Especialização em *Data Science* e Estatística Aplicada

Módulo II - Análise estatística de várias populações

Profa. Dra. Tatiane F N Melo

Goiânia, 2024

IME INSTITUTO DE MATEMÁTICA E ESTATÍSTICA
FEN FACULDADE DE ENFERMAGEM
UFG UNIVERSIDADE FEDERAL DE GOIÁS

Aula 3 - Parte 1

1. Análise de aderência e associação
   - Teste de Independência (*Test of Independence*)

2. Referências Bibliográficas

Teste de Independência

Objetivo

- O teste de independência é usado para verificar se duas variáveis categóricas são independentes ou não.
- Em outras palavras, testar a hipótese nula de que dois critérios de classificação, quando aplicados ao mesmo conjunto de elementos, são independentes.
- Ou seja, temos interesse nas hipóteses:

  H_0: As duas variáveis são independentes.

  H_1: As duas variáveis não são independentes.

Teste de Independência

Tabela de contingência

- A classificação, de acordo com dois critérios, de um conjunto de dados, pode ser mostrada por uma tabela na qual as r linhas representam os vários níveis de um critério de classificação e as c colunas representam os vários níveis do segundo critério.
- Esta tabela é geralmente chamada de tabela de contingência, com dimensão $r \times c$.

Teste de Independência

Tabela de contingência

Segundo critério	Primeiro critério					Total
	1	2	3	...	c	
1	n_{11}	n_{12}	n_{13}	...	n_{1c}	$n_{1.}$
2	n_{21}	n_{22}	n_{23}	...	n_{2c}	$n_{2.}$
3	n_{31}	n_{32}	n_{33}	...	n_{3c}	$n_{3.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	...	$\vdots$	$\vdots$
r	n_{r1}	n_{r2}	n_{r3}	...	n_{rc}	$n_{r.}$
Total	$n_{.1}$	$n_{.2}$	$n_{.3}$	...	$n_{.c}$	n

H_0: Os dois critérios são independentes.

Teste de Independência

Exemplo 1

Suponha que temos interesse em verificar se o status de vacinação é independente da incidência de infecção respiratória.

Status de vacinação	**Incidência de Infecção Respiratória**		Total
	Infectado	Não infectado	
Vacinado	15	20	35
Não vacinado	25	40	65
Total	40	60	100

Teste de Independência

Descrição do teste

Consideremos um experimento aleatório onde:

- r é o número de linhas da tabela de contingência;
- c é o número de colunas da tabela de contingência;
- O_{ij} é a frequência absoluta observada da i-ésima linha e j-ésima coluna;
- E_{ij} é a frequência absoluta esperada da i-ésima linha e j-ésima coluna.

Teste de Independência

Descrição do teste

- Definimos

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim \chi^2_{\phi},$$

onde $\phi = (r-1) \times (c-1)$.

- O valor-p é calculado por:

$$\widehat{\alpha} = P(\chi^2 > \chi^2_{obs} | H_0),$$

onde χ^2_{obs} é o valor que a estatística de teste assume.

Teste de Independência

Frequências esperadas e observadas

- As frequências esperadas e observadas são comparadas.
- Se a discrepância for suficientemente pequena, a hipótese nula é sustentável.
- Se a discrepância for suficientemente grande, a hipótese nula é rejeitada e concluímos que os dois critérios de classificação não são independentes.

Teste de Independência

Cálculo das frequências esperadas

- Em geral, vemos que para obter a frequência esperada para uma determinada célula, multiplicamos o total da linha em que a célula está localizada pelo total da coluna em que a célula está localizada e dividimos o produto pelo total geral.

Teste de Independência

Voltando ao Exemplo 1

Neste caso, temos as hipóteses:

- H_0 : O status de vacinação é independente da incidência de infecção respiratória, ou seja, não há associação entre ser vacinado e a ocorrência de infecção respiratória.
- H_1 : O status de vacinação é dependente da incidência de infecção respiratória, ou seja, existe uma associação significativa entre ser vacinado e a ocorrência de infecção respiratória.

Teste de Independência

Continuação do Exemplo 1

Na tabela abaixo temos as frequências observadas (O_{ij}) e entre parênteses temos as frequências esperadas (E_{ij}). Por exemplo, na célula cuja frequência observada é 15, a frequência esperada é 14, pois $(35 \times 40)/100 = 14$.

Tabela 1: Frequências observadas

	Incidência de Infecção Respiratória		
Status de vacinação	Infectado	Não infectado	Total
Vacinado	15 (14)	20 (21)	35
Não vacinado	25 (26)	40 (39)	65
Total	40	60	100

Teste de Independência

Continuação do Exemplo 1

Neste caso,

$$\chi^2_{obs} = \frac{(15-14)^2}{14} + \frac{(20-21)^2}{21} + \frac{(25-26)^2}{26} + \frac{(40-39)^2}{39} = 0,1831.$$

- Vimos que a estatística de teste tem distribuição χ^2_ϕ com $\phi = (r-1) \times (c-1)$. Aqui, $r = 2$ e $c = 2$. Logo, $\chi \sim \chi^2_1$.
- Calculando o valor-p no `R`:

  $\widehat{\alpha}$ = `pchisq(0.1831, df = 1, lower.tail = FALSE)` = 0,6687226.

Teste de Independência

Continuação do Exemplo 1

Não rejeitamos a hipótese nula ao nível de $\alpha = 1\%$, já que $\widehat{\alpha} < \alpha$. Portanto, concluímos que, para os dados observados, não há uma associação significativa entre "ser vacinado"e a ocorrência de infecção respiratória. As variáveis podem ser consideradas independentes, ao nível de 1%.

Teste de Independência - Aplicação à dados reais

Exemplo 2

- Suponha que queremos saber se a distribuição do tipo de vacina aplicada é independente da faixa etária dos indivíduos.
- Para isso, vamos usar os dados da vacinação contra COVID-19 em Goiânia (Ministério da Saúde - Vacinômetro COVID-19).
- Consideraremos dois tipos de vacinas: Pfizer e Astrazeneca.

Teste de Independência - Aplicação à dados reais

Continuação do Exemplo 2 - Hipóteses

H_0: Não há associação entre a faixa etária e o tipo de vacina administrada. Em outras palavras, a distribuição do tipo de vacina é independente da faixa etária.

H_1: Há uma associação entre a faixa etária e o tipo de vacina administrada. Isto quer dizer que a distribuição do tipo de vacina depende da faixa etária.

Ou seja,

H_0: Faixa etária e tipo de vacina são independentes.

H_1: Faixa etária e tipo de vacina não são independentes.

Teste de Independência - Aplicação à dados reais

Continuação do Exemplo 2

Exemplo 2 no R.

Teste de Independência

Frequências esperadas (pequenas)

- Os problemas de como lidar com pequenas frequências esperadas e pequenos tamanhos totais de amostra podem surgir na análise de tabelas de contingência 2×2.
- Cochran (1954) sugere que o teste χ^2 não deve ser usado se $n < 20$ ou se $20 < n < 40$ e qualquer frequência esperada for menor que 5.
- Quando $n = 40$, uma frequência de célula esperada tão pequena quanto 1 pode ser tolerada.

Teste de Independência

Frequências esperadas (pequenas) - Correção de Yates

- As frequências observadas em uma tabela de contingência são discretas e, portanto, dão origem a uma estatística discreta, χ^2, que é aproximada pela distribuição χ^2, que é contínua.
- Yates (1934) propôs um procedimento para corrigir isso no caso de tabelas 2×2.
- A correção consiste em subtrair metade do número total de observações do valor absoluto da quantidade $ad - bc$ antes do quadrado. Ou seja,

$$\chi^2_{Yates} = \sum_{i=1}^{k} \frac{(|O_{ij} - E_{ij}| - 0,5)^2}{E_{ij}}.$$

Teste de Independência - Aplicação à dados reais

Exemplo 3

Vamos considerar a Base de dados SINASC (Sistema de Informações sobre Nascidos Vivos), no município de São Paulo, em 2023. O interesse é verificar se há independência entre o tipo de parto das mães (1 - vaginal e 2 - cesário) e o local de nascimento dos filhos (1 - hospitalar e 2 - não hospitalar).

Teste de Independência - Aplicação à dados reais

Continuação do Exemplo 3 - Hipóteses

- H_0 : O tipo de parto e o local de nascimento dos filhos são independentes.
- H_1 : O tipo de parto e o local de nascimento dos filhos não são independentes. Ou seja, há associação entre o tipo de parto (cesário e vaginal) e o local de nascimento da criança (hospitalar e não hospitalar).

Teste de Independência - Aplicação à dados reais

Continuação do Exemplo 3

- Exemplo 3 no R.

Teste de Independência - Aplicação à dados reais

Exemplo 4

Aqui também vamos considerar a Base de dados SINASC. Agora o interesse é verificar se há independência entre a raça/cor da mãe (1 - Branca, 2 - Preta, 3 - Amarela, 4 - Parda, 5 - Indígena) e a escolaridade da mãe (1 - Nenhuma, 2 - 1 a 3 anos, 3 - 4 a 7 anos, 4 - 8 a 11 anos, 5 - 12 anos ou mais).

Teste de Independência - Aplicação à dados reais

Continuação do Exemplo 4 - Hipóteses

- H_0 : A raça/cor e a escolaridade da mãe são independentes.
- H_1 : A raça/cor e a escolaridade da mãe não são independentes. Ou seja, há associação entre a raça/cor e escolaridade da mãe.

Teste de Independência - Aplicação à dados reais

Continuação do Exemplo 4

- Exemplo 4 no R.

Características do Teste de Independência

As características de um teste qui-quadrado de independência que o distinguem de outros testes qui-quadrado são as seguintes:

1. Uma única amostra é selecionada de uma população de interesse, e os sujeitos ou objetos são classificados de forma cruzada com base nas duas variáveis de interesse.

2. A justificativa para calcular as frequências de células esperadas é baseada na lei da probabilidade, que afirma que se dois eventos (aqui os dois critérios de classificação) são independentes, a probabilidade de sua ocorrência conjunta é igual ao produto de suas probabilidades individuais.

3. As hipóteses e conclusões são declaradas em termos da independência (ou falta de independência) de duas variáveis.

Referências bibliográficas

1. Cochran, W. G. Some Methods for Strengthening the Common χ^2 Tests, *Biometrics*, 10, 417–451, 1954.

2. Ministério da Saúde - Vacinômetro COVID-19. `https://infoms.saude.gov.br/extensions/SEIDIGI_DEMAS_Vacina_C19/SEIDIGI_DEMAS_Vacina_C19.html`, último acesso: 19/08/2024.

3. Base de dados SINASC, Arquivos de nascidos vivos no município de São Paulo. `https://www.prefeitura.sp.gov.br/cidade/secretarias/saude/epidemiologia_e_informacao/nascidos_vivos/index.php?p=306422`, Ano: 2023.

Referências bibliográficas

4. Vieira, S. Introdução à Bioestatística, 5ª Edição, Elsevier, 2008.

5. Yates, F. Contingency Tables Involving Small Numbers and the χ^2 Tests, *Journal of the Royal Statistical Society*, Supplement, 1, Series B, 217–235, 1934.

Especialização em *Data Science* e Estatística Aplicada

Módulo II - Análise estatística de várias populações

Profa. Dra. Tatiane F N Melo

tmelo@ufg.br

IME INSTITUTO DE MATEMÁTICA E ESTATÍSTICA

FEN FACULDADE DE ENFERMAGEM

UFG UNIVERSIDADE FEDERAL DE GOIÁS