DISCIPLINA: TADI

PROF[a].: Karla Lima

EACH-USP

Aula 7: 27/04/2016

Análise Bidimensional

- Anteriormente, organizamos e resumimos informações com relação a uma única variável (ou conjunto de dados), mas podemos estar 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).

Tabela de dados

Indivíduo	Variável						
	X_1	X_2	X_3	$\ldots$	X_j	$\ldots$	X_p
1	x_{11}	x_{12}	x_{13}	$\ldots$	x_{1j}	$\ldots$	x_{1p}
2	x_{21}	x_{22}	x_{23}	$\ldots$	x_{2j}	$\ldots$	x_{2p}
3	x_{31}	x_{32}	x_{33}	$\ldots$	x_{3j}	$\ldots$	x_{3p}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
i	x_{i1}	x_{i2}	x_{i3}	$\ldots$	x_{ij}	$\ldots$	x_{ip}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	x_{n1}	x_{n2}	x_{n3}	$\ldots$	x_{nj}	$\ldots$	x_{np}

- Nesta tabela, temos p variáveis e n indivíduos.

Exemplos

- Podemos ter um conjunto de dados $\{x_1, x_2, \ldots, x_n\}$, que são as temperaturas na cidade A, durante n meses, e outro conjunto de dados $\{y_1, y_2, \ldots, y_n\}$ que são temperaturas da cidade B, nos mesmos meses.
  - X: temperatura na cidade A.
  - Y: temperatura na cidade B.
- O principal objetivo aqui é explorar relações entre as variáveis. Exemplos:
  1. Existe relação entre a altura de pessoas e o sexo (feminino e masculino) em dada comunidade?
  2. Existe relação entre a renda e o consumo de um certo número de famílias?

Exemplos

Com relação a primeira pergunta (sobre as alturas), pode-se fazer algumas perguntas:

- qual a frequência esperada de uma pessoa dessa população ter, digamos, mais de $170cm$ de altura?
- qual a frequência esperada de uma mulher (ou homem) ter mais de $170cm$ de altura?

Se a resposta para as duas perguntas for a mesma, diríamos que não há associação entre as variáveis altura e sexo, e devemos incorporar esse conhecimento para melhorar o entendimento sobre os comportamentos das variáveis.

Dois conjuntos de dados

Quando consideramos duas variáveis, podemos ter as situações:

1. duas variáveis qualitativas;
2. duas variáveis quantitativas;
3. uma variável qualitativa e outra quantitativa.

Técnicas de análise de dados

1. Quando temos duas variáveis qualitativas, os dados são resumidos em **tabelas de dupla entrada**, onde aparecerão as frequências absolutas que pertencem simultaneamente a categorias de uma e outra variável;
2. Quando as duas variáveis são quantitativas, técnicas como gráficos de dispersão ou de quantis são apropriados;
3. Quando uma variável é qualitativa e a outra é 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.

Variáveis Qualitativas

Exemplo: Suponha que queiramos analisar o comportamento conjunto das variáveis G: grau de instrução e R: região de procedência(dados na Tabela 2.1).

Tabela 1: Distribuição conjunta das frequências das variáveis grau de instrução (G) e região de procedência (R).

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

Fonte: Tabela 2.1

- A linha dos totais fornece a distribuição da variável G.
- A coluna dos totais fornece a distribuição da variável R.
- Estas distribuições são chamadas de **distribuições marginais** e a tabela acima constitui a **distribuição conjunta** de G e R.

Variáveis Qualitativas

Podemos construir tabelas com frequências relativas (proporções).
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.

Em relação ao total geral

Tabela: Distribuição conjunta das proporções (%) em relação ao total geral das variáveis G e R.

R	G			Total
	Ensino Fundamental	Ensino Médio	Superior	
Capital	11%	14%	6%	31%
Interior	8%	19%	6%	33%
Outra	14%	17%	5%	36%
Total	33%	50%	17%	100%

Fonte: Tabela 1

- 11% dos empregados vêm da capital e tem o ensino fundamental;
- 31% dos indivíduos vêm da capital, 33% do interior e 36% de outras regiões.

Em relação ao total de cada coluna

Tabela: Distribuição conjunta das proporções (%) em relação aos totais de cada coluna das variáveis G e R.

R	G			Total
	Ensino Fundamental	Ensino Médio	Superior	
Capital	33%	28%	33%	31%
Interior	25%	39%	33%	33%
Outra	42%	33%	34%	36%
Total	100%	100%	100%	100%

Fonte: Tabela 1

- Entre os empregados com instrução até o ensino fundamental, 33% vêm da capital, enquanto que entre os empregados com ensino médio, 28% vêm da capital.
- Esta tabela serve para comparar a distribuição da procedência dos indivíduos conforme o grau de instrução.
- Analogamente, pode-se construir a distribuição das proporções em relação ao total das linhas.

Em relação ao total de cada linha

Tabela: Distribuição conjunta das proporções (%) em relação aos totais de cada linha das variáveis G e R.

R	G			Total
	Ensino Fundamental	Ensino Médio	Superior	
Capital	36%	46%	18%	100%
Interior	25%	58%	17%	100%
Outra	39%	46%	15%	100%
Total	33%	50%	17%	100%

Fonte: Tabela 1

- Entre os empregados cuja região de procedência é a capital, 36% têm o ensino fundamental, enquanto que entre os empregados do interior, 58% têm ensino médio.

Associação entre Variáveis Qualitativas

- Um dos principais objetivos de se construir uma distribuição conjunta de duas variáveis qualitativas é descrever a associação entre elas.
  - Ou seja, desejamos conhecer o grau de dependência entre elas, de modo que possamos prever melhor o resultado de uma delas quando conhecemos a realização da outra.
  - **Exemplo:** Se quisermos estimar qual a renda média de uma família moradora da cidade de São Paulo, a informação adicional sobre a classe social a que ela pertence nos permite estimar com maior precisão essa renda, pois sabemos que existe uma dependência entre as duas variáveis: renda familiar e classe social.

Associação entre Variáveis Qualitativas

Exemplo: Queremos verificar se existe ou não associação entre o sexo e a carreira escolhida por 200 alunos de Economia e Administração.

Tabela 2: Distribuição conjunta de alunos segundo o sexo X e o curso escolhido (Y).

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

Fonte: Dados hipotéticos

- Primeiramente, verifica-se que fica muito difícil tirar alguma conclusão, devido à diferença entre os totais marginais.
- Devemos construir as proporções segundo as linha ou as colunas para fazermos comparações.

Associação entre Variáveis Qualitativas

Tabela: Distribuição conjunta das proporções (%) de alunos segundo o sexo X e o curso escolhido (Y).

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

Fonte: Tabela 2

- Observamos que, independentemente do sexo, 60% das pessoas preferem Economia e 40% preferem Administração (coluna total).
- Não havendo dependência entre as variáveis, esperaríamos essas mesmas proporções para cada sexo.
- Na tabela, vemos que as proporções do sexo masculino (61% e 39%) e o feminino (58% e 42%) são próximas das marginais (60% e 40%).
    - Esses resultados parecem indicar não haver dependência entre as duas variáveis.
- Podemos concluir que as variáveis sexo e escolha do curso parecem serem não associadas.

Associação entre Variáveis Qualitativas

Considere um problema semelhante, mas envolvendo alunos de Física e Ciências Sociais.

Tabela: Distribuição conjunta das frequências e proporções (%) segundo o sexo X e o curso escolhido (Y).

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

Fonte: Dados hipotéticos

- Independentemente do sexo, observamos uma diferença bem acentuada nas proporções.
- Parece haver maior concentração de homens no curso de Física e de mulheres no de Ciências Sociais.
    - Nesse caso, as variáveis sexo e curso escolhido parecem ser associadas.

Medidas de Associação entre Variáveis Qualitativas

- A quantificação do grau de associação entre duas variáveis é feita pelos chamados *coeficientes de associação ou correlação*.
  - Essas medidas descrevem, por meio de um único número, a associação (ou dependência) entre duas variáveis.

Medidas de Associação entre Variáveis Qualitativas

Queremos verificar se a criação de determinado tipo de cooperativa está associada com algum fator regional. Os dados estão na tabela abaixo.

Tabela 3: 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%)	1551(100%)

Fonte: IBGE

- Observando a tabela acima temos a existência de certa dependência entre as variáveis.
- Se 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.
- Por exemplo, o número esperado de cooperativas de consumidores no Estado de São Paulo seria $648 \times 0,24 = 157$ e no Paraná seria $301 \times 0,24 = 73$, veja na próxima tabela.

Medidas de Associação entre Variáveis Qualitativas

Tabela 4: Valores esperados na tabela anterior assumindo a independência entre as duas variáveis.

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%)	1551(100%)

Fonte: Tabela 3

Medidas de Associação entre Variáveis Qualitativas

Tabela 5: Desvios entre observados e esperados.

Estado	Tipo de Cooperativa			
	Consumidor	Produtor	Escola	Outras
São Paulo	57	−32	−65	40
Paraná	−22	−22	59	−15
Rio G. do Sul	−35	54	6	−2

Fonte: Tabelas 3 e 4

- Comparando as duas tabelas, verificamos as discrepâncias existentes entre os valores observados (Tabela 3) e valores esperados (Tabela 4), caso as variáveis não fossem associadas.
- Na Tabela 5 temos os desvios: valores observados menos valores esperados. Daí, observamos:
  - A soma total dos resíduos é nula (soma de cada linha).
  - A casela Escola-São Paulo é aquela que apresenta o maior desvio da suposição de não-associação (−65). Nessa casela esperávamos 143 casos.

Medidas de Associação entre Variáveis Qualitativas

- A casela Escola-Paraná também tem um desvio alto (59), mas o valor esperado é menor (67).
- Uma maneira de observar esse fato é construir, para cada casela, a medida

$$\frac{(o_i - e_i)^2}{e_i}, \tag{1}$$

onde o_i é o valor observado e e_i é o valor esperado.

Medidas de Associação entre Variáveis Qualitativas

- Usando (1) 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 abaixo indicamos esses valores entre parênteses.
- Uma medida de afastamento global pode ser dada pela soma de todas as medidas (1). Essa medida é denominada χ^2 de Pearson, e no exemplo teríamos:

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

- Um valor muito grande de χ^2 indica associação entre as variáveis, o que parece ser o caso.

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)	−2(8,56)

Medidas de Associação entre Variáveis Qualitativas - Tabela de contingência

- Se os dados forem classificados de acordo com dois critérios (isto é, com duas classificações marginais) temos as tabelas de contingência.
- Com dois critérios de classificação temos uma tabela de contingência bidimensional, se houver mais de dois critérios na classificação, teremos uma tabela de contingência multidimensional.
- Aqui trabalharemos somente com a bidimensional.
    - O aspecto de uma tabela de contingência é o de uma tabela com linhas, correspondentes a um dos critérios, e com colunas, correspondente ao outro critério.

Medidas de Associação entre Variáveis Qualitativas - Tabela de contingência

Tabela: Notação para tabelas de contingência.

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

- 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.

Medidas de Associação entre Variáveis Qualitativas

Aqui, temos

- 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.

Medidas de Associação entre Variáveis Qualitativas

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,

$$\frac{n_{ij}}{n_{.j}} = \frac{n_{i.}}{n}, i = 1, 2, \ldots, r, j = 1, 2, \ldots, s.$$

Daí,

$$n_{ij} = \frac{n_{i.} n_{.j}}{n}, i = 1, 2, \ldots, r, j = 1, 2, \ldots, s \qquad (2)$$

Sob a hipótese de independência, em termos das frequências relativas temos que:

$$f_{ij} = \frac{n_{ij}}{n} = \frac{\frac{n_{i.} n_{.j}}{n}}{n} = \frac{n_{i.} n_{.j}}{n^2} = \frac{n_{i.}}{n} \frac{n_{.j}}{n} = f_{i.} f_{.j}$$

Portanto,

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

Medidas de Associação entre Variáveis Qualitativas

Chamando de frequências esperadas os valores dados pelos segundos membros de (2), e denotando-as por n_{ij}^*, temos que o χ^2 de Pearson se reduz a

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

- Se a hipótese de não-associação for verdadeira, o valor calculado de χ^2 deve estar próximo de zero.
- Se as variáveis forem associadas, o valor de χ^2 deve ser "grande".

Medidas de Associação entre Variáveis Qualitativas

Em termos de frequência relativa temos

$$\chi^2 = n \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(f_{ij} - f_{ij}^*)^2}{f_{ij}^*}.$$

Exercício

A Companhia *A* de dedetização afirma que o processo por ela utilizado garante um efeito mais prolongado do que aquele obtido por seus concorrentes mais diretos. Uma amostra de vários ambientes dedetizados foi escolhida e anotou-se a duração do efeito de dedetização. Os resultados estão na tabela abaixo. Você acha que existe alguma evidência a favor ou contra a afirmação feita pela Companhia *A*?

Companhia	Duração do efeito de dedetização		
	Menos de 4 meses	De 4 a 8 meses	Mais de 8 meses
A	64	120	16
B	104	175	21
C	27	48	5

Associação entre Variáveis Quantitativas

- Quando as duas variáveis são quantitativas podemos usar o mesmo tipo de análise apresentado anteriormente.
  - A distribuição conjunta pode ser resumida em tabelas de dupla entrada e, por meio das distribuições marginais, é possível estudar a associação das variáveis.
  - Para evitar um grande número de entradas, agrupamos os dados em intervalos de classe, de modo semelhante ao resumo feito no caso unidimensional.
- Além das tabelas as variáveis quantitativas são passíveis de procedimentos gráficos.
- **Gráfico de dispersão:** É um dispositivo muito utilizado para se verificar a associação entre duas variáveis quantitativas.

Associação entre Variáveis Quantitativas

- Quando as duas variáveis são quantitativas podemos usar o mesmo tipo de análise apresentado anteriormente.
  - A distribuição conjunta pode ser resumida em tabelas de dupla entrada e, por meio das distribuições marginais, é possível estudar a associação das variáveis.
  - Para evitar um grande número de entradas, agrupamos os dados em intervalos de classe, de modo semelhante ao resumo feito no caso unidimensional.
- Além das tabelas as variáveis quantitativas são passíveis de procedimentos gráficos.
- **Gráfico de dispersão:** É um dispositivo muito utilizado para se verificar a associação entre duas variáveis quantitativas.

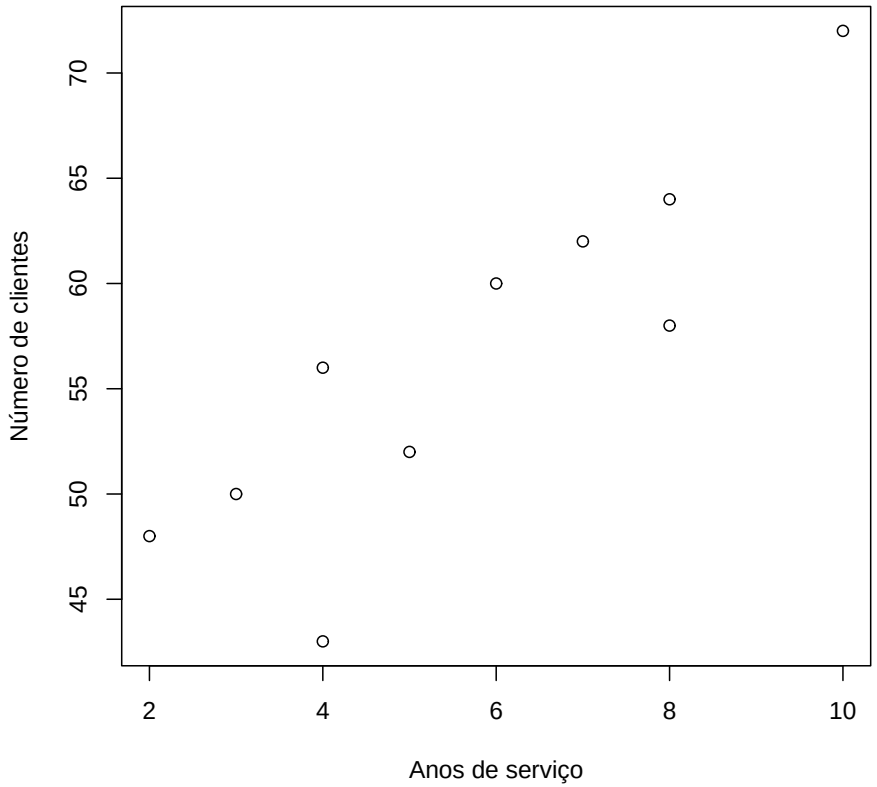
Associação entre Variáveis Quantitativas

Exemplo 1: 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

Associação entre Variáveis Quantitativas

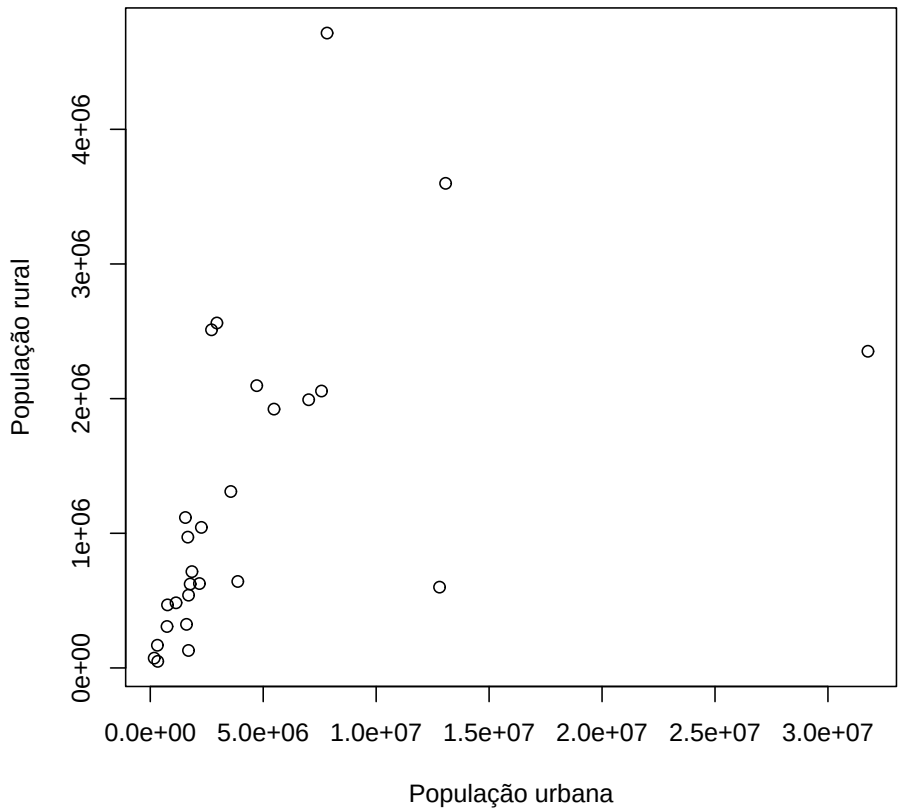
Gráfico de dispersão para as variáveis X: anos de serviço e Y: número de clientes.



- Parece haver associação entre as variáveis, pois o conjunto, à medida que aumenta o tempo de serviço, aumenta o número de clientes.

Associação entre Variáveis Quantitativas

Exemplo 2: Considere o gráfico de dispersão para as variáveis X: população urbana e Y: população rural no Brasil.



- Parece não haver associação entre as variáveis, pois o os pontos não apresentam nenhuma tendência particular.

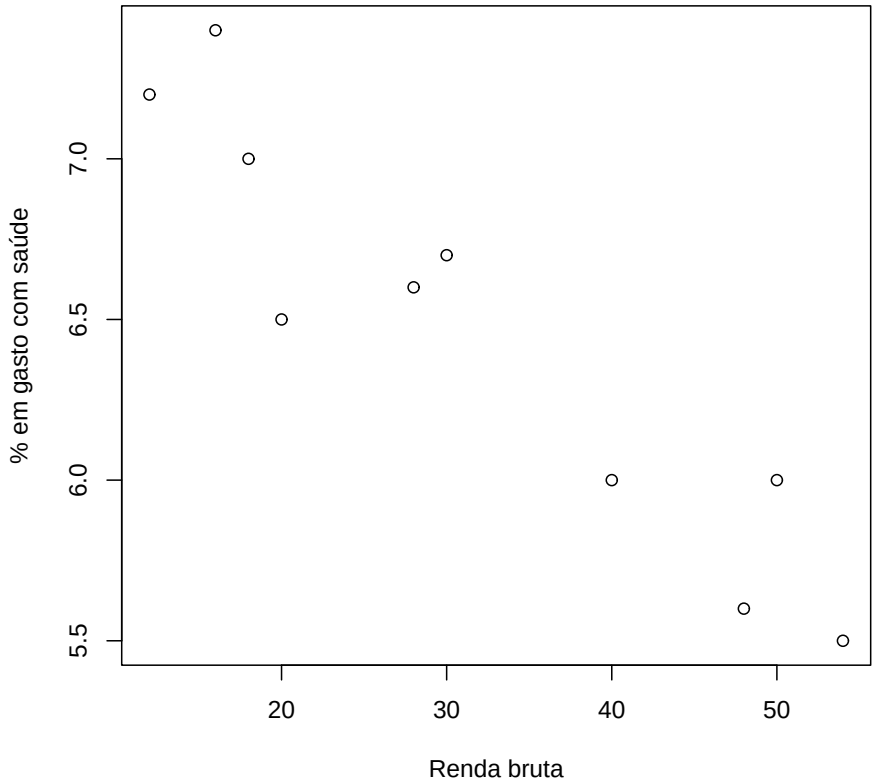
Associação entre Variáveis Quantitativas

Exemplo 3: Renda bruta mensal (X), expressa em número de salários mínimos, e porcentagem da renda bruta anual gasta com assistência médica (Y) para um conjunto de dez famílias.

Família	Renda bruta mensal (X)	% da renda gasta com saúde (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

Associação entre Variáveis Quantitativas

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



- Parece haver uma associação inversa entre as variáveis, isto é, aumentando a renda bruta, diminui a porcentagem sobre ela gasta em assistência médica.

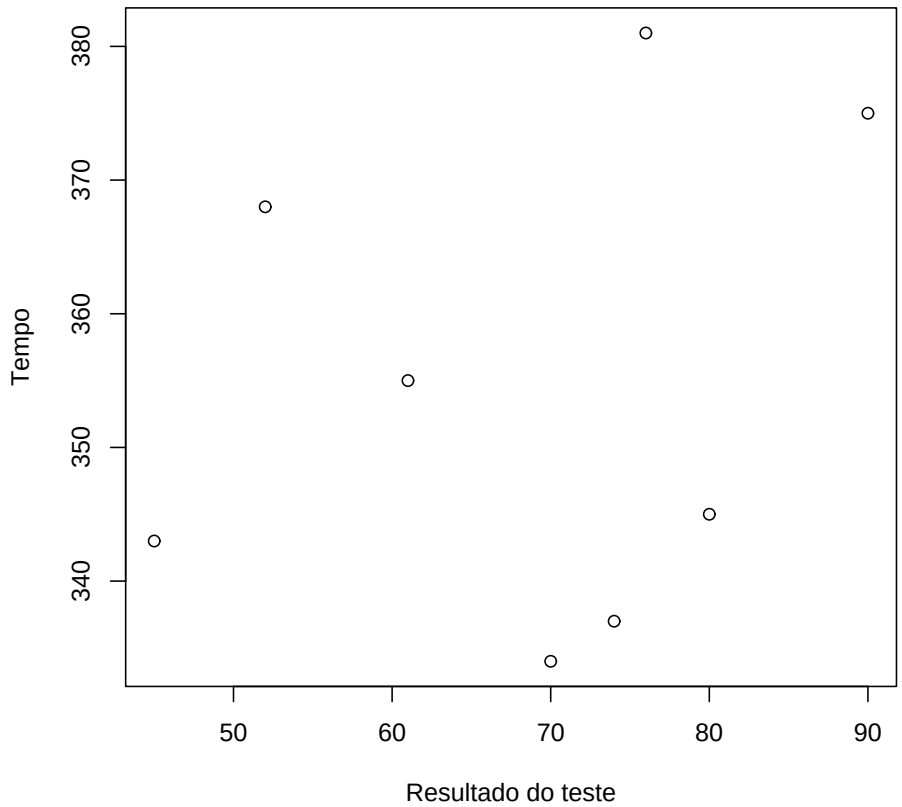
Associação entre Variáveis Quantitativas

Exemplo 4: 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. Aqui temos as variáveis X: resultado obtido no teste (máximo 100 pontos) e Y: tempo, em minutos, necessário para operar a máquina satisfatoriamente.

Indivíduo	Resultado do teste (X)	Tempo de operação de máquina (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

Associação entre Variáveis Quantitativas

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



- Parece não haver associação entre as variáveis.

Associação entre Variáveis Quantitativas

- A visualização dos dados através do gráfico de dispersão é o primeiro passo na análise, mas também desejamos quantificar a intensidade da associação entre duas variáveis.
- Então além do gráfico de dispersão usamos o coeficiente de correlação amostral (avalia o quanto a nuvem de pontos no gráfico de dispersão aproxima-se de uma reta).
- Esta estatística mede o grau de linearidade na relação entre X e Y e é, geralmente, denotado por $Corr(X, Y)$.
  - Temos $-1 \leq Corr(X, Y) \leq 1$.
  - Quando $Corr(X, Y)$ está próximo de 0, existe pouca ou nenhuma relação linear entre X e Y.
  - Quando $Corr(X, Y)$ está próximo de 1, indica uma relação positiva forte.
  - Quando $Corr(X, Y)$ está próximo de -1, indica uma relação negativa forte.

Associação entre Variáveis Quantitativas

Definição: Dados n pares de valores $(x_1, y_1), \ldots, (x_n, y_n)$, chamaremos de coeficiente de correlação linear amostral entre as variáveis X e Y a média dos produtos dos valores padronizados das variáveis, ou seja,

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

Associação entre Variáveis Qualitativas e Quantitativas

Aqui, é comum analisar o que acontece com a variável quantitativa dentro de cada categoria da variável qualitativa.

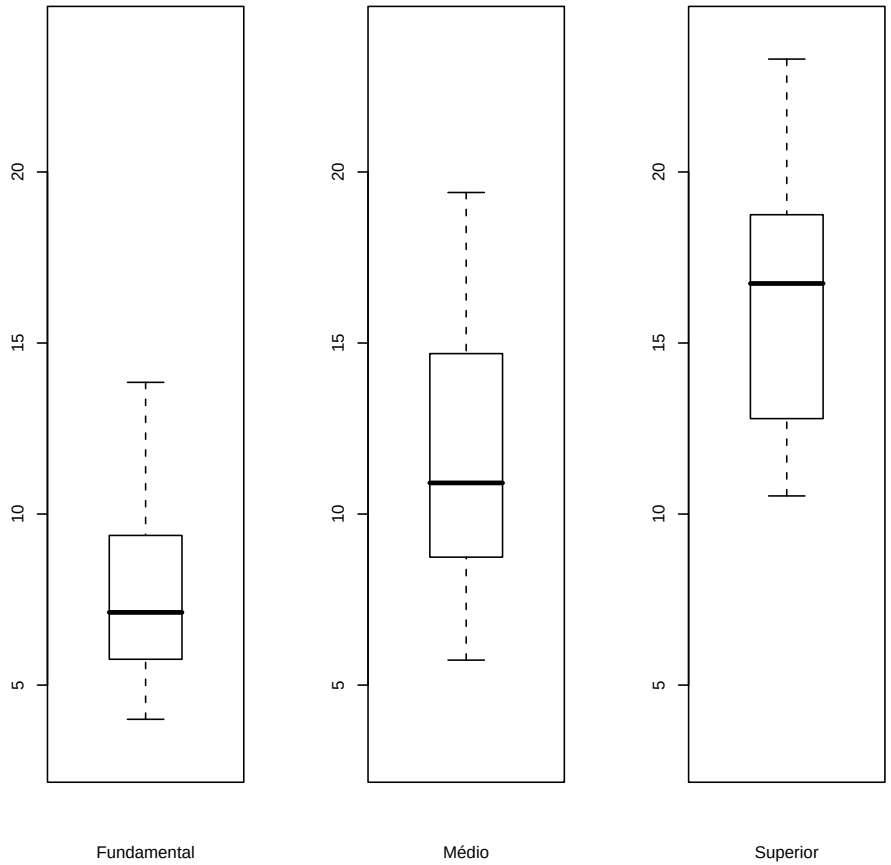
- Essa análise pode ser conduzida por meio de medidas-resumo, histogramas, *box plots* ou ramo-e-folhas.

Exemplo: (Tabela 2.1) Desejamos analisar o comportamento dos salários (S) dentro de cada categoria de grau de instrução (G), ou seja, investigar o comportamento conjunto das variáveis S e G.

- S: Variável quantitativa
- G: Variável qualitativa

Grau	n	$\bar{s}$	$dp(S)$	$var(S)$	$s_{(1)}$	q_1	q_2	q_3	$s_{(n)}$
Fund.	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

Associação entre Variáveis Qualitativas e Quantitativas - *Box-Plots* de salários segundo grau de instrução



- Parece haver dependência dos salários em relação ao grau de instrução: o salário aumenta conforme aumenta o nível de educação do indivíduo.

Medida de associação entre Variáveis Qualitativas e Quantitativas

- Aqui, também é conveniente ter uma medida que quantifique o grau de dependência entre as variáveis.
- 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.

Medida de associação entre Variáveis Qualitativas e Quantitativas

Grau	n	$\bar{s}$	$dp(S)$	$var(S)$	$s_{(1)}$	q_1	q_2	q_3	$s_{(n)}$
Fund.	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	7,55	10,17	14,06	23,30

Região	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	7,55	10,17	14,06	23,30

- Para as variáveis S e G, as variâncias de S dentro das três categorias são menores do que a global (evidência de dependência).
- Para as variáveis S e R, temos duas variâncias de S maiores e uma menor do que a global (não há evidência de dependência).

Medida de associação entre Variáveis Qualitativas e Quantitativas

Usa-se a média das variâncias ponderada pelo número de observações em cada categoria como uma medida-resumo da variância entre as categorias da variável qualitativa. Ou seja,

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

onde k é o número de categorias e $var_i(S)$ denota a variância de S dentro de cada categoria i, $i = 1, 2, \ldots, k$.

Definição: O grau de associação entre duas variáveis (uma qualitativa e outra quantitativa (S)) como o ganho relativo na variância, obtido pela introdução da variável qualitativa é:

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

Aqui, temos $\overline{var(S)} \leq var(S)$ e $0 \leq R^2 \leq 1$.

Medida de associação entre Variáveis Qualitativas e Quantitativas

Nos exemplos anteriores, para a variável S na presença de grau de instrução, temos

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

e

$$var(S) = 20,46.$$

Daí,

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

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

Medida de associação entre Variáveis Qualitativas e Quantitativas

Para a variável S e região de procedência temos

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

Daí,

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

dizemos que $1,3\%$ da variação total do salário é explicada pela região de procedência.

- A comparação desses dois números ($41,5\%$ e $1,3\%$) mostra maior relação entre S e G do que entre S e R.

Gráficos $q \times q$

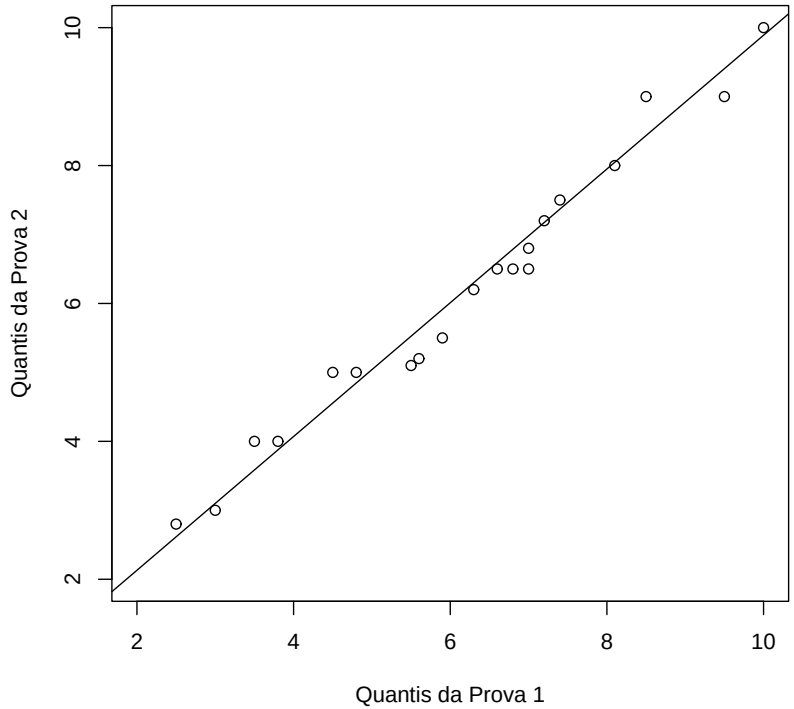
- Outro tipo de representação gráfica que podemos utilizar para duas variáveis é o **gráfico quantis×quantis**.
- Suponha que temos duas variáveis X e Y, cujas medidas estão na mesma unidade.
  - Exemplo: Temperatura de duas cidades; alturas de dois grupos de indivíduos; etc.
- O gráfico $q \times q$ é um gráfico dos quantis de X contra os quantis de Y.
- O gráfico de dispersão fornece uma possível relação global entre as variáveis.
  - O gráfico $q \times q$ mostra se valores pequenos de X estão relacionados com valores pequenos de Y.
  - O gráfico $q \times q$ mostra se valores intermediários de X estão relacionados com valores intermediários de Y.
  - O gráfico $q \times q$ mostra se valores grandes de X estão relacionados com valores grandes de Y.

Gráficos $q \times q$

Exemplo: Na tabela abaixo temos notas de 20 alunos em duas provas de Estatística.

Aluno	Prova 1	Prova 2
1	8,5	8,0
2	3,5	2,8
3	7,2	6,5
4	5,5	6,2
5	9,5	9,0
6	7,0	7,5
7	4,8	5,2
8	6,6	7,2
9	2,5	4,0
10	7,0	6,8
11	7,4	6,5
12	5,6	5,0
13	6,3	6,5
14	3,0	3,0
15	8,1	9,0
16	3,8	4,0
17	6,8	5,1
18	10,0	10,0
19	4,5	5,5
20	5,9	5,0

Gráficos $q \times q$ - Exemplo



- Os pontos estão razoavelmente dispersos ao redor da reta $y = x$, mostrando que as notas dos alunos nas duas provas não são muito diferentes.
- Para as notas abaixo de 5, os alunos tiveram notas maiores na segunda prova.
- Das notas de 5 a 8, os alunos tiveram notas melhores na primeira prova.
- A maioria das notas estão concentradas entre 5 e 8.