



Universidade Federal de Lavras
Departamento de Biologia
Programa em Genética e Melhoramento de Plantas



PGM522: Análise de Experimentos em Genética e Melhoramento de Plantas

Aula 02: Introdução a Estatística

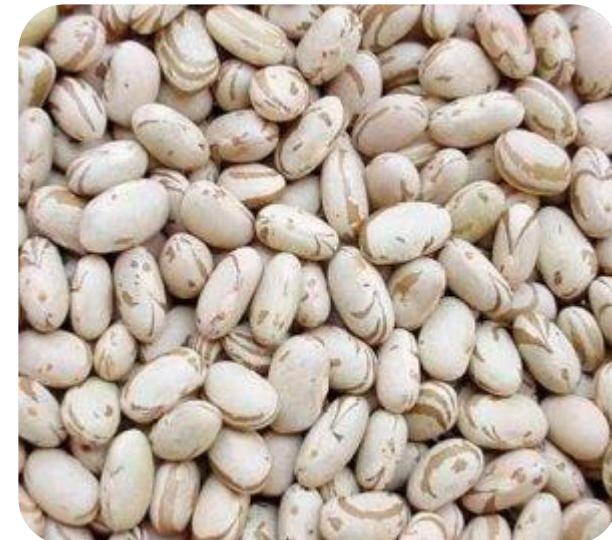
Vinícius Quintão Carneiro
vinicius.carneiro@ufla.br

Conceitos Importantes

Característica, caráter, caractere ou variável: atributo que define uma observação (unidade ou elementos), **que pode ser uma entidade biológica**.

Exemplos:

- Cor da flor
- Aspecto dos grãos
- Produtividade de grãos



Conceitos Importantes

Tipos de variáveis

Qualitativas (categóricas): são variáveis que fornecem dados de natureza não numérica.

- Nominal: não é possível estabelecer uma ordem natural entre seus valores
 - Cor da flor (Branca, rosa ou roxa)
- Ordinal: assume valores que apresentam uma ordenação natural, indicando intensidades crescentes de realização:
 - Severidade de mancha angular (notas de 1 a 9)
 - Arquitetura de plantas (notas de 1 a 9)

Conceitos Importantes

Tipos de variáveis

Quantitativas: são aquelas cujos dados são valores numéricos que expressam quantidades.

- Discretas: são aquelas em que os dados somente podem apresentar determinados valores, em geral, números inteiros (finito e enumerável).

Exemplos:

- Número de plantas (estande)
- Número de vagens
- Número de sementes



Conceitos Importantes

Tipos de variáveis

Contínuas: são aquelas que podem apresentar qualquer valor dentro de um intervalo de variação possível.

Exemplos:

- Produtividade de grãos
- Brix
- Diâmetro (hipocótilo, haste)
- Altura de planta

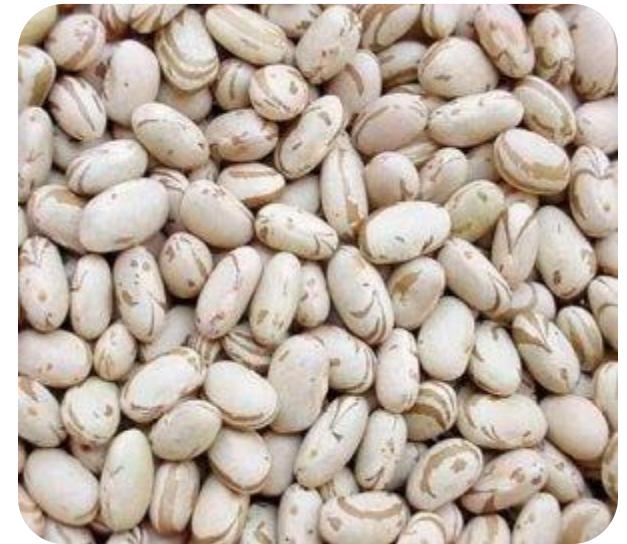


Conceitos Importantes

Fenótipo: particularidade inerente a cada um dos caracteres.

Exemplos:

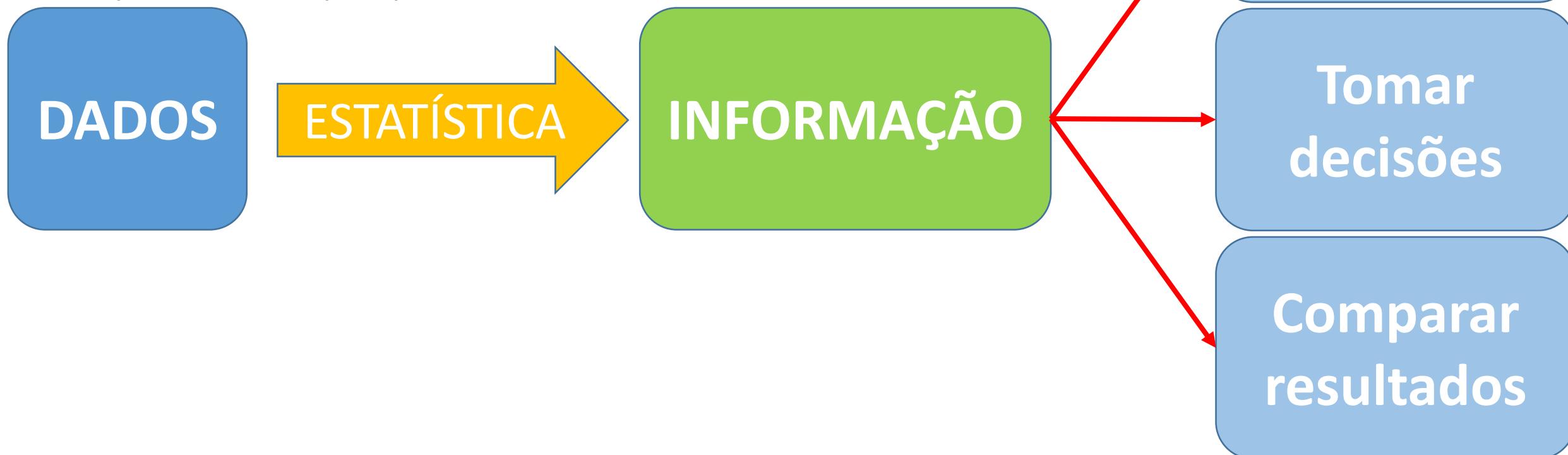
- Cor da flor (branca, rosa ou roxa)
- Produtividade de grãos (500 – 3000 kg/ha)



Conceitos Importantes

Dados: Conjunto de valores (numéricos ou não) que identificam uma observação (unidade ou elemento), **que pode ser uma entidade biológica**.

Informação: conjunto de conhecimentos acumulados sobre certo tema por meio de pesquisa



Como obter os dados?



Industrial Crops & Products 108 (2017) 806–813

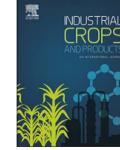


ELSEVIER

Contents lists available at ScienceDirect

Industrial Crops & Products

journal homepage: www.elsevier.com/locate/indcrop



Research Paper

High-performance prediction of macauba fruit biomass for agricultural and industrial purposes using Artificial Neural Networks



Carla Aparecida de O. Castro^a, Rafael T. Resende^{a,*}, Kacilda N. Kuki^c, Vinícius Q. Carneiro^b, Gustavo E. Marcatti^d, Cosme Damião Cruz^b, Sérgio Y. Motoike^c

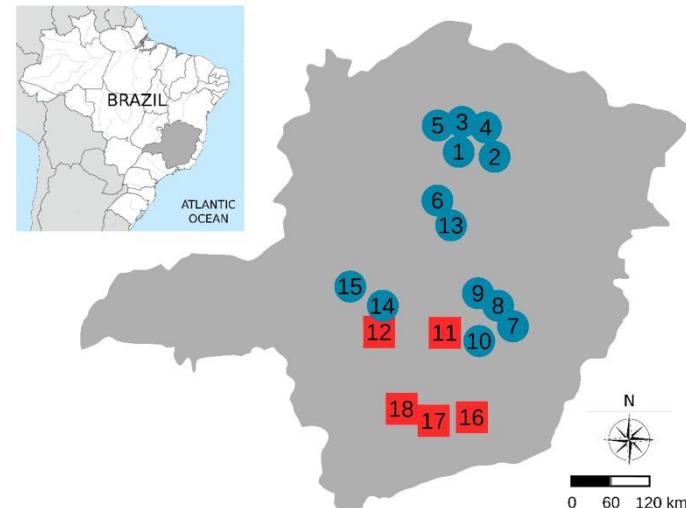


Figura 3. Distribuição das localidades ao longo do território do estado de Minas Gerais (Brasil). As localidades foram agrupadas de acordo com o método k-means e estão relacionadas às coordenadas apresentadas na tabela 1, sendo que os números dentro dos círculos azuis correspondem às localidades do grupo 1 e os números dentro dos quadrados vermelhos são as localidades do grupo 2.



Centro de Desenvolvimento Científico e Tecnológico – Fazenda Muquem - UFLA

Conceitos Importantes

Estatística: Ciência que tem por objetivo obter, organizar, resumir, apresentar, analisar, interpretar dados oriundos de estudos ou experimentos de modo a permitir extrair conclusões relevantes.

Na verdade, mais do que uma sequência de métodos, a estatística é uma forma de pensar e de ver a realidade variável, já que seu conhecimento não apenas fornece um conjunto de técnicas de análise de dados, mas condiciona toda uma postura crítica sobre sua interpretação e a elaboração de conclusões sobre os dados.

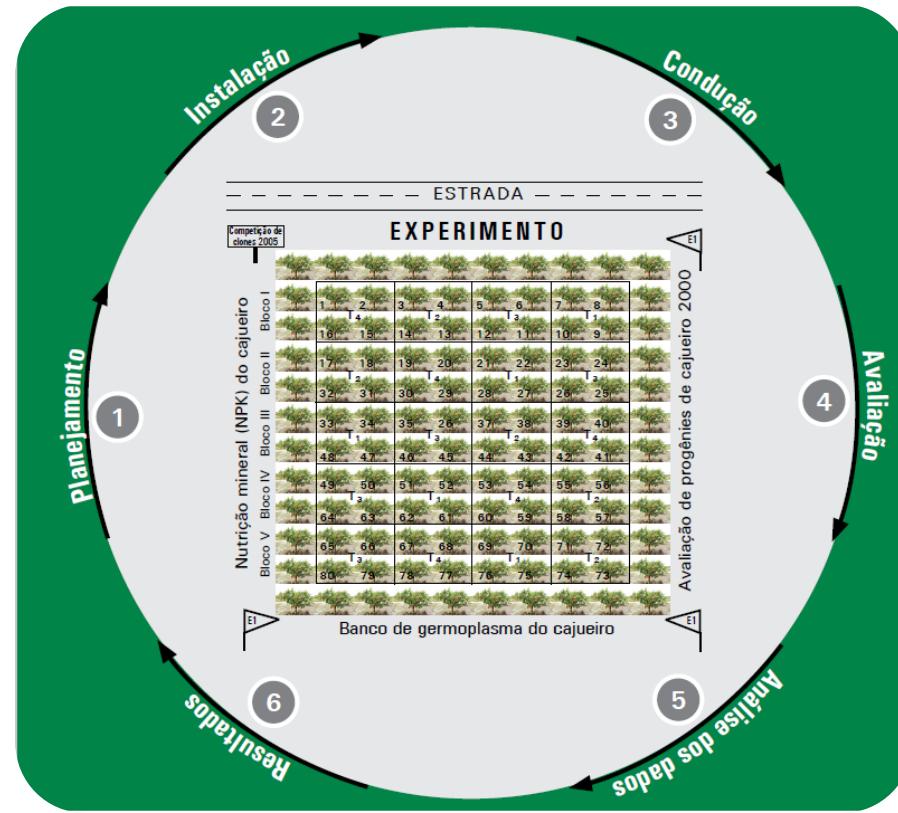
Estatística

- Origem – Latim
 - Estatística – status (Estado);
 - Ocupava-se em descrever quantitativamente os vários aspectos dos assuntos de um governo ou estado.
- Estatística atual – Experimentação
 - Sir Ronald Aylmer Fisher (1890 - 1962);
 - Bases para a experimentação estatisticamente controlada.



Papel da Estatística no Melhoramento de Plantas

- Planejamento experimental: escolha das situações experimentais
- Análise de dados: indica técnicas para resumir e apresentar as informações, bem como para comparar situações de experimentos.



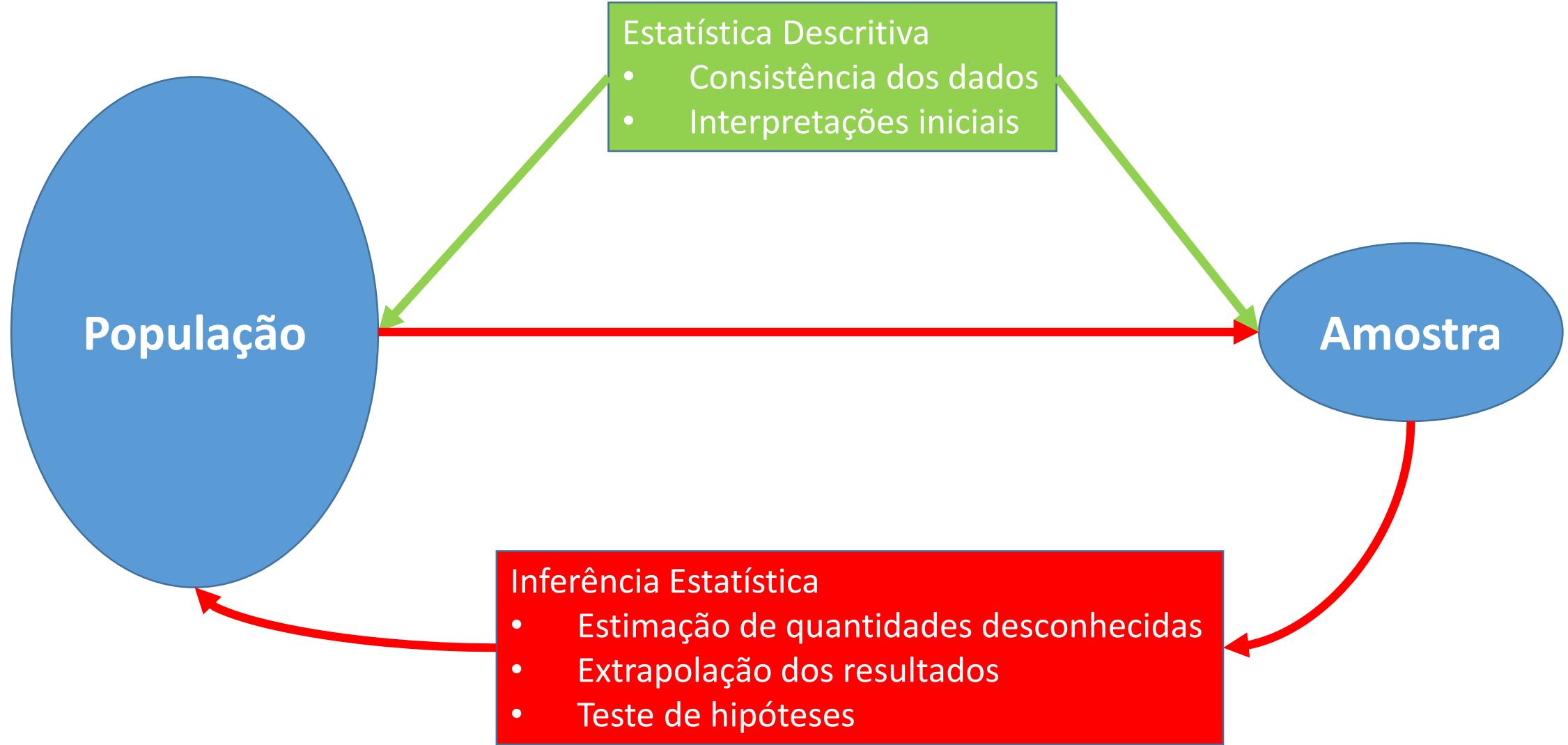
Áreas da estatística

Estatística descritiva: Conjunto de técnicas destinadas a descrever e resumir os dados, a fim de que possamos tirar conclusões.

Inferência estatística: Estudo de técnicas que possibilitam a extração de informações e conclusões, obtidas a partir de subconjuntos de valores (amostras), usualmente de dimensão muito menor.

- População: é uma coleção completa de todos os elementos a serem estudados; ex: População Mundial
- Amostra: é uma subcoleção de elementos extraídos de uma população.

Áreas da Estatística



Estatística Descritiva

Objetivo: Detecção de padrões de interesse nos dados e a sua representação.

- Tabelas
- Gráficos
- Medidas de posição: medidas que representam a tendência central dos dados de modo que são utilizadas para descrever propriedades da população ou amostra.
 - Média, mediana e moda
- Medidas de dispersão: medidas que se aplicam na caracterização de uma distribuição de mensurações.
 - Amplitude, variância, desvio padrão, coeficiente de variação e erro padrão da média.

Medidas de posição

Média Populacional

$$\bar{x} = \mu = \frac{\sum_{i=1}^N x_i}{N} = \frac{x_1 + x_2 + \dots + x_N}{N}$$

$$A_{\text{pop}} = \{10; 20; 20; 50\} \quad \mu = \frac{\sum_{i=1}^4 x_i}{4} = \frac{x_1 + x_2 + x_3 + x_4}{4} = \frac{10 + 20 + 20 + 50}{4} = 25$$

$$\mu = \frac{\sum_{i=1}^{VD} f_i * x_i}{\sum_{i=1}^N f_i} = \frac{f_1 * x_1 + f_2 * x_2 + \dots + f_{VD} * x_{VD}}{1}$$

$$A_{\text{pop}} = \{10; 20; 20; 50\} \quad \mu = \frac{\sum_{i=1}^3 f_i * x_i}{\sum_{i=1}^3 f_i} = \frac{0,25 * 10 + 0,50 * 20 + 0,25 * 50}{0,25 + 0,50 + 0,25} = 25$$

Medidas de posição

Média Amostral

$$\hat{\mu} = \frac{\sum_{i=1}^N x_i}{N} = \frac{x_1 + x_2 + \dots + x_N}{N}$$

$$A_{Am} = \{10; 20; 20\} \quad \hat{\mu} = \frac{\sum_{i=1}^3 x_i}{3} = \frac{x_1 + x_2 + x_3}{3} = \frac{10 + 20 + 20}{3} = 16,67$$

$$\hat{\mu} = \frac{\sum_{i=1}^{VD} f_i * x_i}{\sum_{i=1}^N f_i} = \frac{f_1 * x_1 + f_2 * x_2 + \dots + f_{VD} * x_{VD}}{1}$$

$$A_{Am} = \{10; 20; 20\} \quad \hat{\mu} = \frac{\sum_{i=1}^2 f_i * x_i}{\sum_{i=1}^2 f_i} = \frac{0,33 * 10 + 0,67}{0,33 + 0,67} = 16,67$$

Média - Propriedades

1. A soma dos desvios em relação à média é igual a zero.

$$\sum_{i=1}^N (x_i - \mu) = 0$$

$$A = \{10; 20; 20; 50\}$$

$$\mu = 25$$

$$\sum_{i=1}^N (x_i - \mu) = (10 - 25) + (20 - 25) + (20 - 25) + (50 - 25) = (-15) + (-5) + (-5) + 25 = 0$$

Média - Propriedades

2. A soma ou subtração de uma constante (k) aos dados altera a média de tal forma que a nova média fica adicionada ou subtraída pela constante.

$$k = 10$$

$$A = \{10; 20; 20; 50\}$$

$$\mu = 25$$

Dados Originais	Dados + k	Dados - k
10	20	0
20	30	10
20	30	10
50	60	40
$\mu = 25$	$\mu = 35$	$\mu = 15$

Média - Propriedades

3. A multiplicação ou divisão dos dados por uma constante (k) altera a média de tal forma que a nova média fica multiplicada ou dividida pela constante.

$$k = 10$$

$$A = \{10; 20; 20; 50\}$$

$$\mu = 25$$

Dados Originais	Dados * k	Dados / k
10	100	1
20	200	2
20	200	2
50	500	5
$\mu = 25$	$\mu = 250$	$\mu = 2,5$

Média - Propriedades

4. A média é influenciada por valores extremos

Dados Originais	Novos Dados
10	1000000
20	20
20	20
50	50
$\mu = 25$	$\mu = 250022,5$

Medidas de Posição Mediana

Mediana: valor que ocupa a posição central dos dados ordenados.

n: número de observações

$$\text{Se } n \text{ for ímpar: } md = x_{\left(\frac{n+1}{2}\right)}$$

Dados	Ordenados
10	10
50	20
20	50
$Md = 20$	

$$\text{Se } n \text{ for par: } md = \frac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n+2}{2}\right)}}{2}$$

Dados	Ordenados
10	10
20	20
50	20
20	50
$Md = 20$	

Mediana - Propriedades

1. A soma ou subtração de uma constante k aos dados altera a mediana de tal forma que a nova mediana fica adicionada ou subtraída pela constante

$$k = 10$$

Dados	Ordenados	Dados + k
10	10	20
50	20	30
20	50	60
$Md = 20$		$Md = 30$

Mediana - Propriedades

1. A multiplicação ou divisão dos dados por uma constante k altera a mediana de tal forma que a nova mediana fica multiplicada ou dividida pela constante.

$$k = 10$$

Dados	Ordenados	Dados * k	Dados/k
10	10	100	1
50	20	200	2
20	50	500	5
$Md = 20$		$Md = 200$	$Md = 2$

Medidas de Posição - Moda

Moda: valor mais frequente de um conjunto de dados

Propriedades: $k = 10$

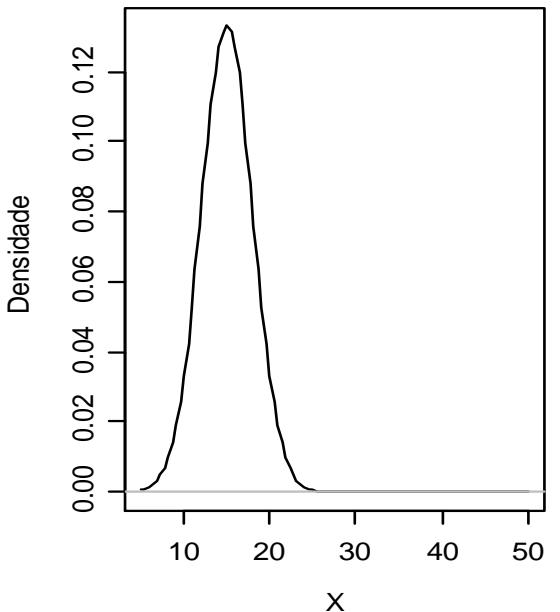
Dados Originais	Dados + k	Dados - k	Dados*k	Dados/k
10	20	0	100	1
20	30	10	200	2
20	30	10	200	2
50	60	50	500	5
Moda = 20	Moda = 30	Moda = 10	Moda = 200	Moda = 2

Medidas de Dispersão

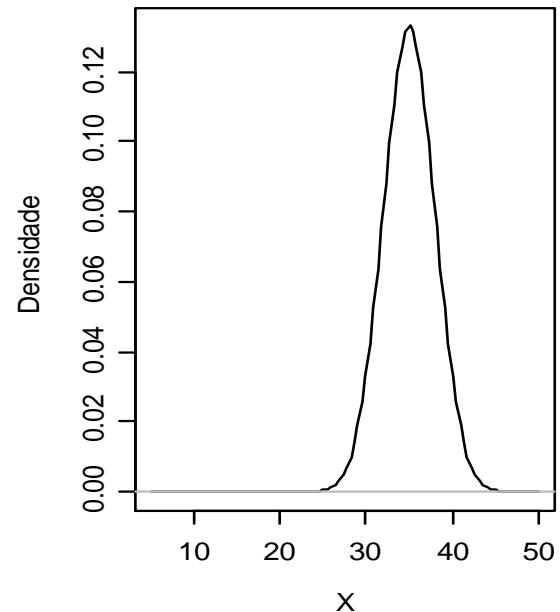
*As **medidas de posição** são importantes para caracterizar um conjunto de mensurações, mas não são suficientes para caracterizar completamente a distribuição dos dados.*

Medidas de dispersão se aplicam na caracterização de uma distribuição de mensurações.

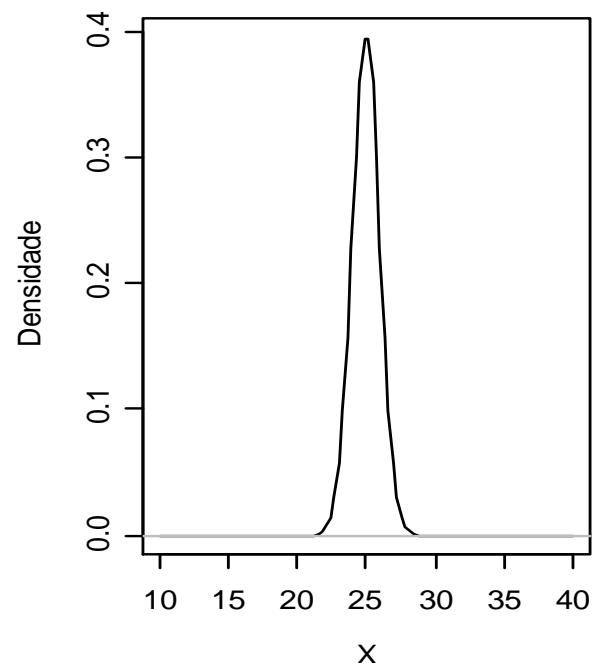
Normal: $\mu = 15$, $\sigma = 3$



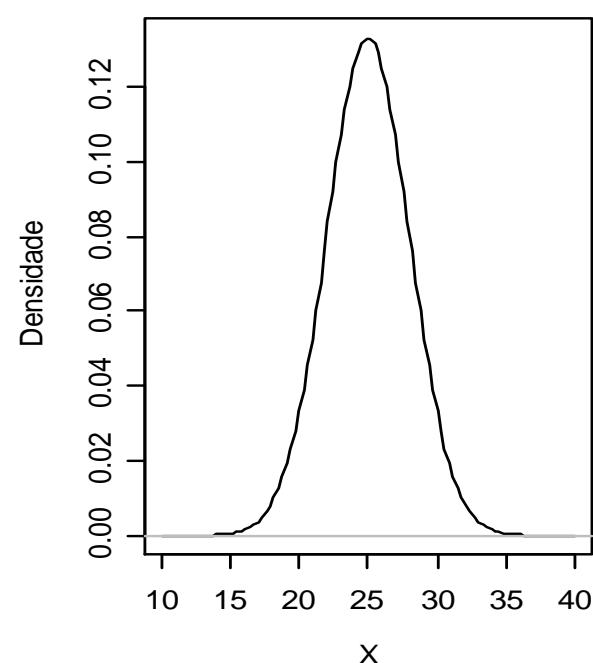
Normal: $\mu = 35$, $\sigma = 3$



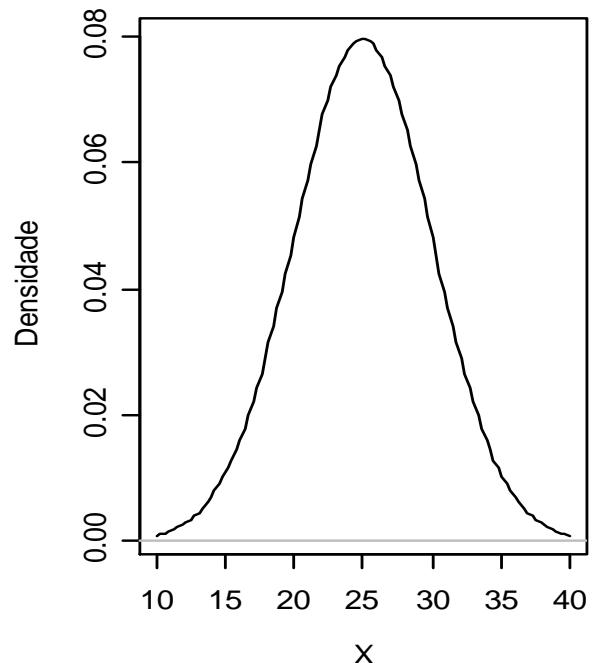
Normal: $\mu = 25$, $\sigma = 1$



Normal: $\mu = 25$, $\sigma = 3$



Normal: $\mu = 25$, $\sigma = 5$



Medidas de Dispersão - Amplitude

Amplitude: diferença entre o maior valor e o menor valor de um conjunto de dados.

Vantagens:

- Rápida e fácil indicação da variabilidade dos dados.

Desvantagens:

- não considera todos os valores;
- Muito influenciada por valores extremos (outliers)

Dados	Dados
10	10
20	1000000
20	20
50	50
Amplitude = 40	
Amplitude = 999990	

Medidas de Dispersão

Variância Populacional

Variância Populacional: medida de dispersão dos valores em torno da média da população.

$$\bar{X} = 0,204$$

$$\sigma^2 = \frac{\sum_{i=1}^N (X_i - \bar{X})^2}{N}$$

$$\sigma^2 = \frac{0,00052}{5} = 0,000104$$

X_i	Desvio ($X_i - \bar{X}$)	$(X_i - \bar{X})^2$
20% = 0,20	$0,20 - 0,204 = -0,004$	0,000016
21% = 0,21	$0,21 - 0,204 = 0,006$	0,000036
22% = 0,22	$0,22 - 0,204 = 0,016$	0,000256
20% = 0,20	$0,20 - 0,204 = -0,004$	0,000016
19% = 0,19	$0,19 - 0,204 = -0,014$	0,000196
Total		0,00052

Medidas de Dispersão

Variância Amostral

Variância Amostral: medida de dispersão dos valores em torno da média da amostra. É obtida pela soma dos quadrados dos desvios em relação à média aritmética, dividida pelo número de graus de liberdade.

$$S^2(X) = \hat{V}(X) = \frac{SQD_X}{n-1} = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1} = \frac{\sum_{i=1}^n X_i^2 - \left(\sum_{i=1}^n X_i\right)^2}{n-1}$$

Dados	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
100	-20	400
150	30	900
130	10	100
100	-20	400
$\bar{x} = 120$	$\sum_{i=1}^N (x_i - \bar{x}) = 0$	$\hat{\sigma}^2 = \frac{\sum_{i=1}^4 (x_i - \bar{x})^2}{4-1} = 600$

Medidas de Dispersão

Variância Amostral

$$S^2(X) = \hat{V}(X) = \frac{SQD_X}{n-1} = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1} = \frac{\sum_{i=1}^n X_i^2 - \frac{\left(\sum_{i=1}^n X_i\right)^2}{n}}{n-1}$$

Amostra A: 4, 8, 3, 9, 7, 5

Amostra B: 1, 5, 2, 14, 3, 11

Para o nosso exemplo, temos:

$$S_A^2 = \frac{244 - \frac{(36)^2}{6}}{6-1} = 5,6$$

$$S_B^2 = \frac{356 - \frac{(36)^2}{6}}{6-1} = 28$$

Medidas de Dispersão - Variância

Graus de Liberdade

É possível demonstrar que, utilizando-se o denominador $n - 1$, obtém-se um estimador não tendencioso da variância populacional, isto é, $E(S^2) = \sigma^2$.

De uma maneira geral, o número de graus de liberdade associados a uma estatística é o número de elementos da amostra, n , menos o número de parâmetros (medidas da população) já estimados. Existem $n - 1$ desvios independentes.

$$S^2(X) = \hat{V}(X) = \frac{SQD_X}{n-1} = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1} = \frac{\sum_{i=1}^n X_i^2 - \frac{\left(\sum_{i=1}^n X_i\right)^2}{n}}{n-1}$$

Variância - Propriedades

Algumas propriedades úteis da variância:

- (i) A variância é sempre maior ou igual a zero, isto é, $S^2(X) \geq 0$;
- (ii) Para $X = k$, sendo k uma constante, $S^2(X) = 0$;
- (iii) Para $Y = X + k$, sendo k uma constante, $S^2(Y) = S^2(X)$;
- (iv) Para $Y = kX$, sendo k uma constante, $S^2(Y) = k^2S^2(X)$.

Variância - Propriedades

(iii) Para $Y = X + k$ ou $Y = X - k$, sendo k uma constante, $\hat{\sigma}^2(Y) = \hat{\sigma}^2(X)$

$$k = 10$$

Dados Originais (X)	Dados + k (Y = X + k)	Dados – k (Y = X – k)
100	110	90
150	160	140
130	140	120
100	110	90
$\hat{\sigma}^2(X) = 600$	$\hat{\sigma}^2(Y) = 600$	$\hat{\sigma}^2(Y) = 600$

Variância - Propriedades

(iv) Para $Y = kX$, sendo k uma constante, $\hat{\sigma}^2(Y) = k^2 \hat{\sigma}^2(X)$

(v) Para $Y = \frac{X}{k}$, sendo k uma constante, $\hat{\sigma}^2(Y) = \frac{\hat{\sigma}^2(X)}{k^2}$

$k = 10$

Dados Originais (X)	Dados * k (Y = X * k)	Dados / k (Y = X / k)
100	1000	10
150	1500	15
130	1300	13
100	1000	10
$\hat{\sigma}^2(X) = 600$	$\hat{\sigma}^2(Y) = 60000$	$\hat{\sigma}^2(Y) = 6$

Variância - Propriedades

Desvantagem:

- A unidade de medida da variância é o quadrado da unidade de medida original.

Medidas de Dispersão Desvio Padrão Amostral

$$S(X) = \sqrt{\hat{V}(X)}$$

Desvio Padrão - Propriedades

$$S(X) = \sqrt{\hat{V}(X)}$$

Algumas propriedades úteis da variância:

- (i) A variância é sempre maior ou igual a zero, isto é, $S^2(X) \geq 0$;
- (ii) Para $X = k$, sendo k uma constante, $S^2(X) = 0$;
- (iii) Para $Y = X + k$, sendo k uma constante, $S^2(Y) = S^2(X)$;
- (iv) Para $Y = kX$, sendo k uma constante, $S^2(Y) = k^2S^2(X)$.

Desvio Padrão - Propriedades

k = 10

Dados Originais (X)	Dados + k (Y = X+k)	Dados – k (Y = X – k)	Dados Originais (X)	Dados * k (Y = X * k)	Dados / k (Y = X / k)
100	110	90	100	1000	10
150	160	140	150	1500	15
130	140	120	130	1300	13
100	110	90	100	1000	10
$\hat{\sigma}^2(X) = 600$	$\hat{\sigma}^2(Y) = 600$	$\hat{\sigma}^2(Y) = 600$	$\hat{\sigma}^2(X) = 600$	$\hat{\sigma}^2(Y) = 60000$	$\hat{\sigma}^2(Y) = 6$
$\hat{\sigma}(X) = 24,49$	$\hat{\sigma}(Y) = 24,49$	$\hat{\sigma}(Y) = 24,49$	$\hat{\sigma}(X) = 24,49$	$\hat{\sigma}(Y) = 244,9$	$\hat{\sigma}(Y) = 2,45$

Medidas de Dispersão

Coeficiente de Variação

Frequentemente, se tem o interesse em comparar variabilidades de diferentes conjuntos de valores. A comparação se torna difícil em situações onde as médias são muito desiguais ou as unidades de medida são diferentes. Nesses casos, o *CV* é indicado por ser uma medida de dispersão relativa.

$$CV(\%) = \frac{S(X)}{\bar{X}} \cdot 100$$

Note que o *CV* é o desvio-padrão expresso em percentagem da média. É uma medida adimensional.

Aplicação:

- Utilizado para avaliação da precisão de experimentos;
- Utilizado para analisar qual amostra é mais homogênea (menor variabilidade). Na situação em que as amostras possuem a mesma média, a conclusão pode ser feita a partir da comparação de suas variâncias. Para amostras com médias diferentes, aquela que apresentar menor CV, é a mais homogênea.

Medidas de Dispersão

Coeficiente de Variação

Qual amostra mais homogênea?

Amostra A	Amostra B
$n_A = 50$	$n_A = 60$
$\bar{X}_A = 65$	$\bar{X}_A = 70$
$S^2(A) = 225$	$S^2(B) = 235$

$$CV(\%) = \frac{S(X)}{\bar{X}} \cdot 100$$

Solução:

$$CV_A = \frac{100\sqrt{225}}{65} = 23,08\%$$

$$CV_B = \frac{100\sqrt{235}}{70} = 21,90\%$$

Medidas de Dispersão

Erro Padrão da Média

- Medida da dispersão das médias amostrais em torno da média da população.
- Estimador da precisão da estimativa de uma média populacional.

É uma medida utilizada para avaliar a precisão da média. É dada por:

$$S(\bar{X}) = \sqrt{\hat{V}(\bar{X})} = \sqrt{\frac{\hat{V}(X)}{n}} = \frac{\sqrt{\hat{V}(X)}}{\sqrt{n}} = \frac{s(X)}{\sqrt{n}}$$

Exemplo:

Considerando $S_A^2 = 5,6$ e $S_B^2 = 28$

$n = 6$

$$S(\bar{X}_A) = \frac{2,3664}{\sqrt{6}} = 0,966$$

$$S(\bar{X}_B) = \frac{5,2915}{\sqrt{6}} = 2,1602$$

Medidas de Dispersão

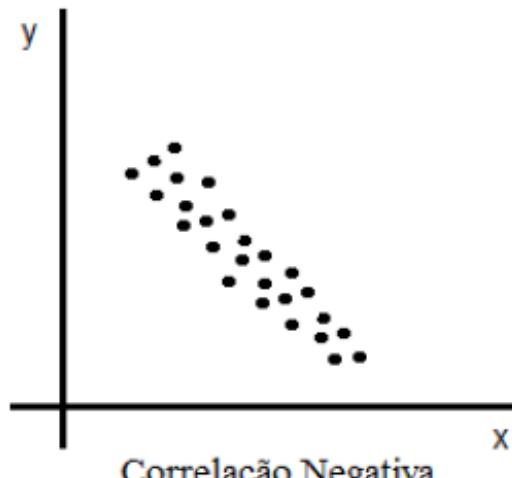
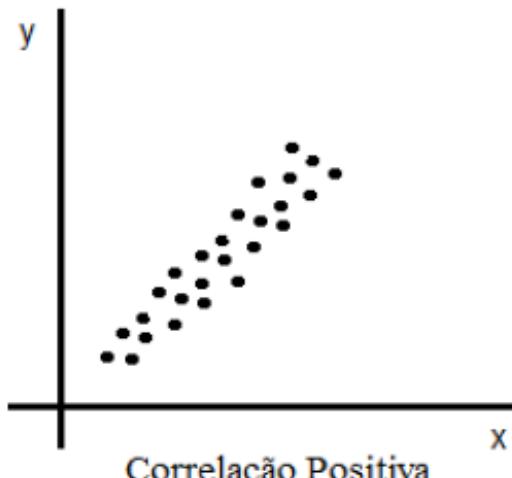
Coeficiente de Correlação

- Medida do grau de associação linear entre duas variáveis.

Sejam duas amostras relativas às variáveis X e Y , dadas a seguir:

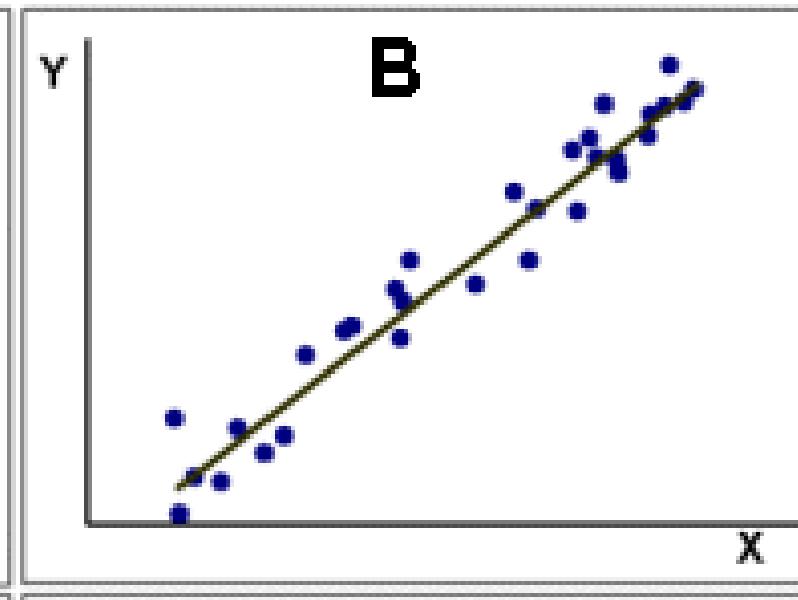
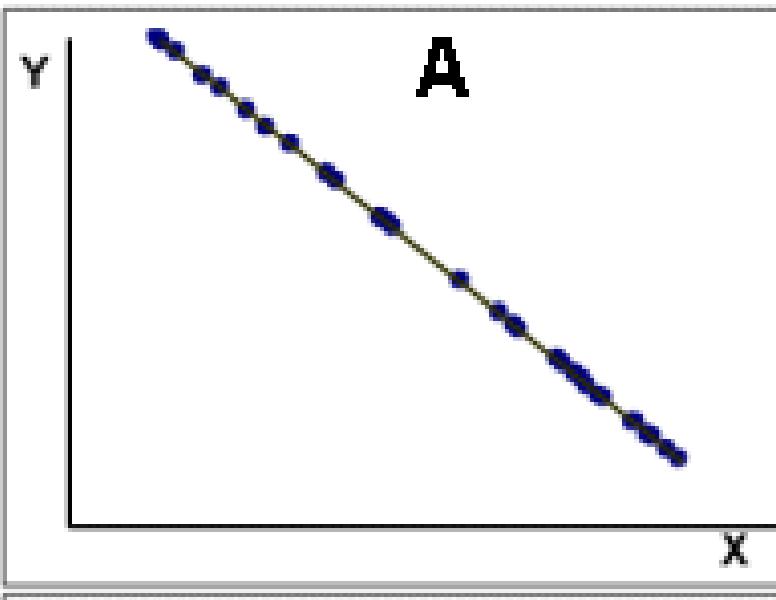
X_i	X_1	X_2	...	X_n
Y_i	Y_1	Y_2	...	Y_n

O coeficiente de correlação entre os valores de X e Y é dado por:



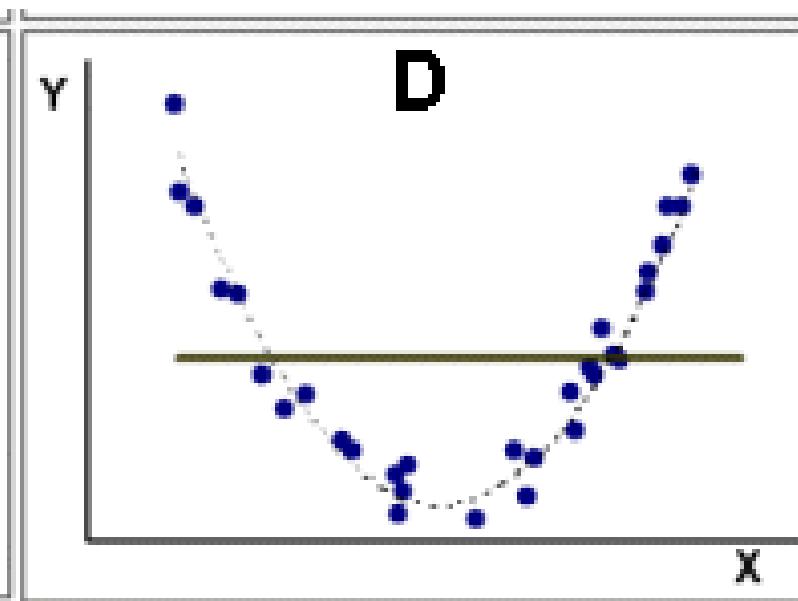
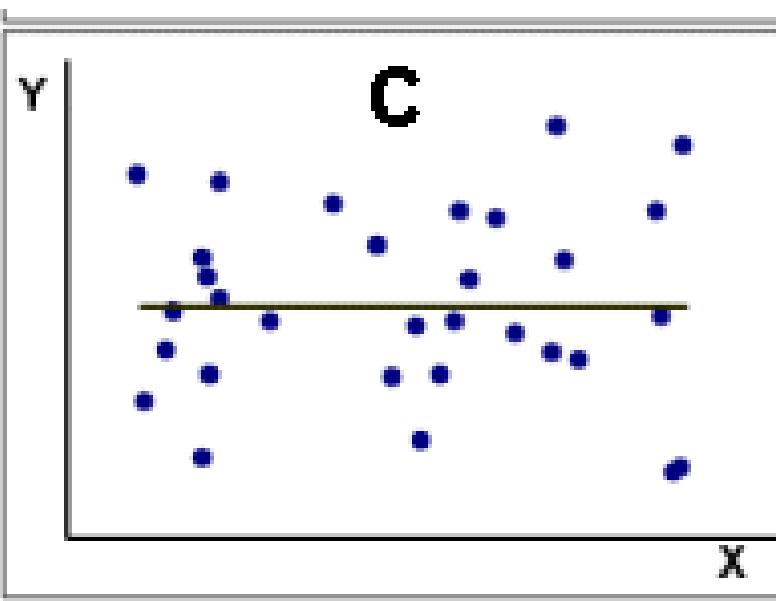
$r < 0$

$r > 0$



$r = 0$

$r = 0$



Medidas de Dispersão

Coeficiente de Correlação

$$r_{XY} = \frac{C\hat{O}V(X, Y)}{\sqrt{\hat{V}(X) \cdot \hat{V}(Y)}} = \frac{\frac{SPD_{XY}}{n-1}}{\sqrt{\frac{SQD_X}{n-1} \cdot \frac{SQD_Y}{n-1}}} \quad -1 \leq r_{XY} \leq 1$$

$$SPD_{XY} = \sum_{i=1}^n X_i Y_i - \frac{\left(\sum_{i=1}^n X_i \right) \left(\sum_{i=1}^n Y_i \right)}{n}$$

$$SQD_X = \sum_{i=1}^n X_i^2 - \frac{\left(\sum_{i=1}^n X_i \right)^2}{n} \quad \text{e} \quad SQD_Y = \sum_{i=1}^n Y_i^2 - \frac{\left(\sum_{i=1}^n Y_i \right)^2}{n}$$

Medidas de Dispersão

Coeficiente de Correlação

Exemplo:

Amostra A	4	8	3	9	7	5
Amostra B	1	5	2	14	3	11

$$SPD_{AB} = \sum_{i=1}^n A_i B_i - \frac{\left(\sum_{i=1}^n A_i\right) \left(\sum_{i=1}^n B_i\right)}{n} = 252 - \frac{(36)(36)}{6} = 36$$

$$SQD_A = \sum_{i=1}^n A_i^2 - \frac{\left(\sum_{i=1}^n A_i\right)^2}{n} = 244 - \frac{(36)^2}{6} = 28$$

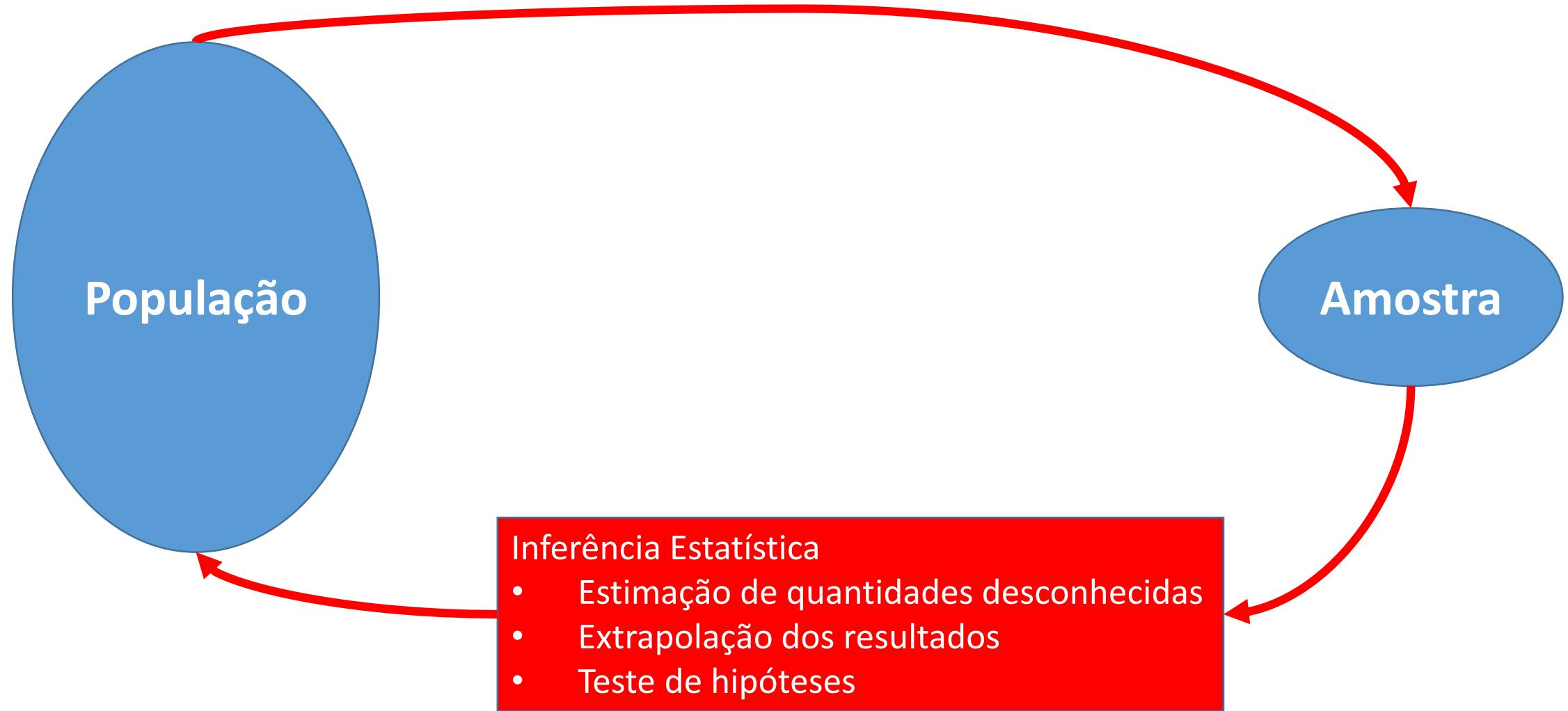
$$SQD_B = \sum_{i=1}^n B_i^2 - \frac{\left(\sum_{i=1}^n B_i\right)^2}{n} = 356 - \frac{(36)^2}{6} = 140$$

$$r_{AB} = \frac{SP_{AB}}{\sqrt{SQD_A \cdot SQD_B}} = \frac{36}{\sqrt{(28) \cdot (140)}} = 0,5750$$

A	A^2	B	B^2	$A * B$
4	16	1	1	4
8	64	5	25	40
3	9	2	4	6
9	81	14	196	126
7	49	3	9	21
5	25	11	121	55

$$\sum_{i=1}^6 A_i = 36 \quad \sum_{i=1}^6 A_i^2 = 244 \quad \sum_{i=1}^6 B_i = 36 \quad \sum_{i=1}^6 B_i^2 = 356 \quad \sum_{i=1}^6 A_i * B_i = 252$$

Inferência Estatística



Inferência Estatística

- Objetivo: Realizar generalizações sobre uma população com base em dados de uma amostra.
- Formas de generalização:
 - **Estimar um único valor – Estimação por ponto;**
 - **Exemplo: Testar o inseticida para saber a dose que mata 90% dos insetos.**
 - **Estimar uma amplitude de valores numéricos – Estimação por intervalo;**
 - **Exemplo: Descobrir um intervalo de doses para matar pelo menos 50% dos insetos.**
 - **Teste de hipótese.**
 - **Exemplo: O inseticida novo é melhor que os existentes no mercado?**

Conceitos importantes

Parâmetro é uma função de valores populacionais, sendo em geral, um valor desconhecido associado à população.

Um estimador de um parâmetro θ é qualquer função das observações da amostra aleatória X_1, X_2, \dots, X_n . Ele representa uma dada fórmula de cálculo que fornecerá valores que serão diferentes, conforme a amostra selecionada.

$$1. \text{ O estimador da média } \mu \text{ é } \hat{\mu} = \bar{X} = \frac{\sum_{i=1}^n X_i}{n}$$

$$2. \text{ O estimador da variância } \sigma^2 \text{ é } \hat{\sigma}^2 = s^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1} = \frac{\sum_{i=1}^n X_i^2 - \frac{(\sum_{i=1}^n X_i)^2}{n}}{n-1}$$

A Estimativa é o valor numérico assumido pelo estimador, quando os valores x_1, x_2, \dots, x_n são considerados.

Exemplos: $\bar{X} = 10,42$ e $s^2 = 4,67$.

Propriedades de Estimadores

1.

O estimador deve ser não viesado.

- Se forem retiradas todas as amostras de tamanho n de uma população, a média de todas as estimativas obtidas em todas as amostras possíveis será igual ao valor do parâmetro que se deseja estimar.

$$A = \{1,2,3\}$$

$$\mu = 2$$

$$\sigma^2 = 0,6666$$

$$\sigma = 0,816496$$

$$E(\bar{X}) = \mu$$

Amostras (n = 2)	$\hat{\mu}$	$\hat{\sigma}^2$	$\hat{\sigma}$
1;1	1	0	0
1;2	1,5	0,5	0,707
1;3	2	2	1,414
2;1	1,5	0,5	0,707
2;2	2	0	0
2;3	2,5	0,5	0,707
3;1	2	2	1,414
3;2	2,5	0,5	0,707
3;3	3	0	0
Média	2	0,6666	0,6284

Propriedades de Estimadores

2. O estimador deve ser consistente.
 - A variância do estimador deve tender a zero, quando n aumenta, tendendo para o infinito. $Var(\bar{X}) = \frac{Var(X)}{n}$
3. O estimador deve ser eficiente:
 - O estimador de maior eficiência é aquele que possui menor variância.

Entre a média amostral (m) e a mediana amostral, qual o estimador mais eficiente para a média populacional μ ?

$$\hat{V}(m) = \frac{\sigma^2}{n} \qquad \qquad \hat{V}(md) = \frac{\pi\sigma^2}{2n}$$

Estimação por Intervalo

Utiliza a informação amostral para obter dois valores, entre os quais se pretende que esteja o parâmetro de interesse, com certo nível de probabilidade ($1 - \alpha$).

Intervalo de confiança para a média populacional

$$P\left(m - t_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}} < \mu < m + t_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}}\right) = 1 - \alpha$$

$$IC = m \pm t_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}}$$

Exemplo

Calcule o intervalo de confiança para a média dos dados de produção de grãos de feijão, g/planta, da geração F₂ do cruzamento entre as cultivares Pompadour x Carioca MG.

5.68	2.95	1.42	12.07	3.55	38.15	11.52	17.2	15.58	5.74
10.99	3.83	33.86	4.18	5.63	18.64	12.98	23.18	24.22	4.26
7.71	1.33	11.11	7.07	3.27	10.25	19.2	16.95	9.29	1.35
0.71	18.77	2.77	2.03	6.45	29.76	3.38	18.73	3.93	6.79
0.95	25.69	17.49	4.06	13.18	3.1	6.43	11.92	12.72	6.22
12.4	10.6	6.21	3.76	3.36	5.32	2.93	13.35	3.96	32.6
21.2	10.63	20.25	0.53	19.72	13.57	21.17	7.15	19.61	11.96
7.71	17.96	5.11	16.74	7.98	45.6	3.48	14.74	24.77	8.25
6.69	15.43	2.6	5.14	5.49	9.39	9.72	3.11	4.37	2.76
11.54	9.9	5.47	7.77	15.27	21.59	4.34	4.72	15.78	24.51

Exemplo

$$IC_{0,95} = m \pm t_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}}$$

$$\hat{\mu} = \frac{\sum_{i=1}^N x_i}{n} = \frac{5,68 + 10,99 + \dots + 24,51}{100} = 11,1645 \text{ g/planta}$$

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n x_i^2 - \frac{(\sum_{i=1}^n x_i)}{n}}{n-1} = \frac{(5,68^2 + 10,99^2 + \dots + 24,51^2) - \frac{(1116,45)^2}{100}}{99} = 77,3901 \left(\frac{g}{planta} \right)^2$$

$$\hat{s}_{(\hat{\mu})} = \sqrt{\frac{s^2}{n}} = \sqrt{\frac{77,3901}{100}} = 0,8797 \text{ g/planta} \quad t_{\frac{\alpha}{2}} = t_{\frac{0,05}{2}} = t_{0,025} = 1,987$$

$$IC_{0,95} = 11,1645 \pm 1,987 * 0,8797 = 11,1645 \pm 1,7480$$

$$IC_{0,95} = [9,4165; 12,9125]$$

Tabela A2 Valores críticos da distribuição t de Student

gl	α Bilateral:	0,40	0,20	0,10	0,05	0,02	0,01	0,001
	α Unilateral:	0,20	0,10	0,05	0,025	0,01	0,005	0,0005
1		1,376	3,078	6,314	12,706	31,821	63,656	636,578
2		1,061	1,886	2,920	4,303	6,965	9,925	31,600
3		0,978	1,638	2,353	3,182	4,541	5,841	12,924
4		0,941	1,533	2,132	2,776	3,747	4,604	8,610
5		0,920	1,476	2,015	2,571	3,365	4,032	6,869
6		0,906	1,440	1,943	2,447	3,143	3,707	5,959
7		0,896	1,415	1,895	2,365	2,998	3,499	5,408
8		0,889	1,397	1,860	2,306	2,896	3,355	5,041
9		0,883	1,383	1,833	2,262	2,821	3,250	4,781
10		0,879	1,372	1,812	2,228	2,764	3,169	4,587
11		0,876	1,363	1,796	2,201	2,718	3,106	4,437
12		0,873	1,356	1,782	2,179	2,681	3,055	4,318
13		0,870	1,350	1,771	2,160	2,650	3,012	4,221
14		0,868	1,345	1,761	2,145	2,624	2,977	4,140
15		0,866	1,341	1,753	2,131	2,602	2,947	4,073
16		0,865	1,337	1,746	2,120	2,583	2,921	4,015
17		0,863	1,333	1,740	2,110	2,567	2,898	3,965
18		0,862	1,330	1,734	2,101	2,552	2,878	3,922
19		0,861	1,328	1,729	2,093	2,539	2,861	3,883
20		0,860	1,325	1,725	2,086	2,528	2,845	3,850
21		0,859	1,323	1,721	2,080	2,518	2,831	3,819
22		0,858	1,321	1,717	2,074	2,508	2,819	3,792
23		0,858	1,319	1,714	2,069	2,500	2,807	3,768
24		0,857	1,318	1,711	2,064	2,492	2,797	3,745
25		0,856	1,316	1,708	2,060	2,485	2,787	3,725
26		0,856	1,315	1,706	2,056	2,479	2,779	3,707
27		0,855	1,314	1,703	2,052	2,473	2,771	3,689
28		0,855	1,313	1,701	2,048	2,467	2,763	3,674
29		0,854	1,311	1,699	2,045	2,462	2,756	3,660
30		0,854	1,310	1,697	2,042	2,457	2,750	3,646
40		0,851	1,303	1,684	2,021	2,423	2,704	3,551
60		0,848	1,296	1,671	2,000	2,390	2,660	3,460
120		0,845	1,289	1,658	1,980	2,358	2,617	3,373
infinito		0,842	1,282	1,645	1,960	2,326	2,576	3,290

O que é uma hipótese? (Estatística)

Suposição quanto ao valor de um parâmetro populacional ou uma afirmação quanto a natureza de uma população.

Exemplos

Ensaios de competição entre linhagens, híbridos, clones.

Ensaios de valor de cultivo e uso (VCU)



Existem linhagens de feijão superiores as demais avaliadas em um ensaio de competição.

Avaliação de progênie visando seleção.

Intuito é estimar GS.



Existe variabilidade devido às causas genéticas entre os genótipos avaliados

Hipótese de Nulidade H_0

É a hipótese estatística a ser testada.

A hipótese H_0 é formulada com o “expresso propósito de ser rejeitada”, e os testes são construídos sob a pressuposição de H_0 ser verdadeira. O teste de hipótese consiste em verificar se a amostra observada difere significativamente do resultado esperado sob H_0 .

$H_0: m_1 = m_2 = \dots = m_t$; todos os possíveis contrastes entre as médias das linhagens avaliadas em um ensaio de competição, são estatisticamente nulos.

Hipótese Alternativa H_a

H_a : É uma hipótese que contraria H_0 , formulada com base no conhecimento prévio do problema, informações de pesquisa, etc.

H_a : pelo menos uma das médias das linhagens avaliadas em um ensaio de competição, é diferente das demais.

Teste de Hipótese

Teste de hipótese é um procedimento que mediante informações obtidas de amostras, permite decidir aceitar ou rejeitar H_0 .

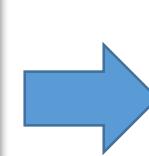
Erro tipo I (α):

O erro tipo I é caracterizado pelo fato de rejeitarmos H_0 quando esta é verdadeira. Designaremos por α a probabilidade de se cometer o erro tipo I. O valor α é chamado nível de significância do teste. Em geral, os valores mais utilizados de α são 1% e 5%.

Erro tipo II (β):

O erro tipo II é caracterizado pelo fato de aceitarmos H_0 quando esta é falsa.

Como Testar as Hipóteses?

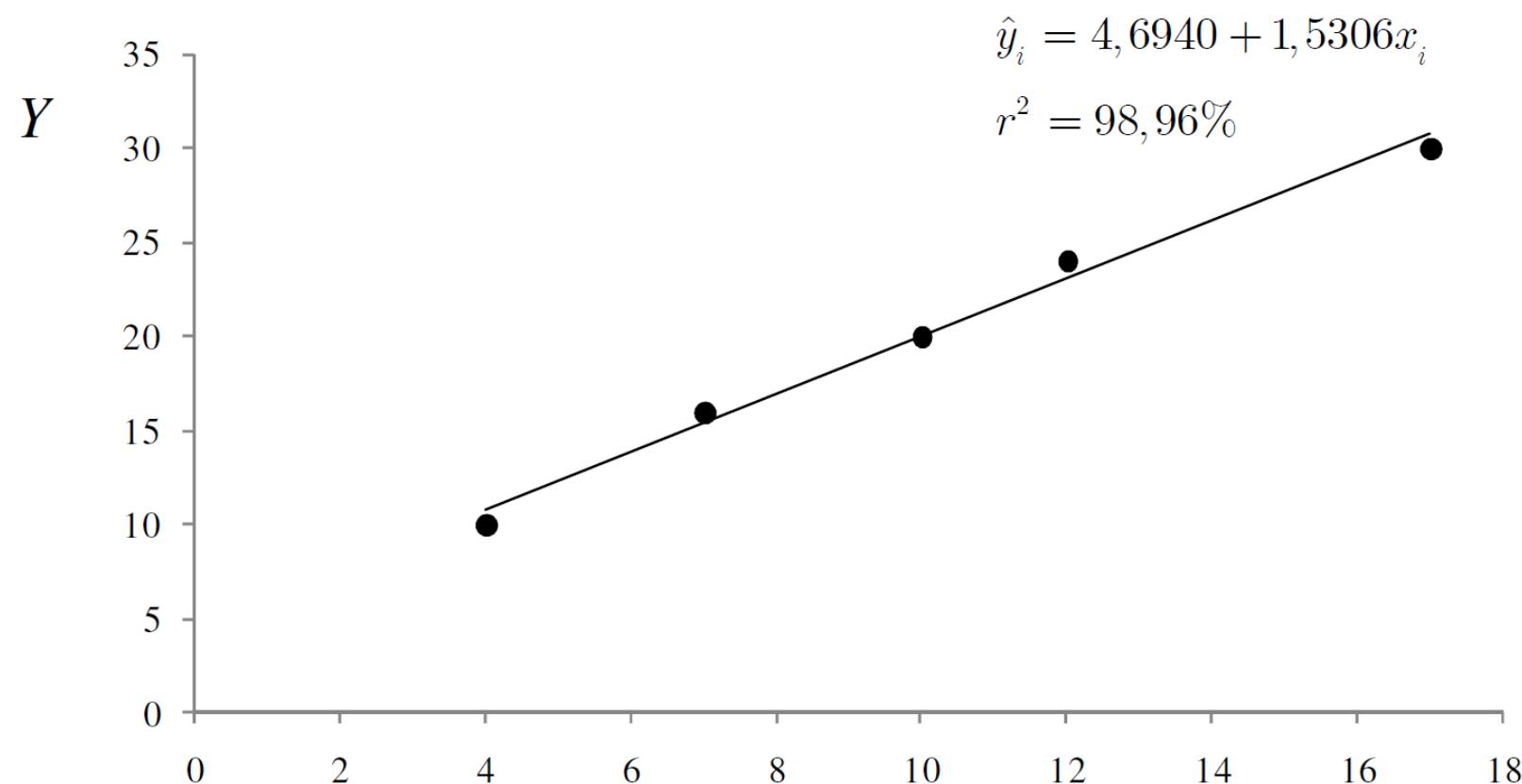


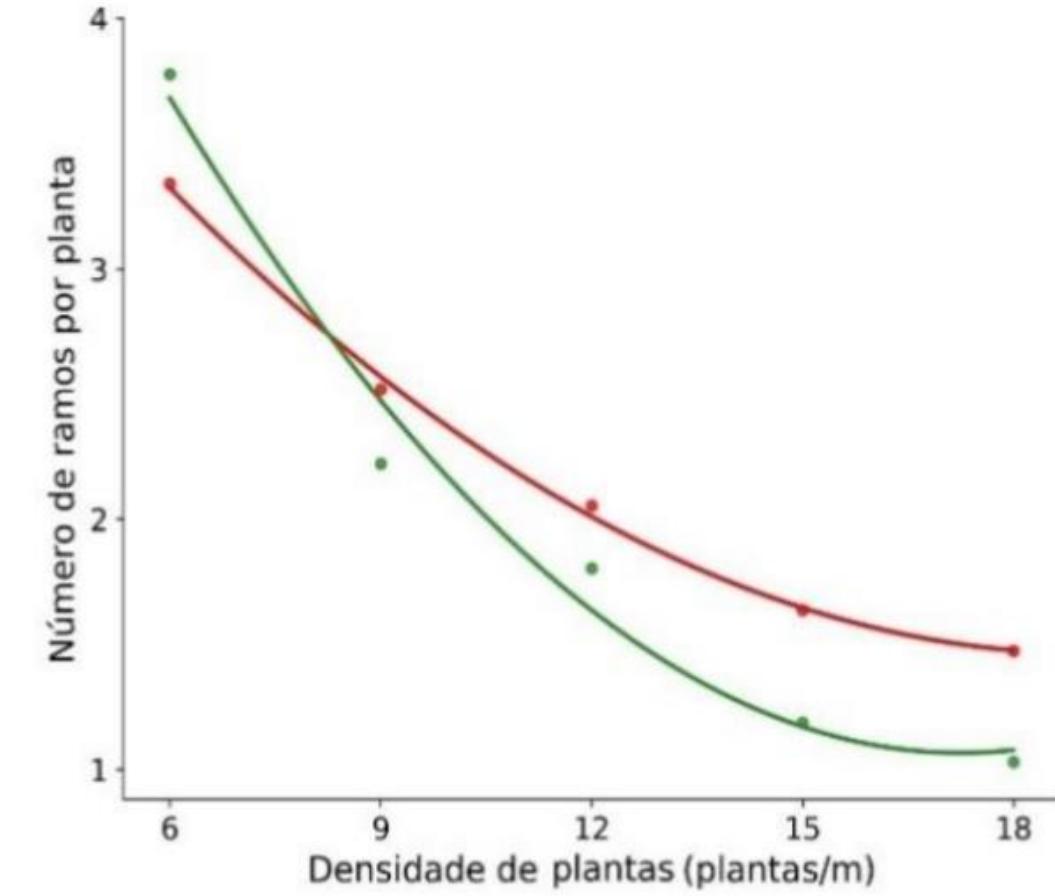
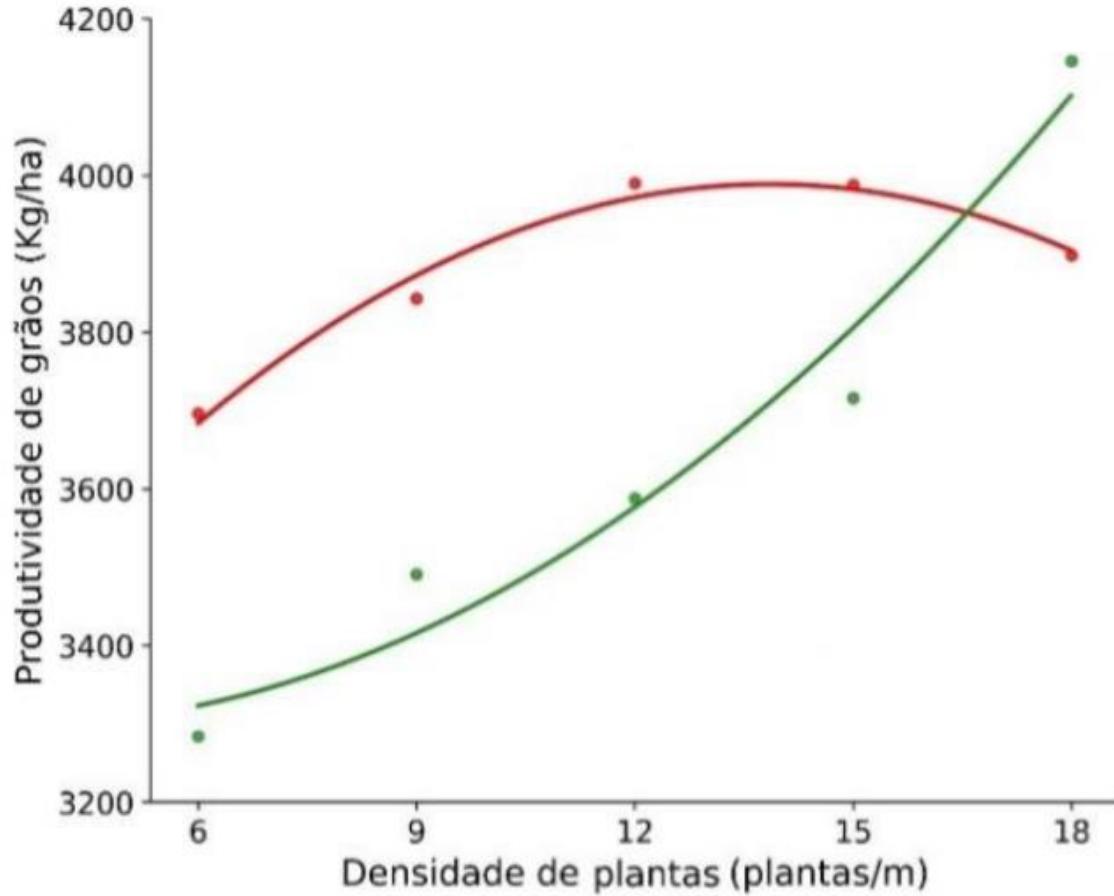
F
t
 χ^2

Procedimentos para realização de um teste de hipótese

1. Enunciar as hipóteses H_0 e H_a ;
2. Fixar o nível de significância α e identificar a estatística do teste;
3. Determinar a região crítica e a região de aceitação em função do nível α pelas Tabelas estatísticas;
4. Por meio dos elementos amostrais, calcular o valor da estatística do teste;
5. Concluir pela rejeição ou não-rejeição de H_0 , caso o valor da estatística obtido no 4º passo pertença ou não pertença, respectivamente, à região crítica determinada no 3º passo.

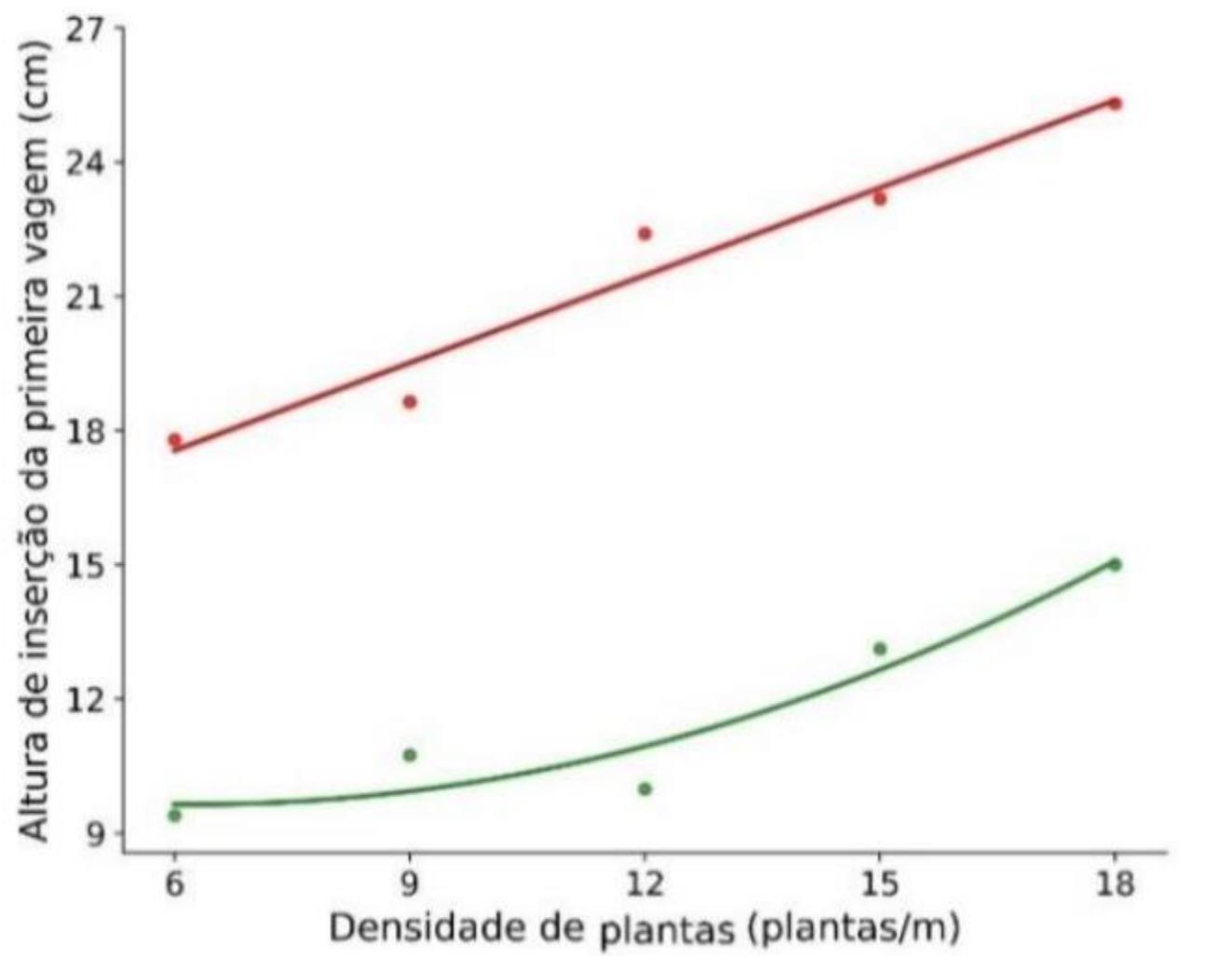
Regressão Linear Simples





A

B



(VR 22) $\hat{Y} = 13,6442 + 0,65171^{**}D; r^2 = 0,9561$
(VR 20) $\hat{Y} = 11,2038 - 0,496698^{ns}D + 0,039553^oD^2; R^2 = 0,9162$

C

Regressão Linear Simples

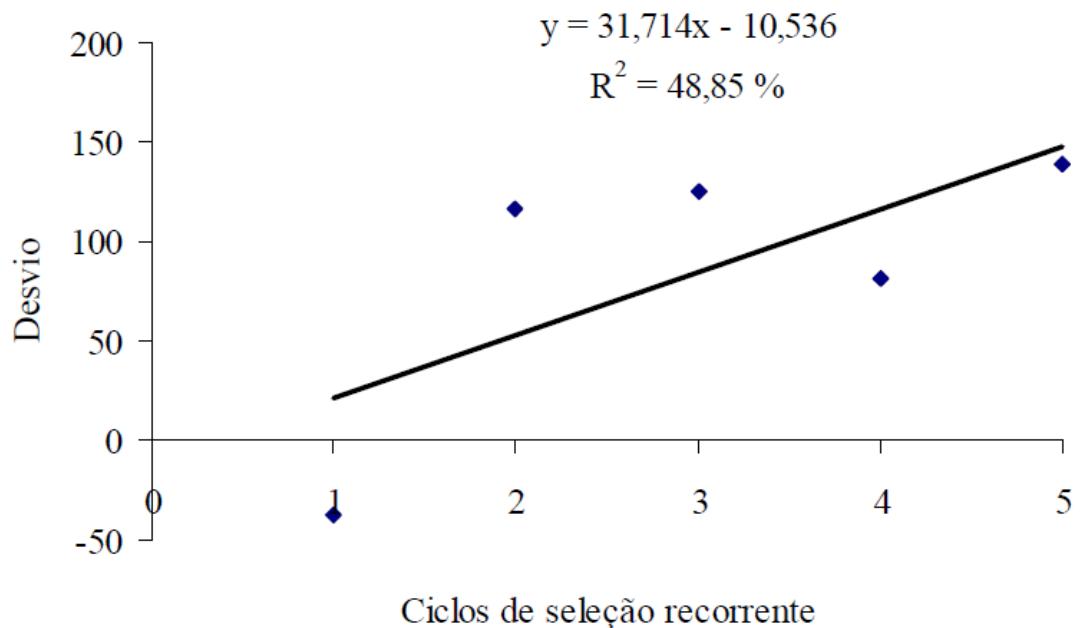


FIGURA 2 Regressão dos desvios das médias das progêniens da geração S_{0:1} em relação às testemunhas, às cultivares Carioca MG e Pérola, para produtividade de grãos (g/2m²) do primeiro ao quinto ciclo de seleção recorrente [$b \neq 0$ ($P \leq 0,01$)].

Regressão Linear Simples

Uma equação de regressão linear simples permite determinar, a partir das estimativas dos parâmetros, como uma variável independente (X) exerce, ou parece exercer, influência sobre outra variável (Y), chamada de variável dependente. Por exemplo, qual a influência do diâmetro à altura do peito (DAP) sobre o volume de árvores de Eucalipto? Esta pergunta poderia ser respondida a partir de uma regressão linear simples entre as variáveis Y (volume das árvores) e X (DAP das árvores). Logicamente, quanto maior o diâmetro, maior o volume, entretanto, é necessário determinar em que proporção isto ocorre e qual o modelo estatístico mais apropriado.

Regressão Linear Simples

O Modelo Estatístico de uma Regressão Linear Simples

$$Y_i = \beta_0 + \beta_1 X_i + e_i \quad e_i \sim NID(0, \sigma^2)$$

Dados n pares de valores de duas variáveis, X_i e Y_i , com $i = 1, 2, \dots, n$, se admitirmos que Y é função linear de X , podemos estabelecer uma regressão linear simples, cujo modelo estatístico é $Y_i = \beta_0 + \beta_1 X_i + e_i$, onde β_0 e β_1 são os parâmetros do modelo, e e_i são os erros aleatórios.

O coeficiente angular da reta (β_1) é também chamado de coeficiente de regressão e o coeficiente linear da reta (β_0) é também conhecido como intercepto sendo o termo constante da equação de regressão.

Regressão Linear Simples

$$Y_i = \beta_0 + \beta_1 X_i + e_i \quad e_i \sim NID(0, \sigma^2)$$

O primeiro passo na análise de regressão linear simples (RLS) é obter as estimativas dos parâmetros β_0 e β_1 . Essas estimativas são obtidas a partir de uma amostra de tamanho n , isto é, a partir de n pares X_i, Y_i .

Regressão Linear Simples

$$Y_i = \beta_0 + \beta_1 X_i + e_i \quad e_i \sim NID(0, \sigma^2)$$

O método usual para a obtenção das estimativas dos parâmetros de um modelo de regressão é o dos Mínimos Quadrados (MMQ). Este método consiste em adotar como estimativas dos parâmetros os valores que minimizam a soma de quadrados dos erros.

Sejam:

$$e_i = Y_i - \beta_0 - \beta_1 X_i$$

$$Z = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 X_i)^2$$

Regressão Linear Simples

A função Z terá mínimo quando suas derivadas parciais em relação a β_0 e a β_1 forem nulas (Observe que Z não tem máximo, por ser uma soma de quadrados). Então,

$$\frac{\partial Z}{\partial \beta_0} = -2 \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 X_i)$$

$$\frac{\partial Z}{\partial \beta_1} = -2 \sum_{i=1}^n X_i(Y_i - \beta_0 - \beta_1 X_i)$$

Regressão Linear Simples

Assim, as estimativas de β_0 e β_1 são dadas por:

$$\begin{cases} \sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0 \\ \sum_{i=1}^n X_i (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0 \end{cases} \quad (2.1)$$

A partir do sistema (2.1), podemos escrever o seguinte sistema de equações normais:

$$\begin{cases} \hat{\beta}_0 n + \hat{\beta}_1 \sum_{i=1}^n X_i = \sum_{i=1}^n Y_i \\ \hat{\beta}_0 \sum_{i=1}^n X_i + \hat{\beta}_1 \sum_{i=1}^n X_i^2 = \sum_{i=1}^n X_i Y_i \end{cases} \quad (2.2)$$

Regressão Linear Simples

Resolvendo este sistema de equações normais obtém-se:

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$
$$\hat{\beta}_1 = \frac{\sum_{i=1}^n X_i Y_i - \frac{\left(\sum_{i=1}^n X_i\right) \left(\sum_{i=1}^n Y_i\right)}{n}}{\sum_{i=1}^n X_i^2 - \frac{\left(\sum_{i=1}^n X_i\right)^2}{n}} = \frac{SPD_{XY}}{SQD_X}$$

A estimativa do coeficiente de regressão $\hat{\beta}_1$ mede o quanto muda na variável dependente \hat{Y} por mudança unitária na variável independente X .

Análise de Variância Regressão Linear Simples

O esquema da análise de variância da regressão linear simples considerando o modelo $Y_i = \beta_0 + \beta_1 X_i + e_i$, é apresentado na Tabela a seguir:

Fontes de Variação	GL	SQ	QM	F
Regressão	1	$SQ_{Regressão}$	$QM_{Regressão} = V_1$	V_1/V_2
Resíduo	$n - 2$	$SQ_{Resíduo}$	$QM_{Resíduo} = V_2$	—
Total	$n - 1$	SQ_{Total}	—	—

Em que, GL é o número de graus de liberdade e $QM_{Regressão} = \frac{SQ_{Regressão}}{1} = SQ_{Regressão}$.
Logo, no caso de uma RLS, $QM_{Regressão} = SQ_{Regressão}$. E, ainda, $QM_{Resíduo} = \frac{SQ_{Resíduo}}{n-2}$

A estatística $F = \frac{V_1}{V_2}$, na Tabela anterior, testa a hipótese $H_0 : \beta_1 = 0$ vs $H_a : \beta_1 \neq 0$.

Sob H_0 verdadeira, esta estatística F tem distribuição F central com 1 e $n - 2$ graus de liberdade.

Regra de Decisão: Se F calculado for $\geq F$ tabelado, rejeita-se H_0 ao nível de significância α . Neste caso, diz-se que o resultado é significativo ($p < \alpha$), e a regressão linear existe. Caso contrário, não se rejeita H_0 , e então, o resultado é não significativo ($p > \alpha$).

SOMA DE QUADRADOS

Regressão Linear Simples

$$SQTotal = \sum_{i=1}^n Y_i^2 - \frac{\left(\sum_{i=1}^n Y_i \right)^2}{n}$$

$$SQRegressão = \frac{\left[\sum_{i=1}^n X_i Y_i - \frac{\left(\sum_{i=1}^n X_i \right) \left(\sum_{i=1}^n Y_i \right)}{n} \right]^2}{\sum_{i=1}^n X_i^2 - \frac{\left(\sum_{i=1}^n X_i \right)^2}{n}} = \frac{(SPD_{XY})^2}{SQD_X} = \hat{\beta}_1 SPD_{XY}$$

Regressão Linear Simples

Coeficiente de Determinação

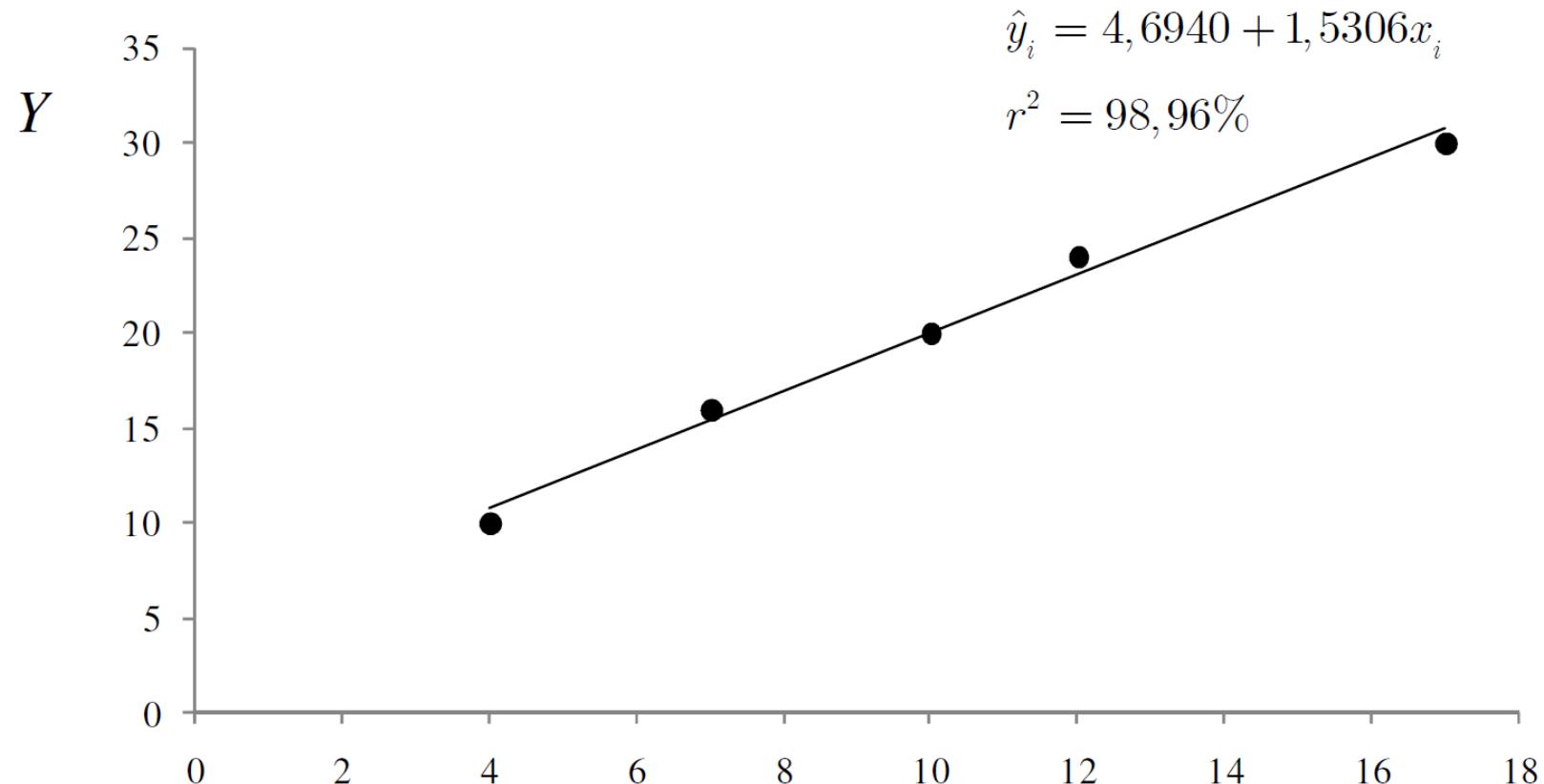
O coeficiente de determinação de uma regressão linear simples, denotado por r^2 e expresso em porcentagem, é dado por:

$$r^2 = \frac{SQ\text{Regressão}}{SQ\text{Total}} \quad 100, \quad 0 \leq r^2 \leq 100\%$$

O r^2 indica a proporção da variação de Y que é explicada pela regressão, ou quanto da $SQ\text{Total}$ está sendo explicada pela regressão, ou quanto da variação na variável dependente Y está sendo explicada pela variável independente X . Quanto maior for o r^2 , melhor. Além do coeficiente de determinação, outros critérios devem ser adotados na escolha de modelos.

Regressão Linear Simples

Coeficiente de Determinação



Regressão Linear Simples

Coeficiente de Determinação

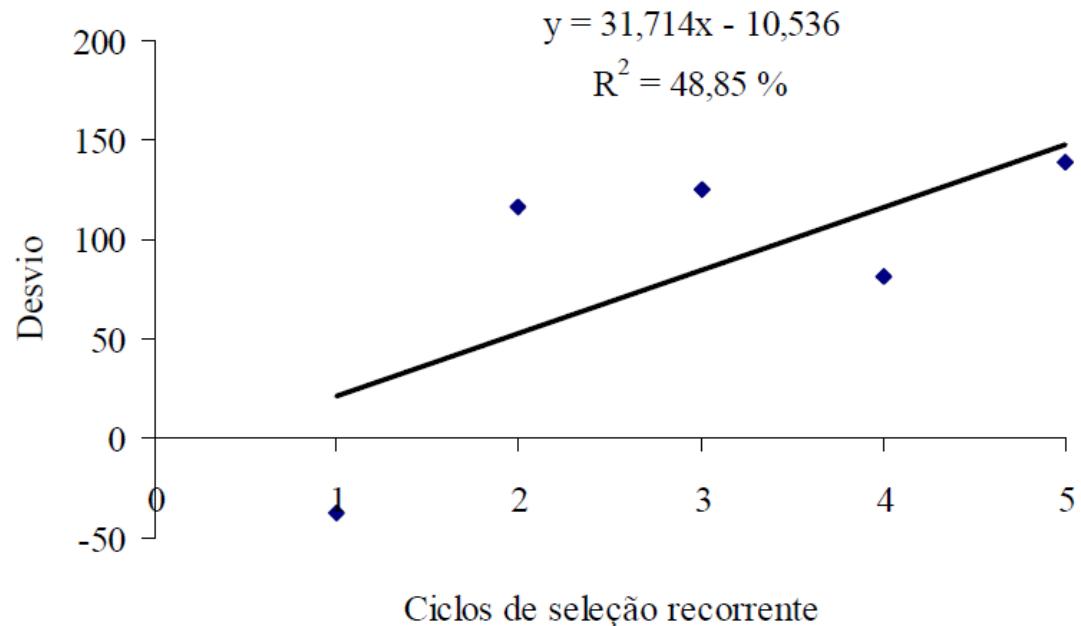


FIGURA 2 Regressão dos desvios das médias das progênies da geração S_{0:1} em relação às testemunhas, às cultivares Carioca MG e Pérola, para produtividade de grãos ($\text{g}/2\text{m}^2$) do primeiro ao quinto ciclo de seleção recorrente [$b \neq 0$ ($P \leq 0,01$)].

Regressão Linear Simples

Exemplo

X	4	7	10	12	17
Y	10	16	20	24	30

$$Y_i = \beta_0 + \beta_1 X_i + e_i$$

$$n = 5, \sum_{i=1}^n X_i = 50, \sum_{i=1}^n X_i^2 = 598, \sum_{i=1}^n Y_i = 100, \sum_{i=1}^n Y_i^2 = 2232, \sum_{i=1}^n X_i Y_i = 1150$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n X_i Y_i - \frac{\left(\sum_{i=1}^n X_i\right) \left(\sum_{i=1}^n Y_i\right)}{n}}{\sum_{i=1}^n X_i^2 - \frac{\left(\sum_{i=1}^n X_i\right)^2}{n}}$$

$$\hat{\beta}_1 = \frac{SPD_{XY}}{SQD_{XY}} = \frac{150}{98} = 1,5306$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} = \frac{100}{5} - 1,5306 \frac{50}{5} = 4,6940$$

Calcule o coeficiente de determinação e interprete.

$$SQTotal = 2232 - \frac{(100)^2}{5} = 232$$

$$SQRegressão = \frac{(150)^2}{98} = 229,5918$$

$$r^2 = \frac{229,5918}{232} 100 = 98,96\%$$

Interpretação: 98,96% da variação observada em Y está sendo “explicada” pela regressão linear ajustada.

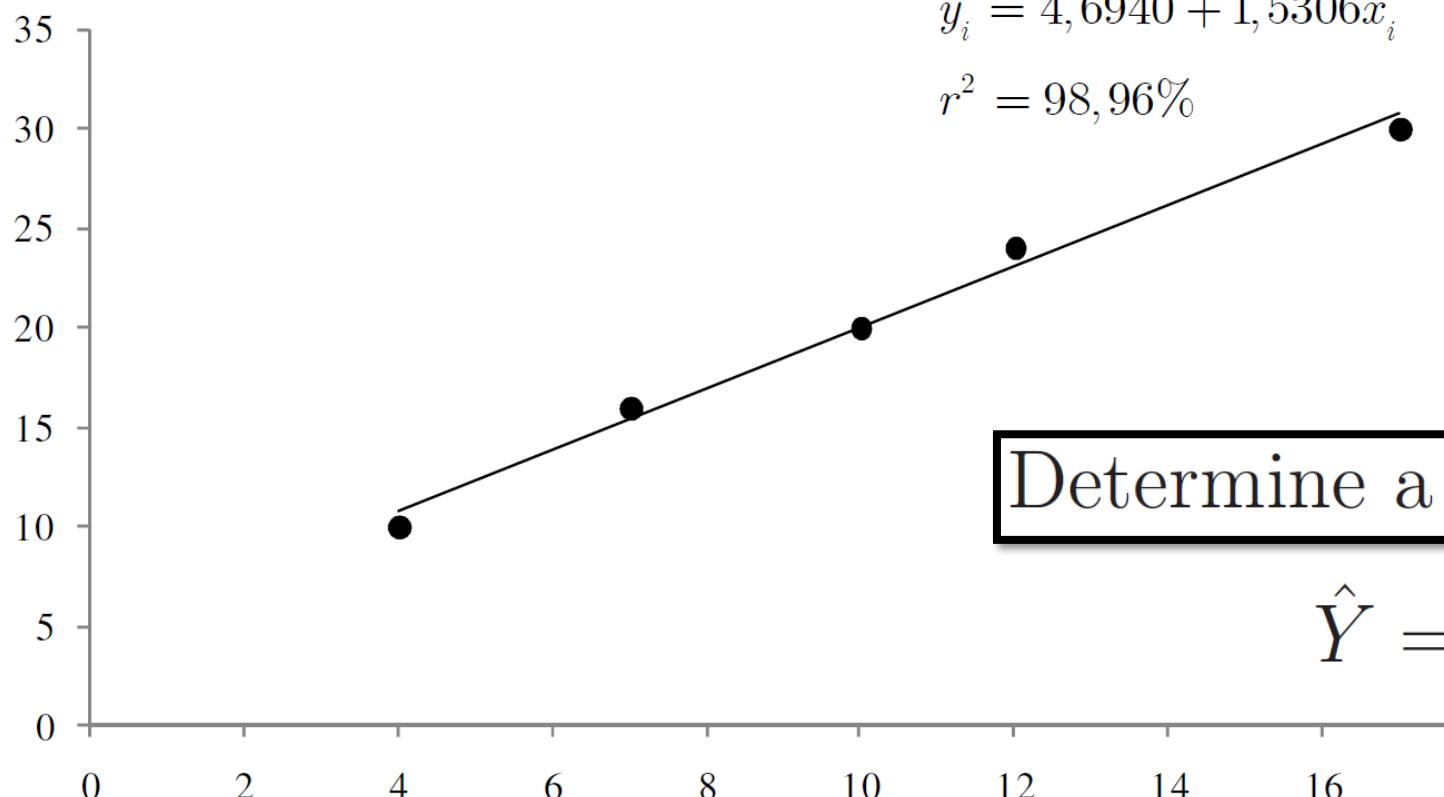
Faça a análise de variância da regressão

	FV	GL	SQ	QM	F
Regressão	1	229,5918	229,5918	286,02**	—
Resíduo	3	2,4082	0,8027	—	—
Total	4	232,0000	—	—	—

** Significativo ao nível de 1% de probabilidade $F_{1\%}(1; 3) = 34,12$.

Com estes resultados pode-se concluir que a hipótese $H_0 : \beta_1 = 0$ foi rejeitada

Logo, a regressão linear existe.



Determine a estimativa de Y para $X = 9$.

$$\hat{Y} = 4,6940 + 1,5306(9) \cong 18,47.$$

$\hat{\beta}_1 = 1,5306$. Assim, para um aumento de uma unidade em X tem-se um acréscimo de 1,5306 em Y , ou melhor, para um aumento de uma unidade em X , estima-se um aumento médio de 1,5306 na variável independente Y .

Interpretação: 98,96% da variação observada em Y está sendo “explicada” pela regressão linear ajustada.

Dúvidas