# Dados Categorizados

As variáveis categóricas têm dois tipos principais de escalas de medição. Muitas escalas categóricas têm uma ordem natural. Os exemplos são atitudes em relação à legalização do aborto (desaprovar em todos os casos, aprovar apenas em certos casos, aprovar em todos os casos), avaliação do nível de estoque de uma empresa (muito baixo, quase certo, muito alto), resposta a um tratamento médico (excelente, bom, regular, ruim) e frequência de sintomas de ansiedade (nunca, ocasionalmente, frequentemente, sempre). Variáveis categóricas com escalas ordenadas são chamadas de variáveis **ordinais**.

Variáveis categóricas com escalas não ordenadas são denominadas variáveis **nominais**. Os exemplos são afiliação religiosa (categorias católica, judaica, protestante, muçulmana, outra), principal meio de transporte para o trabalho (automóvel, bicicleta, ônibus, metrô, caminhada), tipo de música favorita (clássica, country, folclórica, jazz, rock) e o local favorito para fazer compras (shopping local, centro da cidade, Internet, outros).

## Tabela de Contigência

As tabelas de frequência de duas variáveis apresentadas simultaneamente são chamadas tabelas de contingência. As tabelas de contingência são construídas listando todos os níveis de uma variável como linhas em uma tabela e os níveis das outras variáveis como colunas e, em seguida, localizando a frequência de junta ou célula de cada célula. As frequências das células são somadas nas linhas e nas colunas. As somas são colocadas nas margens, cujos valores são chamados de frequências marginais. O valor do canto inferior direito contém a soma das frequências marginais da linha ou da coluna, as quais devem ser iguais a N.

**Exemplo**

| Prática de | Tabagismo | | |
|---|---|---|---|
| Esportes | Presente | Ausente | Total |
| Presente | $a = 50$ | $b = 15$ | $a + b = 65$ |
| Ausente | $c = 10$ | $d = 25$ | $c + d = 35$ |
| Total | $a + c = 60$ | $b + d = 40$ | $n = 100$ |

$H_0$: não há associação entre tabagismo e prática de esportes
$H_A$: há associação entre tabagismo e prática de esportes

### Estrutura de probabilidade para tabelas de contigência

Suponha que haja duas variáveis categóricas, denotadas por X e Y. Deixe eu denotar o número de categorias de X e J o número de categorias de Y. Uma mesa retangular tendo I linhas para as

categorias de colunas X e J para as categorias de Y tem células que mostram as combinações possíveis de resultados.

### *Probabilidades conjuntas, marginais e condicionais*

As probabilidades para tabelas de contingência podem ser de três tipos - conjunta, marginal ou condicional. Suponha primeiro que um sujeito escolhido aleatoriamente da população de interesse é classificado em X e Y. Seja $\pi_i j = P(X = i, Y = j)$ denotar a probabilidade de (X, Y) cai na célula na linha ie coluna j. As probabilidades $\pi_i j$, formam a conjunta distribuição de X e Y. Eles satisfazem i, j $\sum \pi_i j = 1.$ .

### *Sensibilidade e especificidade em testes de diagnóstico*

O teste de diagnóstico é usado para detectar muitas condições médicas. Por exemplo, a mamografia pode detectar câncer de mama em mulheres e o antígeno prostático específico (PSA) pode detectar câncer de próstata em homens. Diz-se que o resultado de um teste de diagnóstico é positivo se afirma que a doença está presente e negativo se afirma que a doença está ausente. A precisão dos testes de diagnóstico é frequentemente avaliada com duas probabilidades condicionais: Dado que um indivíduo tem a doença, a probabilidade de o teste de diagnóstico ser positivo é chamado de sensibilidade. Dado que o sujeito não tem a doença, a probabilidade de o teste ser negativo é chamada de especificidade

$$Sensibilidade = P(Y = 1|X = 1), \quad especificidade = P(Y = 2|X = 2)$$

### Independência

Diz-se que duas variáveis são estatisticamente independentes se a população condicional distribuições de Y são idênticas em cada nível de X. Quando duas variáveis são independentes, a probabilidade de qualquer resultado específico da coluna j é a mesma em cada linha.

### Risco Relativo

Uma diferença entre duas proporções de um determinado tamanho fixo geralmente é mais importante quando ambas as proporções estão próximas de 0 ou 1 do que quando estão próximas do meio do intervalo. Considere uma comparação de dois medicamentos na proporção de indivíduos que tiveram efeitos adversos reações ao usar o medicamento. A diferença entre 0,010 e 0,001 é a mesma como a diferença entre 0,410 e 0,401, ou seja, 0,009. A primeira diferença é mais impressionante, uma vez que 10 vezes mais indivíduos tiveram reações adversas com um medicamento que o de outros. Nesses casos, a razão de proporções é uma medida descritiva mais relevante.

$$Risco = \frac{\pi_1}{\pi_2}$$

## Teste qui quadrado

- **Aderência** :

O objetivo do teste de aderência é Testar a adequabilidade de um modelo probabilístico a um conjunto de dados observados

o teste de aderência pode ser feito de forma semelhante ao de independência.

- **independência**

O teste de independência Qui-Quadrado é usado para descobrir se existe uma associação entre a variável de linha e coluna variável em uma tabela de contingência construído à partir de dados da amostra. A hipótese nula é de que as variáveis não estão associadas, em outras palavras, eles são independentes. A hipótese alternativa é de que as variáveis estão associadas, ou dependentes.

Observações:

Os dados são selecionados aleatoriamente. Todas as frequências esperadas são maiores do que ou igual a 1 (isto é, $E_{ij} > 1$).

Seja $O_{ij} = (O_{11}, O_{12}, \ldots, O_{rc})$ representa o vetor de contagens observadas com distribuição multinomial, $E_{ij}$ representa as frequências esperadas e admintindo válida a hipótese de independência dos critérios de classificação, a estatística

$$Q^2_{obs} = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Assim, dado um nível de significância

$$\alpha$$

, o p-valor é determinado por

$$\text{p-valor} = P[Q^2_{obs} \geq \chi^2_{\alpha;(r-1)(c-1)} | H_0]$$

Tópico anterior

Próximo Tópico