Análise Bidimensional

Prof. Éder Brito - Instituto Federal de Goiás (IFG)

Probabilidade e Estatística - BCC

22 e 26 de março de 2024

Introdução

- É razoável pensar que em diversas situações analisaremos dados onde mais de uma variável é observada para cada unidade amostral.
- Nessa situações também é razoável pensar que um dos objetivos da análise estatística será analisar o comportamento conjunto de duas ou mais dessas variáveis.
- Nesse caso, estaremos realizando uma *análise bidimensional* dos dados. O principal objetivo das análises nessa situação é exploras as relações ou similaridades entre as variáveis.
- Muitas vezes queremos verificar se há uma relação de causa e efeito entre as duas variáveis (se as variáveis são dependentes ou não), se é possível estudar/prever uma das variáveis através da outra (que é mais fácil de medir), ou calcular uma medida de correlação ou de dependência entre as variáveis.

Introdução

De modo geral, os dados podem ser representados matricialmente em formato de Tabela, da seguinte forma:

Indivíduo	Variáveis					
	X_1	X_2	$\cdots$	X_j	$\cdots$	X_p
1	$x_{1,1}$	$x_{1,2}$	$\cdots$	$x_{1,j}$	$\cdots$	$x_{1,p}$
2	$x_{2,1}$	$x_{2,2}$	$\cdots$	$x_{2,j}$	$\cdots$	$x_{2,p}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
i	$x_{i,1}$	$x_{i,2}$	$\cdots$	$x_{i,j}$	$\cdots$	$x_{i,p}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
n	$x_{n,1}$	$x_{n,2}$	$\cdots$	$x_{n,j}$	$\cdots$	$x_{n,p}$

A ideia da análise é, portanto, explorar relações entre as colunas ou, algumas vezes, entre as linhas da Tabela.

Introdução

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. No entanto, 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.

Variáveis qualitativas -Introdução

- Já discutimos anteriormente (e fizemos com os dados de Pobreza) que quando para variáveis qualitativas, os dados podem ser resumidos em *tabelas de dupla entrada (ou de contingência)*, onde aparecerão as frequências absolutas ou contagens de indivíduos que pertencem simultaneamente a categorias de uma e outra variável.

- Também já vimos que a análise desse tipo de dados é melhor realizada com as *frequências relativas (ou proporções, ou porcentagens)*. Aqui existem três possibilidades de expressarmos a proporção de cada casa da tabela:
  - em relação ao total geral
  - em relação ao total de cada linha
  - em relação ao total de cada coluna

  Naturalmente, de acordo com o objetivo de cada problema em estudo, uma delas será a mais conveniente.

Variáveis qualitativas - Exemplo

Lembram-se da Tabela com informações do 36 funcionários que estudamos no início do curso? Vamos analisar a *distribuição conjunta* das frequências das variáveis região de procedência (X) e grau de instrução (Y).

X \ Y	Ensino Fund.	Ensino Médio	Superior	Total
Capital	4	5	2	11
Interior	3	7	2	12
Outra	5	6	2	13
Total	12	18	6	36

Obs.: No R, o comando para essa tabela é simplesmente `table(X,Y)`

Variáveis qualitativas - Distribuições Marginais

- Cada elemento do corpo da tabela dá a frequência observada das realizações simultâneas das variáveis X e Y.
- A *linha* dos totais fornece a distribuição da variável Y, ao passo que a coluna dos totais fornece a distribuição da variável X. Essas distribuições são chamadas de *distribuições marginais* de cada variável.
- Verificar se duas variáveis são independentes ou têm relação nada mais é do que verificar se a distribuição dos valores entre as categorias de uma variável acompanha ou não a distribuição marginal dessa variável.

Variáveis qualitativas - Exemplo

Podemos obter a distribuição conjunta das frequências relativas, expressas como proporções do total geral:

X \ Y	Ensino Fund.	Ensino Médio	Superior	Total
Capital	11%	14%	6%	31%
Interior	8%	19%	6%	33%
Outra	14%	17%	5%	36%
Total	33%	50%	17%	100%

Obs.: No R, o comando para essa tabela é simplesmente `prop.table(table(X,Y))`

Variáveis qualitativas - Exemplo

Suponhamos que queremos verificar se a região de procedência tem relação (ou interfere) com o nível de escolaridade. Para isso, calculamos a distribuição das proporções em relação ao total das colunas:

X \ Y	Ensino Fund.	Ensino Médio	Superior	Total
Capital	33%	28%	33%	31%
Interior	25%	39%	33%	33%
Outra	42%	33%	34%	36%
Total	100%	100%	100%	100%

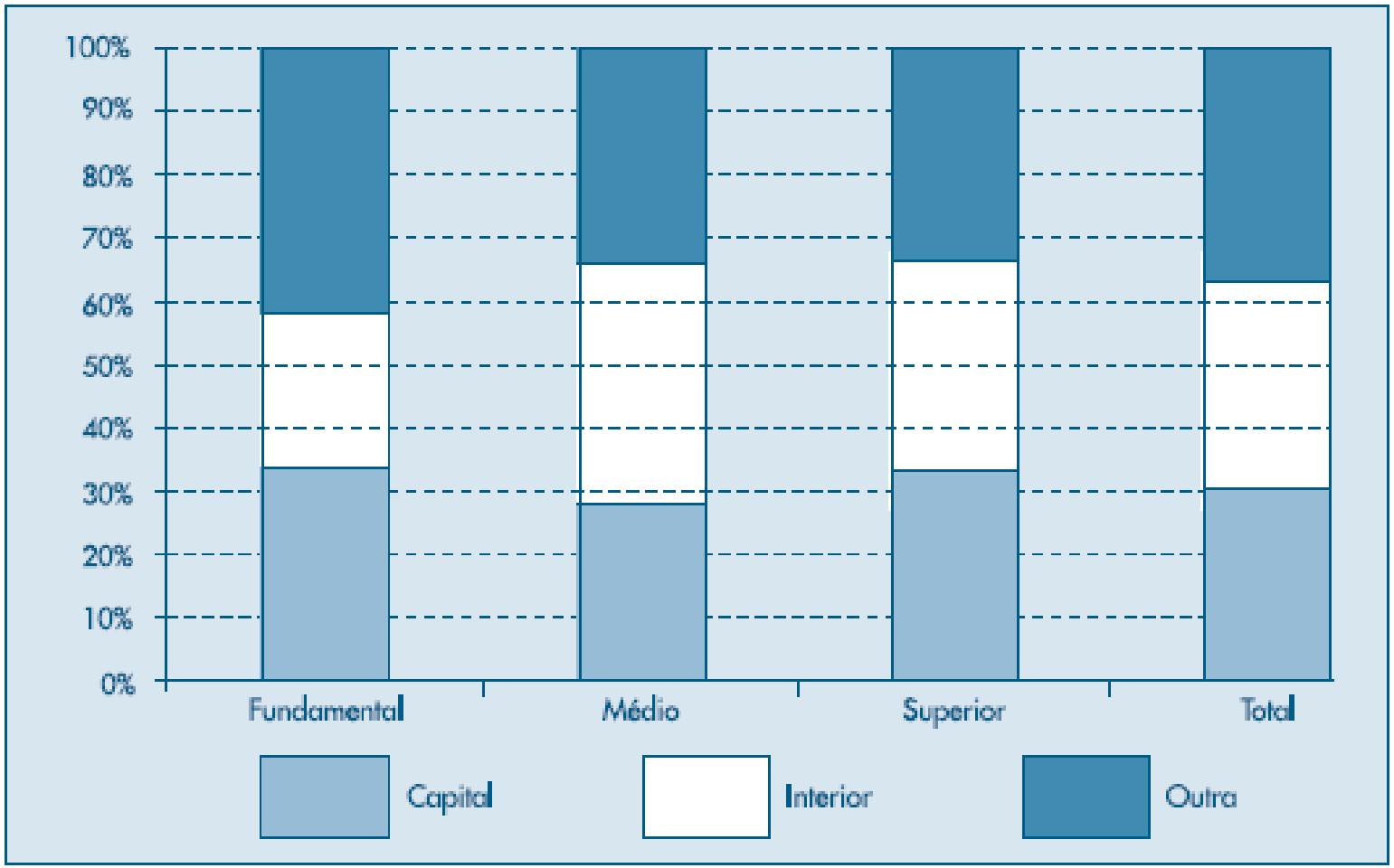
Obs.: No R, o comando para essa tabela é simplesmente `prop.table(table(X,Y), margin=2)`

Variáveis qualitativas - Exemplo

- Pela Tabela anterior, podemos concluir algumas coisas:
  - Se tomarmos um funcionário qualquer da empresa, as chances de ele ser de uma determinada região é muito próxima das outras (31, 33 e 36 porcento).
  - Se soubermos que esse funcionário tem curso superior, nada se altera nessa probabilidade, pois a distribuição das porcentagens é muito próxima da distribuição marginal.
  - Por outro lado, se soubermos que esse funcionário só tem o Ensino Fundamental, há mais chance de ele ser de Outra região do que do Interior. Essa distribuição já se distancia da distribuição marginal, sugerindo a existência de relação de dependência entre a categoria "Ensino Fundamental" e a variável região de procedência.

Variáveis qualitativas - Exemplo

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.

- Isso nos ajudará, por exemplo, a prever melhor o resultado de uma dessas variáveis quando conhecermos a realização da outra.

- Por 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 - Exemplo

Queremos verificar se existe ou não associação entre o sexo (Y) e a carreira (X) escolhida por 200 alunos de Economia e Administração. Esses dados estão na Tabela abaixo:

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

Associação entre variáveis - Exemplo

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

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

Verificamos na coluna Total (marginal) que independentemente do sexo, 60% das pessoas preferem Economia e 40% preferem Administração. Além disso, as proporções dos cursos de acordo com os sexos estão muito próximas das marginais (60 e 40 porcento). Isso nos indica que não parece haver dependência entre as duas variáveis. Nesse caso, a conclusão é de que as variáveis sexo e escolha do curso parecem ser *não associadas*.

Associação entre variáveis - Outro Exemplo

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

X \ Y	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%)

Aqui a dependência das variáveis fica evidente. As proporções dos sexos por curso tem disparidade bem acentuada das proporções marginais. Parece haver maior concentração de homens no curso de Física e de mulheres do curso de Ciências Sociais. Portanto, nesse caso, as variáveis sexo e curso escolhido parecem ser *associadas*.

Medida 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.
- Em geral esses coeficientes variam entre 0 e 1 ou entre -1 e 1, onde a proximidade do zero indica falta de associação.
- A ideia é muito simples: comprar os valores observados nas categorias com os valores esperados de acordo com a distribuição marginal caso não houvesse associação.

Exemplo - Associação de variáveis

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

Estado	Tipo de Cooperativa				Total
	Consumidor	Produtor	Escola	Outras	
SP	214(33%)	237(37%)	78(12%)	119(18%)	648(100%)
PR	51(17%)	102(34%)	126(42%)	22(7%)	301(100%)
RS	111(18%)	304(51%)	139(23%)	45(8%)	602(100%)
Total	376(24%)	643(42%)	343(22%)	189(12%)	1551(100%)

A análise 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.

Exemplo - Associação de variáveis

Podemos então construir uma Tabela com os números esperados de cada tipo de cooperativa por estado, baseado na distribuição marginal das cooperativas:

Estado	Tipo de Cooperativa				Total
	Consumidor	Produtor	Escola	Outras	
SP	157(24%)	269(42%)	143(22%)	79(12%)	648(100%)
PR	73(24%)	124(42%)	67(22%)	37(12%)	301(100%)
RS	146(24%)	250(42%)	133(22%)	73(12%)	602(100%)
Total	376(24%)	643(42%)	343(22%)	189(12%)	1551(100%)

É possível, portanto, comparar as duas tabelas apresentadas anteriormente e obter os *desvios* entre os valores observados e esperados. Com esses desvios construiremos uma medida de associação global das variáveis.

Desvios relativos e medida χ^2 de Pearson

- Os *desvios relativos* de cada entrada da tabela são calculados por

$$\frac{(o-e)^2}{e},$$

onde o representa o valor observado na casa da tabela e e o respectivo valor esperado pela distribuição marginal.

- Uma medida do afastamento global entre os valores observados e esperados pode ser dada pela soma de todos os desvios relativos. Essa medida é denominada χ^2 (qui-quadrado) de Pearson, dada por

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{s}\frac{(o_{i,j}-e_{i,j})^2}{e_{i,j}},$$

onde r é o número de categorias da variável das linhas e s é o número de categorias da variável das colunas.

Medidas de Associação

- Pearson definiu uma medida de associação, baseada na medida χ^2, chamada *coeficiente de contingência*, dado por

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

onde n é número total de dados.

- Contudo, o coeficiente acima não varia entre 0 e 1. 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 $r = s$.

Variáveis quantitativas - Introdução

- Para comparar duas variáveis quantitativas e verificar se existe associação entre elas, um primeiro procedimento que pode ser utilizado é o apresentado anteriormente para variáveis qualitativas, desde que se organize os dados de cada variável em intervalos de classe e, daí, utiliza-se a distribuição marginal como referência para verificar a associação.

- No entanto, existem formas específicas de se avaliar a associação entre duas variáveis quantitativas, seja usando um *gráfico de dispersão* ou um *coeficiente de correlação*.

Associação de variáveis quantitativas - Exemplos

- **Exemplo:** Uma médica ginecologista pode colher dados da idade e do tamanho de um grupo de fetos. Será que existe alguma relação entre essas característica? Se sim, de que forma se dá essa relação? Será possível utilizar informações de uma dessas características para obter informações acerca da outra?

- Outros exemplos
  - Idade e altura das crianças
  - Tempo de prática de esportes e ritmo cardíaco
  - Tempo de estudo e nota na prova
  - Taxa de desemprego e taxa de criminalidade
  - Expectativa de vida e taxa de analfabetismo
  - Memória de um computador e o tempo de execução de um código.

Gráficos de Dispersão

- O primeiro método (e mais intuitivo) de observar uma possível associação entre duas variáveis quantitativas é observar o *gráfico de dispersão* dessas duas variáveis.

- São gráficos que apresentam informação numéricas das duas variáveis, obtidas de uma mesma amostra. Isto é, cada elemento da amostra fornecerá duas informações numéricas, uma para cada variável.

- Essas informações são interpretadas como um par ordenado, e naturalmente, possuem um lugar de representação geométrica em um plano cartesiano.

- Ao analisar o comportamento dos pontos obtidos, podemos ter uma primeira percepção da existência ou não de correlação entre as variáveis e de que tipo ela poderá ser (caso exista).

Exemplo - Associação de variáveis quantitativas

O gráfico a seguir é um exemplo de dados do caso de estudo citado anteriormente. É evidente que existe uma relação entre o tamanho dos fetos e a idade gestacional. E ainda, é fácil concluir que quanto maior a idade, maior o tamanho do feto.

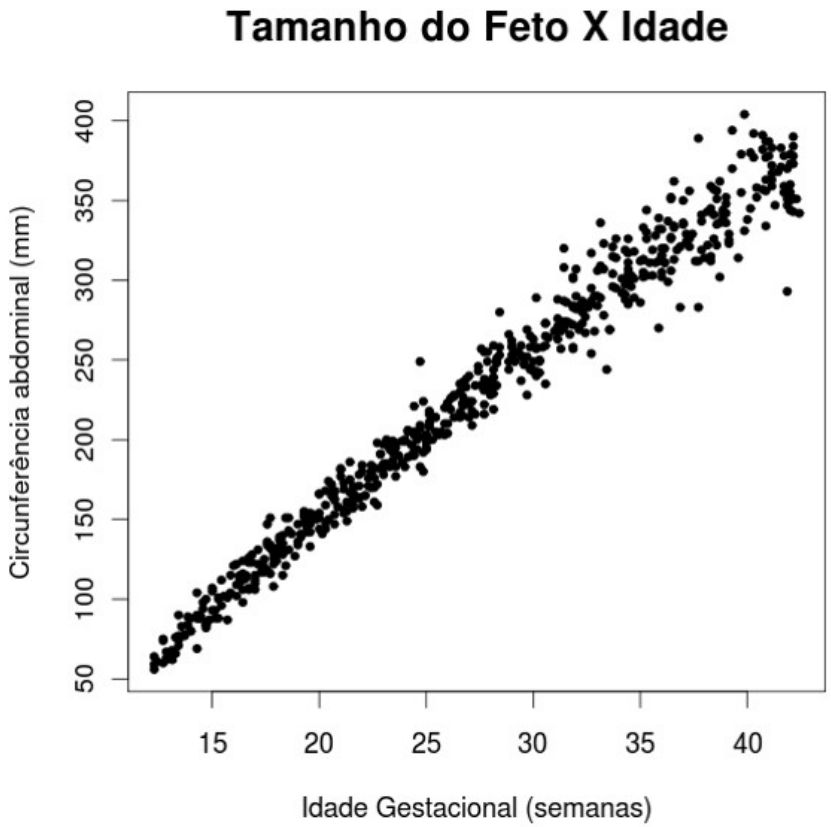


Diagrama de Dispersão

- Para construir um diagrama de dispersão, definimos as duas variáveis (características), denominando-as como variáveis X e Y.

- A variável X é a variável explanatória (independente), enquanto a variável Y é a varável dependente (caso exista de fato a dependência). Essa convenção será útil para o estudos posteriores, porém não faz diferença para a análise da correlação entre as variáveis.

- A seguir, analisemos dois conjuntos de dados e seus respectivos diagramas de dispersão:

Exemplo - Diagrama de Dispersão

Conjunto A

X	Y
1	2
2	0
3	6
4	3
5	9
6	4
7	10
8	8
9	12
10	8

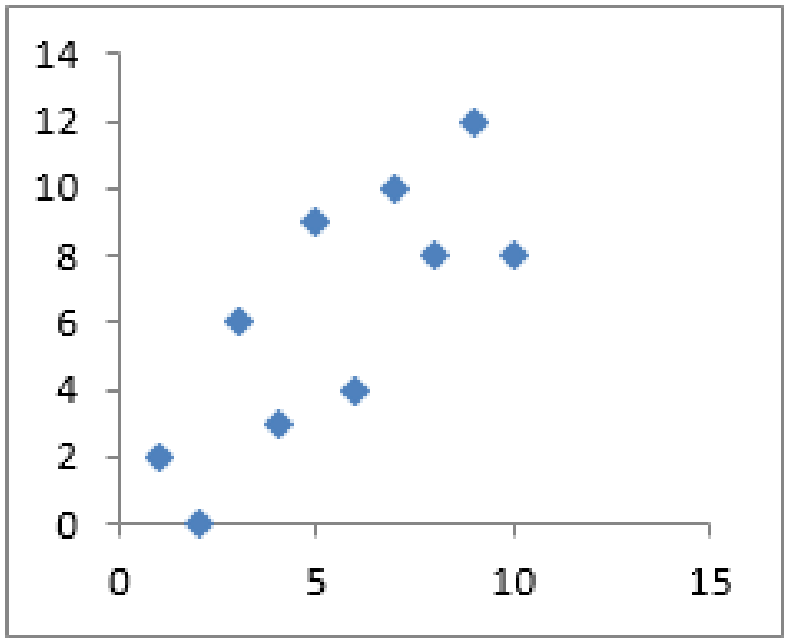


Figura: Diagrama de dispersão do Conjunto A

Exemplo - Diagrama de Dispersão

Conjunto B

X	Y
1	8
2	12
3	8
4	10
5	4
6	9
7	3
8	6
9	0
10	2

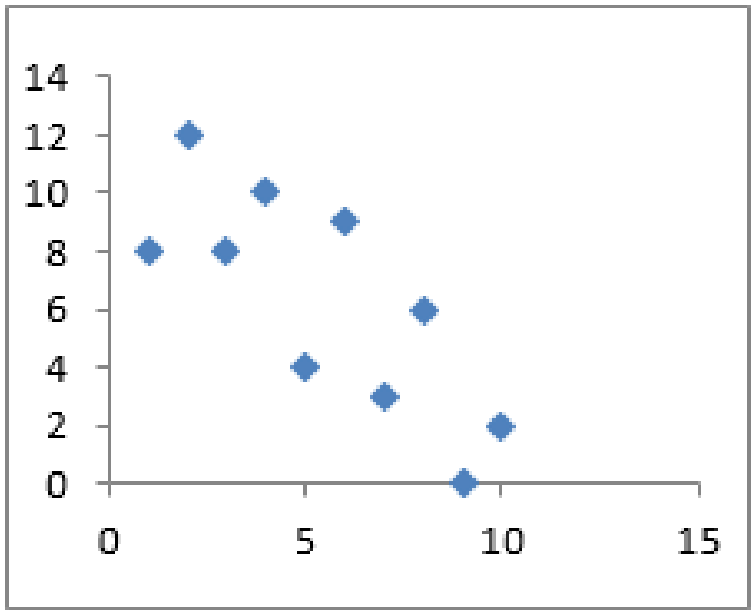


Figura: Diagrama de dispersão do Conjunto B

Análise dos diagramas dos exemplos - Correlação Positiva e Negativa

- Ao analisar tanto os dados numéricos, quanto os diagramas anteriores, é possível verificar que existe relação entre os dados X e Y em ambos conjuntos.

- No conjunto A, a medida que os valores de X aumentam, os valores de Y, em geral, também aumentam. No conjunto B, a medida que os valores de X aumentam, os de Y diminuem.

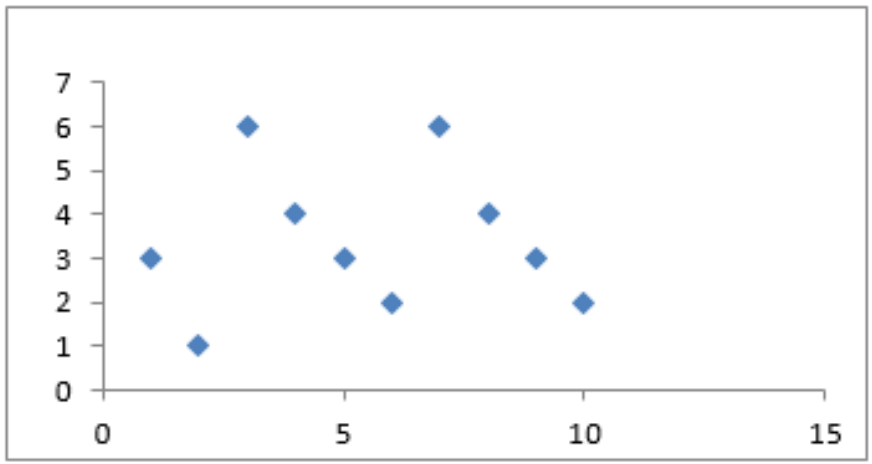
- Por esse motivo, dizemos que no conjunto A, existe **correlação positiva**, enquanto no conjunto B a correlação é uma **correlação negativa**.

Correlação Nula

Quando verificamos que, a medida que X aumenta, os valores de Y não apresentam um comportamento constante, em geral, podemos concluir que não existe correlação entre as variáveis, ou, em outras palavras, dizer que há uma **correlação nula**. Observe o conjunto de dados e o gráfico a seguir:

Conjunto C

X	Y	X	Y
1	3	6	2
2	1	7	6
3	6	8	4
4	4	9	3
5	3	10	2



Tipos de Correlação Linear

- Podemos dizer graficamente que a correlação linear existirá se for possível traçar uma "reta imaginária" que se aproxime dos dados. Essa reta, naturalmente, será crescente ou decrescente, de acordo com a disposição dos dados.

- Além disso, essa reta poderá nos dizer o *grau da correlação*, isto é, a "força" da correlação.

- Observe os exemplos a seguir.

Exemplo - Correlação Fraca

Conjunto *D*

X	Y
1	6
2	3
3	5
4	7
5	2
6	11
7	9
8	3
9	6
10	8

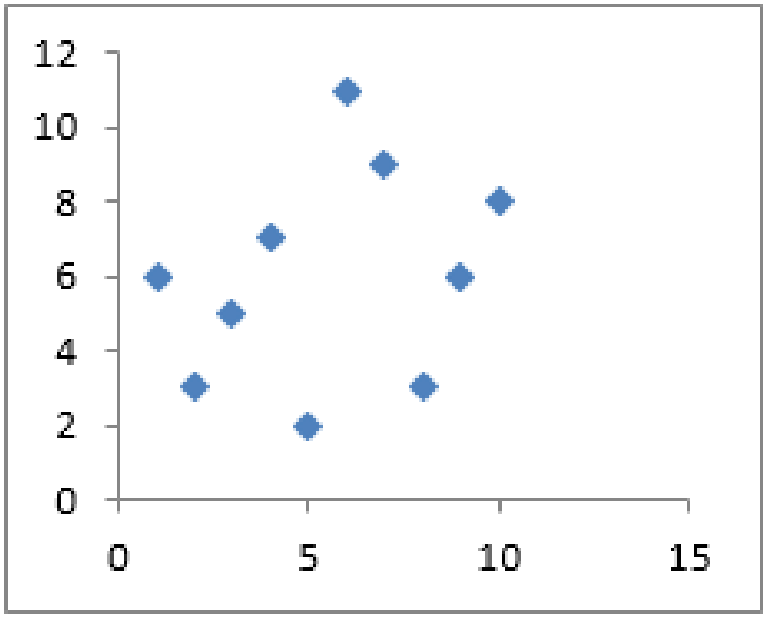


Figura: Diagrama de dispersão do Conjunto D

Exemplo - Correlação Forte

Conjunto E

X	Y
1	2
2	6
3	5
4	8
5	6
6	9
7	10
8	8
9	12
10	10

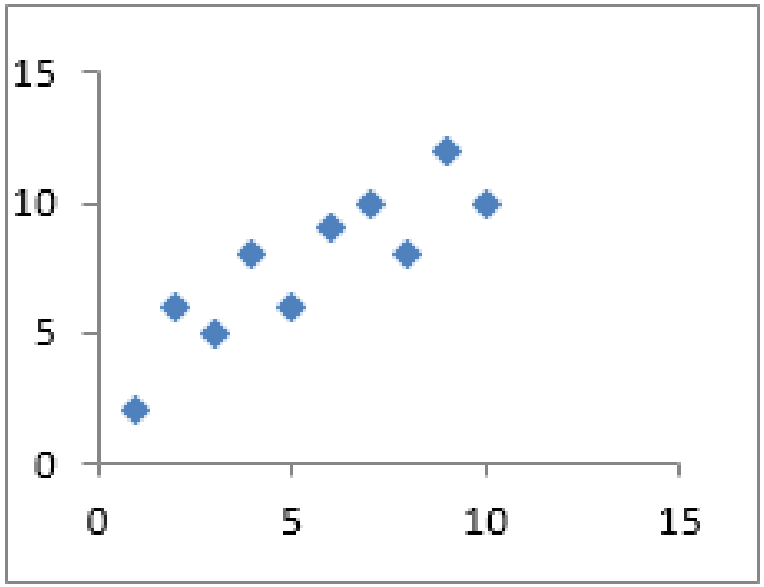


Figura: Diagrama de dispersão do Conjunto E

Exemplo - Correlação Perfeita

Conjunto F

X	Y
1	3
2	4
3	5
4	6
5	7
6	8
7	9
8	10
9	11
10	12

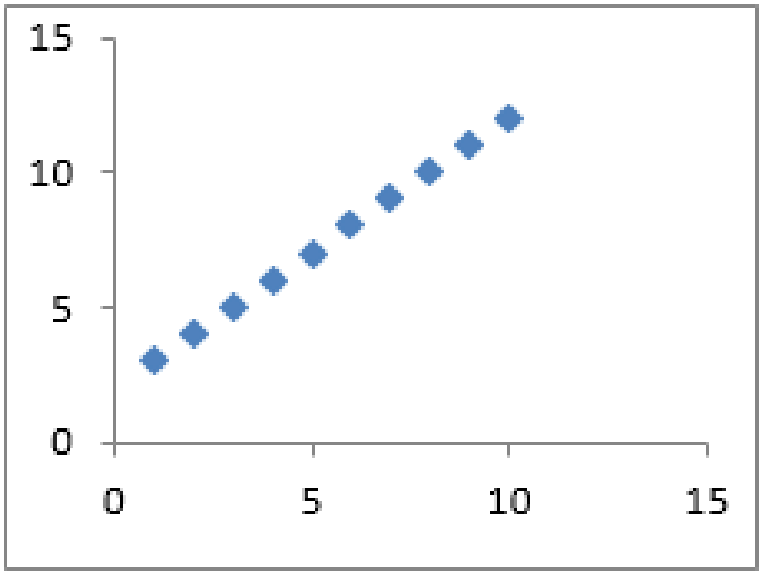
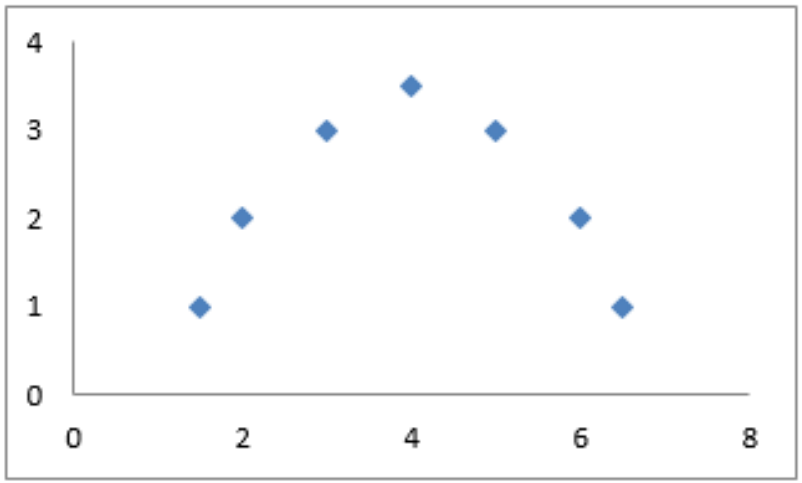


Figura: Diagrama de dispersão do Conjunto F

Correlação Não Linear

Ainda podemos, graficamente, verificar que existe algum tipo de relação entre os dados, porém, uma relação que não seja possível estabelecer a "reta imaginária" como verificamos anteriormente. Essas correlações são ditas *não lineares*. Observe o exemplo a seguir:

X	Y
1.5	1
2	2
3	3
4	3.5
5	3
6	2
6.5	1



Coeficiente de Correlação Linear

Como era de se esperar, a análise gráfica pelos diagramas de dispersão nos ajudam a desconfiar da existência ou não de correlação entre duas variáveis, porém, é insuficiente.

Para termos certeza dessa existência, calcularemos o *coeficiente de correlação linear* (r) entre as variáveis X e Y. Ele é obtido pela relação

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

onde S_X e S_Y são os desvios padrão das variáveis X e Y.

Coeficiente de Correlação - Outras definições

- Desenvolvendo a expressão apresentada anteriormente, podemos reescrever o coeficiente de correlação como:

$$r = \frac{\sum XY - \frac{(\sum X)(\sum Y)}{n}}{\sqrt{\left[\sum X^2 - \frac{(\sum X)^2}{n}\right]\left[\sum Y^2 - \frac{(\sum Y)^2}{n}\right]}}$$

- Também podemos usar a definição de *covariância* entre as duas variáveis S e Y, dada por

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

e definir a correlação entre X e Y como

$$corr(X, Y) = \frac{cov(X, Y)}{S_x S_y}.$$

Coeficiente de Correlação - Interpretação

Observe que $-1 \leq r \leq 1$. Se $r = 0$, indica que não há correlação. O sinal de r está naturalmente relacionado ao tipo de correlação (positiva ou negativa) dos dados.

Além disso, o valor de r indica o grau (força) da correlação:

Se $r = 0$: correlção nula

Se $-0,25 < r < 0$ ou $0 < r < 0,25$: correlação pequena

Se $-0,5 < r < -0,25$ ou $025 < r < 0,5$: correlação fraca

Se $-0,75 < r < -0,5$ ou $0,5 < r < 0,75$: correlação moderada

Se $-1 < r < -0,75$ ou $0,75 < r < 1$: correlação forte

Se $r = -1$ ou $r = 1$: correlação perfeita

Exemplo - Conjunto A

X	Y	X^2	Y^2	XY
1	2	1	4	2
2	0	4	0	0
3	6	9	36	18
4	3	16	9	12
5	9	25	81	45
6	4	36	16	24
7	10	49	100	70
8	8	64	64	64
9	12	81	144	108
10	8	100	64	80
$\sum X = 55$	$\sum Y = 62$	$\sum X^2 = 385$	$\sum Y^2 = 518$	$\sum XY = 423$

Exemplo - Conjunto A

$$r = \frac{\sum XY - \frac{(\sum X)(\sum Y)}{n}}{\sqrt{\left[\sum X^2 - \frac{(\sum X)^2}{n}\right]\left[\sum Y^2 - \frac{(\sum Y)^2}{n}\right]}}$$

$$= \frac{423 - \frac{(55)(62)}{10}}{\sqrt{\left[385 - \frac{(55)^2}{10}\right]\left[518 - \frac{(62)^2}{10}\right]}}$$

$$= \frac{423 - 341}{\sqrt{(82.5)(133.6)}}$$

$$= 0.781$$

Gráficos $q \times q$

- Outro tipo de representação gráfica que podemos utilizar para duas variáveis numéricas é o *gráfico quantis* $\times$ *quantis*, também conhecidos como *QQ-Plots*.

- Suponha que temos valores $x_1, \ldots, x_n$ da variável X e valores $y_1, \ldots, y_m$ da variável Y, todos medidos pela mesma unidade (note que não é necessário a mesma quantidade de dados, ou seja, n pode ser diferente de m).

- O gráfico $q \times q$ é um gráfico dos quantis de X contra os quantis de Y.
  - Se $n = m$, o gráfico compara os dados de cada vetor.
  - Se $n \neq m$, utiliza-se os dados do menor vetor e a mesma quantidade de quantis do outro vetor.

Gráficos $q \times q$

- Enquanto um 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, se valores intermediários de X estão relacionados com valores intermediários de Y e se valores grandes de X estão relacionados com valores grandes de Y.

- Num gráfico de dispersão podemos ter $x_1 < x_2$ e $y_1 > y_2$, o que não pode acontecer num gráfico q $\times$ q, pois os valores em ambos os eixos estão ordenados, do menor para o maior.

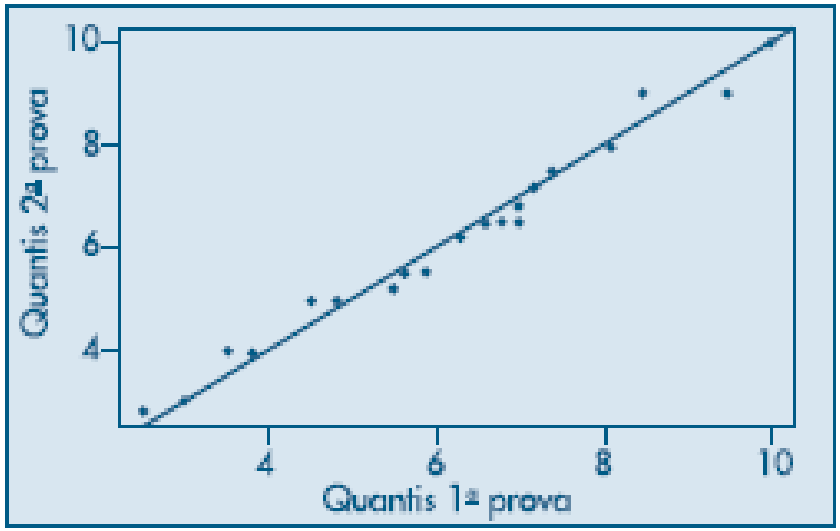
Exemplo - Gráfico $q \times q$

Considere as notas de 20 alunos em duas provas de Estatística, de acordo com a Tabela abaixo.

Aluno	Prova 1	Prova 2	Aluno	Prova 1	Prova 2
1	8.5	8.0	11	7.4	6.5
2	3.5	2.8	12	5.6	5.0
3	7.2	6.5	13	6.3	6.5
4	5.5	6.2	14	3.0	3.0
5	9.5	9,0	15	8.1	9.0
6	7.0	7.5	16	3.8	4,0
7	4.8	5.2	17	6.8	5.5
8	6.6	7.2	18	10.0	10.0
9	2.5	4.0	19	4.5	5.5
10	7.0	6.8	20	5.9	5.0

Exemplo - Gráfico $q \times q$

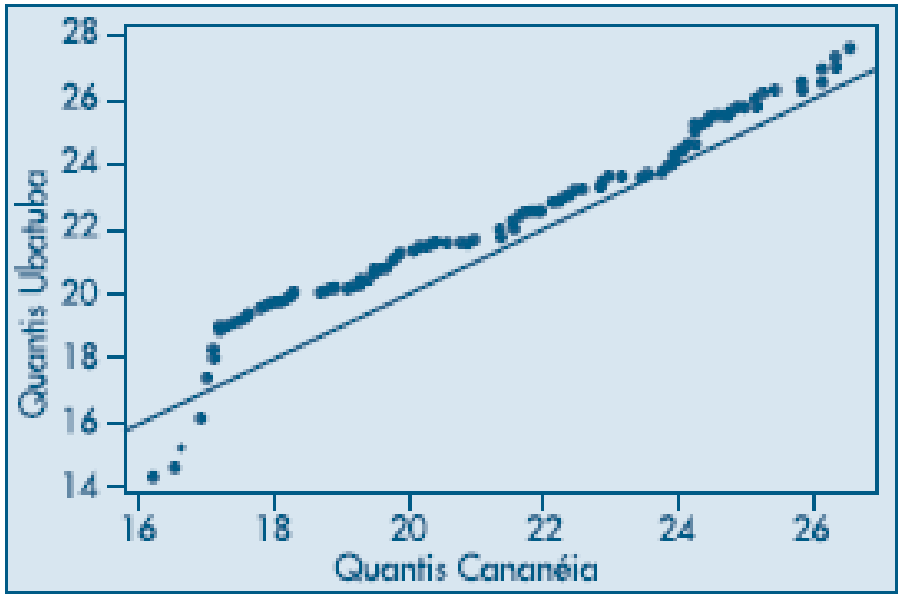
O gráfico q $\times$ q está plotado abaixo.



Podemos perceber que os pontos estão razoavelmente dispersos ao redor da reta $x = y$, mostrando que as notas dos alunos nas duas provas não são muito diferentes. Mas podemos notar que, para notas abaixo de cinco, os alunos tiveram notas maiores na segunda prova, ao passo que, para notas de cinco a oito, os alunos tiveram notas melhores na primeira prova. A maioria das notas estão concentradas entre cinco e oito.

Exemplo - Gráfico $q \times q$

Observe esse gráfico q $\times$ q que compara temperaturas de Ubatuba e Cananéia (cidades litorâneas de São Paulo).



Observamos que a maioria dos pontos está acima da reta $y = x$, mostrando que as temperaturas de Ubatuba são, em geral, maiores do que as de Cananéia, para valores maiores do que 17 graus

Associação entre variáveis qualitativas e quantitativas - Introdução

- Quando queremos verificar a existência de associação entre uma variável qualitativa e quantitativa, a análise é bastante simples: basta analisarmos o que acontece com a variável quantitativa quando os dados são categorizados de acordo com os diversos atributos da variável qualitativa.

- Essa análise pode ser conduzida por meio de medidas-resumo, histogramas ou *boxplots*.

- Claro, ainda é possível categorizar a variável quantitativa (em intervalos de classe) e realizar a análise como no primeiro caso (variáveis qualitativas).

Exemplo - Associação quali-quanti

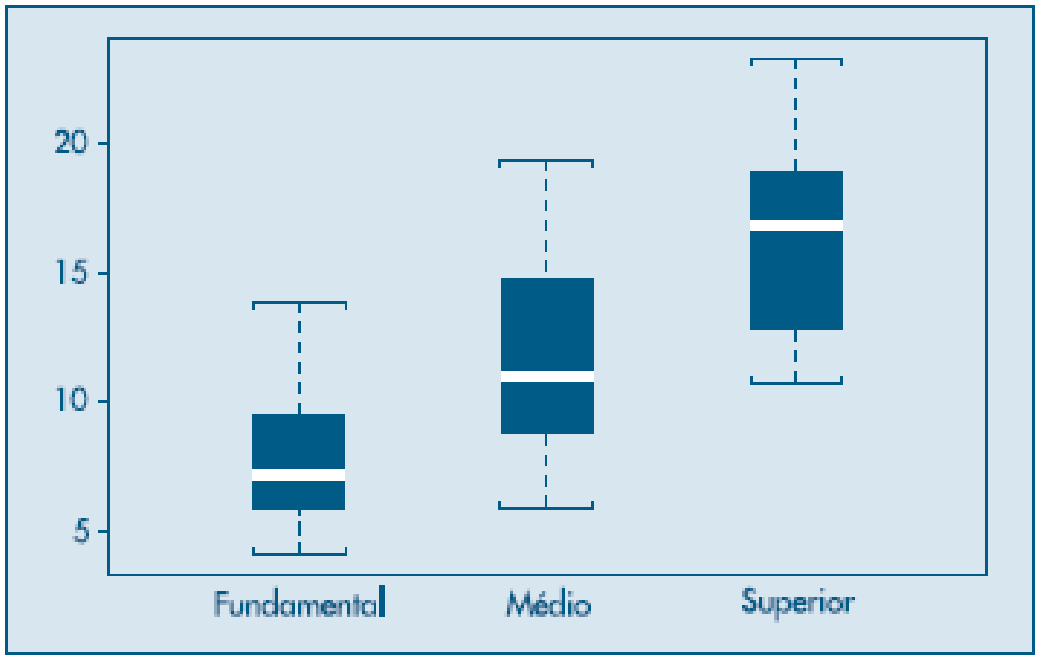
- Voltemos à Tabela inicial do curso, dos 36 funcionários da empresa. Agora desejamos analisar o comportamento dos salários dentro de cada categoria de grau de instrução.

- Como dito anteriormente, calcularemos as medidas resumo da variável numérica (salário - X) para os dados referentes a cada categoria da variável qualitativa (grau de instrução - Y). Os resultados estão dispostos na tabela a seguir:

GI.	n	$\bar{x}$	S_X	Var_X	x_{min}	q_1	q_2	q_3	x_{max}
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	4.00	7.55	10.17	14.06	23.30

Exemplo - Associação quali-quanti

Para auxiliar visualmente a análise de associação, podemos também construir boxplots comparativos:



A análise desses resultados sugere uma dependência dos salários em relação ao grau de instrução

Exemplo - Associação quali-quanti

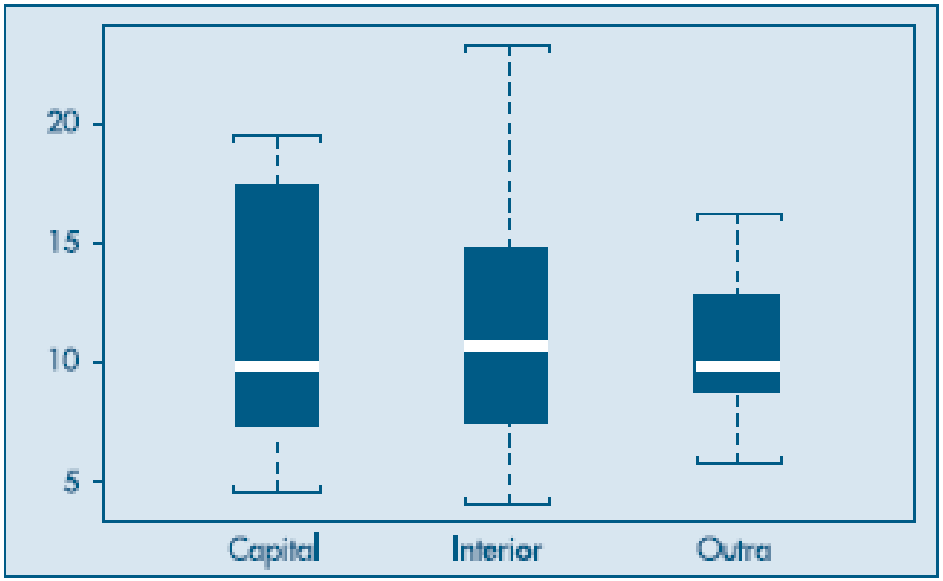
- Comparemos agora a variável quantitativa salário (X) com a variável qualitativa região de procedência (Y).

- O procedimento é o mesmo anterior: obter as medidas resumo por categoria da variável região de procedência:

Região	n	$\bar{x}$	S_X	Var_X	x_{min}	q_1	q_2	q_3	x_{max}
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

Exemplo - Associação quali-quanti

Boxplots comparativos:



A análise desses resultados mostra 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.

Medida de associação de variáveis quali-quanti

- Para construir a medida de associação, necessita-se de uma medida-resumo da variância entre as categorias da variável qualitativa. Essa medida pode ser uma média ponderada pelo número de observações em cada categoria:

$$\overline{Var(X)} = \frac{\sum_{i=1}^{k} n_i Var_i(X)}{\sum_{i=1}^{k} n_i},$$

onde k é o número de categorias e $var_i(X)$ denota a variância da variável numérica X dentro da categoria i, com $i = 1, \ldots, k$.

- O grau de associação entre as duas variáveis será dado por

$$R^2 = \frac{Var(X) - \overline{Var(X)}}{Var(X)} = 1 - \frac{\overline{Var(X)}}{Var(X)}.$$

Exemplo 01 - Quali *vs* Quali

Uma amostra de 200 habitantes de uma cidade foi escolhida para declarar sua opinião sobre um certo projeto governamental. O resultado foi o seguinte:

Opinião	Local de residência			Total
	Urbano	Suburbano	Rural	
A favor	30	35	35	100
Contra	60	25	15	100
Total	90	60	50	200

(a) Calcule as proporções em relação ao total das colunas.

(b) Você diria que a opinião independe do local de residência?

(c) Encontre uma medida de dependência entre as variações.

Exemplo 02 - Quanti *vs* Quanti

Muitas vezes a determinação da capacidade de produção instalada para certo tipo de indústria em certas regiões é um processo difícil e custoso. Como alternativa, pode-se estimar a capacidade de produção através da escolha de uma outra variável de medida mais fácil e que esteja linearmente relacionada com ela.
Suponha que foram observados os valores para as variáveis: capacidade de produção instalada (X, em toneladas), potência instalada (Y, em 1000 kW) e área construída (Z, em 100m). Com base num critério estatístico, qual das variáveis você escolheria para estimar a capacidade de produção instalada?

X	4	5	4	5	8	9	10	11	12	12
Y	1	1	2	3	3	5	5	6	6	6
Z	6	7	10	10	11	9	12	10	11	14

Exemplo 03 - Quanti *vs* Quanti: Dispersão e QQ-plot

Utilizar o banco de dados `salários` disponibilizado no *Moodle*.

O banco de dados trata de salários, em 1979 (em francos suíços), para quatro profissões (professor secundarista, mecânico, adminsitrador e engenheiro eletricista), em 30 cidades de diferentes países.
Fonte: "Prices and Salaries Around The World", 1979/1980. União dos Bancos Suíços, Zurique.

(a) Primeiro, verificar se existe correlação entre cada par de profissões.
(b) Depois, verificar entre cada par de profissões se a distribuição dos salários é relacionada.

Exemplo 04 - Quali *vs* Quanti

Utilizar o banco de dados `veiculos` disponibilizado no *Moodle*.

O banco de dados trata de informações sobre 30 veículos novos, nacionais (N) e importados (I) em março de 1999. As variáveis são Preço em dólares, comprimento em metros e motor em CV.
Fonte: Folha de S. Paulo, 14/3/1999

(a) Primeiro, verificar se existe correlação entre as variáveis numéricas.

(b) Depois, dividir o banco de dados em termos da variável categórica (Origem), obter as medidas resumo e plotar os boxplots comparativos para cada variável numérica.

(c) Calcular o coeficiente R^2 para a variação de cada variável numérica e interpretar a associação dessas variáveis com a Origem dos veículos.

Até a Próxima!!!