DESCRIÇÃO DE UMA NUMÉRICA, SEGUNDO UMA CATEGÓRICA

Por fazer:
- clarificar fragmentos pouco claros, e corrigir problemas de redação
- completar e desenvolver quando for necessário
- pensar que faltaria incluir (exemplos, etc.)
- o trabalho com o apêndice pode ficar para um próximo passo

Principal referência (até agora):
- OpenIntro (ler seções correspondentes)

Também
- Bryman: Capítulo 15 na edição 5th: Quantitative Data Analysis
- https://oli.cmu.edu/jcourse/workbook/activity/page?context=90d50f4a80020ca601343b8eb604eef1

Nesta secção exploramos algumas opções estudar a relação entre uma variável independente categórica e uma variável dependente igualmente categórica

C -> C

Também aqui, o procedimento consiste em enfocar na variável dependente aplicando as mesmas ferramentas da análise univariada, mas dividendo a amostra segundo os diferentes grupos da variável categórica.

1. Antes da Análise: pergunta e dados

Pergunta de investigação

> *As ministras mulheres têm mais expertise tecnica dos ministros homens?*

A pergunta inclui 1 variável independente categórica (variável sexo, com 2 categorias) e uma variável dependente categórica (especialistas vs. não-especialistas).

Dados

Continuamos trabalhando com a base BaseTutorial_1, agora focando na variável sexo e especialistas.

Controlamos a classe:

```
class(B$especialista)
```

```
[1]"factor"
```

Revisamos as categorias

```
table(B$especialista)
```

```
      especialista não-especialista
                 7                 9
```

Vemos que a variável regista 7 casos de especialistas e 9 casos de não-especialistas.

Controlamos pela existência de valores em falta:

```
table (is.na(B$especialista))
```

```
```
FALSE
  16

2. Dados 'crus' (desagregados)

Revisamos as categorias

```
B [ , c("sexo","especialista")]
```

Description: df [16 × 2]

sexo <fctr>	especialista <fctr>
female	não-especialista
male	especialista
male	especialista
female	não-especialista
female	não-especialista
male	especialista
female	especialista
male	especialista
male	não-especialista
male	especialista
female	especialista
female	não-especialista
male	não-especialista
male	não-especialista
female	não-especialista
female	não-especialista

16 rows

Como vimos nos capítulos anteriores, os dados aparecem por sujeito. No primeiro caso temos uma ministra, do sexo feminino, que não é especialista. No segundo caso, temos um ministro, do sexo masculino, que é especialista.

3. Análise

Interessa-nos saber se a proporção de especialistas é significativamente diferente entre os dois grupos da variável **sexo**.

Tabela de contingência

As tabelas de dupla entrada, também chamadas de tabelas de contingência ou tabelas de cruzamento, permitem analisar como as observações se distribuem entre as categorias de duas variáveis simultaneamente.

Frequências absolutas:

```
t2 <- table (B$sexo,B$especialista)
t2
addmargins(t2)
```

```
        especialista não-especialista
female             2                6
male               5                3

        especialista não-especialista Sum
female             2                6   8
male               5                3   8
Sum                7                9  16
```

*A tabela apresenta as categorias da variável **sexo** nas linhas e as categorias da variável **especialista** nas colunas. As células na interseção mostram o número de indivíduos que combinam essas categorias.*

- *Por exemplo, a célula (1,1) indica que **2 mulheres** são especialistas, enquanto a célula (2,2) indica que **3 homens** não são especialistas.*
- *A soma horizontal revela que a amostra é composta por **8 mulheres** e **8 homens**.*
- *A soma vertical mostra que, no total, há **7 especialistas** e **9 não-especialistas**.* [1]

Um olhar mais atento ao cruzamento das categorias revela:

- ***2 mulheres** são especialistas e **6 mulheres** não o são.*
- ***5 homens** são especialistas e **3 homens** não o são.*

[1] En las tablas de contingencia de R, las celdas se leen como (fila, columna):

- El primer número indica la línea (ou seja, a categoria da variável listada nas linhas).
- O segundo número indica a coluna (ou seja, a categoria da variável listada nas colunas).

Por exemplo:

- (1,1) corresponde à interseção da primeira linha com a primeira coluna.
- (1,2) corresponde à interseção da primeira linha com a segunda coluna.
- (2,2) corresponde à interseção da segunda linha com a segunda coluna.

Frequências relativas e percentagens condicionais:

```
addmargins (prop.table (t2,1)*100,2)
```

```
             especialista não-especialista   Sum
female               25.0             75.0 100.0
male                 62.5             37.5 100.0
```

A tabela de percentagens condicionais evidencia a relação entre as variáveis:

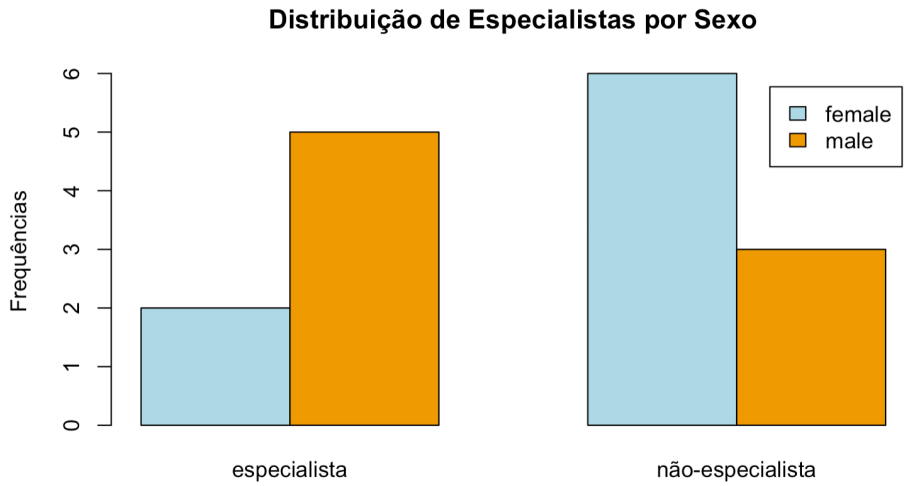
- *Entre as **mulheres**, **25%** são especialistas, enquanto **75%** não o são.*
- *Entre os **homens**, **62,5%** são especialistas e **37,5%** não o são.*

*Estas percentagens são condicionais, pois refletem a distribuição da variável dependente (**especialista**) em função das categorias da variável independente (**sexo**).*
***Nota:** Quando a variável independente é colocada nas linhas, as percentagens condicionais são chamadas de **percentagens por linha** (row percents), calculadas de forma independente para cada linha.*

Barplot

A representação gráfica ajuda a interpretar os dados de forma mais intuitiva. O código abaixo gera um gráfico de barras que destaca as proporções em cada categoria:

```
barplot(t2,
        beside = TRUE,                    # Barras agrupadas lado a lado
        col = c("lightblue", "orange2"), # Cores por sexo
        legend.text = rownames(t2),    # legenda com as categorias de sexo
        main = "Distribuição de Especialistas por Sexo",
        ylab = "Frequências")
```



4. Teste Estatístico

Quando analisamos dados categóricos, nem sempre as diferenças observadas entre as variáveis são suficientes para indicar uma relação real entre elas. É possível que as diferenças sejam fruto do acaso, sem refletir um padrão consistente na população. Para avaliar se as associações observadas têm um significado real, ou seja, se são pouco prováveis de ocorrer por acaso, utilizamos testes estatísticos.

Um desses testes é o **teste qui-quadrado de Pearson (χ2)**, que basicamente calcula se as diferenças entre os dados observados e os valores esperados, caso as variáveis sejam independentes, são suficientemente grandes para serem consideradas significativas. Se o valor do χ2 for elevado, isso sugere que há uma relação entre as variáveis; se for baixo, não há evidências suficientes para afirmar que existe uma associação.

Este tipo de análise é fundamental para evitar conclusões precipitadas. Mesmo que as porcentagens entre homens e mulheres pareçam diferentes, o teste estatístico pode mostrar que essas diferenças não são fortes o suficiente para descartar a possibilidade de acaso. Assim, a análise ajuda a garantir que nossas conclusões sejam baseadas em evidências robustas, em vez de impressões ou coincidências.

Calcular a estatística do qui-quadrado e o valor-p.

```
chisq.test(t2)
```

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

data:  t2
X-squared = 1.0159, df = 1, p-value = 0.3135
```

As Frequências Esperadas (se não houvesse associação entre as variáveis)

```
chisq.test(t2)$expected
```

```
        especialista não-especialista
female          3.5              4.5
male            3.5              4.5
```

O valor de χ2 obtido foi 1,0159, com um valor-p de 0,3135. Isso significa que, sob a hipótese nula (que assume que não há associação entre as variáveis), há uma probabilidade de aproximadamente 31,35% de observarmos uma diferença tão extrema quanto a observada, apenas por acaso. Como o valor-p é maior que 0,05, não rejeitamos a hipótese nula (H_0), ou seja, não há evidências suficientes para afirmar que existe uma associação significativa entre as variáveis. [2]

[2] Ao interpretar testes estatísticos, lembrar-se de:
a) Conferir sempre o valor-p para determinar se a associação é significativa.
b) Analisar as frequências esperadas para garantir que os dados atendem às condições do teste (e.g., valores esperados maiores que 5 em todas as células).

CALCULO DO TESTE

Condições para aplicar o teste qui-quadrado:
Versão 1

a) **Independência das observações**: As observações (indivíduos ou casos) devem ser independentes. Não deve haver repetição ou dependência entre os dados.
b) **Tamanho da amostra adequado**: O número de observações em cada célula da tabela de contingência deve ser suficiente para garantir a validade do teste. A condição mais importante aqui é que **os valores esperados em cada célula da tabela de contingência devem ser maiores que 5**.
    - Se algumas células tiverem valores esperados menores que 5, é comum usar o **teste exato de Fisher**, que não exige essa condição.

Versão 2:

Condições para a aplicabilidade do teste de qui-quadrado, de acordo com esta fonte:

1. **Nenhum dos valores esperados é menor ou igual a 1:**
    - Este es un requisito más estricto que el de "mayores que 5". En lugar de decir que todos los valores esperados deben ser mayores que 5, esta condición establece que **ningún valor esperado puede ser menor o igual a 1**. Esto se hace para garantizar que las celdas con valores muy pequeños no sesguen el resultado del test.
2. **O total de valores esperados com valor inferior ou igual a 5 representa menos de 20% do total:**
    - Esta es una condición adicional más flexible. Si algunos valores esperados son menores que 5, la suma total de estos valores debe ser inferior al 20% de todos los valores esperados. Este enfoque permite cierto margen de flexibilidad si solo unos pocos valores son pequeños, pero de todos modos asegura que no haya una gran cantidad de valores pequeños que puedan afectar la fiabilidad del test.

PASSOS

1. Construção da tabela de contingência:

```
```
t2 <- table (B$sexo,B$especialista)
```
```

Sexo / Especialista	Especialista	Não Especialista	Total
Mulher	3	5	8
Homem	4	4	8
Total	7	9	16

2. Cálculo das frequências esperadas:
Calcular as **frequências esperadas** para cada célula, com base na hipótese nula (que assume que as variáveis são independentes).
A fórmula para calcular as frequências esperadas é:

$$E_{ij} = \frac{(F_{i\text{total}}) \times (F_{\text{j total}})}{F_{\text{total}}}$$

Onde:

- E_{ij} é a frequência esperada na célula i, j.
- $F_{i\text{total}}$ é a soma das frequências observadas da linha i.
- $F_{\text{j total}}$ é a soma das frequências observadas da coluna j.
- F_{total} é o número total de observações.

No exemplo, se quisermos calcular a frequência esperada para as mulheres especialistas:

$$E_{\text{mulheres, especialistas}} = \frac{(8 \text{ mulheres}) \times (7 \text{ especialistas})}{16 \text{ total}}$$

$$E_{\text{mulheres, especialistas}} = \frac{8 \times 7}{16} = 3.5$$

- **Frequência esperada para mulheres especialistas:**

$$E_{\text{mulheres, especialistas}} = \frac{(8 \text{ mulheres}) \times (7 \text{ especialistas})}{16 \text{ total}} = \frac{8 \times 7}{16} = 3.5$$

- **Frequência esperada para mulheres não especialistas:**

$$E_{\text{mulheres, não especialistas}} = \frac{(8 \text{ mulheres}) \times (9 \text{ não especialistas})}{16 \text{ total}} = \frac{8 \times 9}{16} = 4.5$$

- **Frequência esperada para homens especialistas:**

$$E_{\text{homens, especialistas}} = \frac{(8 \text{ homens}) \times (7 \text{ especialistas})}{16 \text{ total}} = \frac{8 \times 7}{16} = 3.5$$

- **Frequência esperada para homens não especialistas:**

$$E_{\text{homens, não especialistas}} = \frac{(8 \text{ homens}) \times (9 \text{ não especialistas})}{16 \text{ total}} = \frac{8 \times 9}{16} = 4.5$$

Assim, as frequências esperadas são:

Sexo / Especialista	Especialista	Não Especialista	Total
Mulher (feminino)	3.5	4.5	8
Homem (masculino)	3.5	4.5	8
Total	7	9	16

Com R

```
chisq.test(t2)$expected
```

Se cumple a condição >5?

3. Cálculo da estatística chi-quadrado:
Calcular a estatística do teste de chi-quadrado utilizando a fórmula:

$$\chi^2 = \sum_i \sum_j \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Ou seja, para cada célula da tabela, subtraímos a frequência observada da frequência esperada, elevamos ao quadrado essa diferença e dividimos pela frequência esperada. Em seguida, somamos esses valores para todas as células.

Agora vamos aplicar essa fórmula às células da tabela:

- **Para mulheres especialistas:**

$$\frac{(O_{\text{mulheres, especialistas}} - E_{\text{mulheres, especialistas}})^2}{E_{\text{mulheres, especialistas}}} = \frac{(3 - 3.5)^2}{3.5} = \frac{(-0.5)^2}{3.5} = \frac{0.25}{3.5} \approx 0.0714$$

- **Para mulheres não especialistas:**

$$\frac{(O_{\text{mulheres, não especialistas}} - E_{\text{mulheres, não especialistas}})^2}{E_{\text{mulheres, não especialistas}}} = \frac{(5 - 4.5)^2}{4.5} = \frac{(0.5)^2}{4.5} = \frac{0.25}{4.5} \approx 0.0556$$

- **Para homens especialistas:**

$$\frac{(O_{\text{homens, especialistas}} - E_{\text{homens, especialistas}})^2}{E_{\text{homens, especialistas}}} = \frac{(4 - 3.5)^2}{3.5} = \frac{(0.5)^2}{3.5} = \frac{0.25}{3.5} \approx 0.0714$$

- **Para homens não especialistas:**

$$\frac{(O_{\text{homens, não especialistas}} - E_{\text{homens, não especialistas}})^2}{E_{\text{homens, não especialistas}}} = \frac{(4 - 4.5)^2}{4.5} = \frac{(-0.5)^2}{4.5} = \frac{0.25}{4.5} \approx 0.0556$$

Agora, somamos todos esses valores para obter o valor de χ^2:

$$\chi^2 = 0.0714 + 0.05\text{E} \downarrow \cdot 0.0714 + 0.0556 \approx 0.2540$$

4. Graus de liberdade:
Os **graus de liberdade** (df) são calculados pela fórmula:

$$df = (r - 1)(c - 1)$$

Onde:

- r é o número de linhas (neste caso, 2: mulheres e homens).
- c é o número de colunas (neste caso, 2: especialistas e não especialistas).

Portanto, neste exemplo:

$$df = (2 - 1)(2 - 1) = 1$$

Graus de Liberdade: Conceito e Ilustração com Dados
Os **graus de liberdade** no teste qui-quadrado indicam quantos valores na tabela podem variar livremente, dado que os totais marginais (as somas das linhas e colunas) já estão fixados.

Exemplo com os nossos dados:
Estamos analisando a relação entre **sexo** (*mulheres*, *homens*) e **especialidade** (*especialista*, *não especialista*). Nossa tabela tem o seguinte formato:

Sexo	Especialista	Não Especialista	Total
Mulher	3	5	8
Homem	4	4	8
Total	7	9	16

Essa é uma tabela 2×2, pois temos duas categorias para cada variável.

Os graus de liberdade são calculados como:

$$df = (2-1)(2-1) = 1$$

Isso significa que, em nossa tabela, apenas **1 célula** pode variar livremente, dado que os totais das linhas e colunas já estão determinados.

Entendendo com as frequências esperadas:
As frequências esperadas para cada célula são calculadas com a fórmula:

$$\text{Esperado} = \frac{\text{Total da linha} \times \text{Total da coluna}}{\text{Total geral}}$$

Para nossa tabela:

- **Célula 1,1** (*mulheres, especialistas*):

$$\text{Esperado} = \frac{8 \times 7}{16} = 3,5$$

- **Célula 1,2** (*mulheres, não especialistas*):

$$\text{Esperado} = \frac{8 \times 9}{16} = 4,5$$

- **Célula 2,1** (*homens, especialistas*):

$$\text{Esperado} = \frac{8 \times 7}{16} = 3,5$$

- **Célula 2,2** (*homens, não especialistas*):

$$\text{Esperado} = \frac{8 \times 9}{16} = 4,5$$

Portanto, a tabela com os valores esperados seria:

Sexo	Especialista	Não Especialista	Total
Mulher	3,5	4,5	8
Homem	3,5	4,5	8

Total	7	9	16

O que significa 1 grau de liberdade?
No contexto dessa tabela2×2:

1. Podemos ajustar **livremente apenas 1 célula** da tabela. Por exemplo, se escolhermos o valor da célula 1 1,1 (*mulheres, especialistas*), os demais valores serão automaticamente determinados pelos totais das linhas e colunas.
2. O grau de liberdade reflete a quantidade de informação que realmente "varia" antes de sermos limitados pelos totais marginais.

Por que isso é importante?
O grau de liberdade é essencial para determinar o valor crítico da estatística qui-quadrado. Em nosso exemplo, com df=1, o valor crítico para um nível de significância de 5% (p=0,05) é aproximadamente 3,841. Se o valor calculado de $\chi 2$ for maior que esse valor, rejeitamos a hipótese nula de independência. Caso contrário, não rejeitamos a hipótese nula.

4. Valor-p e Decisão

Depois de calcular a estatística chi-quadrado, comparamos o valor calculado com o valor crítico da distribuição chi-quadrado correspondente ao nível de significância (α) (por exemplo, 0,05) e aos graus de liberdade.

- Se o valor de $\chi 2$ calculado for **maior que o valor crítico**, rejeitamos a hipótese nula e concluímos que há uma **associação significativa** entre as variáveis.
- Se o valor de $\chi 2$ calculado for **menor que o valor crítico**, **não rejeitamos** a hipótese nula, o que significa que não há evidências suficientes para afirmar que existe uma associação entre as variáveis.

Com a estatística chi-quadrado ($\chi 2 \approx 0.25409$ e os graus de liberdade (df = 1), podemos comparar o valor calculado com o valor crítico da distribuição chi-quadrado para o nível de significância de 0,05 (que corresponde a um valor crítico de aproximadamente 3,841).

- Como $\chi 2 = 0.2540$ é **menor que 3.841**, **não rejeitamos** a hipótese nula.

Dado que o valor de $\chi 2$ calculado é menor que o valor crítico, **não rejeitamos a hipótese nula** de independência entre as variáveis sexo e especialização. Isso significa que não há **evidência suficiente** para sugerir que o sexo e a especialização dos ministros estão **associados**.

Com r

```
chisq.test(t2)
```

O valor de $\chi 2$ obtido foi 1,0159, com um valor-p de 0,3135. Isso significa que, sob a hipótese nula (que assume que não há associação entre as variáveis), há uma probabilidade de aproximadamente 31,35% de observarmos uma diferença tão extrema quanto a observada, apenas por acaso. Como o valor-p é maior que 0,05, não rejeitamos a hipótese nula (H_0), ou seja, não há evidências suficientes para afirmar que existe uma associação significativa entre as variáveis. [3]

[3] Ao interpretar testes estatísticos, lembrar-se de:

a) Conferir sempre o valor-p para determinar se a associação é significativa.
b) Analisar as frequências esperadas para garantir que os dados atendem às condições do teste (e.g., valores esperados maiores que 5 em todas as células).

5. EXERCÍCIOS

Secção PANORAMA

1) Identificar e escolher 5 artigos vinculados a um tópico do seu interesse, que utilizem análise estatística.
2) Para cada um dos 5 artigos escolhidos, identificar que tipo de análise aplicam (descritivo, explicativo) e que técnicas e ferramentas (coeficinetes, regressões, boxplots, etc.).
3) Para 2 dos 5 artigos escolhidos,
   a) identificar e descrever a base de dados utilizados (unidade de observação, número de observações, variáveis) e
   b) reconstruir o codebook incluindo pelo menos 5 variáveis.

Secção UMA QUANTITATIVA

4) Para 1 dos 5 artigos escolhidos, identificar a descrição de uma variável numérica. Transcrever essa descrição.

Secção UMA QUALITATIVA

5) Para 1 dos 5 artigos escolhidos, identificar a descrição de uma variável qualitativa. Transcrever essa descrição.

Secção ANALISE BIVARIADA

6) Para 1 dos 5 artigos escolhidos, identifique a variável independente e a variável dependente. Classifique-as (e.g., categórica, numérica).
7) Agora escolha um artigo com uma variável independente categórica e uma variável dependente categórica. Defina os grupos de cada variável categórica.
8) Faça uma tabela de contingência e analise de que forma as observações se distribuem pelas diferentes células.
9) Por fim, realize um teste estatístico para determinar a existência de associação entre as duas variáveis.

6. Referencias

Openintro
Bryman
xxxx

APÊNDICE (R)
[Ver markdown]