MAT02018 - Estatística Descritiva

Distribuições bidimensionais

Rodrigo Citton P. dos Reis
citton.padilha@ufrgs.br

Universidade Federal do Rio Grande do Sul
Instituto de Matemática e Estatística
Departamento de Estatística

Porto Alegre, 2021

Variáveis quantitativas

Variáveis quantitativas

- Quando as variáveis envolvidas são ambas quantitativas, podemos "categorizá-las".
- Ou seja, agrupar as observações em **intervalos de classe** para cada uma das variáveis[1] e assim, analisá-las como variáveis qualitativas, apresentando a distribuição conjunta em tabelas de dupla entrada[2].
- Mas, além desse tipo de análise, as variáveis quantitativas são passíveis de procedimentos analíticos e gráficos mais refinados.
- O **gráfico de dispersão** é provavelmente o mais comum destes procedimentos .

---

[1]Da mesma forma que foi apresentado e discutido nas notas de aula de *Distribuição de Frequências*.

[2]Se as variáveis de interesse não apresentam um número muito grande de valores distintos, uma alternativa seria não agrupar em intervalos de classe, e simplesmente utilizar uma tabela de dupla entrada considerando os próprios valores da variável.

Variáveis quantitativas

- **Exemplo:** suponha que vinte e cinco pacientes são atendidos em uma clínica oftalmológica e os seguintes valores são registrados para **pressão intraocular** (PIO) e **idade**.

Table 1: Tabela de dados brutos.

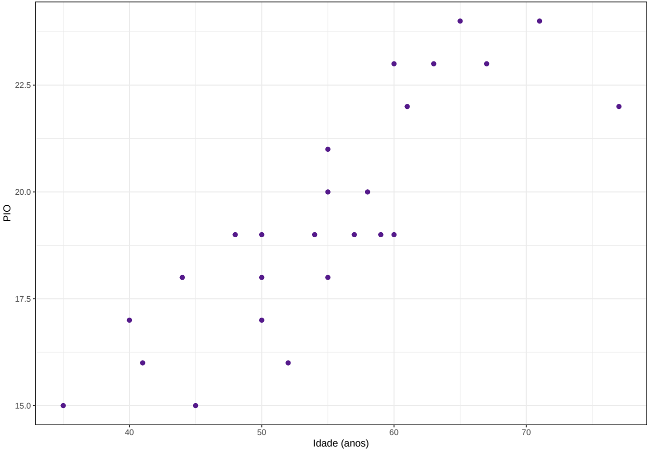
ID	Idade	PIO
1	35	15
2	40	17
3	41	16
4	44	18
5	45	15
6	48	19
7	50	19
8	50	18
9	50	17
10	52	16
11	54	19
12	55	18

Variáveis quantitativas

13	55	21
14	55	20
15	57	19
16	58	20
17	59	19
18	60	23
19	60	19
20	61	22
21	63	23
22	65	24
23	67	23
24	71	24
25	77	22

Variáveis quantitativas

- A figura abaixo apresenta a distribuição conjunta das variáveis idade e pressão intraocular por meio de um gráfico de dispersão.



Variáveis quantitativas

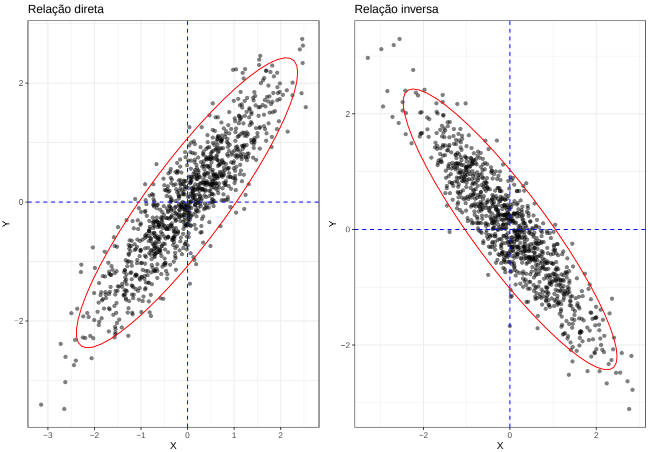
- O gráfico de dispersão é construído pelo conjunto de pontos $\{(x_i, y_i); i = 1, 2, \ldots, n\}$ em que x representa a variável do eixo horizontal (no exemplo, a variável idade) e y representa a variável do eixo vertical (no exemplo, a variável pressão intraocular).
- Note que cada ponto corresponde aos valores observados para um indivíduo específico.
- Através do diagrama de dispersão é possível observar que, em geral, valores de idade mais altos são associados com valores de pressão intraocular mais altos (as variáveis parecem relacionadas).

Variáveis quantitativas

- **Perguntas:**
  - Qual o tipo da relação entre as variáveis Idade e PIO?
  - Qual a força desta relação?

Tipos de relação

- A relação entre duas variáveis quantitativas pode ser descrito como **positivo (direta)** ou **negativo (inversa)**, e **linear** ou **não-linear**.



Tipos de relação

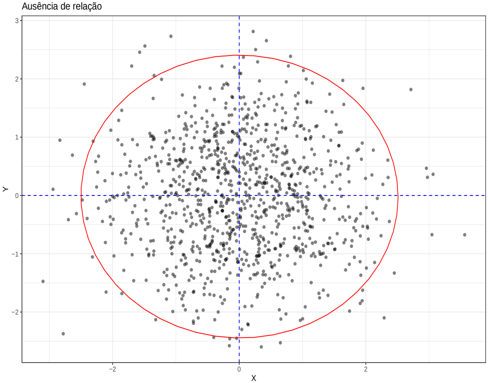
- O gráfico de dispersão da esquerda mostra uma relação direta ou positiva entre as variáveis X e Y, tendência destacada pela declividade positiva da elipse tracejada.
- Perceba também, que o gráfico foi dividido em quatro "quadrantes".
- A grande maioria dos pontos está situada no primeiro e terceiro 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.

Tipos de relação

- De forma análoga, o gráfico de dispersão da direita mostra uma relação inversa ou negativa, tendência também destacada pela declividade negativa da elipse tracejada, e procedendo-se como anteriormente, a soma dos produtos das coordenadas será negativa.

Tipos de relação

- A gráfico a seguir apresenta um exemplo de **ausência de associação** entre as variáveis.
- A soma dos produtos das coordenadas será zero, pois cada resultado positivo tem um resultado negativo simétrico, anulando-se na soma[3].

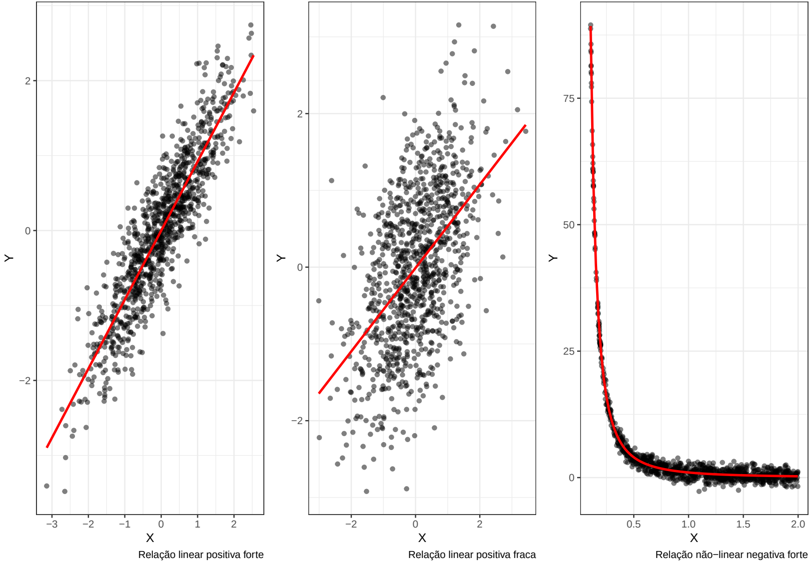


[3]Note que a soma dos produtos das coordenadas depende, e muito, do número de pontos. Além disso, a unidade de medida também pode influenciar. E é por isso, que logo em seguida, vamos considerar uma medida de correlação **padronizada**.

Tipos de relação

- A "força" da relação se refere à proximidade dos pontos na nuvem.
- Se pudéssemos traçar uma curva ou uma reta subjacente, teríamos a força da relação associada a proximidade dos pontos com respeito a esta curva.

Tipos de relacão



Coeficiente de correlação linear

Coeficiente de correlação linear

- A **força** de uma associação entre duas variáveis quantitativas pode ser medida por um **coeficiente de correlação**.
- O **coeficiente de correlação produto-momento** é uma medida da intensidade de **associação linear** existente entre duas variáveis quantitativas.
- Também é conhecido como **coeficiente de correlação de Pearson**, pois sua fórmula de cálculo foi proposta por Karl Pearson em 1896.
- O coeficiente de correlação de Pearson é denominado por ρ na população e r na amostra.

Definição

- Dados n pares **(na amostra)** 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

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

em que $\bar{x}$ e $\bar{y}$ são as médias de X e Y, e s_x e s_y são os desvios padrões de X e Y. Ou seja, o coeficiente de correlação é **a média dos produtos dos valores padronizados das variáveis**.

Propriedades

- O coeficiente de correlação pode variar entre -1 e 1.
- Valores negativos de r indicam uma correlação do tipo inversa (*negativa*):
  - Quando x aumenta, y em média diminui (ou vice-versa).
- Valores positivos para r ocorrem quando a correlação é direta (*positiva*):
  - x e y variam no mesmo sentido.

Propriedades

- O valor máximo (tanto $r = 1$ como $r = -1$) é obtido quando todos os pontos do diagrama de dispersão estão em uma linha reta inclinada (**correlação perfeita**).
- Quando não existe correlação ($r = 0$) entre x e y, os pontos se distribuem em nuvens circulares.
- Quando os pontos formam uma nuvem cujo eixo principal é uma curva (relação não-linear), o valor de r **não mede corretamente a associação** entre as variáveis.

Propriedades

- Da definição do coeficiente de correlação obtemos a seguinte formulação alternativa[4]:

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right) = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sqrt{\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2 \right) \left(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2 \right)}}.$$

- O numerador da expressão acima, que mede o total da concentração dos pontos pelos quatro quadrantes, dá origem a uma medida bastante usada e que definimos a seguir.

Propriedades

Dados n pares (na amostra) de valores $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, chamaremos de **covariância** entre as duas variáveis X e Y a

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

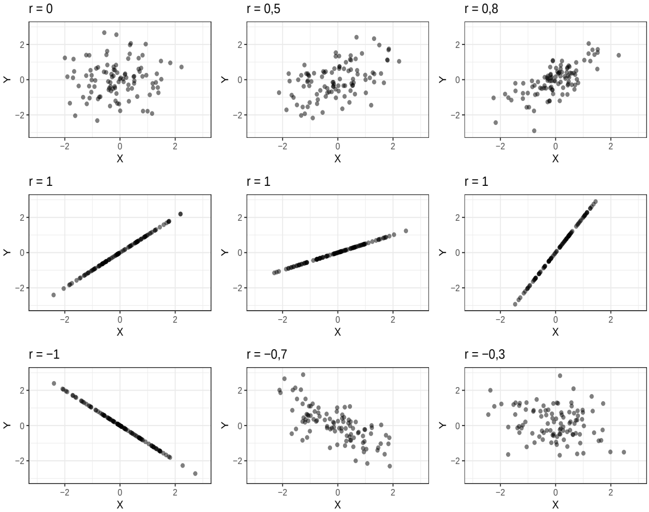
ou seja, a média ("estimada") dos produtos dos valores centrados das variáveis.

- Com essa definição, obtemos a seguinte relação

$$r = \frac{cov_{xy}}{s_x \times s_y}.$$

[4] E dependendo da situação, mais prática.

Exemplos



Exemplos

- Retomando o exemplo da pressão intraocular (PIO) e idade de 25 pacientes atendidos em uma clínica oftalmológica, podemos organizar a seguinte tabela para calcularmos o coeficiente de correlação entre estas duas variáveis.

Table 2: Cálculo do coeficiente de correlação.

ID	Idade (x)	PIO (y)	$x - \bar{x}$	$y - \bar{y}$	$\frac{x-\bar{x}}{s_x} = z_x$	$\frac{y-\bar{y}}{s_y} = z_y$	$z_x \cdot z_y$
1	35	15	-19,88	-4,44	-2,01	-1,63	3,28
2	40	17	-14,88	-2,44	-1,51	-0,90	1,35
3	41	16	-13,88	-3,44	-1,41	-1,26	1,78
4	44	18	-10,88	-1,44	-1,10	-0,53	0,58
5	45	15	-9,88	-4,44	-1,00	-1,63	1,63
6	48	19	-6,88	-0,44	-0,70	-0,16	0,11
7	50	19	-4,88	-0,44	-0,49	-0,16	0,08
8	50	18	-4,88	-1,44	-0,49	-0,53	0,26
9	50	17	-4,88	-2,44	-0,49	-0,90	0,44
10	52	16	-2,88	-3,44	-0,29	-1,26	0,37
11	54	19	-0,88	-0,44	-0,09	-0,16	0,01
12	55	18	0,12	-1,44	0,01	-0,53	-0,01
13	55	21	0,12	1,56	0,01	0,57	0,01
14	55	20	0,12	0,56	0,01	0,21	0,00

Exemplos

	15	57	19	2,12	-0,44	0,21	-0,16	-0,03
	16	58	20	3,12	0,56	0,32	0,21	0,06
	17	59	19	4,12	-0,44	0,42	-0,16	-0,07
	18	60	23	5,12	3,56	0,52	1,31	0,68
	19	60	19	5,12	-0,44	0,52	-0,16	-0,08
	20	61	22	6,12	2,56	0,62	0,94	0,58
	21	63	23	8,12	3,56	0,82	1,31	1,07
	22	65	24	10,12	4,56	1,03	1,67	1,72
	23	67	23	12,12	3,56	1,23	1,31	1,60
	24	71	24	16,12	4,56	1,63	1,67	2,73
	25	77	22	22,12	2,56	2,24	0,94	2,11
Soma	325	1372	486	0,00	0,00			20,28

Exemplos

E assim, temos que:

- $\bar{x} = 1372/25 = 54,88$ anos e $s_x = 9,87$ anos.
- $\bar{y} = 486/25 = 19,44$ mmHg e $s_y = 2,73$ mmHg.
- O coeficiente de correlação é $r = 20,28/25 = 0,81$, uma relação direta (positiva) e relativamente forte[5].

[5] Maior a idade $\Rightarrow$ Maior a pressão intraocular.

Próxima aula

- Distribuições bivariadas: uma variável qualitativa e uma variável quantitativa.
- ComplementaR.

Por hoje é só!

Bons estudos!