LADE

LABORATÓRIO DE ANÁLISES DE DADOS EDUCACIONAIS E ESTATÍSTICA APLICADA

IFCE - CAMPUS FORTALEZA

Estatística Descritiva com Python

Lembrete

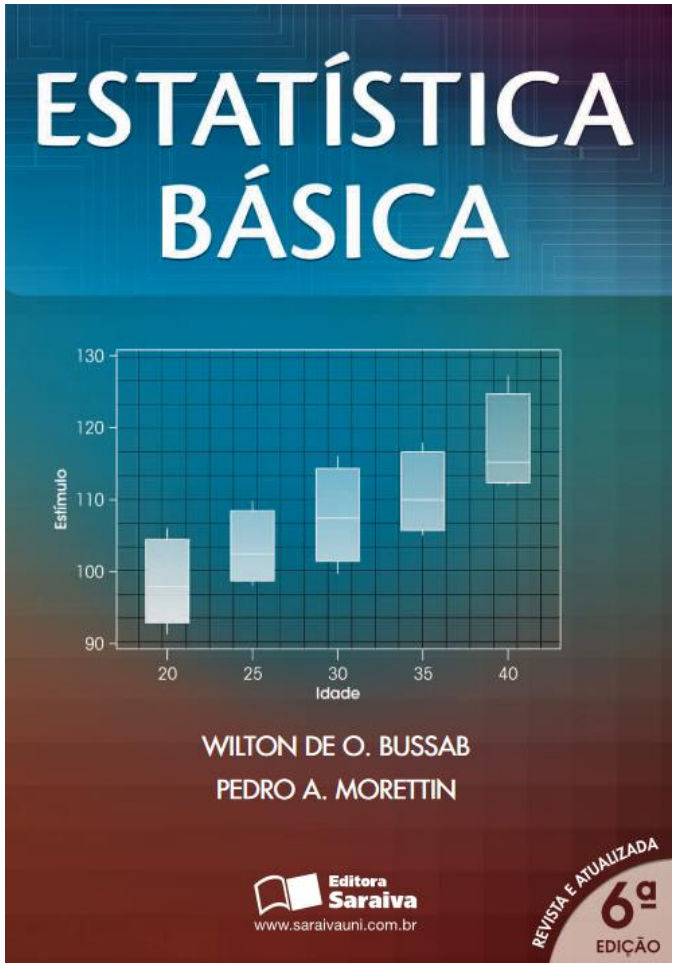
Tipos de Variáveis

Qualitativas

- Nominal
  - Profissão
  - Sexo
  - Religião
- Ordinal
  - Escolaridade
  - Estágio da doença
  - Classe social

Quantitativas

- Discreta
  - Nº de filhos
  - Nº de acessos à plataforma
- Contínua
  - Altura
  - Peso
  - Salário



LADE

Até agora vimos como organizar e resumir informações pertinentes a uma única variável (ou a um conjunto de dados), mas frequentemente estamos interessados em analisar o comportamento conjunto de duas ou mais variáveis aleatórias. Os dados aparecem na forma de uma matriz, usualmente com as colunas indicando as variáveis e as linhas os indivíduos (ou elementos). A Tabela abaixo mostra a notação de uma matriz com **p variáveis $X_1, X_2, \ldots, X_p$ e n indivíduos, totalizando np dados.** (BUSSAB, W. O.; MORETTIN, P. A. – Estatística Básica. Editora Saraiva, São Paulo, 2010.)

Tabela de dados.

Indivíduo	Variável					
	X_1	X_2	...	X_j	...	X_p
1	x_{11}	x_{12}	...	x_{1j}	...	x_{1p}
2	x_{21}	x_{22}	...	x_{2j}	...	x_{2p}
⋮	⋮	⋮		⋮		⋮
i	x_{i1}	x_{i2}	...	x_{ij}	...	x_{ip}
⋮	⋮	⋮		⋮		⋮
n	x_{n1}	x_{n2}	...	x_{nj}	...	x_{np}

LADE

Quando consideramos duas variáveis (ou dois conjuntos de dados), podemos ter três situações:

(a) as duas variáveis são qualitativas;
(b) as duas variáveis são quantitativas; e
(c) uma variável é qualitativa e outra é quantitativa.

As técnicas de análise de dados nas três situações são diferentes.

Quando as variáveis são qualitativas, os dados são resumidos em tabelas de dupla entrada (ou de contingência), onde aparecerão as freqüências absolutas ou contagens de indivíduos que pertencem simultaneamente a categorias de uma e outra variável.

Quando as duas variáveis são quantitativas, as observações são provenientes de mensurações, e técnicas como gráficos de dispersão ou de quantis são apropriadas.

Quando temos uma variável qualitativa e outra quantitativa, em geral analisamos o que acontece com a variável quantitativa quando os dados são categorizados de acordo com os diversos atributos da variável qualitativa.

Mas podemos ter também o caso de duas variáveis quantitativas agrupadas em classes. Por exemplo, podemos querer analisar a associação entre renda e consumo de certo número de famílias e, para isso, agrupamos as famílias em classes de rendas e classes de consumo. Desse modo, recaímos novamente numa tabela de dupla entrada.

Contudo, em todas as situações, o objetivo é encontrar as possíveis relações ou associações entre as duas variáveis. Essas relações podem ser detectadas por meio de métodos gráficos e medidas numéricas

LADE

Variáveis Qualitativas

Suponha que queiramos analisar o comportamento conjunto das variáveis **Y: grau de instrução e V: região de procedência**, cujas observações estão contidas na Tabela Abaixo. A distribuição de frequências é representada por uma tabela de dupla entrada e está na Tabela Abaixo.

Cada elemento do corpo da tabela dá a **frequência observada das realizações simultâneas de Y e V.** Assim, observamos quatro indivíduos da capital com ensino fundamental, sete do interior com ensino médio etc.

A linha dos totais fornece a distribuição da variável Y, ao passo que a **coluna** dos totais fornece a **distribuição da variável V.** As distribuições assim obtidas são chamadas tecnicamente de **distribuições marginais**, enquanto a Tabela Abaixo constitui a **distribuição conjunta de Y e V.**

Distribuição conjunta das freqüências das variáveis grau de instrução (Y) e região de procedência (V).

V \ Y	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

Em vez de trabalharmos com as frequências absolutas, podemos construir tabelas com as frequências relativas (proporções), como foi feito no caso unidimensional. Mas aqui existem três possibilidades de expressarmos a proporção de cada casela:

(a) em relação ao total geral;
(b) em relação ao total de cada linha;
(c) ou em relação ao total de cada coluna

De acordo com o objetivo do problema em estudo, uma delas será a mais conveniente.

Distribuição conjunta das proporções (em porcentagem) em relação ao total geral das variáveis Y e V definidas no texto.

V \ Y	Fundamental	Médio	Superior	Total
Capital	11%	14%	6%	31%
Interior	8%	19%	6%	33%
Outra	14%	17%	5%	36%
Total	33%	50%	17%	100%

A Tabela Abaixo apresenta a distribuição conjunta das frequências relativas, **expressas como proporções do total geral.** Podemos, então, afirmar que **11% dos empregados vêm da capital e têm 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. Observe que, devido ao problema de aproximação das divisões, a distribuição das proporções introduz algumas diferenças não existentes.

Distribuição conjunta das freqüências das variáveis grau de instrução (Y) e região de procedência (V).

V \ Y	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

Distribuição conjunta das proporções (em porcentagem) em relação ao total geral das variáveis Y e V definidas no texto.

V \ Y	Fundamental	Médio	Superior	Total
Capital	11%	14%	6%	31%
Interior	8%	19%	6%	33%
Outra	14%	17%	5%	36%
Total	33%	50%	17%	100%

$$\frac{4}{36} = 0{,}1111\ ou\ 11{,}11\%$$

$$\frac{11}{36} = 0{,}305555\ ou\ 31\%$$

LADE

A Tabela Abaixo apresenta a **distribuição das proporções em relação ao total das colunas.** 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.

Distribuição conjunta das proporções (em porcentagem) em relação aos totais de cada coluna das variáveis Y e V definidas no texto.

V \ Y	Fundamental	Médio	Superior	Total
Capital	33%	28%	33%	31%
Interior	25%	39%	33%	33%
Outra	42%	33%	34%	36%
Total	100%	100%	100%	100%

Distribuição conjunta das proporções (em porcentagem) em relação ao total geral das variáveis Y e V definidas no texto.

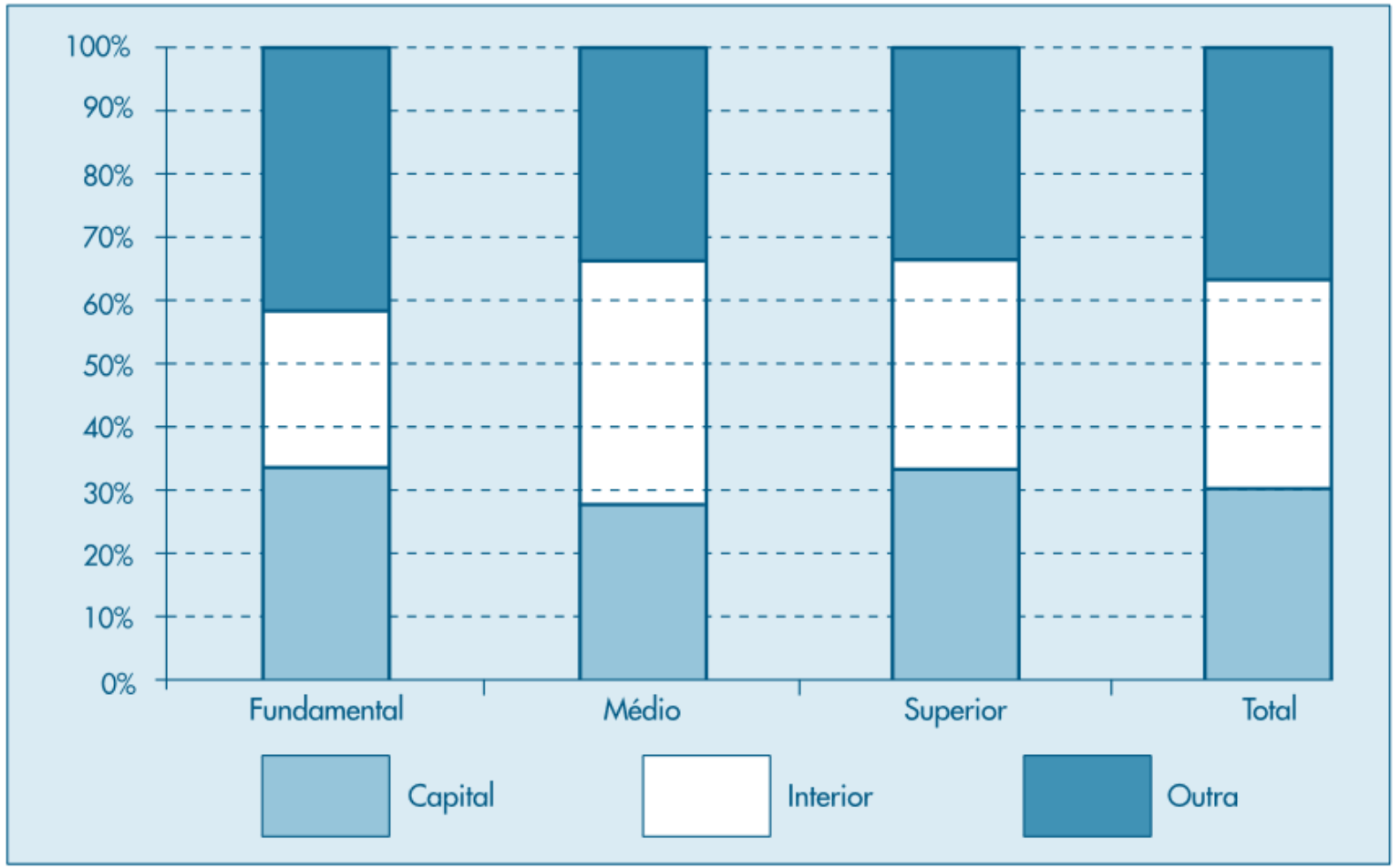
V \ Y	Fundamental	Médio	Superior	Total
Capital	11%	14%	6%	31%
Interior	8%	19%	6%	33%
Outra	14%	17%	5%	36%
Total	33%	50%	17%	100%

Exercício : Construir a distribuição conjunta das proporções em relação as linhas.

LADE

A comparação entre as duas variáveis também pode ser feita utilizando-se representações gráficas.

Distribuição da região de procedência por grau de instrução.



Associação entre Variáveis Qualitativas

Um dos principais objetivos de se construir uma distribuição conjunta de duas variáveis qualitativas é descrever a associação entre elas, isto é, queremos conhecer **o grau de dependência entre elas**, de modo que possamos prever melhor o resultado de uma delas quando conhecermos a realização da outra.

Exemplo: Queremos verificar se existe ou não associação entre o sexo e a carreira escolhida por 200 alunos de Economia e Administração. Esses dados estão na Tabela 1.

Tabela 1

Distribuição conjunta de alunos segundo o sexo (X) e o curso escolhido (Y).

Y \ X	Masculino	Feminino	Total
Economia	85	35	120
Administração	55	25	80
Total	140	60	200

Tabela 2

Distribuição conjunta das proporções (em porcentagem) de alunos segundo o sexo (X) e o curso escolhido (Y).

Y \ X	Masculino	Feminino	Total
Economia	61%	58%	60%
Administração	39%	42%	40%
Total	100%	100%	100%

Exemplo: Queremos verificar se existe ou não associação entre o sexo e a carreira escolhida por 200 alunos de Economia e Administração. Esses dados estão na Tabela 1.

Inicialmente, verificamos que fica muito difícil tirar alguma conclusão, devido à diferença entre os totais marginais. Devemos, pois, construir as proporções segundo as linhas ou as colunas para podermos fazer comparações. **Fixemos os totais das colunas**; a distribuição está na Tabela 2.

Tabela 1

Distribuição conjunta de alunos segundo o sexo (X) e o curso escolhido (Y).

Y \ X	Masculino	Feminino	Total
Economia	85	35	120
Administração	55	25	80
Total	140	60	200

Tabela 2

Distribuição conjunta das proporções (em porcentagem) de alunos segundo o sexo (X) e o curso escolhido (Y).

Y \ X	Masculino	Feminino	Total
Economia	61%	58%	60%
Administração	39%	42%	40%
Total	100%	100%	100%

A partir dessa tabela podemos observar que, independentemente do sexo, **60% das pessoas preferem Economia e 40% preferem Administração (observe na coluna de total).**

Não havendo dependência entre as variáveis, esperaríamos essas mesmas proporções para cada sexo.

Observando a tabela, vemos que as proporções do **sexo masculino (61% e 39%) e do sexo feminino (58% e 42%) são próximas das marginais (60% e 40%).**

Esses resultados parecem indicar não haver dependência entre as duas variáveis, para o conjunto de alunos considerado. Concluímos então que, neste caso, **as variáveis sexo e escolha do curso parecem ser não associadas.**

Distribuição conjunta das proporções (em porcentagem) de alunos segundo o sexo (X) e o curso escolhido (Y).

Y \ X	Masculino	Feminino	Total
Economia	61%	58%	60%
Administração	39%	42%	40%
Total	100%	100%	100%

Vamos considerar, agora, um problema semelhante, mas envolvendo alunos de Física e Ciências Sociais, cuja distribuição conjunta está na Tabela Abaixo

Distribuição conjunta das freqüências e proporções (em porcentagem), segundo o sexo (X) e o curso escolhido (Y).

Y \ X	Masculino	Feminino	Total
Física	100 (71%)	20 (33%)	120 (60%)
Ciências Sociais	40 (29%)	40 (67%)	80 (40%)
Total	140 (100%)	60 (100%)	200 (100%)

Inicialmente, convém observar que, para economizar espaço, **resumimos duas tabelas numa única**, indicando as proporções em relação **aos totais das colunas entre parênteses**. Comparando agora a distribuição das proporções pelos cursos, **independentemente do sexo (coluna de totais),** com as distribuições diferenciadas por sexo (colunas de masculino e feminino), **observamos uma disparidade bem acentuada nas proporções**. **Parece, pois, haver maior concentração de homens no curso de Física e de mulheres no de Ciências Sociais.** Portanto, nesse caso, as variáveis sexo e curso escolhido parecem ser associadas. **Quando existe associação entre variáveis, sempre é interessante quantificar essa associação, e isso será objeto da próxima seção.** Antes de passarmos a discutir esse aspecto, convém observar que teríamos obtido as mesmas conclusões do Exemplo anterior se tivéssemos calculado as proporções, mantendo constantes os totais das linhas.

LADE

Medidas de Associação entre Variáveis Qualitativas

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 (ou dependência) entre duas variáveis.** Para maior facilidade de compreensão, esses coeficientes usualmente variam entre **0 e 1, ou entre –1 e +1, e a proximidade de zero indica falta de associação.** Existem muitas medidas que quantificam a associação entre variáveis qualitativas, apresentaremos apenas duas delas: **o chamado coeficiente de contingência, devido a K. Pearson e uma modificação desse.** **(BUSSAB, W. O.; MORETTIN, P. A. – Estatística Básica. Editora Saraiva, São Paulo, 2010.)**

LADE

Exemplo: Queremos verificar se a criação de determinado tipo de cooperativa está associada com algum fator regional. Coletados os dados relevantes, obtemos a Tabela Abaixo.

Cooperativas autorizadas a funcionar por tipo e estado, junho de 1974.

Estado	Tipo de Cooperativa				Total
	Consumidor	Produtor	Escola	Outras	
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 (51%)	139 (23%)	48 (8%)	602 (100%)
Total	376 (24%)	643 (42%)	343 (22%)	189 (12%)	1.551 (100%)

Fonte: Sinopse Estatística do Brasil — IBGE, 1977.

Tabela 1

A análise da tabela mostra a existência de certa dependência entre as variáveis. Caso não houvesse associação, esperaríamos que em cada **estado tivéssemos 24% de cooperativas de consumidores, 42% de cooperativas de produtores, 22% de escolas** e **12% de outros tipos.** Então, por exemplo, o número esperado de cooperativas de consumidores no Estado de São Paulo seria 648 × 0,24 = 157 e no Paraná seria 301 × 0,24 = 73 (ver Tabela a seguir).

A análise da tabela mostra a existência de certa dependência entre as variáveis. Caso não houvesse associação, esperaríamos que em cada **estado tivéssemos 24% de cooperativas de consumidores, 42% de cooperativas de produtores, 22% de escolas** e **12% de outros tipos.** Então, por exemplo, o número esperado de cooperativas de consumidores no Estado de São Paulo seria 648 × 0,24 = 157 e no Paraná seria 301 × 0,24 = 73 (ver Tabela a seguir).

Estado	Tipo de Cooperativa				Total
	Consumidor	Produtor	Escola	Outras	
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. do Sul	146 (24%)	250 (42%)	133 (22%)	73 (12%)	602 (100%)
Total	376 (24%)	643 (42%)	343 (22%)	189 (12%)	1.551 (100%)

Tabela 2

Desvios entre observados e esperados.

Estado	Tipo de Cooperativa			
	Consumidor	Produtor	Escola	Outras
São Paulo	57 (20,69)	–32 (3,81)	–65 (29,55)	40 (20,25)
Paraná	–22 (6,63)	–22 (3,90)	59 (51,96)	–15 (6,08)
Rio G. do Sul	–35 (8,39)	54 (11,66)	6 (0,27)	–25 (8,56)

Tabela 3

Comparando as duas tabelas, podemos verificar as discrepâncias existentes entre os valores observados (Tabela 1) e os valores esperados (Tabela 2), caso as variáveis não fossem associadas. Na Tabela 3 resumimos os desvios: valores observados menos valores esperados. Observando essa tabela podemos tirar algumas conclusões:

Valores observados (Tabela 1)

Cooperativas autorizadas a funcionar por tipo e estado, junho de 1974.

Estado	Tipo de Cooperativa				Total
	Consumidor	Produtor	Escola	Outras	
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 (51%)	139 (23%)	48 (8%)	602 (100%)
Total	376 (24%)	643 (42%)	343 (22%)	189 (12%)	1.551 (100%)

Fonte: Sinopse Estatística do Brasil — IBGE, 1977.

Valores esperados (Tabela 2)

Estado	Tipo de Cooperativa				Total
	Consumidor	Produtor	Escola	Outras	
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. do Sul	146 (24%)	250 (42%)	133 (22%)	73 (12%)	602 (100%)
Total	376 (24%)	643 (42%)	343 (22%)	189 (12%)	1.551 (100%)

Valores observados menos valores esperados (Tabela 3)

Desvios entre observados e esperados.

Estado	Tipo de Cooperativa			
	Consumidor	Produtor	Escola	Outras
São Paulo	57 (20,69)	–32 (3,81)	–65 (29,55)	40 (20,25)
Paraná	–22 (6,63)	–22 (3,90)	59 (51,96)	–15 (6,08)
Rio G. do Sul	–35 (8,39)	54 (11,66)	6 (0,27)	–25 (8,56)

LADE

Valores observados (Tabela 1)

Cooperativas autorizadas a funcionar por tipo e estado, junho de 1974.

Estado	Tipo de Cooperativa				Total
	Consumidor	Produtor	Escola	Outras	
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 (51%)	139 (23%)	48 (8%)	602 (100%)
Total	376 (24%)	643 (42%)	343 (22%)	189 (12%)	1.551 (100%)

Fonte: Sinopse Estatística do Brasil — IBGE, 1977.

Valores esperados (Tabela 2)

Estado	Tipo de Cooperativa				Total
	Consumidor	Produtor	Escola	Outras	
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. do Sul	146 (24%)	250 (42%)	133 (22%)	73 (12%)	602 (100%)
Total	376 (24%)	643 (42%)	343 (22%)	189 (12%)	1.551 (100%)

Construção dos Valores esperados (Tabela 2)

Estado	Tipo de Cooperativa				Total
	Consumidor	Produtor	Escola	Outras	
São Paulo					648 (100%)
Paraná					301 (100%)
Rio G. do Sul					602 (100%)
Total	376 (24%)	643 (42%)	343 (22%)	189 (12%)	1.551 (100%)

Construção dos Valores esperados (Tabela 2)

Estado	Tipo de Cooperativa				Total
	Consumidor	Produtor	Escola	Outras	
São Paulo	a11	a12	a13	a14	648 (100%)
Paraná	a21	a22	a23	a24	301 (100%)
Rio G. do Sul	a31	a32	a33	a34	602 (100%)
Total	376 (24%)	643 (42%)	343 (22%)	189 (12%)	1.551 (100%)
Total	376	643	343	189	1.551 (100%)
Total	(376/1551)*100 = 24,24%	(643/1551)*100 = 41,46%	(343/1551)*100 = 22,11%	(189/1551)*100 = 12,19%	1.551 (100%)

24,2 + 41,5 + 22,1 + 12,2 = 100 %

a11 =648*24,24% = 157
a21 = 301*24,24% = 73
a31= 602*24,24% = 146
Total = 376

a13 = 648*22,11% = 143
a23 = 301*22,11% = 67
a33 = 602*22,11% = 133
Total= 343

a11+a12+a13+a14= 157+269+143+79=648
a21+a22+a23+a24= 73+125 + 67+37=302
a31+a32+a33+a34= 146+250+133+73 =602

a21+a22+a23+a24= 73 + 124 + 67+37 = 301

a12 = 648*41,46%= 269
a22 = 301*41,46%= **125 --- 124**
a32 = 602*41,46%= 250
Total = 644
643

a14 = 648*12,19%=79
a24 = 301*12,19%=37
a34 = 602*12,19%=73
Total = 189

Construção dos Valores esperados (Tabela 2)

Estado	Tipo de Cooperativa				Total
	Consumidor	Produtor	Escola	Outras	
São Paulo	157	269	143	79	648 (100%)
Paraná	73	124	67	37	301 (100%)
Rio G. do Sul	146	250	133	73	602 (100%)
Total	376	643	343	189	1.551 (100%)

Total	(376/1551)*100 = 24,24%	(643/1551)*100 = 41,46%	(343/1551)*100 = 22,11%	(189/1551)*100 = 12,19%	1.551 (100%)

24,24+41,46+22,11+12,19=100

a11 =648*24,24% = 157
a21 = 301*24,24% = 73
a31= 602*24,24% = 146
Total = 376

a13 = 648*22,11% = 143
a23 = 301*22,11% = 67
a33 = 602*22,11% = 133
Total= 343

a11+a12+a13+a14= 157+269+143+79=648
a21+a22+a23+a24= 73+125 + 67+37=302
a31+a32+a33+a34= 146+250+133+73 =602

a21+a22+a23+a24= 73+124 + 67+37=301

a12 = 648*41,46%= 269
a22 = 301*41,46%= **125 --- 124**
a32 = 602*41,46%= 250
Total = 644
643

a14 = 648*12,19%=79
a24 = 301*12,19%=37
a34 = 602*12,19%=73
Total = 189

Valores observados (Tabela 1)

Cooperativas autorizadas a funcionar por tipo e estado, junho de 1974.

Estado	Tipo de Cooperativa			
	Consumidor	Produtor	Escola	Outras
São Paulo	214	237	78	119
Paraná	51	102	126	22
Rio G. do Sul	111	304	139	48

Valores esperados (Tabela 2)

Estado	Tipo de Cooperativa				Total
	Consumidor	Produtor	Escola	Outras	
São Paulo	157	269	143	79	648 (100%)
Paraná	73	124	67	37	301 (100%)
Rio G. do Sul	146	250	133	73	602 (100%)
Total	376	643	343	189	1.551 (100%)

Valores observados menos valores esperados (Tabela 3)

Desvios entre observados e esperados.

Estado	Tipo de Cooperativa			
	Consumidor	Produtor	Escola	Outras
São Paulo	57 (20,69)	–32 (3,81)	–65 (29,55)	40 (20,25)
Paraná	–22 (6,63)	–22 (3,90)	59 (51,96)	–15 (6,08)
Rio G. do Sul	–35 (8,39)	54 (11,66)	6 (0,27)	–25 (8,56)

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

$$214 - 157 = 57\,, \quad \frac{(214-157)^2}{157} = 20{,}69$$

(i) A soma total dos resíduos é nula. Isso pode ser verificado facilmente somando-se cada linha.

(ii) A casela Escola-São Paulo é aquela que apresenta o maior desvio da suposição de não-associação (–65). Nessa casela esperávamos 143 casos. A casela Escola Paraná também tem um desvio alto (59), mas o valor esperado é bem menor (67).

Portanto, se fôssemos considerar os desvios relativos, aquele correspondente ao segundo caso seria bem maior. Uma maneira de observar esse fato é construir, para cada casela, a medida

$$\frac{(o_i - e_i)^2}{e_i} \quad *$$

no qual o_i é o valor observado e e_i é o valor esperado.

Usando (*) para a casela Escola-São Paulo obtemos $(-65)^2/143 = 29{,}55$ e para a casela Escola-Paraná obtemos $(59)^2/67 = 51{,}96$, o que é uma indicação de que o desvio devido a essa última casela é "maior" do que aquele da primeira. Na **Tabela 3** indicamos entre parênteses esses valores para todas as caselas.
Uma medida do afastamento global pode ser dada pela soma de todas as medidas (*). Essa medida é denominada $\chi 2\ (qui - quadrado)$ de Pearson, e no nosso exemplo teríamos

$$\chi 2 = 20{,}69 + 6{,}63 + \ldots + 8{,}56 = 171{,}76.$$

Simplificando

A fórmula para calcular a frequência esperada (E) em uma tabela de contingência. (Não se preocupe com os arredondamentos, coloque com 2 casa decimais)

$$E_{ij} = \frac{n_{i.} \times n_{.j}}{n}$$

Onde:

	Coluna 1	Coluna 2	Total
Linha 1	$(n_{1.} \times n_{.1})/n$	$(n_{1.} \times n_{.2})/n$	$n_{1.}$
Linha 2	$(n_{2.} \times n_{.1})/n$	$(n_{2.} \times n_{.2})/n$	$n_{2.}$
Total	$n_{.1}$	$n_{.2}$	n

- E_{ij} é a frequência esperada na célula ij da tabela de contingência,
- $n_{i.}$ é o total da linha i,
- $n_{.j}$ é o total da coluna j,
- n é o total geral de observações na tabela.

Um valor grande de $\chi 2$ indica associação entre as variáveis, o que parece ser o caso.

Antes de dar uma fórmula geral para essa medida de associação, vamos introduzir, na Tabela Abaixo, uma notação geral para tabelas de dupla entrada

Notação para tabelas de contingência.

X \ Y	B_1	B_2	...	B_j	...	B_s	Total
A_1	n_{11}	n_{12}	...	n_{1j}	...	n_{1s}	$n_{1.}$
A_2	n_{21}	n_{22}	...	n_{2j}	...	n_{2s}	$n_{2.}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
A_i	n_{i1}	n_{i2}	...	n_{ij}	...	n_{is}	$n_{i.}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
A_r	n_{r1}	n_{r2}	...	n_{rj}	...	n_{rs}	$n_{r.}$
Total	$n_{.1}$	$n_{.2}$	...	$n_{.j}$	...	$n_{.s}$	$n_{..}$

Suponha que temos duas variáveis qualitativas X e Y, classificadas em **r categorias** $A_1, A_2, \ldots, A_r$ para X e ***s* categorias** $B_1, B_2, \ldots, B_s$, para Y.

Notação para tabelas de contingência.

X \ Y	B_1	B_2	...	B_j	...	B_s	Total
A_1	n_{11}	n_{12}	...	n_{1j}	...	n_{1s}	$n_{1.}$
A_2	n_{21}	n_{22}	...	n_{2j}	...	n_{2s}	$n_{2.}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
A_i	n_{i1}	n_{i2}	...	n_{ij}	...	n_{is}	$n_{i.}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
A_r	n_{r1}	n_{r2}	...	n_{rj}	...	n_{rs}	$n_{r.}$
Total	$n_{.1}$	$n_{.2}$	...	$n_{.j}$	...	$n_{.s}$	$n_{..}$

Na tabela, temos:

Notação para tabelas de contingência.

X \ Y	B_1	B_2	...	B_j	...	B_s	Total
A_1	n_{11}	n_{12}	...	n_{1j}	...	n_{1s}	$n_{1.}$
A_2	n_{21}	n_{22}	...	n_{2j}	...	n_{2s}	$n_{2.}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
A_i	n_{i1}	n_{i2}	...	n_{ij}	...	n_{is}	$n_{i.}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
A_r	n_{r1}	n_{r2}	...	n_{rj}	...	n_{rs}	$n_{r.}$
Total	$n_{.1}$	$n_{.2}$	...	$n_{.j}$	...	$n_{.s}$	$n..$

n_{ij} = número de elementos pertencentes à i-ésima categoria de X e j-ésima categoria de Y;

$n_{i.} = \sum_{j=1}^{s} n_{ij}$ = número de elementos da i-ésima categoria de X;

$n_{.j} = \sum_{i=1}^{r} n_{ij}$ = número de elementos da j-ésima categoria de Y;

$n.. = n = \sum_{i=1}^{r} \sum_{j=1}^{s} n_{ij}$ = número total de elementos.

Sob a hipótese de que as variáveis X *e* Y não sejam associadas (comumente dizemos independentes), temos que

$$\frac{n_{i1}}{n_{.1}} = \frac{n_{i2}}{n_{.2}} = \ldots = \frac{n_{is}}{n_{.s}}, \; i = 1, 2, \ldots, r$$

ou ainda

$$\frac{n_{ij}}{n_{.j}} = \frac{n_{i.}}{n}, \quad i = 1, \ldots, r, \; j = 1, \ldots, s$$

de onde se deduz, finalmente, que

$$n_{ij} = \frac{n_{i.}\, n_{.j}}{n}, \qquad i = 1, \ldots, r, \; j = 1, \ldots, s. \qquad **$$

Portanto, sob a hipótese de independência, de (**) segue que, em termos de frequências relativas, podemos escrever

$$f_{ij} = f_{i.}\, f_{.j}\,.$$

Chamando de frequências esperadas os valores dados pelos segundos membros de (**), e denotando-as por n^*_{ij}, temos que o qui-quadrado de Pearson pode ser escrito

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(n_{ij} - n^*_{ij})^2}{n^*_{ij}}, \qquad ***$$

onde n_{ij} são os valores efetivamente observados. Se a hipótese de não-associação for verdadeira, o valor calculado de (***) deve estar próximo de zero. Se as variáveis forem associadas, o valor de χ2 deve ser grande.

Pearson definiu uma medida de associação, baseada em (***), chamada coeficiente de contingência, dado por

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}}.$$

Contudo, o coeficiente acima não varia entre **0 e 1.** O valor máximo de **C** depende de **r e s.** Para evitar esse inconveniente, costuma-se definir um outro coeficiente, dado por

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

que atinge o máximo igual a 1 se **r = s.**

Lembrando: Suponha que temos duas variáveis qualitativas X e Y, classificadas em **r categorias** $A_1, A_2, \ldots, A_r$ para X e ***s* categorias** $B_1, B_2, \ldots, B_s$ para Y

Colocando parâmetros para afirmação abaixo

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(n_{ij} - n^*_{ij})^2}{n^*_{ij}}, \qquad ***$$

Se a hipótese de não-associação for verdadeira, o valor calculado de (***) deve estar próximo de zero. Se as variáveis forem associadas, o valor de χ2 deve ser grande.

Casos Particulares

Tabela 2x2

Para determinar o valor crítico do qui-quadrado para uma tabela 2x2 com um nível de significância de **0,05 (ou 5%),** você pode consultar uma tabela de valores críticos do qui-quadrado. No entanto, vou calcular isso para você.
Para uma tabela de contingência 2 x 2, o grau de liberdade é

grau de liberdade = (número de linhas−1) × (número de colunas−1) = (2−1) × (2−1) = 1

Com um nível de significância de 0,05 e **1 grau de liberdade**, você pode usar um software estatístico ou calculadora online para encontrar o valor crítico do qui-quadrado. Este valor é aproximadamente 3,8415.
Portanto, para um nível de significância de 0,05, o valor crítico do qui-quadrado para uma tabela 2x2 é aproximadamente **3,8415.**

Tabela 3x2

Para uma tabela de contingência 3 x 2, o número de graus de liberdade é dado por

grau de liberdade = (número de linhas−1) × (número de colunas−1) = (3−1) × (2−1) = 2

Com um nível de significância **0,05 (ou 5%)** e **2 graus de liberdade**, você pode consultar uma tabela de valores críticos do qui-quadrado ou usar um software estatístico para encontrar o valor crítico do qui-quadrado.
O valor crítico do qui-quadrado para um nível de significância de 0,05 e 2 graus de liberdade é aproximadamente 5,991.
Portanto, para uma tabela 3x2 e um nível de significância de 0,05, o valor crítico do qui-quadrado é aproximadamente **5,991**.

Conclusão:

Quando o valor calculado do qui-quadrado é maior que o valor crítico correspondente para um determinado nível de significância, você pode rejeitar a hipótese nula de independência entre as variáveis na tabela de contingência. Em outras palavras, você pode afirmar que há uma associação significativa entre as variáveis.

Portanto, se o valor calculado do qui-quadrado for maior que os valores críticos que mencionamos anteriormente **(3,8415 para uma tabela 2x2 e 5,991 para uma tabela 3x2, ambos para um nível de significância de 0,05),** então você teria evidências estatisticamente significativas para afirmar que há uma associação entre as variáveis na tabela de contingência.

Associação entre Variáveis Quantitativas

Quando as variáveis envolvidas são ambas do tipo quantitativo, pode-se usar o mesmo tipo de análise apresentado nas seções anteriores e exemplificado com variáveis qualitativas. De modo análogo, a distribuição conjunta pode ser resumida em tabelas de dupla entrada e, por meio das distribuições marginais, é possível estudar a associação das variáveis. Algumas vezes, para evitar um grande número de entradas, agrupamos os dados marginais em intervalos de classes, de modo semelhante ao resumo feito no caso unidimensional.

Mas, além desse tipo de análise, as variáveis quantitativas são passíveis de procedimentos analíticos e gráficos mais refinados. Um dispositivo bastante útil para se verificar a associação entre duas variáveis quantitativas, ou entre dois conjuntos de dados, **é o gráfico de dispersão**, que vamos introduzir por meio de exemplos.

Número de anos de serviço (X) por número de clientes (Y) de agentes de uma companhia de seguros.

Agente	Anos de serviço (X)	Número de clientes (Y)
A	2	48
B	3	50
C	4	56
D	5	52
E	4	43
F	6	60
G	7	62
H	8	58
I	8	64
J	10	72

Nesse tipo de gráfico temos os possíveis pares de valores (x, y), na ordem que aparecem. Para o exemplo, vemos que parece haver uma associação entre as variáveis, porque no conjunto, à medida que aumenta o tempo de serviço, aumenta o número de clientes.

Gráfico de dispersão para as variáveis X: anos de serviço e Y: número de clientes.

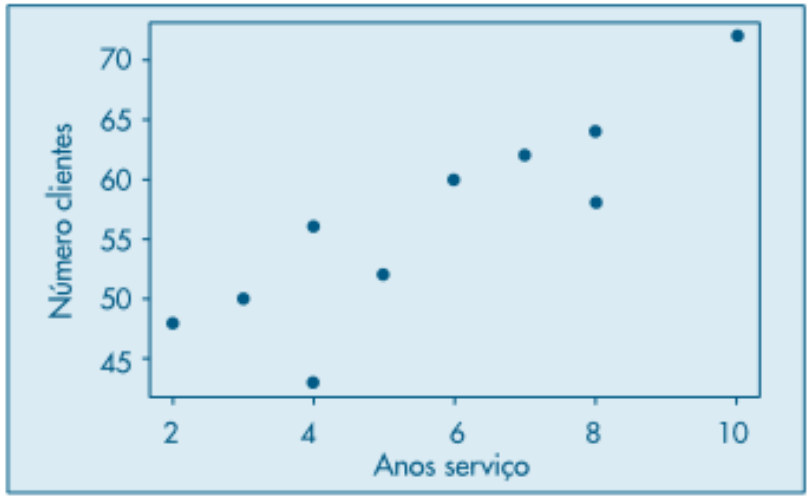
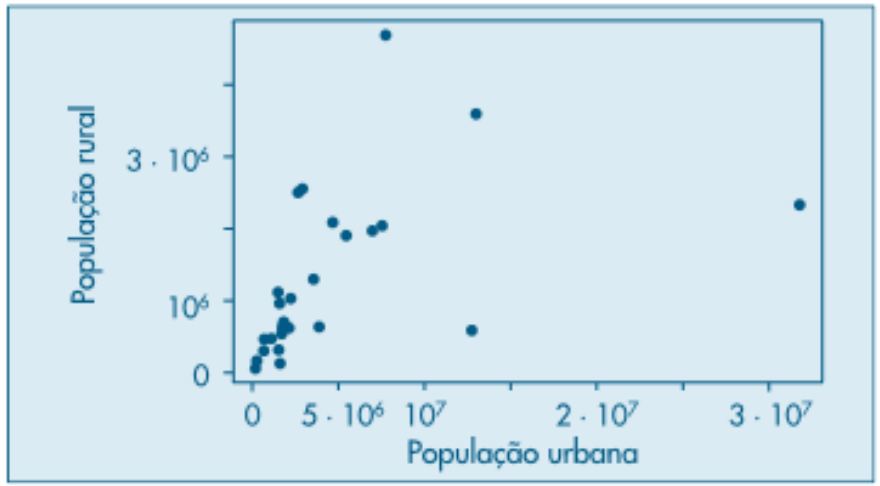


Gráfico de dispersão para as variáveis X: população urbana e Y: população rural.

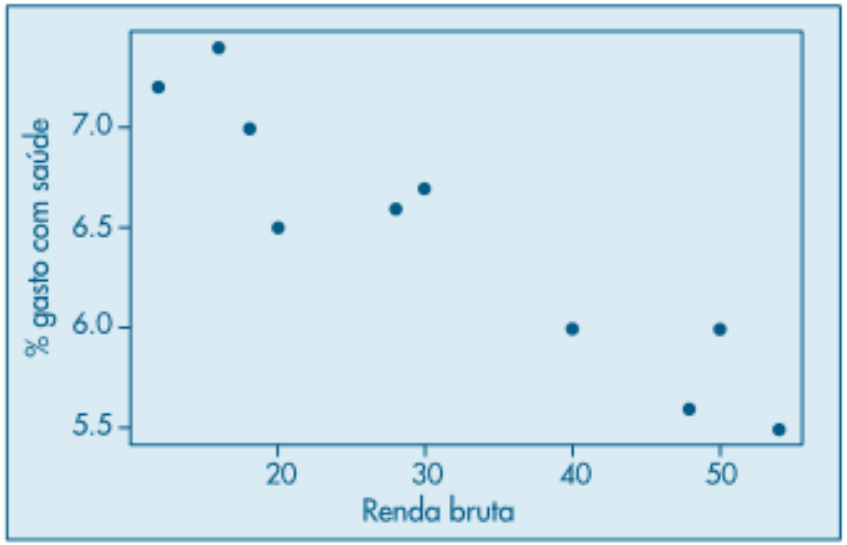


Vemos que parece não haver associação entre as variáveis, pois os pontos não apresentam nenhuma tendência particular.

Renda bruta mensal (X) e porcentagem da renda gasta em saúde (Y) para um conjunto de famílias.

Família	X	Y
A	12	7,2
B	16	7,4
C	18	7,0
D	20	6,5
E	28	6,6
F	30	6,7
G	40	6,0
H	48	5,6
I	50	6,0
J	54	5,5

Gráfico de dispersão para as variáveis X: renda bruta e Y: % renda gasta com saúde.



Observando o gráfico de dispersão , vemos que existe uma associação **“inversa”,** isto é, aumentando a renda bruta, diminui a porcentagem sobre ela gasta em assistência médica.

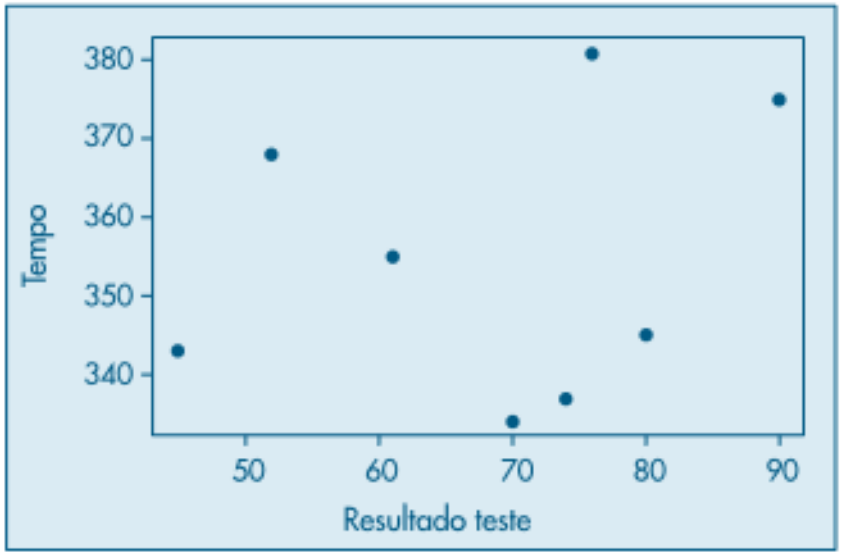
Oito indivíduos foram submetidos a um teste sobre conhecimento de língua estrangeira e, em seguida, mediu-se o tempo gasto para cada um aprender a operar uma determinada máquina. As variáveis medidas foram:

X: resultado obtido no teste (máximo = 100 pontos);
Y: tempo, em minutos, necessário para operar a máquina satisfatoriamente.

Resultado de um teste (X) e tempo de operação de máquina (Y) para oito indivíduos.

Indivíduo	X	Y
A	45	343
B	52	368
C	61	355
D	70	334
E	74	337
F	76	381
G	80	345
H	90	375

Gráfico de dispersão para as variáveis X: resultado no teste e Y: tempo de operação.

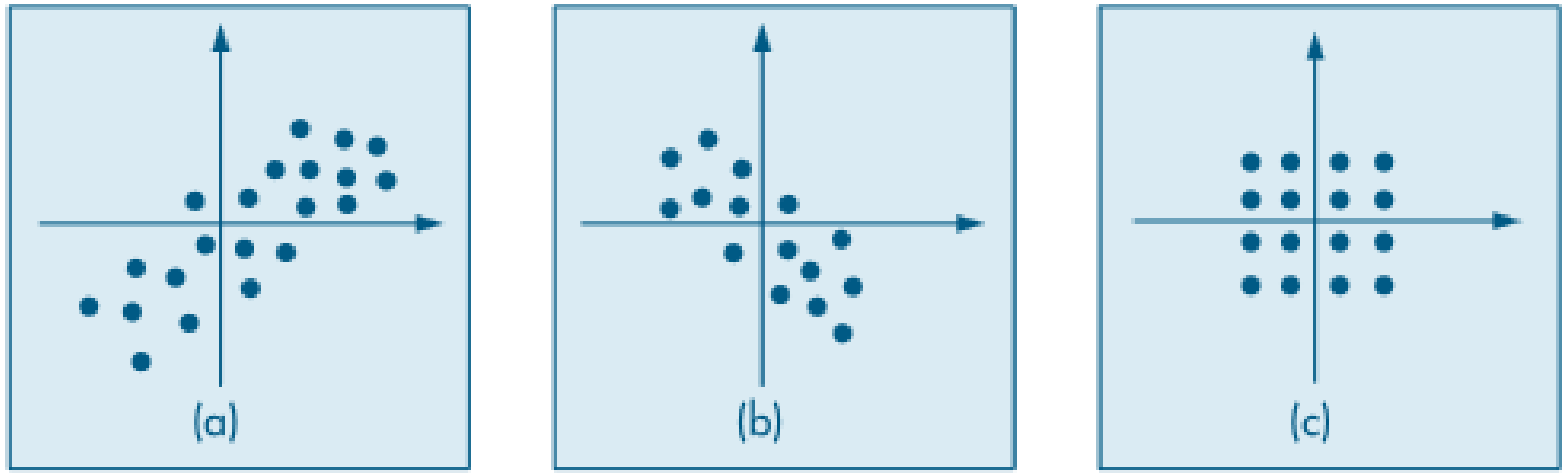


Os dados estão na Tabela acima. Do gráfico de dispersão concluímos que parece não haver associação entre as duas variáveis, **pois conhecer o resultado do teste não ajuda a prever o tempo gasto para aprender a operar a máquina.**

LADE

A partir dos gráficos apresentados, verificamos que a representação gráfica das variáveis quantitativas ajuda muito a compreender o comportamento conjunto das duas variáveis quanto à existência ou não de associação entre elas.
Contudo, é muito útil quantificar esta associação. Existem muitos tipos de associações possíveis, e **aqui iremos apresentar o tipo de relação mais simples, que é a linear.** Isto é, iremos definir uma medida que avalia o quanto a nuvem de pontos no gráfico de dispersão aproxima-se de uma reta. Esta medida será definida de modo a variar num intervalo finito, especificamente, **de –1 a +1.**

Tipos de associações entre duas variáveis.



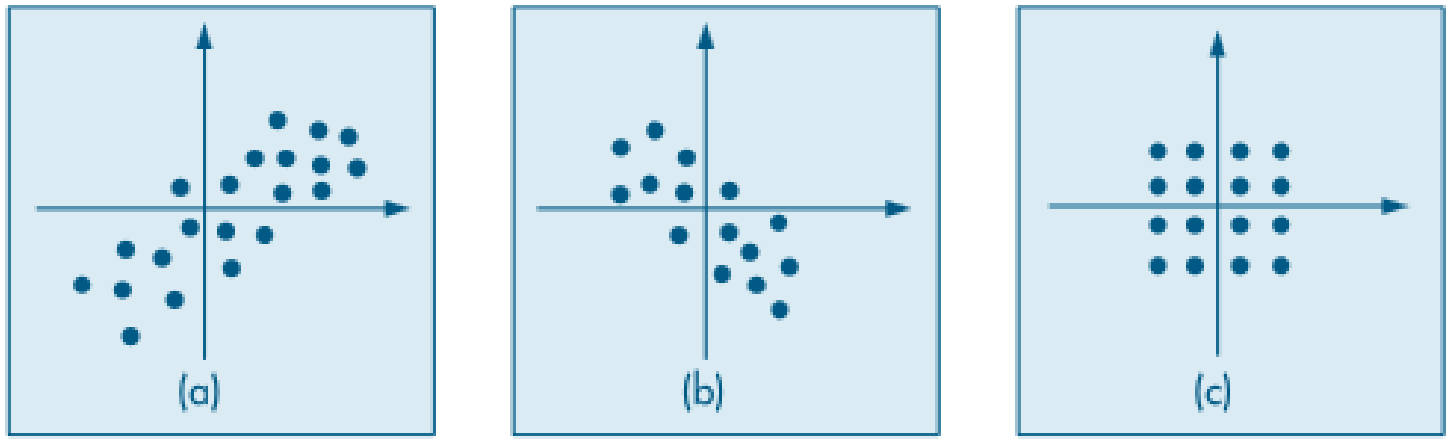
Consideremos um gráfico de dispersão como o da **Figura (a)** no qual, por meio de uma transformação conveniente, a origem foi colocada no centro da nuvem de dispersão.

Aqueles dados possuem uma associação linear direta (ou positiva) e notamos que a grande maioria dos pontos está situada no primeiro e terceiro quadrantes. Nesses quadrantes as coordenadas dos pontos têm o mesmo sinal, e, portanto, o produto delas será sempre positivo. Somando-se o produto das coordenadas dos pontos, o resultado será um número positivo, pois existem mais produtos positivos do que negativos.

Para a dispersão da **Figura (b)**, observamos uma dependência linear inversa (ou negativa) e, procedendo-se como anteriormente, a soma dos produtos das coordenadas será negativa.

Finalmente, para a **Figura (c),** a soma dos produtos das coordenadas será zero, pois cada resultado positivo tem um resultado negativo simétrico, anulando-se na soma. Nesse caso não há associação linear entre as duas variáveis. Em casos semelhantes, quando a distribuição dos pontos for mais ou menos circular, a soma dos produtos será aproximadamente zero.

Tipos de associações entre duas variáveis.



Coeficiente de correlação (linear)

Número de anos de serviço (X) por número de clientes (Y) de agentes de uma companhia de seguros.

Agente	Anos de serviço (X)	Número de clientes (Y)
A	2	48
B	3	50
C	4	56
D	5	52
E	4	43
F	6	60
G	7	62
H	8	58
I	8	64
J	10	72

Cálculo do coeficiente de correlação.

Agente	Anos x	Clientes y	$x-\bar{x}$	$y-\bar{y}$	$\frac{x-\bar{x}}{dp(x)}=z_x$	$\frac{y-\bar{y}}{dp(y)}=z_y$	$z_x \cdot z_y$
A	2	48	–3,7	–8,5	–1,54	–1,05	1,617
B	3	50	–2,7	–6,5	–1,12	–0,80	0,846
C	4	56	–1,7	–0,5	–0,71	–0,06	0,043
D	5	52	–0,7	–4,5	–0,29	–0,55	0,160
E	4	43	–1,7	–13,5	–0,71	–1,66	1,179
F	6	60	0,3	3,5	0,12	0,43	0,052
G	7	62	1,3	5,5	0,54	0,68	0,367
H	8	58	2,3	1,5	0,95	0,19	0,181
I	8	64	2,3	7,5	0,95	0,92	0,874
J	10	72	4,3	15,5	1,78	1,91	3,400
Total	57	565	0	0			8,769

$\bar{x} = 5{,}7$, $dp(X) = 2{,}41$, $\bar{y} = 56{,}5$, $dp(Y) = 8{,}11$

Grau de associação linear = 8,769/10 = 0,8769

Portanto, para esse exemplo, o grau de associação linear está quantificado **por 87,7%**

Definição: Dados n pares de valores $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, chamaremos de coeficiente de correlação entre as duas variáveis X e Y a

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

ou seja, a média dos produtos dos valores padronizados das variáveis.

Associação entre Variáveis Qualitativas e Quantitativas

Como mencionado na introdução deste capítulo, 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, box plots ou ramo-e-folhas.** Vamos ilustrar com um exemplo.

Retomemos os dados da Tabela Abaixo, para os quais desejamos analisar agora o comportamento dos salários dentro de cada categoria de grau de instrução, ou seja, investigar o comportamento conjunto das variáveis S e Y.

Informações sobre estado civil, grau de instrução, número de filhos, salário (expresso como fração do salário mínimo), idade (medida em anos e meses) e procedência de 36 empregados da seção de orçamentos da Companhia MB.

Nº	Estado civil	Grau de instrução	Nº de filhos	Salário (× sal. mín.)	Idade		Região de procedência
					anos	meses	
1	solteiro	ensino fundamental	—	4,00	26	03	interior
2	casado	ensino fundamental	1	4,56	32	10	capital
3	casado	ensino fundamental	2	5,25	36	05	capital
4	solteiro	ensino médio	—	5,73	20	10	outra
5	solteiro	ensino fundamental	—	6,26	40	07	outra
6	casado	ensino fundamental	0	6,66	28	00	interior
7	solteiro	ensino fundamental	—	6,86	41	00	interior
8	solteiro	ensino fundamental	—	7,39	43	04	capital
9	casado	ensino médio	1	7,59	34	10	capital
10	solteiro	ensino médio	—	7,44	23	06	outra
11	casado	ensino médio	2	8,12	33	06	interior
12	solteiro	ensino fundamental	—	8,46	27	11	capital
13	solteiro	ensino médio	—	8,74	37	05	outra
14	casado	ensino fundamental	3	8,95	44	02	outra
15	casado	ensino médio	0	9,13	30	05	interior
16	solteiro	ensino médio	—	9,35	38	08	outra
17	casado	ensino médio	1	9,77	31	07	capital
18	casado	ensino fundamental	2	9,80	39	07	outra
19	solteiro	superior	—	10,53	25	08	interior
20	solteiro	ensino médio	—	10,76	37	04	interior
21	casado	ensino médio	1	11,06	30	09	outra
22	solteiro	ensino médio	—	11,59	34	02	capital
23	solteiro	ensino fundamental	—	12,00	41	00	outra
24	casado	superior	0	12,79	26	01	outra
25	casado	ensino médio	2	13,23	32	05	interior
26	casado	ensino médio	2	13,60	35	00	outra
27	solteiro	ensino fundamental	—	13,85	46	07	outra
28	casado	ensino médio	0	14,69	29	08	interior
29	casado	ensino médio	5	14,71	40	06	interior
30	casado	ensino médio	2	15,99	35	10	capital
31	solteiro	superior	—	16,22	31	05	outra
32	casado	ensino médio	1	16,61	36	04	interior
33	casado	superior	3	17,26	43	07	capital
34	solteiro	superior	—	18,75	33	07	capital
35	casado	ensino médio	2	19,40	48	11	capital
36	casado	superior	3	23,30	42	02	interior

Retomemos os dados da Tabela Abaixo, para os quais desejamos analisar agora o comportamento dos salários dentro de cada categoria de grau de instrução, ou seja, investigar o comportamento conjunto das variáveis S e Y.

Informações sobre estado civil, grau de instrução, número de filhos, salário (expresso como fração do salário mínimo), idade (medida em anos e meses) e procedência de 36 empregados da seção de orçamentos da Companhia MB.

Nº	Estado civil	Grau de instrução	Nº de filhos	Salário (× sal. mín.)	Idade		Região de procedência
					anos	meses	
1	solteiro	ensino fundamental	—	4,00	26	03	interior
2	casado	ensino fundamental	1	4,56	32	10	capital
3	casado	ensino fundamental	2	5,25	36	05	capital
4	solteiro	ensino médio	—	5,73	20	10	outra
5	solteiro	ensino fundamental	—	6,26	40	07	outra
6	casado	ensino fundamental	0	6,66	28	00	interior
7	solteiro	ensino fundamental	—	6,86	41	00	interior
8	solteiro	ensino fundamental	—	7,39	43	04	capital
9	casado	ensino médio	1	7,59	34	10	capital
10	solteiro	ensino médio	—	7,44	23	06	outra
11	casado	ensino médio	2	8,12	33	06	interior
12	solteiro	ensino fundamental	—	8,46	27	11	capital
13	solteiro	ensino médio	—	8,74	37	05	outra
14	casado	ensino fundamental	3	8,95	44	02	outra
15	casado	ensino médio	0	9,13	30	05	interior
16	solteiro	ensino médio	—	9,35	38	08	outra
17	casado	ensino médio	1	9,77	31	07	capital
18	casado	ensino fundamental	2	9,80	39	07	outra
19	solteiro	superior	—	10,53	25	08	interior
20	solteiro	ensino médio	—	10,76	37	04	interior
21	casado	ensino médio	1	11,06	30	09	outra
22	solteiro	ensino médio	—	11,59	34	02	capital
23	solteiro	ensino fundamental	—	12,00	41	00	outra
24	casado	superior	0	12,79	26	01	outra
25	casado	ensino médio	2	13,23	32	05	interior
26	casado	ensino médio	2	13,60	35	00	outra
27	solteiro	ensino fundamental	—	13,85	46	07	outra
28	casado	ensino médio	0	14,69	29	08	interior
29	casado	ensino médio	5	14,71	40	06	interior
30	casado	ensino médio	2	15,99	35	10	capital
31	solteiro	superior	—	16,22	31	05	outra
32	casado	ensino médio	1	16,61	36	04	interior
33	casado	superior	3	17,26	43	07	capital
34	solteiro	superior	—	18,75	33	07	capital
35	casado	ensino médio	2	19,40	48	11	capital
36	casado	superior	3	23,30	42	02	interior

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

Grau de instrução	n	$\bar{s}$	dp(S)	var(S)	$s_{(1)}$	q_1	q_2	q_3	$s_{(n)}$
Fundamental	12	7,84	2,79	7,77	4,00	6,01	7,13	9,16	13,65
Médio	18	11,54	3,62	13,10	5,73	8,84	10,91	14,48	19,40
Superior	6	16,48	4,11	16,89	10,53	13,65	16,74	18,38	23,30
Todos	36	11,12	4,52	20,46	4,00	7,55	10,17	14,06	23,30

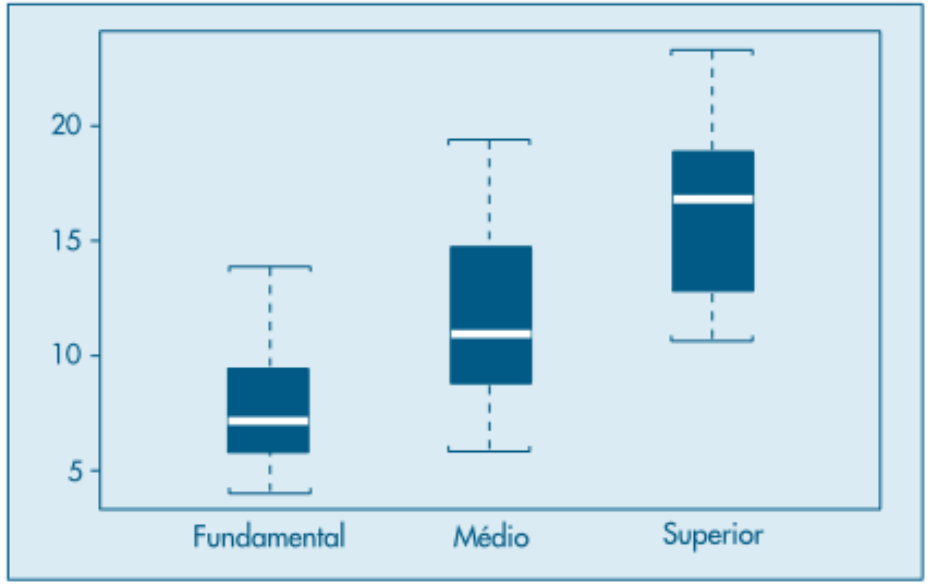
LADE

Comecemos a análise construindo a Tabela Abaixo, que contém medidas-resumo da variável S para cada categoria de Y.
A seguir, na Figura Abaixo , apresentamos uma visualização gráfica por meio de **box plots.**

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

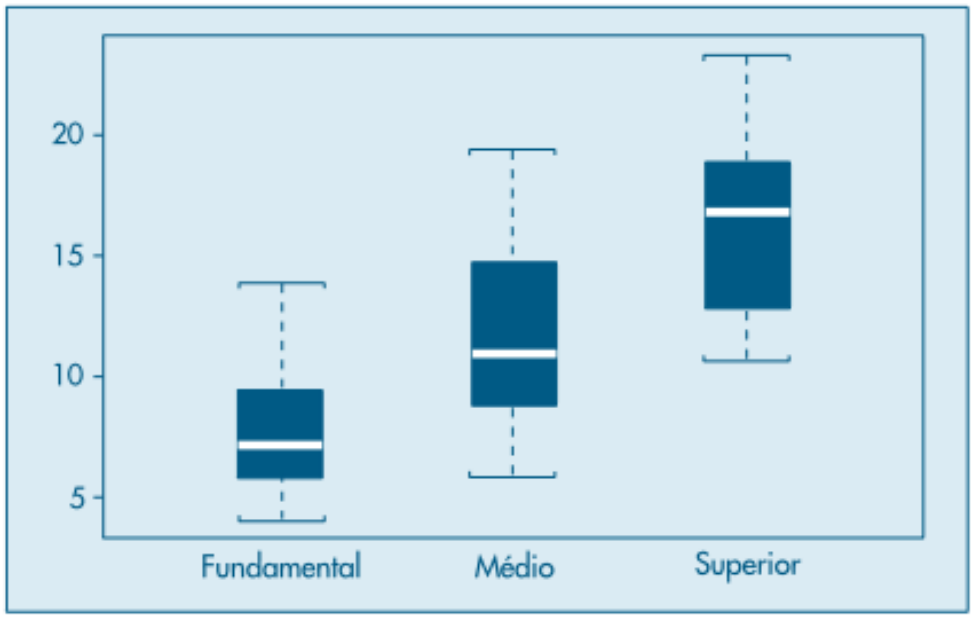
Grau de instrução	n	$\bar{s}$	dp(S)	var(S)	$s_{(1)}$	q_1	q_2	q_3	$s_{(n)}$
Fundamental	12	7,84	2,79	7,77	4,00	6,01	7,13	9,16	13,65
Médio	18	11,54	3,62	13,10	5,73	8,84	10,91	14,48	19,40
Superior	6	16,48	4,11	16,89	10,53	13,65	16,74	18,38	23,30
Todos	36	11,12	4,52	20,46	4,00	7,55	10,17	14,06	23,30

Box plots de salário segundo grau de instrução.



A leitura desses **resultados sugere uma dependência dos salários em relação ao grau de instrução**: o salário aumenta conforme aumenta o nível de educação 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.**

Box plots de salário segundo grau de instrução.

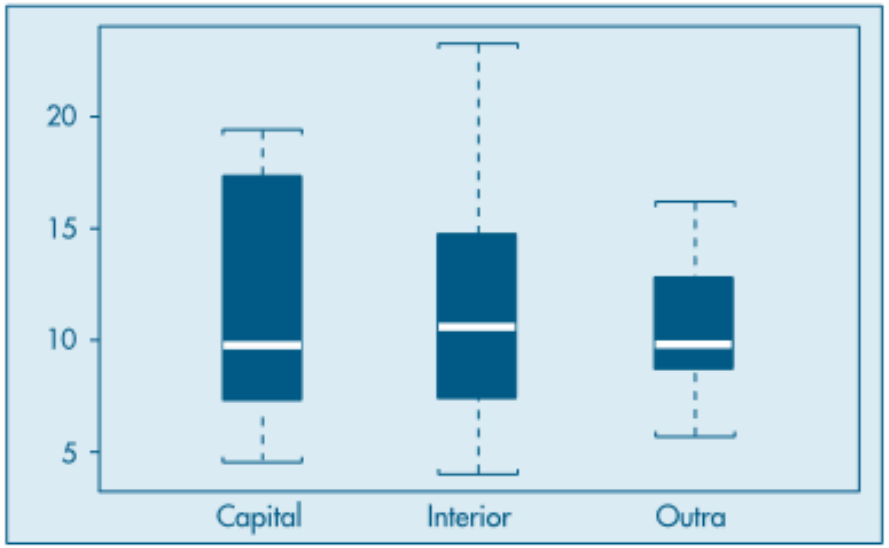


Na Tabela Abaixo e Figura Abaixo temos os resultados da análise dos salários em função da região de procedência (V), que mostram a inexistência de uma relação melhor definida entre essas duas variáveis. Ou, ainda, os salários estão mais relacionados com o grau de instrução do que com a região de procedência

Medidas-resumo para a variável salário segundo a região de procedência, na Companhia MB.

Região de procedência	n	$\bar{s}$	dp(S)	var(S)	$s_{(1)}$	q_1	q_2	q_3	$s_{(n)}$
Capital	11	11,46	5,22	27,27	4,56	7,49	9,77	16,63	19,40
Interior	12	11,55	5,07	25,71	4,00	7,81	10,64	14,70	23,30
Outra	13	10,45	3,02	9,13	5,73	8,74	9,80	12,79	16,22
Todos	36	11,12	4,52	20,46	4,00	7,55	10,17	14,06	23,30

Box plots de salário segundo região de procedência.



Como nos casos anteriores, **é conveniente poder contar com uma medida que quantifique o grau de dependência entre as variáveis.** Com esse intuito, convém observar que **as variâncias podem ser usadas como insumos para construir essa medida.** Sem usar a informação da variável categorizada, a variância calculada para a variável quantitativa 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 da quantitativa e portanto existe uma relação entre as duas variáveis.

Observe que, **para as variáveis S e Y**, **as variâncias de S dentro das três categorias são menores do que a global**. Já para **as variáveis S e V, temos duas variâncias de S maiores e uma menor do que a global**, o que corrobora a afirmação acima.

Necessita-se, então, de uma **medida-resumo da variância entre as categorias da variável qualitativa**. Vamos usar a média das variâncias, porém ponderada pelo número de observações em cada categoria, ou seja,

$$\overline{\text{var}(S)} = \frac{\sum_{i=1}^{k} n_i \text{var}_i(S)}{\sum_{i=1}^{k} n_i},$$

no qual k é o número de categorias (k = 3 nos dois exemplos acima) e $var_i(S)$ denota a variância de S dentro da categoria $i, i = 1, 2, \ldots, k$.

$$\overline{\text{var}(S)} = \frac{\sum_{i=1}^{k} n_i \text{var}_i(S)}{\sum_{i=1}^{k} n_i},$$

Pode-se mostrar que $\overline{var(S)} \leq var(S)$, 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,

$$R^2 = \frac{\text{var}(S) - \overline{\text{var}(S)}}{\text{var}(S)} = 1 - \frac{\overline{\text{var}(S)}}{\text{var}(S)}.$$

Note que $0 \leq R^2 \leq 1$. O símbolo R^2 é usual em análise de variância e regressão.

Voltando aos dados do Exemplo , vemos que para a variável S na presença de grau de instrução, tem-se

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

Grau de instrução	n	$\bar{s}$	dp(S)	var(S)	$s_{(1)}$	q_1	q_2	q_3	$s_{(n)}$
Fundamental	12	7,84	2,79	7,77	4,00	6,01	7,13	9,16	13,65
Médio	18	11,54	3,62	13,10	5,73	8,84	10,91	14,48	19,40
Superior	6	16,48	4,11	16,89	10,53	13,65	16,74	18,38	23,30
Todos	36	11,12	4,52	20,46	4,00	7,55	10,17	14,06	23,30

$$\overline{\text{var}(S)} = \frac{12(7,77) + 18(13,10) + 6(16,89)}{12 + 18 + 6} = 11,96,$$

de modo que

$$R^2 = 1 - \frac{11,96}{20,46} = 0,415,$$

$$\text{var}(S) = 20,46,$$

e dizemos que **41,5% da variação total do salário** é explicada pela variável grau de instrução.

Para S e região de procedência temos

Medidas-resumo para a variável salário segundo a região de procedência, na Companhia MB.

Região de procedência	n	$\bar{s}$	dp(S)	var(S)	$s_{(1)}$	q_1	q_2	q_3	$s_{(n)}$
Capital	11	11,46	5,22	27,27	4,56	7,49	9,77	16,63	19,40
Interior	12	11,55	5,07	25,71	4,00	7,81	10,64	14,70	23,30
Outra	13	10,45	3,02	9,13	5,73	8,74	9,80	12,79	16,22
Todos	36	11,12	4,52	20,46	4,00	7,55	10,17	14,06	23,30

$$\overline{\text{var}(S)} = \frac{11(27{,}27) + 12(25{,}71) + 13(9{,}13)}{11 + 12 + 13} = 20{,}20,$$

e, portanto,

$$R^2 = 1 - \frac{20{,}20}{20{,}46} = 0{,}013,$$

de modo que apenas **1,3% da variabilidade dos salários é explicada pela região de procedência**. A comparação desses dois números mostra maior relação entre S e Y do que entre S e V

Usando o Python

Montando a tabela para Salário segundo a Região de Procedência

```
CompanhiaMB['Região de Procedência'].value_counts()
```

```
Região de Procedência
outra       13
interior    12
capital     11
Name: count, dtype: int64
```

```
CompanhiaMB['Salario (x Sal Min)'].groupby(CompanhiaMB['Região de Procedência']).mean()
```

```
Região de Procedência
capital     11.455455
interior    11.550000
outra       10.445385
Name: Salario (x Sal Min), dtype: float64
```

LADE

```
CompanhiaMB['Salario (x Sal Min)'].groupby(CompanhiaMB['Região de Procedência']).std()
```

```
Região de Procedência
capital     5.476653
interior    5.296055
outra       3.145453
Name: Salario (x Sal Min), dtype: float64
```

```
CompanhiaMB['Salario (x Sal Min)'].groupby(CompanhiaMB['Região de Procedência']).var()
```

```
Região de Procedência
capital     29.993727
interior    28.048200
outra        9.893877
Name: Salario (x Sal Min), dtype: float64
```

LADE

```
print(CompanhiaMB['Salario (x Sal Min)'].mean())
print(CompanhiaMB['Salario (x Sal Min)'].std())
print(CompanhiaMB['Salario (x Sal Min)'].var())
```

```
11.122222222222222
4.587457503803861
21.044766349206352
```

```
Região_Procedencia={'capital':11,'interior':12,'outra':13}
Media_Procedencia={'capital':11.455455,'interior':11.550000,'outra':10.445385}
Std_Procedencia ={'capital':5.476653,'interior':5.296055,'outra':3.145453}
var_Procedencia={'capital':29.993727,'interior':28.048200,'outra':9.893877}

df = pd.DataFrame({'Região': Região_Procedencia,
                   'Média': Media_Procedencia,
                   'Desvio Padrão': Std_Procedencia,
                   'Variância': var_Procedencia})

# Exibindo o DataFrame
print(df)
```

```
          Região      Média  Desvio Padrão  Variância
capital       11  11.455455       5.476653  29.993727
interior      12  11.550000       5.296055  28.048200
outra         13  10.445385       3.145453   9.893877
```

LADE

```
            Região      Média  Desvio Padrão  Variância
capital         11  11.455455       5.476653  29.993727
interior        12  11.550000       5.296055  28.048200
outra           13  10.445385       3.145453   9.893877
```

```
] Total=[36,11.1222, 4.5874,21.0447]
  df.loc['Total']=Total
```

```
df
```

	Região	Média	Desvio Padrão	Variância
capital	11.0	11.455455	5.476653	29.993727
interior	12.0	11.550000	5.296055	28.048200
outra	13.0	10.445385	3.145453	9.893877
Total	36.0	11.122200	4.587400	21.044700

LADE

LABORATÓRIO DE ANÁLISES DE DADOS EDUCACIONAIS E ESTATÍSTICA APLICADA

IFCE - CAMPUS FORTALEZA

OBRIGADO!!!