MAT02018 - Estatística Descritiva

Distribuições bidimensionais

Rodrigo Citton P. dos Reis
citton.padilha@ufrgs.br

Universidade Federal do Rio Grande do Sul
Instituto de Matemática e Estatística
Departamento de Estatística

Porto Alegre, 2022

Apresentação

Apresentação

- Até o presente momento vimos como organizar, resumir e apresentar informações referentes a uma **única variável**.
- Muito do interesse em se coletar diversas variáveis em um levantamento estatístico está em analisar o **comportamento conjunto** de **duas ou mais variáveis**.
- Algumas questões de pesquisa, relacionadas a este objetivo, que podem ser listadas:
  - As variáveis estão relacionadas **(associadas, correlacionadas)**?
  - Como caraterizar esta realação?
  - Qual a força desta relação?
- Nestas breves notas de aula vamos apresentar formas de organização e apresentação de **distribuições bidimensionais**[1] de frequências, assim como o cálculo e interpretação de medidas resumo.

[1] Distribuições de duas variáveis. Nestas notas de aula nos concentraremos no caso de duas variáveis, mas tais ideias podem ser generalizados para o caso de mais que duas variáveis.

Introdução

Introdução

- Quando consideramos duas variáveis[2], podemos ter três situações:
  - As duas variáveis são qualitativas;
  - As duas variáveis são quantitativas;
  - Uma variável é qualitativa e outra é quantitativa.
- As técnicas de análise de dados nas três situações são diferentes.

[2] Ou ainda, dois grupos em que as mesmas variáveis foram mensuradas, formando dois conjuntos de dados. Exemplo: a atividade física é mensurada (por algum instrumento de medida: questionário de hábitos de vida; acelerômetro) em dois grupos de indivíduos: fumantes e não-fumantes.

Introdução

- Quando as variáveis são qualitativas, os dados são resumidos e apresentados em **tabelas de dupla entrada**[3].
- Quando as duas variáveis são quantitativas, **gráficos de dispersão** são apropriados.
- Quando temos uma variável qualitativa e outra quantitativa, em geral, analisamos o que acontece com a variável quantitativa agrupada em classes[4].

[3] Também conhecidas como **tabelas de contingência**.

[4] Ou seja, a variável qualitativa é interpretada como uma variável de grupo de observações.

Introdução

- Contudo, em todas as situações, o objetivo é encontrar as possíveis relações **(associações)** entre as duas variáveis.
- Essas relações podem ser detectadas por meio de métodos gráficos e medidas resumo.
- Interpretaremos a existência de associação como uma **"mudança"** de opinião sobre o comportamento de uma variável na presença de informação sobre a segunda variável.

Introdução

Exemplificando

Os clientes de um serviço de **Streaming** têm a **comédia** como gênero preferido de filme. Se estratificamos os clientes entre *jovens* e *idosos*, esta preferência se mantém a mesma nos dois grupos?

- Se a preferência por gênero de filme muda entre as faixas etárias, então temos uma **associação** entre as variáveis **idade** e **preferência por gênero de filme**.

Tal informação pode ser utilizada para a definição de campanhas de *marketing* específicas para cada faixa etária, com sugestões "mais precisas".

Variáveis qualitativas

Variáveis qualitativas

- Considere como exemplo o conjunto de dados coletados de empregados da seção de orçamentos da Companhia MB (BUSSAB; MORETTIN, 2017). Suponha que queiramos analisar o comportamento conjunto das variáveis **grau de instrução** e **região de procedência**.

Tabela 1: Tabela de dados brutos.

ID	Região de Procedência	Grau de Instrução
1	interior	ensino fundamental
2	capital	ensino fundamental
3	capital	ensino fundamental
4	outra	ensino médio
5	outra	ensino fundamental
6	interior	ensino fundamental
7	interior	ensino fundamental
8	capital	ensino fundamental
9	capital	ensino médio
10	outra	ensino médio

Variáveis qualitativas

11	interior	ensino médio
12	capital	ensino fundamental
13	outra	ensino médio
14	outra	ensino fundamental
15	interior	ensino médio
16	outra	ensino médio
17	capital	ensino médio
18	outra	ensino fundamental
19	interior	superior
20	interior	ensino médio
21	outra	ensino médio
22	capital	ensino médio
23	outra	ensino fundamental
24	outra	superior
25	interior	ensino médio
26	outra	ensino médio
27	outra	ensino fundamental
28	interior	ensino médio
29	interior	ensino médio
30	capital	ensino médio
31	outra	superior
32	interior	ensino médio

Variáveis qualitativas

33	capital	superior
34	capital	superior
35	capital	ensino médio
36	interior	superior

Variáveis qualitativas

- A **distribuição conjunta de frequências** é apresentada logo a seguir, na tabela de dupla entrada:

Tabela 2: Distribuição conjunta das frequências das variáveis grau de instrução e região de procedência.

Região de Procedência	Grau de Instrução			
	ensino fundamental	ensino médio	superior	Total
capital	4	5	2	11
interior	3	7	2	12
outra	5	6	2	13
Total	12	18	6	36

Variáveis qualitativas

Note que cada elemento do corpo da tabela fornece a frequência observada das **realizações simultâneas** de **Região de Procedência** (X) e **Grau de Instrução** (Y).

- Assim, observamos quatro indivíduos da capital com ensino fundamental, sete do interior com ensino médio, etc.
- A **linha dos totais** fornece a **distribuição (unidimensional)** da variável **Grau de Instrução**, ao passo que a **coluna dos totais** fornece a **distribuição** da variável **Região de Procedência**.
- As distribuições assim obtidas são chamadas tecnicamente de **distribuições marginais**, enquanto a tabela do slide anterior constitui a **distribuição conjunta**[5] de X e Y.

[5] Podemos concluir que a partir da distribuição conjunta das variáveis (n_{ij}) é possível obter as distribuições marginais de cada uma das variáveis, somando os elementos da linha ($n_{i\cdot} = \sum_{j=1}^{J} n_{ij},\ i = 1, \ldots, I$), ou da coluna ($n_{\cdot j} = \sum_{i=1}^{I} n_{ij},\ j = 1, \ldots, J$). Por outro lado, não é possível obter a distribuição conjunta a partir das distribuições marginais.

Variáveis qualitativas

- Assim como no caso unidimensional, podemos ter interesse não só em **frequências absolutas**, mas também em **frequências relativas** e **porcentagens**.
- Mas aqui existem três possibilidades de expressarmos a proporção de cada casela:
  - Em relação ao total geral;
  - Em relação ao total de cada linha;
  - Em relação ao total de cada coluna.
- De acordo com o objetivo do problema em estudo, uma delas será a mais conveniente.

Variáveis qualitativas

Table 1. Socio-demographics and clinical characteristics of individuals with self-reported diabetes overall and according to glycated hemoglobin (A1C) level: Laboratory subsample, National Health Survey—Brazil, 2014–2015.

Characteristic	Overall	A1C (%)				*p* Value †
		<7	7 to <8	8 to <9	≥9	
Overall, *n* (%)	465 (100)	225 (46.0)	73 (19.8)	60 (14.1)	107 (20.1)	
Sex, *n* (%)						0.12
Female	308 (61.5)	158 (49.9)	40 (15.9)	37 (12.8)	73 (21.4)	
Male	157 (38.5)	67 (39.6)	33 (26.0)	23 (16.3)	34 (18.1)	
Age group (years), *n* (%)						0.005
18–44	53 (11.5)	25 (25.3)	6 (26.7)	4 (15.7)	18 (32.3)	
45–59	168 (38.1)	70 (40.4)	24 (16.4)	18 (13.7)	56 (29.5)	
60+	244 (50.4)	130 (54.9)	43 (20.8)	38 (14.1)	33 (10.2)	
Race, *n* (%)						0.31
White	191 (52.1)	96 (48.3)	36 (22.4)	23 (12.5)	36 (16.9)	
Black or Brown	274 (47.9)	129 (43.5)	37 (17.0)	37 (15.9)	71 (23.6)	

Fonte: dos Reis, R.C.P.; Duncan, B.B.; Szwarcwald, C.L.; Malta, D.C.; Schmidt, M.I. Control of Glucose, Blood Pressure, and Cholesterol among Adults with Diabetes: The Brazilian National Health Survey. *J. Clin. Med.* **2021**, *10*, 3428. https://doi.org/10.3390/jcm10153428

Variáveis qualitativas

Table 1

Sociodemographic and clinical characteristics of adult participants in the *Brazilian National Health Survey*, 2013 and 2019.

Characteristic	2013				2019			
	n	Total n (%)	Men n (%)	Women n (%)	n	Total n (%)	Men n (%)	Women n (%)
Age (years)	60,202				88,531			
18-24		7,823 (15.9)	3,467 (16.7)	4,356 (15.3)		8,145 (13.9)	3,864 (14.9)	4,281 (13.0)
25-34		13,923 (21.7)	5,877 (22.4)	8,046 (21.0)		15,970 (18.1)	7,606 (18.8)	8,364 (17.5)
35-44		12,817 (19.2)	5,545 (18.9)	7,272 (19.4)		18,033 (20.2)	8,735 (20.2)	9,298 (20.3)
45-54		10,246 (17.5)	4,633 (17.5)	5,613 (17.5)		15,885 (17.8)	7,608 (17.7)	8,277 (18.0)
55-64		7,681 (13.4)	3,276 (13.1)	4,405 (13.7)		14,572 (15.0)	6,857 (14.8)	7,715 (15.3)
≥ 65		7,712 (12.3)	3,122 (11.4)	4,590 (13.0)		15,926 (14.9)	6,992 (13.6)	8,934 (16.0)
Race/Skin color	60,199				88,522			
White		24,106 (47.6)	10,226 (47.0)	13,880 (48.1)		32,409 (43.3)	15,126 (42.5)	17,283 (43.9)
Black		5,631 (9.1)	2,525 (9.1)	3,106 (9.2)		10,132 (11.5)	4,943 (11.6)	5,189 (11.4)
Mixed-race		29,512 (42.0)	12,796 (42.7)	16,716 (41.3)		44,646 (43.8)	20,950 (44.3)	23,696 (43.3)
Asian		533 (0.9)	203 (0.8)	330 (1.0)		665 (0.9)	308 (0.9)	357 (0.9)
Indigenous		417 (0.4)	169 (0.4)	248 (0.5)		670 (0.5)	329 (0.6)	341 (0.5)
Education level	60,202				88,531			
Incomplete elementary		24,083 (39.0)	10,834 (39.9)	13,249 (38.2)		35,572 (34.8)	17,857 (35.5)	17,715 (34.1)
Complete elementary		9,215 (15.5)	4,062 (16.5)	5,153 (14.6)		12,005 (14.5)	5,933 (15.7)	6,072 (13.4)
Complete high shool		19,149 (32.8)	7,916 (32.2)	11,233 (33.3)		27,337 (34.9)	12,342 (34.6)	14,995 (35.2)
Complete higher education		7,755 (12.7)	3,108 (11.4)	4,647 (13.9)		13,617 (15.8)	5,530 (14.2)	8,087 (17.3)
Body mass index (kg/m²)	59,402				87,678			
Low/Normal (< 25)		25,446 (43.1)	11,455 (44.5)	13,991 (41.8)		36,908 (42.0)	17,271 (40.8)	19,637 (43.2)
Overweight (25-29.9)		21,598 (36.1)	10,095 (38.7)	11,503 (33.8)		32,744 (36.7)	16,739 (40.0)	16,005 (33.8)
Obesity (≥ 30)		12,358 (20.8)	4,370 (16.8)	7,988 (24.4)		18,026 (21.2)	7,652 (19.2)	10,374 (23.1)

Fonte: Reis, R. C. P. dos, Duncan, B. B., Malta, D. C., Iser, B. P. M., & Schmidt, M. I. (2022). Evolution of diabetes in Brazil: prevalence data from the 2013 and 2019 Brazilian National Health Survey. *Cadernos de Saúde Pública*, *38*(suppl 1). https://doi.org/10.1590/0102-311x00149321

Variáveis qualitativas

- A tabela a seguir apresenta a distribuição conjunta das frequências relativas, expressas como proporções do total geral.
- Ou seja, cada elemento da tabela abaixo é obtido pela divisão da frequência absoluta (apresenta na tabela anterior) pelo total de observações, $n = 36$ (e multiplicação por 100).

Tabela 3: Distribuição conjunta das frequências relativas (em porcentagem) em relação ao total geral das variáveis grau de instrução e região de procedência.

Região de Procedência	Grau de Instrução			
	ensino fundamental	ensino médio	superior	Total
capital	11	14	6	31
interior	8	19	6	33
outra	14	17	6	36
Total	33	50	17	100

Variáveis qualitativas

- Podemos, então, afirmar que **11%** dos funcionários **vêm da capital e possuem o ensino fundamental**.
- Os totais nas margens fornecem as distribuições unidimensionais de cada uma das variáveis.
  - Por exemplo, 31% dos indivíduos vêm da capital, 33% do interior e 36% de outras regiões.

Variáveis qualitativas

- A tabela seguinte apresenta a distribuição das proporções em relação ao total das colunas.
- Os elementos da coluna de *ensino fundamental* foram obtidos por dividir as frequências absolutas (4, 3 e 5) por 12, o total desta coluna.

Tabela 4: Distribuição conjunta das frequências relativas (em porcentagem) em relação aos totais de colunas das variáveis grau de instrução e região de procedência.

Região de Procedência	Grau de Instrução			
	ensino fundamental	ensino médio	superior	Total
capital	33	28	33	94
interior	25	39	33	97
outra	42	33	33	108
Total	100	100	100	300

Variáveis qualitativas

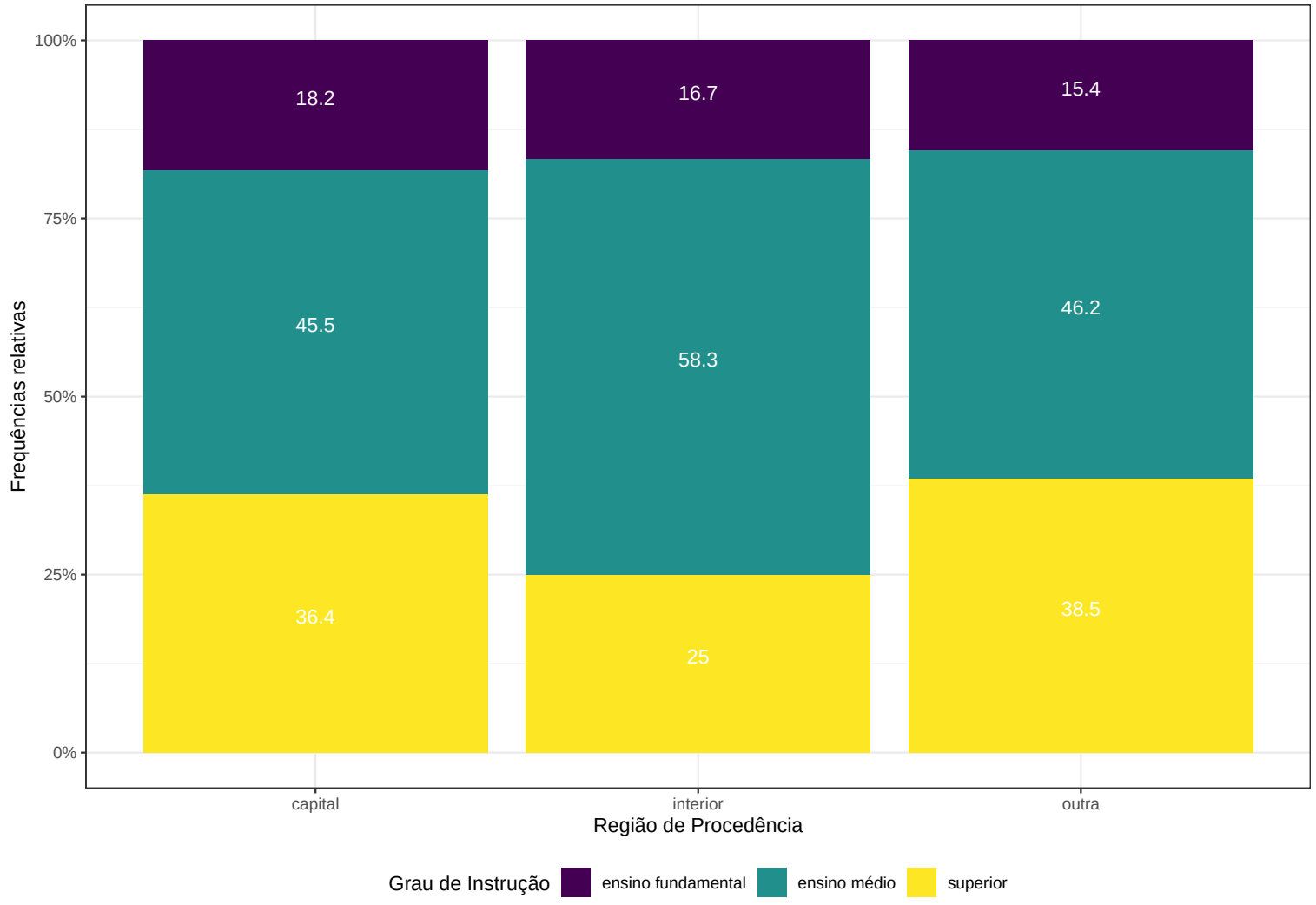
- Podemos dizer que, entre os empregados com instrução até o ensino fundamental, 33% vêm da capital, ao passo que entre os empregados com ensino médio, 28% vêm da capital.
- Esse tipo de tabela serve para comparar a distribuição da procedência dos indivíduos conforme o grau de instrução.
- Se tivéssemos interesse em comparar a distribuição do grau de instrução dos indivíduos conforme a procedência, então calcularíamos as frequências relativas em relação aos totais das linhas[6].

[6] **Sua vez:** construa a distribuição conjunta de frequências relativas (em porcentagem) em relação aos totais de linhas.

Variáveis qualitativas

- Uma forma alternativa de apresentação da distribuição conjunta de duas variáveis qualitativas é feita por meio de gráficos.
- Em geral, utiliza-se o **gráfico de barras**.

Variáveis qualitativas



Variáveis qualitativas

- Uma pergunta frequente de pesquisadores e usuários de Estatística é sobre a associação entre duas variáveis.
- Buscar explicar como se comporta uma variável em função da distribuição de outra tem sido o objetivo de vários estudos que utilizam a Estatística como ferramenta auxiliar.
- Em outras palavras, a distribuição de frequências da variável Y muda de acordo com o nível (categoria) da variável X?
- Em caso afirmativo, diremos que as variáveis são **associadas**, ou dependem uma da outra.
- Caso contrário, diremos que as variáveis são **independentes**, e portanto, não há associação entre as variáveis.

Variáveis qualitativas

Table 2. Sociodemographic and social distancing data of case-patients and controls, Porto Alegre, Brazil, April–June 2020*

Characteristic	Case-patients, n = 271	Controls, n = 1,396
Sex		
M	119 (43.9)	537 (38.5)
F	152 (56.1)	859 (61.5)
Adherence to social distancing		
Very little	32 (11.8)	56 (4.0)
Little	32 (11.8)	81 (5.8)
Moderate—some	43 (15.9)	260 (18.6)
High—a great deal	88 (32.5)	651 (46.6)
Practically isolated from everybody	76 (28.0)	348 (24.9)
Daily routine		
Go out every day, all day, to work or other regular activity	118 (43.5)	251 (18.0)
Go out every day for some activity	12 (4.4)	102 (7.3)
Go out from time to time to shop and stretch my legs	30 (11.1)	192 (13.8)
Go out only for essential things like buying food	74 (27.3)	696 (49.9)
Stay at home all the time	37 (13.7)	155 (11.1)
Social distancing		
Least	44 (16.2)	100 (7.2)
Little	72 (26.6)	216 (15.5)
Much	98 (36.2)	729 (52.2)
Most	57 (21.0)	351 (25.1)
Mask use§		
No	14 (7.1)	5 (1.2)
Sometimes	NA	10 (2.4)
Always	184 (92.9)	405 (96.4)

Fonte: Gonçalves M, dos Reis R, Tólio R, Pellanda L, Schmidt M, Katz N, et al. Social Distancing, Mask Use, and Transmission of Severe Acute Respiratory Syndrome Coronavirus 2, Brazil, April–June 2020. Emerg Infect Dis. 2021;27(8):2135-2143. https://doi.org/10.3201/eid2708.204757

Estudo da UFRGS mostra que uso de máscara reduz em 87% a chance de contrair covid-19

Variáveis qualitativas

- **Exemplo:** em setembro de 2019, o jornalismo esportivo divulgava os resultados das partidas disputas pelo Grêmio.
  - Uma questão era observada pelos especialistas: sem o jogador Maicon, o Grêmio tinha apresentado um desempenho pior.
  - Na tabela a seguir são apresentadas as frequências relativas (em porcentagem) de 52 partidas disputadas até aquele momento.

Tabela 5: Resultados dos jogos do Grêmio em 2019.

	Vitórias	Empates	Derrotas	Total
Sem Maicon	43,48	39,13	17,39	100
Com Maicon	58,62	24,14	17,24	100
Total	51,92	30,77	17,31	100

Variáveis qualitativas

- Note que as frequências são relativas a linha.
- Assim, observamos que sem considerar que Maicon jogou, o jogo resulta em vitória do Grêmio em aproximadamente 52% das vezes.
  - Mas, quando Maicon está em campo esta porcentagem muda para aproximadamente 59%.
  - Quando ele não está em campo, a porcentagem de vitórias cai para 43%.

Variáveis qualitativas

- Assim, poderíamos concluir que a presença de Maicon no jogo foi importante para o desempenho do time do Grêmio no ano de 2019.
- Em outras palavras, as variáveis *resultado do jogo do Grêmio* e *escalação do time do Grêmio* são associadas (ou dependentes), pois a distribuição de frequências da variável *resultado do jogo do Grêmio* muda conforme o nível (categoria) da variável *escalação do time do Grêmio*[7].

[7] Um caminho natural a partir deste resultado seria quantificar o grau de dependência entre estas variáveis. A medida resumo mais utilizada é **coeficiente de contingência**, mais conhecido como o χ^2 de Pearson. Não apresentaremos nestas notas o coeficiente de contingência, mas recomendamos a leitura complementar deste.

Próxima aula

- Distribuições bivariadas: variáveis quantitativas.

Por hoje é só!

Bons estudos!

BUSSAB, W. de O.; MORETTIN, P. A. **Estatística básica**. 9. ed. São Paulo: Saraiva, 2017.