Análise de Dados Categorizados - Aula4

Márcia D Elia Branco

Universidade de São Paulo
Instituto de Matemática e Estatística
www.ime.usp.br/∼mbranco - mdbranco@usp.br

Inferência em Tabelas 2×2

Covariável X	Resposta Y		Totais
	j=1	j=2	
i=1	n_{11}	n_{12}	n_{1+}
i=2	n_{21}	n_{22}	n_{2+}
Totais	n_{+1}	n_{+2}	n

- Na última aula vimos algumas medidas de associação: Risco Atibuivel, Risco Relativo e Razão de Chances.
- Vamos estudar hoje Testes de Hipóteses baseados na aproximação qui-quadrado e o Teste Exato de Fisher.
- Também veremos, como usar a abordagem bayesiana para fazer inferência em tabelas 2×2.

Testes Qui-quadrado

Considere o modelo produto de binomiais com os totais das linhas previamente fixados pelo plano amostral.

Neste caso, o interesse é testar a hipótese de homogeneidade

$$H_0 : p_{(1)1} = p_{(2)1} \quad \textit{versus} \quad H_a : p_{(1)1} \neq p_{(2)1}$$

A estatística de Pearson é dada por

$$Q_P = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(n_{ij} - e_{ij})^2}{e_{ij}}$$

em que

$$e_{ij} = \frac{n_{i+} n_{+j}}{n}$$

Teste da Razão de Verossimilhança para homogeneidade.

A função de verossimilhança

$$L(p_{(1)1}, p_{(2)1}) \propto p_{(1)1}^{n_{11}} p_{(1)2}^{n_{12}} p_{(2)1}^{n_{21}} p_{(2)2}^{n_{22}}$$

Sob a hipótese alternativa, o ponto de máximo é o emv sem restrição. Assim, a verossimilhança no ponto de máximo é

$$L_1 \propto \left(\frac{n_{11}}{n_{1+}}\right)^{n_{11}} \left(\frac{n_{12}}{n_{1+}}\right)^{n_{12}} \left(\frac{n_{21}}{n_{2+}}\right)^{n_{21}} \left(\frac{n_{22}}{n_{2+}}\right)^{n_{22}}$$

Sob a hipótese nula, o ponto de máximo é obtido maximizando-se a função restrita a H_0.
Considere $p_{+1} = p_{(1)1} = p_{(2)1}$, temos a função de verossimilhança

$$L_0(p_{+1}) \propto p_{+1}^{n_{11}+n_{21}}(1-p_{+1})^{n_{12}+n_{22}}$$

Derivando a log-verossimilhança e igualando a zero obtemos o ponto de máximo $\hat{p}_{+1} = \frac{n_{+1}}{n}$.

Então

$$L_0 \propto \left(\frac{n_{+1}}{n}\right)^{n_{+1}} \left(\frac{n_{+2}}{n}\right)^{n_{+2}}$$

Usando a notação $e_{ij} = \frac{n_{i+}n_{+j}}{n}$, resulta em

$$\frac{L_1}{L_0} = \left(\frac{n_{11}}{e_{11}}\right)^{n_{11}} \left(\frac{n_{12}}{e_{12}}\right)^{n_{12}} \left(\frac{n_{21}}{e_{21}}\right)^{n_{21}} \left(\frac{n_{22}}{e_{22}}\right)^{n_{22}}.$$

Portanto, o valor da estatística do teste de RV é dada por:

$$2log(\frac{L_1}{L_0}) = 2\sum_{i=1}^{2}\sum_{j=1}^{2} n_{ij} \log\left(\frac{n_{ij}}{e_{ij}}\right)$$

Os graus de liberdades associados a distribuição assintótica dessa estatística são $\nu = 2 - 1 = 1$.

Para os dados do exemplo 2 (ultima aula) vamos testar a homogeneidade entre as populações de indivíduos com câncer e controle.

$$H_0 : p_{1(1)} = p_{1(2)} \quad versus \quad H_a : p_{1(1)} \neq p_{1(2)}$$

Usando $\alpha = 0.05$ e $\nu = 1$ a região critica é $RC = \{Q_{RV} > 3.84\}$.

$$Q_{RV} = 2\left[688\log(\frac{688}{669}) + 650\log(\frac{650}{669}) + 21\log(\frac{21}{40}) + 59\log(\frac{59}{40})\right]$$

$$Q_{RV} = 19.88 \in RC \qquad Valor - P < 10^{-5}$$

Rejeita-se a hipótese de homogeneidade.

Exemplo 3: Durante 18 semanas de determinado ano foi contado o número de acidentes de carros registrados na Suécia, avaliando-se o tipo de estrada e o fato de haver ou não um limite de velocidade. O objetivo é avaliar se o limite de velocidade influencia de maneira diferente o número de acidentes dependendo do tipo de estrada.

Limite de velocidade	Tipo de estrada		Totais
	Auto-estrada	Outra	
Sim	8	42	50
Não	57	106	163
Totais	65	148	213

Vamos realizar um teste de multiplicatividade considerando o modelo produto de Poisson.

Hipóteses :

$$H_0 : \mu_{11} = \frac{\mu_{1+}\mu_{+1}}{\mu}; \mu_{12} = \frac{\mu_{1+}\mu_{+2}}{\mu}; \mu_{21} = \frac{\mu_{2+}\mu_{+1}}{\mu}; \mu_{22} = \frac{\mu_{2+}\mu_{+2}}{\mu}$$

H_a : Existe pelo menos uma diferente.

- É possível usar a estatística de Pearson Q_P .
- Alternativamente, podemos usar a estatística baseada na Razão de Verossimilhança Q_{RV}.
- É possivel mostrar que $Q_{RV} = 2 \sum_{i=1}^{2} \sum_{j=1}^{2} n_{ij} \log(\frac{n_{ij}}{e_{ij}})$ (exercício).

  Com graus de liberdades $\nu = 4 - 3 = 1$.

Regiões Criticas dos testes :

$RC = \{Q_{RV} > 3.84\}$ e $RC = \{Q_P > 3.84\}$.

Valores esperados das caselas são $e_{11} = 15.26$, $e_{12} = 34.74$, $e_{21} = 49.74$ e $e_{22} = 113.26$.

Resultando em $Q_P = 6.49$ e $Q_{RV} = 7.09$.

Em ambos os casos rejeita-se a hipótese de multiplicidade e conclui-se pela dependência entre as variáveis. O limite de velocidade influencia de forma diferente o número de acidentes dependendo do tipo de estrada.

Teste Exato de Fisher

- O Teste exato de Fisher é um teste não paramétrico usado para pequenas amostras.
- A hipótese nula consiste em não associação entre as variáveis. A hipótese alternativa pode ser unilateral ou bilateral.
- Para construção do teste ambas marginais da tabelas são supostas conhecidas.
- É preciso calcular as probabilidades associada a ocorrência da tabela observada e de outras tabelas "mais extremas"do que essa.

Teste Exato de Fisher

Suponha que $N_{11} \mid n_{1+} \sim Binomial(n_{1+}, p_{(1)1})$ e $N_{21} \mid n_{2+} \sim Binomial(n_{2+}, p_{(2)1})$, independentes.

Se condicionarmos aos totais das colunas, temos que :

(i) apenas uma variável é livre, a segunda fica totalmente determinada pelo conhecimento da primeira.
(ii) a distribuição condicional $N_{11} \mid n_{1+}, n_{+1}, n$ é uma hipergeométrica.
(ii) a probabilidade de ocorrer o resultado de uma particular tabela é equivalente a $P(N_{11} = n_{11})$ e dada por

$$\frac{C_{n_{11}}^{n1+} \times C_{n_{+1}-n_{11}}^{n-n_{1+}}}{C_{n_{+1}^{n}}}$$

Teste Exato de Fisher

Exemplo 4: O proplema proposto por Fisher em 1935 consiste em comprovar ou refutar a afirmação feita por uma senhora que diz ser capaz de diferenciar, pelo paladar, a ordem que foi adicionado o leite à sua xicará de chá.
O experimento consistiu em considerar 8 xícaras de chá. Em 4 delas o leite foi colocado primeiro e nas outras 4, o chá foi colocado antes do leite. Antes da senhora fazer as provas, foi informado a ela que haviam 4 xicaras de cada tipo.

Ordem correta	Resposta		Totais
	Leite	Chá	
Leite	3	1	4
Chá	1	3	4
Totais	4	4	8

Teste Exato de Fisher

H_0 : Não há associação entre Resposta e Ordem Correta.
H_a : Existe associação e o sentido desta é de que a senhora tem a sensibilidade anunciada.

Quais seriam as tabelas mais extremas no sentido de H_a ?

Ordem correta	Resposta		Totais
	Leite	Chá	
Leite	4	0	4
Chá	0	4	4
Totais	4	4	8

Teste Exato de Fisher

Calculando as probabilidades das duas tabelas

$$P(N_{11} = 3) = \frac{C_3^4 \times C_1^4}{C_4^8} = 0.229$$

$$P(N_{11} = 4) = \frac{C_4^4 \times C_0^4}{C_4^8} = 0.014$$

O valor-P associado ao teste é dado por 0.243 . Não rejeita-se H_0. Conclusão: Não há evidências de que a Senhora tenha a sensibilidade anunciada.

- Para conduzir um teste bilateral, deve-se calcular as probabilidades de tabelas mais extremas nos dois sentidos.
- Se iniciarmos com o modelo Multinomial ou Produto de Poisson, também é possível mostrar que condicionando aos totais marginais obtemos a mesma distribuição hipergeométrica.

Covariável X	Resposta Y		Totais
	j=1	j=2	
i=1	n_{11}	n_{12}	n_{1+}
i=2	n_{21}	n_{22}	n_{2+}
Totais	n_{+1}	n_{+2}	n

- Intervalos de credibilidade $1 - \alpha$ são obtidos usando os quantis da distribuição *a posteriori*.
- Existem dois tipos de IC: caudas iguais e HPD .
- Testes de hipóteses podem ser feitos considerando-se as probabilidades das hipóteses serem verdadeiras ou utilizando os Intervalos de Credibilidade.

Intervalo de credibilidade de caudas iguais de probabilidade $1-\alpha$ para $d = p_{(1)1} - p_{(2)1}$

Considere as distribuições *a priori* : $p_{(1)1} \sim Beta(a_1, b_1)$ e $p_{(2)1} \sim Beta(a_2, b_2)$ independentes.

Usando a fórmula de Bayes, obtemos que as distribuições *a posteriori* também são Betas independentes com parâmetros:

$$A_1 = a_1 + n_{11} \ , \ B_1 = b_1 + n_{12} \ , \ A_2 = a_2 + n_{21} \ , \ B_2 = b_2 + n_{22}.$$

Para construção do IC para d precisamos da sua distribuição *a posteriori* . Não conseguimos obter uma distribuição de probabilidades conhecida para diferença de Betas.

- A distribuição *a posteriori* $f(d \mid n_{11}, n_{21})$ pode ser aproximada via simulação. Método de Monte Carlo.
- O método consiste em simular de cada uma das Betas de forma independente. Para cada par de valores simulados, obter o valor de d.
- Se simularmos uma grande quantidade de valores, as estatísticas amostrais devem se aproximar dos parâmetros dessa distribuição.
- Usamos os quantis da amostra de Monte Carlo para aproximar os quantis populacionais de ordem $\alpha/2$ e $1 - \alpha/2$ e obter o IC aproximado.

Exemplo 1: O interesse é comparar dois vermífugos. Modelo Produto de Binomiais.

Vermífugo	Verminose		Totais
	Sim	Não	
1	48	152	200
2	68	132	200
Totais	116	284	400

$$\hat{d} = \frac{48}{200} - \frac{68}{200} = -0.10 \quad \text{e} \quad E[d \mid n_{11}, n_{22}] = \frac{49}{202} - \frac{69}{202} \approx -0.10$$

As distribuições *a posteriori* são $p_{(1)1} \mid n_{11} \sim Beta(49, 153)$ e $p_{(2)1} \mid n_{21} \sim Beta(69, 133)$, usando distribuições *a priori* uniformes.

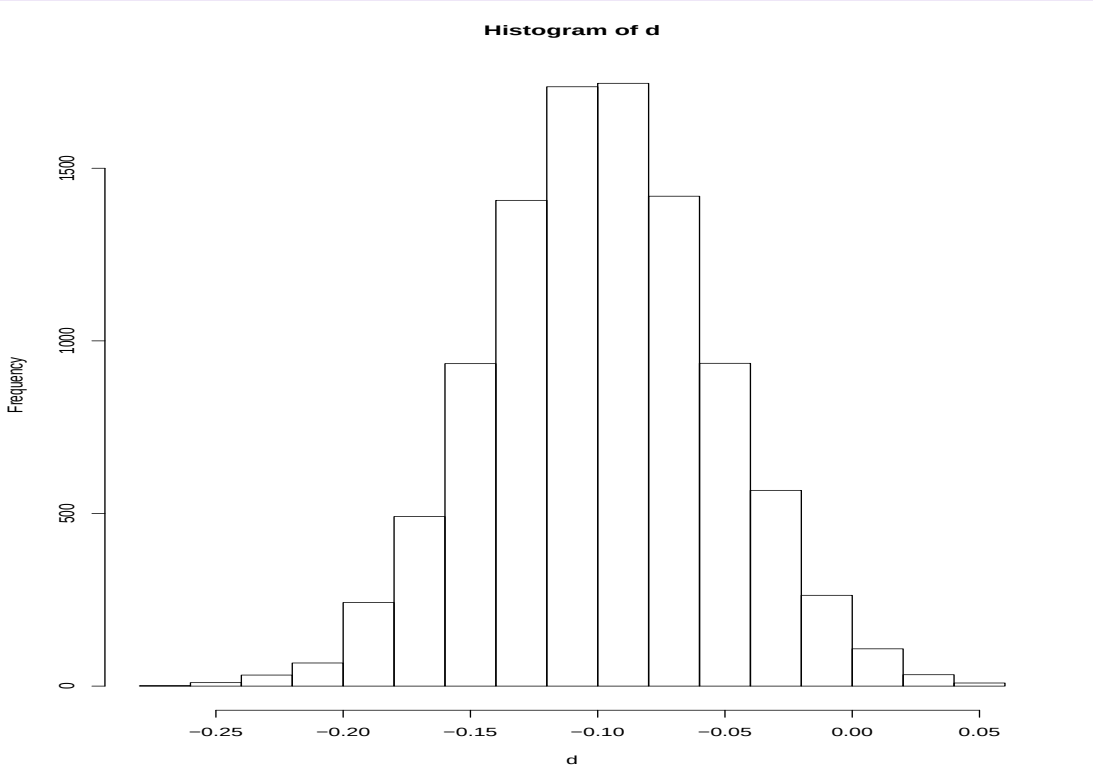
Intervalo de probabilidade 0.90 para d é $[-0.172; -0.024]$.

$P(p_{(1)1} > p_{(2)1} \mid n_{11}, n_{21}) \approx 0.015.$

O codigos em **R** usados foram:

$> p1 = rbeta(10000, 49, 153)$
$> p2 = rbeta(10000, 69, 133)$
$> d = p1 - p2$
$> quantile(d, c(0.05, 0.95))$
$> mean(p1 > p2)$

Histograma dos valores simulados (M=10000) da distribuição *a posteriori* de $d = p_{(1)1} - p_{(2)1}$.



Exemplo 2: O proplema proposto por Fisher em 1935 consiste em comprovar ou refutar a afirmação feita por uma senhora que diz ser capaz de diferenciar, pelo paladar, a ordem que foi adicionado o leite à sua xicará de chá.
O experimento consistiu em considerar 8 xícaras de chá. Em 4 delas o leite foi colocado primeiro e nas outras 4, o chá foi colocado antes do leite. Antes da senhora fazer as provas, foi informado a ela que haviam 4 xicaras de cada tipo.

Ordem correta	Resposta		Totais
	Leite	Chá	
Leite	3	1	4
Chá	1	3	4
Totais	4	4	8

(a) Construir o intervalo de credibilidade de 0.90 para a diferença $d = p_{(1)1} - p_{(2)1}$.

(b) Testar as hipóteses $H_0 : p_{(1)1} \leq p_{(2)1}$ $\times$ $H_a : p_{(1)1} > p_{(2)1}$.

Solução:

Usando simulação, temos

(a) (-0.096 , 0.716)

(b) $P(p_{(1)1} \leq p_{(2)1}) = 0.097$

Note que a probabilidade de H_a ser verdadeira é 0.903, evidência em favor da suposição de que a senhora tem a sensibilidade indicada.

Distribuição *a posteriori* **aproximada para** *OR* **.**

- A IB também tem uma teoria para grandes amostras.
- Mas, diferente da Clássica, a aproximação normal não é obtida para a distribuição do estimador e sim para a distribuição *a posteriori* do parâmetro θ.
- Sob condições de regularidade $f(\theta \mid x)$ é aproximada por uma $N(Mo, V)$ em que Mo é a moda da posteriori e V é o negativo da segunda derivada da log posteriori no ponto Mo.
- Inicialmente, vamos obter a aproximação para o logaritmo da chance $\theta = \log\left(\frac{\pi}{1-\pi}\right)$.

A função de verossimilhança associada a uma amostra da binomial é proporcional à

$$\pi^x(1-\pi)^{n-x}$$

Vamos considerar *a priori* $Beta(a, b)$ então *a posteriori* é proporcional a

$$\pi^{x+a-1}(1-\pi)^{n-x+b-1}.$$

Logo, $\pi \mid x \sim Beta(A, B)$ com $A = x + a$ e $B = n - x + b$.

No entanto, o nossa parâmetro de interesse é $\theta = \log\left(\frac{\pi}{1-\pi}\right)$.

Para obter a distribuição de θ podemos usar o método Jacobiano de transformação de variáveis.

Fazendo a transformação inversa temos que:

$$\pi = \frac{e^{\theta}}{1+e^{\theta}} \text{ e } \; 1-\pi = \frac{1}{1+e^{\theta}}.$$

O Jacobiano da transformação é

$$\frac{d\pi}{d\theta} = \frac{e^{\theta}}{(1+e^{\theta})^2}.$$

Assim

$$f(\theta \mid x) \propto \frac{e^{\theta}}{(1+e^{\theta})^2} \left[\frac{e^{\theta}}{(1+e^{\theta})}\right]^{A-1} \left[\frac{1}{(1+e^{\theta})}\right]^{B-1}$$

e

$$\log f(\theta \mid x) = C + A\theta + (A+B)\log(1+e^{\theta}).$$

Derivando $\log f(\theta \mid x)$ e igualando a zero, obtemos

$$Mo = \log\left(\frac{A}{B}\right) = \log\left(\frac{a+x}{b+n-x}\right).$$

Fazendo a segunda derivada da log-posteriori e substituindo θ por Mo e alterando o sinal, temos

$$V = \frac{1}{A} + \frac{1}{B}.$$

Resulta que

$$\theta \mid x \approx N(Mo, V)$$

Resultado: Sob o modelo produto de binomiais e com prioris Betas independentes, a distribuição *a posteriori* aproximada para $\log(OR)$ é $N(m_{OR}, v_{OR})$ em que

$$m_{OR} = \log\left(\frac{a_1 + n_{11}}{b_1 + n_{12}}\right) - \log\left(\frac{b_2 + n_{22}}{a_2 + n_{21}}\right)$$

$$v_{OR} = \frac{1}{a_1 + n_{11}} + \frac{1}{b_1 + n_{12}} + \frac{1}{a_2 + n_{21}} + \frac{1}{b_2 + n_{22}}$$

Prova: Para mostrar o resultado temos que usar o resultado anterior e o fato que diferença de duas v.a. normais independentes é também normal com a média dada pela diferença das médias e variância pela soma das variâncias.

- Usando IB podemos incluir informações prévias sobre a quantidade de interesse usando uma distribuição *a priori* informativa.

- Para o problema de comparação de duas populações binomiais, uma maneira alternativa de especificar a distribuição *a priori* é considerar uma normal bivariada nas tranformações $logito(\pi) = \log\left(\frac{\pi}{1-\pi}\right)$. Esta abordagem permite incluír a dependência entre $p_{(1)1}$ e $p_{(2)1}$.