MÉTODOS ESTATÍSTICOS

Conteúdo programático

1. Revisão de conceitos fundamentais da Estatística
2. Correlação linear
3. Análise de regressão linear simples
4. Análise de regressão múltipla (2 variáveis)
5. Análise de dados de classificação simples e dupla

Medidas descritivas

- Medidas de localização ou tendência central
  - Média aritmética
  - Mediana
  - Moda
- Medidas separatrizes
  - Quantis
  - Mediana
  - Quartis
- Medidas de variação ou dispersão
  - Amplitude total
  - Variância
  - Desvio padrão
  - Coeficiente de variação
- Medidas de formato
  - Coeficiente de assimetria
  - Coeficiente de curtose

Medidas descritivas

- Medidas de localização ou tendência central
  - Média aritmética
- Medidas separatrizes
- Medidas de variação ou dispersão
  - Variância
  - Desvio padrão
- Medidas de formato
- Medidas de associação
  - Covariância
  - Coeficiente de correlação linear de Pearson

Tópico 2. Correlação linear

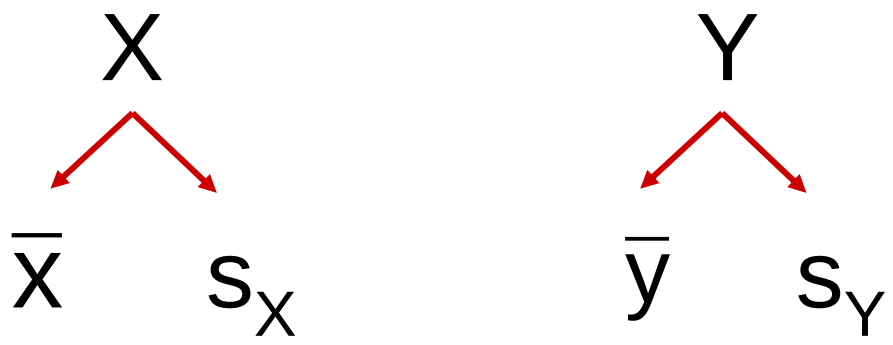
2.1 Medidas de associação: covariância e coeficiente de correlação linear

2.2 Estimação do coeficiente de correlação linear de Pearson

2.3 Inferências sobre o coeficiente de correlação linear de Pearson

Associação entre variáveis

Duas ou mais variáveis



→ Se elas não se relacionam, podem ser estudadas individualmente

→ Se elas estão correlacionadas, devem ser estudadas conjuntamente

Associação entre variáveis

Duas ou mais variáveis

Objetivos:

→ analisar o **comportamento conjunto** dessas variáveis

→ identificar se existe algum tipo de **associação entre elas**

- valores altos (ou baixos) de uma variável ocorrem junto com valores altos (ou baixos) da outra variável?

Exemplos:

✓ relação entre a altura dos pais e a altura dos filhos

✓ relação entre peso de uma pessoa e sua pressão arterial

✓ relação entre peso da carga de caminhões e seu consumo de combustível

Associação entre variáveis

- Na engenharia de recursos hídricos também há interesse em conhecer o grau de associação entre duas ou mais variáveis, como por exemplo:
  - ✓ relação entre as vazões médias anuais e as áreas de drenagem;
  - ✓ relação entre as alturas anuais de precipitação e as altitudes dos postos pluviométricos;
  - ✓ relação entre as intensidades, as durações e as frequências das precipitações intensas.

- O primeiro objetivo é o de analisar o comportamento simultâneo das variáveis, tomadas duas a duas, verificando se a variação positiva (ou negativa) de uma delas está associada a uma variação positiva (ou negativa) da outra, ou se não há nenhuma forma de associação entre elas.

Diagrama de dispersão bivariada

Primeira abordagem exploratória → diagrama de dispersão de pontos

Permite visualizar o grau de associação e a tendência de variação conjunta

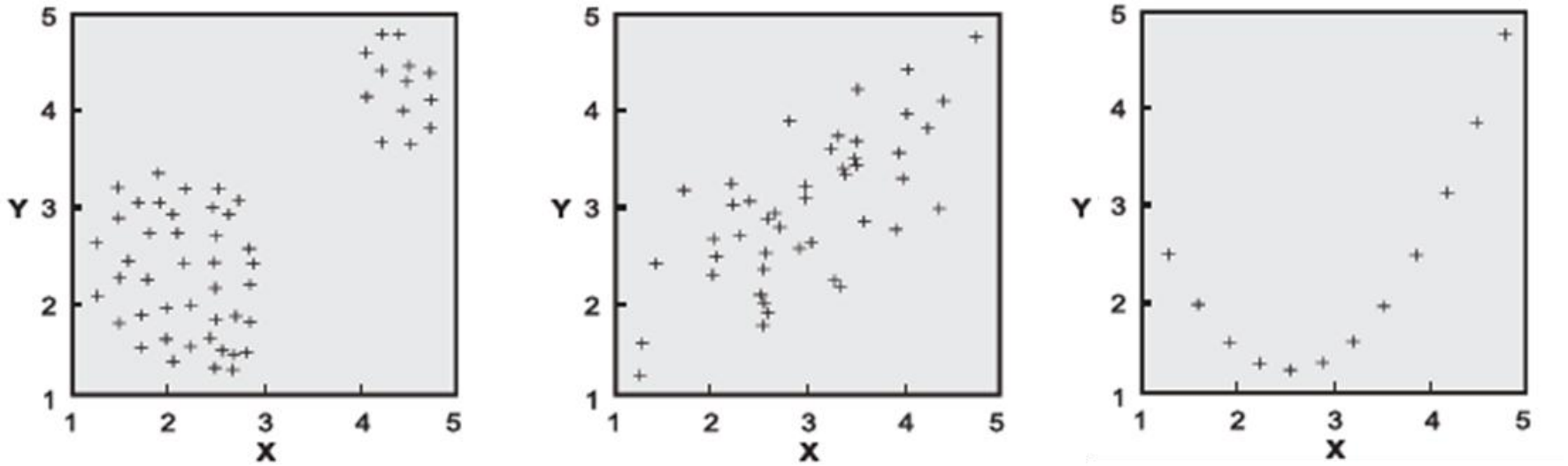


Figura 1. Exemplos de variação conjunta entre duas variáveis.
Fonte: Naghettini e Pinto (2007).

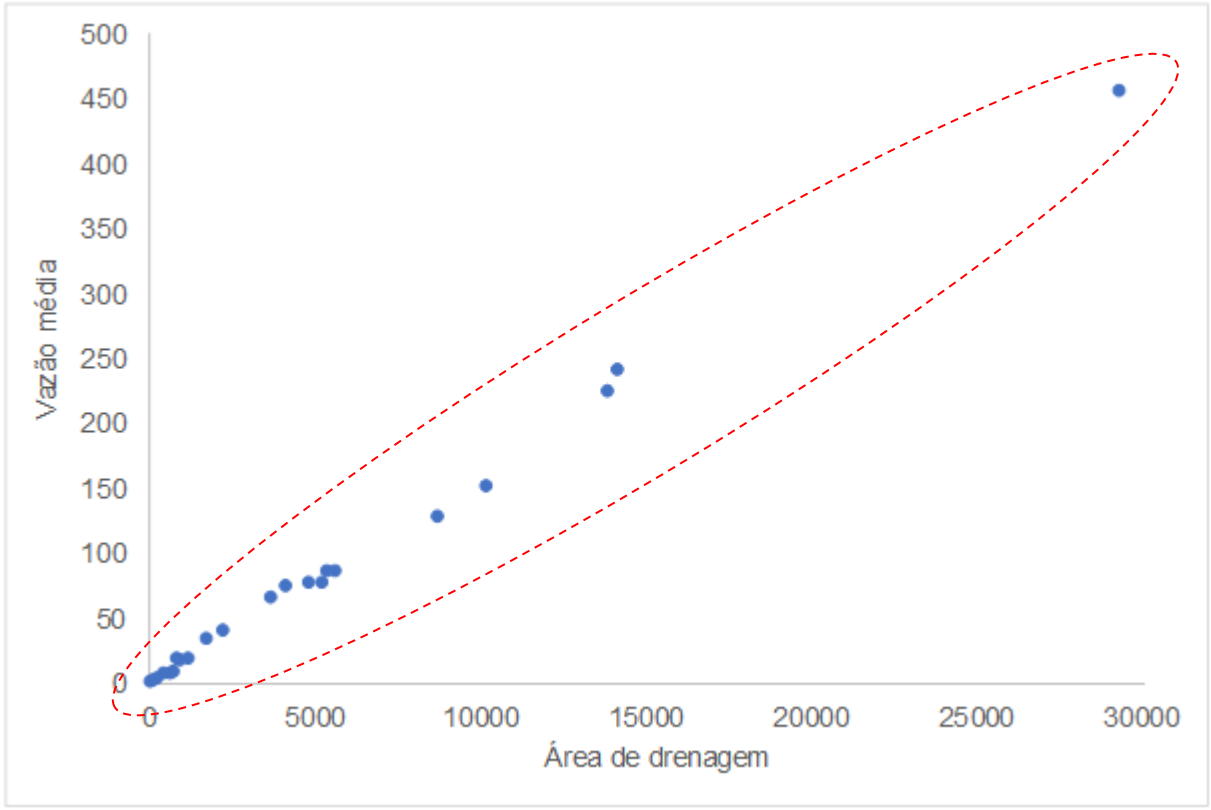
Exemplo:

Um engenheiro está estudando a bacia do rio São Francisco. Um dos objetivos da pesquisa é verificar se existe correlação entre a área de drenagem e a vazão média de longo termo, observadas em 22 estações fluviométricas do alto rio. Os valores observados foram os seguintes:

Tabela 1. Área de drenagem e vazão média (Q) de 22 estações fluviométricas da bacia do alto rio São Francisco.

Estação	Área	Q (m^3/s)	Estação	Área	Q (m^3/s)
1	83,9	1,32	12	3727,4	65,30
2	188,3	2,29	13	4142,9	75,00
3	279,4	4,24	14	4874,2	77,20
4	481,3	7,34	15	5235,0	77,50
5	675,7	8,17	16	5414,2	86,80
6	769,7	8,49	17	5680,4	85,70
7	875,8	18,90	18	8734,0	128,00
8	964,2	18,30	19	10191,5	152,00
9	1206,9	19,30	20	13881,8	224,00
10	1743,5	34,20	21	14180,1	241,00
11	2242,4	40,90	22	29366,2	455,00

Diagrama de dispersão



O diagrama de dispersão fornece uma idéia do tipo de relacionamento entre as duas variáveis que, em geral, são denotadas por X e Y

→ à medida que aumenta a área de drenagem observa-se um aumento da vazão média

Medidas de associação

- A forma de medir a intensidade da relação entre variáveis depende do tipo de variáveis e da escala em que elas são medidas.

- Existem muitos tipos de correlação: simples, múltipla, parcial, canônica, etc.

- Será tratado apenas do tipo mais comum: a **correlação linear simples**

- A intensidade da associação linear simples entre duas variáveis **numéricas contínuas** (X e Y) pode ser mensurada por meio de duas medidas:

  - ✓ **Covariância**
  - ✓ **Coeficiente de correlação linear de Pearson**

Covariância

- É uma medida do grau de correlação linear entre duas variáveis
  - → expressa a variação conjunta das variáveis

- É denotada por s_{xy} e definida como a **média dos produtos dos desvios de X pelos desvios de Y**

$$s_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

← **Fórmula de definição**

- Pode assumir valores positivos (indicando relação positiva entre X e Y) ou negativos (indicando relação negativa)

- Sua desvantagem é a **dificuldade de interpretação**, pois seus valores podem variar de $-\infty$ a $+\infty$.

Variância: média dos quadrados dos desvios

$$s_x^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}$$

$$s_y^2 = \frac{\sum(y_i - \bar{y})^2}{n-1}$$

Covariância: média dos produtos dos desvios de X e de Y

$$s_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

Covariância

Soma dos produtos dos desvios de X e Y

$$s_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1} = \frac{SPXY}{n-1} = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{n-1}$$

Fórmula de definição

Fórmula prática

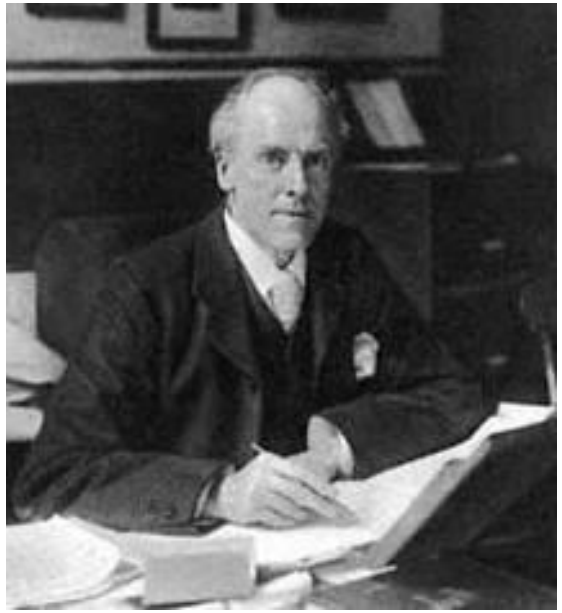
Demonstração

$$\begin{aligned}
\sum (x_i - \bar{x})(y_i - \bar{y}) &= \sum \left[x_i (y_i - \bar{y}) - \bar{x}(y_i - \bar{y}) \right] \\
&= \sum x_i (y_i - \bar{y}) - \sum \bar{x}(y_i - \bar{y}) \\
&= \sum (x_i y_i - x_i \bar{y}) - \bar{x} \sum (y_i - \bar{y}), \quad \text{sendo } \sum (y_i - \bar{y}) = 0 \\
&= \sum x_i y_i - \sum x_i \bar{y} \\
&= \sum x_i y_i - n\bar{x}\bar{y}
\end{aligned}$$

Medidas de associação

- A forma de medir a intensidade da relação entre variáveis depende do tipo de variáveis e da escala em que elas são medidas.
- Existem muitos tipos de correlação: simples, múltipla, parcial, canônica, etc.
- Será tratado apenas do tipo mais comum: a **correlação linear simples**
- A intensidade da associação linear simples entre duas variáveis **numéricas contínuas** (X e Y) pode ser mensurada por meio de duas medidas:
  - ✓ **Covariância**
  - ✓ **Coeficiente de correlação linear de Pearson**

Coeficiente de Correlação Linear de Pearson



Karl Pearson
(1857-1936)

O nome deste coeficiente é uma homenagem ao trabalho pioneiro do matemático britânico Karl Pearson, que desenvolveu um grande número de métodos estatísticos e, em 1901, fundou a revista *Biometrika*.

Trouxe contribuições extremamente importantes para o desenvolvimento da teoria da análise de regressão e de correlação, bem como do teste de qui-quadrado.

Foi um dos fundadores da Estatística moderna.

Coeficiente de correlação linear de Pearson

- É uma medida da correlação linear entre duas variáveis
- É preferível à covariância por ser **mais precisa** e ser **independente das unidades de medida** de X e de Y, variando de -1 a 1
- É denotado por **r** ou r_{xy} e é definido pela seguinte expressão:

$$r = \frac{s_{xy}}{s_x s_y} = \frac{\frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1}}{\sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}} \sqrt{\frac{\sum (y_i - \bar{y})^2}{n-1}}}$$

← **Fórmula de definição**

- Esta expressão **não é muito prática** para cálculo manual

⇨ É possível simplificar a expressão matemática de r, obtendo outra mais conveniente para o cálculo manual

$$r = \frac{s_{xy}}{s_x s_y} = \frac{\frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1}}{\sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}\sqrt{\frac{\sum (y_i - \bar{y})^2}{n-1}}} = \frac{\frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1}}{\frac{\sqrt{\sum (x_i - \bar{x})^2}}{\sqrt{n-1}} \times \frac{\sqrt{\sum (y_i - \bar{y})^2}}{\sqrt{n-1}}} = \frac{\frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1}}{\frac{\sqrt{\sum (x_i - \bar{x})^2}\sqrt{\sum (y_i - \bar{y})^2}}{n-1}}$$

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2}\sqrt{\sum (y_i - \bar{y})^2}} = \frac{SPXY}{\sqrt{SQX.SQY}} = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sqrt{\left(\sum x_i^2 - n\bar{x}^2\right)\left(\sum y_i^2 - n\bar{y}^2\right)}}$$

$$SPXY = \sum (x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - n\bar{x}\bar{y}$$

$$SQX = \sum (x_i - \bar{x})^2 = \sum x_i^2 - n\bar{x}^2$$

$$SQY = \sum (y_i - \bar{y})^2 = \sum y_i^2 - n\bar{y}^2$$

Fórmulas práticas

Medidas de associação

Covariância

$$s_{xy} = \frac{SPXY}{n-1} = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{n-1}$$

Coeficiente de correlação linear de Pearson

$$r = \frac{s_{xy}}{s_x s_y} = \frac{SPXY}{\sqrt{SQX.SQY}} = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sqrt{\left(\sum x_i^2 - n\bar{x}^2\right)\left(\sum y_i^2 - n\bar{y}^2\right)}}$$

SPXY → Soma dos produtos dos desvios de X e Y

SQX → Soma dos quadrados dos desvios de X

SQY → Soma dos quadrados dos desvios de Y

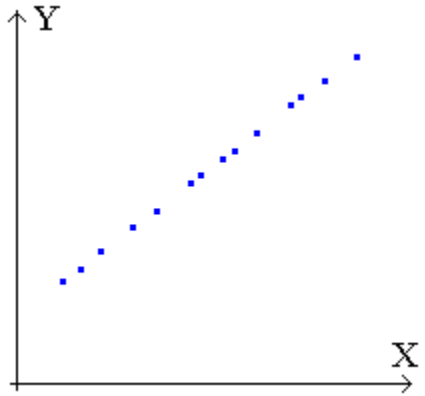
Interpretação do coeficiente de correlação

- Apesar de ser um valor adimensional, ele não é uma taxa e, portanto, o resultado não deve ser expresso em percentagem
- É uma medida da relação linear entre as duas variáveis e não tem sentido quando a relação é não linear

A interpretação dos valores é a seguinte:

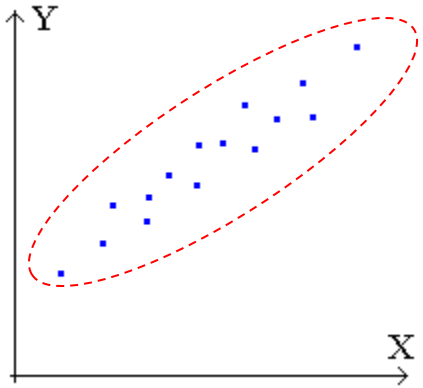
- valores próximos de **+1** indicam uma correlação **positiva** e **forte** entre x e y
- valores próximos de **-1** indicam uma correlação **negativa** e **forte** entre x e y
- valores próximos de **0** indicam uma correlação **fraca** entre x e y

Interpretação do coeficiente de correlação



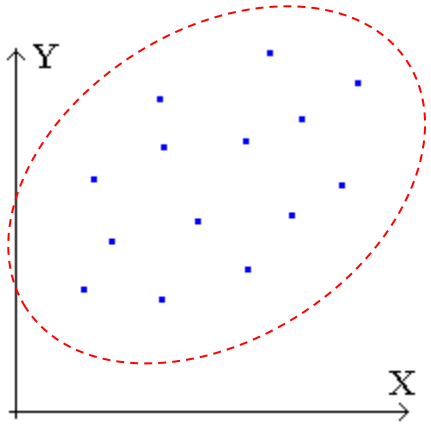
Positiva perfeita

r = 1



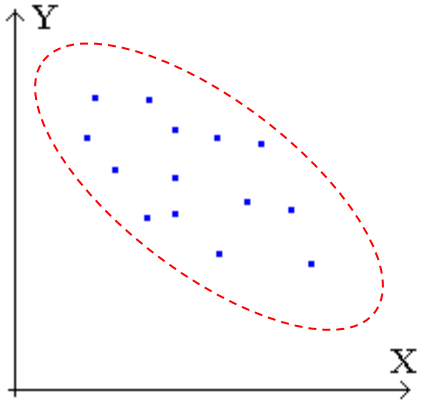
Positiva elevada

r = 0,8



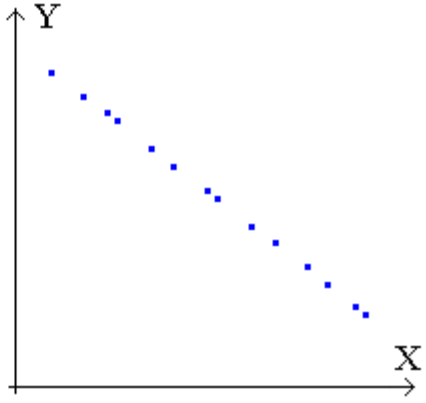
Positiva baixa

r = 0,1



Negativa média

r = -0,5

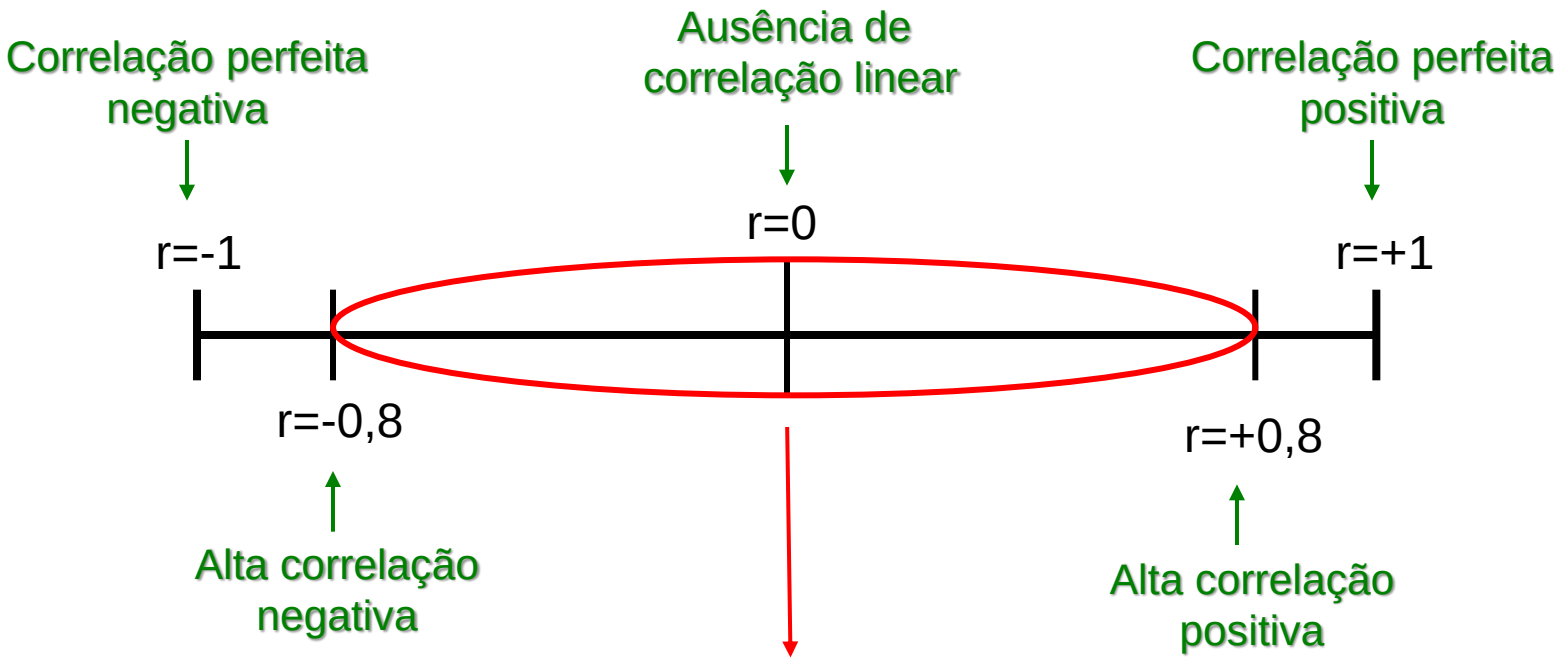


Negativa perfeita

r = -1

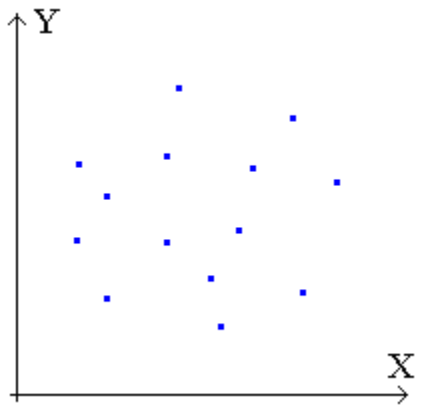
Interpretação do coeficiente de correlação linear

Na prática, considera-se como **alta correlação** entre as variáveis **X** e **Y**, quando **|r| ≥ 0,8**.

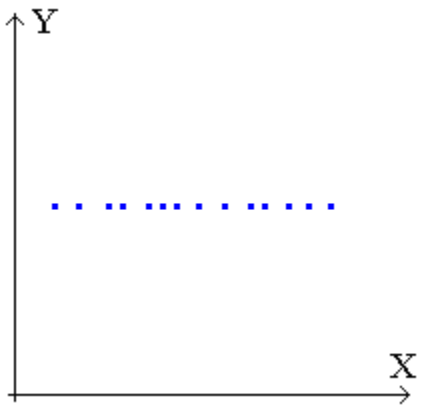


Valores de **r** dentro da elipse indicam **correlação linear de moderada a fraca, na medida em que se aproximam de zero**

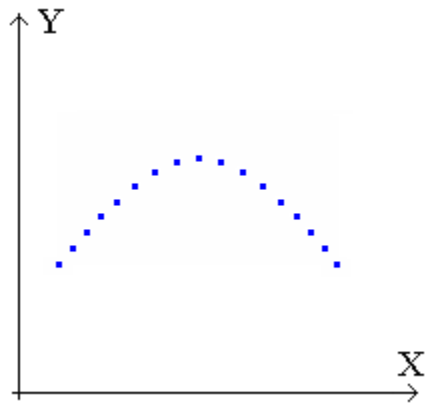
Interpretação do coeficiente de correlação linear



r = 0



r = 0



r = 0

Não tem sentido quando a relação é não linear

Importante:

⇨ Coeficiente de correlação linear igual a zero não indica que as variáveis são independentes, mas que **não existe relação linear** entre elas.

⇨ O coeficiente de correlação não indica, necessariamente, relação de causa e efeito entre as variáveis consideradas e, sim, que as duas variáveis variam conjuntamente.

Correlação e causalidade

- O coeficiente de correlação **não mede** a relação de **causa e efeito** entre as variáveis, ainda que essa relação possa estar presente.

  Exemplo: Existe forte correlação positiva entre as vendas anuais de chicletes e a taxa de criminalidade nos EUA.

- Mas não podemos concluir que há relação de causa e efeito entre as variáveis.

- O que se observa é que as duas variáveis são dependentes do **tamanho da população**, e é essa relação mútua com a terceira variável (tamanho da população) que produz a correlação forte e positiva entre a venda de chicletes e a incidência de crimes nos EUA.

Exemplo:

Um engenheiro está estudando a bacia do rio São Francisco. Um dos objetivos da pesquisa é verificar se existe correlação entre as variáveis área de drenagem e vazão média de longo termo, observadas em 22 estações fluviométricas do alto rio. Os valores observados foram os seguintes:

Tabela 1. Área de drenagem e vazão média (Q) de 22 estações fluviométricas da bacia do alto rio São Francisco.

Estação	Área	Q (m^3/s)	Estação	Área	Q (m^3/s)
1	83,9	1,32	12	3727,4	65,30
2	188,3	2,29	13	4142,9	75,00
3	279,4	4,24	14	4874,2	77,20
4	481,3	7,34	15	5235,0	77,50
5	675,7	8,17	16	5414,2	86,80
6	769,7	8,49	17	5680,4	85,70
7	875,8	18,90	18	8734,0	128,00
8	964,2	18,30	19	10191,5	152,00
9	1206,9	19,30	20	13881,8	224,00
10	1743,5	34,20	21	14180,1	241,00
11	2242,4	40,90	22	29366,2	455,00

Calcule para esses dados o coeficiente de correlação linear de Pearson (r_{xy}).

No exemplo: Tabela auxiliar

i	x	y	x^2	y^2	xy
1	83,9	1,3	7039,2	1,74	110,75
2	188,3	2,3	35456,9	5,24	431,21
3	279,4	4,2	78064,4	17,98	1184,66
4	481,3	7,3	231649,7	53,88	3532,74
5	675,7	8,2	456570,5	66,75	5520,47
6	769,7	8,5	592438,1	72,08	6534,75
7	875,8	18,9	767025,6	357,21	16552,62
8	964,2	18,3	929681,6	334,89	17644,86
9	1206,9	19,3	1456607,6	372,49	23293,17
10	1743,5	34,2	3039792,3	1169,64	59627,70
11	2242,4	40,9	5028357,8	1672,81	91714,16
12	3727,4	65,3	13893510,8	4264,09	243399,22
13	4142,9	75,0	17163620,4	5625,00	310717,50
14	4874,2	77,2	23757825,6	5959,84	376288,24
15	5235,0	77,5	27405225,0	6006,25	405712,50
16	5414,2	86,8	29313561,6	7534,24	469952,56
17	5680,4	85,7	32266944,2	7344,49	486810,28
18	8734,0	128,0	76282756,0	16384,00	1117952,00
19	10191,5	152,0	103866672,3	23104,00	1549108,00
20	13881,8	224,0	192704371,2	50176,00	3109523,20
21	14180,1	241,0	201075236,0	58081,00	3417404,10
22	29366,2	455,0	862373702,4	207025,00	13361621,00
$\sum$	**114938,8**	**1831,0**	**1592726109,2**	**395628,62**	**25074635,69**

Exemplo: n=22 $\bar{x} = 5224{,}49$ $\bar{y} = 83{,}22$

$$\sum x_i^2 = 1592726109{,}2 \qquad \sum y_i^2 = 395628{,}62 \qquad \sum x_i y_i = 25074635{,}69$$

$$SPXY = \sum x_i y_i - n\bar{x}\bar{y} = 25074635{,}69 - 22 \times 5224{,}49 \times 83{,}22 = 15509115{,}28$$

$$SQX = \sum x_i^2 - n\bar{x}^2 = 1592726109{,}2 - 22 \times 5224{,}49^2 = 992229393{,}5$$

$$SQY = \sum y_i^2 - n\bar{y}^2 = 395628{,}62 - 22 \times 83{,}22^2 = 243256{,}13$$

$$r_{XY} = \frac{SPXY}{\sqrt{SQX.SQY}} = \frac{15509115{,}28}{\sqrt{992229393{,}5 \times 243256{,}13}} = 0{,}9983$$

Interpretação: Correlação forte e positiva entre as variáveis área de drenagem e vazão média, ou seja, existe uma forte tendência de valores altos de vazão média estarem associados a valores altos de área.

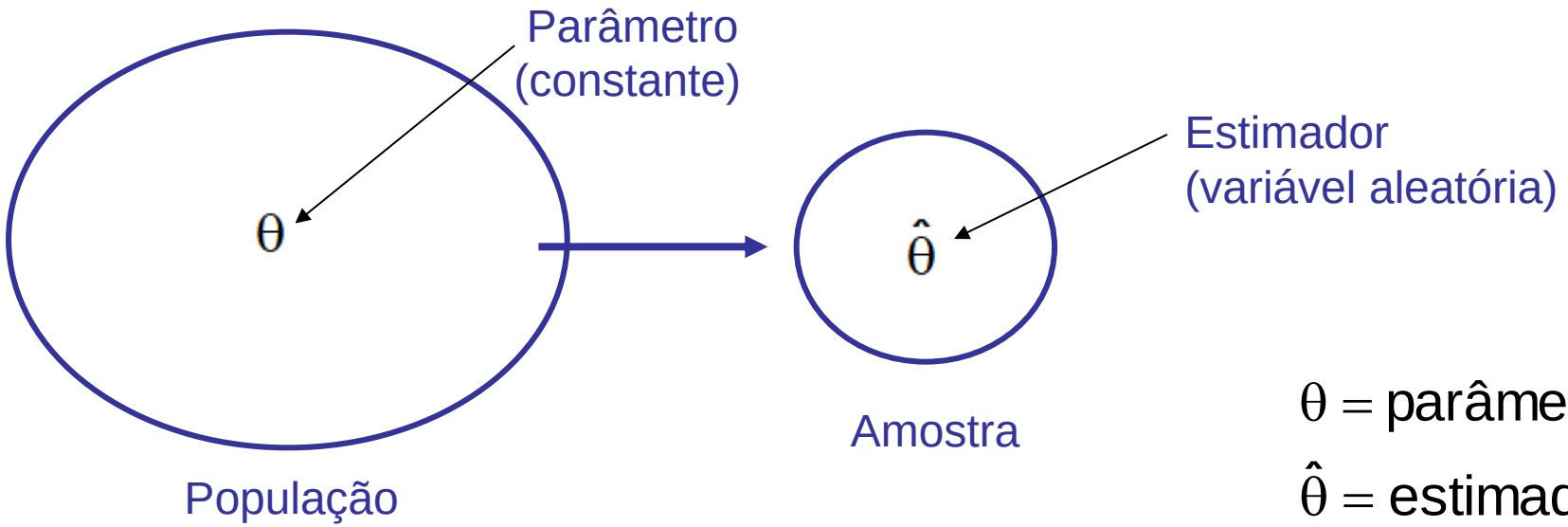
Tópico 2. Correlação linear

2.1 Medidas de associação: covariância e coeficiente de correlação linear

2.2 Estimação do coeficiente de correlação linear de Pearson

2.3 Inferências sobre o coeficiente de correlação linear de Pearson

2.2. Estimação do coeficiente de correlação



θ = parâmetro

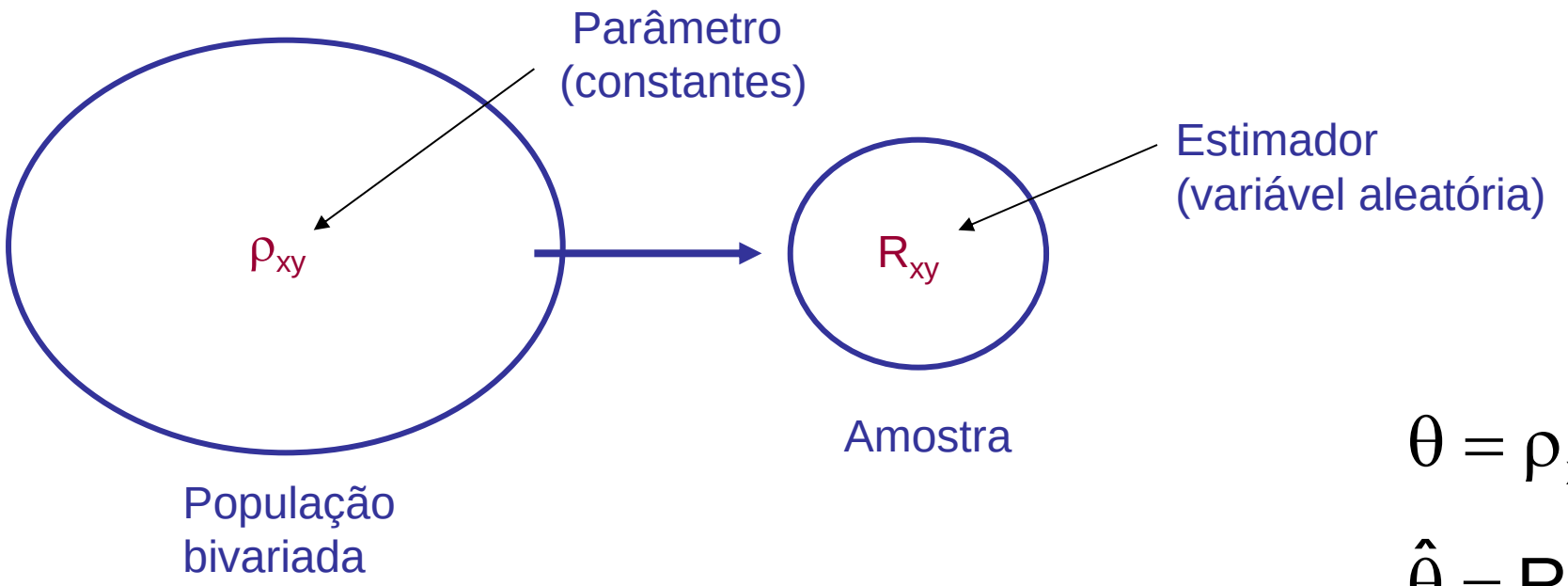
$\hat{\theta}$ = estimador

Parâmetro é um valor populacional que desejamos conhecer.

Estimador é o valor que calculamos na amostra para obter informação sobre o parâmetro. Todo estimador é uma variável aleatória, pois pode assumir diferentes valores dependendo da configuração da amostra.

Estimativa é um valor particular que o estimador assume em uma amostra.

2.2. Estimação do coeficiente de correlação

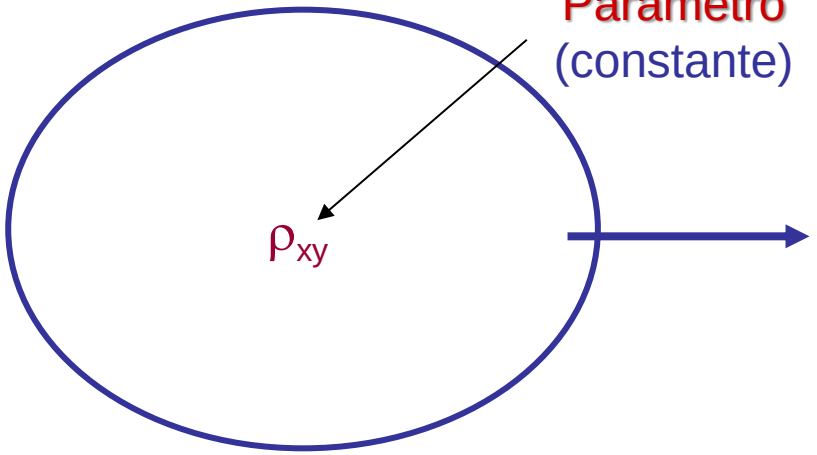


$$\theta = \rho_{xy}$$

$$\hat{\theta} = R_{xy}$$

$\rho_{xy} \rightarrow$ Coeficiente de correlação linear populacional

2.2. Estimação do coeficiente de correlação



Amostra aleatória

$$\{(X_1,Y_1), (X_2,Y_2),\ldots,(X_n,Y_n)\}$$

$\downarrow$

R_{xy} ←Variável aleatória

Estimativas de ρ_{XY}

Estimador

$$R = \frac{S_{xy}}{S_x S_y}$$

r_1

r_2

r_3

...

r_k

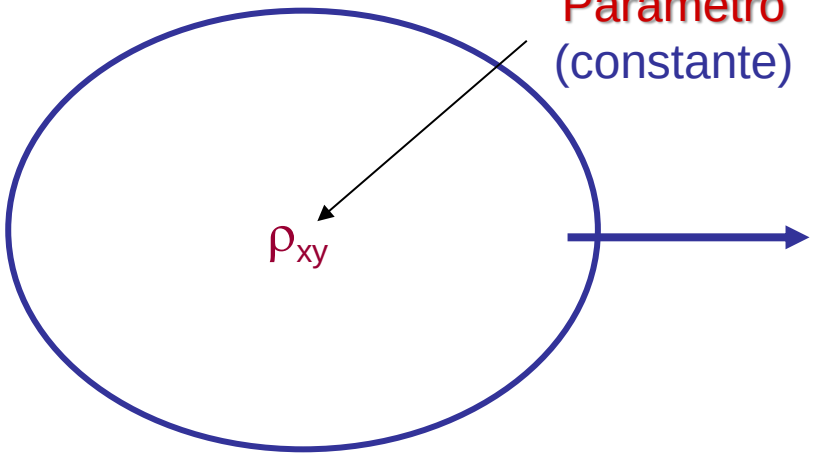
Exemplo: Um engenheiro está estudando a bacia do rio São Francisco. Um dos objetivos da pesquisa é verificar se existe correlação entre a área de drenagem e a vazão média de longo termo, observadas em 22 estações fluviométricas do alto rio. Os valores observados foram os seguintes:

Tabela 1. Área de drenagem e vazão média (Q) de 22 estações fluviométricas da bacia do alto rio São Francisco.

Estação	Área	Q (m^3/s)	Estação	Área	Q (m^3/s)
1	83,9	1,32	12	3727,4	65,30
2	188,3	2,29	13	4142,9	75,00
3	279,4	4,24	14	4874,2	77,20
4	481,3	7,34	15	5235,0	77,50
5	675,7	8,17	16	5414,2	86,80
6	769,7	8,49	17	5680,4	85,70
7	875,8	18,90	18	8734,0	128,00
8	964,2	18,30	19	10191,5	152,00
9	1206,9	19,30	20	13881,8	224,00
10	1743,5	34,20	21	14180,1	241,00
11	2242,4	40,90	22	29366,2	455,00

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = 0,9983$$

2.2. Estimação do coeficiente de correlação



Amostra aleatória

$$\{(X_1,Y_1), (X_2,Y_2),\ldots,(X_n,Y_n)\}$$

$\downarrow$

R_{xy} ←Variável aleatória

Estimativas de ρ_{XY}

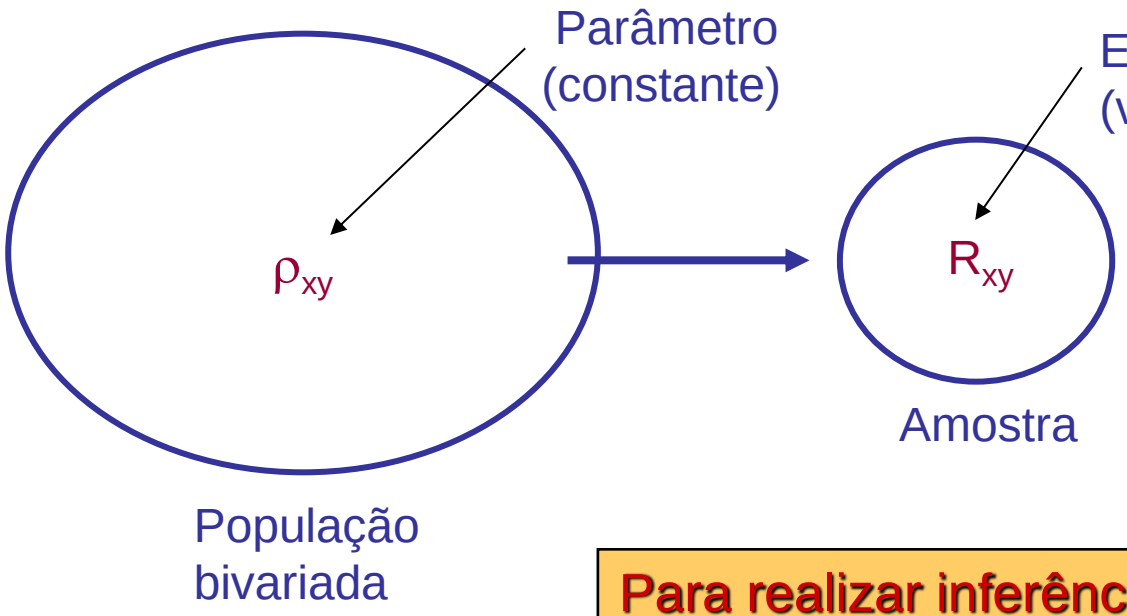
Estimador

$$R = \frac{S_{xy}}{S_x S_y}$$

r_1, r_2, r_3, ..., r_k

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = 0{,}9983$$

2.2. Estimação do coeficiente de correlação



Como o estimador R_{xy} é uma variável aleatória ele tem uma distribuição de probabilidade, chamada distribuição amostral.

Para realizar inferências sobre o parâmetro ρ_{xy} é necessário conhecer a distribuição amostral do estimador R_{xy}

Métodos de inferência

- Testes de hipótese para ρ_{xy} → $H_0: \rho_{xy} = 0$
- Intervalo de confiança para ρ_{xy} → $LI < \rho_{xy} < LS$

2.3 Inferências sobre o coeficiente de correlação

Caso geral: Dada a amostra aleatória da população estatística bivariada de (X,Y), correspondente a n unidades de observação e representada por (x_1,y_1), (x_2,y_2),...,(x_n,y_n), o coeficiente de correlação ρ_{xy} é usualmente estimado por

$$R_{xy} = \frac{S_{xy}}{S_x\, S_y} \rightarrow \text{V.A.} \begin{cases} E(R_{xy}) \\ V(R_{xy}) \end{cases}$$

Se a amostra é aleatória, independentemente da distribuição de probabilidade de (X,Y), o coeficiente R_{xy} tem propriedades interessantes como estimador de ρ_{xy}:

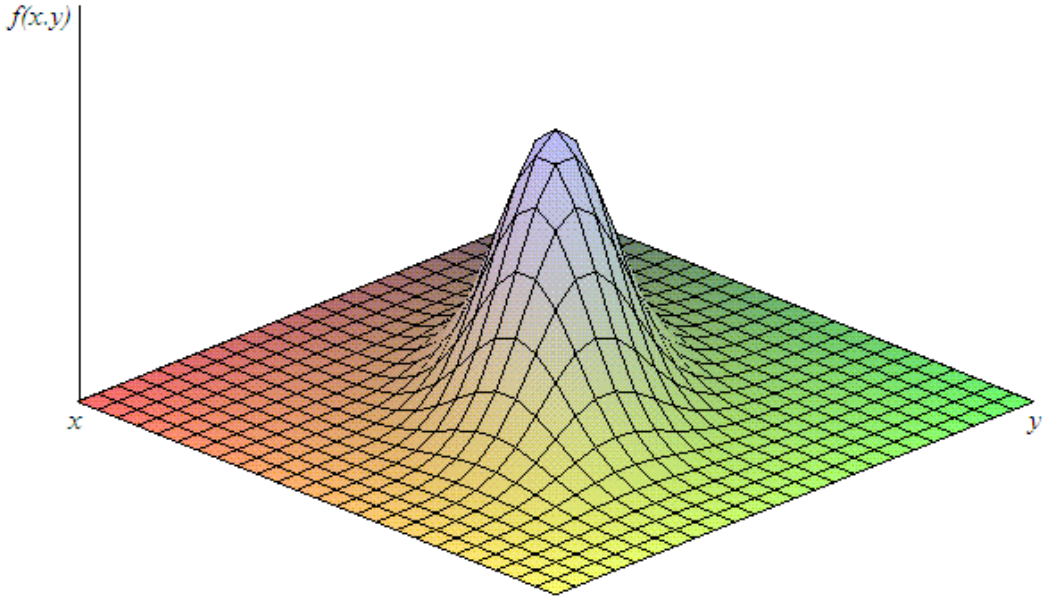
1. Imparcialidade, pois $E(R_{xy}) \cong \rho_{xy}$
2. $-1 \leq R_{xy} \leq 1$
3. R_{xy} é um estimador consistente para ρ_{xy}

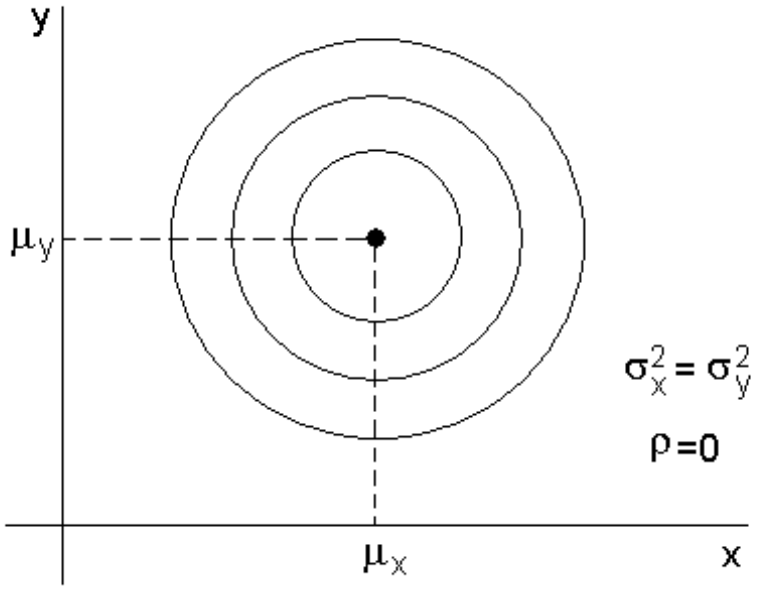
$$\hat{\theta} \xrightarrow{n \rightarrow N} \theta$$

2.3 Inferências sobre o coeficiente de correlação

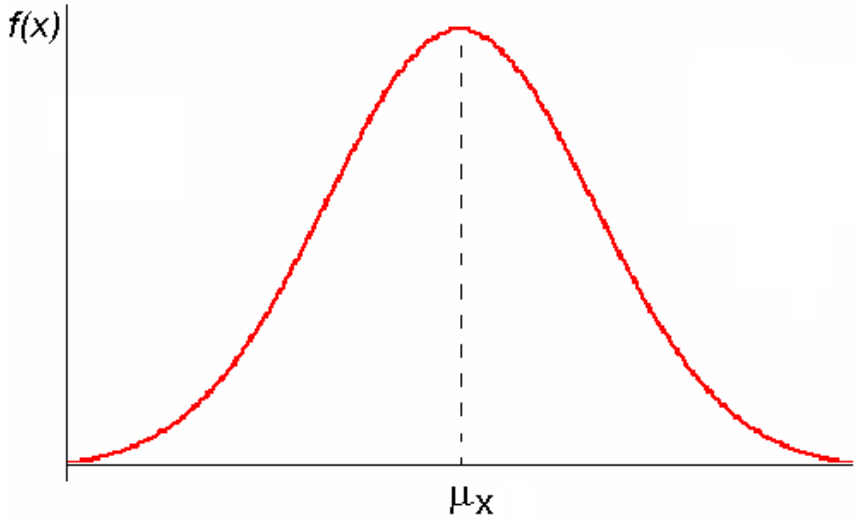
Caso normal: a distribuição conjunta de (X,Y) é a distribuição normal bivariada.

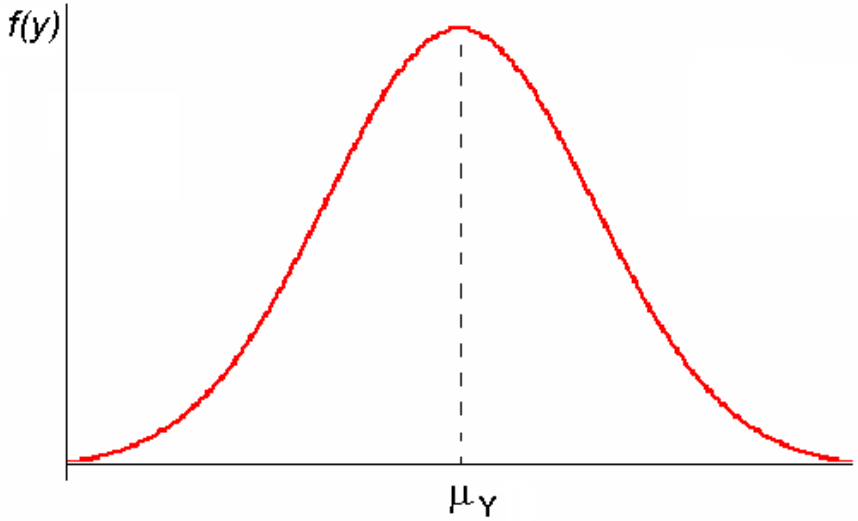
Nesse caso, X e Y são variáveis aleatórias independentes se, e somente se, $\rho_{xy}=0$.





Função densidade de probabilidade da normal individual para X e Y





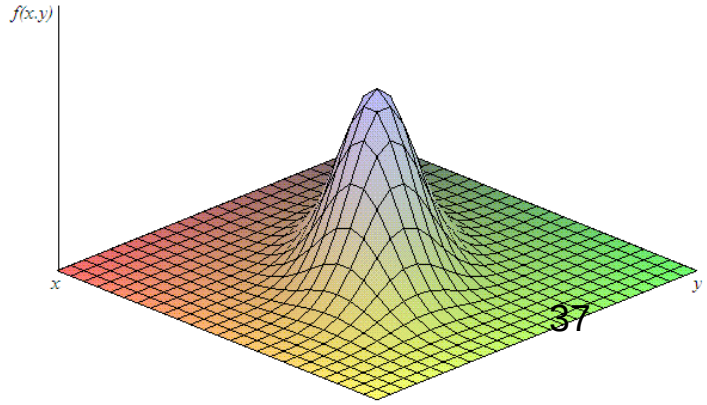
$$f(x) = \frac{1}{2\pi\sqrt{\sigma_x}} e^{-\frac{1}{2}\left(\frac{x-\mu_x}{\sqrt{\sigma_x}}\right)^2}$$

$$f(y) = \frac{1}{2\pi\sqrt{\sigma_y}} e^{-\frac{1}{2}\left(\frac{y-\mu_y}{\sqrt{\sigma_y}}\right)^2}$$

Parâmetros da normal: μ e σ

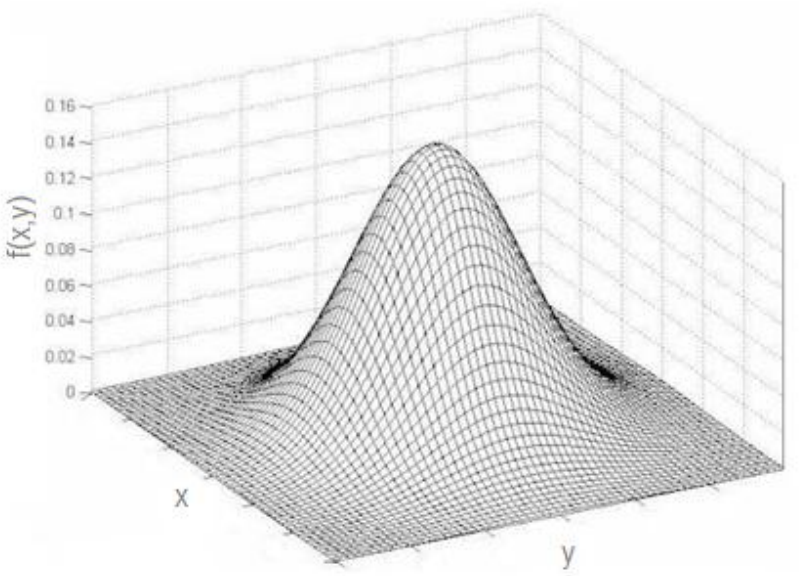
Função densidade de probabilidade da normal bivariada

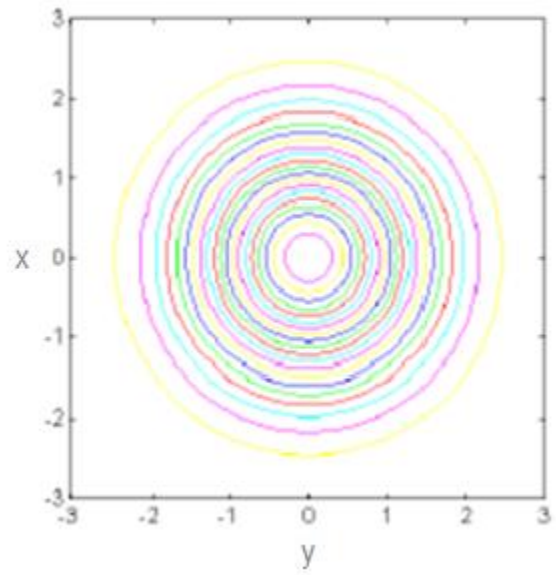
$$f(x,y) = \frac{1}{2\pi\sqrt{\sigma_x\sigma_y\left(1-\rho_{xy}^2\right)}} e^{\left\{-\frac{1}{2\left(1-\rho_{xy}^2\right)}\left[\left(\frac{x-\mu_x}{\sqrt{\sigma_x}}\right)^2+\left(\frac{y-\mu_y}{\sqrt{\sigma_y}}\right)^2-2\rho_{xy}\left(\frac{x-\mu_x}{\sqrt{\sigma_x}}\right)\left(\frac{y-\mu_y}{\sqrt{\sigma_y}}\right)\right]\right\}}$$



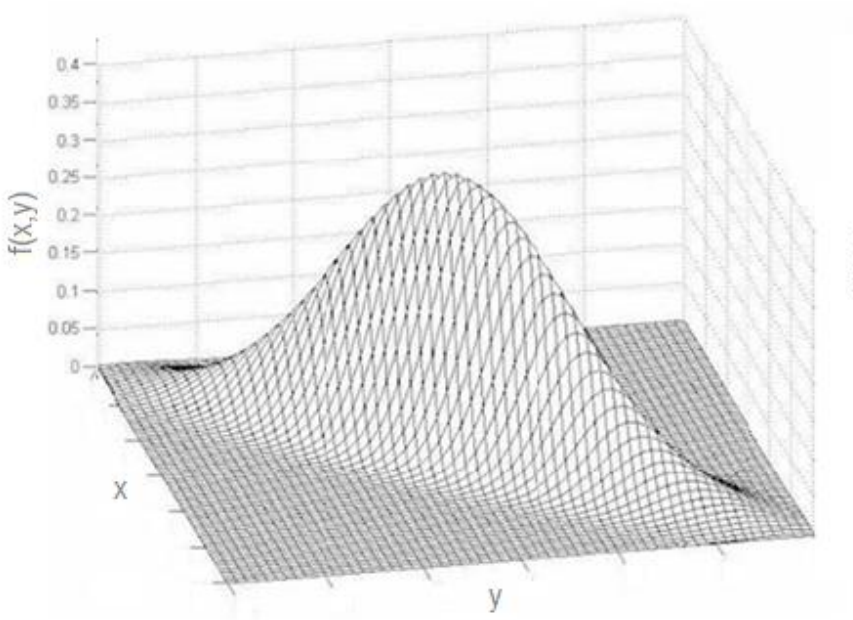
Parâmetros da normal bivariada: μ_x, σ_x, μ_y, σ_y e ρ_{xy}

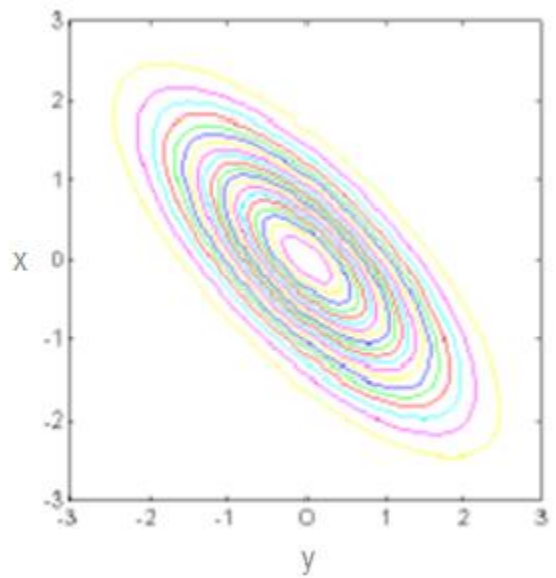
Distribuição bivariada com ρ= 0



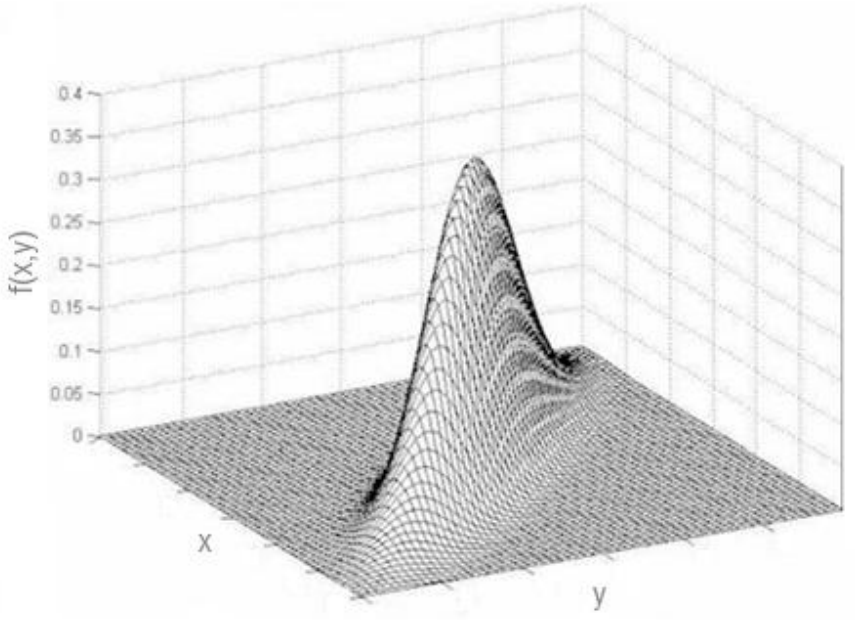


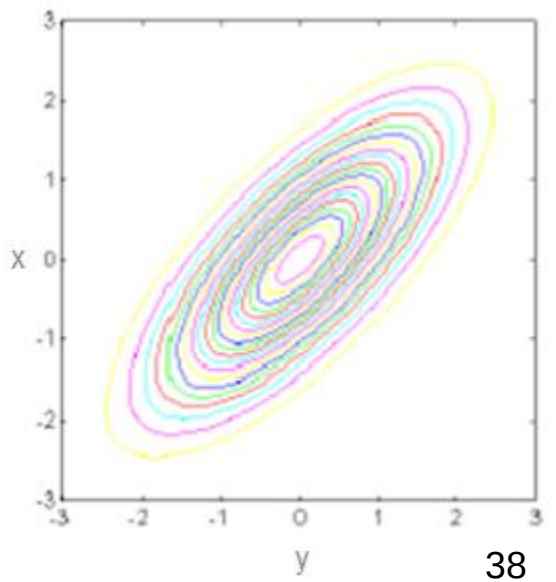
Distribuição bivariada com ρ= -0,9



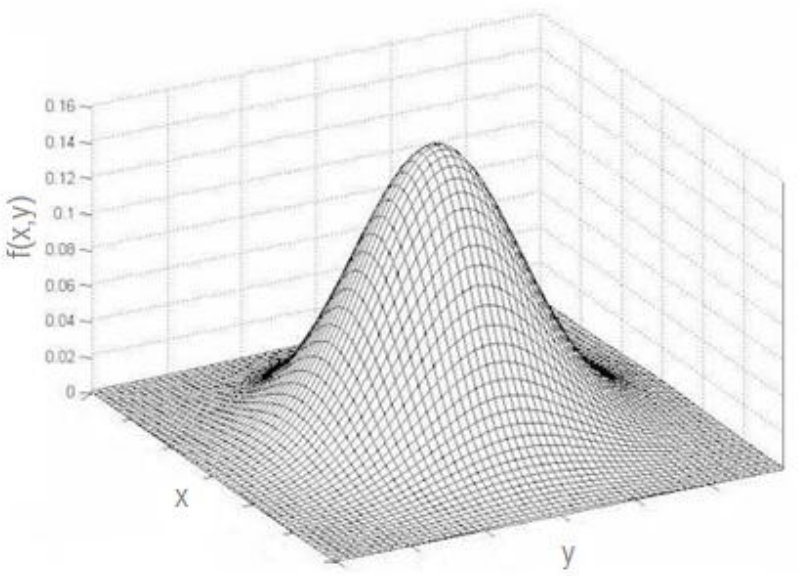


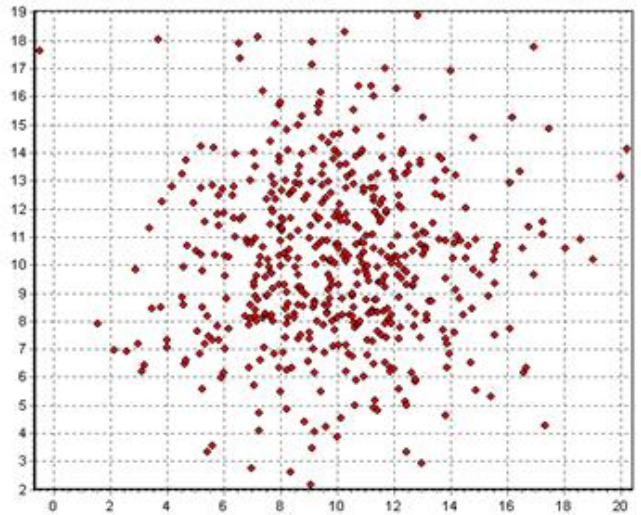
Distribuição bivariada com ρ=0,9





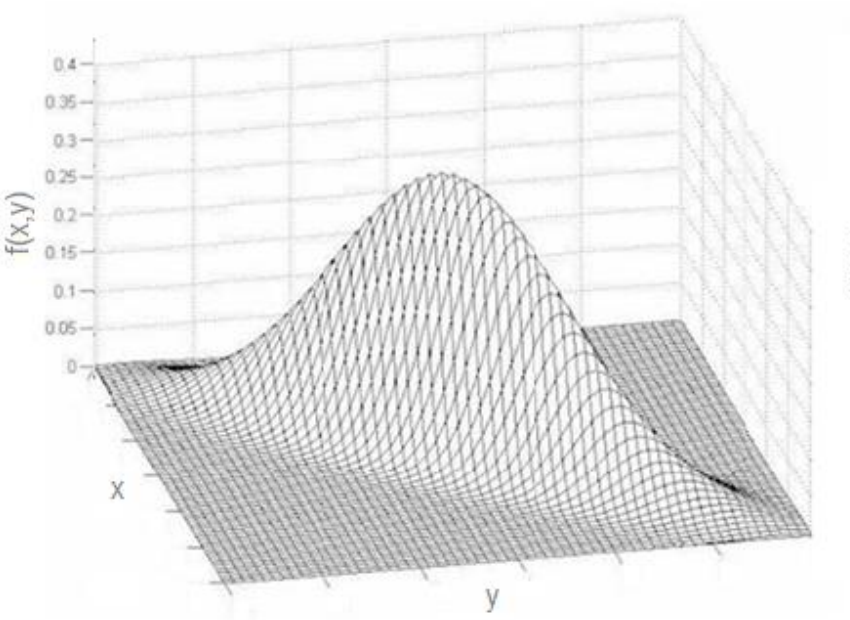
Distribuição bivariada com ρ= 0

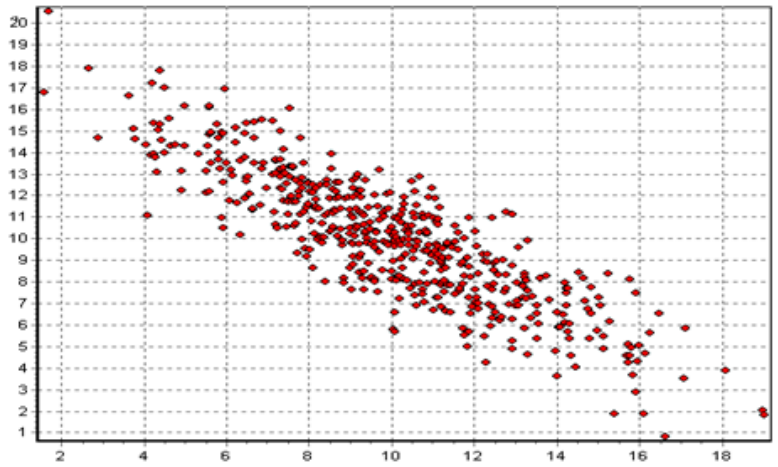




r = 0

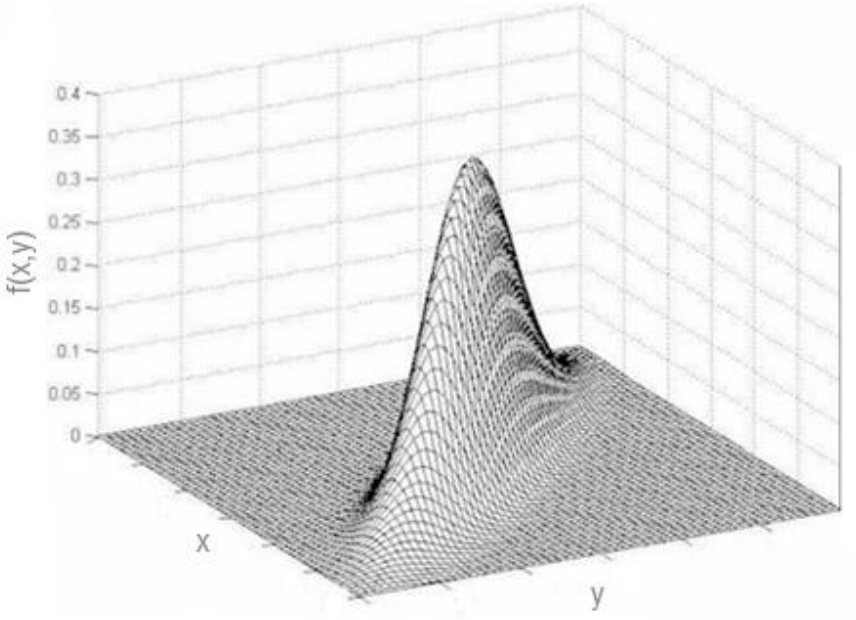
Distribuição bivariada com ρ=-0,9

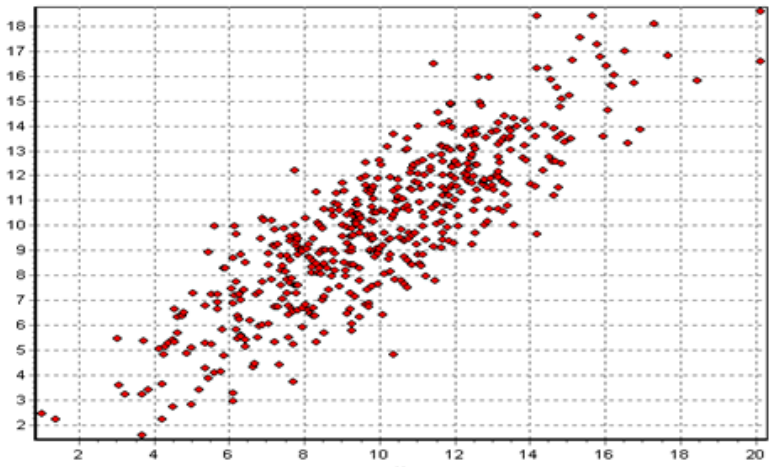




r < 0

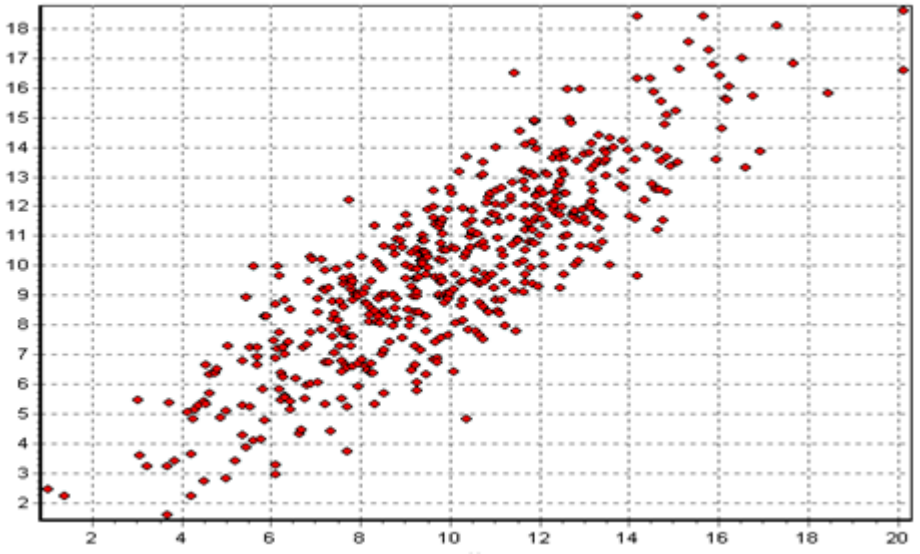
Distribuição bivariada com ρ=0,9



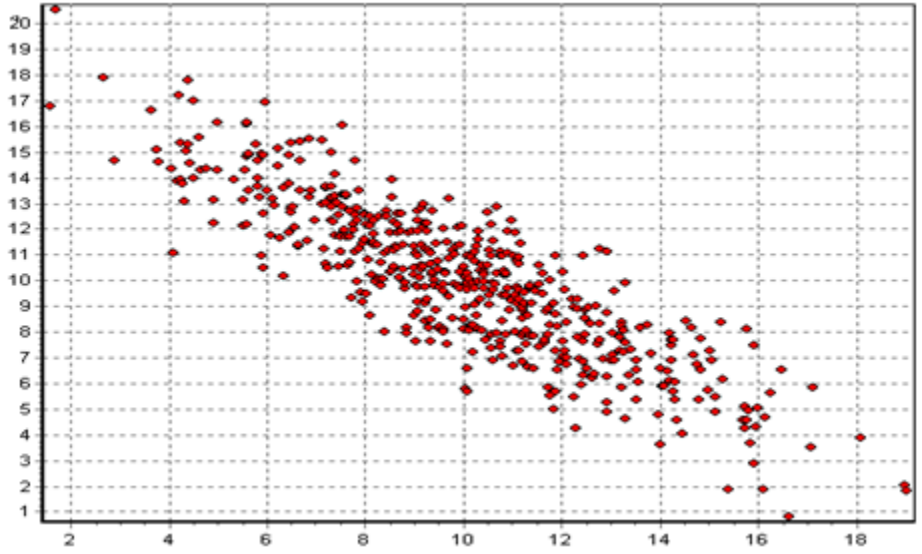


r > 0

Casos em que existe associação linear simples

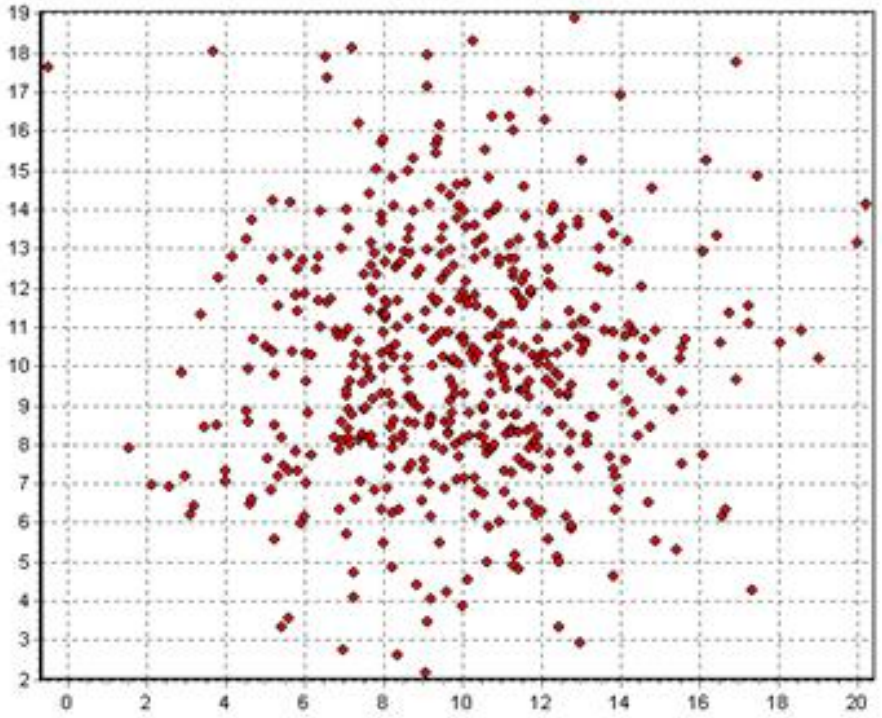


$\rho > 0$

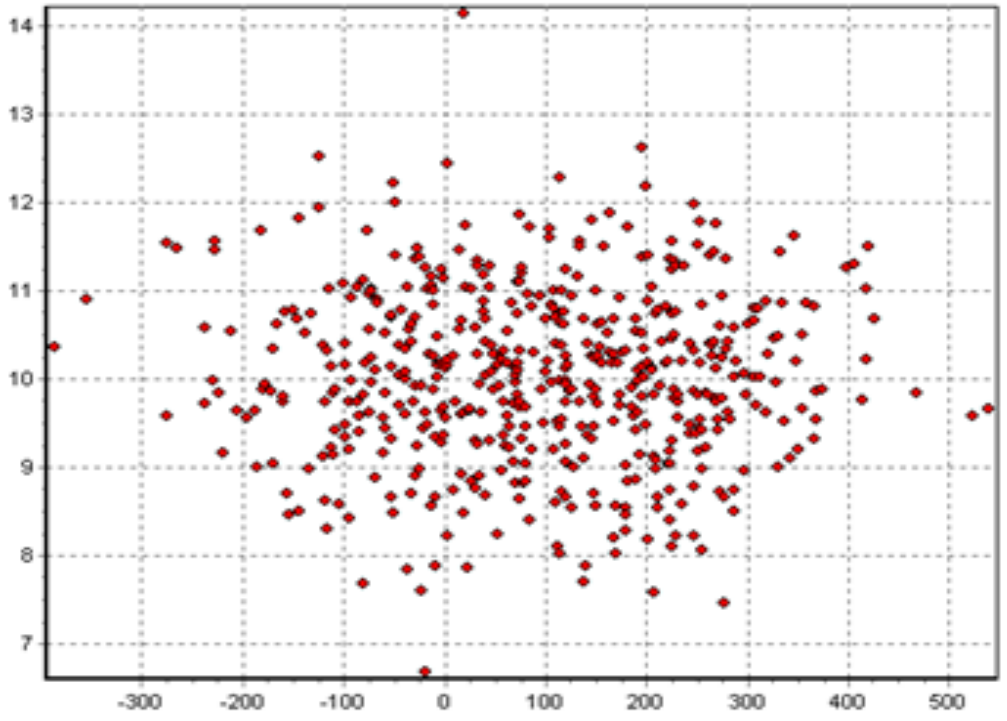


$\rho < 0$

Casos em que não existe associação linear entre as variáveis



$\rho = 0,\ \sigma_x = \sigma_y$



$\rho = 0,\ \sigma_x > \sigma_y$

2.3 Inferências sobre o coeficiente de correlação

Vamos considerar duas situações do caso normal: $\rho_{xy} = 0$ e $\rho_{xy} \neq 0$.

Função densidade de probabilidade da normal bivariada

Se $\rho_{xy} = 0 \rightarrow$

$$f(x,y) = \frac{1}{2\pi\sqrt{\sigma_x \sigma_y}} e^{\left\{-\frac{1}{2}\left[\left(\frac{x-\mu_x}{\sqrt{\sigma_x}}\right)^2 + \left(\frac{y-\mu_y}{\sqrt{\sigma_y}}\right)^2\right]\right\}}$$

Se $\rho_{xy} \neq 0 \rightarrow$

$$f(x,y) = \frac{1}{2\pi\sqrt{\sigma_x \sigma_y \left(1-\rho_{xy}^2\right)}} e^{\left\{-\frac{1}{2\left(1-\rho_{xy}^2\right)}\left[\left(\frac{x-\mu_x}{\sqrt{\sigma_x}}\right)^2 + \left(\frac{y-\mu_y}{\sqrt{\sigma_y}}\right)^2 - 2\rho_{xy}\left(\frac{x-\mu_x}{\sqrt{\sigma_x}}\right)\left(\frac{y-\mu_y}{\sqrt{\sigma_y}}\right)\right]\right\}}$$

2.3 Inferências sobre o coeficiente de correlação

Vamos considerar duas situações: $\rho_{xy} = 0$ e $\rho_{xy} \neq 0$.

Situação 1: Caso de independência entre X e Y ⇨ $\rho_{xy} = 0$

A distribuição amostral do coeficiente de correlação R_{xy} foi obtida por Fisher em 1915 e tem forma relativamente simples se $\rho_{xy}=0$.

Nessa situação, tem-se também que a variavel T, definida por

$$T = \frac{\hat{\theta} - \theta}{S(\hat{\theta})} = \frac{R_{xy} - \rho_{xy}}{S(R_{xy})} = \frac{R_{xy}}{\sqrt{\frac{1 - R_{xy}^2}{n-2}}},$$

← Estimador

← Desvio padrão do estimador

tem distribuição t de Student com n-2 graus de liberdade.

Esta variável pode ser utilizada para testar a hipótese de nulidade:

$$H_0 : \rho_{xy} = 0$$

2.3 Inferências sobre o coeficiente de correlação

Situação 2: Caso de dependência entre X e Y ⇨ $\rho_{xy} \neq 0$

Nesta situação, a forma da distribuição do coeficiente R_{xy} é muito complicada, o que levou Fisher a propor uma aproximação normal.
Assim, para tamanho de amostra razoável (≥25), a variável aleatória

$$W = \frac{1}{2}\ln\left(\frac{1+R_{xy}}{1-R_{xy}}\right) \qquad W \sim N(\mu_W; \sigma_W)$$

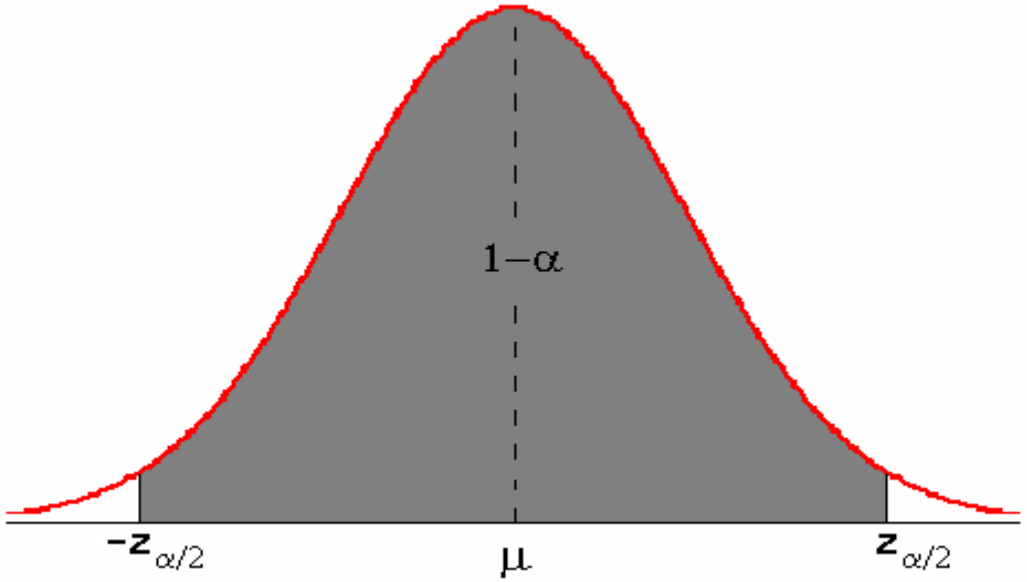
tem distribuição aproximadamente normal, onde:

$$\mu_W = \frac{1}{2}\left[\ln\left(\frac{1+\rho_{xy}}{1-\rho_{xy}}\right) + \frac{\rho_{xy}}{n-1}\right] \quad \text{e} \quad \sigma_W = \frac{1}{\sqrt{n-3}}$$

Deste modo, para construir o intervalo de confiança para ρ_{xy}, ao nível de confiança 1- α, podemos usar como pivô a variável Z, que tem distribuição normal padrão.

$$Z = \frac{W - \mu_W}{\sigma_W}$$

Z ~ N (0, 1)



$$P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) = 1 - \alpha$$

Inicialmente substituímos Z pela expressão $\frac{W - \mu_W}{\sigma_W}$

$$P\left(-z_{\alpha/2} \leq \frac{W - \mu_W}{\sigma_W} \leq z_{\alpha/2}\right) = 1 - \alpha$$

Isolando-se μ_w nesta expressão, obtém-se:

$$P(W - z_{\alpha/2}\sigma_w \leq \mu_w \leq W + z_{\alpha/2}\sigma_w) = 1 - \alpha$$

$$W = \frac{1}{2}\ln\left(\frac{1+R_{xy}}{1-R_{xy}}\right) \qquad \mu_W = \frac{1}{2}\left[\ln\left(\frac{1+\rho_{xy}}{1-\rho_{xy}}\right) + \frac{\rho_{xy}}{n-1}\right] \qquad \sigma_W = \frac{1}{\sqrt{n-3}}$$

Fazendo as substituições e considerando $\frac{\rho_{xy}}{n-1}$ desprezível, temos

$$P\left(\frac{1}{2}\ln\left(\frac{1+R_{xy}}{1-R_{xy}}\right) - \frac{z_{\alpha/2}}{\sqrt{n-3}} \leq \frac{1}{2}\ln\left(\frac{1+\rho_{xy}}{1-\rho_{xy}}\right) \leq \frac{1}{2}\ln\left(\frac{1+R_{xy}}{1-R_{xy}}\right) + \frac{z_{\alpha/2}}{\sqrt{n-3}}\right) = 1-\alpha$$

Para obter os limites do intervalo de confiança, é necessário resolver as inequações para ρ_{xy}.

$$P\left[\frac{\exp\left\{2\left[\frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) - \frac{z_{\alpha/2}}{\sqrt{n-3}}\right]\right\}-1}{\exp\left\{2\left[\frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) - \frac{z_{\alpha/2}}{\sqrt{n-3}}\right]\right\}+1} \leq \rho_{xy} \leq \frac{\exp\left\{2\left[\frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) + \frac{z_{\alpha/2}}{\sqrt{n-3}}\right]\right\}-1}{\exp\left\{2\left[\frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) + \frac{z_{\alpha/2}}{\sqrt{n-3}}\right]\right\}+1}\right] = 1-\alpha$$

Limite inferior

Limite superior

O intervalo de confiança para ρ_{xy} também pode ser escrito da seguinte forma:

$$IC(\rho;1-\alpha): \left[\frac{\exp\left\{2\left[\frac{1}{2}\ln\left(\frac{1+r}{1-r}\right)-\frac{z_{\alpha/2}}{\sqrt{n-3}}\right]\right\}-1}{\exp\left\{2\left[\frac{1}{2}\ln\left(\frac{1+r}{1-r}\right)-\frac{z_{\alpha/2}}{\sqrt{n-3}}\right]\right\}+1}; \frac{\exp\left\{2\left[\frac{1}{2}\ln\left(\frac{1+r}{1-r}\right)+\frac{z_{\alpha/2}}{\sqrt{n-3}}\right]\right\}-1}{\exp\left\{2\left[\frac{1}{2}\ln\left(\frac{1+r}{1-r}\right)+\frac{z_{\alpha/2}}{\sqrt{n-3}}\right]\right\}+1} \right]$$

Limite inferior — Limite superior

onde $z_{\alpha/2}$ é o valor de z que delimita a área $\alpha/2$.

Obs.: Essa transformação não é válida para $\rho=-1$ ou $\rho=+1$ e para $\rho=0$. Nesses casos, a distribuição exata t deve ser preferida.

Exemplo: Um engenheiro está estudando a bacia do rio São Francisco. Um dos objetivos da pesquisa é verificar se existe correlação entre a área de drenagem e a vazão média de longo termo, observadas em 22 estações fluviométricas do alto rio. Os valores observados foram os seguintes:

Tabela 1. Área de drenagem e vazão média (Q) de 22 estações fluviométricas da bacia do alto rio São Francisco.

Estação	Área	Q (m^3/s)	Estação	Área	Q (m^3/s)
1	83,9	1,32	12	3727,4	65,30
2	188,3	2,29	13	4142,9	75,00
3	279,4	4,24	14	4874,2	77,20
4	481,3	7,34	15	5235,0	77,50
5	675,7	8,17	16	5414,2	86,80
6	769,7	8,49	17	5680,4	85,70
7	875,8	18,90	18	8734,0	128,00
8	964,2	18,30	19	10191,5	152,00
9	1206,9	19,30	20	13881,8	224,00
10	1743,5	34,20	21	14180,1	241,00
11	2242,4	40,90	22	29366,2	455,00

a) Verifique se a correlação entre as variáveis x e y é significativa, ou seja, teste a hipótese de que $\rho=0$.

b) Construa o intervalo de confiança, ao nível de 95%, para o coeficiente de correlação linear populacional entre as variáveis área e vazão.

Resolução: a) Teste de hipótese

1. Pressuposições:

- Distribuição da variável (X,Y) é normal bivariada;

2. Hipóteses estatísticas:

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

3. Erro de conclusão: $\alpha = 0{,}05$ (probabilidade de erro tipo I)

Erro tipo I = rejeitar H_0 quando ela é verdadeira.

4. Estatística do teste:

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = 0{,}9983$$

$$T = \frac{R_{xy}}{\sqrt{\frac{1 - R_{xy}^2}{n-2}}} \rightarrow t = \frac{0{,}9983}{\sqrt{\frac{1 - 0{,}9983^2}{22-2}}} = 75{,}99$$

5. Decisão e conclusão:

$\nu = 20 \rightarrow t_{\alpha/2(20)}$

Tabela II. Limites da distribuição t de Student.

Graus de Liberdade (ν)	**Limites bilaterais**: P(|t| > $t_{\alpha/2}$)							
	Nível de Significância (α)							
	0,50	0,20	0,10	0,05	0,025	0,02	0,01	0,005
1	1,000	3,078	6,314	12,706	25,542	31,821	63,657	127,320
2	0,816	1,886	2,920	4,303	6,205	6,965	9,925	14,089
3	0,715	1,638	2,353	3,183	4,177	4,541	5,841	7,453
4	0,741	1,533	2,132	2,776	3,495	3,747	4,604	5,598
5	0,727	1,476	2,015	2,571	3,163	3,365	4,032	4,773
6	0,718	1,440	1,943	2,447	2,969	3,143	3,707	4,317
7	0,711	1,415	1,895	2,365	2,841	2,998	3,500	4,029
8	0,706	1,397	1,860	2,306	2,752	2,896	3,355	3,833
9	0,703	1,383	1,833	2,262	2,685	2,821	3,250	3,690
10	0,700	1,372	1,813	2,228	2,634	2,764	3,169	3,581
11	0,697	1,363	1,796	2,201	2,503	2,718	3,106	3,497
12	0,695	1,356	1,782	2,179	2,560	2,681	3,055	3,428
13	0,694	1,350	1,771	2,160	2,533	2,650	3,012	3,373
14	0,692	1,345	1,761	2,145	2,510	2,624	2,977	3,326
15	0,691	1,341	1,753	2,132	2,490	2,602	2,947	3,286
16	0,690	1,337	1,746	2,120	2,473	2,583	2,921	3,252
17	0,689	1,333	1,740	2,110	2,458	2,567	2,898	3,223
18	0,688	1,330	1,734	2,101	2,445	2,552	2,878	3,197
19	0,688	1,328	1,729	2,093	2,433	2,539	2,861	3,174
20	0,687	1,325	1,725	2,086	2,423	2,528	2,845	3,153
21	0,686	1,323	1,721	2,080	2,414	2,518	2,831	3,135
22	0,686	1,321	1,717	2,074	2,406	2,508	2,819	3,119
23	0,685	1,319	1,714	2,069	2,398	2,500	2,807	3,104
24	0,685	1,318	1,711	2,064	2,391	2,492	2,797	3,091
25	0,684	1,316	1,708	2,060	2,385	2,485	2,787	3,078
26	0,684	1,315	1,706	2,056	2,379	2,479	2,779	3,067
27	0,684	1,314	1,703	2,052	2,373	2,473	2,771	3,057
28	0,683	1,313	1,701	2,048	2,369	2,467	2,763	3,047
29	0,683	1,311	1,699	2,045	2,364	2,462	2,756	3,038
30	0,683	1,310	1,697	2,042	2,360	2,457	2,750	3,030
40	0,681	1,303	1,684	2,021	2,329	2,423	2,705	2,971
60	0,679	1,296	1,671	2,000	2,299	2,390	2,660	2,915
120	0,677	1,289	1,658	1,980	2,270	2,358	2,617	2,860
...	0,674	1,282	1,645	1,960	2,241	2,326	2,576	2,807
	0,25	0,10	0,05	0,025	0,0125	0,01	0,005	0,0025
Graus de Liberdade (ν)	Nível de Significância (α)							
	Limites unilaterais: P(|t| > t_{α})							

Resolução: a) Teste de hipótese

1. Pressuposições:

- Distribuição da variável (X,Y) é normal bivariada;

2. Hipóteses estatísticas:

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

3. Erro de conclusão: $\alpha = 0{,}05$ (probabilidade de erro tipo I)

Erro tipo I = rejeitar H_0 quando ela é verdadeira.

4. Estatística do teste:

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = 0{,}9983$$

$$T = \frac{R_{xy}}{\sqrt{\frac{1-R_{xy}^2}{n-2}}} \rightarrow t = \frac{0{,}9983}{\sqrt{\frac{1-0{,}9983^2}{22-2}}} = 75{,}99$$

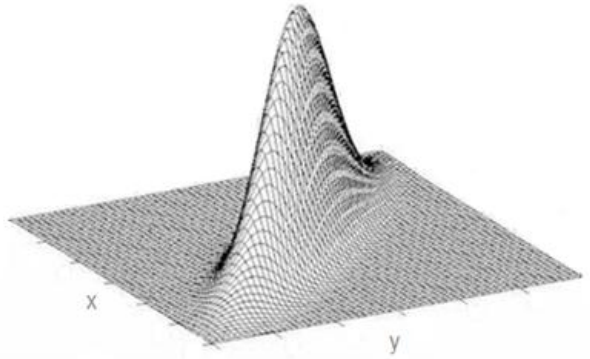
5. Decisão e conclusão:

$\nu=20 \rightarrow t_{\alpha/2(20)}=2{,}086$

$t = 75{,}99 > t_{\alpha/2(20)}=2{,}086 \rightarrow$ Rejeitamos H_0

Concluímos, ao nível de 5% de significância, que o coeficiente de correlação populacional difere de zero. Portanto, existe uma correlação linear positiva entre a área de drenagem e a vazão média do rio São Francisco. Valores altos de área estão associados a valores altos de vazão média.

Resolução: b) Intervalo de confiança



1. Pressuposições:

- Distribuição da variável (X,Y) é normal bivariada;
- As variáveis são correlacionadas, ou seja, $\rho \neq 0$.

2. Obtenção da estimativa pontual:

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = 0{,}9983$$

3. Cálculo dos limites do intervalo:

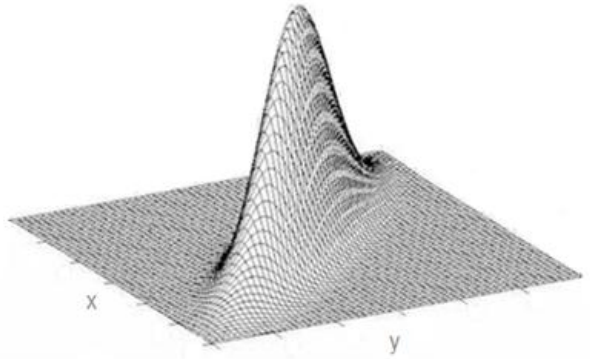
$n = 22$

$z_{\alpha/2} =$

Tabela I. Área sob a curva normal padrão de 0 a z, $P(0 \leq Z \leq z)$.

z	0	1	2	3	4	5	6	7	8	9
0,0	0,0000	0,0040	0,0080	0,0120	0,0160	0,0199	0,0239	0,0279	0,0319	0,0359
0,1	0,0398	0,0438	0,0478	0,0517	0,0557	0,0596	0,0636	0,0675	0,0714	0,0754
0,2	0,0793	0,0832	0,0871	0,0910	0,0948	0,0987	0,1026	0,1064	0,1103	0,1141
0,3	0,1179	0,1217	0,1255	0,1293	0,1331	0,1368	0,1406	0,1443	0,1480	0,1517
0,4	0,1554	0,1591	0,1628	0,1664	0,1700	0,1736	0,1772	0,1808	0,1844	0,1879
0,5	0,1915	0,1950	0,1985	0,2019	0,2054	0,2088	0,2133	0,2157	0,2190	0,2224
0,6	0,2258	0,2291	0,2324	0,2357	0,2389	0,2422	0,2454	0,2486	0,2518	0,2549
0,7	0,2580	0,2612	0,2642	0,2673	0,2704	0,2734	0,2764	0,2794	0,2823	0,2852
0,8	0,2881	0,2910	0,2939	0,2967	0,2996	0,3023	0,3051	0,3078	0,3106	0,3133
0,9	0,3159	0,3186	0,3212	0,3238	0,3264	0,3289	0,3315	0,3340	0,3365	0,3389
1,0	0,3413	0,3438	0,3461	0,3485	0,3508	0,3531	0,3554	0,3577	0,3599	0,3621
1,1	0,3643	0,3665	0,3686	0,3708	0,3729	0,3749	0,3770	0,3790	0,3810	0,3830
1,2	0,3849	0,3869	0,3888	0,3907	0,3925	0,3944	0,3962	0,3980	0,3997	0,4015
1,3	0,4032	0,4049	0,4066	0,4082	0,4099	0,4115	0,4131	0,4147	0,4162	0,4177
1,4	0,4192	0,4207	0,4222	0,4236	0,4251	0,4265	0,4279	0,4292	0,4306	0,4319
1,5	0,4332	0,4345	0,4357	0,4370	0,4382	0,4394	0,4406	0,4418	0,4429	0,4441
1,6	0,4452	0,4463	0,4474	0,4484	0,4495	0,4505	0,4515	0,4525	0,4535	0,4545
1,7	0,4554	0,4564	0,4573	0,4582	0,4591	0,4599	0,4608	0,4616	0,4625	0,4633
1,8	0,4641	0,4649	0,4656	0,4664	0,4671	0,4678	0,4686	0,4693	0,4699	0,4706
1,9	0,4713	0,4719	0,4726	0,4732	0,4738	0,4744	0,4750	0,4756	0,4761	0,4767
2,0	0,4772	0,4778	0,4783	0,4788	0,4793	0,4798	0,4803	0,4808	0,4812	0,4817
2,1	0,4821	0,4826	0,4830	0,4834	0,4838	0,4842	0,4846	0,4850	0,4854	0,4857
2,2	0,4861	0,4864	0,4868	0,4871	0,4875	0,4878	0,4881	0,4884	0,4887	0,4890
2,3	0,4893	0,4896	0,4898	0,4901	0,4904	0,4906	0,4909	0,4911	0,4913	0,4916
2,4	0,4918	0,4920	0,4922	0,4925	0,4927	0,4929	0,4931	0,4932	0,4934	0,4936
2,5	0,4938	0,4940	0,4941	0,4943	0,4945	0,4946	0,4948	0,4949	0,4951	0,4952
2,6	0,4953	0,4955	0,4956	0,4957	0,4959	0,4960	0,4961	0,4962	0,4963	0,4964
2,7	0,4965	0,4966	0,4967	0,4968	0,4969	0,4970	0,4971	0,4972	0,4973	0,4974
2,8	0,4974	0,4975	0,4976	0,4977	0,4977	0,4978	0,4979	0,4979	0,4980	0,4981
2,9	0,4981	0,4982	0,4982	0,4983	0,4984	0,4984	0,4985	0,4985	0,4986	0,4986
3,0	0,4987	0,4987	0,4987	0,4988	0,4988	0,4989	0,4989	0,4989	0,4990	0,4990
3,1	0,4990	0,4991	0,4991	0,4991	0,4992	0,4992	0,4992	0,4992	0,4993	0,4993
3,2	0,4993	0,4993	0,4994	0,4994	0,4994	0,4994	0,4994	0,4995	0,4995	0,4995
3,3	0,4995	0,4995	0,4995	0,4996	0,4996	0,4996	0,4996	0,4996	0,4996	0,4997
3,4	0,4997	0,4997	0,4997	0,4997	0,4997	0,4997	0,4997	0,4997	0,4997	0,4998
3,5	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998
3,6	0,4998	0,4998	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999
3,7	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999
3,8	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999
3,9	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000

Resolução: b) Intervalo de confiança



1. Pressuposições:

- Distribuição da variável (X,Y) é normal bivariada;
- As variáveis são correlacionadas, ou seja, $\rho \neq 0$.

2. Obtenção da estimativa pontual:

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = 0{,}9983$$

3. Cálculo dos limites do intervalo:

$$n = 22$$

$$z_{\alpha/2} = 1{,}96$$

$$IC(\rho; 0{,}95): \left[\frac{\exp\left\{2\left[\frac{1}{2}\ln\left(\frac{1+0{,}9983}{1-0{,}9983}\right) - \frac{1{,}96}{\sqrt{22-3}}\right]\right\}-1}{\exp\left\{2\left[\frac{1}{2}\ln\left(\frac{1+0{,}9983}{1-0{,}9983}\right) - \frac{1{,}96}{\sqrt{22-3}}\right]\right\}+1}; \frac{\exp\left\{2\left[\frac{1}{2}\ln\left(\frac{1+0{,}9983}{1-0{,}9983}\right) + \frac{1{,}96}{\sqrt{22-3}}\right]\right\}-1}{\exp\left\{2\left[\frac{1}{2}\ln\left(\frac{1+0{,}9983}{1-0{,}9983}\right) + \frac{1{,}96}{\sqrt{22-3}}\right]\right\}+1} \right]$$

$$IC(\rho; 0{,}95): [0{,}9958; 0{,}9993]$$

4. Conclusão: Concluímos com 95% de confiança que os limites 0,9958 e 0,9993 compreendem o coeficiente de correlação populacional.

Exercício resolvido

O proprietário de uma grande cadeia de sorveterias gostaria de estudar os efeitos da temperatura atmosférica (x) sobre as vendas (y) durante a temporada do verão. Uma amostra de 21 dias consecutivos foi selecionada, com os seguintes resultados:

i	x	y	i	x	y
1	17,22	2933,6	12	23,89	3705,6
2	21,11	3242,4	13	36,67	6562,0
3	22,78	3474,0	14	37,78	6330,4
4	23,89	3956,5	15	33,33	6118,1
5	26,67	4554,8	16	30,56	5461,9
6	27,78	4342,5	17	28,89	4979,4
7	29,44	5172,4	18	31,11	5519,8
8	31,11	5597,0	19	26,67	4361,8
9	32,22	6060,2	20	27,78	4130,2
10	32,78	5905,8	21	24,44	3821,4
11	33,33	6253,2			

A partir dos dados acima, pede-se:

a) Construir o diagrama de dispersão (X,Y).

b) Traçar uma reta para indicar a média de x e uma reta para indicar a média de y , no diagrama de dispersão, e comentar sobre a dispersão dos pontos.

c) Calcular a covariância amostral (s_{xy}). O que ela indica?

d) Calcular o coeficiente de correlação linear da amostra (r) e interpretar seu significado.

e) Teste a hipótese de que $\rho=0$.

f) Construa o intervalo de confiança, ao nível de 95%, para o coeficiente de correlação linear.

b) Traçar uma reta para indicar X e uma reta para indicar Y, no diagrama de dispersão, e comentar sobre a dispersão dos pontos;

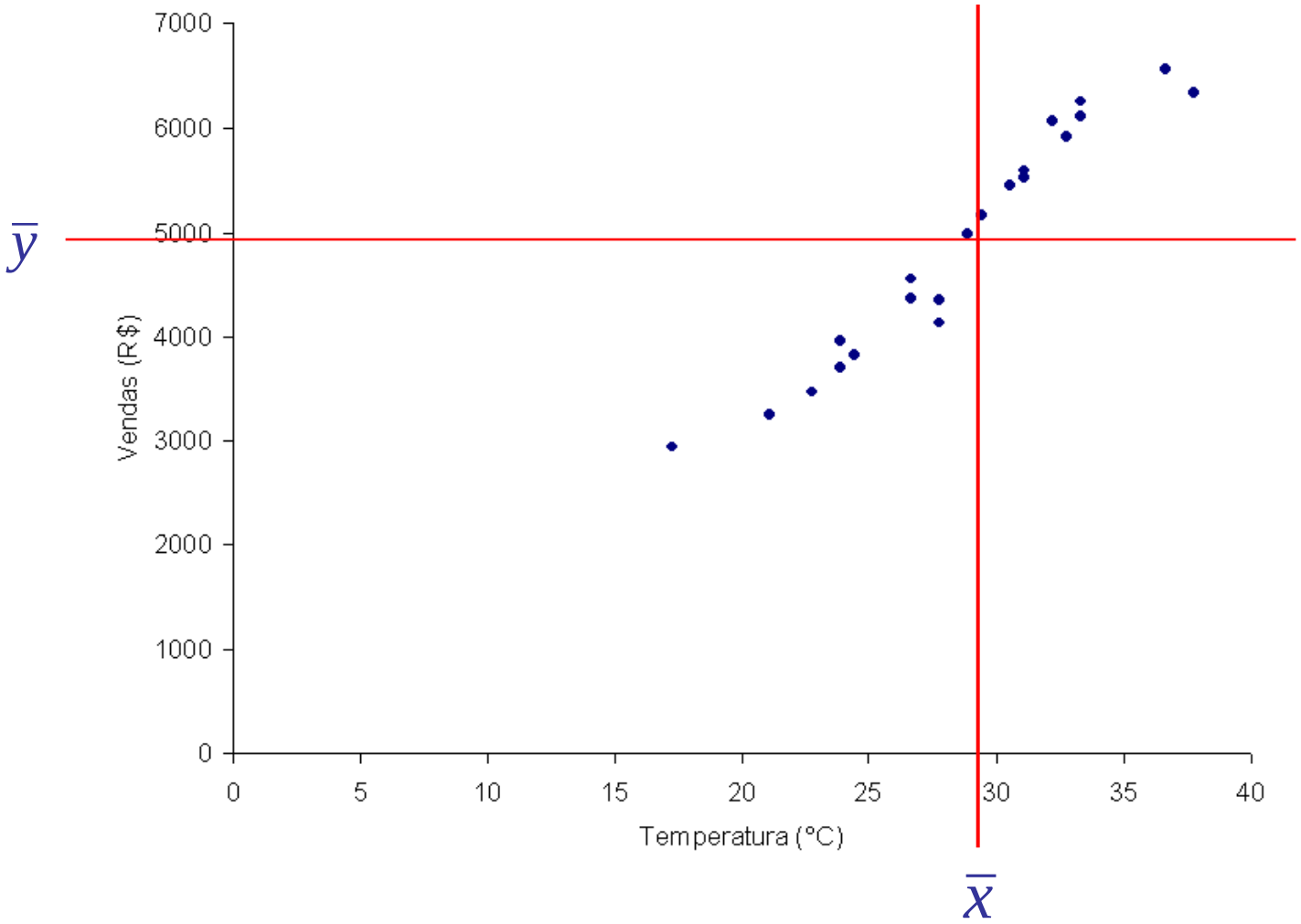


Tabela auxiliar

i	x	y	x^2	y^2	xy
1	17,22	2933,6	296,53	8606008,96	50516,59
2	21,11	3242,4	445,63	10513157,76	68447,06
3	22,78	3474,0	518,93	12068676	79137,72
4	23,89	3956,5	570,73	15653892,25	94520,79
5	26,67	4554,8	711,29	20746203,04	121476,52
6	27,78	4342,5	771,73	18857306,25	120634,65
7	29,44	5172,4	866,71	26753721,76	152275,46
8	31,11	5597,0	967,83	31326409	174122,67
9	32,22	6060,2	1038,13	36726024,04	195259,64
10	32,78	5905,8	1074,53	34878473,64	193592,12
11	33,33	6253,2	1110,89	39102510,24	208419,16
12	23,89	3705,6	570,73	13731471,36	88526,78
13	36,67	6562,0	1344,69	43059844	240628,54
14	37,78	6330,4	1427,33	40073964,16	239162,51
15	33,33	6118,1	1110,89	37431147,61	203916,27
16	30,56	5461,9	933,91	29832351,61	166915,66
17	28,89	4979,4	834,63	24794424,36	143854,87
18	31,11	5519,8	967,83	30468192,04	171720,98
19	26,67	4361,8	711,29	19025299,24	116329,21
20	27,78	4130,2	771,73	17058552,04	114736,96
21	24,44	3821,4	597,31	14603097,96	93395,02
Soma	599,45	102483	17643,28	525310727,3	3037589,2
Media	28,55	4880,14			

c) Calcular a covariância amostral (s_{xy}). O que ela indica?

$$s_{xy} = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{n-1} = \frac{SPXY}{n-1} = 5609{,}38\ ^{o}C.R\$$$

Indica a intensidade da associação entre a temperatura e a venda de sorvetes. O sinal positivo indica que a associação é positiva.

d) Calcular o coeficiente de correlação linear da amostra (r) e interpretar seu significado.

$$r_{xy} = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sqrt{\left(\sum x_i^2 - n\bar{x}^2\right)\left(\sum y_i^2 - n\bar{y}^2\right)}} = \frac{SPXY}{\sqrt{SQX.SQY}} = 0{,}9695$$

Existe forte correlação linear positiva (r>0,8) entre temperatura e vendas

Significado: valores altos de temperatura estão associados a valores altos de vendas e vice-versa.

e) Teste de hipótese

1. Pressuposições:

- Distribuição da variável (X,Y) é normal bivariada;

2. Hipóteses estatísticas:

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

3. Erro de conclusão: $\alpha = 0{,}05$ (probabilidade de erro tipo I)

Erro tipo I = rejeitar H_0 quando ela é verdadeira.

4. Estatística do teste:

$$r_{xy} = \frac{SPXY}{\sqrt{SQX.SQY}} = 0{,}9695$$

$$T = \frac{R_{xy}}{\sqrt{\frac{1-R_{xy}^2}{n-2}}} \rightarrow t = \frac{0{,}9695}{\sqrt{\frac{1-0{,}9695^2}{21-2}}} = 17{,}23$$

5. Decisão e conclusão:

$\nu = 19 \rightarrow t_{\alpha/2(19)} = 2{,}093$

$t = 17{,}23 > t_{\alpha/2(19)} = 2{,}093 \rightarrow$ Rejeitamos H_0

Concluímos, ao nível de 5% de significância, que o coeficiente de correlação populacional difere de zero. Portanto, existe uma correlação linear positiva entre a temperatura e a venda de sorvetes. Valores altos de temperatura estão associados a valores altos de vendas e vice-versa.

f) Intervalo de confiança

1. Pressuposições:

- Distribuição da variável (X,Y) é normal bivariada;
- As variáveis são correlacionadas, ou seja, $\rho \neq 0$.

2. Obtenção da estimativa pontual:

$$r_{xy} = \frac{SPXY}{\sqrt{SQX.SQY}} = 0{,}9695$$

3. Cálculo dos limites do intervalo :

$$n = 21$$

$$z_{\alpha/2} = z_{0{,}475} = 1{,}96$$

$$IC(\rho; 0{,}95): \left[\frac{\exp\left\{2\left[\frac{1}{2}\ln\left(\frac{1+0{,}9695}{1-0{,}9695}\right) - \frac{1{,}96}{\sqrt{21-3}}\right]\right\}-1}{\exp\left\{2\left[\frac{1}{2}\ln\left(\frac{1+0{,}9695}{1-0{,}9695}\right) - \frac{1{,}96}{\sqrt{21-3}}\right]\right\}+1}; \frac{\exp\left\{2\left[\frac{1}{2}\ln\left(\frac{1+0{,}9695}{1-0{,}9695}\right) + \frac{1{,}96}{\sqrt{21-3}}\right]\right\}-1}{\exp\left\{2\left[\frac{1}{2}\ln\left(\frac{1+0{,}9695}{1-0{,}9695}\right) + \frac{1{,}96}{\sqrt{21-3}}\right]\right\}+1} \right]$$

$$IC(\rho; 0{,}95): [0{,}9248; 0{,}9878]$$

4. Conclusão: Concluímos com 95% de confiança que os limites 0,9248 e 0,9878 compreendem o coeficiente de correlação populacional.

Bibliografia consultada

BARBETTA, P.A. **Estatística Aplicada às Ciências Sociais**. 9ª Ed. Florianópolis: Editora da UFSC, 2014. 320p.

LEVIN, J.; FOX, J.A.; FORDE, D.R. **Estatística para Ciências Humanas**. 11ª Ed. São Paulo: Editora Pearson, 2012. 458p.

MONTGOMERY, D.C.; RUNGER, G.C.; HUBELE. N.F. **Estatística Aplicada à Engenharia**. 2 ed. Rio de Janeiro: Editora LTC. 2004. 335p.

NAGHETTINI, M.; PINTO, E.J. de A. **Hidrologia estatística**. Belo Horizonte: CPRM, 2007. 552 p.

Sistema Galileu de Educação Estatística. Disponível em: http://www.galileu.esalq.usp.br