Tarefa 4
Análise Biavariada de dados Qualitativos

Douglas Rodrigues
Karina Yaginuma

Universidade Federal Fluminense

- Leia o conteúdo e faça os exercícios dos slides.
- Você deve entregar:
  1. Um relatório, em fomato pdf, contendo todos os gráficos e planilhas elaboradas, com as respostas dos questionamentos feitos ao longo da tarefa.
  2. Os comandos utilizados.

Análise Bivariada - Variáveis Qualitativas

- Agora considere que queremos estudar a relação entre variáveis qualitativas.
- Podemo utilizar as mesma metodologia aplicada às variáveis quantitativas?

Análise Bivariada - Variáveis Qualitativas

- Agora considere que queremos estudar a relação entre variáveis qualitativas.
- Podemo utilizar as mesma metodologia aplicada às variáveis quantitativas?
- Não!!
- Como proceder neste caso?

Tabela de Contingência

- Para entender como duas variáveis qualitativas se relacionam, podemos construri uma tabela de contingência.
- Contém a informação da distribuição de frequência conjunta de duas variáveis qualitativas.
- Também chamada de **distribuição de frequência conjunta**.
- Para construir uma tabela de contigência no **R**, podemos utilizar a função `table()`.

Exercício 1

- Importe os dados do arquivo `funcionarios.xlsx`. Faça os ajustes nos parâmetros necessários para importar os dados.
- Utilize a função `table()` para contruir a tabela de contingência das variáveis Estado civil e Grau de instrução.

```
> tabela <- table(funcionarios$‘Estado civil‘,
funcionarios$‘Grau de instrução‘)
```

Estado Civil\ Grau de instrução	F	M	S	Total
casado	5	12	3	20
solteiro	7	6	3	16
Total	12	18	6	36

Tabela: Tabela de contingência: frequência absoluta

Exercício 1

- Note que se utilizar o comando as.data.frame, não conseguimos a tabela desejada para exportação.

```
> tabela1 <-
as.data.frame(table(funcionarios$‘Estado
civil‘,funcionarios$‘Grau de instrução‘))
```

- Solução:

```
> tabela2 <- matrix(tabela[1:6], nrow=2, ncol=3)
# cria a matriz de frequência
> colnames(tabela2) <- c("fundamental", "médio",
"superior") # nomeia as colunas
> row.names(tabela2) <- c("casado", "solteiro")
textcolorgreen# nomeia as linhas
```

Exportando dados - pacote `xlsx`

- Uma maneira de exportar dados direto em formato `xlsx` é utilizando a função `write.xlsx` do pacote `xlsx`.

```
> write.xlsx(tabela2, file = "tabelas.xlsx",
sheetName = "EstadoCivil_Instrução", append = F)
```

- O comando gera um arquivo "tabelas.xlsx":
  - `sheetName`: nome da planilha;
  - `append`: se `FALSE` cria um arquivo novo, se `TRUE` exporta a informação para um arquivo já existente com nome definido pelo argumento `file`.

Exercício 2

- Instale o pacote `xlsx`.
- Utilize a função `write.xlsx` para exportar os dados da `tabela2`, nomeio o arquivo `tabelas.xlsx`.

Tabela de Contingência

- Ao invés de trabalharmos com as frequências absolutas, podemos construir tabelas com as frequências relativas (proporções), como foi feito para o caso de uma única variável.
- Note que temos três possibilidades de expressarmos a proporção de cada célula (de acordo com o objetivo de cada pesquisa, uma delas será mais conveniente):
  - em relação ao total geral;
  - em relação ao total de cada linha;
  - em relação ao total de cada coluna.

Exemplo

Tabela: frequência relativa em relação ao total geral

E. Civil\ G. Instrução	F	M	S	Total
casado	14%	33.3%	8.3%	55.6%
solteiro	19.4%	16.7%	8.3%	44.4%
Total	33.4%	50%	16.6%	100%

Tabela: frequência relativa em relação ao total da linha

E. Civil\ G. Instrução	F	M	S	Total
casado	25%	60%	15%	100%
solteiro	43.8%	37.5%	18.8%	100%
Total	33.4%	50%	16.6%	100%

Exemplo

Tabela: frequência relativa em relação ao total da coluna

E. Civil\ G. Instrução	F	M	S	Total
casado	41.7%	66.7%	50%	55.6%
solteiro	58.3%	33.3%	50%	44.4%
Total	100%	100%	100%	100%

- Note que as interpretações de cada tabela são diferentes. Por exemplo,
  - na primeira tabela, 14% dos funcionários são casados e possuem ensino fundamental;
  - na segunda tabela, 25% dos funcionários casados possuem ensino fundamental;
  - na terceira tabela, 58.3% dos funcionários que possuem ensino fundamental são casados.

Função tabpct()

- A função tabpct() gera tabelas de contingências (proporção em relação a linha e coluna) e automaticamente um gráfico mosaico.
- A função faz parte do pacote epiDisplay.

```
> tabela3 <- tabpct(funcionarios$‘Estado
civil‘,funcionarios$‘Grau de instrução‘,
main="Tabela de contingência", ylab="Grau de
instrução", xlab = "Estado civil")
```

Exercício 3

- Utilize a função `tabpct` para construir as tabelas de contigências com relação ao total das linhas e das colunas.
- Utilize a função `write.xlsx` para exportar as duas tabelas para o arquivo `tabelas.xlsx`, criada anteriormente.

Observações: Note que o objeto `tabela3` é uma lista, para acessar a primeira tabela utilize o comando `tabela3[[1]]`. E para acessar as segunda tabela `tabela3[[2]]`.

Gráfico de Mosaico

- Representação gráfica das proporções de cada categoria de uma variável separada pelas categorias de outra variável.

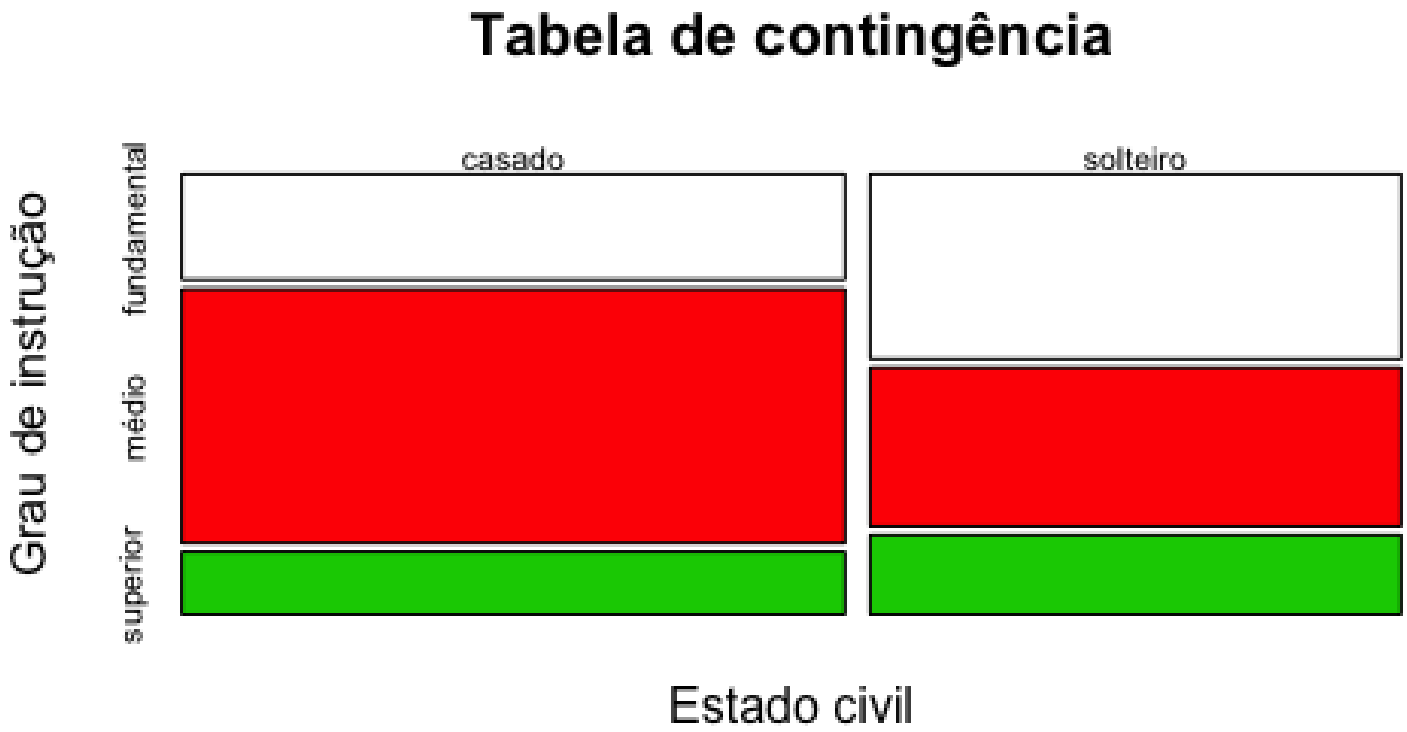

Gráfico de Mosaico

- O gráfico é gerado automaticamente pela função `tabpct`. Também podemos usar a função `mosaicplot`.

```
> mosaicplot(table(funcionarios$‘Estado civil‘,
funcionarios$‘Grau de instrução‘), col =
c("green", "blue", "red"), main = "Gráfico de
Mosaico")
```

- O primeiro argumento deve ser a tabela de contingência das variáveis de interesse.

Exercício 4

- Gere utilizando a função `mosaicplot`, o gráfico de mosaico das variávies Estado Civil e Grau de instrução.
- O que podemos concluir do gráfico?

Associação entre variáveis qualitativas

- Um dos principais objetivos da distribuição conjunta é descrever a relação existente entre as variáveis.
- Como no caso das variáveis quantitativas, queremos determinar o grau de relação entre as variáveis.

Motivação

- Por exemplo, suponhamos que sorteamos uma pessoa ao acaso da população da cidade de São Paulo, e devemos adivinhar qual o sexo desta pessoa.

- Como aproximadamente a metade da população é sexo femino e a outra metade é do sexo masculino, não temos preferência em sugerir qualquer um dos dois sexos.

Motivação

- Se a mesma pergunta fosse feita, e nos fosse dito que a pessoa sorteada trabalha na indústria siderúrgica, qual seria a sua resposta agora?

- **Resposta:** Seríamos inclinados a sugerir que a pessoa é do sexo masculino, pois há uma predominância deste sexo neste ramo de ocupação.

Motivação

- Se a informação adicional fosse que a pessoa sorteada leciona no ensino fundamental, qual seria a sua resposta?

- **Resposta:** A nossa sugestão seria modificada, pois a grande maioria dos professores do ensino fundamental é do sexo feminino.

Comentários

- Isso porque existe um grau de relação entre as variáveis sexo e ramo de ocupação.
- Vejamos, agora, como podemos identificar a existência de uma relação ou não entre duas variáveis, através da distribuição conjunta (tabela de contingência).

Exemplo

Observando a tabela, é possível verificar algum tipo de relação entre as variáveis?

Curso \ Sexo	Masculino	Feminino	Total
Economia	85(61%)	35(58%)	120(60%)
Administração	55(39%)	25(42%)	80(40%)
Total	140(100%)	60(100%)	200(100%)

Exemplo

- Podemos observar que as proporções do sexo masculino (61% e 39%) e do sexo feminino (58% e 42%) são próximas das proporções geral (60% e 40%).
- Estes resultados indicam que não há relação entre as duas variáveis, ou seja, as variáveis sexo e curso parecem ser independentes.

Exemplo

- Agora considere os cursos de física e ciências sociais, cuja distribuição conjunta é dada pela seguinte tabela.

Curso \ Sexo	Masculino	Feminino	Total
Física	100(71%)	20(33%)	120(60%)
Ciênciais Sociais	40(29%)	40(67%)	80(40%)
Total	140(100%)	60(100%)	200(100%)

Exemplo

- Comparando a distribuição das proporções pelos cursos, independente do sexo (coluna de total), com as distribuições diferenciadas por sexo (coluna de sexo masculino e de sexo feminino), observamos uma disparidade bem acentuada nas proporções.
- Note que há uma maior concentração de homens no curso de física e de mulheres no curso de ciências sociais.
- Portanto, neste caso, existem evidências de uma relação entre as variáveis sexo e curso.

Observação

- Convém observar que teríamos obtido as mesmas conclusões se tivéssemos calculado as proporções, mantendo constantes os totais das linhas.
- **Como quantificar essa relação?**

Qui-Quadrado

- Queremos quantificar a relação entre duas variáveis qualitativas utilizando a tabela de contingência.
- Considere o exemplo anterior.

Curso \ Sexo	Masculino	Feminino	Total
Física	100(71%)	20(33%)	120(60%)
Ciênciais Sociais	40(29%)	40(67%)	80(40%)
Total	140(100%)	60(100%)	200(100%)

- Concluímos que há evidência de uma relação entre as duas variáveis.
- Caso não houvesse, esperaríamos que a distribuição da variável curso, independente do sexo, fosse 60% para física e 40% para ciências sociais.

- Se assumirmos indepêndencia, esperaríamos a seguinte distribuição conjunta:

Curso \ Sexo	Masculino	Feminino	Total
Física	84(60%)	36(60%)	120(60%)
Ciênciais Sociais	56(40%)	24(40%)	80(40%)
Total	140(100%)	60(100%)	200(100%)

- Valores esperados quando supomos independência.

Qui-Quadrado

- Comparando os valores observados e esperados, verificamos discrepâncias entre os valores.
- Podemos quantificar essas discrepâncias através dos desvios entre observados (o_i) e esperados (e_i).
- Desvio: $d_i = o_i - e_i$.
  - o_i: representa o valor observado;
  - e_i: representa o valor esperado;
  - i : índice que representa a célula.

Tabela: Desvio

Curso \ Sexo	Masculino	Feminino	Total
Física	16	-16	0
Ciênciais Sociais	-16	16	0
Total	0	0	0

- Note que a soma total dos desvios é nula.
- **Solução:** desvios ao quadrado, ou seja, desvio$^2 = (o_i - e_i)^2$.

- Note que todos os desvios são idênticos.
- Mas é evidente que estes desvios possuem pesos diferentes.
- Para lidar com este problema, vamos utilizar o desvio relativo:

$$dr_i = \frac{(o_i - e_i)^2}{e_i}.$$

Tabela: Desvio relativo

Curso \ Sexo	Masculino	Feminino	Total
Física	3.05	7.11	10.16
Ciênciais Sociais	4.57	10.67	15.24
Total	7.62	17.78	25.4

Qui-Quadrado

- Uma medida de afastamento global é dada pela soma dos desvios relativos.
- Essa medida é conhecida como medida Qui-Quadrado:

$$\mathcal{X}^2 = \sum_i \frac{(o_i - e_i)^2}{e_i},$$

onde o somatório é estendido a todas as células da tabela.

- No exemplo anterior, temos que $\mathcal{X}^2 = 25.4$.

Qui-Quadrado

- Note que, quanto maior for o valor de Qui-Quadrado, maior será o grau de relação existente entre as duas variáveis.
- A medida Qui-Quadrado, satisfaz

$$0 \leq \mathcal{X}^2 < \infty.$$

- Portanto, fica muito difícil, baseando-se na magnitude do valor Qui-Quadrado julgar se a relação é alta ou não.

Coeficiente de Contingência

- K. Pearson propôs o chamado coeficiente de contingência C, definido por

$$C = \sqrt{\frac{\mathcal{X}^2}{\mathcal{X}^2 + n}},$$

onde n é o número de observações.

- Teoricamente,

$$0 \leq C \leq 1.$$

- Esse coeficiente será nulo quando as variáveis não são associadas e portanto $\mathcal{X}^2 = 0$.

Coeficiente de Contingência

- Entretanto, mesmo quando existe uma relação perfeita, C não é necessariamente igual a 1.
- Para resolver este problema, podemos definir o **coeficiente de contingência modificado**, dado por

$$C^* = C\sqrt{\frac{t}{t-1}},$$

onde $t = \min\{\text{números de linhas, números de colunas}\}$, da tabela de contingência. Temos que

$$0 \leq C^* \leq 1.$$

Coeficiente de contingência modificado

- Usualmente C^* acima de 0.5 indicaria uma relação moderada para forte, o que bastaria para considerar que existe relação entre as variáveis.
- No exemplo, temos

$$\begin{aligned} \mathcal{X}^2 &= 25.4 \\ C &= \sqrt{\frac{25.4}{25.4+200}} = 0.33 \\ C^* &= 0.33\sqrt{\frac{2}{1}} = 0.47. \end{aligned}$$

Cálculo no **R**

- Para calcular o Qui-Quadrado utilize a função `chisq.test` (a função realiza o teste qui-quadrado de Pearson, que será estudado mais adiante no curso).

```
> q <- chisq.test(funcionarios$‘Estado
civil‘,funcionarios$‘Grau de instrução‘)
> q$statistic # fornece o valor da estatística
Qui-Quadrado
```

- Para calcular o Coeficiente de Contingência, precisamos instalar o pacote `DescTools`, e usar o comando `ContCoef()`.

```
> install.packages("DescTools")
> require("DescTools")
> ContCoef(funcionarios$‘Estado civil‘,
funcionarios$‘Grau de instrução‘)
```

Exercício 5

- Calcule a medida Qui-Quadrado para as variáveis Estado Civil e Grau de instrução.
- Calcule também os coeficientes de contingência e de contingência modificado.
- Com base nestes valores o que podemos concluir?