Para saber mais: teste de independência qui-quadrado

O teste de independência qui-quadrado é uma técnica estatística não paramétrica usada para determinar se **existe ou não uma associação significativa entre duas variáveis categóricas**. Variáveis de natureza categóricas são observações que não podem corresponder a um valor numérico, representando categorias ou grupos distintos, cujas observações não têm uma ordem específica.

Para exemplificação, vamos analisar alguns casos em que temos variáveis de natureza categórica:

* Uma pesquisa de opinião que busca investigar se a preferência política em uma cidade está associada à faixa etária dos eleitores;
* Um estudo entre funcionários para verificar se há uma associação entre a satisfação no trabalho e a categoria profissional dos funcionários; ou
* Em uma instituição de ensino, coletar dados para analisar se o desempenho acadêmico está associado ao método de ensino utilizado.

Em todos os exemplos comparamos duas variáveis categóricas: preferência política com faixa etária; satisfação no trabalho com a categoria profissional; e desempenho acadêmico com o método de ensino. Para nos ajudar a entender como formular o teste, vamos utilizar o exemplo da instituição de ensino.

Digamos que na coleta de dados para avaliar a associação entre o desempenho acadêmico e o método de ensino obtemos um conjunto de dados com 50 valores, os cinco primeiros podem ser observados abaixo:

|  | **Desempenho** | **Metodo\_Ensino** |
| --- | --- | --- |
| 0 | Regular | Moderno |
| 1 | Bom | Tradicional |
| 2 | Bom | Tradicional |
| 3 | Excelente | Moderno |
| 4 | Bom | Moderno |

Para iniciar a avaliação, iremos formular uma hipótese.

#### **Etapa 1: Formulando hipótese**

Aplicamos o teste qui-quadrado para avaliar se as variáveis categóricas são independentes entre elas. Então, as hipóteses formuladas vão sempre buscar verificar essa independência, mantendo a seguinte estrutura:

* Hipótese Nula (H0): As variáveis são independentes.
* Hipótese Alternativa (H1): As variáveis não são independentes.

Para nosso exemplo, podemos definir as seguintes hipóteses:

* H0: o desempenho acadêmico e o método de ensino são independentes.
* H1: o desempenho acadêmico e o método de ensino *não* são independentes.

Com a hipótese formulada, podemos ir para o próximo passo: criação da tabela de contingência

#### **Etapa 2: construção da tabela de contingência**

Uma tabela de contingência é uma tabela que mostra a distribuição conjunta de duas ou mais variáveis categóricas, permitindo entender a relação entre essas variáveis.

A tabela é organizada em linhas e colunas, onde as linhas representam uma variável categórica e as colunas representam outra. Cada célula da tabela contém o número de observações que se enquadram em uma determinada combinação de categorias.

No nosso exemplo, podemos colocar as categorias de desempenho nas linhas e o modelo de ensino nas colunas. Observando o resultado, conseguimos verificar que existem 4 amostras com a situação de “desempenho bom” e “modelo tradicional”, enquanto existem 8 amostras com a situação de “desempenho bom” e “modelo moderno”.

Ao final da contagem, temos a seguinte tabela de contingência:

| **Desempenho** | **Moderno** | **Tradicional** |
| --- | --- | --- |
| Bom | 8 | 4 |
| Excelente | 6 | 6 |
| Muito Bom | 7 | 6 |
| Regular | 8 | 4 |
| Ruim | 8 | 3 |

No Python, conseguimos calcular essa tabela pelo método [crosstabdo Pandas](https://pandas.pydata.org/docs/reference/api/pandas.crosstab.html). Podemos transformar nossos dados em um DataFrame e utilizar o crosstab para obtermos a tabela de contingência de forma automática.

import pandas as pd

# Dados do exemplo

df = pd.DataFrame({'Desempenho': {0: 'Regular', 1: 'Bom', 2: 'Bom', 3: 'Excelente', 4: 'Bom', 5: 'Bom', 6: 'Regular', 7: 'Regular', 8: 'Excelente', 9: 'Ruim', 10: 'Bom', 11: 'Regular', 12: 'Muito Bom', 13: 'Bom', 14: 'Bom', 15: 'Regular', 16: 'Muito Bom', 17: 'Bom', 18: 'Muito Bom', 19: 'Excelente', 20: 'Ruim', 21: 'Excelente', 22: 'Ruim', 23: 'Ruim', 24: 'Regular', 25: 'Excelente', 26: 'Ruim', 27: 'Excelente', 28: 'Excelente', 29: 'Regular', 30: 'Regular', 31: 'Ruim', 32: 'Regular', 33: 'Regular', 34: 'Muito Bom', 35: 'Excelente', 36: 'Regular', 37: 'Excelente', 38: 'Excelente', 39: 'Muito Bom', 40: 'Ruim', 41: 'Ruim', 42: 'Ruim', 43: 'Ruim', 44: 'Muito Bom', 45: 'Regular', 46: 'Excelente', 47: 'Muito Bom', 48: 'Muito Bom', 49: 'Muito Bom', 50: 'Muito Bom', 51: 'Muito Bom', 52: 'Bom', 53: 'Ruim', 54: 'Muito Bom', 55: 'Bom', 56: 'Bom', 57: 'Excelente', 58: 'Bom', 59: 'Muito Bom'},

'Metodo\_Ensino': {0: 'Moderno', 1: 'Tradicional', 2: 'Tradicional', 3: 'Moderno', 4: 'Moderno', 5: 'Moderno', 6: 'Moderno', 7: 'Moderno', 8: 'Moderno', 9: 'Tradicional', 10: 'Tradicional', 11: 'Moderno', 12: 'Tradicional', 13: 'Moderno', 14: 'Tradicional', 15: 'Tradicional', 16: 'Moderno', 17: 'Moderno', 18: 'Tradicional', 19: 'Tradicional', 20: 'Moderno', 21: 'Tradicional', 22: 'Moderno', 23: 'Moderno', 24: 'Tradicional', 25: 'Tradicional', 26: 'Moderno', 27: 'Moderno', 28: 'Moderno', 29: 'Moderno', 30: 'Moderno', 31: 'Moderno', 32: 'Tradicional', 33: 'Moderno', 34: 'Tradicional', 35: 'Tradicional', 36: 'Tradicional', 37: 'Moderno', 38: 'Tradicional', 39: 'Moderno', 40: 'Tradicional', 41: 'Moderno', 42: 'Tradicional', 43: 'Moderno', 44: 'Moderno', 45: 'Moderno', 46: 'Moderno', 47: 'Moderno', 48: 'Moderno', 49: 'Tradicional', 50: 'Moderno', 51: 'Moderno', 52: 'Moderno', 53: 'Moderno', 54: 'Tradicional', 55: 'Moderno', 56: 'Moderno', 57: 'Tradicional', 58: 'Moderno', 59: 'Tradicional'}})

# Tabela de contingência

tab\_contingencia = pd.crosstab(df['Desempenho'], df['Metodo\_Ensino'])

Após a formulação da tabela podemos aplicar o teste.

#### **Etapa 4: Teste de independência qui-quadrado**

O Teste qui-quadrado utiliza dos dados da tabela de contingência para calcular o valor da Estatística Chi2 (*qui-quadrado*) e o p-valor. No python, conseguimos aplicar a função [chi2\_contingency da biblioteca SciPy](https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.chi2_contingency.html), precisando enviar apenas a tabela de contingência como argumento. A saída de chi2\_contingency são quatro valores: Estatística Chi2, p-valor, graus de liberdade e frequências esperadas. Os dois últimos não são muito necessários para nossa tomada de decisão atual, então nesse exemplo iremos coletar apenas os dois resultados iniciais.

from scipy.stats import chi2\_contingency

chi2, p\_valor, \_, \_ = chi2\_contingency(tab\_contingencia)

# Resultados iniciais

print(f'Estatística chi2: {chi2}')

print(f'p-valor: {p\_valor}')

# Tomada de decisão

nivel\_significancia = 0.05

if p\_valor < nivel\_significancia:

conclusao = 'Rejeitar a hipótese nula'

else:

conclusao = 'Não rejeitar a hipótese nula'

print('Conclusão:', conclusao)

**Saída:**

Estatística chi2: 1.850394024307068

p-valor: 0.7632507187470916

Conclusão: Não rejeitar a hipótese nula

Ao não rejeitar a hipótese nula, entendemos que as variáveis de desempenho acadêmico e o método de ensino são independentes, ou seja, não há uma associação significativa entre elas.

Por fim vale destacar algumas das condições específicas para a aplicação do teste qui-quadrado:

1. **Repetições:** As frequências esperadas em cada célula da tabela de contingência devem ser suficientemente grandes, geralmente maiores que 5, para que a aproximação seja válida (alguns autores dizem que em pelo menos 80% das células devemos ter 5 ou mais registros);
2. **Tamanho da amostra:** A amostra deve ser suficientemente grande, um total de observações maior que 20.