BIS0005 - Bases Computacionais da Ciência

Aula 03 - Noções de Estatística, Correlação e Regressão

Saul Leite
Centro de Matemática, Computação e Cognição
Universidade Federal do ABC

Q2 2018

Estatística

- A Estatística é o ramo da ciência interessada em métodos para coleta, análise e apresentação de dados experimentais;

- Está presente em toda a ciência experimental pois fornece diretrizes para coleta de dados e métodos para sua análise

Estatística

A análise estatística pode ser dividida em duas áreas:

1. **Estatística descritiva**: Esta interessada na redução, análise e interpretação dos dados. Pode incluir a construção de gráficos, tabelas, e computação de várias medidas:medidas de tendência central (ex. a média), de dispersão (ex. a variância), de frequência (ex. percentagem), etc.
2. **Estatística indutiva / inferencial**: Voltada a realizar estimativas a partir de uma amostra ou testar ideias teóricas (hipóteses) com dados experimentais.

Estatística

Estatística Descritiva:

- O número de acidentes (frequência) nas rodovias antes e depois da Lei Seca;
- Gráfico com a distribuição da idade dos ingressantes nos bacharelados interdisciplinares da UFABC.

Estatística Indutiva/Inferencial:

- Estimação da porcentagem da população que votará para um determinado candidato em uma eleição, junto com uma margem de erro ("intervalo de confiança");
- Teste estatístico de tendência de queda nas populações de atum-rabilho a partir de observações sistemáticas;

Conceitos Básicos

População

População pode ser definido como o conjunto de elementos que têm em comum uma determinada característica.

- indivíduos com dengue;
- aparelhos de televisão fabricados em uma fábrica;

Amostra

Uma amostra é um subgrupo de uma população, ou seja, é todo subconjunto não vazio e com menor número de elementos que o conjunto definido como população.

Conceitos Básicos

Variáveis

É toda característica que, observada em uma unidade experimental, pode variar de um indivíduo para outro.

- Exemplo: gênero, cor da pele, altura, idade, salário, nível de hemoglobina no sangue.
- Podem ser divididas em: **Qualitativas** e **Quantitativas**

Conceitos Básicos: Variáveis

Qualitativas:

- **Nominal**: Uma categoria se diferencia da outra somente pela denominação que recebem. Ex.: gênero feminino ou masculino, estado civil, nacionalidade.
- **Ordinal**: É possível reconhecer graus de intensidade entre as categorias. Ex.: nível sócio-econômico (baixo, médio, alto), avaliação de um serviço (ruim, regular, bom).

Conceitos Básicos: Variáveis

Quantitativas:

- **Discretas**: São aquelas em que os dados somente podem apresentar determinados valores, em geral, números inteiros. Ex.: número de filhos, número de baixas hospitalares.
- **Contínuas**: Aquelas cujos dados somente podem apresentar qualquer valor dentro de um intervalo de variação possível. Ex.: altura, peso, níveis de hemoglobina no sangue.

Conceitos Básicos: Variáveis

Exemplo:

	mpg	cyl	hp	wt	vs	am
Mazda RX4	21.0	6	110	2.620	0	1
Mazda RX4 Wag	21.0	6	110	2.875	0	1
Datsun 710	22.8	4	93	2.320	1	1
Hornet 4 Drive	21.4	6	110	3.215	1	0
Hornet Sportabout	18.7	8	175	3.440	0	0
Valiant	18.1	6	105	3.460	1	0

vs: motor (0 = em V, 1 = em linha)

am: transmissão (0 = automático, 1 = manual)

Medidas Estatísticas

É conveniente dispor de medidas que informem sobre a amostra de maneira mais resumida do que os dados brutos são capazes de fazer. Dão uma visão global dos dados, podendo ser:

Tendência central

São aquelas que produzem um valor em torno do qual os dados observados se distribuem, e que visam sintetizar em um único número o conjunto de dados.

Ex: Média aritmética, Mediana e Moda.

Dispersão

É a variabilidade que os dados apresentam entre si.

Ex: Variância e Desvio-padrão.

Medidas de Tendência central

Média: É a soma de todas as observações dividida pelo número de observações:

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{X_1 + X_2 + X_3 + X_4 + \ldots + X_n}{n},$$

em que X_i representa as observações.

Medidas de Tendência central

Exemplo do cálculo da **média** em R:

```
x <- c(1,1,2,6,5,3,1,3,4)
mean(x)
```

```
## [1] 2.888889
```

Equivalente a:

```
x <- c(1,1,2,6,5,3,1,3,4)
sum(x)/length(x)
```

```
## [1] 2.888889
```

Medidas de Tendência central

Mediana: valor central do conjunto que divide os dados (*ordenados*) em duas partes iguais (mesmo número de escores abaixo e acima do valor).

Atenção: os dados devem estar ordenados

Medidas de Tendência central

Exemplo do cálculo da **mediana** em R:

```
x <- c(1,1,2,6,5,3,1,3,4)
median(x)
```

```
## [1] 3
```

Note que é o elemento na posição $\frac{n+1}{2}$ do vetor **ordenado**:

```
x <- c(1,1,2,6,5,3,1,3,4)
sort(x)
```

```
## [1] 1 1 1 2 3 3 4 5 6
```

Medidas de Tendência central

Quando temos um número **par** de elementos, a **mediana** é a média dos dois valores centrais:

```
x <- c(1,1,2,6,5,3,1,3)
sort(x)
```

```
## [1] 1 1 1 2 3 3 5 6
```

Temos que mediana $= (2+3)/2$:

```
x <- c(1,1,2,6,5,3,1,3)
median(x)
```

```
## [1] 2.5
```

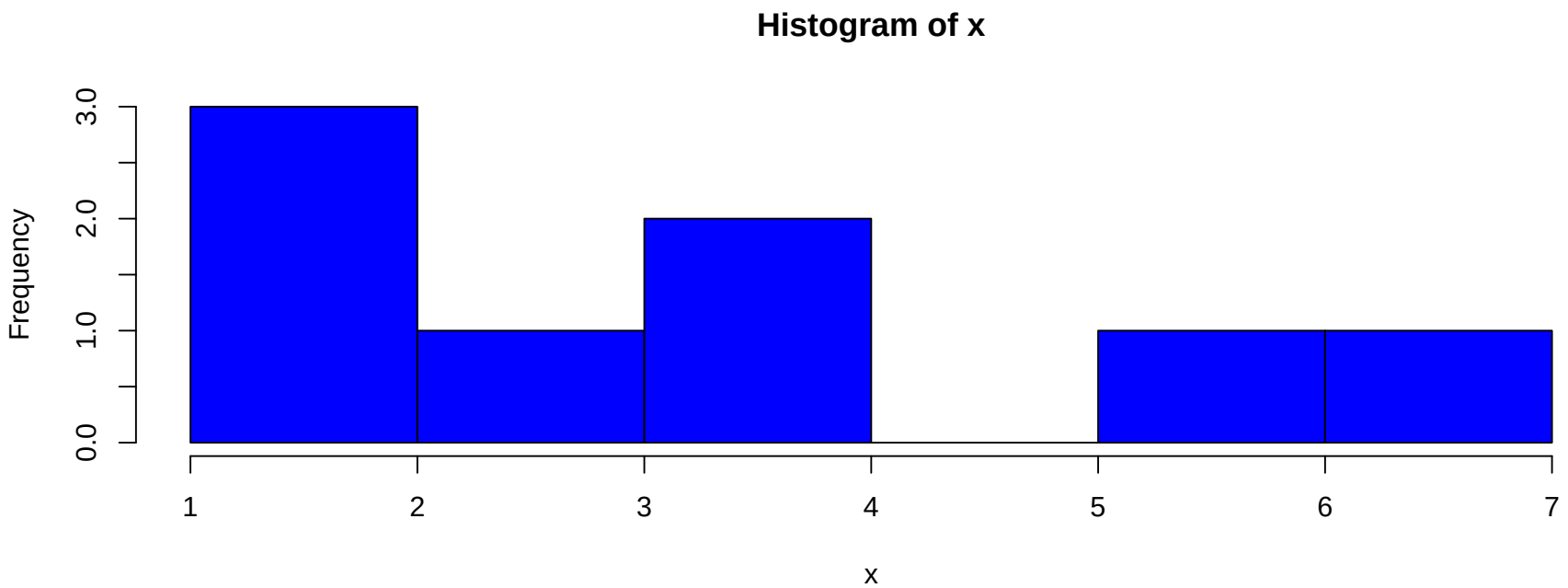
Medidas de Tendência central

Moda: é o valor que ocorre com mais frequência entre todas as observações.

Medidas de Tendência central

Exemplo do cálculo da moda:

```
x <- c(1,1,2,6,5,3,1,3)
hist(x,breaks=c(1,2,3,4,5,6,7),right=FALSE,col="blue")
```



Portanto a moda seria 1.

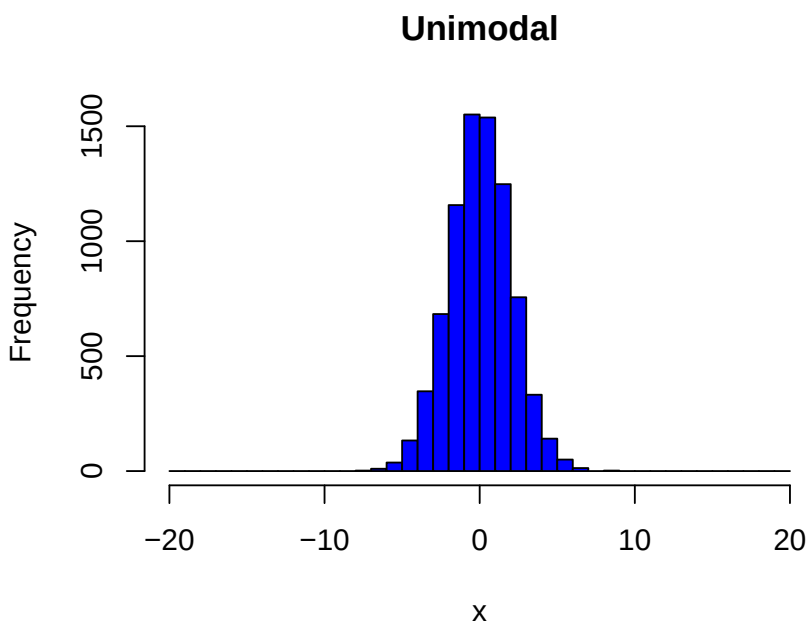
Medidas de Tendência central

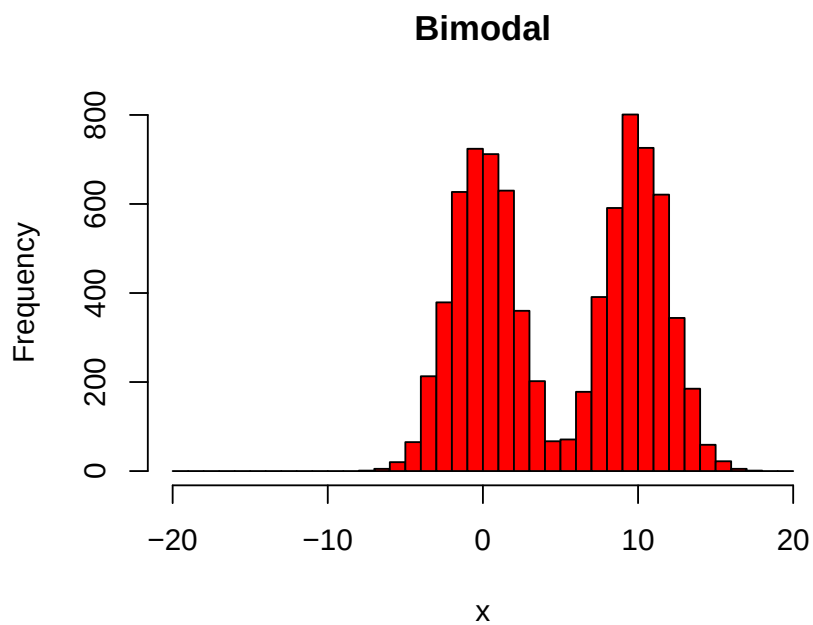
A **moda** pode não existir ou não ser única, como ilustrado nos exemplos abaixo:

$x = (1, 1, 3, 3, 5, 7, 7, 7, 11, 13) \rightarrow$ moda $= 7$

$x = (1, 1, 3, 3, 3, 4, 5, 5, 7, 7, 7, 11, 13) \rightarrow$ duas modas 3 e 7

$x = (5, 8, 9, 11, 13) \rightarrow$ não possui moda





Histogramas

Podemos fazer histogramas em R usando o comando **hist**, como ilustrado nos slides anteriores. Ele possui os seguintes parâmetros:

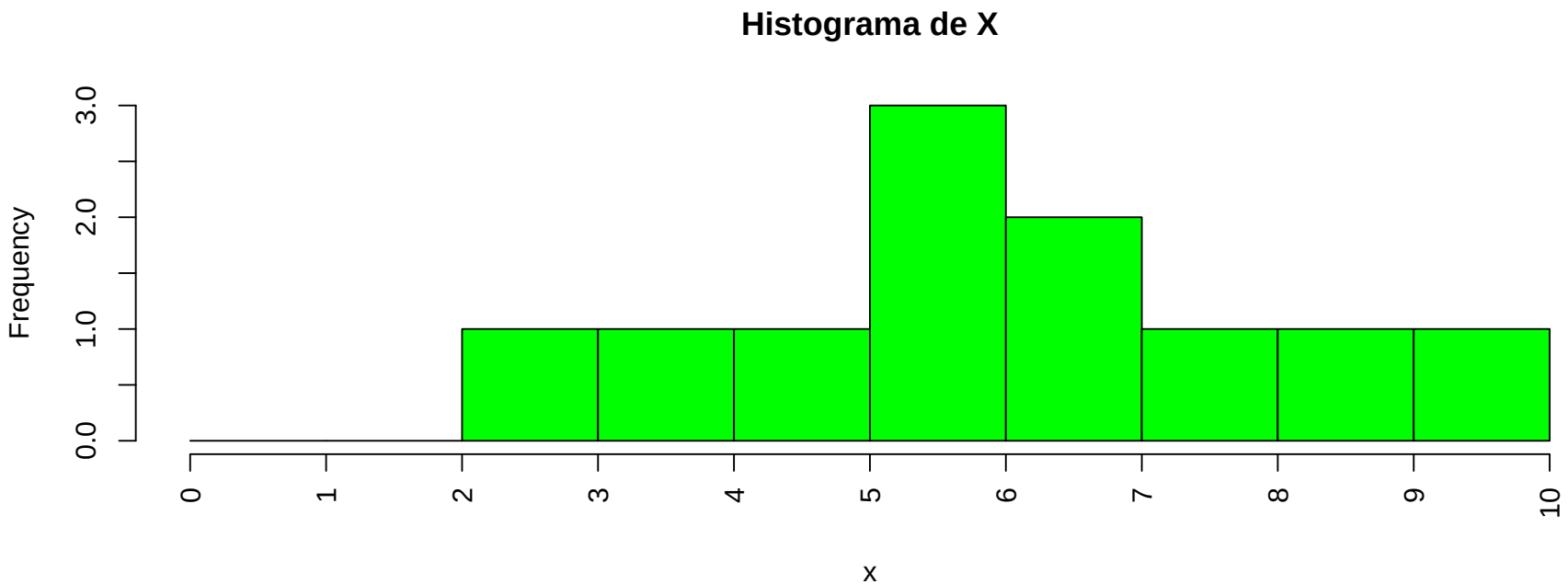
- **x**: dados para gerar o histograma.
- **breaks**: indica os pontos de divisão para os intervalos.
- **right**: indica se os intervalos são fechados para a direita ou esquerda.
- **freq**: se o resultado deve ser mostrado em frequências ou frequências relativas (proporções) (pode ser TRUE ou FALSE).

Além dos parâmetros usuais de gráficos, como **col**, **main**, **xlab**, **ylab**, **xlim**, etc. . .

Histogramas

Exemplo:

```
x <- c(3, 5, 6, 4, 5, 8, 9, 6, 2, 7, 5)
hist(x,breaks=seq(0,10,1),right=FALSE,
     col="green",main="Histograma de X",xaxt="n")
axis(1,at=seq(0,10,1),las=2)
```



Histogramas

O comando **axis** usado no exemplo anterior é usado para alterar os valores numéricos que aparecem nos eixos. Ele recebe os seguintes parâmetros:

- **side**: toma valor 1 se for para o eixo x e 2 para o eixo y.
- **at**: pontos onde as marcas serão exibidas.
- **las**: toma valor 1 se os números nos eixos devem ser exibidos horizontalmente e 2 se devem ser exibidos perpendicularmente ao eixo.

Atenção: para não ter as marcas sobrepostas nos eixos, desabilite as marcas dos eixo geradas automaticamente usando xaxt $=$ "n" ou yaxt $=$ "n", para o eixo x ou y, respectivamente.

Medidas Estatísticas

O processo de trabalhar com dados introduz uma variabilidade nos resultados obtidos, pois cada individuo de uma amostra vai ter características ligeiramente diferentes.

Essa variabilidade é medida através das **medidas de dispersão**.

Dentre as medidas de dispersão tem-se:

- Variância
- Desvio-padrão

Medidas de Dispersão

Variância (amostral): É a soma dos desvios ao quadrado dividido pelo número de elementos menos 1.

Valores X	Desvios $(X - \bar{X})$	Desvios ao Quadrado $(X - \bar{X})^2$
0	-5	25
4	-1	1
6	1	1
8	3	9
7	2	4
$\bar{X} = 5$	$\sum(X - \bar{X}) = 0$	$\sum(X - \bar{X})^2 = 40$

Logo temos que:

$$S^2 = \frac{\sum(X - \bar{X})^2}{n - 1} = \frac{40}{4} = 10$$

Medidas de Dispersão

Desvio Padrão (amostral): É definido como a raiz quadrada da variância. Ou seja:

$$S = \sqrt{\frac{\sum (X - \bar{X})^2}{n - 1}}$$

Medidas de Dispersão

Exemplo do cálculo da média e variância no R:

```
x <- c(0,4,6,8,7)
#variância
var(x)
```

```
## [1] 10
```

```
#desvio pardrão
sd(x)
```

```
## [1] 3.162278
```

Medidas de Dispersão

Como exemplo, considere os seguintes dados da circunferência, altura, e volume de 31 árvores Cerejeiras-Negras:

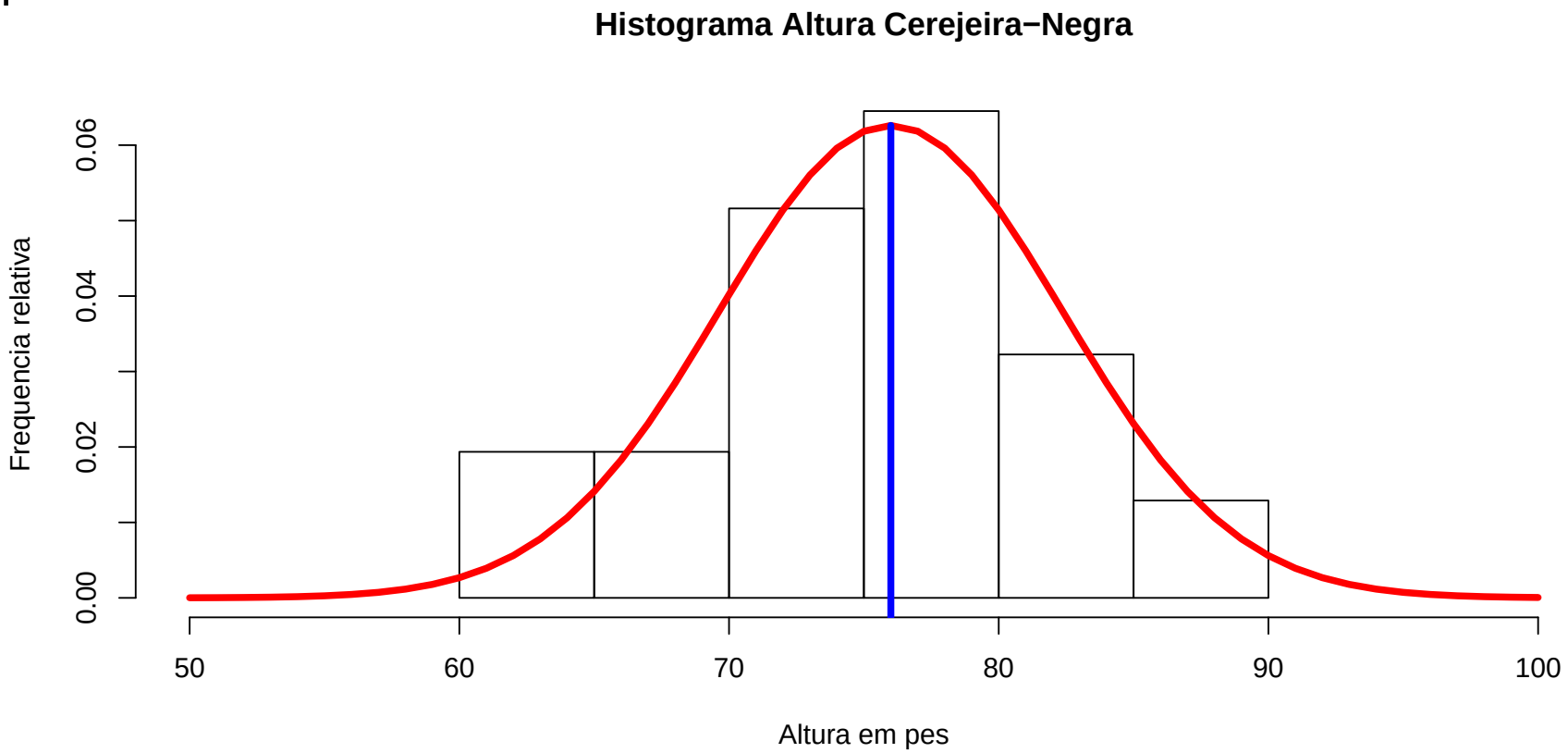
Circ.	Altura	Vol.
8.3	70	10.3
8.6	65	10.3
8.8	63	10.2
10.5	72	16.4
10.7	81	18.8
10.8	83	19.7

Média Altura $= 76$

Desvio Padrão altura $= 6.371813$

Medidas de Dispersão

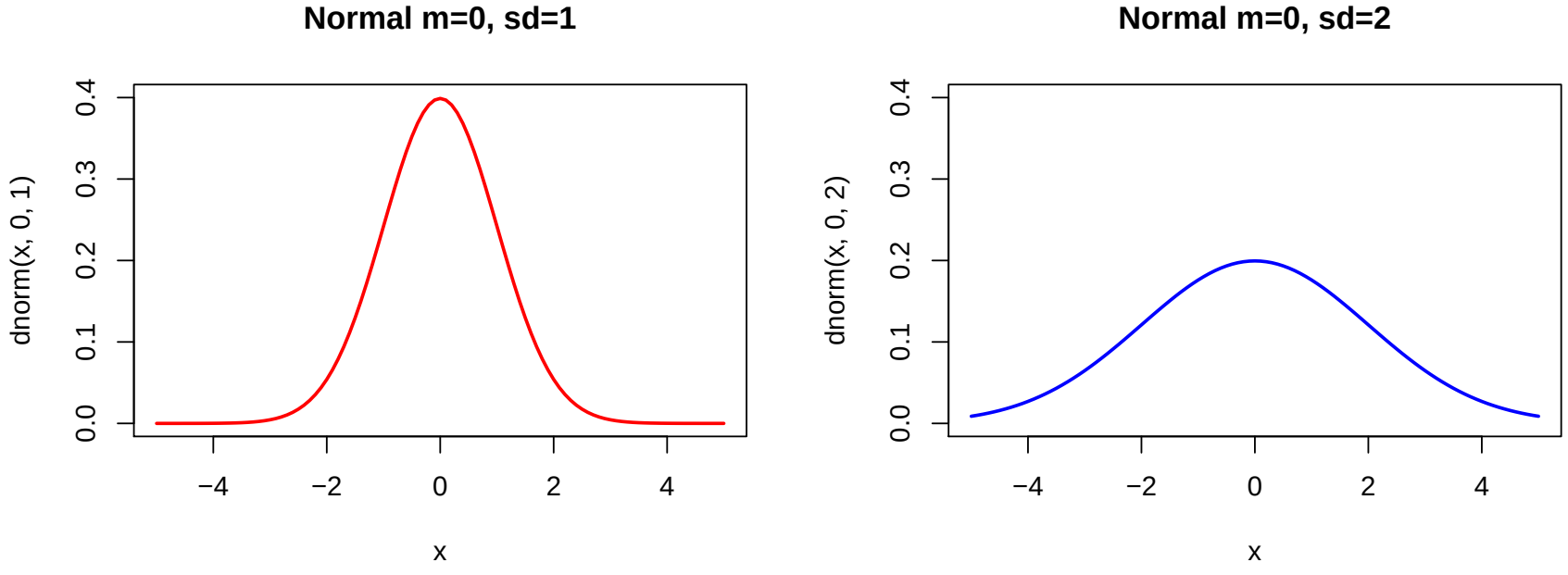
Em vermelho uma distribuição normal com média 76 e desvio padrão 6.371813:



A linha em azul representa o valor da **média**.

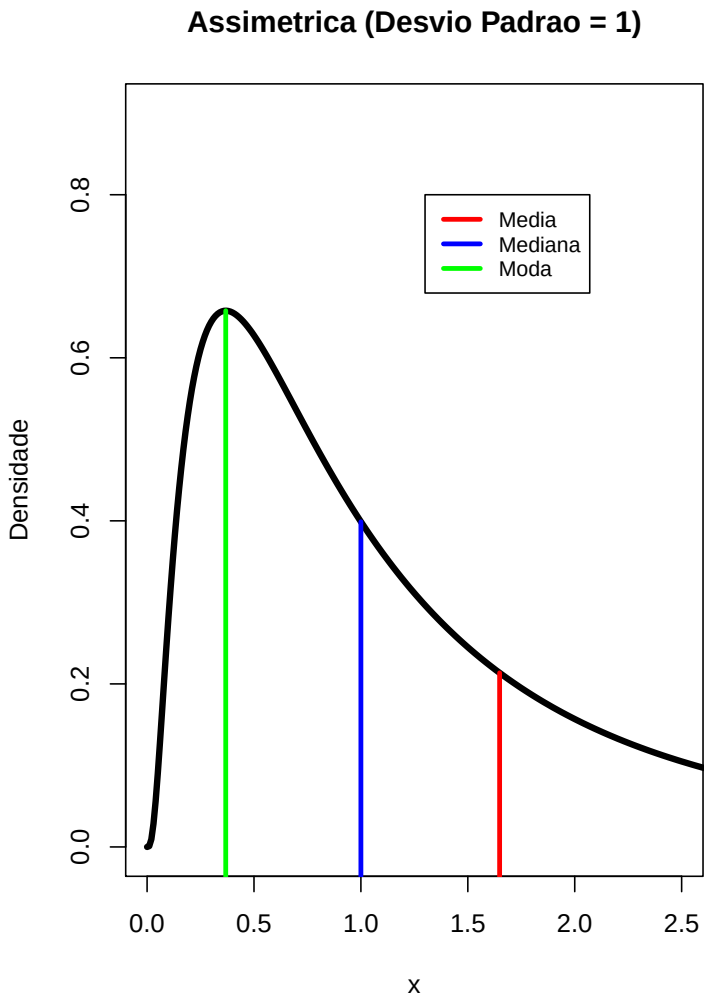
Distribuição Normal (Gaussiana)

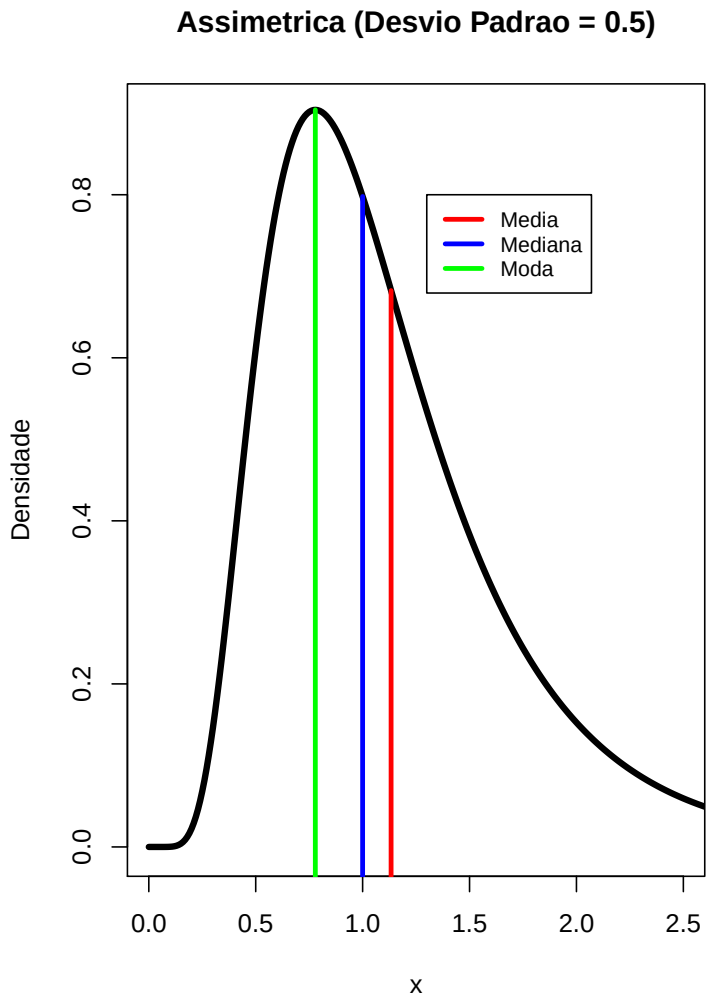
Determinada por dois parâmetros: **Média** (posição central) **Desvio padrão** (largura)



A distribuição normal é **simétrica** e **unimodal**. Por causa da simetria, valores de média, mediana e moda são iguais.

Exemplo distribuições assimétricas





Correlação e Regressão

Correlação e Regressão

As técnicas de correlação e regressão analisam dados amostrais, procurando determinar como duas (ou mais) variáveis estão relacionadas umas com as outras.

Exemplos:

Variável independente	**Variável dependente**
Horas de treinamento	Número de acidentes
Altura da pessoa	Número do sapato
Cigarros por dia	Capacidade pulmonar
Meses do ano	Volume de vendas
Peso da pessoa	QI

Variáveis dependentes/independentes

Independente: Valores manipulados ou selecionados pelo pesquisador (altura, idade, mês). Podem ser ou não a "causa" da variável dependente.

Dependente: Valores observados, contados, ou medidos, que não estejam sob controle direto do pesquisador (velocidade, taxa de câmbio). Podem ser "causadas" ou não pela variável independente.

Variáveis dependentes/independentes

Independente: Valores manipulados ou selecionados pelo pesquisador (altura, idade, mês). Podem ser ou não a "causa" da variável dependente.

Dependente: Valores observados, contados, ou medidos, que não estejam sob controle direto do pesquisador (velocidade, taxa de câmbio). Podem ser "causadas" ou não pela variável independente.

OBS.: Quando não há relação causal óbvia entre duas ou mais variáveis, qual é 'independente ' ou 'dependente' é uma questão de rótulo.

Correlação e Regressão

A análise de **correlação** tem como resultado um número que expressa o grau de relacionamento entre duas variáveis.

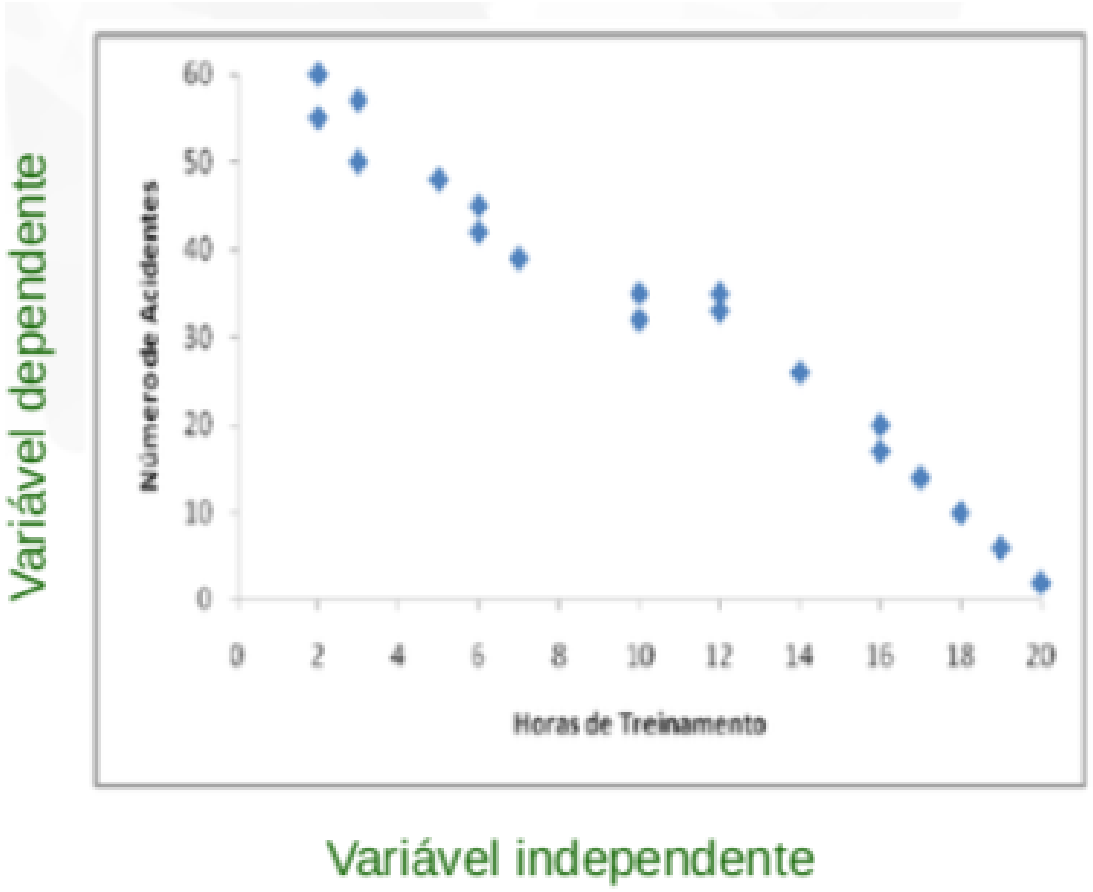
A análise de **regressão** expressa o resultado em uma equação matemática, descrevendo o relacionamento.

Ambas análises, geralmente utilizadas em pesquisas exploratórias.

Correlação: Gráfico de Disperssão

A análise gráfica do comportamento entre as variáveis mostra a existência de **correlação negativa**, pois à medida que X cresce, Y decresce.

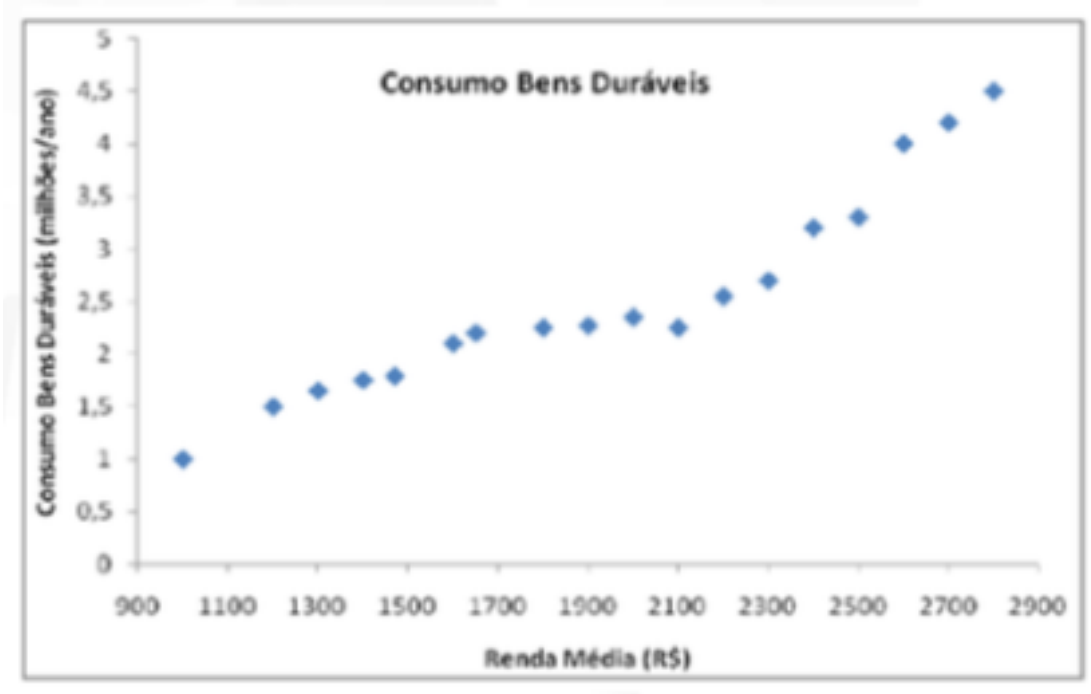
O gráfico mostra que a empresa, ao investir em treinamento, reduz o número de acidentes na fábrica



Correlação: Gráfico de Disperssão

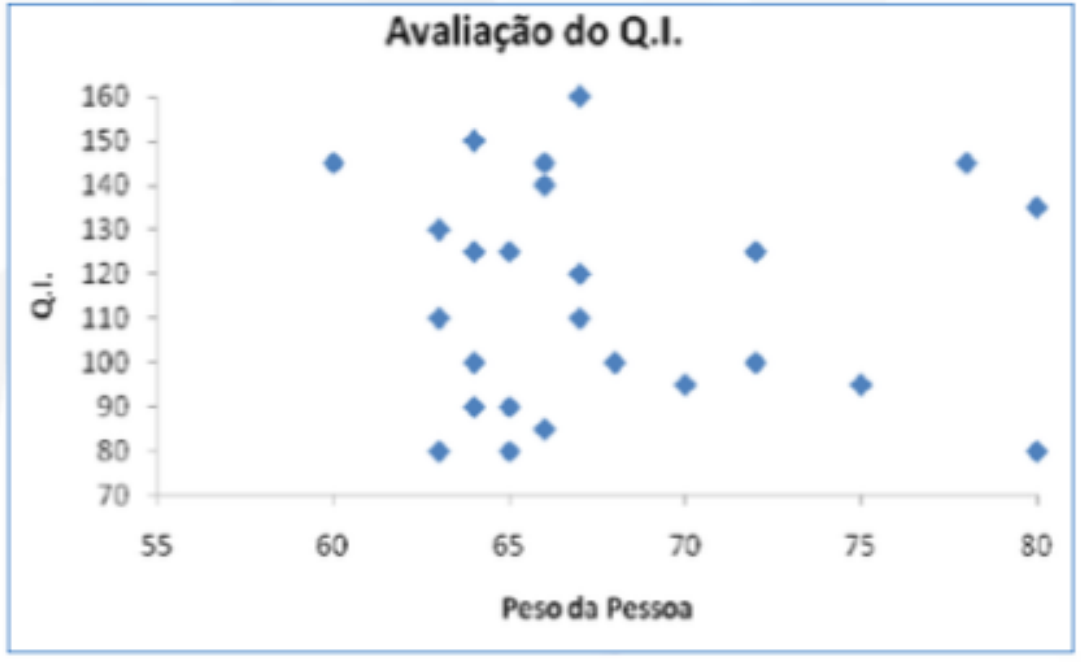
A análise gráfica do comportamento entre as variáveis mostra a existência de **correlação positiva**, pois à medida que X cresce, Y também cresce.

O gráfico mostra que, com o aumento médio da renda da população, o consumo de bens duráveis aumenta.



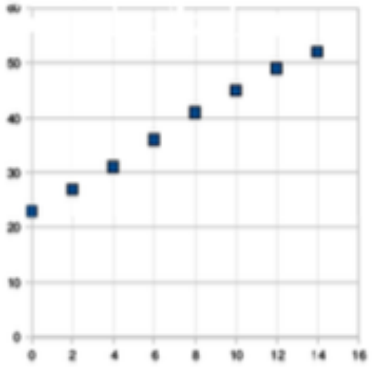
Correlação: Gráfico de Dispersão

Não há correlação linear, o gráfico mostra que não existe evidência de alguma relação entre o peso de uma pessoa com seu Q.I.



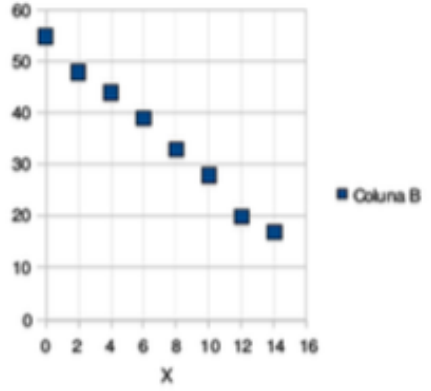
Correlação: Gráfico de Disperssão

Correlação Linear Positiva



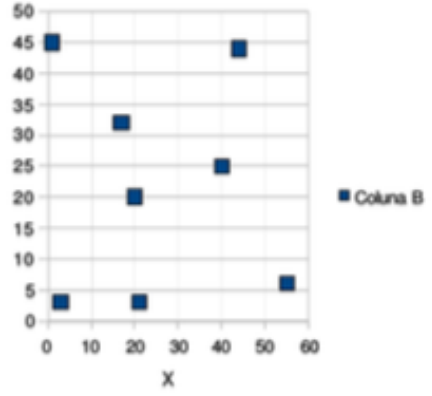
À medida que x cresce, y tende a crescer.

Correlação Linear Negativa

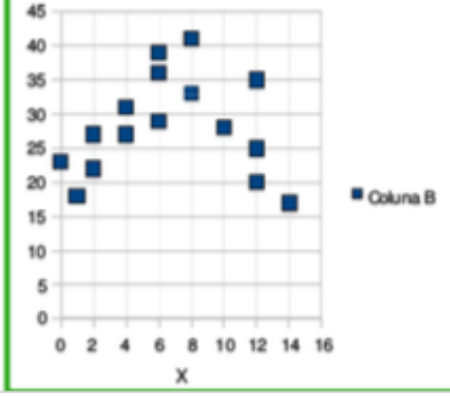


À medida que x cresce, y tende a decrescer.

Não há Correlação



Correlação Não Linear

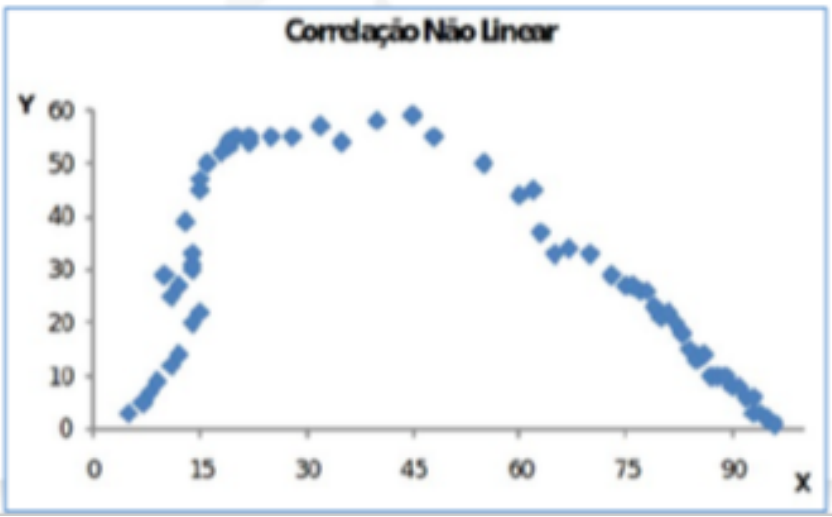


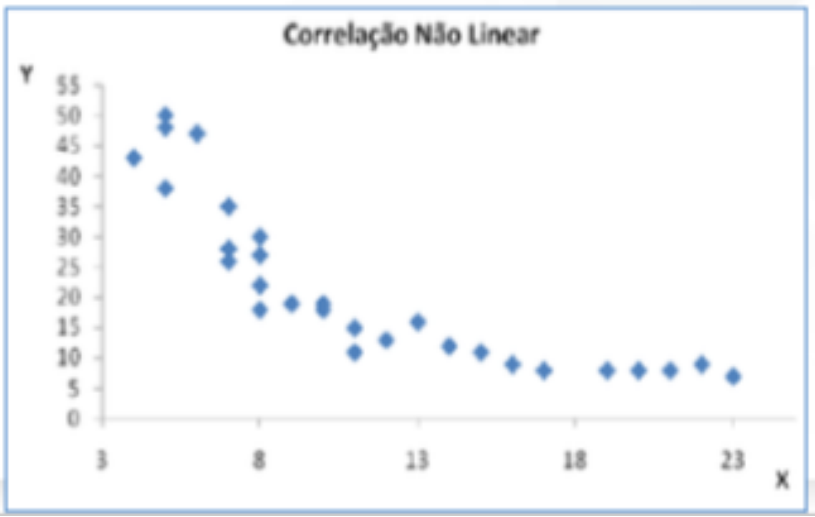
Correlação: Gráfico de Dispersão

Podemos ter dois tipos de correlação entre as variáveis:

Correlação linear: em que a relação entre as duas variáveis é expressa adequadamente por uma reta.

Correlação não-linear: apesar de existir uma relação clara entre as variáveis, esta não pode ser modelada por uma reta.





Correlação: Coeficiente de Correlação

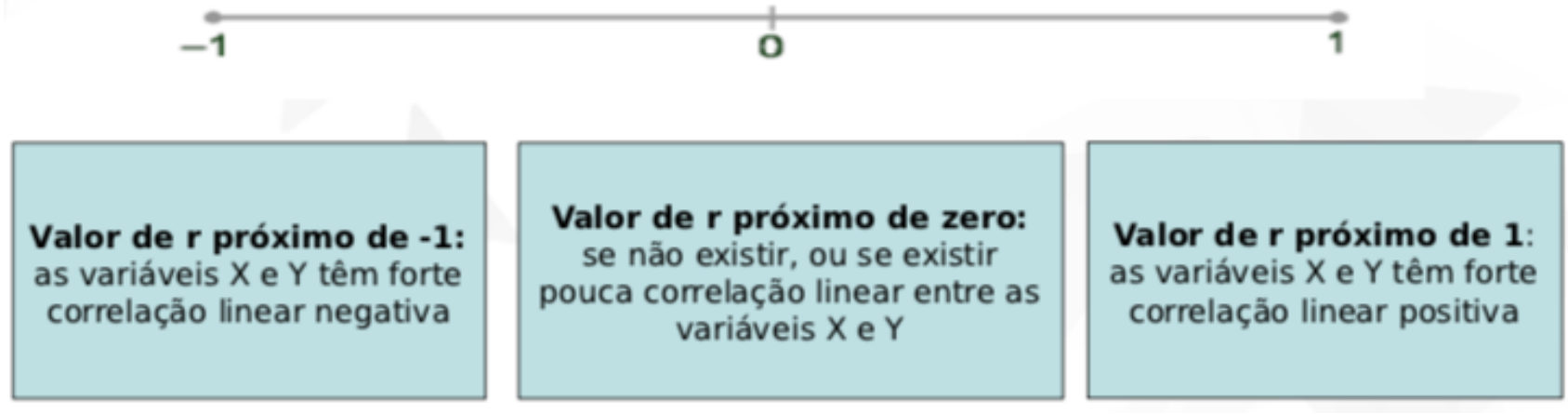
Utilizar apenas o gráfico de dispersão para interpretar a existência de uma correlação pode ser uma tarefa bastante subjetiva.

Como medida mais objetiva, utiliza-se o **coeficiente de correlação** para medir o grau e o tipo de uma correlação linear entre duas variáveis.

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

Correlação: Coeficiente de Correlação

O intervalo de variação do coeficiente de correlação r está ente -1 à 1.



Correlação: Coeficiente de Correlação

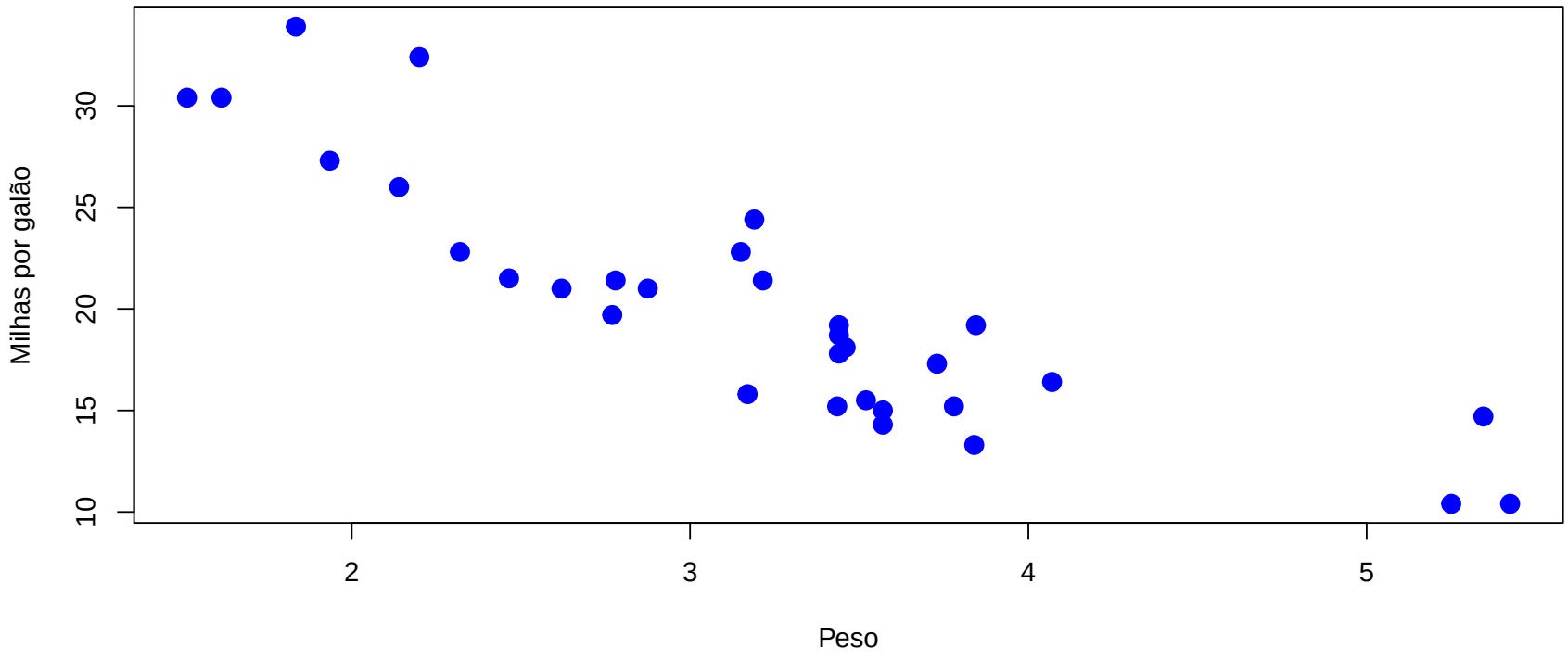
Exemplo 1:

	mpg	cyl	hp	wt	vs	am
Mazda RX4	21.0	6	110	2.620	0	1
Mazda RX4 Wag	21.0	6	110	2.875	0	1
Datsun 710	22.8	4	93	2.320	1	1
Hornet 4 Drive	21.4	6	110	3.215	1	0
Hornet Sportabout	18.7	8	175	3.440	0	0
Valiant	18.1	6	105	3.460	1	0

Vejamos se há correlação linear entre peso (**wt**) (em milhares de libras) e milhas por galão (**mpg**) de carros.

Correlação: Coeficiente de Correlação

O Gráfico de dispersão é dado por:



O coeficiente de correlação é dado por: $r = -0.8676594$.

Correlação: Coeficiente de Correlação

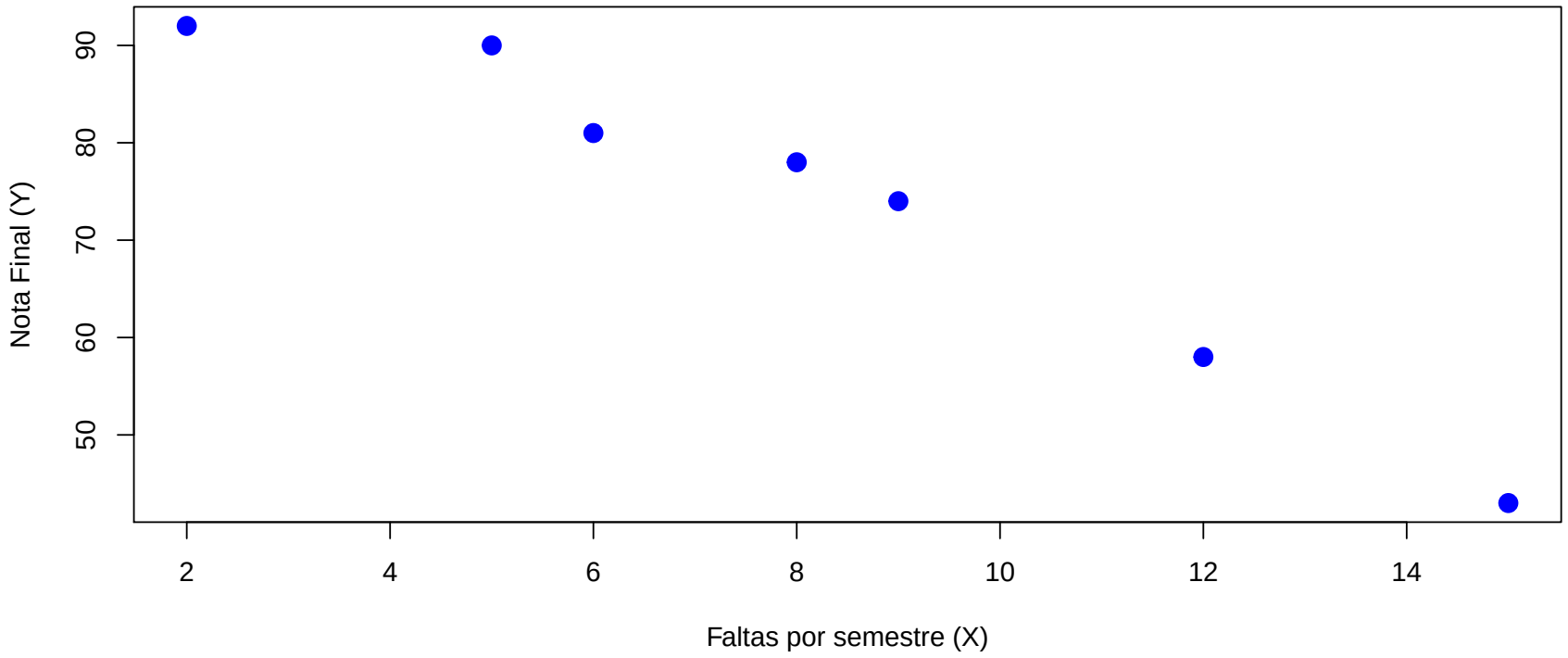
Exemplo 2:

Faltas por semestre (X)	Nota Final (Y)
8	78
2	92
5	90
12	58
15	43
9	74
6	81

Vejamos a correlação entre número de faltas e nota final.

Correlação: Coeficiente de Correlação

Gráfico de Dispersão



O coeficiente de correlação é dado por: $r = -0.9747632$.

Correlação: Coeficiente de Correlação

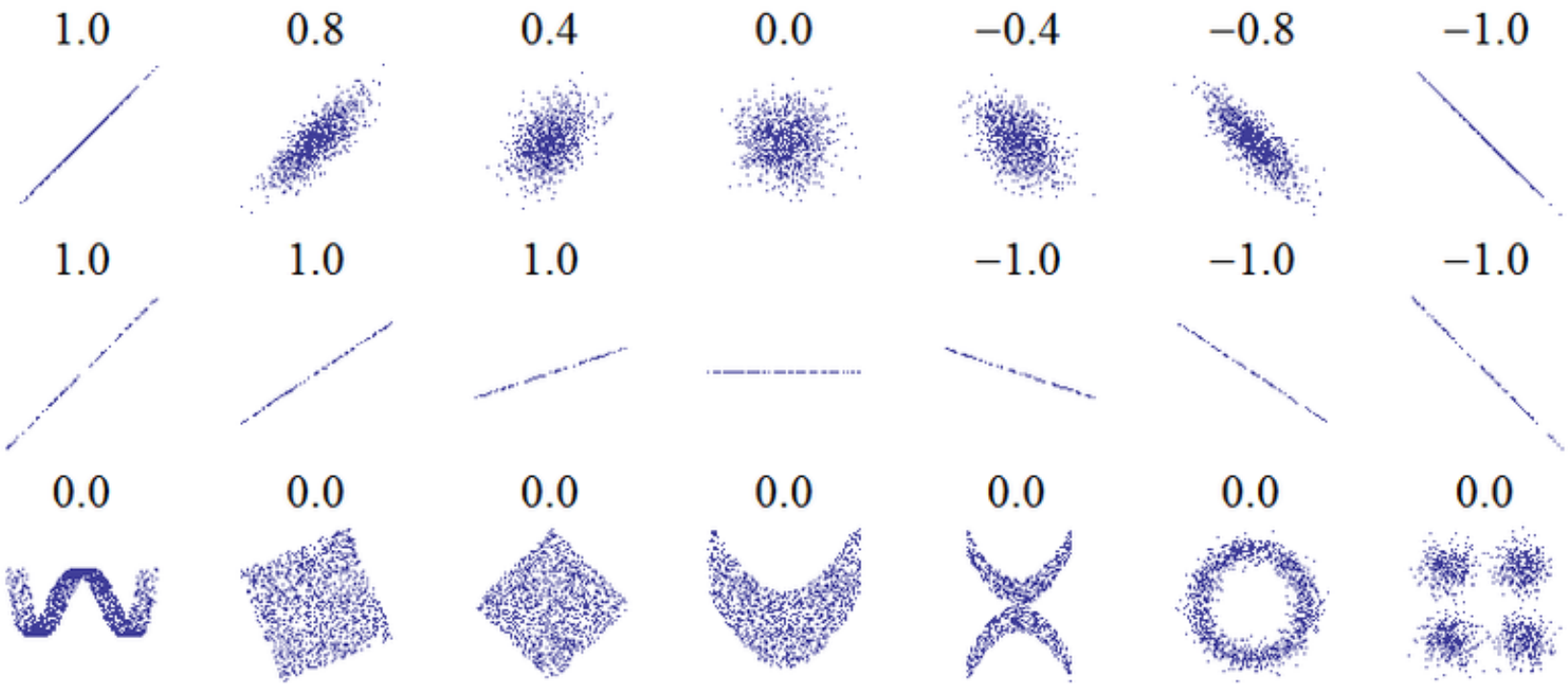
Como calcular usando o R:

```
x <- c(8,2,5,12,15,9,6)
y <- c(78,92,90,58,43,74,81)
cor(x,y) #calcula o coef. de correlacao entre x e y
```

```
## [1] -0.9747632
```

OBS.: Note que os valores dos vetores x e y são os mesmos dados pela tabela para o exemplo 2.

Correlação: Coeficiente de Correlação



Correlação e Causalidade

Correlação **não** necessariamente implica em causalidade.

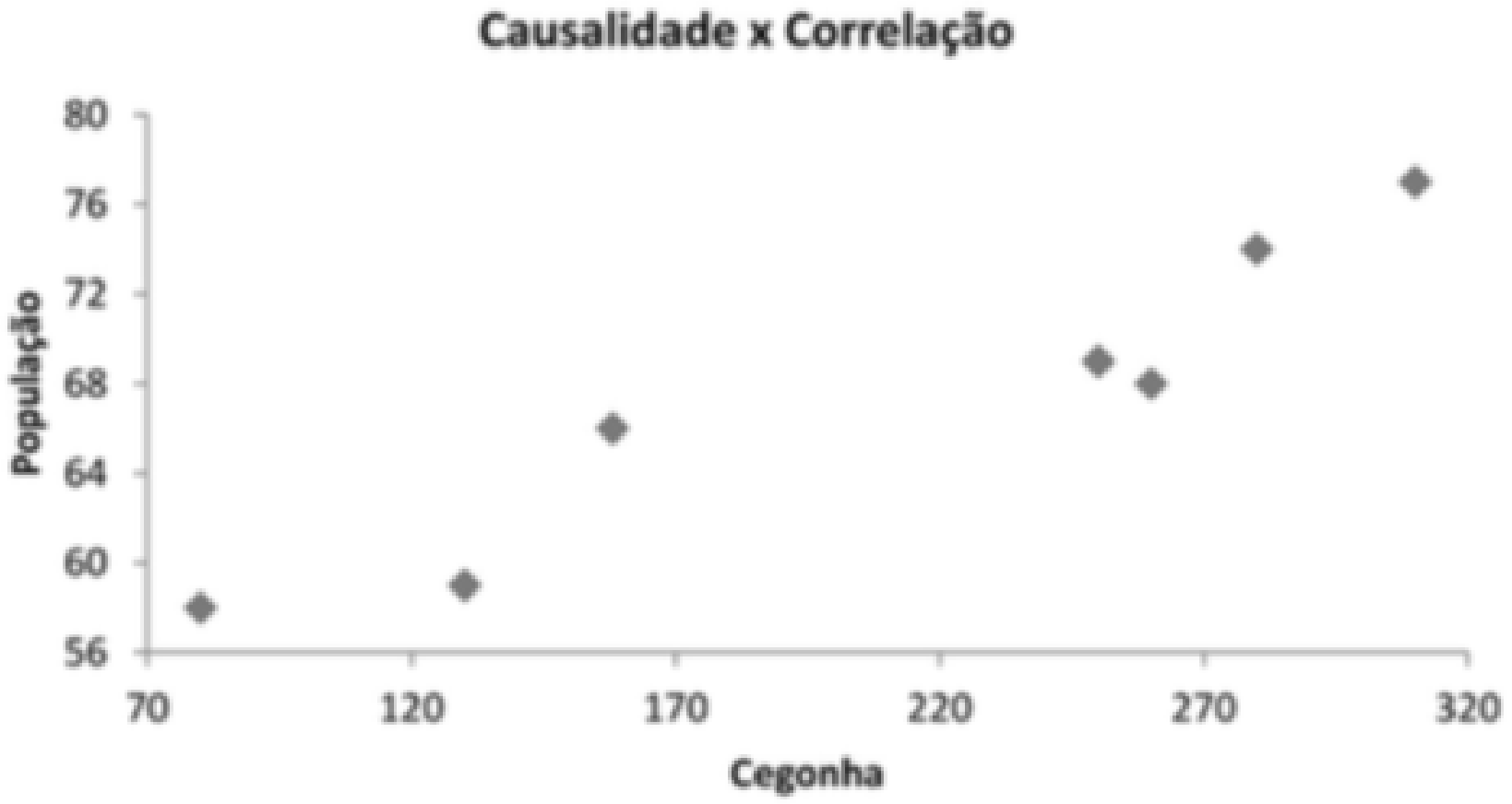
O número de pessoas usando óculos-de-sol e a quantidade de sorvete consumido em um particular dia são altamente correlacionados.

Isto não significa que usar óculos-de-sol causa a compra de sorvetes ou vice-versa!

É extremamente difícil estabelecer relações causais a partir de dados observacionais. Precisamos realizar experimentos para obter mais evidências de uma relação causal.

Correlação e Causalidade

Correlação não necessariamente implica em causalidade.



Reta de regressão linear

Depois de constatar que existe uma correlação linear significante, é possível escrever uma equação que descreva a relação linear entre as variáveis X e Y.

Essa equação chama-se **reta de regressão**.

Pode-se escrever a equação de uma reta como sendo:

$$f(x) = mx + b,$$

em que:

- m é a inclinação.
- b é o intercepto-y.

Reta de regressão linear

Para a reta de regressão, m é dado por:

$$m = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

e b é dado por:

$$b = \bar{Y} - m\bar{X}$$

Reta de regressão linear

Podemos calcular a regressão linear em R com o comando **lm** ("*Linear model*"):

```
x <- c(8,2,5,12,15,9,6)
y <- c(78,92,90,58,43,74,81)
#a expressão 'y ~ x' indica que desejamos
# explicar y com a variável x.
lm(y ~ x)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Coefficients:
## (Intercept)            x
##     105.668       -3.924
```

Reta de regressão linear

Podemos ver que $b = 105.668$ e $m = -3.924$. Vamos definir a função e fazer o gráfico dos dados:

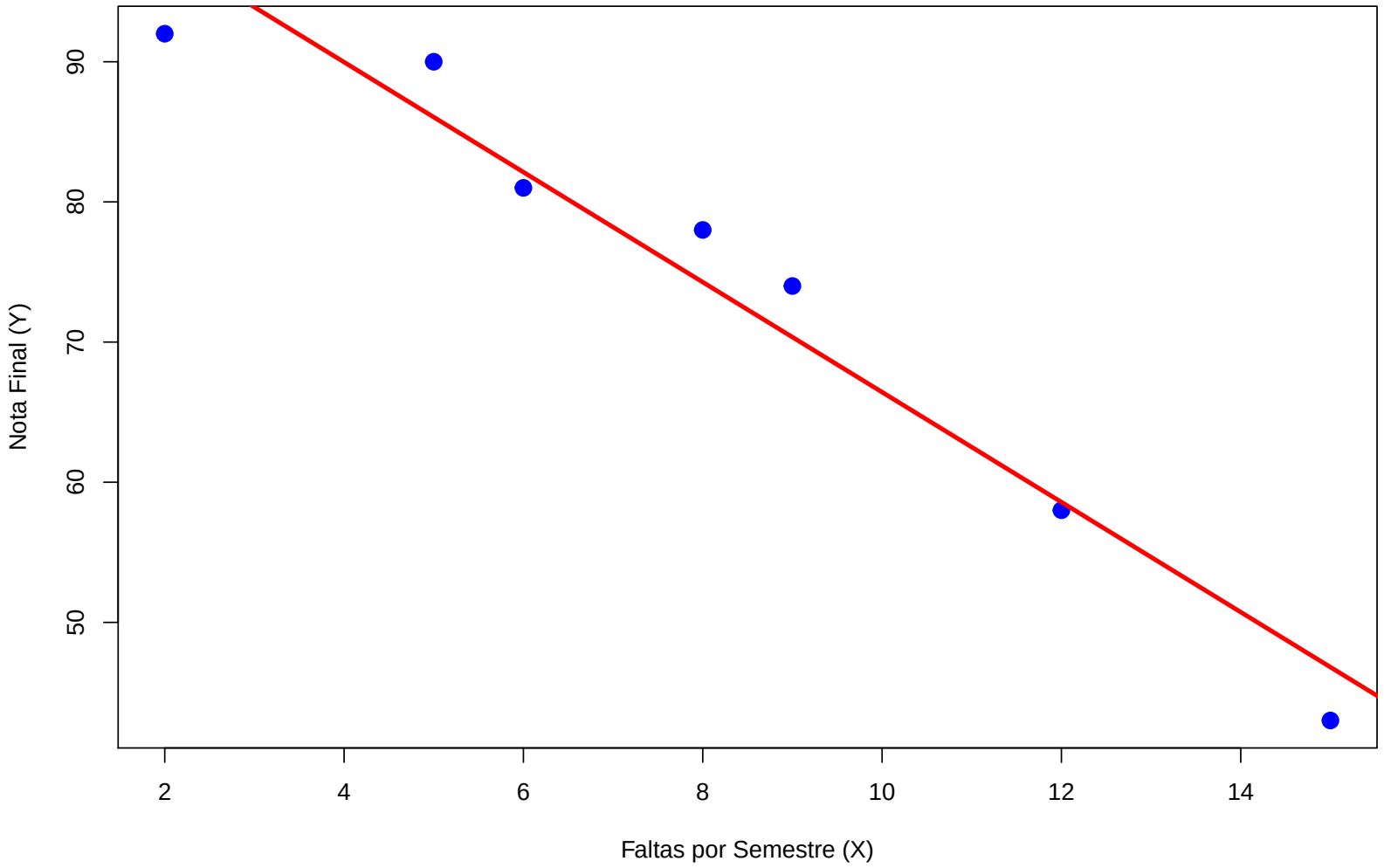
```
x <- c(8,2,5,12,15,9,6)
y <- c(78,92,90,58,43,74,81)

#Gráfico de dispersão dos pontos
plot(x,y,ylab="Nota Final (Y)",
     xlab="Faltas por Semestre (X)",pch=19,col="blue")

#Fazendo o gráfico da reta
f <- function(x) 105.668 - 3.924*x
xp <-seq(0,20,1)
lines(xp,f(xp),col="red",lwd=3)
```

Reta de regressão linear

Gráfico Resultante:



Reta de regressão linear

De forma alternativa, podemos usar o comando **abline** para fazer gráficos de retas. Seus parâmetros são o **intercepto-**y (b) e a **inclinação da reta** (m):

```
x <- c(8,2,5,12,15,9,6)
y <- c(78,92,90,58,43,74,81)

#Gráfico de dispersão dos pontos
plot(x,y,ylab="Nota Final (Y)",
     xlab="Faltas por Semestre (X)",pch=19,col="blue")

#Fazendo o gráfico da reta
abline(105.668, -3.924,col="red",lwd=3)
```

ATIVIDADE EM SALA

Exercício 1

Considere os dados abaixo, que representam o diâmetro e o altura de conchas (trabalho de Carlos Marcellari 1984).

Concha	Diametro	Altura
1	185	78
2	194	65
3	173	77
4	200	76
5	179	72
6	213	76
7	134	75
8	191	77
9	177	69
10	199	65

Exercício 1

Responda:

1. Calcule a média e mediana da altura e diâmetro das conchas;
2. Determine a variância e desvio padrão da altura e diâmetro;
3. Através do gráfico de dispersão, determine se existe uma correlação linear. Confirme sua resposta calculando o coeficiente de correlação.

OBS.: Para facilitar, copie os dados abaixo:

```
##  [1] 185 194 173 200 179 213 134 191 177 199
```

```
##  [1] 78 65 77 76 72 76 75 77 69 65
```

Exercício 2

Considere os dados abaixo, que representam o tempo médio de estudo (em anos) vs. renda anual (em dólares) (dados de 1971).

	Educacao	Renda
biologists	15.09	8258
architects	15.44	14163
civil.engineers	14.52	11377
mining.engineers	14.64	11023
computer.programers	13.83	8425
librarians	14.15	6112
secondary.school.teachers	15.08	8034
pharmacists	15.21	10432
radio.tv.announcers	12.71	7562
bookkeepers	11.32	4348

Exercício 2

Responda:

1. Através do gráfico de dispersão, determine se existe uma correlação linear.
2. Faça a regressão linear para este exemplo e adicione ao gráfico da dispersão o gráfico da reta de regressão.

OBS.: Para facilitar, copie os dados abaixo:

```
##  [1] 15.09 15.44 14.52 14.64 13.83 14.15
##  [7] 15.08 15.21 12.71 11.32
```

```
##  [1]  8258 14163 11377 11023  8425  6112
##  [7]  8034 10432  7562  4348
```

Atividades para casa

Atividades para fazer até a próxima aula:

- Fazer todos os exercícios do capítulo 3 do livro usando R.
- Ler o **Capítulo 4** "Bases de Dados" do livro.

Referências

- Aulas dos Profs. David Correa Martins Jr, Wagner Tanaka Botelho, Jesús P. Mena-Chalco, Fernanda Almeida e Carolina Benetti.
- Livro Bases Computacionais da Ciência.