Inferência para frequencias: testes qui-quadrado

Objetivos:

Descrição de dados categóricos por tabelas e gráficos
Teste qui-quadrado de aderência
Teste qui-quadrado de independência
Teste qui-quadrado de homogeneidade

DESCRIÇÃO DE DADOS CATEGÓRICOS

O que são dados categóricos?

São dados decorrentes da observação de variáveis categóricas, ou seja, aqueles que identificam para cada caso uma categoria.

As categorias podem ser derivadas de variáveis qualitativas (nominais ou ordinais) ou quantitativas.

DESCRIÇÃO DE DADOS CATEGÓRICOS

O que fazer para descrever dados categóricos?

1. **Faça uma figura:** A exibição dos seus dados irá revelar coisas que você talvez não veja numa tabela numérica. Auxilia no planejamento de sua abordagem analítica e contribui para pensar claramente sobre os padrões e relacionamentos que podem estar escondidos nos seus dados.
2. **Faça uma figura:** Uma exibição bem projetada irá executar muito do trabalho da análise dos seus dados. A figura pode revelar padrões, valores atípicos ou valores extraordinários que ressaltam erros.
3. **Faça uma figura:** Uma figura bem escolhida ajuda a relatar aos outros o que você encontrou nos seus dados em sua análise.

DESCRIÇÃO DE DADOS CATEGÓRICOS

4. Faça uma tabela de frequências: Uma tabela de frequências ou de frequências relativas é o primeiro passo para se obter uma visualização preliminar quantitativa sobre as variáveis e possíveis associações entre elas.

5. Faça uma tabela de contingência: Tabela de dupla entrada que mostra como as frequências das categorias de uma variável se distribuem ao longo das categorias de outra variável.

6. Analise as distribuições condicionais: Uma distribuição condicional mostra a distribuição de uma variável apenas para aqueles casos que satisfazem uma condição em outra.

DESCRIÇÃO DE DADOS CATEGÓRICOS

7. Verifique se há independência entre as variáveis: Numa tabela de contingência quando a distribuição de uma variável é a mesma para todas as categorias da outra, dizemos que as variáveis são independentes.

8. Descreva a associação entre as variáveis por meio de coeficientes adequados: Uma das medidas mais usuais é o qui-quadrado dePearson.

$$\chi = \sum_i \sum_j \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

TESTE QUI-QUADRADO

- **Teste de Aderência**: *Utilizado para verificar se a distribuição de frequências de uma variável categórica se distribui de acordo com um modelo.*
- **Teste de Homogeneidade**: *Utilizado para comparar a distribuição de frequências para dois ou mais grupos da mesma variável categórica.*
- **Teste de Independência**: *Utilizado para verificar se duas variáveis categóricas são independentes ou não associadas.*

TESTE QUI-QUADRADO DE ADERÊNCIA

Seja $X_1, X_2, ..., X_n$ uma amostra aleatória que caracteriza uma população P e queremos testar a hipótese

$H_0 : P = P_0$

onde P_0 tem uma distribuição de probabilidades específica e de tal modo que podemos escrever a hipótese H_0 como :

$H_0 : p_1 = p_{10}, p_2 = p_{20}, \ldots p_s = p_{s0}$

onde p_{i0} são proporções especificadas que caracterizam P_0.

Pergunta : **A população segue o modelo(padrão) especificado?**

Suposições e condições:

1. Os dados devem ser derivados de contagem(frequências) para as categorias da variável categórica;
2. As frequências das células da tabela de dupla entrada devem ser independentes umas das outras;
3. Os sujeitos contados na tabela devem ser de uma amostra aleatória extraída de alguma população;
4. Devemos ter dados suficientes;

5. Devemos esperar que a frequência seja de pelo menos 5 elementos em cada célula da tabela;

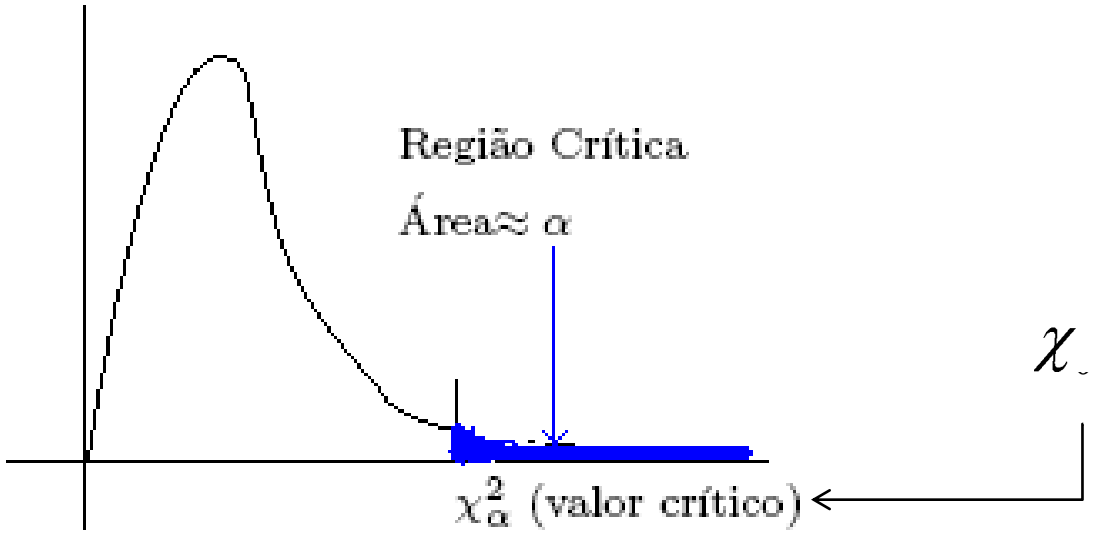
6. Parte-se de um *modelo probabilístico* considerado satisfatório para descrever o comportamento da população, ou seja, quando se tem uma teoria de que as proporções deveriam ocorrer em cada categoria e que acreditam que a sua teoria é verdadeira

$$H_0 : p_1 = \ ?_{10}; p_2 = \ ?_{20}; \ldots; p_s = \ ?_{s0}$$

Estatística de teste:

$$\chi_{\cdot} = \sum_{i=1}^{s} \frac{(o_i - e_i)^2}{e_i}$$

Regra de decisão:
Considerando um nível de significância α, rejeitar a hipótese nula se:

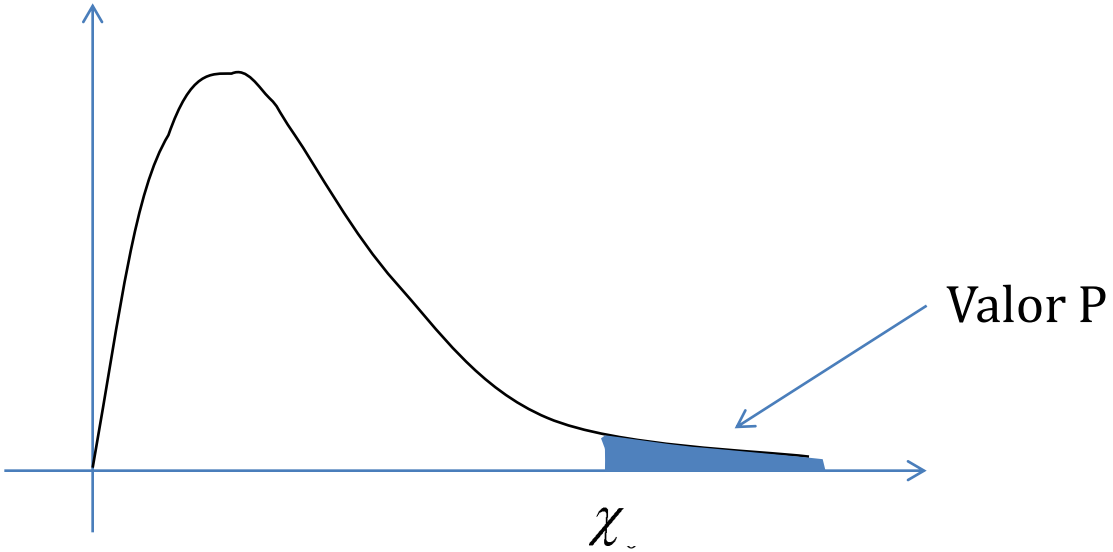


$\chi_{\cdot} > \chi_{(\cdot\ -\ )}$; s é o número de células.

A regra de decisão pode ser baseada também no Valor P:

$$P = P(\chi^2_{(s-1)} > \chi^2_0) \approx \text{área sob a curva à direita de } \chi^2_0$$

Se, para α fixado, obtemos $P < \alpha$, rejeitamos a hipótese nula



Obs.: o teste de aderência nos permite, na melhor das hipóteses, ressaltar se os dados são consistentes ou não com a teoria proposta. O teste não prova se o modelo teórico é verdadeiro.

Exemplo1:
A Mars, Inc. diz que as cores de seus chocolates M&M® são 14% amarelos, 13% vermelhos, 20% laranjas, 24% azuis, 16% verdes e 13% marrons. Um consumidor curioso resolveu verificar se essas proporções se verificam de fato. Tomando um saquinho de M&M, ele observou que continha 29 amarelos, 23 vermelhos, 12 laranjas, 14 azuis, 8 verdes e 20 marrons.
Esta amostra é consistente com as proporções anunciadas? O que deveria concluir o consumidor sobre sua suspeita?

Exemplo2:
A política de uma empresa exige que os espaços do estacionamento sejam designados aleatoriamente a cada pessoa, mas você suspeita que não seja bem assim. Existem 3 áreas de igual tamanho com 80 vagas cada: área A, próximo ao prédio; área B, um pouco mais longe; e área C, do outro lado da rua. Você coleta dados sobre os funcionários de gerência de nível médio e acima para ver quantos foram designados para cada área. 18 estão na área A; 13 estão na área B e 8 estão na área C. Os dados estão consistentes com a proposta da empresa?

TESTE QUI-QUADRADO DE HOMOGENEIDADE

Sejam $X_{11}, X_{12}, ..., X_{1s}$ e $X_{21}, X_{22}, ..., X_{2s}$

Duas amostras aleatórias que caracterizam as populações P_1 e P_2 quanto à variável categórica X. Se desejamos testar a hipótese de que as frequências são as mesmas para ambas populações das quais as amostras foram extraídas, estamos diante de um teste de homogeneidade cuja hipótese nula é formulada como segue:

$$H_0 : P_1 = P_2$$

Pergunta: **As populações são homogêneas ?**

Estatística de teste:

$$\chi^2 = \sum_i \sum_j \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

onde:

$$e_{ij} = \frac{n_{i.} \times n_{.j}}{n}$$

$n_{i.}$ é o número total de elementos na amostra extraída de P_i e

$n_{.j}$ são frequências totais observadas para cada categoria da variável X.

Populações	X				Total
	1	2	3	4	
P_1	O_{11}	O_{21}	O_{13}	O_{14}	$n_{1.}$
P_2	O_{21}	O_{22}	O_{23}	O_{24}	$n_{2.}$
Total	$n_{.1}$	$n_{.2}$	$n_{.3}$	$n_{.4}$	n

Suposições e condições:

1. Os dados devem ser derivados de contagem(frequencias) para as categorias da variável categórica;
2. As frequências das células da tabela de dupla entrada devem ser independentes umas das outras;
3. Os sujeitos contados na tabela **devem ser de amostras aleatórias extraídas de populações distintas** – condição de aleatoriedade
4. Devemos ter dados suficientes
5. A frequência esperada deve ser de **pelo menos 5 elementos em cada célula** da tabela;
6. O teste é aplicável tanto para variável qualitativa como quantitativa, desde que essa última seja categorizada.

Exemplo:
Sabe-se que, embora os usuários da Internet gostem da conveniência das compras *on-line*, eles realmente têm preocupações em relação à privacidade e a segurança. Um estudo buscou saber se essa preocupação é a mesma entre homens e mulheres. Utilizando uma amostra de 825 mulheres e 775 homens usuários da Internet, questionou-se o grau de concordância com a declaração: " Eu não gosto de fornecer o número de meu cartão de crédito ou informações pessoais *on-line".*
Os dados obtidos estão na tabela abaixo:

	Concorda fortemente	Concorda	Discorda	Discorda fortemente	Total
Mulheres	268	276	216	65	825
Homens	358	234	118	65	775
Total	626	510	324	130	00

O que se pode interpretar dos resultados?

TESTE QUI-QUADRADO DE INDEPENDÊNCIA

Sejam $X_1, X_2, ..., X_n$ e $Y_1, Y_2, ..., Y_m$ duas amostras aleatórias que descrevem as variáveis X e Y. Se desejarmos testar a hipótese de independên cia ou não associação entre X e Y, estamos diante de um teste de independên cia, onde a hipótese nula é formulada como segue :

H_0 : As variáveis X e Y são independen tes

H_1 : As variáveis não são independen tes

Pergunta: **As variáveis X e Y são independentes ?**

Estatística de teste: $\chi^2 = \sum_i \sum_j \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$

onde:

$e_{ij} = \frac{n_{i.} \times n_{.j}}{n}$

$n_{i.}$ são frequências marginais relativas a variável X e $n_{.j}$ são frequências marginais relativas à variável Y.

X	Y				Total
	1	2	3	4	
A	O_{11}	O_{21}	O_{13}	O_{14}	$n_{1.}$
B	O_{21}	O_{22}	O_{23}	O_{24}	$n_{2.}$
Total	$n_{.1}$	$n_{.2}$	$n_{.3}$	$n_{.4}$	n

Suposições e condições:

1. Os dados devem ser derivados de contagem(frequencias) para as categorias das variáveis categóricas
2. As frequências das células da tabela de dupla entrada devem ser independentes umas das outras
3. Os sujeitos contados na tabela devem ser de uma amostra aleatória extraída de uma única população
4. Devemos ter dados suficientes
5. Devemos esperar que a frequência seja de pelo menos 5 elementos em cada célula da tabela
6. O teste é aplicável tanto para variável qualitativa como quantitativa

Exemplo:
Uma grande empresa do nordeste dos Estados Unidos que compra peixes de pescadores locais e os distribui para grandes firmas e restaurantes está estudando o lançamento de uma nova campanha publicitária sobre os benefícios do peixe para a saúde. Como evidência, eles gostariam de citar o seguinte estudo. Pesquisadores médicos acompanharam 6272 homens suecos durante 30 anos para ver se havia alguma associação entre a quantidade de peixe na sua dieta e câncer de próstata (“ Fatty Fish Consumption and Risk of Prostate Cancer, Lancet, Junho 2001”).

Consumo de peixe	Sem câncer	Com câncer
Nunca/raramente	110	14
Pequena parte da dieta	2420	201
Parte moderada	2769	209
Grande parte	507	42

a) Isto é um levantamento de dados, um estudo prospectivo ou um experimento?
b) Há evidência de uma associação entre a quantidade de peixe na dieta de um homem e o risco de desenvolver câncer de próstata?
c) Este estudo prova que comer peixe não previne do câncer de próstata? Explique.