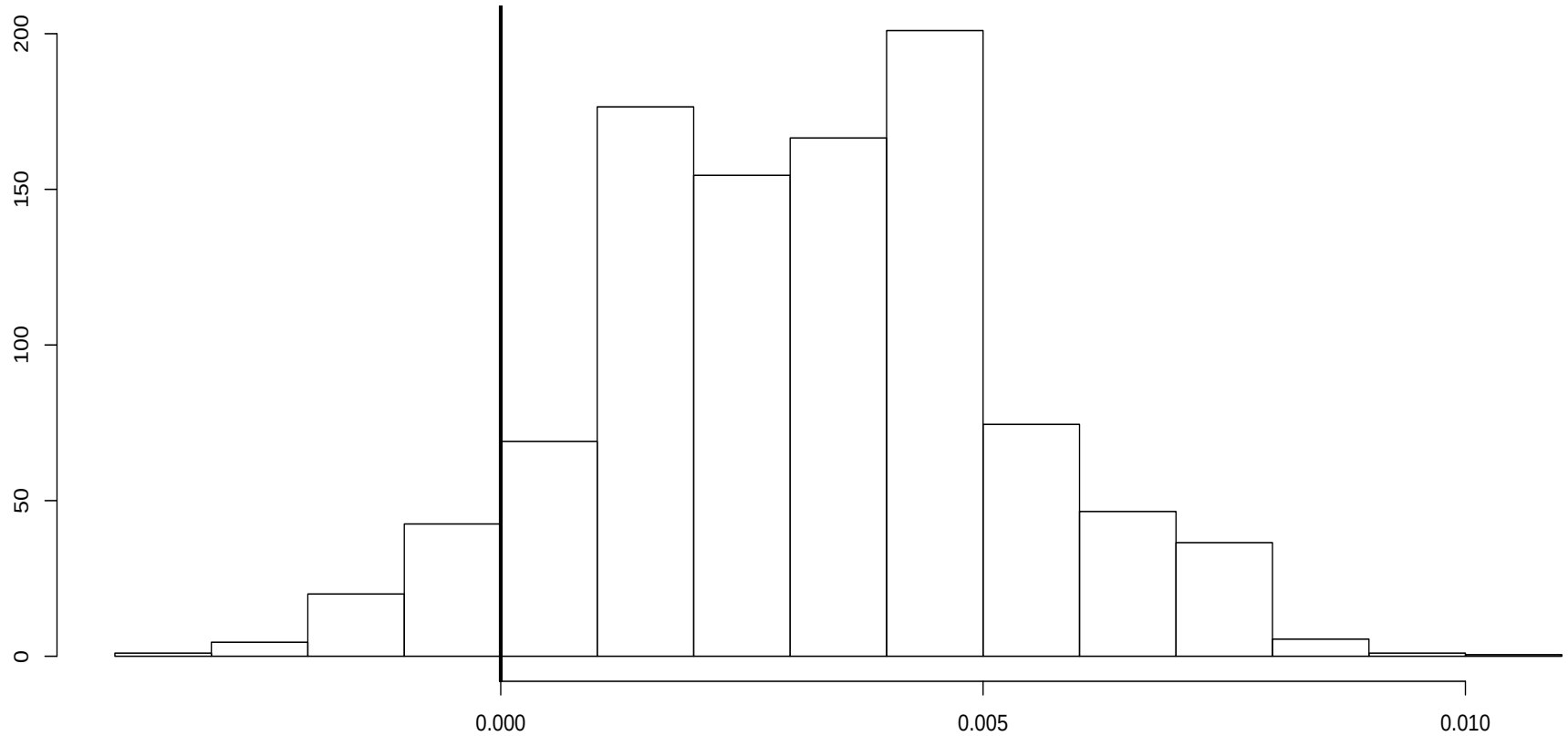
Medidas para variáveis ordinais

- Quando ambas as variáveis são ordinais, outras medidas podem ser mais apropriadas, principalmente dependendo das hipóteses de interesse.
- Em geral, nesses casos, está-se mais interessado em medir concordância do que dependência, embora tais conceitos possam estar relacionados, como já vimos.
- A idéia é comparar a quantidade de observações concordantes com as discordantes.

Medidas para variáveis ordinais

- Defina
  - C: número de pares concordantes.
  - D: número de pares discordantes.
- Coeficiente τ-b de Kendall : $\tau_b = \frac{C-D}{n_{..}(n_{..}-1)/2}$.
- Coeficiente τ-c de Kendall: $\tau_c = \frac{C-D}{n_{..}^2(min(r,s)-1)/(2min(r,s))}$.
- Podemos usar um algoritmo semelhante ao caso anterior, mas agora obtendo as ditribuições dos coeficientes acima sem restringir à H_0.

Histograma das distribuição exata do coeficiente τ_b obtidas via simulação $IC(\tau_b, 0,95) = [-0,001; 0,007]$



Comentários

- Pelo comportamento do histograma e do intervalo de confiança, temos indícios de que a concordância plena é praticamente nula.
- No entanto, podem existir outros padrões de concordância (p.e., concordância marginal).
- Os coeficientes τ_b e τ_c são mais apropriados para tabelas quadradas e não quadradas, respectivamente.