Especialização em *Data Science* e Estatística Aplicada

Módulo IV - Análise de Dados Categóricos

Prof Dr Márcio Augusto Ferreira Rodrigues

Goiânia, 2025

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

Conteúdo Programático

- Dados categorizados. Tabelas de contingência bidimensionais.
- Tabelas de contingência tridimensionais.
- Testes para dados categorizados: qui-quadrado, exato de Fisher e razão de verossimilhança.
- Modelo de regressão logística binária.
- Aplicações no software R.

Conteúdo - Aula 1

1. Introdução

2. Tabelas de Contingência 2×2
   - Introdução
   - Medidas de Associação

3. Tabelas de Contingência $r \times s$
   - Medidas de Associação

4. Tabelas de contingência 2 x 2 x K
   - Teste de Mantel-Haenszel

Análise de Dados Categorizados

A análise de experimentos em que a variável resposta é por natureza categórica é denominada análise de dados categóricos ou, também, análise de dados discretos, isto porque distribuições discretas de probabilidade encontram-se associadas às variáveis resposta.

A importância desta área da Estatística está relacionada com o fato de seu objetivo surgir frequentemente no mais variados campos científicos (Agronomia, Ciências Biomédicas, Biologia, Genética, Psicologia, Psicologia, Economia, etc).

Análise de Dados Categorizados

Estaremos interessandos em situações em que a variável resposta é:

1. Categórica
   - Nominal: gênero, raça, religião, status (doente/saúdável), etc
   - Ordinal: IMC (eutrófico, sobrepeso, obeso), Infecção (sem, mono ou poli), et.
2. Discreta (contagem)
   - número de cáries por pacientes;
   - número de casos de COVID;
   - número de ovos por volume de fezes, etc;

Análise de Dados Categorizados

Pesquisa Científica

1. Pergunta de Interesse;
2. Desenho do Estudo/Coleta dos Dados
3. Análise Estatística: Modelar/Predizer:
    - Conhercer o bando de Dados;
    - Análise Descritiva (cada variável separadamente);
    - Análise Bivariada (resposta vs cada covariável);
    - Modelo de Regressão (paramétrico ou não-paramétrico);
    - Inferência;
    - Resposta da Pergunta/ Interpretação dos Resultados.

Análise de Dados Categorizados

Pergunta de Interesse

- Comparação de Grupos.
- Identificação de Fatores de Risco ou Prognóstico.
- Estimação/Predição.

Análise de Dados Categorizados

Desenho do Estudo

1. Tipos de Desenho de Estudo.
2. Efeito Transversal vs Longitudinal
3. Tipo de Viés.
4. Validade do estudo.

Análise de Dados Categorizados

Algumas questões que merecem ser observadas

- Os grupos são comparáveis?
- As variáveis de confusão foram medidas/controladas?
- É possível alocar tratamento às unidades amostrais de forma aleatória?
- Os erros de medição podem ser medidos e controlados?
- As perdas (dados perdidos) podem viciar os resultados?
- Podemos estender os resultados para outros estudos?

Análise de Dados Categorizados

Tipos de Estudos

1. Estudos Transersais
2. Estudos Longitudianis
   - 2.1 Obervacionais;
     - Coorte (prospetivo ou histórico);
     - Caso-controle(retrospectivo);
   - 2.2 Experimentais: Ensaio Clínico (Cross-over)

Análise de Dados Categorizados

Estudo Transversal ou de Prevalência

Características Básicas

- Amostra tomada em um tempo pré-determinado
- Causalidade reversa (impossível determinar causa e efeito).
- Não é apropriado para estudar doenças raras e nem de curta duração.

Análise de Dados Categorizados

Estudo de Coorte

Características Básicas

- Estudos observacionais;
- Grupos de comparação (braços da coorte): usualmente definidos pela presença ou não de uma exposição de interesse;
- Podem ser prospectivos (forma mais comum) ou retrospectivo/histórico.

Análise de Dados Categorizados

Estudo Caso-Controle

Características Básicas

- Estudos observacionais e retrospectivos;
- Grupos de comparação: definidos pela presença ou não de uma doença de interesse.

Análise de Dados Categorizados

Estudo Clínico Aleatorizado

Características Básicas

- Presença de grupos de comparação
- Estudos experimentais, isto é, existe a intervenção do investigador, que consiste em aleatorizar indivíduo ao grupo;
- Vantagem: controla por fatores de confusão medidos e não medidos.

Análise de Dados Categorizados

Viés

- Desvio da verdade por defeito no delineamento ou na condução de um estudo.
- Erro sistemático no delineamento, condução e análise de um estudo, resultando em erro na estimativa da magnitude da associção entre variável explicativa e a resposta de interesse.

Análise de Dados Categorizados

Fontes de Viés

1. Fatores de confusão.
2. Viés de Seleção: alocação das unidades de análise privilegia subgrupos com probabilidade diferenciada de apresentar a resposta. Exemplo: Perda de acompanhamento em estudos longitudinais.
3. Viés de Informação: erro sistemático na classificação das variáveis sob estudo.
4. Outros: viés de publicação, etc.

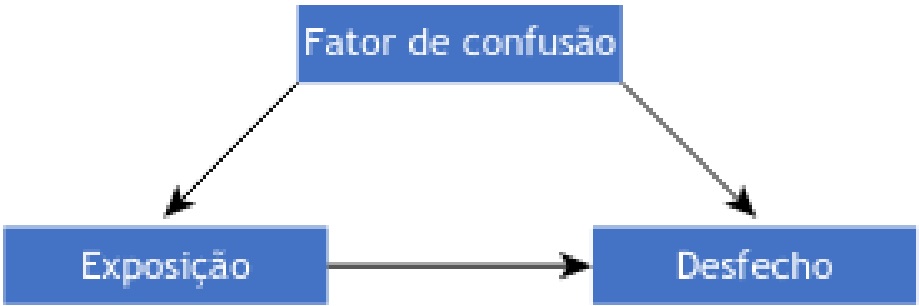
Análise de Dados Categorizados

Fator de Confusão

Definição 1
Um terceiro fator que está associado tanto com a exposição/covariável quanto com a resposta/doença, mas não se encontra no ele causal entre eles.

Análise de Dados Categorizados

Figura 1: Esquema padrão de confundimento: a variável é relacionada à exposição, é causa do desfecho e não está na via causal entre a exposição principal e o desfecho de interesse



Fonte: Fumo-dos-Santos, C.,Ferreira, J.C. Lidando com fatores de confusão em estudos observacionais. J Bras Pneumol. 2023.

Análise de Dados Categorizados

Fator de Confusão

Duas condições para caracterizar um fator de confusão:

1. Ser associado com a covariável/exposição sem ser sua consequência.
2. Estar associado com a resposta/desfecho independente da exposição.

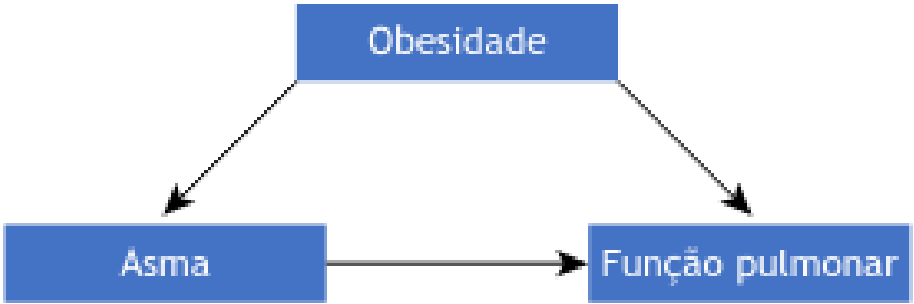
Análise de Dados Categorizados

Exemplos de Confundimento

1. Idade na associação entre fumo e câncer de estômago.
2. Fumo na associação entre consumo de café e câncer de pulmão.
3. Contra-exemplo: Colesterol na associação entre dieta e infarto.

Análise de Dados Categorizados

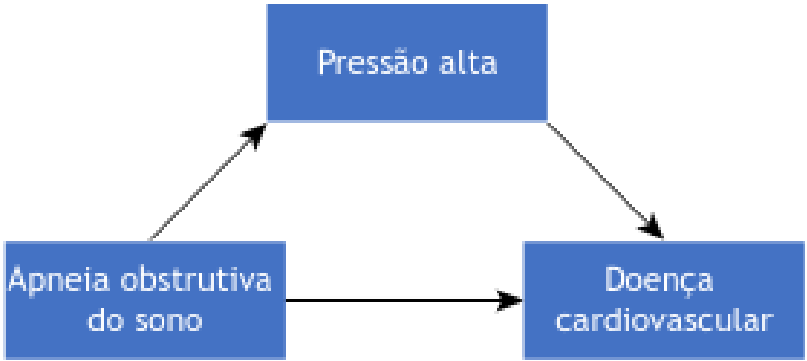
Figura 2: A obesidade é um fator de confusão na relação entre asma e função pulmonar, uma vez que a obsidade pode piorar a asma e causar redução da função pulmonar.



Fonte: Fumo-dos-Santos, C.,Ferreira, J.C. Lidando com fatores de confusão em estudos observacionais. J Bras Pneumol. 2023.

Análise de Dados Categorizados

Figura 3: Esquema de um efeito de mediação: a apneia obstruiva do sono pode levar a doenças cardioasculares (efeito direto), mas a apneia obstrutuiva do sono também pode levar à hipertensão arterial, que causa doenças cardioasculares (efeito indireto)



Fonte: Fumo-dos-Santos, C.,Ferreira, J.C. Lidando com fatores de confusão em estudos observacionais. J Bras Pneumol. 2023.

Tabelas de Contingência

Tabelas de Contingência

- Os dados dispersos ou o rol de dados, após serem classificados originam as tabelas de frequências uni, bi ou multivariadas dependendo do número de variáveis tratadas simultaneamente.
- A análise bivariada trata a distribuição conjunta de duas variáveis, enquanto a multivariada trata a distribuição conjunta de três ou mais variáveis.
- Em qualquer das tabelas explicitam-se as categorias de cada variável e o número ou a porcentagem de casos que nelas estão contidos.
- As caselas (ou células) exibem a contagem das frequências n_{ij} resultantes de uma amostra aleatória, por isso a tabela diz-se de **Contingência** ou de classificação cruzada.

Tabelas de Contingência

Tabela de Contingência 2 $\times$ 2

X	Y		Total
	$j=1$	$j=2$	
$i=1$	n_{11}	n_{12}	n_{1+}
$i=2$	n_{21}	n_{22}	n_{2+}
Total	n_{+1}	n_{+2}	n

Em que :

- n_{ij} é o número de sujeitos da amostra classificados como pertencendo simultaneamente ao nível i da variável X e ao nível j da variável Y.
- n é a dimensão da amostra tal que $\sum_i \sum_j n_{ij} = n$.

Tabelas de Contingência

- Os totais marginais representam a análise de cada variável isoladamente.
- Os respondentes escolhidos aleatoriamente de uma população têm uma distribuição de probabilidade $p_{ij} = P(X = i, Y = j)$ que é a **distribuição conjunta** das variáveis X e Y.
- A probabilidade de um indivíduo apresentar a categoria j de Y, dado que pertence à categoria i de X é denotada por $p_{(i)j} = P(Y = j|X = i)$.

Tabelas de Contingência

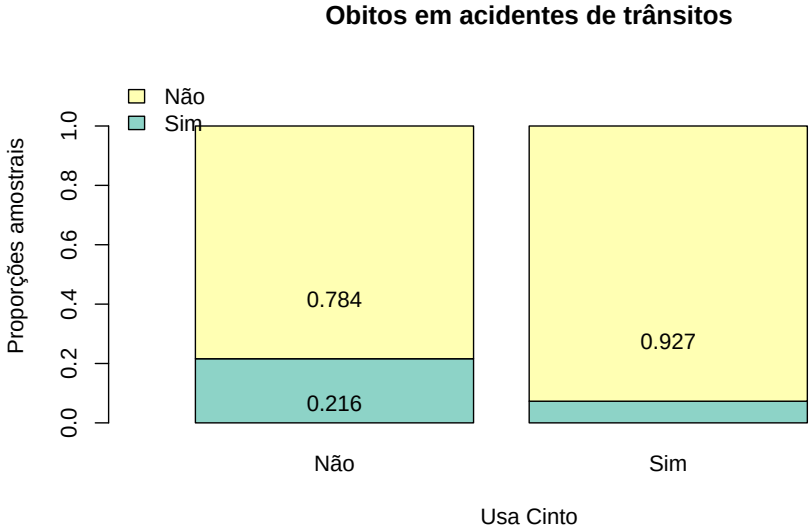
- As **distribuições marginais** correspondem à distribuição de cada uma das variáveis, dada pelos totais de linha ou de coluna, resultantes da soma da probabilidades conjuntas.
- A soma das probabilidades de uma linha é $p_{i+} = \sum_j p_{ij}$ e a soma das probabilidades de uma coluna é $p_{+j} = \sum_i p_{ij}$.
- Tanto a soma das probabilidades de todas as caselas como a soma das probabilidades marginais da linha ou coluna são iguais a 1, isto é,
$\sum_i \sum_j p_{ij} = n = \sum_i p_{ij} = \sum_j p_{ij} = 1.$

Tabelas de Contingência

- A questão mais importante numa tabela de Contingência é saber se as variáveis são ou não **independentes** e, caso se relacionam, qual o grau e o sentido dessa **associação**.

Exemplo 1
Os dados abaixo são dados de um estudo que pretende analisar o efeito do uso do cinto de segurança em acidentes de trânsito.

Usa Cinto	Morte		Total
	Sim	Não	
Não	16 $(21,6\%)$	58 $(78,4\%)$	74 (100%)
Sim	6 $(7,3\%)$	76 $(92,7\%)$	82 (100%)
Total	22 $(14,1\%)$	134 $(85,9\%)$	156 (100%)


Obitos em acidentes de trânsitos
Não
Sim
Proporções amostrais
0.0
0.2
0.4
0.6
0.8
1.0
0.784
0.216
0.927
Não
Sim
Usa Cinto

Tabelas de Contingência

- Muitas questões sobre dados categóricos podem ser respondidas estabelecendo hipóteses de associação.

Testes de Independência ou Associação

H_0 : Não existe associação entre as variáveis (são independentes)

H_a : Existe associação (são dependentes)

X	Y		Total
	$j=1$	$j=2$	
$i=1$	n_{11}	n_{12}	n_{1+}
$i=2$	n_{21}	n_{22}	n_{2+}
Total	n_{+1}	n_{+2}	n

Tabelas de Contingência

Teste Qui-Quadrado (Estatística do teste)
Supondo H_0 verdadeira,

$$Q = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{(n_{ij} - \hat{m}_{ij})^2}{\hat{m}_{ij}} \sim \chi^2_1$$

- Sob H_0, $\hat{m}_{ij} = \dfrac{n_{i+}n_{+j}}{n}$ e Q tem distribuição aproximada de Qui-quadrado com 1 grau de liberdade.
- A soma é feita em todas as caselas da tabela de contingência;
- Quando H_0 é verdadeira, n_{ij} e $\hat{m}_{ij}$ tendem a estar próximos para cada casela e assim, Q é pequeno;
- Se H_0 é falsa, pelo menos alguns valores de n_{ij} e $\hat{m}_{ij}$ tendem a não estar próximos, levando a valores maiores de $(n_{ij} - \hat{m}_{ij})$ e uma estatística de teste grande.
- **Quanto maior o valor Q, maior a evidência contra H_0: independência**

Tabelas de Contingência

Exemplo 2

Os dados abaixo trata-se do primeiro relato de um ensaio clínico que comprovou a eficácia da Zidovudina (AZT) para prolongar a vida de pacientes com AIDS. O estudo teve a duração de oito semanas.

Grupo	Situação		Total
	Vivo	Morto	
AZT	144	1	145
Placebo	121	16	137
Total	265	17	282

Existe evidência da eficácia do AZT?

Tabelas de Contingência

Código em R

```r
## exemplo1
dados <- matrix(c(144, 1, 121, 16), byrow = T, ncol=2)
Q <- chisq.test(dados)
Q
```

```
   Pearson's Chi-squared test with Yates' continuity correction

data:  dados

X-squared = 13.139, df = 1, p-value = 0.0002891
```

Tabelas de Contingência

Teste Razão de Verossimilhanças

- Alternativamente ao teste qui-quadrado de Pearson, podemos usar o teste Razão de Verossimilhança.

Estatística do Teste Razão de Verossimilhanças

$$G^2 = 2\sum_{i=1}^{2}\sum_{j=1}^{2} n_{ij}\log(\frac{n_{ij}}{\hat{m}_{ij}})$$

Tabelas de Contingência

Observações: Testes qui-quadrado e razão de verossimilhança

- A aproximação é válida para a distribuição qui-quadrado se todas as frequências esperadas forem maior que 5.
- O que fazer quando a aproximação não é adequada?
  - **Teste Exato de Fisher**.

Medindo a Associação em uma Tabela de Contingência

As principais perguntas normalmente feitas na análise de uma tabela de contingência são:

- **Existe uma associação?** O teste qui-quadrado ou o teste da razão de verossimilhança responde a essa pergunta.
- **Quão forte é a associação?** Para resumir isto usamos uma estatística para estimar a força da associação na população.
- **Como os dados diferem do que a independência prevê?** Os resíduos padronizados destacam as células que apresentam valores maiores ou menores do que o esperado sob a hipótese de independência.

> **Definição 2**
> **Medidas de associação** é uma estatística ou um parâmetro que resume a força da dependência entre duas variáveis.

Medindo a Associação em uma Tabela de Contigência

Classificação cruzada da opinião sobre uniões civis do mesmo sexo por raça

Caso A

Raça	Opinião		Total
	A favor	Contra	
Branca	360	240	600
Negra	240	160	400
Total	600	400	1000

Nenhuma Associação

- Apresenta independência estatística, representanado a associação mais fraca possível;
- Negros e brancos têm 60% a favor e 40% contra uniões civis;
- A opinião não esta associada a raça.

Caso B

Raça	Opinião		Total
	A favor	Contra	
Branca	600	0	600
Negra	0	400	400
Total	600	400	1000

Associação Máxima

- Apresenta a associação mais forte possível;
- Todos os brancos são favoráveis a uniões civis, enquanto todos os negros são contrários;
- A opinião é completamente dependente da raça.
- Para essa população, se conhecemos a sua raça saberemos sua opinião.

Medindo a Associação em uma Tabela de Contigência

O qui-quadrado não mensura a associação

- Um valor alto para o χ^2 no teste de Independência sugere que as variáveis estão associadas.
- Isto não implica que as variáveis tenham uma associação forte;
- Essa estatística simplesmente indica **quanta evidência existe** de que as variávies sejam dependentes e **não quão forte é a dependência**.
- Para uma associação dada, valores altos de χ^2 ocorrem para tamanhos de amostra maiores.

Medindo a Associação em uma Tabela de Contigência

Principais medidas de associação

- Diferença de proporções Risco Atribuível)
- Odds Ratio (Razão de chances)
- Risco relativo

Medindo a Associação em uma Tabela de Contigência

1. Diferença de proporções (ou Risco Atribuível)

- Muitas tabelas tabelas 2 x 2 comparam dois grupos em uma variável binária.
- Nesses casos, uma medida de associação útil é a diferença entre as proporções para uma categoria de resposta;
- Dada uma tabela de contingência

X	Y		Total
	$j=1$	$j=2$	
$i=1$	n_{11}	n_{12}	n_{1+}
$i=2$	n_{21}	n_{22}	n_{2+}
Total	n_{+1}	n_{+2}	n

- A diferença de proporção é estimado por

$$\hat{d} = \frac{n_{11}}{n_{1+}} - \frac{n_{21}}{n_{2+}}$$

Medindo a Associação em uma Tabela de Contigência

1. Diferença de proporções (ou Risco Atribuível)

- $d = 0$ não há diferença entre os grupos.
- $d > 0$ os indivíduos expostos ao fator de risco tem maior probabilidade de apresentar o desfecho.
- $d < 0$ os indivíduos não expostos ao fator de risco tem maior probabilidade de apresentar o desfecho.
- Por exemplo, podemos medir a diferença entre as proporções de brancos e negros que são a favor da permissão de uniões civis entre o mesmo sexo:
  - No caso A, temos:

$$\frac{360}{600} - \frac{240}{400} = 0,60 - 0,60 = 0$$

  - No caso B, temos:

$$\frac{600}{600} - \frac{0}{400} = 1$$

Medindo a Associação em uma Tabela de Contigência

1. Diferença de proporções (ou Risco Atribuível)

- A diferença das proporções populacionais é 0 sempre que as distribuições condicionais sejam idênticas, isto é, quando as variáveis são independentes.
- A diferença é 1 ou -1 para associação mais forte possível.
- Quanto maior a associação maior o valor absoluto da diferença das proporções.

Medindo a Associação em uma Tabela de Contigência

Risco $\times$ Chance

Para uma variável resposta binária, usamos sucesso para representar o resultado de interesse e fracasso o outro resultado: Assim:

- O **risco** de ocorrer o sucesso é a probabilidade do sucesso ocorrer.
- A **chance (Odds)** de ocorrer o sucesso é a razão da probabilidade do sucesso pela probabilidade do fracasso ocorrer.

Assim, se o sucesso tem probabilidade π de ocorrer, temos:

$$Risco = \pi$$

e

$$Chance = \frac{\pi}{1-\pi}$$

Medindo a Associação em uma Tabela de Contigência

Obs:
Uma odds igual a 2 significa que o sucesso é duas vezes mais possível de ocorrer que o fracasso, enquanto que uma odds de 0,25 significa que o fracasso é quatro vezes mais possível do que o sucesso.

Medindo a Associação em uma Tabela de Contingência

2. Odds Ratio (Razão de chances)

- Ao comparar a resposta de duas populações independentes, por exemplo, casos/controles, com/sem um fator prognóstico ou comparar dois tratamentos suas odds (chances) são comparadas.
- A **Razão de chances ou Odds ratio** é uma razão entre a odds (chance) de ocorrência do sucesso em grupo e a odds (chance) de ocorrência do sucesso em outro grupo.
- Se π_1 e π_2 são as probabilidades de sucesso de duas populações, então a Odds ratio é

$$\theta = \frac{Odds_1}{Odds_2} = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)}$$

- A Odds ratio é mais informativa para comparar π_1 e π_2 que sua diferença.

Medindo a Associação em uma Tabela de Contingência

2. Odds Ratio (Razão de chances)

- Por exemplo, para $\pi_1 = 0,9$, $\pi_2 = 0,8$ e $\pi_1 = 0,6$, $\pi_2 = 0,5$ temos que em ambos $\pi_1 - \pi_2 = 0,1$ enquanto suas odds ratio são 2,25 e 1,5, respectivamente.
- Em termos da distribuição conjunta de uma tabela de contingência 2×2 , θ é equivalentemente definida por

$$\theta = \frac{\pi_{11}/\pi_{12}}{\pi_{21}/\pi_{22}} = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}}$$

- $\theta = 1$ é equivalente a $\pi_1 = \pi_2$, isto é, as variáveis são independentes.
- $\theta > 1$ ou $\theta < 1$ corresponde a uma dependência positiva ou negativa, respectivamente.
- A dependência se torna mais forte a medida que θ se afasta de 1.

Medindo a Associação em uma Tabela de Contingência

2. Odds Ratio (Razão de chances)

- A Odds ratio não depende das distribuições marginais das variáveis e por isso é uma boa medida de associação.
- A odds ratio amostral é

$$\hat{\theta} = \frac{p_{11}p_{22}}{p_{12}p_{21}} = \frac{n_{11}n_{22}}{n_{12}n_{21}}$$

- $\hat{\theta}$ toma valores no intervalo $[0, \infty]$;
- $\hat{\theta} = 0$ ou $\hat{\theta} = \infty$ quando frequência de alguma casela é igual a zero no numerador ou denominador, respectivamente;
- $\hat{\theta} = 0$ é indefinido quando ambas as caselas de uma linha ou coluna são nulas. Uma forma clássica de tratar essa situação é adicionar 0,5 as frequências das caselas.

Medindo a Associação em uma Tabela de Contingência

2. Odds Ratio (Razão de chances)

Intervalo de Confiança para Odds Ratio

- Pode se provar que, para uma amostra aleatória, $\ln \hat{\theta}$ tem uma aproximação normal melhor do que $\hat{\theta}$;
- Assim, a inferência é feita em termos de $\ln \theta$;
- Em particular, podemos provar que, assintoticamente

$$\ln \hat{\theta} \sim N\left(\ln \theta, \frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}\right)$$

- Além disso, na escala logarítmica, a interpretação é mais simples:
  - $\ln \theta = 0$ corresponde a independência;
  - Valores positivo (negativo) de $\ln \theta$ corresponde a dependência positiva (negativa);
  - A força da associação é o incremento em $|\theta|$

Medindo a Associação em uma Tabela de Contingência

2. Odds Ratio (Razão de chances)

Intervalo de Confiança para Odds Ratio

O intervalo de confiança assintótico $(1-\alpha)100\%$ para θ fica

$$\left(e^{\ln\hat{\theta} - z_{\frac{\alpha}{2}}\sqrt{var[\ln\hat{\theta}]}}; e^{\ln\hat{\theta} + z_{\frac{\alpha}{2}}\sqrt{var[\ln\hat{\theta}]}} \right)$$

em que:

- $z_{\frac{\alpha}{2}}$ é percentil de ordem $\frac{\alpha}{2}$ da distribuição normal padrão;
- $$var[\ln\hat{\theta}] = \frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}$$

Medindo a Associação em uma Tabela de Contingência

2. Odds Ratio (Razão de chances)

Intervalo de Confiança para Odds Ratio

- Se o intervalo de confiança de θ inclui o 1 **não há associação** estatisticamente significante entre a exposição e o desfecho.
- Se o intervalo de confiança de θ não inclui o 1 **há associação** estatisticamente significante entre a exposição e o desfecho.
  - Se o intervalo está todo **acima de 1**, a exposição é um fator de risco.
  - Se o intervalo está todo **abaixo de 1** a exposição é um fator de proteção.

Medindo a Associação em uma Tabela de Contingência

2. Odds Ratio (Razão de chances)

Exemplo 3

Exemplo Ferimentos graves em crianças envolvidas em acidentes automobilísticos (análise de 413 acidentes automobilísticos envolvendo crianças)

Cinto	Ferimentos Graves		Total
	Sim	Não	
Não	50	240	290
Sim	16	107	123
Total	66	347	413

$$\hat{\theta} = \frac{n_{11}n_{22}}{n_{12}n_{21}} = \frac{50 \times 107}{240 \times 16} = 1,40$$

Não usar o cinto de segurança aumenta em cerca de 40% a chance de ferimentos graves em acidentes automobilísticos com crianças

Medindo a Associação em uma Tabela de Contingência

3. Risco Relativo(RR)

Risco Relativo
O Risco Relativo (RR) é uma razão entre o risco de ocorrência do sucesso em um grupo e o risco de ocorrência de sucesso em outro grupo. Isto é, é a probabilidade que um indivíduo do grupo A ter o evento relativa a probabilidade de que um indivíduo do grupo B desenvolver o evento.

O Risco relativo amostral é

$$\widehat{RR} = \frac{\frac{n_{11}}{n_{1+}}}{\frac{n_{21}}{n_{2+}}} = \frac{n_{11}n_{2+}}{n_{21}n_{1+}}$$

Medindo a Associação em uma Tabela de Contingência

3. Risco Relativo(RR)

Intervalo de Confiança para Risco Relativo

- Pode se provar que, para uma amostra aleatória, $\ln \widehat{RR}$ tem uma aproximação normal melhor do que $\widehat{RR}$;
- Assim, a inferência é feita em termos de $\ln RR$;
- Em particular, podemos provar que, assintoticamente

$$\ln \widehat{RR} \sim N\left(\ln RR, \frac{1}{n_{11}} - \frac{1}{n_{1+}} + \frac{1}{n_{21}} - \frac{1}{n_{2+}}\right)$$

- Assim, o intervalo de confiança assintótico $(1-\alpha)100\%$ para RR fica

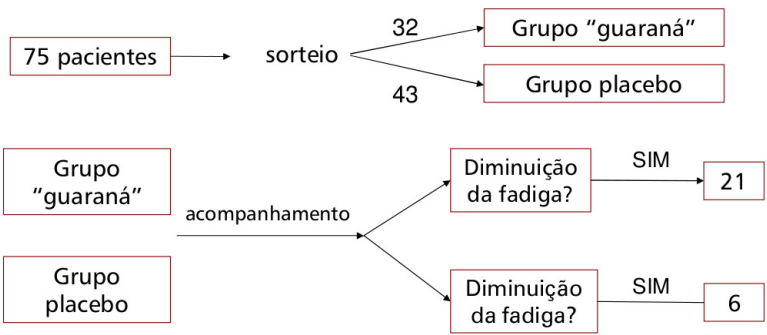
$$\left(e^{\ln \widehat{RR} - z_{\frac{\alpha}{2}}\sqrt{var[\ln \widehat{RR}]}}; e^{\ln \widehat{RR} + z_{\frac{\alpha}{2}}\sqrt{var[\ln \widehat{RR}]}}\right)$$

Medindo a Associação em uma Tabela de Contingência

3. Risco Relativo(RR)

Exemplo 4

Estudo para verificar se a ingestão de extrato de guaraná tem efeito sobre a fadiga em pacientes tratados com quimioterapia (Hospital Albert Einstein e Faculdade de Medicina do ABC, 2010).



Medindo a Associação em uma Tabela de Contigência

3. Risco Relativo(RR)

Exemplo - continuação

Grupo	Diminuição da Fadiga		Total
	Não	Sim	
Guaraná	11	21	32
Placebo	37	06	43
Total	48	27	75

$$\widehat{RR}_{G/P} = \frac{21/32}{6/43} = \frac{903}{192} = 4,70$$

Assim, o risco de diminuir a fadiga no grupo que toma guaraná é 4,70 vezes do que o grupo que toma placebo.

Medindo a Associação em uma Tabela de Contingência

3. Risco Relativo(RR)

- $RR \approx 1 \Longrightarrow$ **Ausência** de risco
- $0 < RR << 1 \Longrightarrow$ **diminui** o risco do desfecho entre aqueles do Primeiro Grupo. **(Fator de proteção)**
- $RR >> 1 \Longrightarrow$ **aumenta** o risco do desfecho entre aqueles do Primeiro Grupo. **(Fator de risco)**

Nota

- Estimativas de risco só podem ser feitas quando partimos da exposição e observamos o evento.
- Isto impossibilita o uso do RR em estudos que partem da ocorrência do evento e depois observam a exposição.

Medindo a Associação em uma Tabela de Contigência

Risco relativo(RR) vs Odds ratio

- O risco relativo (RR) e Odds ratio (OR) são duas medidas de associação diferentes que quantifica a associação entre duas variáveis qualitativas.
- O risco relativo é mais intuitivo, enquato que a Odds ratio tem outras propriedades desejaveis:
  - O risco relativo não pode ser estimado em estudo de caso-controle, de maneira geral, não pode ser estimado em estudos que parte da ocorrência do evento e depois observam a exposição.
  - A Odds ratio pode ser estimada em qualquer tipo de estudos (caso-controle, coorte, ensaios clínicos, estudos transversais);
  - A distribuição amostral da OR é mais simples que do RR, geralmente é preferível trabalhar com a odds ratio.

Medindo a Associação em uma Tabela de Contigência

Risco relativo(RR) vs Odds ratio

- Para doenças raras (eventos raros), a Odds ratio é uma boa aproximação para o risco relativo. Isto é, quando a probabilidade da doença é muito baixa ($< 10\%$) nos dois grupos comparados, $OR \approx RR$.
- A estimativa da Odds ratio é mais fácil de ser calculada, no entanto, não definida quando a tabela possui caselas nulas.
- Neste caso, usa se o procedimento usual de somar 0,5 a todas as caselas da tabela.
- RR não exibe boas propriedades matemáticas, ao contrario de RR:
- OR é invariante na rotação da tabela, isto é, ao mudar a categoria de referencia o novo OR fica $\frac{1}{OR}$.
- O RR não possui essa propriedade.

Tabelas de Contingência r $\times$ s

Tabelas de Contingência r $\times$ s

- Tabelas 2 $\times$ 2 são estendidas naturalmente para tabelas de dimensões maiores, chamadas r $\times$ s.
- As estatísticas qui-quadrado, sob H_0, têm distribuição qui-quadrado com $(r-1) \times (s-1)$ graus de liberdade.
- Em geral, os dados referem-se a mensurações de duas características (X e Y) feitas em n unidades experimentais, que são apresentadas conforme a seguinte tabela:

Tabelas de Contingência r $\times$ s

Variável(X)	Variável(Y)				Totais X
	1	2	$\cdots$	s	
1	n_{11}	n_{12}	$\cdots$	n_{1s}	n_{1+}
2	n_{21}	n_{22}	$\cdots$	n_{2s}	n_{+}
$\vdots$	$\cdots$	$\cdots$	$\vdots$	$\cdots$	$\cdots$
r	n_{r1}	n_{r2}	$\cdots$	n_{rs}	n_{r+}
Totais Y	n_{+1}	n_{+2}	$\cdots$	n_{+s}	n

- O teste procura verificar se os níveis da variável X exerce alguma influência sobre os níveis da variável Y.

Tabelas de Contingência r $\times$ s

Estatística do teste
Supondo H_0 verdadeira,

$$Q = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{(n_{ij} - \hat{m}_{ij})^2}{\hat{m}_{ij}} \sim \chi^2_q$$

- Sob H_0, $\hat{m}_{ij} = \dfrac{n_{i+}n_{+j}}{n}$ e Q tem distribuição aproximada de Qui-quadrado com $q = (r-1) \times (s-1)$ graus de liberdade.
- A soma é feita em todas as caselas da tabela de contingência;
- Quando H_0 é verdadeira, n_{ij} e $\hat{m}_{ij}$ tendem a estar próximos para cada casela e assim, Q é pequeno;
- Se H_0 é falsa, pelo menos alguns valores de n_{ij} e $\hat{m}_{ij}$ tendem a não estar próximos, levando a valores maiores de $(n_{ij} - \hat{m}_{ij})$ e uma estatística de teste grande.
- **Quanto maior o valor Q, maior a evidência contra H_0: independência**

Tabelas de Contingência r $\times$ s

Exemplo 5

A reação ao tratamento por quimioterapia está sendo estudada em quatro grupos de pacientes com câncer. Deseja-se investigar se todos os tipos reagem da mesma maneira. Uma amostra de pacientes de cada grupo foi escolhida ao acaso e classificou-se a reação em três categorias:

Câncer	Reação			Totais
	Pouca	Média	Alta	
Tipo I	51	33	16	100
Tipo II	58	29	13	100
Tipo III	48	42	30	120
Tipo IV	26	38	16	80
Totais	183	142	75	400

Tabelas de Contingência r $\times$ s

Medidas de Associação: Odds Ratio

- A razão de chances (Odds ratio), θ, é uma medida de associação para uma tabela 2×2 de grande importância.
- É também a base para detectar estruturas de associação em tabelas $r \times s$.
- Para isso, é necessário a decomposição da tabela tabela $r \times s$ em um conjunto de tabelas 2×2.
- Em geral, para uma tabela $r \times s$ temos um conjunto de $(r-1)(s-1)$ tabelas 2×2, e as razões de chances correspondentes descrevem a associação subjacente.

Tabelas de Contingência r $\times$ s

Medidas de Associação: Odds Ratio

- Para variáveis nominais este conjunto de tabelas 2×2 é definido em termos de uma categoria de referência, geralmente a casela (r, s).

- Então as tabelas 2×2 formadas possuem em sua casela superior da diagonal a casela (i, j) e na casela inferior da diagonal a casela de referência (r, s).

- Assim, as **razões de chances nominais** são definidas como

$$\theta_{ij}^{rs} = \frac{n_{ij} n_{rs}}{n_{rj} n_{is}}, \qquad i = 1, ..., r-1 \qquad j = 1, ..., s-1.$$

Tabelas de Contingência r $\times$ s

Medidas de Associação: Odds Ratio

- Ao obervar uma amostra, a razão de chances nominal da amostra é dada por

$$\hat{\theta}_{ij}^{rs} = \frac{n_{ij}n_{rs}}{n_{rj}n_{is}}, \qquad i = 1, ..., r-1 \qquad j = 1, ..., s-1.$$

- Qualquer casela (a, b) da tabela poderia servir como categoria de referência e as razões de chances nominais são então definidas de forma análoga.

Tabelas de Contingência r $\times$ s

Medidas de Associação: Odds Ratio

- Diferentes tipos de razão de chances são adequados para variáveis ordinais.
- Fixar uma casela de referência não é adequado.
- Uma escolha mais natural é comparar cada nível da variável ordinal com o próximo imediado, ou cada nível com os eventos que estão no mesmo nível ou acima dele.
- As tabelas 2×2 são formadas por duas linhas sucessivas i e $i+1$ e duas colunas sucessivas j e $j+1$. $i = 1, ..., r-1 \quad j = 1, ..., s-1$.

Tabelas de Contingência r $\times$ s

Medidas de Associação: Odds Ratio

- Dessa forma são formadas $(r-1)(s-1)$ tabelas locais e as odds ratios correspondentes são chamadas de **odds ratios locais**, dadas por

$$\theta_{ij}^{L} = \frac{\pi_{ij}\pi_{i+1,j+1}}{\pi_{i+1,j}\pi_{i,j+1}}, \quad i = 1, ..., r-1 \quad j = 1, ..., s-1.$$

- Ao obervar uma amostra, a razão de chances local da amostra é

$$\widehat{\theta}_{ij}^{L} = \frac{n_{ij}n_{i+1,j+1}}{n_{i+1,j}n_{i,j+1}}, \quad i = 1, ..., r-1 \quad j = 1, ..., s-1.$$

- Para uma tabela de contingência r $\times$ s a hipótese de independência é equivalente à hipótese de que todas as razões de chances são iguais a 1.
- Em termos das odds ratios locais, é equivalente a

$$\theta_{ij}^{L} = 1, \quad i = 1, ..., r-1 \quad j = 1, ..., s-1.$$

Tabelas de contingência 2 x 2 x K

Tabelas de contingência 2 x 2 x K

Exemplo: Dados hipotético referente a resposta (Y) a um tratamento (X) em duas clínicas diferentes (Z)

Clínica (Z)	Medicamento (X)	Resposta (Y)	
		Sucesso	Fracasso
1	A	18	12
	B	12	8
2	A	2	8
	B	8	32

Tabelas de contingência 2 x 2 x K

Tabela parcial XY dado Z = 1

Medicamento (X)	Resposta (Y)	
	Sucesso	Fracasso
A	18	12
B	12	8

Tabela parcial XY dado Z = 2

Medicamento (X)	Resposta (Y)	
	Sucesso	Fracasso
A	2	8
B	8	32

Tabela marginal XY, ignorando o efeito de Z

Medicamento (X)	Resposta (Y)	
	Sucesso	Fracasso
A	20	20
B	20	40

Tabelas de contingência 2 x 2 x K

- Considere uma tabela de contingência $2 \times 2 \times K$, com duas variáveis binárias, X e Y, classificadas entre os K níveis de uma variável explicativa Z.

- Cada tabela parcial XY, em cada nível k de Z, tem-se as probabilidade condicionais $\pi_{ij|k}$ associada.

- A razão de chance (odds ratio) pode ser definida para cada uma das tabelas de probabilidades condicionais, como

$$\theta_{(k)}^{XY} = \frac{\pi_{11k}\pi_{22k}}{\pi_{12k}\pi_{21k}}, \qquad k = 1, ..., K$$

- Esta razão de chance é chamada de razão de chance condicional.

Tabelas de contingência 2 x 2 x K

- A razão de chance (odds ratio) correspondente a tabela de probabilidade marginal, (π_{ij+}), é chamada de razão de chance marginal, e dada portanto

$$\theta^{XY} = \frac{\pi_{11+}\pi_{22+}}{\pi_{12+}\pi_{21+}}$$

- As razões de chance marginais e condicionais expressam a associação entre as variáveis e são estimadas pelas razões de chance amostrais correspondentes, dadas por:

$$\hat{\theta}^{XY}_{(k)} = \frac{n_{11k}n_{22k}}{n_{12k}n_{21k}}, \qquad \hat{\theta}^{XY} = \frac{n_{11+}n_{22+}}{n_{12+}n_{21+}}$$

Tabelas de contingência 2 x 2 x K

- A razão de chance condicional expressa a associação parcial XY controlando o nível de Z.
- A razão de chance marginal expressa a associação marginal XY, ignorando os efeitos de Z.
- As razões de chance condicionais podem diferir substancialmente em k, mesmo na direção da associação.
- Nesse caso, a razão de chance marginal será enganosa para descrever a associação XY, que deve ser captada pelas razões de chances condicionais, levando em consideração a variável explicativa Z.

Tabelas de contingência 2 x 2 x K

- X e Y são **condicionalmente independente** dado Z se elas são independentes em cada tabela parcial

- No caso de tabelas $2 \times 2 \times K$ isto significa

$$\theta_{(1)}^{XY} = \theta_{(2)}^{XY} = ... = \theta_{(k)}^{XY} = 1$$

- A condição anterior geralmente não implica $\hat{\theta}^{XY} = 1$ (razão de chance marginal)

- $\hat{\theta}^{XY} = 1$ corresponde a independência marginal de X e Y.

Tabelas de contingência 2 x 2 x K

- Se X e Y tem uma associação idêntica e todos os níveis de Z, dizemos que X e Y têm **associação homogênea** dado Z.

- No caso de tabelas $2 \times 2 \times K$ isto significa que todas as tabelas parciais a mesma razão de chances,

$$\theta_{(1)}^{XY} = \theta_{(2)}^{XY} = ... = \theta_{(k)}^{XY}$$

- Independência condicional é um caso especial de associação homogênea.

Teste de Mantel-Haenszel

Teste de Mantel-Haenszel

- O objetivo é testar a associação entre duas variáveis, digamos X e Y, controlando uma terceira variável, Z.
- A terceira variável define os estratos e o teste de Mantel-Haenszel os combina em um único teste e um valor de Odds ratio.
- Esse teste é conhecido como teste de Independência Condicional: X independe de Y, dado Z.

Teste de Mantel-Haenszel

Teste de Mantel-Haenszel

- X e Y são variáveis binárias
- Z é um variável categórica ou categorizada com k níveis

Tabela 1: Tabela parcial XY dado $Z = i$

X	Y		Total
	1	2	
1	n_{11i}	n_{12i}	n_{1+i}
2	n_{21i}	n_{22i}	n_{2+i}
Total	n_{+1i}	n_{+2i}	n_i

Teste de Mantel-Haenszel

Para testar

- $H_0 : XY$ são condicionalmente Independente em todos os níveis de Z
- $H_1 : XY$ não são Independentes em pelo menos um nível de Z

A estatística de Mantel-Haenszel (MH), para k tabelas, é dada por:

$$MH = \frac{(|\sum_{i=1}^{k}(n_{11i} - \mu_{11i})| - 0,5)^2}{\sum_{i=1}^{k}\sigma^2_{11i}} \sim \chi^2_1$$

em que:

$\mu_{11i} = \hat{E}(n_{11i}) = \frac{n_{1+i}n_{+1i}}{n_i}$ e $\sigma^2_{11i} = v\hat{a}r(n_{11i}) = \frac{n_{1+i}n_{+1i}n_{2+i}n_{+2i}}{n_i^2(n_i-1)}$

Teste de Mantel-Haenszel

- Esta é a versão do teste de MH com correção de continuidade.
- Basta retirar o termo 0,5 do numerador para ter a versão usual.
- Se todos os $\theta_{(k)}^{XY} = 1$ então MH é pequeno (perto de zero)
- Se todos os $\theta_{(k)}^{XY} < 1$ ou $\theta_{(k)}^{XY} > 1$ então MH é grande
- Se para alguns $\theta_{(k)}^{XY} > 1$ e para outros $\theta_{(k)}^{XY} < 1$ o teste de MH não é apropriado.
- O teste funciona bem e é mais poderoso quando $\theta_{(k)}^{XY}$ estão na mesma direção e são de tamanhos comparáveis.

Teste de Mantel-Haenszel

- Para tabelas 2×2, quando $\theta_{(1)}^{XY} = \theta_{(2)}^{XY} = ... = \theta_{(k)}^{XY}$ um estimador de Mantel-Haenszel de um valor comum para a razão de chances é

$$\hat{\theta}_{MH} = \frac{\sum_{i=1}^{K} \frac{n_{11i} n_{22i}}{n_i}}{\sum_{i=1}^{K} \frac{n_{12i} n_{21i}}{n_i}}$$

- $\hat{\theta}_{MH}$ é chamada de razão de chances combinado para a associação entre X e Y, ou simplesmente de razão de chances de Mantel-Haenszel.

Especialização em *Data Science* e Estatística Aplicada

Módulo IV - Análise de Dados Categóricos

Prof Dr Márcio Augusto Ferreira Rodrigues

marcioaugusto@ufg.br

IME INSTITUTO DE MATEMÁTICA E ESTATÍSTICA

FEN FACULDADE DE ENFERMAGEM

UFG UNIVERSIDADE FEDERAL DE GOIÁS