

Tabelas Bidimensionais

Unidade I

Tabelas Bidimensionais

Exemplo 1

- Deseja-se estudar se existe associação entre gênero e crença em vida após a morte. É mais provável um dos gêneros acreditar em vida após a morte do que o outro, ou a crença em vida após a morte independe do gênero?

Uma amostra de 1091 americanos foi selecionada e os mesmos foram classificados de acordo com seu gênero e sua opinião sobre vida após a morte. Dentre as mulheres da amostra, por exemplo, 435 disseram acreditar em vida após a morte e 147 disseram que não ou que não sabem.

Gênero	Acredita em vida após a morte		Total
	Sim	Não ou não sabe	
Feminino	435	147	582
Masculino	375	134	509
Total	810	281	1091

Fonte: General Social Survey - 1991

Tabelas Bidimensionais

Exemplo 2

- Deseja-se estudar a relação entre uso de aspirina e infarto do miocárdio. O uso de aspirina reduz o risco de infarto do miocárdio?

A pesquisa consistia de um estudo aleatorizado de 5 anos. Uma amostra de 11034 médicos tomaram placebo e 11037 tomaram aspirina. Dia sim, dia não, os médicos que participavam do estudo tomavam ou uma aspirina ou um placebo. O estudo era cego – os médicos que participavam do estudo não sabiam o tipo de pílula que estavam tomando.

Grupo	Infarto do Miocárdio		Total
	Sim	Não ou não sabe	
Placebo	189	10845	11034
Aspirina	104	10933	11037
Total	293	21778	22071

Estrutura de Probabilidade

Sejam X e Y duas variáveis respostas categorizadas:

X com I níveis ($i=1,2,\dots,I$)

Y com J níveis ($j=1,2,\dots,J$)

- Quando se classifica os elementos com relação a ambas as variáveis, existem IJ combinações possíveis.
- As respostas (X, Y) de um elemento selecionado aleatoriamente de alguma população tendo uma *distribuição de probabilidade*.
- Esta distribuição é apresentada em uma tabela retangular com I linhas para as categorias de X e J colunas para as categorias de Y .

$\pi_{ij} \rightarrow$ probabilidade de que (X, Y) caia na célula da linha i , coluna j

$X \backslash Y$	Y_1	Y_2	...	Y_J
X_1	π_{11}	π_{12}		π_{1J}
X_2	π_{21}	π_{22}		π_{2J}
...				
X_I	π_{I1}	π_{I2}		π_{IJ}

Distribuição Conjunta de X e Y

$\{\pi_{ij}\}$ onde $\pi_{ij} = P(X=i, Y=j)$ probabilidade de (X, Y) pertencer a célula (i, j)
e $\sum_{i,j} \pi_{ij} = 1$.

Distribuição Marginais de X e de Y

$\{\pi_{i+}\}$ \Rightarrow distribuição marginal de X (*variável na linha*) onde

$$\pi_{i+} = P(X=i) = \sum_j \pi_{ij}, \quad i=1,2,\dots,I$$

$\{\pi_{+j}\}$ \Rightarrow distribuição marginal de Y (*variável na coluna*) onde

$$\pi_{+j} = P(Y=j) = \sum_i \pi_{ij}, \quad j=1,2,\dots,J$$

e satisfazem

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

Distribuições Condicionais de Y dado o nível i da variável X

$$\pi_{ji|i} = P(Y=j | X=i) = \frac{P(X=i, Y=j)}{P(X=i)} = \frac{\pi_{ij}}{\pi_{i+}}, \quad j=1,2,\dots,J$$

- Observe que nesta definição i é mantido fixo. Ou seja, para cada i , temos uma distribuição condicional de Y dado $X=i$.

Distribuições Amostrais

A análise de dados categorizados, como de qualquer tipo de dado, requer suposições sobre o mecanismo aleatório que gerou os dados.

- Quando o tamanho total da amostra é fixado em uma tabela, mas os totais das linhas e das colunas não. \rightarrow *modelo de amostragem multinomial*
- Quando as linhas de uma tabela de contingência se referem a I grupos, os tamanhos das amostras daqueles grupos são freqüentemente fixados pelo delineamento amostral.
 - Se Y tem 2 categorias \rightarrow *modelo de amostragem binomial para cada grupo*
 - Se Y tem J categorias \rightarrow *modelo de amostragem multinomial para cada grupo*

Tabelas de Contingência

Suponha que observações independentes do vetor (X, Y) são feitas em n unidades amostrais.

$X \backslash Y$	Y_1	Y_2	...	Y_J	Total
X_1	n_{11}	n_{12}		n_{1J}	n_{1+}
X_2	n_{21}	n_{22}		n_{2J}	n_{2+}
...					
X_I	n_{I1}	n_{I2}		n_{IJ}	n_{I+}
Total	n_{+1}	n_{+2}	...	n_{+J}	n

Onde $n_{ij} \rightarrow$ freqüência observada da célula (i, j) e $\sum_{j=1}^J \sum_{i=1}^I n_{ij} = n$

$$n_{i+} = \sum_{j=1}^J n_{ij} \rightarrow \text{total da linha } i \text{ de } X$$

$$n_{+j} = \sum_{i=1}^I n_{ij} \rightarrow \text{total da coluna } j \text{ de } Y$$

COMPARAÇÃO DE PROPORÇÕES EM TABELAS 2 x 2:

- Variáveis respostas tendo duas categorias são chamadas variáveis binárias.
- Muitos estudos comparam dois grupos sobre uma variável resposta binária, Y.
- Os dados podem ser dispostos em uma tabela de contingência 2 x 2, na qual as linhas são os dois grupos e as colunas são os níveis de resposta Y.

Grupo	Y		Total
	Sucesso	Insucesso	
1	n_{11}	n_{12}	n_{1+}
2	n_{21}	n_{22}	n_{2+}
Total	n_{+1}	n_{+2}	n

Estimador de $\pi_{ij} : p_{ij} = \frac{n_{ij}}{n}, \begin{cases} i=1,2,\dots,I \\ j=1,2,\dots,J \end{cases}$

Estimador de $\pi_{i+} : p_{i+} = \frac{n_{i+}}{n}, i=1,2,\dots,I$

Estimador de $\pi_{+j} : p_{+j} = \frac{n_{+j}}{n}, j=1,2,\dots,J$

Estimador de $\pi_{j|i} : p_{ji} = \frac{n_{ij}}{n_{i+}}, j=1,2,\dots,J$

Exemplo 2

- Deseja-se estudar a relação entre uso de aspirina e infarto do miocárdio. O uso de aspirina reduz o risco de infarto do miocárdio?

A pesquisa consistia de um estudo aleatorizado de 5 anos. Uma amostra de 11034 médicos tomaram placebo e 11037 tomaram aspirina. Dia sim, dia não, os médicos que participavam do estudo tomavam ou uma aspirina ou um placebo. O estudo era cego – os médicos que participavam do estudo não sabiam o tipo de pílula que estavam tomando.

Grupo	Infarto do Miocárdio		Total
	Sim	Não ou não sabe	
Placebo	189	10845	11034
Aspirina	104	10933	11037
Total	293	21778	22071

Diferença de Proporções

- Considere Y uma variável resposta binária com categorias designadas **sucesso** e **insucesso**, especificamente, observadas nos dois grupos a serem comparados.

- Para os indivíduos na **linha 1**, sejam:

π_1 - a probabilidade de sucesso

$1 - \pi_1$ - a probabilidade de insucesso

Distribuição condicional de Y na linha 1.

- Para os indivíduos na **linha 2**, sejam:

π_2 - a probabilidade de sucesso

$1 - \pi_2$ - a probabilidade de insucesso

Distribuição condicional de Y na linha 2.

A **diferença de proporções** $\pi_1 - \pi_2$ compara as probabilidades de sucesso dos dois grupos (linhas).

Observe que:

$$-1 \leq \pi_1 - \pi_2 \leq 1$$

Temos que $\pi_1 - \pi_2 = 0$ quando $\pi_1 = \pi_2$, isto é, quando a resposta é **independente** do grupo de classificação.

Sejam p_1 e p_2 as proporções amostrais de sucesso para os dois grupos (linhas).

A diferença de proporções amostrais $p_1 - p_2$ é um estimador de $\pi_1 - \pi_2$.

Considerando que para grandes amostras, os estimadores

$$p_i \sim N\left(\pi_i, \frac{\pi_i(1-\pi_i)}{n_i}\right)$$

temos que:

$$p_1 - p_2 \sim N\left(\pi_1 - \pi_2, \frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}\right)$$

O **intervalo de confiança** $(1 - \alpha)$ para $\pi_1 - \pi_2$ considerando grandes amostras é dado por :

$$(p_1 - p_2) \mp z_{\alpha/2} \hat{\sigma}(p_1 - p_2)$$

sendo:

■ $z_{\alpha/2}$ - percentil da $N(0,1)$ tendo cauda a direita igual a $\alpha/2$.

■ $\hat{\sigma}(p_1 - p_2)$ - o erro padrão de $p_1 - p_2$ dado por:

$$\hat{\sigma}(p_1 - p_2) = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

Exemplo

Considerando o problema apresentado no **Exemplo 2** determine o intervalo de 95% de confiança para diferença de proporção de infartos do miocárdio entre os médicos que tomaram placebo e os que tomaram aspirina.

O intervalo obtido indica que o uso de aspirina reduz o risco de infarto do miocárdio?

Grupo	Infarto do Miocárdio		Total
	Sim	Não ou não sabe	
Placebo	189	10845	11034
Aspirina	104	10933	11037
Total	293	21778	22071

Risco Relativo

Em tabelas 2 x 2, o **risco relativo** é a razão das probabilidades de sucesso dos dois grupos,

$$\frac{\pi_1}{\pi_2}$$

O risco relativo pode ser qualquer número real não negativo.

■ O risco relativo é igual 1 quando $\pi_1 = \pi_2$, isto é, quando a resposta é independente do grupo.

■ Dois grupos com proporções amostrais p_1 e p_2 tem um risco relativo amostral dado por:

$$\frac{p_1}{p_2}$$

■ Sua distribuição amostral pode ser altamente assimétrica a menos que os tamanhos das amostras sejam muito grandes, de modo que seu intervalo de confiança é mais complexo.

Risco Relativo

- **Intervalo de Confiança $(1-\alpha)$ para *risco relativo*:**

- Um *intervalo de confiança para o log do risco relativo* é dado por:

$$\log\left(\frac{p_1}{p_2}\right) \pm z_{\alpha/2} \sqrt{\frac{1-p_1}{n_1 p_1} + \frac{1-p_2}{n_2 p_2}}$$

- O *antilog dos limites deste intervalo* produz um *intervalo de confiança para o real risco relativo*.
- Usar apenas diferença de proporções para comparar dois grupos pode ser algo enganoso quando ambas as proporções são próximas de zero.

Exemplo

Considerando o problema apresentado no **Exemplo 2** estime o risco relativo de infarto do miocárdio e determine o respectivo intervalo de 95% de confiança.

O intervalo obtido indica que o uso de aspirina reduz o risco de infarto do miocárdio?

Grupo	Infarto do Miocárdio		Total
	Sim	Não ou não sabe	
Placebo	189	10845	11034
Aspirina	104	10933	11037
Total	293	21778	22071