Noções de Estatística

Aula 3

- Introdução à Análise Bidimensional
- Medidas de Associação entre Variáveis
- Regressão Linear

3.1 Análise Bidimensional

- Como avaliar a relação existente entre duas variáveis?

Até agora, nós estudamos alguns métodos para resumir e extrair informações de um estudo através das análises individuais de cada variável presente, mas frequentemente nos deparamos com situações em que gostaríamos de analisar o comportamento conjunto de duas variáveis. Para entender melhor o que isso significa, imagine a seguinte situação:

Você está realizando um estudo sobre obesidade infantil. Para isso, você coleta dados sobre uma série de crianças entrevistadas.

Tabela 3.1 – Pesquisa sobre obesidade infantil

Criança	Idade	Peso	Altura	Sexo
1	6	19	1.12	F
2	10	25	1.50	M
3	5	17	1.10	F
4	4	15	1.05	M
5	7	22	1.12	F

Observe o peso da criança número 2 em comparação com as demais. Você diria que essa criança está obesa?

Quando observamos o peso da criança número 2, vemos que ele é o maior de todos. Essa análise pode nos levar a conclusões precipitadas e consequentemente errôneas. Temos que levar em consideração outras informações. Intuitivamente, sabemos que algumas variáveis podem influenciar diretamente no peso de uma pessoa, por exemplo, ***altura***. Faz sentido pensarmos que quanto mais alta uma pessoa, mais pesada ela será. Outro exemplo: ***Sexo***. Se pensarmos em um homem e uma mulher da mesma altura, é muito provável que o homem terá mais músculos que a mulher, sendo assim mais pesado. Outra variável que não podemos deixar de lado no nosso estudo sobre crianças é a variável ***idade***. Por estarem em fase de crescimento, é claro que crianças mais velhas apresentarão maior altura e, consequentemente, maior peso.

Você concorda que todas essas informações estão intimamente relacionadas? Se você não for cauteloso na sua análise, pode deixar essas informações te confundirem.

É por esta razão que precisamos estudar técnicas adequadas para avaliar a relação existente entre duas (ou mais) variáveis.

Quando consideramos duas variáveis, podemos estar diante de três situações:

- As duas variáveis são qualitativas
- As duas variáveis são quantitativas
- Uma variável é qualitativa e a outra quantitativa

É muito importante saber identificar corretamente o tipo de uma variável, pois, como veremos a seguir, as

técnicas utilizadas são diferentes para cada uma destas situações.

3.2 Tabelas de Distribuição Conjunta

- Associação entre duas Variáveis Qualitativas

Exemplo: Podemos buscar compreender a relação entre o grau de instrução de uma pessoa e a sua respectiva região de procedência. Para isso, uma tabela de distribuição conjunta é uma forma conveniente de observarmos simultaneamente as duas variáveis.

Tabela 3.2.1 Distribuição **conjunta** das frequências das variáveis Grau de instrução e Região de procedência

	Fundamental	Médio	Superior	Total
Capital	4	5	2	11
Interior	3	7	2	12
Outra	5	6	2	13
Total	12	18	6	36

Outra forma ainda melhor de observarmos a relação entre as variáveis é trabalharmos com as suas frequências **relativas**, ou seja, proporções, que podem ser consideradas em relação:

1. Ao total geral
2. Ao total de cada linha
3. Ao total de cada coluna

De acordo com o objetivo do estudo, uma forma de visualização pode ser mais conveniente que a outra.

Veja abaixo duas possibilidades:

Tabela 3.2.2 Distribuição conjunta das proporções das variáveis Grau de instrução e Região de procedência em relação ao total geral das variáveis

	Fundamental	Médio	Superior	Total
Capital	11%	14%	6%	31%
Interior	8%	19%	6%	33%
Outra	14%	17%	6%	36%
Total	33%	50%	17%	100%

Tabela 3.2.3 Distribuição conjunta das proporções das variáveis Grau de instrução e Região de procedência em relação ao total de cada coluna

	Fundamental	Médio	Superior	Total
Capital	33%	28%	33%	31%
Interior	25%	39%	33%	33%
Outra	42%	33%	33%	36%
Total	100%	100%	100%	100%

Com esses exemplos, vimos que a construção das tabelas de Distribuição Conjunta pode ser muito útil para a nossa análise. Com essa ferramenta, conseguimos ter uma maior visibilidade com relação ao ***grau de dependência*** entre as nossas variáveis, de modo que podemos prever melhor o resultado de uma delas quando conhecemos a realização da outra.

3.2.1 Medidas de associação entre variáveis

Outra forma muito importante de avaliarmos o grau de dependência entre variáveis é através das Medidas de Associação. De modo geral, a quantificação do grau de associação entre duas variáveis é feita pelos chamados coeficientes de associação ou correlação. Essas são medidas que descrevem, por meio de um único número, a associação entre duas variáveis.

- Associação entre duas Variáveis Qualitativas

A seguir, veremos algumas medidas de associação possíveis para o caso de uma análise bidimensional entre duas variáveis **qualitativas**.

Exemplo: Vamos verificar se o cultivo de um determinado tipo de fruta está associado com algum fator regional.

Como fazemos isso?

Imagine que soubéssemos como se distribuem os tipos de cultivos de frutas em todo o Brasil. E que soubéssemos que o cultivo das frutas se distribui da seguinte maneira:

24% são do tipo Laranja
42% são do tipo Maçã
22% são do tipo Pêra
12% são outros tipos de frutas

Porém, imagine que eu observo a seguinte tabela:

Tabela 3.2.4 Tipos de fazendas de cultivo de frutas por Estado

Estado	Tipo de Fazenda				
	1.Laranja	2.Maçã	3.Pêra	4.Outras	Total
São Paulo	214 33%	237 37%	78 12%	119 18%	648 100%
Paraná	51 17%	102 34%	126 42%	22 7%	301 100%
Rio G do Sul	111 18%	304 50%	139 23%	48 8%	602 100%
Total	376 (24%)	643 (42%)	343 (22%)	189 (12%)	1551 (100%)

Observe atentamente a tabela.

Uma análise cautelosa nos indica uma certa dependência entre as variáveis. Caso não houvesse associação, esperaríamos que em cada estado observássemos 24% de fazendas produtoras de laranja, 42% de fazendas produtoras de maçã, 22% de fazendas produtoras de Pêra e 12% de fazendas produtoras de outros tipos de frutas. Ou seja, esperaríamos uma tabela assim:

Tabela 3.2.4 Tipos de Fazenda de cultivo de frutas por Estado assumindo **independência** entre as duas variáveis.

Estado	Tipo de Fazenda				
	1.Laranja	2.Maçã	3.Pêra	4.Outras	Total
São Paulo	157 24%	269 42%	143 22%	79 12%	648 100%
Paraná	73 24%	124 42%	67 22%	37 12%	301 100%
Rio G S	146 24%	250 42%	133 22%	73 12%	602 100%
Total	376 (24%)	643 (42%)	343 (22%)	189 (12%)	1551 (100%)

Com isso concluímos que há uma certa diferença entre a nossa expectativa e o que foi de fato observado. Essa diferença é chamada de **resíduo**.

Resíduo = (Valor observado – Valor esperado)

Uma das formas de avaliarmos a associação entre as duas variáveis é através do cálculo da medida de associação de χ^2 (qui-quadrado de Pearson). Um valor grande de χ^2 em geral indica associação entre as variáveis.

$$\chi^2 = \sum_{i=1}^{n} \frac{(valor\ observado_i - valor\ esperado_i)^2}{valor\ esperado_i}$$

Vamos compreender melhor essa fórmula.

Primeiramente, calculamos a diferença entre o que observamos e a nossa expectativa em cada casela da tabela. Em seguida, calculamos essa diferença ao quadrado, assim nos livramos de valores negativos (e também da possibilidade de, futuramente, obter valores próximos de zero ao somarmos todas as diferenças). Depois, dividimos esse resultado pelo valor esperado, para levar em conta contagens maiores e menores (não queremos uma fórmula que nos dê um qui quadrado grande apenas porque estamos trabalhando com um banco de dados grande). Por fim, somamos esses resultados para todas as caselas da tabela.

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - e_i)^2}{e_i}$$

O coeficiente de Pearson é bastante utilizado. Porém vemos que ele varia de 0 a infinito, pois trata-se de uma soma de valores sempre positivos. Por esse motivo, apenas o fato da nossa amostra ser grande já pode influenciar no nosso coeficiente. Pode ser mais interessante utilizar uma fórmula mais restrita e, consequentemente, mais interpretável.

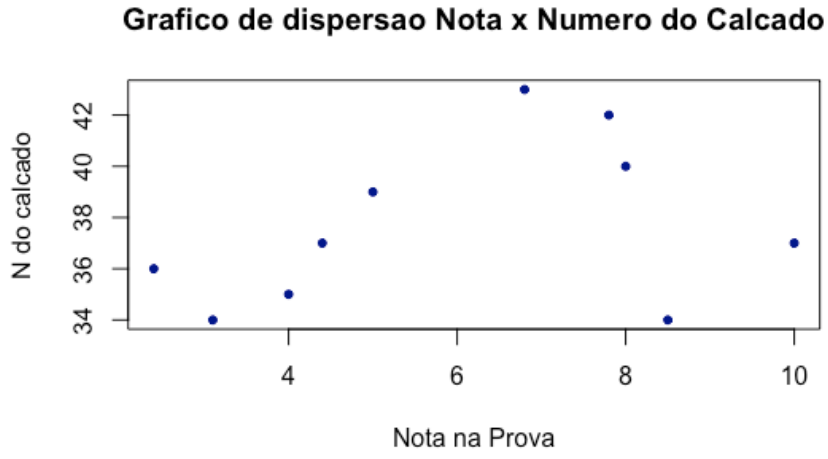
$$T = \sqrt{\frac{\frac{\chi^2}{n}}{(r-1)(s-1)}}$$

Sendo r o número de linhas e s o número de colunas da tabela de Distribuição conjunta. (Ou o número de categorias de cada uma das variáveis). Tal coeficiente

varia entre [0;1]. Valores próximos a 0 indicam falta de associação e próximos a 1 indicam alta associação.

- Associação entre duas Variáveis Quantitativas

Uma ferramenta muito útil para se verificar a associação entre variáveis quantitativas é o gráfico de dispersão. Observe os exemplos abaixo:



Qual a diferença entre os gráficos?

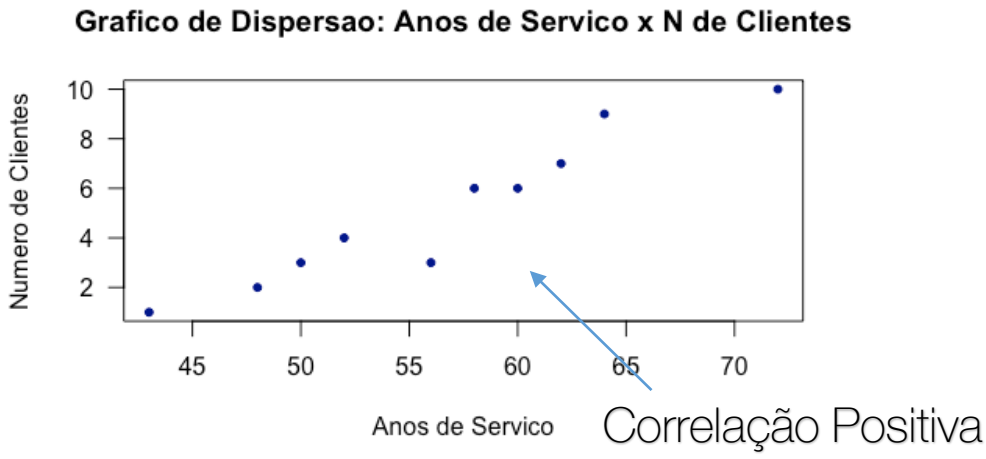
No primeiro gráfico observamos um comportamento sistemático das observações. Vemos que há uma tendência sendo obedecida. No segundo gráfico, vemos que não há um comportamento específico sendo obedecido. É assim que podemos notar graficamente a influência (ou a ausência de influência) de uma variável sobre a outra.

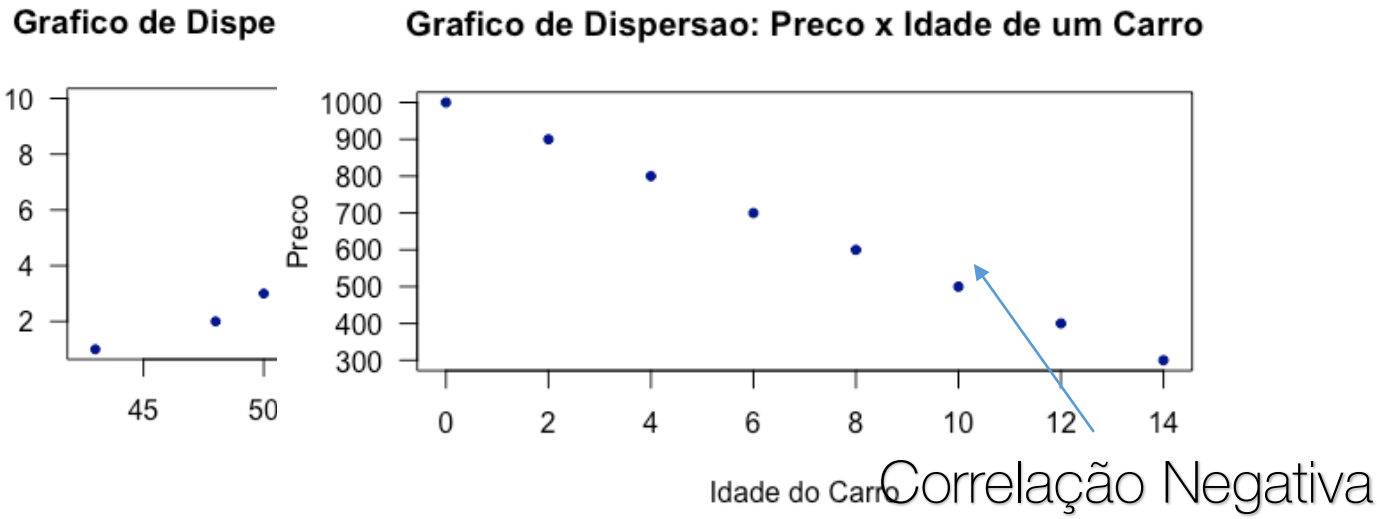
Para calcularmos uma medida de associação para o caso de duas variáveis quantitativas, utilizamos o cálculo da correlação, definido por:

$$corr(X,Y) = \frac{1}{n}\sum_{i=1}^{n}\frac{(x_i - \bar{x})}{dp(x)}.\frac{(y_i - \bar{y})}{dp(y)}$$

- O coeficiente de correlação é um valor entre [-1;1].
- Quanto mais perto das pontas, mais correlacionados estão os dados.
- Quanto mais próximo de 0, menos correlacionados estão os dados.

A correlação pode ser negativa ou positiva, dependendo da relação existente entre as variáveis.





- Associação entre uma Variável Quantitativa e uma variável qualitativa

Nessas situações, é comum analisarmos o que acontece com a variável quantitativa em cada categoria da variável qualitativa.

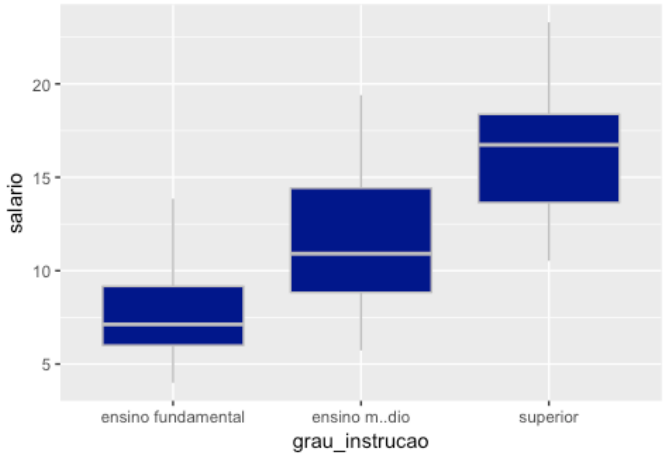
Inicialmente, podemos fazer uma análise exploratória das medidas resumo da variável quantitativa dentro de cada categoria da variável qualitativa. Vejamos um exemplo de tabela de medidas-resumo:

Medidas-resumo para a variável Salário, segundo o grau de instrução em uma determinada companhia.

	n	Min.	Q1	Mediana	Média	Q3	Max.	dp(S)	Var(S)
Fundamental	12	4	6.01	7.12	7.84	9.16	13.85	2.83	8.01
Medio	18	5.73	8.84	10.91	11.53	14.42	19.4	3.61	13.04
Superior	6	10.53	13.65	16.74	16.48	18.38	23.3	4.11	16.89
Todos	36	4	7.55	10.16	11.12	14.06	23.3	4.52	20.46

Também é interessante construir gráficos para a variável quantitativa para cada categoria da variável qualitativa. Exemplo:

Boxplots para a variável Salário para cada região de procedência.



Como vimos nos últimos casos, também é interessante poder contar com uma medida que quantifique o grau de dependência entre as variáveis.

Para isso, a medida utilizada será o R^2, ou coeficiente de determinação.

Como funciona o R^2?

Primeiramente, precisamos entender a variável Salário. No nosso exemplo podemos notar que a variável salário possui uma certa variabilidade, certo? Sim, porque cada pessoa possui um salário diferente. Essa variabilidade pode existir por diversas razões. As pessoas possuem trabalhos diferentes, idades diferentes, posições diferentes, trabalham em setores diferentes, em empresas diferentes e assim por diante. Podemos imaginar uma infinidade de razões para os salários das pessoas serem diferentes. Observe novamente o nosso boxplot Salário x Grau de Instrução.

Você concorda que o grau de instrução de uma pessoa parece influenciar o seu respectivo salário? Em outras palavras, saber o grau de instrução de uma pessoa pode ser uma informação útil para tentarmos prever o salário da mesma.

O R^2 é um valor entre 0 e 1 e ele responde à seguinte pergunta:

Qual porcentagem da variabilidade da variável Salário pode ser explicada pela variabilidade em grau de instrução?

Ou seja, alterando (variando) os valores da variável Grau de instrução, eu provoco uma variabilidade na variável Salário. Porém sabemos que existem várias razões para a variável Salário variar além de grau de instrução.

Por isso, existe um percentual que é motivado pela variável Grau de instrução e um restante que é explicado por outras variáveis.

Portanto, o R^2 representa uma porcentagem.

A fórmula para o cálculo do R^2 é:

$$1 - \frac{SS_{res}}{SS_{Tot}}$$

Em que:

$$SS_{res} = \sum_i (y_i - f_i)^2$$

e:

$$SS_{Tot} = \sum_{i} (y_i - \bar{y})^2$$

Conclusão:

Para cada combinação de tipo de variável há uma medida de associação correspondente e também métodos de visualização adequados.
