1) Exploração dos dados no BD

Cap. 4

DADOS JUNTOS AO RH DA EMPRESA

Nº	Estado civil X_1	Grau de instrução X_2	Nº de filhos X_3	Salário (× sal. mín.) X_4	Idade		Região de procedência X_6
					anos X_5	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

RELEMBRANDO:

Variável
- ◇ QUALITATIVA
  - ORDINAL (ESCOLARIDADE)
  - NOMINAL (ESTADO CIVIL, REGIÃO)
- △ QUANTITATIVA (Nº filhos / Idade / salário)

Int outra C

S □ □ □

C □ □ □

Posso fazer:

X_1 = Est. civil
X_2 = Escolari.
X_3 = Nº filhos
X_4 = Idade (anos)
X_5 = Procedência
X_6 = salário

→ genérico

	X_1	X_2	X_3	X_4	X_5	...	X_m
1	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	...	x_{1m}
2	x_{21}						
3	x_{31}						
4	x_{41}						
⋮	⋮						
i	x_{i1}	x_{i2}					
⋮	⋮						
j	x_{j1}	x_{j2}					x_{jm}

$\sum_{t=1}^{j} x_{t1}$

$\sum_{t=1}^{j}\sum_{k=1}^{m} x_{tk}$

↳ somatória facilita entender nossas contas

2) P/ ANALISAR O BD → precisamos investigar a RELAÇÃO ENTRE NOSSAS VARIÁVEIS.

⇒ Aqui temos que considerar, então se as variáveis são DEPENDENTES ou INDEPENDENTES e se são Qualitativas (ORDINAL/NOMINAL) ou Quantitativa (Discretas/contínuas)

Situação 1: (Qualitativa e Qualitativa)

Exemplo: Região de Procedência x Estado civil

X_1 \ X_5	capital	Interior	Outra	TOTAL
Solteiro	3	4	4	11
Casado	4	3	2	9
	7	7	6	20

→ Distribuição conjunta entre X_1 e X_5 dist(X,Y)

Posso montar tabelas ≠s c/ os dados

→ Proporção em relação ao total geral

→ Proporção a uma variável específica → { linhas ou coluna }

	C	I	O	T
S	3/20 (15%)	4/20 (20%)	4/20 (20%)	11/20
C	4/20	3/20	2/20	9/20
	7/20	7/20	6/20	20/20

Em relação ao TOTAL geral

	C	T	O	
S	3/7	4/7	4/6	11/20
C	4/7	3/7	2/6	9/20
	100%	100%	100%	100%

As variáveis são dependentes ou Independentes?

3) Vamos pensar num caso prático

Tamanho da empresa (X_1)
- G //
- P //

Tipo de auditoria (X_2)
- Big Four //
- outras //

Tabela rápido

X_1 \ X_2	Big Four	Outra	TOTAL
G	100 (71%)	20 (33%)	120 (60%)
P	40 (29%)	40 (67%)	80 (40%)
	140 (100%)	100% (60)	200 (100%)

Parece haver associação //
- Baixa freq. nas pequenas
- Alta freq. nas gdes

Como medir essa associação?

- C = coeficiente de contingência (de Pearson)
- $\tilde{C}$ = ou coeficiente de contingência modificado

Vamos aos dados

observado

X_1 \ X_2	Big Four	Outra	total
G	100 (83%)	20 (17%)	120 (100%)
P	40 (50%)	40 (50%)	80 (100%)
TOTAL	140 (70%)	60 (30%)	200 (100%)

Linha

4) Se não houvesse dependência, empresas cerca de 70% de G e P contratariam Big four e 30% outras. Isto é:

→ Esperado

	Big	Outros	TOTAL
G	84 (70%)	36 (30%)	120 100%
P	56 (70%)	24 (30%)	80 100%
TOTAL	140 (70%)	60 (30%)	200 100%

Desvio entre obs e esperado

$e_i = (o_i - o_e)$

	Big	OUTRA
G	100-84 = 16 (3,05)	20-36 = -16 (7,11)
P	40-56 = -16	40-24 = 16 (10,6)

Veja que $\sum_{i=1}^{4} e_i = 0$ → (4,57)

$\frac{(o_i - o_e)^2}{o_e} = \frac{(16)^2}{84} = 3,05$

$\frac{|100-84|^2}{84} = 3,05$

Fazendo: $3,05 + 7,11 + 4,57 + 10,6 = 25,33$

χ^2 (qui-quadrado) de Pearson

→ Valores altos de χ^2 ⇒ Associação

Agora queremos C e $\tilde{C}$

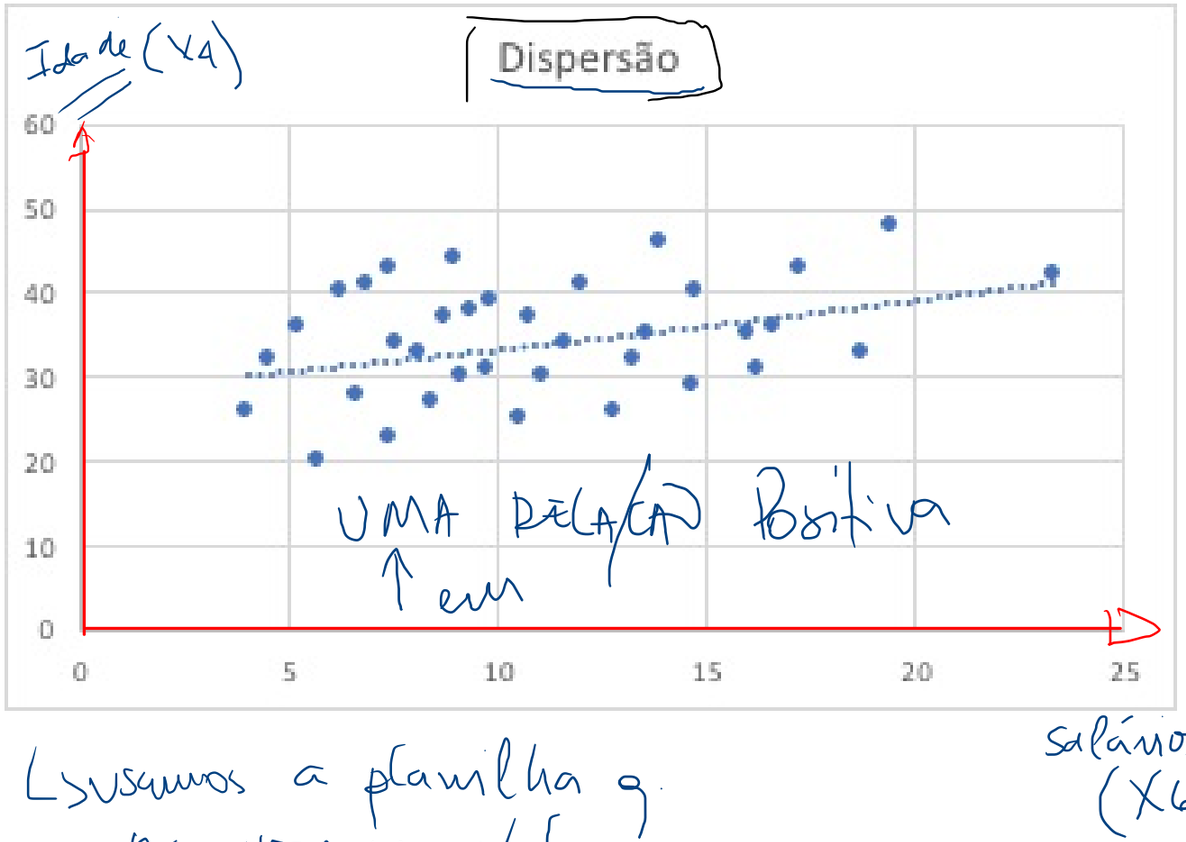
$C = \sqrt{\frac{\chi^2}{\chi^2 + n}} \simeq 0,33$ | $\sqrt{\frac{25,33}{25,33+200}}$

$\tilde{C} = \sqrt{\frac{\chi^2/n}{(n-1)(s-1)}}$ → 0,36

onde n = nº de linhas (2)
s = nº de colunas (2)

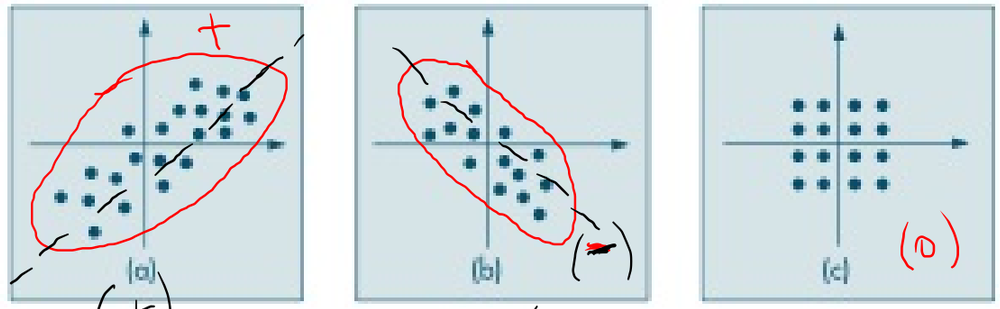
5) Se tivermos X_4 e X_6 quantitativas?

Podemos fazer uma relação entre Idade e salário.



Lsusamos a planilha q. vcs usaram p/ fazer a lista //

RELAÇÕES Possíveis $\frac{dx}{dy} > 0$



$\frac{dx}{dy}$

A medida + tradicional é o Ind. de correlação

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

c/ $(-1 < \rho < 1)$

Veja que precisamos dos seguintes dados

$\bar{X}_4$; $\bar{X}_6$; $dp(x_4)$; $dp(x_6)$; n ; $(X_{4i} - \bar{X}_4)$; $(X_{6i} - \bar{X}_6)$

p/ os dados da planilha: $\rho \cong 0{,}3$

Raciocínio

Se Salário e Idade são associados?

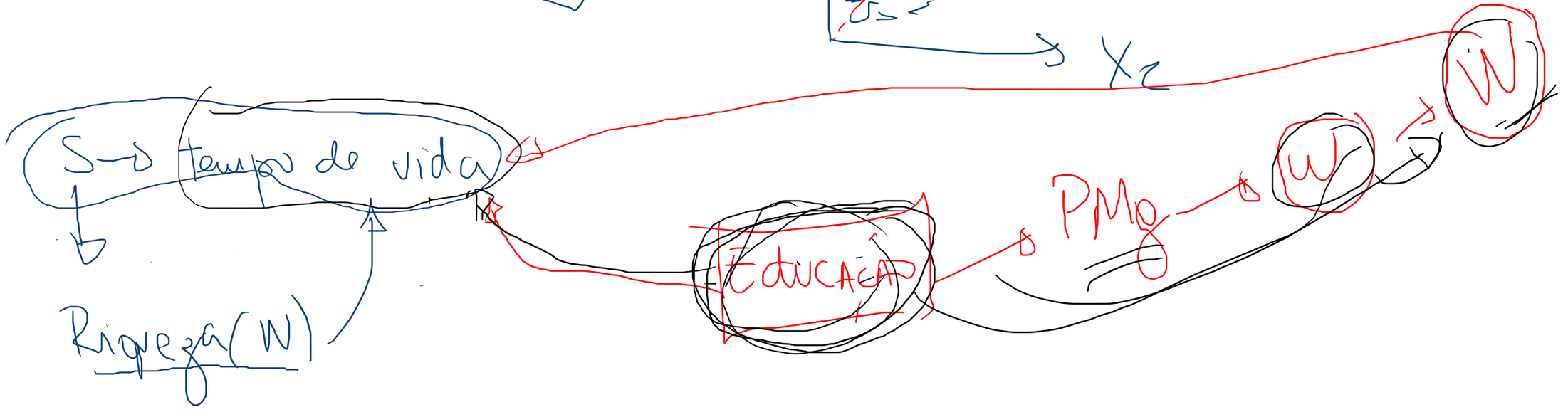
Salário → Quant

Idade → Quant

Y1

dispersão

X2



b) P/ os dados abaixo, podemos fazer:

Vendedor	tempo de casa (X)	Vendas ($) (Y)
1	2	48
2	3	50
3	4	56
4	5	52
5	4	43
6	6	60
7	7	62
8	8	58
9	8	64
10	10	72

$\sum_{i=1}^{10} x_i = 57$ $\quad \sum_{i=1}^{10} y_i = 565$

$\bar{X} = \frac{1}{n}\sum_{1}^{10} x_i = 5,7 \quad \bar{Y} = 56,5$

$(X-\bar{X})$	$(Y-\bar{Y})$	$\frac{x_i-\bar{X}}{dp(x)}$ (Z_x)	$\frac{y_i-\bar{Y}}{dp(y)}$ (Z_y)	$Z_x \cdot Z_y$
$(2-5,7) = -3,7$	$-8,5$	$-3,7/2,41 = -1,54$	$-1,05$	1,617
$= -2,7$	$-6,5$	$-1,12$	$-0,8$	0,846
$= -1,7$	$-0,5$	$-0,71$	$-0,06$	0,043
$= -0,7$	$-4,5$	$-0,29$	$-0,55$	0,160
$= -1,7$	$-13,5$	$-0,71$	$-1,66$	1,179
$= 0,3$	$3,5$	$0,12$	$0,43$	0,052
$= 1,3$	$5,5$	$0,54$	$0,68$	0,367
$= 2,3$	$1,5$	$0,95$	$0,19$	0,181
$= 2,3$	$7,5$	$0,95$	$0,92$	0,874
$= 4,3$	$15,5$	$1,78$	$1,91$	3,4

$\sum_{i=1}^{10} Z_x \cdot Z_y$

$Var(X) = \frac{\sum_{1}^{10}(x_i - \bar{X})^2}{n} \Rightarrow dp(X) = 2,41$

$dp(Y) = 8,11$

$\rho = 0,87$

8,769/10

7) Uma forma alternativa:

$$\rho(x,y) = \frac{1}{n} \sum \left(\frac{x_i - \bar{x}}{dp(x)}\right)\left(\frac{y_i - \bar{y}}{dp(y)}\right)$$

$$= \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sqrt{\left(\sum_1^n x_i^2 - n\bar{x}^2\right)\left(\sum_1^n y_i^2 - n\bar{y}^2\right)}}$$

Outra medida: $\underset{\text{covariância}}{cov}(x,y) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n}$

Assim:

$$\rho = \frac{cov(x,y)}{dp(x) \cdot dp(y)}$$