Estudos de correlação

Prof. Luiz R. Nakamura

Departamento de Informática e Estatística
Universidade Federal de Santa Catarina

luiz.nakamura@ufsc.br

Florianópolis – SC

Introdução

Análise bidimensional

Verificar o relacionamento entre duas variáveis em estudo e, para isso, essas variáveis devem ser observadas e analisadas simultaneamente

Coeficiente de correlação linear de Pearson

Resume o relacionamento entre duas variáveis quantitativas em apenas um número

Modelo de regressão linear simples

Descreve o relacionamento entre duas variáveis por meio de uma equação

Conceitos introdutórios

Quando estudamos o relacionamento entre duas variáveis, estamos, na realidade, estudando a relação ou estrutura de dependência ou associação dessas variáveis

- uma variável será chamada de **independente** e será representada pela letra X. Outras terminologias utilizadas: explicativa, explanatória, feature, ...
- a outra variável em estudo será chamada de **dependente** e será denotada por Y. Outras terminologias utilizadas: resposta, alvo, target, ...

Observações

- Para que seja possível realizar uma análise de correlação e/ou regressão, os dados devem provir de observações emparelhadas e em condições semelhantes
- Tamanho da amostra utilizada deve ser razoável ($n \geq 30$) para realizar conclusões

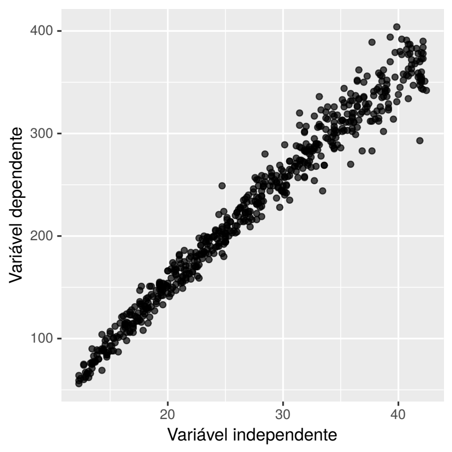
Diagrama de dispersão

Objetivos

Conceber uma ideia inicial de como duas variáveis quantitativas estão relacionadas

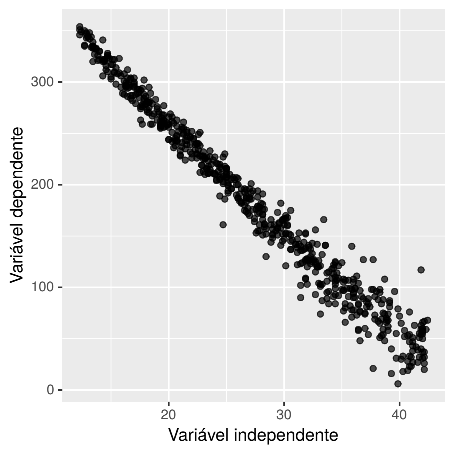
- A **direção** dessa relação: o que acontece com Y quando X aumenta?
- A **força** dessa relação: a qual "taxa" os valores de Y aumentam ou diminuem em função de X
- A **natureza** dessa relação: qual o tipo de relacionamento entre as duas variáveis? Podemos descrevê-lo com uma reta, parábola, exponencial etc.

Exemplos



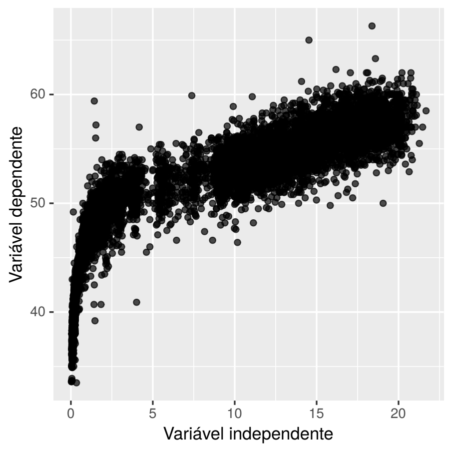
- **Direção:** à medida que a variável X aumenta, os valores de Y tendem a aumentar também
- **Força:** a taxa de crescimento é constante ao longo de todo eixo X
- **Natureza:** seria possível ajustar uma reta crescente que passasse por entre os pontos
- **Conclusão:** há correlação linear forte e positiva

Exemplos



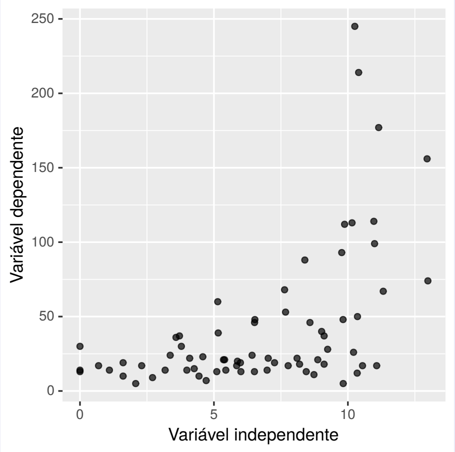
- **Direção:** à medida que a variável X aumenta, os valores de Y tendem a diminuir
- **Força:** a taxa de decrescimento é constante ao longo de todo eixo X
- **Natureza:** seria possível ajustar uma reta decrescente que passasse por entre os pontos
- **Conclusão:** há correlação linear forte e negativa

Exemplos



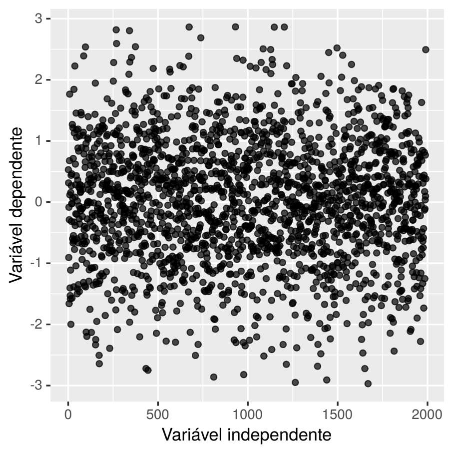
- **Direção:** à medida que a variável X aumenta, os valores de Y tendem a aumentar também
- **Força:** para valores muito pequenos de X a taxa de aumento em Y é muito alta. Posteriormente essa taxa é bem baixa, isto é, Y cresce de maneira extremamente suave.
- **Natureza:** não seria razoável ajustar uma reta que passasse por entre os pontos. Uma opção seria, por exemplo, utilizar uma função logarítmica.
- **Conclusão:** há forte correlação entre as variáveis, porém ela não é linear.

Exemplos



- **Direção:** à medida que a variável X aumenta, os valores de Y tendem a aumentar também
- **Força:** para valores pequenos de X a taxa de aumento em Y é quase nula. Posteriormente essa taxa aumenta, isto é, Y cresce de maneira mais acentuada.
- **Natureza:** não seria razoável ajustar uma reta que passasse por entre os pontos. Uma opção seria, por exemplo, utilizar uma função exponencial.
- **Conclusão:** há uma baixa ou moderada correlação entre as variáveis e ela não é linear

Exemplos



- **Direção:** não há um padrão aparente nos pontos
- **Força:** os pontos parecem se distribuir de maneira aleatória
- **Natureza:** não é possível considerar qualquer função para representar as observações
- **Conclusão:** não há um relacionamento aparente entre as duas variáveis

Coeficiente de correlação linear de Pearson

Objetivo

Os objetivos do coeficiente de correlação linear de Pearson são o de mensurar, por meio de um único valor, o grau de relacionamento entre duas variáveis quantitativas, bem como indicar a direção dessa relação

Notação

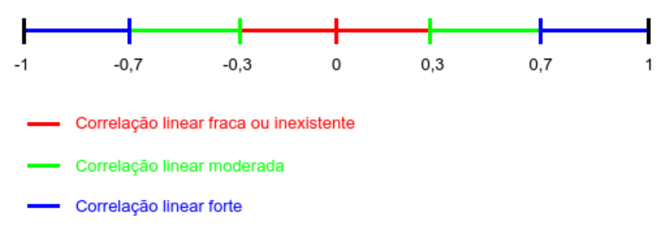
- O coeficiente de correlação linear de Pearson **populacional** é definido pela letra ρ
- O coeficiente de correlação linear de Pearson **amostral** é definido pela letra r

Definição

O coeficiente de correlação linear amostral de Pearson pode ser expresso como:

$$r = \frac{\sum_{i=1}^{n} X_i Y_i - n\bar{x}\bar{y}}{\sqrt{\sum_{i=1}^{n} X_i^2 - n(\bar{x})^2}\sqrt{\sum_{i=1}^{n} Y_i^2 - n(\bar{y})^2}}$$

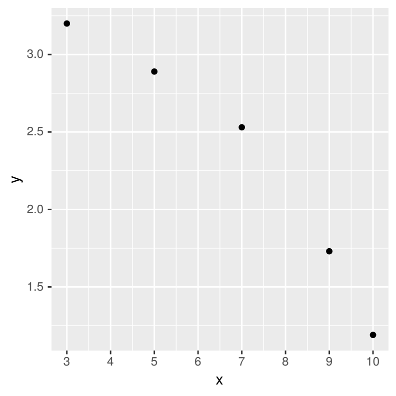
Interpretação



Exemplo

Considere as variáveis X e Y como dispostas a seguir. Construa o diagrama de dispersão e calcule o coeficiente de correlação linear de Pearson. Interprete seu resultado.

X	3	5	7	9	10
Y	3,20	2,89	2,53	1,73	1,19



$$r = \frac{\sum_{i=1}^{n} X_i Y_i - n\bar{x}\bar{y}}{\sqrt{\sum_{i=1}^{n} X_i^2 - n(\bar{x})^2}\sqrt{\sum_{i=1}^{n} Y_i^2 - n(\bar{y})^2}}$$

	X	Y	XY	X^2	Y^2
	3	3,20	9,60	9	10,24
	5	2,89	14,45	25	8,35
	7	2,53	17,71	49	6,40
	9	1,73	15,57	81	2,99
	10	1,19	11,90	100	1,42
Σ	34	11,54	69,23	264	29,40

$$\bar{x} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{34}{5} = 6,8$$

$$\bar{y} = \frac{\sum_{i=1}^{n} Y_i}{n} = \frac{11,54}{5} = 2,31$$

$$\begin{aligned} r &= \frac{69,23 - 5 \times 6,8 \times 2,31}{\sqrt{264 - 5 \times (6,8)^2}\sqrt{29,40 - 5 \times (2,31)^2}} \\ &= -0,97 \end{aligned}$$

A correlação linear entre as variáveis X e Y é forte e negativa. A medida que o valor de X aumenta, o valor de Y diminui

Exemplo

Vamos avaliar as idades de 12 mulheres, relacionando-as com suas pressões arteriais. Construa um diagrama de dispersão e calcule o coeficiente de correlação linear de Pearson para os dados a seguir. Interprete os resultados.

Tabela: Idade e pressão

Idade	Pressão
56	147
42	125
72	160
36	118
47	128
55	150
49	145
38	115
42	140
68	152
60	155
63	149

Utilizando o software R

```
> x = c(56, 42, 72, 36, 47, 55,
      49, 38, 42, 68, 60, 63)

> y = c(147, 125, 160, 118, 128,
      150, 145, 115, 140, 152,
      155, 149)

> plot(x, y)

> cor(x, y)
```

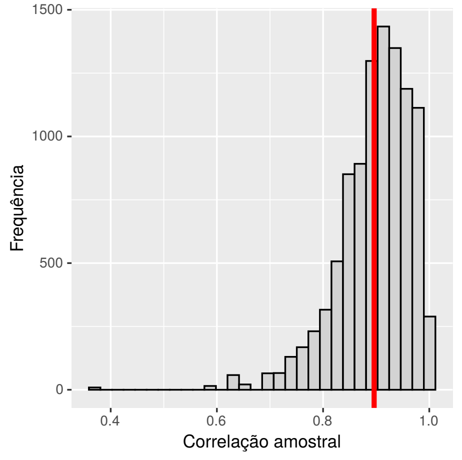
Inferência sobre o parâmetro ρ

- Intervalo de confiança
- Teste de hipótese

Intervalo de confiança (revisão)

- O que é?
- Como é construído?
  - $\mu = \bar{x} + \text{erro}$
- Ideia via simulação

Exemplo: 10.000 amostras



A distribuição amostral do coeficiente de correlação de Pearson é assimétrica!

Construção de um intervalo de confiança para ρ

Como vimos, a distribuição amostral do coeficiente de correlação amostral r não é simétrica. Assim, a construção do intervalo de confiança é baseada em uma transformação do coeficiente r.

Passo 1: Transformar o coeficiente r

$$z_r = \frac{1}{2} \ln \left(\frac{1+r}{1-r} \right)$$

Passo 2: Calcular os limites do coeficiente transformado

$$a = LI_{z_r} = z_r - z_{\frac{\alpha}{2}} \sqrt{\frac{1}{n-3}}$$

$$b = LS_{z_r} = z_r + z_{\frac{\alpha}{2}} \sqrt{\frac{1}{n-3}}$$

$z_{\frac{\alpha}{2}}$ é o valor obtido a partir da tabela da distribuição normal padrão

Passo 3: Os limites do intervalo de confiança para ρ são dados por

$$LI_r = \frac{\exp\{2a\} - 1}{\exp\{2a\} + 1}$$

$$LS_r = \frac{\exp\{2b\} - 1}{\exp\{2b\} + 1}$$

Exemplo

Estamos avaliando as médias de 15 estudantes no ensino médio, relacionando-as com os índices dos mesmos estudantes nos seus cursos universitários. Sabendo que o coeficiente de correlação linear entre essas duas variáveis é igual à $r = 0,90$. Encontre e interprete o intervalo de 90% de confiança para o verdadeiro coeficiente de correlação populacional ρ

Passo 1

$$z_r = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) = \frac{1}{2}\ln\left(\frac{1+0,9}{1-0,9}\right) = 1,4722$$

Passo 2

$$a = LI_{z_r} = z_r - z_{\frac{\alpha}{2}}\sqrt{\frac{1}{n-3}} = 1,4722 - 1,64\sqrt{\frac{1}{15-3}} = 0,9988$$
$$b = LS_{z_r} = z_r + z_{\frac{\alpha}{2}}\sqrt{\frac{1}{n-3}} = 1,4722 + 1,64\sqrt{\frac{1}{15-3}} = 1,9456$$

Passo 3: Os limites do intervalo de confiança para ρ são dados por

$$LI_r = \frac{\exp\{2a\} - 1}{\exp\{2a\} + 1} = \frac{\exp\{2 \times 0,9988\} - 1}{\exp\{2 \times 0,9988\} + 1} = 0,76$$

$$LS_r = \frac{\exp\{2b\} - 1}{\exp\{2b\} + 1} = \frac{\exp\{2 \times 1,9456\} - 1}{\exp\{2 \times 1,9456\} + 1} = 0,96$$

Conclusão

Com 90% de confiança, o intervalo I.C.(ρ; 90%) $= [0,76; 0,96]$ contém o verdadeiro valor do coeficiente de correlação linear entre as variáveis em estudo. Isto é, como todos os valores presentes no intervalo são superiores a 0,70, existem evidências de que a correlação entre as notas no ensino médio e os índices na universidade é forte.

Observação

Se o valor zero estiver no intervalo de confiança calculado, existem evidências de que as variáveis em estudo são independentes

Teste de hipótese

Hipóteses (Teste bilateral)

As hipóteses a serem testadas são:

$$\begin{cases} H_0 : & \rho = 0 \\ H_1 : & \rho \neq 0 \end{cases}$$

Interpretação das hipóteses

Isto é, deseja-se testar se as variáveis X e Y são não correlacionadas (H_0) contra a hipótese de que elas são correlacionadas (H_1)

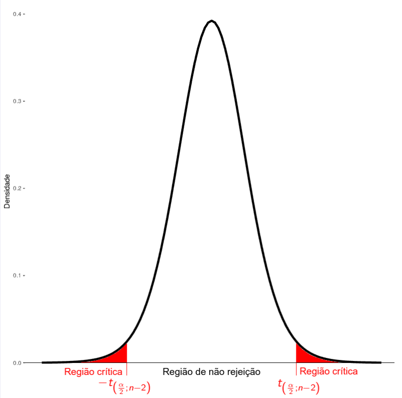
Estatística do teste

$$t_{\text{calc}} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

Valor crítico e região crítica

A estatística do teste deve ser comparada com o valor obtido na tabela da distribuição t de Student: $t_{\left(\frac{\alpha}{2};n-2\right)}$. Assim, a região crítica do teste é dada por:

$$t_{\text{calc}} \leq -t_{\left(\frac{\alpha}{2};n-2\right)} \text{ e } t_{\text{calc}} \geq t_{\left(\frac{\alpha}{2};n-2\right)}$$



Exemplo

Estamos avaliando as médias de 15 estudantes no ensino médio, relacionando-as com os índices dos mesmos estudantes nos seus cursos universitários. Sabendo que o coeficiente de correlação linear entre essas duas variáveis é igual à $r = 0,90$. Pede-se: verifique, ao nível de 10% de significância se as variáveis são, de fato, correlacionadas

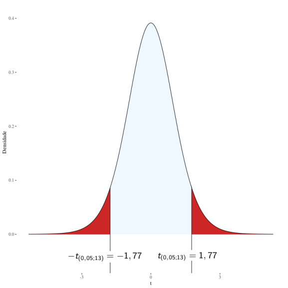
Hipóteses testadas

$$\begin{cases} H_0: & \rho = 0 \\ H_1: & \rho \neq 0 \end{cases}$$

Estatística do teste

$$\begin{aligned} t_{\text{calc}} &= \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \\ &= \frac{0,90\sqrt{15-2}}{\sqrt{1-0,90^2}} \\ &= 7,445 \end{aligned}$$

Região crítica e conclusões



Como $t_{calc} = 7,445 \geq t_{(0,05;13)} = 1,77$, ao nível de 5% de significância, rejeitamos a hipótese nula. Assim, existem evidências de que $\rho \neq 0$, ou seja, existe correlação entre as médias no ensino médio e os indíces na universidade.

Valor p

Podemos obter as mesmas conclusões baseado no valor p que, neste exemplo, é igual a $0,00000487$. Como o valor $p < 0,10$ então a hipótese nula é rejeitada.

Exemplo

Vamos avaliar as idades de 12 mulheres, relacionando-as com suas pressões arteriais. Sabemos que o coeficiente de correlação linear amostral entre as variáveis é $r = 0,897$. Pede-se: verifique, ao nível de 5% de significância se as variáveis são, de fato, correlacionadas

Tabela: Idade e pressão

Idade	Pressão
56	147
42	125
72	160
36	118
47	128
55	150
49	145
38	115
42	140
68	152
60	155
63	149

Cálculo de r, I.C. e T.H.

```
> x = c(56, 42, 72, 36, 47, 55, 49,
        38, 42, 68, 60, 63)

> y = c(147, 125, 160, 118, 128, 150,
        145, 115, 140, 152, 155, 149)

> cor.test(x, y)
```