Análise de Dados Categóricos
Resposta multicategórica
Seções 3.1 e 3.2

Fernando Lucambio

Departamento de Estatística
Universidade Federal do Paraná

Outubro, 2023

- **3.1 Distribuição multinomial**
- **3.2 Tabelas de contingência $I \times J$**
- **3.2.1 Uma distribuição multinomial**
- **3.2.2 I distribuições multinomiais**
- **3.2.3 Teste de independência**

O Capítulo 3 generaliza isso para uma configuração em que o valor da variável de resposta é escolhido a partir de um conjunto fixo de mais de duas opções.

Por exemplo, as opções de resposta podem ter o formato:

- **1. Escala Likert de cinco níveis - Discordo totalmente, discordo, neutro, concordo ou concordo totalmente,**
- **2. Compostos químicos em experimentos de descoberta de medicamentos - Positivo, bloqueador ou nenhum,**
- **3. Colocação de prateleiras de cereais em uma mercearia - fundo, meio ou topo,**
- **4. Afiliação a partidos políticos canadenses - Conservador, Novo Democrata, Liberal, Bloc Quebecois ou Verde, e**
- **5. Graus de carne bovina — Prime, Choice, Select, Standard, Utility e Commercial.**

Para esses exemplos, algumas respostas são ordinais, por exemplo, escala Likert e outras não, por exemplo, compostos químicos. Investigaremos respostas multicategorias ordinais e nominais, não ordenadas neste capítulo.

Em cada um dos exemplos acima, uma unidade observada se encaixa exatamente em uma categoria. Por exemplo, um composto químico não pode ser tanto positivo quanto bloqueador. Existem outras situações em que uma unidade pode se encaixar simultaneamente em mais de uma categoria, como nas perguntas da pesquisa “escolha todas as que se aplicam”. Investigaremos esses problemas de “resposta múltipla” separadamente na Seção 6.4.

A distribuição de probabilidade multinomial é a extensão da distribuição binomial para situações em que há mais de duas categorias para uma resposta. Seja Y a variável aleatória resposta categórica com níveis $j = 1, \cdots, J$, onde cada categoria tem probabilidade $\pi_j = P(Y = j)$ tal que $\sum_{j=1}^{J} \pi_j = 1$.

Se houverem n tentativas idênticas com respostas $Y_1, \cdots, Y_n$, então podemos definir variáveis aleatórias N_j, $j = 1, \cdots, J$, tal que N_j conte o número de tentativas respondendo com a categoria j. Ou seja, $N_j = \sum_{i=1}^{n} \boldsymbol{I}(Y_i = j)$; onde $\boldsymbol{I}(\cdot) = 1$ quando a condição entre parênteses for verdadeira e = 0 caso contrário.

Seja $n_1, \cdots, n_J$ denotando a contagem de respostas observadas para a categoria j com $\sum_{j=1}^{J} n_j = n$. A função de probabilidade (PMF) para observar um determinado conjunto de contagens $n_1, \cdots, n_J$ é

$$P(N_1 = n_1, \cdots, N_J = n_j) = \frac{n!}{\prod_{j=1}^{J} n_j!} \prod_{j=1}^{J} \pi_j^{n_j},$$

conhecida como distribuição de probabilidade multinomial. Observe que quando $J = 2$, a distribuição simplifica para a distribuição binomial, onde $n_1 = w$, $n_2 = n - w$, $\pi_1 = \pi$ e $\pi_2 = 1 - \pi$ na notação dessa seção. Usamos a estimação por máxima verossimilhança para obter estimativas de $\pi_1, \cdots, \pi_J$. A função de verossimilhança é simplesmente a equação acima e o MLE para cada π_j é $\widehat{\pi}_j = n_j/n$, ou seja, a proporção observada para cada categoria.

A Seção 1.2 discute como as contagens de duas distribuições binomiais podem ser analisadas na forma de uma tabela de contingência 2×2. Expandimos esta discussão agora para permitir mais de duas linhas e/ou colunas.

Assumimos que duas variáveis categóricas, X e Y, são medidas em cada unidade com níveis de $i = 1, \cdots, I$ para X e $j = 1, \cdots, J$ para Y.

Nossa amostra é composta por n unidades classificadas cruzadamente de acordo com seus níveis de X e Y. As contagens de unidades para as quais a combinação $(X = i; Y = j)$ é observada são denotadas por n_{ij} e representam observações das variáveis aleatórias N_{ij}, $i = 1, \cdots, I, j = 1, \cdots, J$.

Nosso objetivo é usar as contagens observadas para fazer inferências sobre a distribuição do N_{ij} e identificar quaisquer relações entre X e Y.

Para isso, podemos criar uma tabela de contingência resumindo essas contagens, sendo X a variável de linha e Y a variável de coluna.

A tabela abaixo fornece a tabela geral de contingência. Observe que usamos um subscrito + para indicar uma soma; por exemplo, $n_{+j} = \sum_{i=1}^{I} n_{ij}$ é o total da coluna j. Vamos denotar n_{++} simplesmente por n.

Tabela de contingência $I \times J$

		Y				
		1	2	$\cdots$	J	Total
X	1	n_{11}	n_{12}	$\cdots$	n_{1J}	n_{1+}
	2	n_{21}	n_{22}	$\cdots$	n_{2J}	n_{2+}
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
	I	n_{I1}	n_{I2}	$\cdots$	n_{IJ}	n_{2+}
	Total	n_{+1}	n_{+2}	$\cdots$	n_{+J}	n

Primeiro, considere a situação em que um tamanho de amostra fixo de n unidades é amostrado de uma grande população. Definimos $\pi_{ij} = P(X = i; Y = j)$.

Esta é a mesma configuração multinomial da Seção 3.1 com n tentativas e $I \times J$ respostas possíveis, exceto que agora cada resposta possível é uma combinação de duas variáveis em vez de apenas uma.

Assumimos que cada unidade amostrada tem uma e apenas uma combinação de categorias X e Y, de modo que

$$\sum_{i=1}^{I} \sum_{j=1}^{J} \pi_{ij} = 1.$$

O estimador de máxima verossimilhança para essas probabilidades procede de forma análoga ao que foi feito na Seção 3.1, com pequenas modificações na notação para levar em conta a segunda variável.

A função de probabilidades para $N_{11}, \cdots, N_{IJ}$ torna-se

$$P(N_{11} = n_{11}, \cdots, N_{IJ} = n_{IJ}) = \frac{n!}{\prod_{i=1}^{I}\prod_{j=1}^{J} n_{ij}!} \prod_{i=1}^{I}\prod_{j=1}^{J} \pi_{ij}^{n_{ij}},$$

que também é a função de verossimilhança para uma amostra de tamanho n. O MLE de π_{ij} é a proporção estimada $\widehat{\pi}_{ij} = n_{ij}/n$.

Distribuições marginais para X e para Y também podem ser encontradas. A probabilidade marginal para o nível i de X é

$$\pi_{i+} = P(X = i)$$

para $i = 1, \cdots, I$. A distribuição marginal de X é, portanto, multinomial com n tentativas e probabilidades $\pi_{1+}, \cdots, \pi_{I+}$ e resulta nas contagens marginais $n_{1+}, \cdots, n_{I+}$. Da mesma forma, a probabilidade marginal para a categoria j de Y é $\pi_{+j} = \sum_{i=1}^{I} \pi_{ij}$ para $j = 1, \cdots, J$.

A distribuição marginal de Y é multinomial com n tentativas, probabilidades $\pi_{+1}, \cdots, \pi_{+J}$, resultando nas contagens marginais $n_{+1}, \cdots, n_{+J}$. Observe que $\sum_{i=1}^{I} \pi_{i+} = 1$ e $\sum_{j=1}^{J} \pi_{+j} = 1$. Os estimadores de máxima verossimilhança (MLEs) de π_{i+} e π_{+j} são as proporções de linha e coluna correspondentes, $\widehat{\pi}_{i+} = n_{i+}/n$ e $\widehat{\pi}_{+j} = n_{+j}/n$, respectivamente.

Quando o resultado de X não afeta as probabilidades dos resultados de Y , dizemos que Y é independente de X. Como resultado da independência, a probabilidade de qualquer resultado conjunto de fatores $(X = i, Y = j)$ nas probabilidades marginais para $X = i$ e $Y = j$: $\pi_{ij} = \pi_{i+}\pi_{+j}$.

A independência simplifica a estrutura das probabilidades dentro de uma tabela de contingência, reduzindo o número de parâmetros desconhecidos para $(I - 1) + (J - 1) = I + J - 2$ probabilidades marginais. Assim, há uma redução de $(IJ-1)-(I+J-2) = (I-1)(J-1)$ parâmetros em comparação com a mesma tabela sem independência. Examinaremos em breve como realizar um teste de hipótese para independência.

Um modelo alternativo é necessário quando amostras de tamanhos n_{i+}, $i = 1, \cdots, I$ sou deliberadamente tiradas de cada um dos diferentes grupos.

Neste caso, as contagens marginais n_{i+} são fixas por planejamento, então temos uma distribuição multinomial de J categorias separada em cada um dos I grupos, onde $n = \sum_{i=1}^{I} n_{i+}$. Cada uma dessas distribuições tem seu próprio conjunto de parâmetros de probabilidade.

Definindo $P(Y = j \,|\, X = i) = \pi_{j\,|\,i}$ como a probabilidade condicional de observar a categoria de resposta j dado que uma unidade é do grupo i. Observe que $\sum_{j=1}^{J} \pi_{j\,|\,i} = 1$ para cada $i = 1, \cdots, I$.

A distribuição conjunta condicional de $N_{i1}, \cdots, N_{iJ}$ tem função de probabilidade

$$P(N_{i1} = n_{i1}, \cdots, N_{iJ} = n_{iJ} \,|\, N_{i+} = n_{i+}) = \frac{n_{i+!}}{\prod_{j=1}^{J} n_{ij}!} \prod_{j=1}^{J} \pi_{j\,|\,i}^{n_{ij}},$$

para cada $i = 1, \cdots, I$. Assumindo que I amostras diferentes são independentes, a verossimilhança é o produto de I distribuições multinomiais,

$$\prod_{i=1}^{I} \frac{n_{i+!}}{\prod_{j=1}^{J} n_{ij}!} \prod_{j=1}^{J} \pi_{j\,|\,i}^{n_{ij}}.$$

Como resultado, este modelo é muitas vezes referido como o modelo produto de multinomiais. O estimador de máxima verossimilhança (MLE) de $\pi_{j\,|\,i}$ é $\widehat{\pi}_{j\,|\,i} = n_{ij}/n_{i+}$.

Um teste de independência,

$$\begin{aligned} H_0 &: \pi_{ij} = \pi_{i+}\pi_{j+} \quad \text{para cada } i,j \\ H_1 &: \pi_{ij} \neq \pi_{i+}\pi_{j+} \quad \text{para algum } i,j \end{aligned}$$

pode ser realizado usando um teste qui-quadrado de Pearson e um teste da razão de verossimilhanças (LRT).

Esses testes já foram mostrados como testes para a igualdade de duas probabilidades de sucesso binomial, o que equivale a um teste de independência em um modelo produto de multinomiais com $I = J = 2$. A estatística do teste qui-quadrado de Pearson é novamente formado pela soma

$$\frac{(\text{contagem observada - contagem esperada estimada})^2}{(\text{contagem esperada estimada})}$$

em todas as células da tabela de contingência.

A contagem observada é n_{ij}. A contagem esperada sob independência é $n\widehat{\pi}_{i+}\widehat{\pi}_{+j} = n_{i+}n_{+j}/n$ para o modelo multinomial, ou equivalentemente $n_{i+}\widehat{\pi}_{+j} = n_{i+}n_{+j}/n$ sob o modelo produto de multinomiais. Isso leva à estatística de teste

$$X^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{\left(n_{ij} - n_{i+}n_{+j}/n\right)^2}{n_{i+}n_{+j}/n}.$$

A estatística X^2 tem uma distribuição aproximada de $\chi^2_{(I-1)(J-1)}$ em grandes amostras quando a hipótese nula é verdadeira.

Quando a hipótese nula é falsa, esperamos grandes desvios entre as contagens observadas e esperadas em relação ao tamanho de $n_{i+}n_{+j}/n$, que levam a grandes valores da estatística X^2. Portanto, rejeitamos a hipótese nula de independência entre X e Y quando $X^2 > \chi^2_{(I-1)(J-1),1-\alpha}$.

A razão de verossimilhança é formada da maneira usual como

$$\Lambda = \frac{\text{Máximo da função de verossimilhança sob } H_0}{\text{Máximo da função de verossimilhança sob } H_0 \text{ ou } H_1}.$$

O cálculo de Λ é baseado em $\widehat{\pi}_{ij} = \widehat{\pi}_{i+}\widehat{\pi}_{+j} = n_{i+}n_{+j}/n$ no numerador e $\widehat{\pi}_{ij} = n_{ij}/n$ no denominador para estimar cada π_{ij}. Aplicando a transformação usual com Λ, temos

$$-2\log(\Lambda) = 2\sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij}\left(\frac{n_{ij}}{n_{i+}n_{+j}/n}\right),$$

onde assumimos $0 \times \log(0) = 0$ por convenção. Assim como X^2, a estatística LRT transformada tem uma distribuição amostral em amostras grandes $\chi^2_{(I-1)(J-1)}$ quando H_0 é verdadeiro e usa a mesma regra de rejeição.

Os graus de liberdade usados para o teste de independência são encontrados calculando

$$(\text{No. de parâmetros sob } H_1) - (\text{No. de parâmetros sob } H_0).$$

Esta é uma maneira geral de encontrar graus de liberdade para qualquer teste de comparação de modelos. Conforme mostrado na Seção 3.2.1, precisamos estimar os $I - J - 2$ parâmetros quando a hipótese nula de independência é válida e $IJ-1$ quando não. Assim, os graus de liberdade para o teste de independência são $(IJ - 1) - (I + J - 2) = (I - 1)(J - 1)$.

Tanto o teste LRT quanto o teste qui-quadrado de Pearson geralmente fornecem resultados semelhantes em amostras grandes. No entanto, seus valores podem diferir consideravelmente em amostras menores, levando a ambiguidade se seus valores estiverem em lados opostos da região de rejeição.

Tem havido uma série de recomendações sobre o que constitui uma amostra “grande o suficiente” para obter uma boa aproximação χ^2.

Os critérios mais comuns são $n_{i+}n_{+j}/n > 1$ ou > 5 para todas as células da tabela de contingência. Esses critérios podem não ser atendidos quando há contagens de células muito pequenas em muitas células da tabela. Por exemplo, uma linha para a qual a contagem marginal n_{i+} não é muito maior do que o número de colunas não pode ter contagens de células esperadas “grandes” em todas as colunas. Nesses casos, uma simulação de Monte Carlo ou os procedimentos de teste exatos descritos na Seção 6.2 podem fornecer uma avaliação visual para determinar se a aproximação à distribuição $\chi^2_{(I-1)(J-1)}$ é apropriada. Esta é a abordagem preferida sempre que houver alguma dúvida em relação à aproximação χ^2 e fornecemos um exemplo de sua implementação.

Exemplo 3.3: Crackers enriquecidos com fibra.