Universidade Federal de São Paulo

Boxplot e Medidas de associação entre variáveis

Professor

Julio Cezar

Aula de hoje

- Construção do gráfico Box Plot;
- Discusão sobre outliers;
- Discusão sobre as formas da distribuição dos dados;
- Medidas de associação entre variáveis.

Box plot

- Aprender como é construído um box plot;
- Compreender como se interpreta um box plot;
- Comparar grupos através de box plots.

UNIFESP
Universidade Federal de São Paulo
1933

BOX PLOT

- O box plot, ou diagrama de caixa, é um gráfico que capta importantes aspectos de um conjunto de dados através do seu resumo dos seguintes valores: valor mínimo, primeiro quartil, segundo quartil, terceiro quartil e valor máximo.

- O conjunto destas medidas fornece evidência acerca da posição, dispersão, assimetria e valores extremos (atípicos).

- De modo geral, um ponto será considerado atípico quando estiver fora do intervalo denotado por $(LI; LS)$, onde

$$LI = \mathbf{q_1 - (1{,}5) \times AIQ}, \quad LS = \mathbf{q_3 + (1{,}5) \times AIQ} \quad \text{e} \quad \mathbf{AIQ = q_3 - q_1}.$$

Obs.: Amplitude interquartil (AIQ)= q_3 - q_1

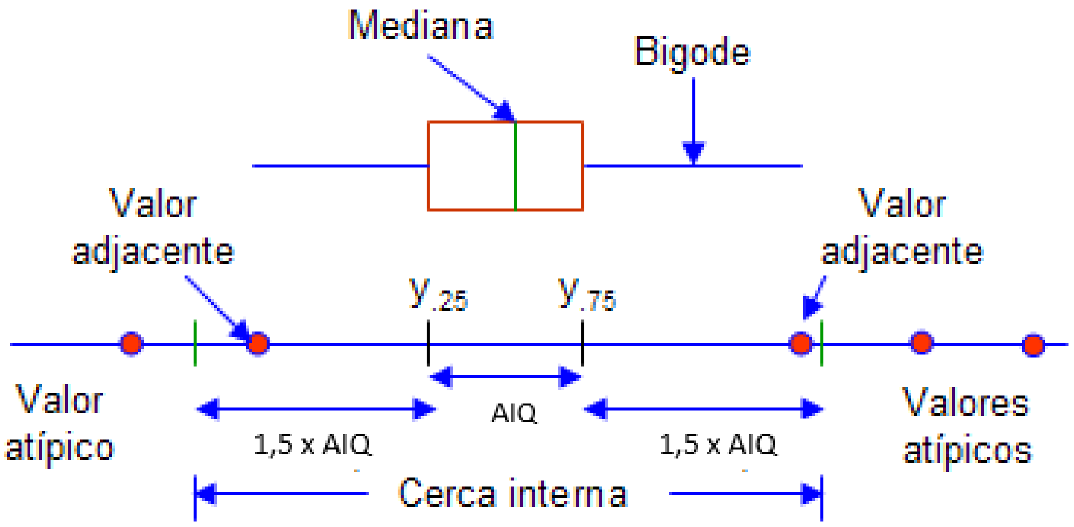
UNIFESP

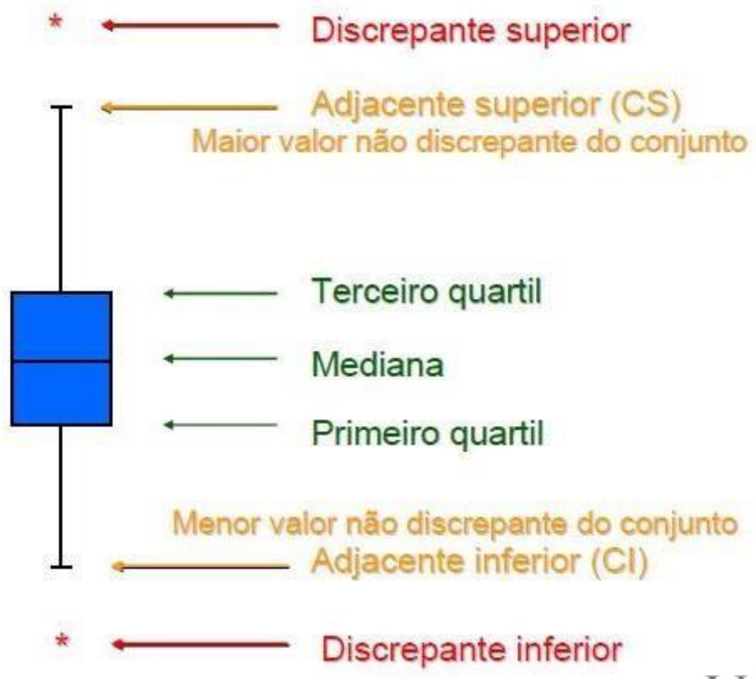
Universidade Federal de São Paulo

1933

Box plot

O objetivo deste gráfico é fornecer informações **sobre a variabilidade dos dados e valores atípicos** que podem influenciar o cálculo de medidas como a média aritmética, por exemplo.





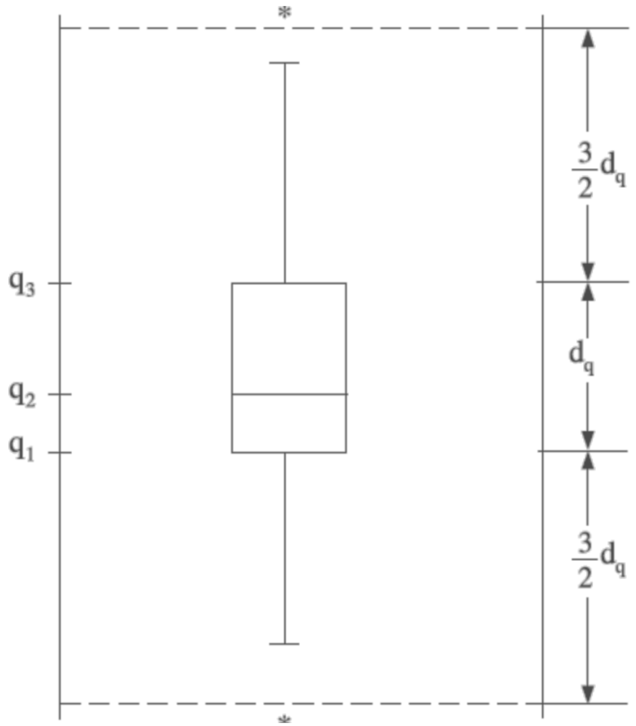
UNIFESP
Universidade Federal de São Paulo
1933

Box plot

Representa os dados em um retângulo construído com base nos quartis e fornece informações sobre valores extremos (atípicos ou outliers).

Box plot

Representa os dados em um retângulo construído com base nos quartis e fornece informações sobre valores extremos (atípicos ou outliers).

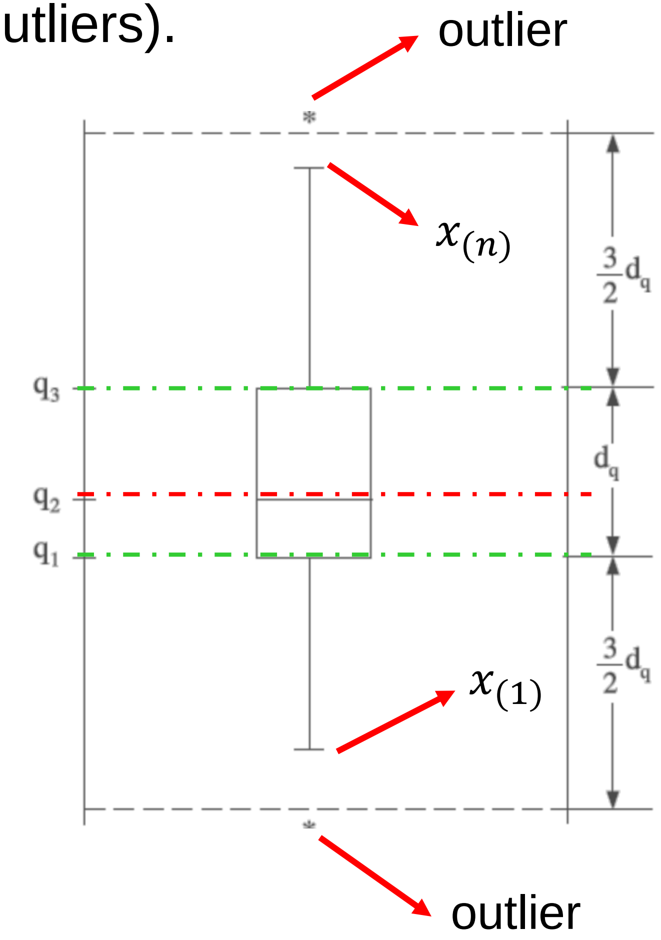


Box plot

Representa os dados em um retângulo construído com base nos quartis e fornece informações sobre valores extremos (atípicos ou outliers).

$$q_3 = q(0,75)$$

$$q_1 = q(0,25)$$



$$LS = q_3 + 1,5d_q$$

LS : limite superior

$$LI = q_1 - 1,5d_q$$

LS : limite inferior

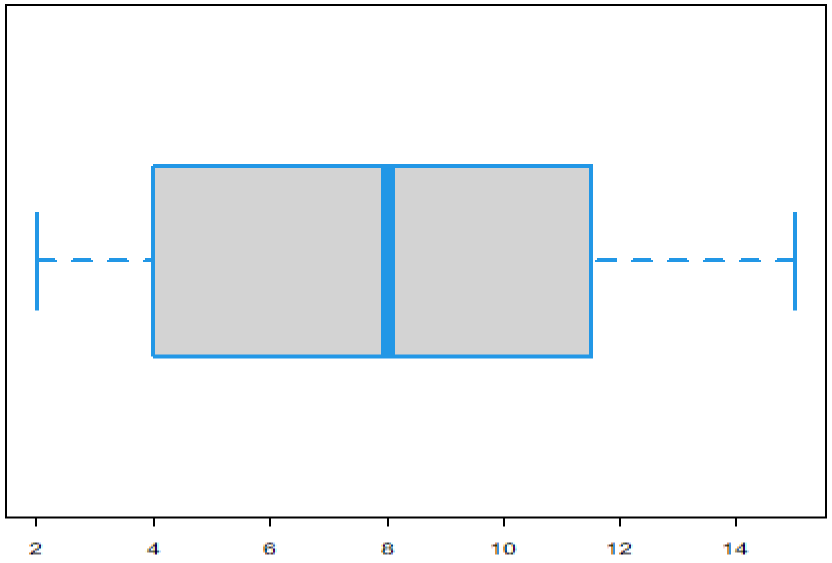
Exemplo

$x_{(1)} = 2, \boxed{x_{(2)} = 3, x_{(3)} = 5}, x_{(4)} = 7, \boxed{x_{(5)} = 8}, x_{(6)} = 10, \boxed{x_{(7)} = 11, x_{(8)} = 12}, x_{(9)} = 15$

1º Parte *2º Parte* *3º Parte* *4º Parte*

$q_1 = \mathbf{4}$ $Md = \mathbf{8}$ $q_3 = \mathbf{11,5}$

$\text{AIQ} = 11{,}5 - 4 = \mathbf{7,5}$ $LI = 4 - 1{,}5 \text{ x } (7{,}5) = \mathbf{-7,25}$ $LS = 11{,}5 + 1{,}5 \text{ x } (7{,}5) = \mathbf{22,75}$



UNIFESP
UNIVERSIDADE FEDERAL DE SÃO PAULO
1933

Exemplo

Exemplo: Tempo de vida útil de máquinas. Construa um boxplot.

18	21	21	23	23	25
27	29	30	31	32	32
32	34	35	36	38	41
42	42	43	44	45	46
46	47	48	50	54	56
57	58	60	61	98	116

UNIFESP
UNIVERSIDADE FEDERAL DE SÃO PAULO
1933

EXEMPLO

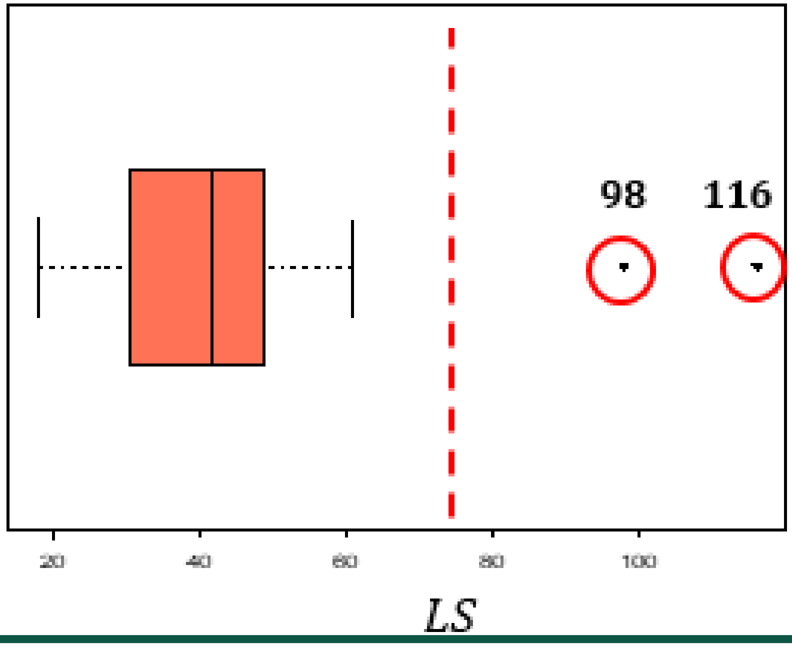
Exemplo: Tempo de vida útil de máquinas. Construa um boxplot.

18	21	21	23	23	25
27	29	30	31	32	32
32	34	35	36	38	41
42	42	43	44	45	46
46	47	48	50	54	56
57	58	60	61	98	116

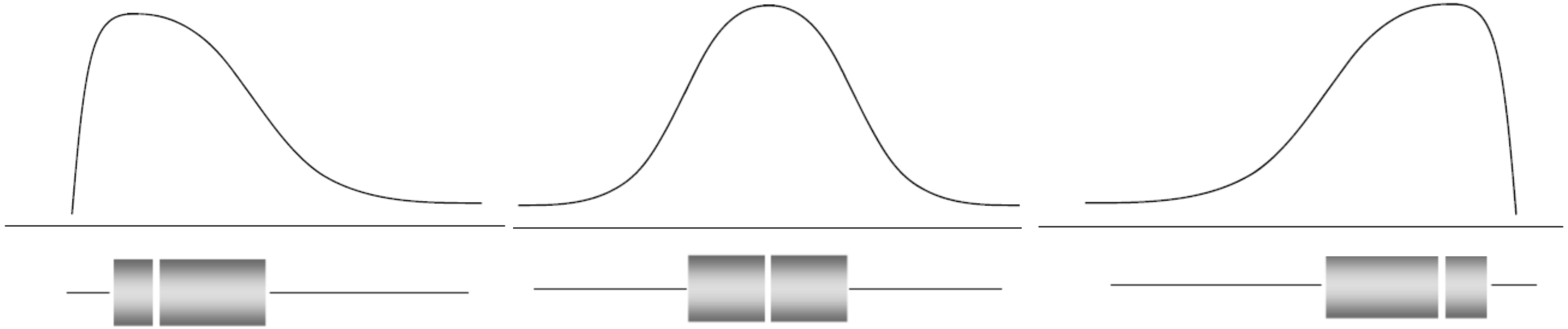
$x_{(1)} = \mathbf{18} \quad x^*_{(34)} = \mathbf{61} \quad Md = \mathbf{41,50}$

$q(0{,}25) = \mathbf{30,5} \quad q(0{,}75) = \mathbf{49} \quad d_q = 49 - 30{,}5 = \mathbf{18,5}$

$LI = 30{,}5 - 1{,}5(18{,}5) = \mathbf{2,75} \quad LS = 49 + 1{,}5(18{,}5) = \mathbf{76,75}$



Formas da distribuição



Os cinco valores $x_{(1)}, q_1$, q_2, q_3 e $x_{(n)}$, são importantes para se ter uma boa ideia da assimetria da distribuição dos dados. Para uma distribuição simétrica ou aproximadamente simétrica, temos

$$q_2 - x_{(1)} \cong x_{(n)} - q_2;$$
$$q_2 - q_1 \cong q_3 - q_2;$$
$$q_1 - x_{(1)} \cong x_{(n)} - q_3;$$

E distâncias entre mediana e q_1, q_3 menores do que a distância entre os extremos e q_1, q_3.

Variável padronizada: o valor Z

- Diferença entre um valor e a média, dividida pelo desvio-padrão;

$$Z = \frac{X - \overline{X}}{S}$$

- Medida de distância da média (por exemplo, um valor Z igual a 2 significa que o valor está a 2 desvios da média);
- Um valor Z acima de 3 ou abaixo de -3 é considerado um valor extremo (outlier). Esse é um dos **critérios para encontrar outlier**, mas não significa que o valor seja um erro ou que não deveria fazer parte dos dados. Significa que deve ser examinado. Usar medidas baseadas em ordenamento podem ser mais adequadas para dados com muitos valores extremos.

UNIFESP
UNIVERSIDADE FEDERAL DE SÃO PAULO
1933

Exemplo

- Se a média é 14 e o desvio-padrão é 3, qual é o valor Z para o valor 18,5?

$$Z = \frac{X - \overline{X}}{S} = \frac{18{,}5 - 14}{3} = 1{,}5$$

- O valor 18,5 está a 1,5 desvio-padrão acima da média;

- Um valor Z negativo significa que o valor é menor que a média.

MEDIDAS NUMÉRICAS DESCRITIVAS PARA UMA POPULAÇÃO

- Medidas numéricas para uma população são chamadas **parâmetros;**
- A média da população é a soma dos valores que compõem a população, dividida pelo tamanho da população (**N**)

$$\mu = \frac{\sum_{i=1}^{N} X_i}{N} = \frac{X_1 + X_2 + \cdots + X_N}{N}$$

onde

μ = média da população

N = tamanho da população

X_i = *i*-ésimo valor de X

Variância da População

Mede a dispersão em torno da média

Variância da População:

$$\sigma^2 = \frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N}$$

onde

μ = média da população

N = tamanho da população

X_i = *i*-ésimo valor de X

UNIFESP
Universidade Federal de São Paulo
1933

Desvio-Padrão da População

- Medida de variação mais utilizada;
- Mostra a variação em torno da média;
- Raiz quadrada da variância;
- Tem a mesma unidade dos dados originais.

Desvio-Padrão da população:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N}}$$

MEDIDAS DE ASSOCIAÇÃO ENTRE VARIÁVEIS

As medidas de associação entre:

- Duas variáveis qualitativas → Qui-Quadrado de Pearson (χ^2) e T;
- Duas variáveis quantitativas → r de Pearson;
- Variáveis qualitativa e quantitativa → Coeficiente de determinação (R^2).

UNIFESP
UNIVERSIDADE FEDERAL DE SÃO PAULO
1933

Motivação

Existe uma relação entre a altura de pessoas e o sexo em dada comunidade?

P1: Qual a frequência esperada de uma pessoa dessa população ter mais de 170 cm?

P2: Qual a frequência esperada de uma mulher (ou homem) ter mais de 170 cm?

- Se a frequência esperada é a mesma: não há associação entre as variáveis altura e sexo.
- Caso contrário, existe uma provável associação.

UNIFESP
Universidade Federal de São Paulo
1933

Associação entre variáveis

Quando consideramos duas variáveis (ou dois conjuntos de dados), podemos ter 3 situações e as técnicas de análise são diferentes.

(a) As duas **qualitativas** (**tabela de contingência**).
(b) As duas **quantitativas** (**gráficos de dispersão**).
(c) **Qualitativa** e **quantitativa** (**tabela de contingência**).

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 dependência entre duas variáveis. Esses coeficientes geralmente variam de 0 a 1 ou -1 a +1, e a proximidade do zero indica falta de associação.

ASSOCIAÇÃO ENTRE VARIÁVEIS

- Quando estamos interessados no comportamento conjunto de duas variáveis, os dados podem ser resumidos em **tabelas de dupla entrada** (ou **contingência**).

Exemplo 1: Uma pesquisa é feita entre alunos do primeiro ano da faculdade e perguntou-se aos alunos se *trabalhavam* **(variável X)** e o *número de vestibulares prestados* **(variável Y).**

Neste caso, cada elemento do corpo da tabela dá a frequência observada das realizações simultâneas das duas variáveis.

UNIFESP
UNIVERSIDADE FEDERAL DE SÃO PAULO
1933

Tabela de frequências conjunta

Tabela 1: Frequências absolutas conjunta das variáveis X (trabalhavam) e Y (n° de vestibulares prestados:

Tabela de frequência marginal de X

X\Y	1	2	3	Total
sim	4	2	2	8
não	5	6	1	12
Total	9	8	3	20

Tabela de frequência marginal de Y

TABELA DE FREQUÊNCIAS CONJUNTA

- Tabelas de frequências marginal ou individual

X	freq
sim	8
não	12
Total	20

Y	1	2	3	Total
freq	9	8	3	20

ASSOCIAÇÃO ENTRE VARIÁVEIS QUALITATIVAS

Exemplo 2: Deseja-se determinar se a criação de determinado tipo de cooperativa está associada a um fator regional:

Tabela 2: Frequências absolutas por Tipo de Cooperativa (X) e Estado (Y).

Estado	Tipo de Cooperativa				
	Consumidor	Produtor	Escola	Outros	Total
SP	214	237	78	119	648
PR	51	102	126	22	301
RS	111	304	139	48	602
Total	376	643	343	189	1551

ASSOCIAÇÃO ENTRE VARIÁVEIS QUALITATIVAS

Exemplo 2: Deseja-se determinar se a criação de determinado tipo de cooperativa está associada a um fator regional:

As **porcentagens** correspondem à distribuição do **Total Geral** (=1551) por Tipo de Cooperativa

Tabela 2: Frequências absolutas por Tipo de Cooperativa (X) e Estado (Y).

Estado	Tipo de Cooperativa				
	Consumidor	Produtor	Escola	Outros	Total
SP	214	237	78	119	648
PR	51	102	126	22	301
RS	111	304	139	48	602
Total	376	643	343	189	1551
	24,%	42%	22%	12%	100%

UNIFESP
UNIVERSIDADE FEDERAL DE SÃO PAULO
1933

ASSOCIAÇÃO ENTRE VARIÁVEIS QUALITATIVAS

Exemplo 2: Deseja-se determinar se a criação de determinado tipo de cooperativa está associada a um fator regional:

As **porcentagens** correspondem à distribuição do Total de cada Estado (**TOTAL POR LINHA**) por Tipo de Cooperativa

Tabela 2: Frequências absolutas (relativas) por Tipo de Cooperativa (X) e Estado (Y).

Estado	Tipo de Cooperativa				
	Consumidor	Produtor	Escola	Outros	Total
SP	214	237	78	119	648
	33%	**37%**	**12%**	**18%**	**100%**
PR	51	102	126	22	301
	17%	**34%**	**42%**	**7%**	**100%**
RS	111	304	139	48	602
	18%	**51%**	**23%**	**8%**	**100%**
Total	376	643	343	189	1551
	24%	42%	22%	12%	100%

UNIFESP
UNIVERSIDADE FEDERAL DE SÃO PAULO
1933

Associação entre variáveis qualitativas

Exemplo 2: Deseja-se determinar se a criação de determinado tipo de cooperativa está associada a um fator regional:

Tabela 2: Frequências absolutas (relativas) por Tipo de Cooperativa (X) e Estado (Y).

Estado	Tipo de Cooperativa				
	Consumidor	Produtor	Escola	Outros	Total
SP	214	237	78	119	648
	33%	37%	12%	18%	100%
PR	51	102	126	22	301
	17%	34%	42%	7%	100%
RS	111	304	139	48	602
	18%	51%	23%	8%	100%
Total	376	643	343	189	1551
	24%	42%	22%	12%	100%

Há INDÍCIOS de associação entre as variáveis de interesse Estado e Tipo de Cooperativa com base nesta análise feita sobre a distribuição de frequências por estado

ASSOCIAÇÃO ENTRE VARIÁVEIS QUALITATIVAS

Exemplo 2: Deseja-se determinar se a criação de determinado tipo de cooperativa está associada a um fator regional:

Tabela 2: Frequências absolutas (relativas) por Tipo de Cooperativa (X) e Estado (Y).

Estado	Tipo de Cooperativa				
	Consumidor	Produtor	Escola	Outros	Total
SP	214	237	78	119	648
	33%	37%	12%	18%	100%
PR	51	[illegible]	[illegible]	[illegible]	301
	17%	[illegible]	[illegible]	[illegible]	00%
RS	111	[illegible]	[illegible]	[illegible]	602
	18%	51%	23%	8%	100%
Total	376	643	343	189	1551
	24%	42%	22%	12%	100%

Mas como mensurar essa associação??

Há INDÍCIOS de associação entre as variáveis de interesse Estado e Tipo de Cooperativa com base nesta análise feita sobre a distribuição de frequências por estado

UNIFESP
UNIVERSIDADE FEDERAL DE SÃO PAULO
1933

Associação entre variáveis qualitativas

Se não houvesse **associação** (dependência), esperaríamos que em cada estado tivesse 24% de cooperativas de consumidores, 42% de cooperativas de produtores, 22% de escolas e 12% outros tipos.

Assim, o número esperado de cooperativas de consumidores em SP seria **648*0,24 = 157**, e assim por diante.

Tabela 3: Frequências esperadas, assumindo independência entre as 2 variáveis.

Estado	Tipo de Cooperativa								
	Consumidor		Produtor		Escola		Outros		Total
SP	157	24%	269	42%	143	22%	79	12%	648
PR	73	24%	125	42%	67	22%	37	12%	301
RS	146	24%	250	42%	133	22%	73	12%	602
Total	376	24%	643	42%	343	22%	189	12%	1551

UNIFESP
Universidade Federal de São Paulo
1933

Associação entre variáveis qualitativas

Comparando as duas tabelas, podemos verificar a discrepância existente entre os valores observados e os valores esperados, caso as variáveis forem independentes.

Tabela 4: **Desvios** entre frequências observadas e **esperadas**.

Estado	**Tipo de Cooperativa**				
	Consumidor	**Produtor**	**Escola**	**Outros**	**Total**
SP	57	-32	-65	40	0
PR	-22	-23	59	-15	0
RS	-35	54	6	-25	0
Total	0	0	0	0	0

QUI-QUADRADO DE PEARSON

Para comparar os desvios é interessante padronizá-los e transformá-los em positivos. E, então, obter o **coeficiente de contingência**:

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

onde o_{ij} são as frequências observadas da *i*-ésima categoria de **X** e *j*-ésima categoria de **Y** e e_{ij} são as frequências esperadas.

Um valor grande de χ^2 indica associação entre as variáveis. Como interpretar quão grande?!!

UNIFESP
UNIVERSIDADE FEDERAL DE SÃO PAULO
1933

QUI-QUADRADO DE PEARSON PARA O EXEMPLO 1

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

Estado
i = 1, 2, 3

r=3 linhas

Cooperativa
j = 1, 2, 3, 4

s=4 colunas

o_{ij}

Tabela 2: Frequências **observadas** por Tipo de Cooperativa e Estado.

Estado (i = 1, 2, 3)	Tipo de Cooperativa (j = 1, 2, 3, 4) Consumidor		Produtor		Escola		Outros		Total	
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%**	48	**8%**	602	**100%**
Total	376	24%	643	42%	343	22%	189	12%	1551	100%

o_{23}

e_{ij}

Tabela 3: Frequências **esperadas**, assumindo independência entre as 2 variáveis.

Estado (i = 1, 2, 3)	Tipo de Cooperativa (j = 1, 2, 3, 4) **Consumidor**		**Produtor**		**Escola**		**Outros**		**Total**
SP	**157**	**24%**	**269**	**42%**	**143**	**22%**	**79**	**12%**	**648**
PR	**73**	**24%**	**125**	**42%**	**67**	**22%**	**37**	**12%**	
RS	**146**	**24%**	**250**	**42%**	**133**	**22%**	**73**	**12%**	
Total	**376**	**24%**	**643**	**42%**	**343**	**22%**	**189**	**12%**	

e_{23}

UNIFESP
UNIVERSIDADE FEDERAL DE SÃO PAULO
1933

Qui-Quadrado de Pearson para o Exemplo 1

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{s}\frac{(o_{ij}-e_{ij})^2}{e_{ij}} = \frac{(214-157)^2}{157} + \frac{(237-269)^2}{269} + \ldots + \frac{(48-73)^2}{73} = 171{,}75.$$

Para facilitar a interpretação da associação definiu-se o coeficiente de contingência corrigido, que assume valores entre 0 e 1:

$$T = \sqrt{\frac{\frac{\chi^2}{n}}{(r-1)(s-1)}} = \sqrt{\frac{\frac{171{,}75}{1551}}{(3-1)(4-1)}} = 0{,}136$$

Quanto mais próximo de 1 maior é associação entre a criação de cooperativas e algum fator regional. Como o valor de T = 0,136 (bem próximo de 0) conclui-se que não há associação entre os estados e tipo de cooperativas.

UNIFESP
Universidade Federal de São Paulo
1933

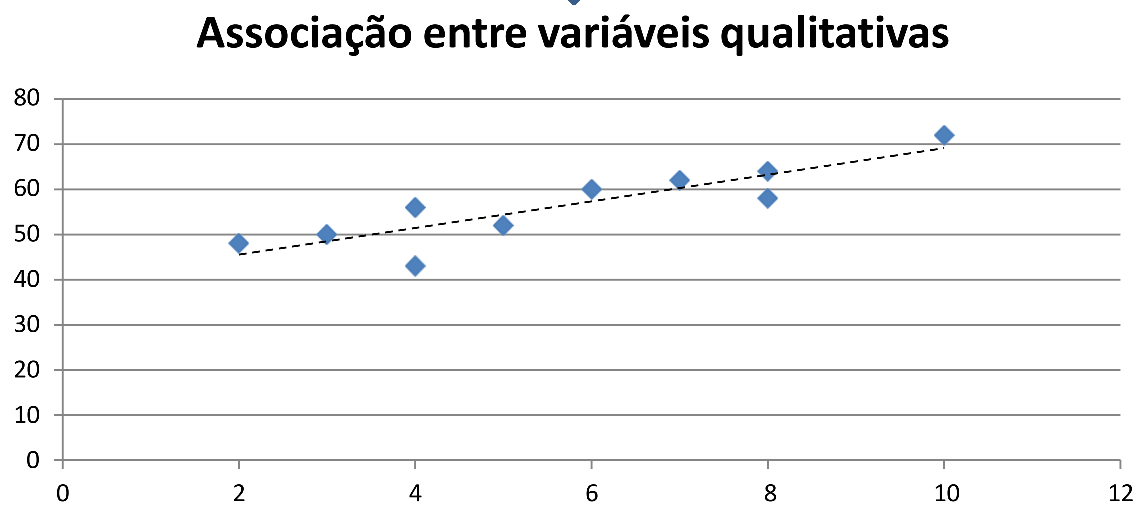
Associação entre variáveis quantitativas

Exemplo 2: Existe associação entre o número de clientes e o tempo de serviço de agentes de uma companhia de seguros?

Ind.	Anos de serviço (X)	N. 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

Uma forma bastante útil de verificar a associação entre variáveis quantitativas é

pelo gráfico de dispersão.



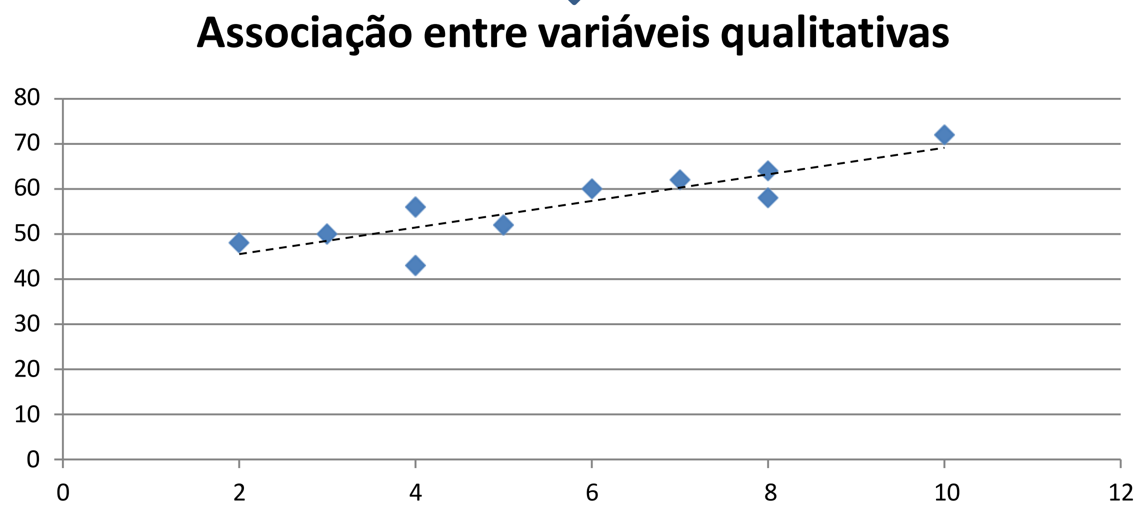
ASSOCIAÇÃO ENTRE VARIÁVEIS QUANTITATIVAS

Exemplo 2: Existe associação entre o número de clientes e o tempo de serviço de agentes de uma companhia de seguros?

Ind.	Anos de serviço (X)	N. 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

Uma forma bastante útil de verificar a associação entre variáveis quantitativas é

pelo gráfico de dispersão.



Resp: Parece que sim, pois à medida que aumenta o tempo de serviço, o número de clientes também aumenta.

UNIFESP
UNIVERSIDADE FEDERAL DE SÃO PAULO
1933

CORRELAÇÃO E COVARIÂNCIA

- A medida que se utiliza com mais frequência para quantificar o grau de uma **associação linear**, é o **coeficiente de correlação de Pearson (r).**

  Esta medida avalia o quanto a nuvem de pontos do gráfico de dispersão se aproxima de uma reta.

- O coeficiente de correlação de Pearson (r), também chamado de correlação linear ou r de Pearson, é um grau de relação entre duas variáveis quantitativas.

- Na definição do coeficiente de correlação de pares de variáveis, está implícita a definição de uma medida que dá uma ideia da ***variabilidade conjunta*** existente entre as variáveis e que é a **covariância amostral.**

UNIFESP
UNIVERSIDADE FEDERAL DE SÃO PAULO
1933

Coeficiente de correlação de Pearson

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

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

$$= \sum_{i=1}^{n}\frac{z_x z_y}{n}$$

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

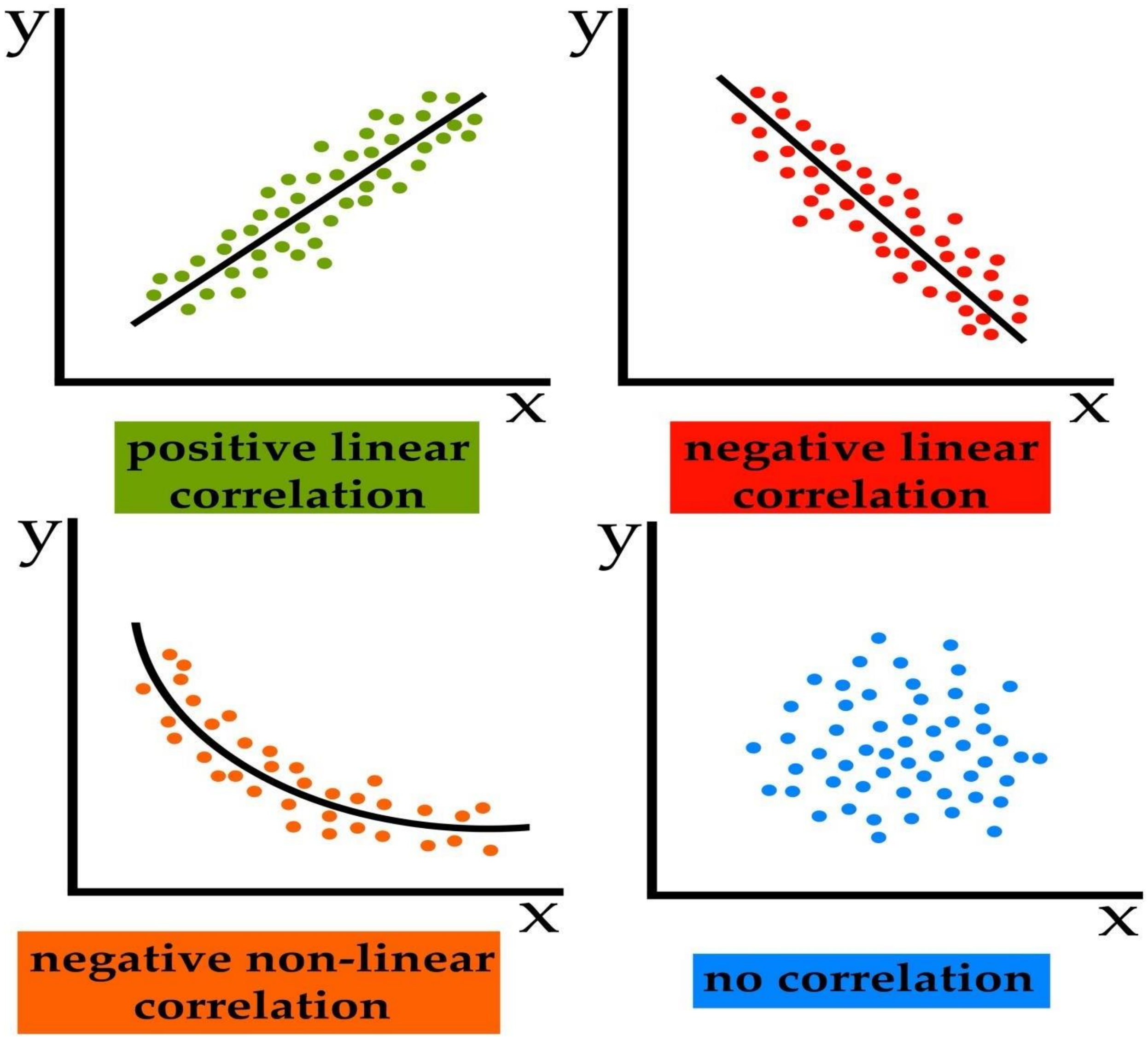
O coeficiente de correlação linear satisfaz:

$$-1 \leq r \leq 1$$

→ Quanto mais próximo dos extremos, -1 ou 1, maior é o grau de associação entre as variáveis de interesse.

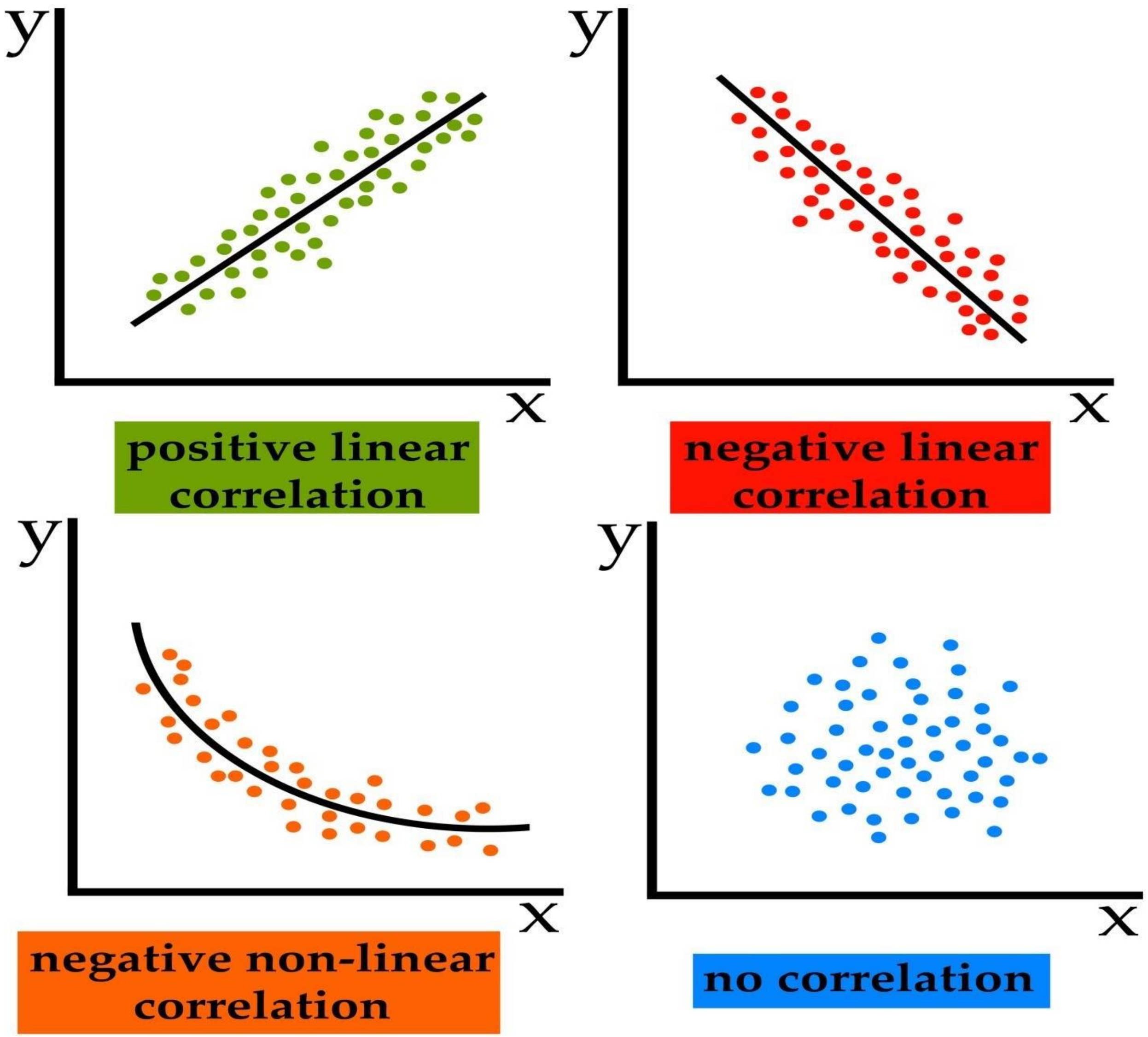
UNIFESP
UNIVERSIDADE FEDERAL DE SÃO PAULO
1933

ASSOCIAÇÕES ENTRE 2 VARIÁVEIS QUANTITATIVAS



UNIFESP
UNIVERSIDADE FEDERAL DE SÃO PAULO
1933

ASSOCIAÇÕES ENTRE 2 VARIÁVEIS QUANTITATIVAS



- *<u>Correlação positiva</u>*:

  corr(X,Y) ≈ 1

- *<u>Correlação negativa</u>*

  ou inversa:

  corr(X<Y) ≈-1

- *<u>Não há correlação:</u>*

  corr(X,Y) ≈ 0.

- *<u>Correlação moderada:</u>*

  |corr(X,Y)| ≈ 0,5.

UNIFESP

UNIVERSIDADE FEDERAL DE SÃO PAULO

1933

ASSOCIAÇÕES ENTRE 2 VARIÁVEIS QUANTITATIVAS

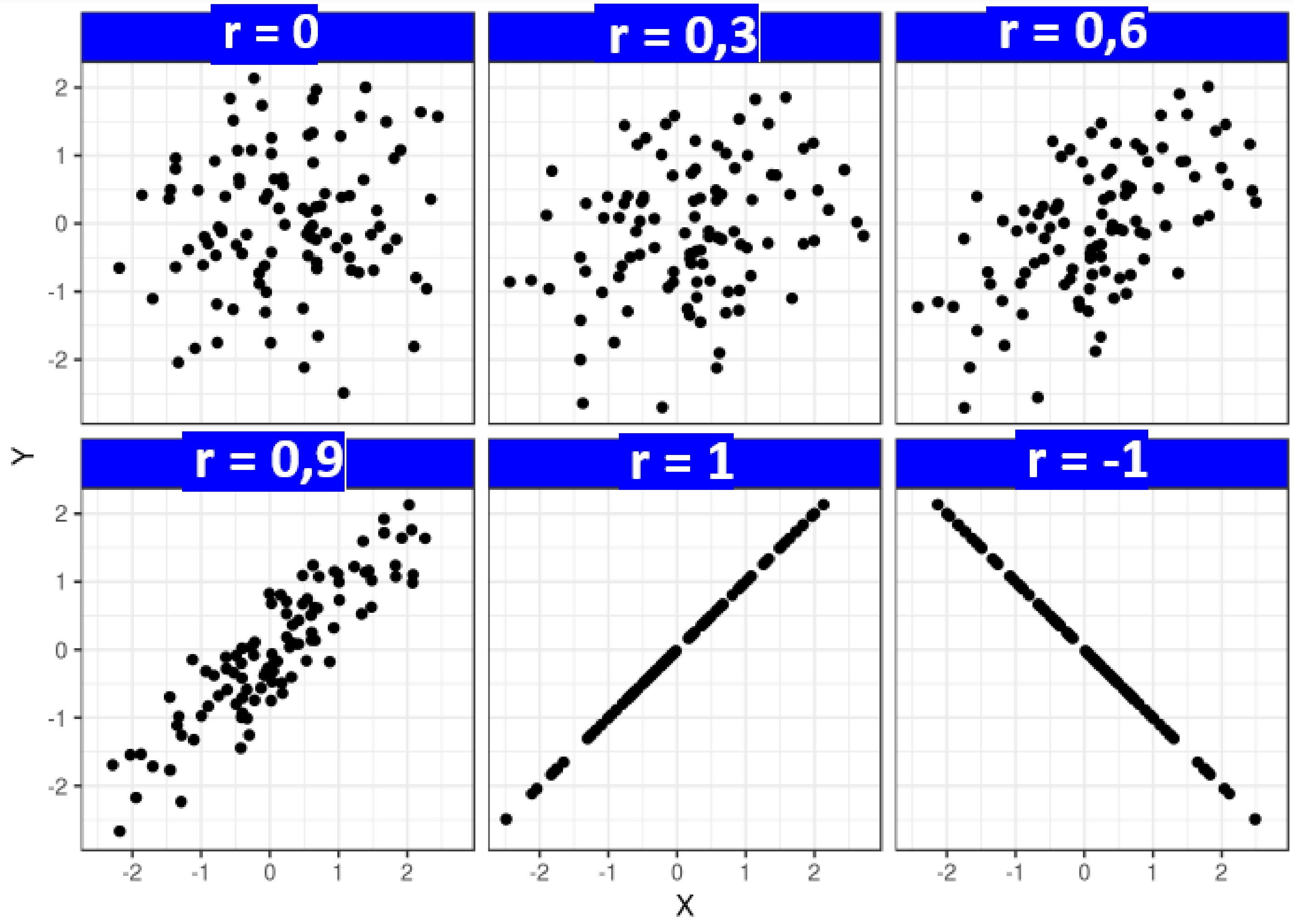
Coeficiente de correlação de Pearson (r)

- Se $r = 0$ correlação nula;
- Se $0 < |r| \leq 0{,}3$ correlação fraca;
- Se $0{,}3 < |r| \leq 0{,}6$ correlação moderada;
- Se $0{,}6 < |r| \leq 0{,}9$ correlação forte;
- Se $0{,}9 < |r| < 1$ correlação muito forte;
- Se $|r| = 1$ correlação perfeita.

UNIFESP
UNIVERSIDADE FEDERAL DE SÃO PAULO
1933

Associações entre 2 variáveis quantitativas

Coeficiente de correlação de Pearson (r)



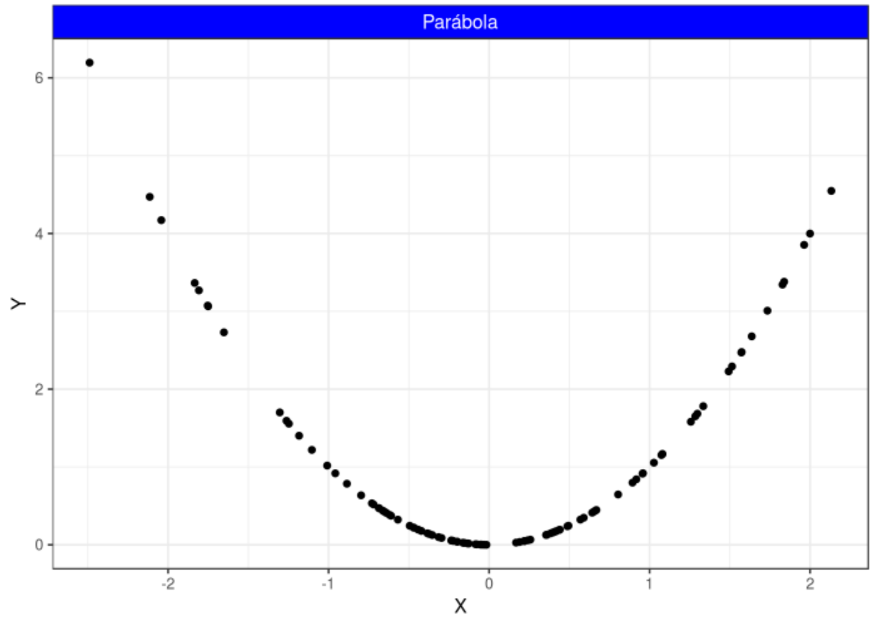
CORRELAÇÃO NÃO IMPLICA NECESSARIAMENTE CAUSALIDADE

Ao encontrarmos uma correlação entre eventos, buscamos estabelecer uma relação de causalidade entre eles.

No entanto, se duas variáveis têm correlação não nula, não podemos já inferir que uma causa a outra!

Correlação não implica necessariamente causalidade

Nota-se claramente que existe uma relação entre as variáveis Y e X, só que a relação não parece com o formato de uma reta, mas sim, com o formato de uma parábola, e assim, o coeficiente de correlação não é indicado para medir essa correlação, e por curiosidade, a correlação resultante foi igual a 0. Para isso, existem outros procedimentos estatísticos capazes de resolver tal situação, tal como modelos GAM e modelos GAMLSS.



Covariância

É uma medida equivalente que mede **a associação entre duas variáveis quantitativas**.

Definição: Dados n pares de valores $(x_1, y_1), \ldots, (x_n, y_n)$, a **covariância** entre as duas variáveis X e Y é:

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

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

Além disso, *corr(X,Y)* pode ser escrito como:

$$corr(X,Y) = \frac{cov(X,Y)}{dp(X)dp(Y)}$$

UNIFESP

Universidade Federal de São Paulo

1933

ASSOCIAÇÃO ENTRE VARIÁVEL QUALITATIVA E QUANTITATIVA

Neste caso, analisa-se o que acontece com a variável quantitativa dentro de cada nível da variável qualitativa.

Exemplo: Considere a tabela descritiva da variável Salário dos funcionários de uma empresa abaixo.

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

	n	Média	dp(X)	var(x)	$X_{(1)}$	q_1	q_2	q_3	$X_{(n)}$
Fundam.	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

dp = desvio padrão; $x_{(1)}$=mínimo; q_1=primeiro quartil; q_2=segundo quartil (mediana); q_3=terceiro quartil; $x_{(n)}$=máximo

UNIFESP
UNIVERSIDADE FEDERAL DE SÃO PAULO
1933

ASSOCIAÇÃO ENTRE VARIÁVEL QUALITATIVA E QUANTITATIVA

Exemplo – continuação...

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

	n	Média	dp(X)	var(x)	$X_{(1)}$	q_1	q_2	q_3	$X_{(n)}$
Fundam.	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

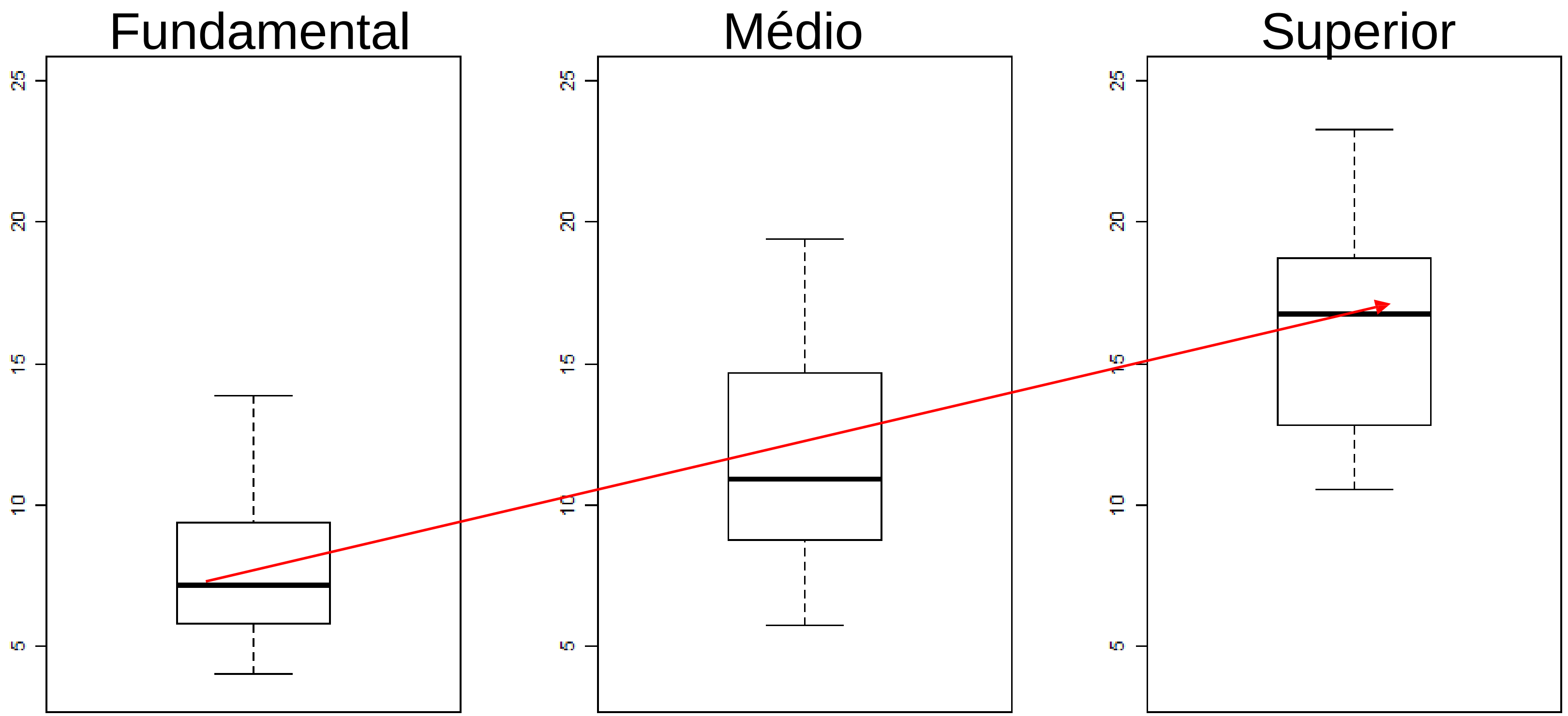
dp = desvio padrão; $x_{(1)}$=mínimo; q_1=primeiro quartil; q_2=segundo quartil (mediana); q_3=terceiro quartil; $x_{(n)}$=máximo

Análise Descritiva:
Há indícios de que quanto maior o grau de instrução, maior é o salário.

Ou seja, **há indícios** de associação entre as variáveis!!

UNIFESP
UNIVERSIDADE FEDERAL DE SÃO PAULO
1933

Box-plots de salário segundo grau de instrução



<u>Análise gráfica:</u>
O salário aumenta conforme aumenta o nível de educação do indivíduo → sugere dependência entre as variáveis.

→ **Mas como mensurar essa associação??**

COEFICIENTE DE DETERMINAÇÃO

- O grau de associação entre as duas variáveis é definido como **o ganho relativo na variância, devido à introdução da variável qualitativa**, é dado por:

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

em que, $0 \leq R^2 \leq 1$ e

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

onde k é o número de categorias, $Var_i(X)$ denota a variância de **X** dentro da categoria *i* e N é o número total de dados.

UNIFESP
UNIVERSIDADE FEDERAL DE SÃO PAULO
1933

Coeficiente de determinação

Voltando ao exemplo (grau de instrução, na Companhia MB)

$$\overline{Var(X)} = \frac{12(7.77)+18(13{,}10)+6(16{,}89)}{12+8+6 = 36} = 11{,}96$$

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

→ Portanto, podemos dizer que 41,5% da variação total do salário é **explicada** pela variável grau de instrução.

COEFICIENTE DE DETERMINAÇÃO

OBSERVAÇÕES:

1. $\overline{Var(X)} \leq Var(X)$;

2. Se as variâncias dentro das classes for menor que a global, significa que a variável qualitativa melhora a capacidade de previsão da quantitativa e portanto existe uma relação entre as duas variáveis.

OUTRAS MEDIDAS DE ASSOCIAÇÃO

→Para conhecimento (não é foco deste curso)!!

Dependendo do tipo de variáveis envolvidas existem outras medidas de associação mais indicadas. Maiores detalhes podem ser vistos em:

https://journals.sagepub.com/doi/pdf/10.1177/8756479308317006

RESUMO DA AULA

- Construção do gráfico Box Plot;
- Discusão sobre outliers;
- Discusão sobre as formas da distribuição dos dados;
- Medidas de associação entre variáveis.

PRÓXIMA AULA

- Teoria de Probabilidade;
- Propriedades;
- Probabilidade condicional;
- Teorema de Bayes.

UNIFESP
UNIVERSIDADE FEDERAL DE SÃO PAULO
1933

Referências

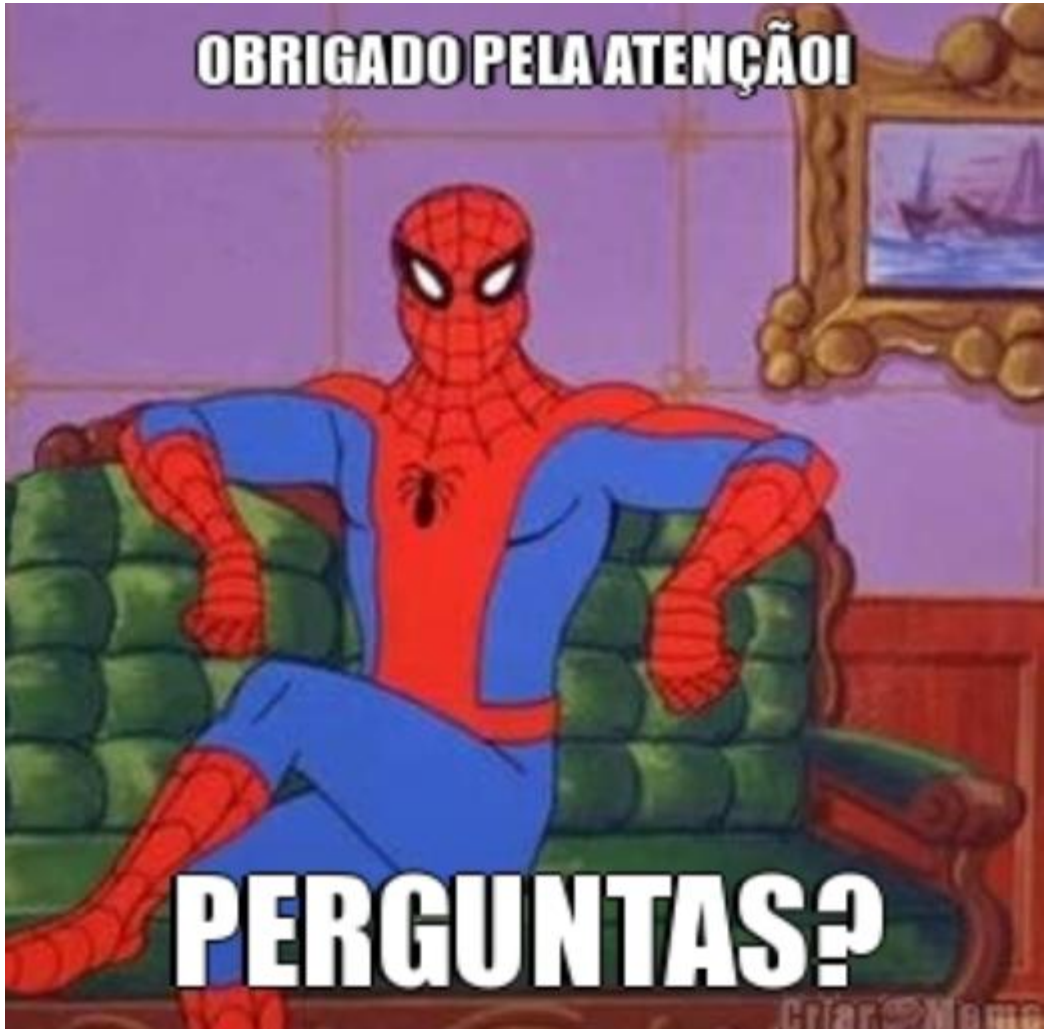
MONTGOMERY, D. C.; RUNGER, G. C. **Estatística aplicada e probabilidade para engenheiros**. 6 ed. Rio de Janeiro: LTC, 2018. 628p.

MORETTIN, P. A.; BUSSAB, W. O. **Estatística básica**. 9 ed. São Paulo: Saraiva, 2017. 554p.

MOORE, D. S. NOTZ, W. I.; FLIGNER, M. A. **A estatística básica e sua prática**. 7 ed. Rio de Janeiro: LTC, 2017. 628p.

UNIFESP
Universidade Federal de São Paulo
1933

Class finished



UNIFESP
Universidade Federal de São Paulo
1933