# 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

# Uma variável qualitativa e uma variável quantitativa

# Uma variável qualitativa e uma variável quantitativa

- Como mencionado anteriormente, é comum nessas situações analisar o que acontece com a variável quantitativa dentro de cada categoria da variável qualitativa.
- Essa análise pode ser conduzida por meio de **medidas-resumo**, **histogramas** ou ***boxplots***.
- Consideremos novamente o exemplo dos empregados da seção de orçamentos da Companhia MB, mas agora com o interesse de avaliar a associação entre a variável **salário** e **grau de instrução**.

# Uma variável qualitativa e uma variável quantitativa

| ID | Salario (x Sal Min) | Grau de Instrução |
|---|---|---|
| 1 | 4,00 | ensino fundamental |
| 2 | 4,56 | ensino fundamental |
| 3 | 5,25 | ensino fundamental |
| 5 | 6,26 | ensino fundamental |
| 6 | 6,66 | ensino fundamental |
| 7 | 6,86 | ensino fundamental |
| 8 | 7,39 | ensino fundamental |
| 12 | 8,46 | ensino fundamental |
| 14 | 8,95 | ensino fundamental |
| 18 | 9,80 | ensino fundamental |
| 23 | 12,00 | ensino fundamental |
| 27 | 13,85 | ensino fundamental |
| 4 | 5,73 | ensino médio |
| 9 | 7,59 | ensino médio |
| 10 | 7,44 | ensino médio |
| 11 | 8,12 | ensino médio |
| 13 | 8,74 | ensino médio |
| 15 | 9,13 | ensino médio |

# Uma variável qualitativa e uma variável quantitativa

| ID | Salario (x Sal Min) | Grau de Instrução |
|---|---|---|
| 16 | 9,35 | ensino médio |
| 17 | 9,77 | ensino médio |
| 20 | 10,76 | ensino médio |
| 21 | 11,06 | ensino médio |
| 22 | 11,59 | ensino médio |
| 25 | 13,23 | ensino médio |
| 26 | 13,60 | ensino médio |
| 28 | 14,69 | ensino médio |
| 29 | 14,71 | ensino médio |
| 30 | 15,99 | ensino médio |
| 32 | 16,61 | ensino médio |
| 35 | 19,40 | ensino médio |
| 19 | 10,53 | superior |
| 24 | 12,79 | superior |
| 31 | 16,22 | superior |
| 33 | 17,26 | superior |
| 34 | 18,75 | superior |
| 36 | 23,30 | superior |

# Tabela de medidas resumo

- Vamos começar organizando em uma tabela as principais medidas resumo da variável salário para cada um dos grupos de escolaridade.
  - Ou seja, vamos calcular a **média**, o **desvio padrão** e demais medidas resumo considerando apenas os indivíduos com ensino superior.
  - Logo em seguida, calcularemos as mesmas medidas considerando apenas os indivíduos com ensino médio.
  - E por fim, calcularemos as mesmas medidas considerando apenas os indivíduos com ensino fundamental.

# Tabela de medidas resumo

- Assim, a variável grau de instrução define grupos de indivíduos, e descreveremos a distribuição da variável salário para cada um dos grupos.

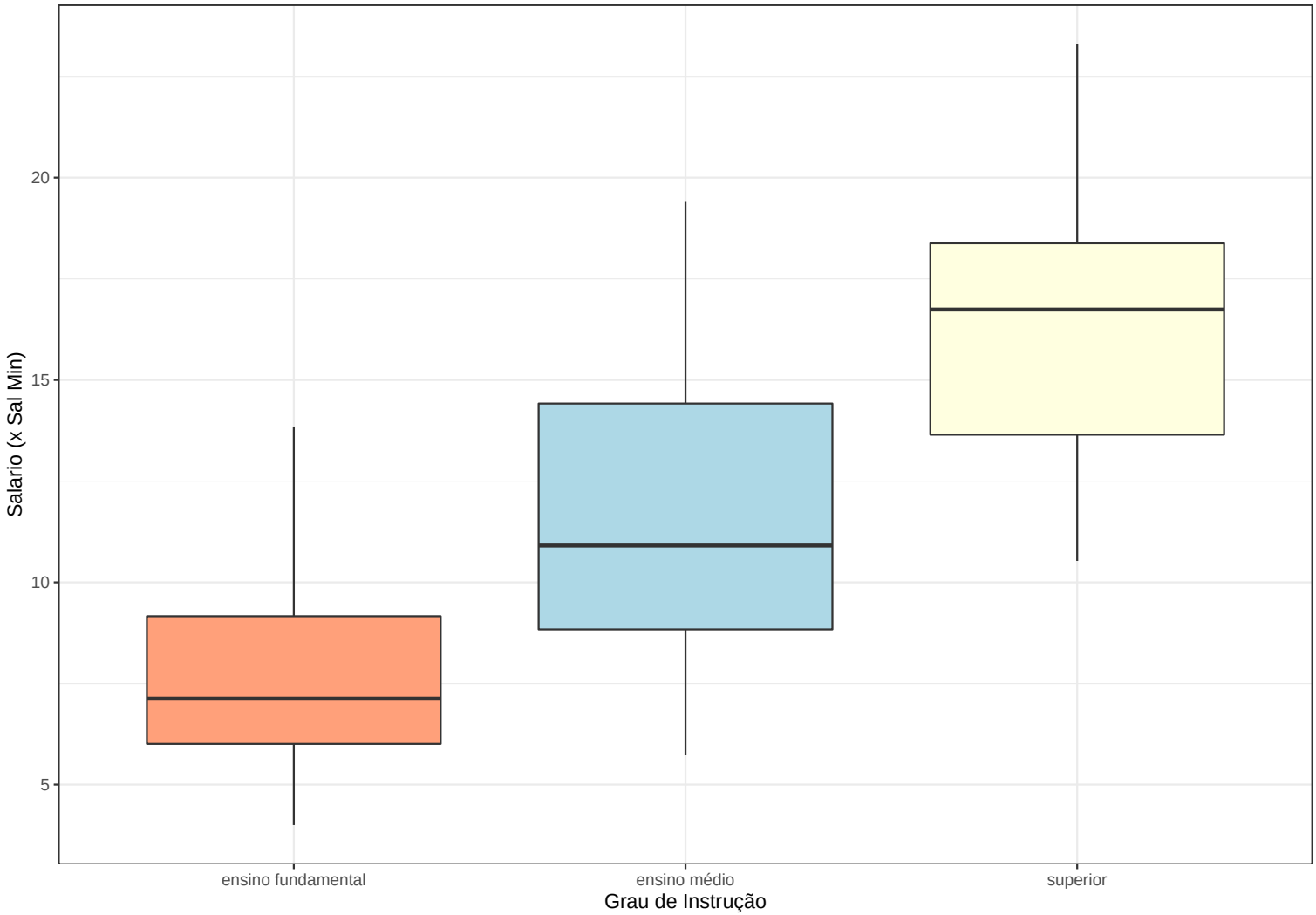
**Tabela 1:** Medidas-resumo para a variável salário, segundo o grau de instrução, na Companhia MB.

| Grau de Instrução | n | Média | D. Padrão | Variância | Min | Q1 | Q2 | Q3 | Max |
|---|---|---|---|---|---|---|---|---|---|
| ensino fundamental | 12 | 7,84 | 2,96 | 8,74 | 4,00 | 6,01 | 7,12 | 9,16 | 13,85 |
| ensino médio | 18 | 11,53 | 3,72 | 13,80 | 5,73 | 8,84 | 10,91 | 14,42 | 19,40 |
| superior | 6 | 16,48 | 4,50 | 20,27 | 10,53 | 13,65 | 16,74 | 18,38 | 23,30 |
| Global | 36 | 11,12 | 4,59 | 21,04 | 4,00 | 7,55 | 10,16 | 14,06 | 23,30 |

# Gráficos da distribuição

- Uma outra forma de apresentarmos a distribuição de salário por grupo de instrução é através do gráfico de *boxplot*.
- Novamente devemos construir cada uma das caixas considerando os dados dos indivíduos de cada um dos grupos[1].

# Gráficos da distribuição




[1] Como se cada um dos grupos formasse um conjunto de dados.

# Associação entre variáveis

- Estes resultados sugerem uma **dependência** dos salários em relação ao grau de instrução:
  - o salário aumenta conforme aumenta o nível de escolaridade do indivíduo.
- O salário médio de um funcionário é $11,12$ (salários mínimos), já para um funcionário com curso superior o salário médio passa a ser $16,48$, enquanto funcionários com o ensino fundamental completo recebem, em média, $7,84$.

# Associação entre variáveis

- Como nos casos anteriores, é conveniente poder contar com uma medida que quantifique o **grau de dependência** entre as variáveis.
- Tal medida pode ser construída a partir das **variâncias de grupo** e a **variância global**.
- Sem usar a informação da variável categorizada[2], a variância calculada para a variável quantitativa[3] para todos os dados mede a dispersão dos dados globalmente.
- **Se a variância dentro de cada categoria for pequena e menor do que a global**, significa que a variável qualitativa **melhora a capacidade de previsão**[4] da quantitativa **e portanto existe uma relação entre as duas variáveis**.

[2] Em nosso exemplo, a variável grau de instrução.

[3] Em nosso exemplo, a variável salário.

[4] Ao saber o grau de instrução conseguimos falar de maneira mais "precisa" sobre o salário dos empregados da Companhia MB?

# Associação entre variáveis

- Observe que, para as variáveis salário (denotaremos por $Y$) e grau de instrução, as variâncias de $Y$ dentro das três categorias são menores do que a global[5].
- Necessita-se, então, de uma medida resumo da variância entre as categorias da variável qualitativa.
- Vamos usar a **média das variâncias** ponderada pelo número de observações em cada categoria, ou seja,

$$\bar{s^2}_Y = \frac{\sum_{j=1}^{k} s_j^2(Y) n_j}{\sum_{j=1}^{k} n_j},$$

em que $k$ é o número de categorias ($k = 3$ em nosso exemplo) e $s_j^2(Y)$ denota a variância de $Y$ dentro da categoria $j$, $j = 1, 2, \ldots, k$.

[5]Lembrando: para o grupo ensino fundamental, $s_Y^2 = 8,74$; para o grupo ensino médio, $s_Y^2 = 13,80$; para o grupo ensino superior, $s_Y^2 = 20,27$; e por fim, para todo o grupo de empregados, $s_Y^2 = 21,04$.

# Associação entre variáveis

- Pode-se mostrar que a média das variâncias é menor ou igual a variância global[6], de modo que podemos definir o **grau de associação** entre as duas variáveis **como o ganho relativo na variância**, obtido pela introdução da variável qualitativa. Explicitamente,

$$R^2 = \frac{s_Y^2 - \bar{s^2}_Y}{s_Y^2} = 1 - \frac{\bar{s^2}_Y}{s_Y^2}.$$

[6] $\bar{s^2}_Y \leq s_Y^2$.

# Associação entre variáveis

- Note que $0 \leq R^2 \leq 1$.
- Quanto maior a capacidade preditiva da variável de grupo, menor será a média das variâncias ($\bar{s^2}_Y$) em relação a variância global, e portanto a razão $\frac{\bar{s^2}_Y}{s^2_Y}$ será próxima de zero, enquanto que $R^2$ tenderá a 1.
- Já se a média das variâncias é próxima da variância global, então a razão $\frac{\bar{s^2}_Y}{s^2_Y}$ será próxima de um, enquanto que $R^2$ tenderá a 0.

Concluímos então, que:

- $R^2$ alto (valores próximos de 1) indicam forte dependência (associação) entre a variável qualitativa e a variável quantitativa.
- $R^2$ baixo (valores próximos de 0) indicam fraca dependência (associação), ou ausência de associação, entre a variável qualitativa e a variável quantitativa.

# Associação entre variáveis

- Retomando o exemplo da Companhia MB, temos que

$$\bar{s^2}_Y = \frac{\sum_{j=1}^{3} s_j^2(Y) n_j}{\sum_{j=1}^{3} n_j} = \frac{12 \times 8,74 + 18 \times 13,8, +6 \times 20,27}{12 + 18 + 6} = 13,19,$$

de modo que

$$R^2 = 1 - \frac{\bar{s^2}_Y}{s_Y^2} = 1 - \frac{13,19}{21,04} = 0,3731.$$

e dizemos que 37,31% da variância global do salário é explicada pela variável grau de instrução[7].

[7] Uma dependência relativamente forte, levando em consideração que dificilmente um $R^2$ próximo de 1 é observado.

# Exercício

# Exercício

A tabela a seguir apresenta os dados referentes a uma amostra de 30 flores.

**Tabela 2:** Tabela de dados brutos.

| comprimento da sépala | largura da sépala | comprimento da pétala | largura da pétala | espécie |
|---|---|---|---|---|
| 5,1 | 3,5 | 1,4 | 0,2 | setosa |
| 4,9 | 3,0 | 1,4 | 0,2 | setosa |
| 4,7 | 3,2 | 1,3 | 0,2 | setosa |
| 4,6 | 3,1 | 1,5 | 0,2 | setosa |
| 5,0 | 3,6 | 1,4 | 0,2 | setosa |
| 5,4 | 3,9 | 1,7 | 0,4 | setosa |
| 4,6 | 3,4 | 1,4 | 0,3 | setosa |
| 5,0 | 3,4 | 1,5 | 0,2 | setosa |
| 4,4 | 2,9 | 1,4 | 0,2 | setosa |
| 4,9 | 3,1 | 1,5 | 0,1 | setosa |
| 7,0 | 3,2 | 4,7 | 1,4 | versicolor |
| 6,4 | 3,2 | 4,5 | 1,5 | versicolor |
| 6,9 | 3,1 | 4,9 | 1,5 | versicolor |
| 5,5 | 2,3 | 4,0 | 1,3 | versicolor |
| 6,5 | 2,8 | 4,6 | 1,5 | versicolor |
| 5,7 | 2,8 | 4,5 | 1,3 | versicolor |
| 6,3 | 3,3 | 4,7 | 1,6 | versicolor |
| 4,9 | 2,4 | 3,3 | 1,0 | versicolor |

# Exercício

| | | | | |
|---|---|---|---|---|
| 6,6 | 2,9 | 4,6 | 1,3 | versicolor |
| 5,2 | 2,7 | 3,9 | 1,4 | versicolor |
| 6,3 | 3,3 | 6,0 | 2,5 | virginica |
| 5,8 | 2,7 | 5,1 | 1,9 | virginica |
| 7,1 | 3,0 | 5,9 | 2,1 | virginica |
| 6,3 | 2,9 | 5,6 | 1,8 | virginica |
| 6,5 | 3,0 | 5,8 | 2,2 | virginica |
| 7,6 | 3,0 | 6,6 | 2,1 | virginica |
| 4,9 | 2,5 | 4,5 | 1,7 | virginica |
| 7,3 | 2,9 | 6,3 | 1,8 | virginica |
| 6,7 | 2,5 | 5,8 | 1,8 | virginica |
| 7,2 | 3,6 | 6,1 | 2,5 | virginica |

# Exercício

- Crie uma nova variável a partir da variável comprimento da pétala, atribuindo valores: **pequeno** se o comprimento da pétala é menor ou igual que 4,5 e **normal** se o comprimento da pétala é superior a 4,5. Chame esta variável de *comprimento da pétala categorizado*.
- Construa a tabela de dupla entrada da distribuição conjunta de frequências das variáveis *espécie* e *comprimento da pétala categorizado*. Você diria que estas variáveis são associadas?
- Construa o gráfico de dispersão das variáveis *comprimento da sépala* e *largura da sépala*. Calcule o coeficiente de regressão. O que você diria sobre a associação entre estas variáveis?
- Construa o gráfico de dispersão das variáveis *comprimento da sépala* e *comprimento da pétala*. Calcule o coeficiente de regressão. O que você diria sobre a associação entre estas variáveis?

# Exercício

A partir dos seguintes exercícios, conclua a respeito da associação entre as variáveis *largura da pétala* e *espécie*.

- Organize uma tabela com o cálculo de diferentes medidas resumo para descrever a distribuição da variável *largura da pétala* entre as diferentes *espécies*.
- Construa um gráfico para apresentar a distribuição a distribuição da variável *largura da pétala* entre as diferentes *espécies*.
- Calcule o $R^2$ para avaliar a associação entre as variáveis *largura da pétala* e *espécie*.

# Próxima aula

- Atividade de Avaliação III.

# Por hoje é só!

**Bons estudos!**