Statistics for Business and Economics

Newbold

Capítulo 2 – Aula 03

Descrevendo dados: Numericamente

Objetivos

Calcular e interpretar a média, mediana e moda para um conjunto de dados

Encontrar o intervalo, variância, desvio padrão e coeficiente de variação e compreender seu significado

Aplicar a regra empírica para descrever a variação dos valores populacionais em torno da média

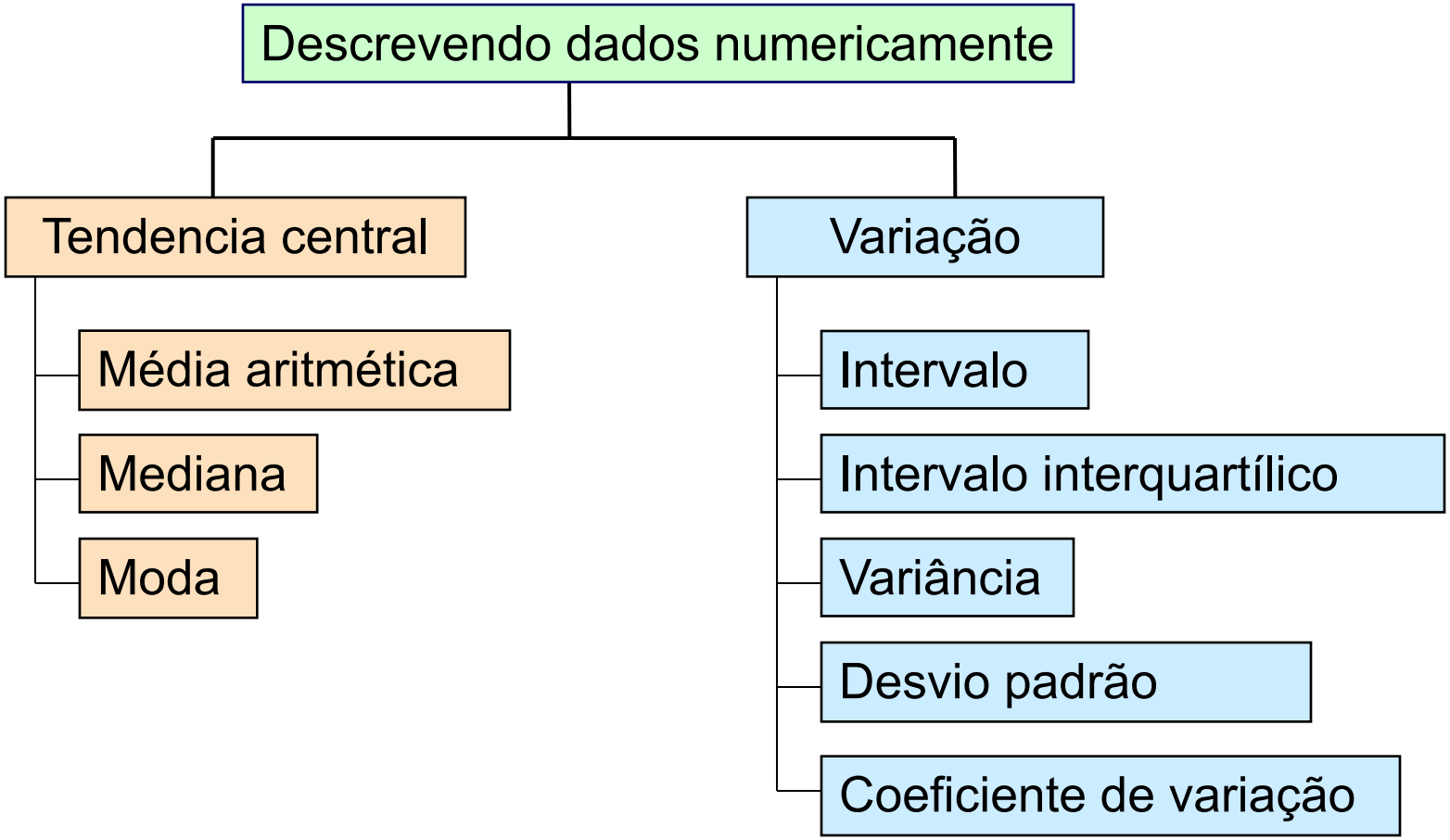
Explicar a média ponderada e saber quando usá-la

Explicar como uma linha de regressão de mínimos quadrados estima uma relação linear entre duas variáveis

Tópicos

- Medidas de tendência central, variação e forma
- Média, mediana, moda, média geométrica
- Quartis
- Amplitude, amplitude interquartil, variância e desvio padrão, coeficiente de variação
- Distribuições simétricas e assimétricas
- Medidas de resumo da população
- Média, variância e desvio padrão
- A regra empírica e a regra de Bienaymé-Chebyshev
- Resumo de cinco números e gráficos de box-and-whisker plots
- Covariância e coeficiente de correlação
- Armadilhas em medidas descritivas numéricas e considerações éticas

Descrevendo dados numericamente



2.1

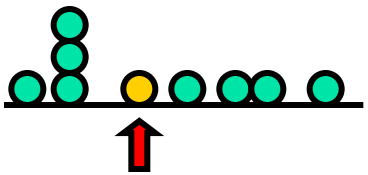
Medidas de Tendencia central

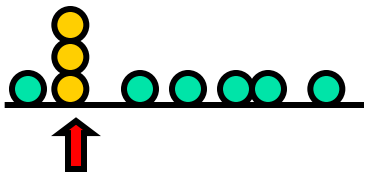
Visão geral

Tendencia central

- **Média**
- **Mediana**
- **Moda**

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$





Média aritmética

Ponto médio dos valores classificados

Valor observado com mais frequência

Média aritmética

- A média aritmética (mean) é a medida mais comum de tendência central

Para uma população de N valores:

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

Valores da população

Tamanho da população

- For a sample of size n:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

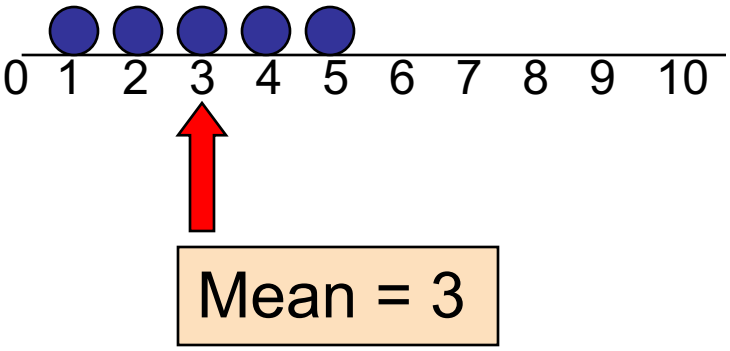
Valores observados

Tamanho da amostra

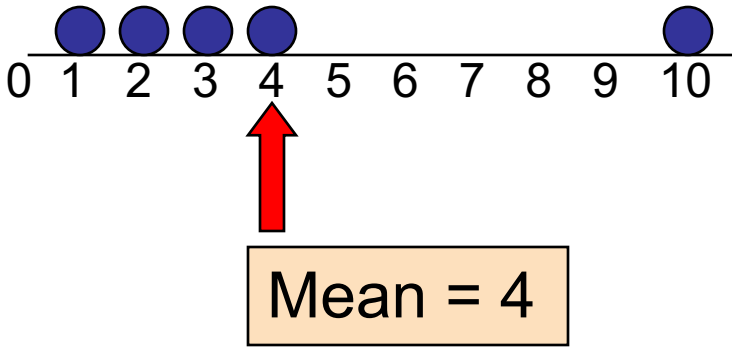
Média aritmética

(cont)

- A medida de tendência central mais comum
- Média = soma dos valores dividida pelo número de valores
- Afetado por valores extremos (outliers)



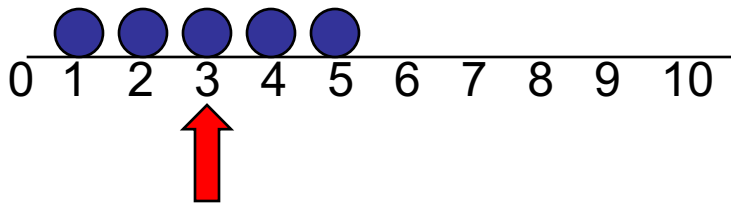
$$\frac{1+2+3+4+5}{5} = \frac{15}{5} = 3$$



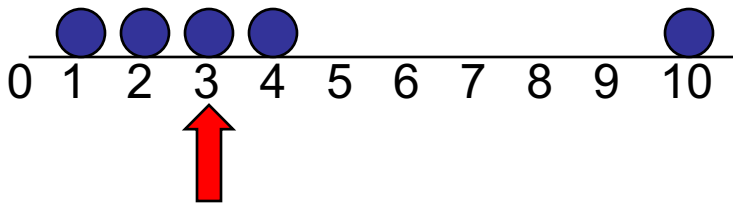
$$\frac{1+2+3+4+10}{5} = \frac{20}{5} = 4$$

Mediana

- Em uma lista ordenada, a mediana é o número “meio” (50% acima, 50% abaixo)



Mediana = 3



Mediana = 3

- Não é afetada por valores extremos

Encontrando a Mediana

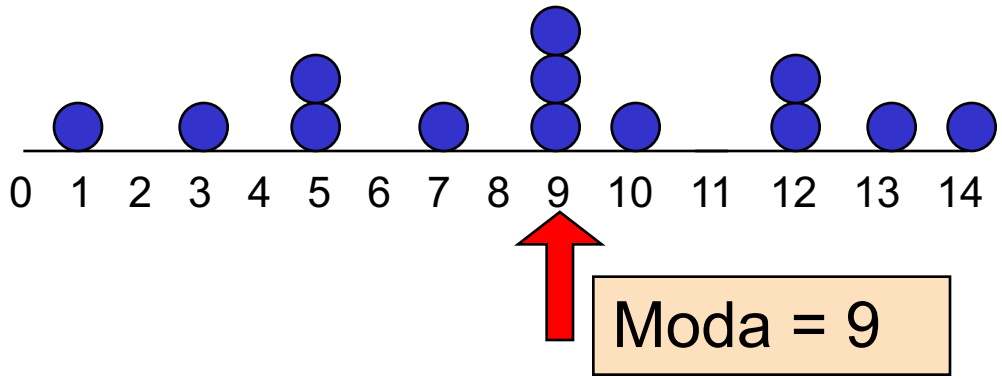
- A localização da Mediana:

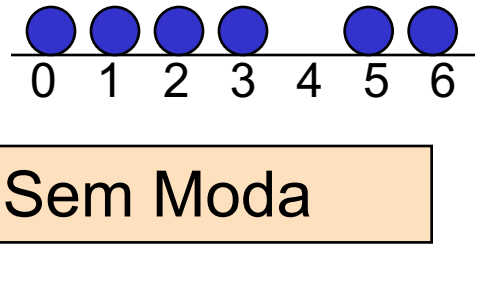
- $Posição\ da\ Mediana = \frac{n+1}{2}$ posição em dados ordenados
  - Se o número de valores for ímpar, a mediana é o número do meio
  - Se o número de valores for par, a mediana é a média dos dois números do meio

- Note que $\frac{n+1}{2}$ não é o *valor* da Mediana, é a *posição* da Mediana nos dados ordenados

Moda

- Uma medida de tendência central
- Valor que ocorre com mais frequência
- Não é afetado por valores extremos
- Usado para dados numéricos ou categóricos
- pode não haver moda
- Pode haver vários modas



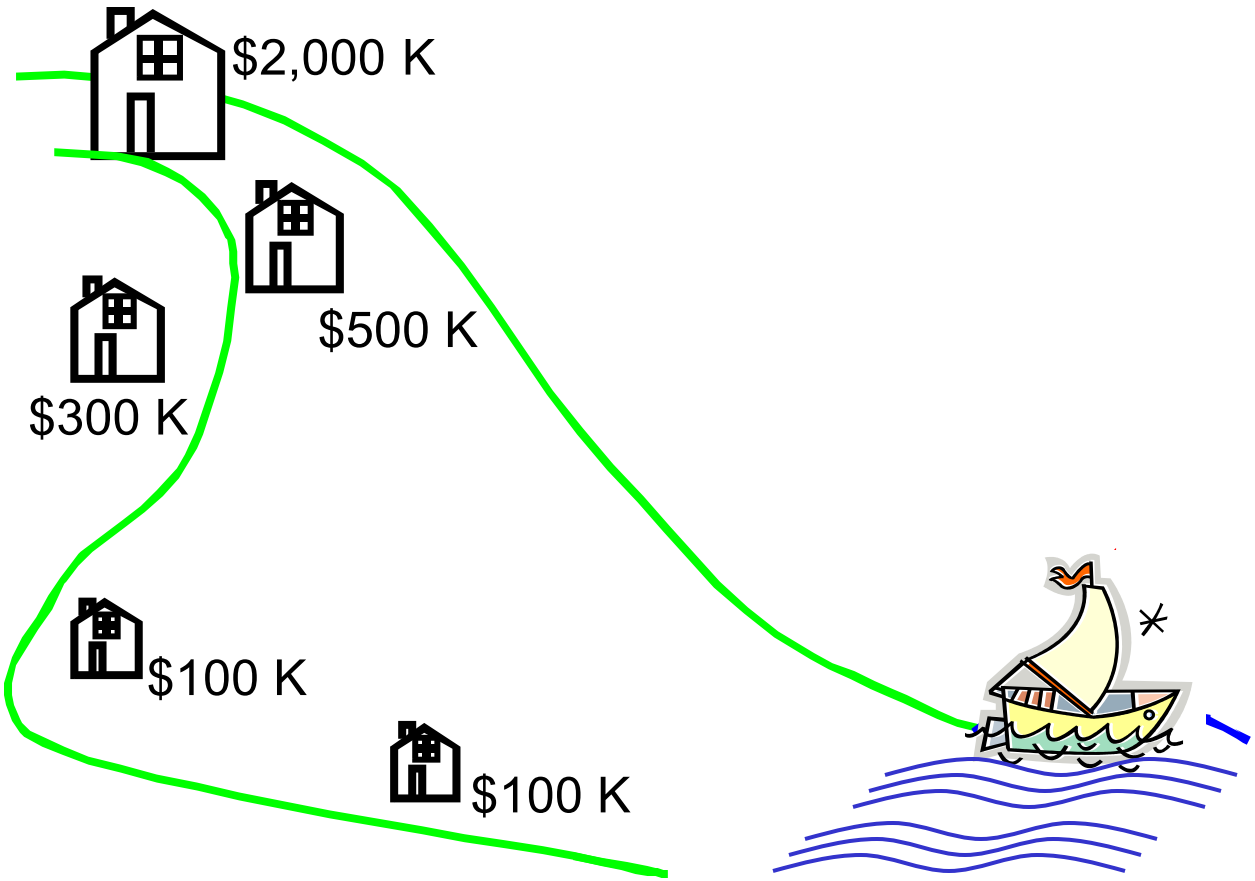


Exemplo de avaliação

- Cinco casas da colina para a praia

Preço das casas:

$2,000,000
500,000
300,000
100,000
100,000



Exemplo de avaliação : Resumo das estatísticas

Preços das casas:

$2,000,000	
500,000	
300,000	
100,000	
100,000	
Total	3,000,000

- **Média:** ($3,000,000/5)
  = **$600,000**
- **Mediana:** valor do meio dos dados ordenados
  = **$300,000**
- **Moda:** valor mais frequente
  = **$100,000**

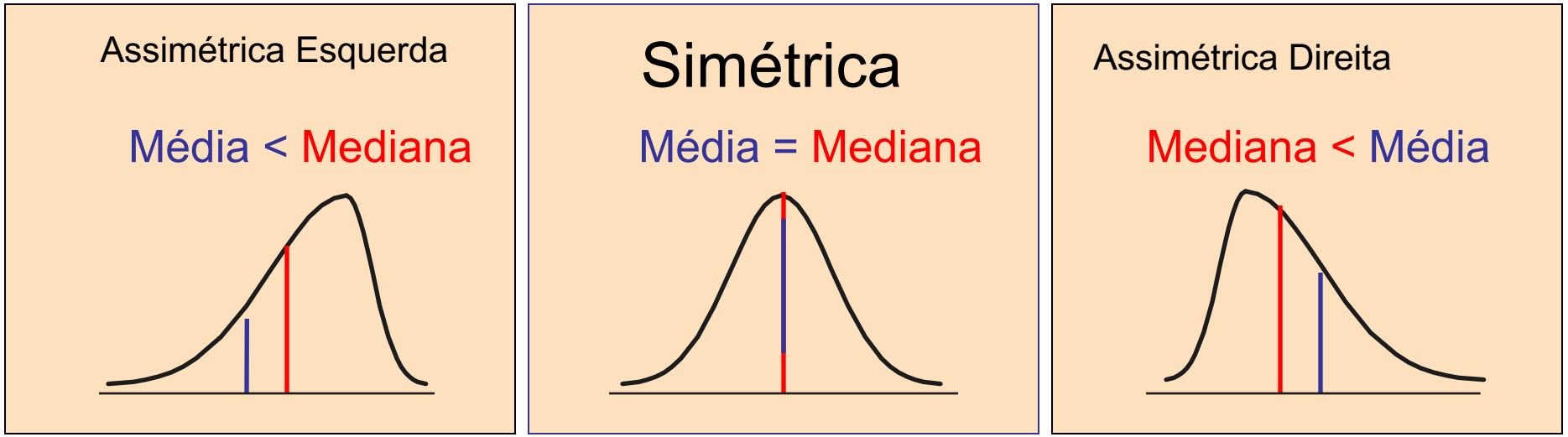
Qual a melhor medida de posição central?

- **A média é geralmente usada, a menos que existam valores extremos (outliers). . .**
- **Então a mediana é freqüentemente usada, já que a mediana não é sensível a valores extremos.**

Exemplo: os preços medianos das casas podem ser relatados para uma região – menos sensíveis a valores discrepantes

Forma de uma distribuição

- Descreve como os dados são distribuídos
- medidas de forma
  - Simétrica ou enviesada



Média Geométrica

- Média Geométrica
  - Usado para medir a taxa de mudança de uma variável ao longo do tempo

$$\bar{x}_g = \sqrt[n]{(x_1 \times x_2 \times \cdots \times x_n)} = (x_1 \times x_2 \times \cdots \times x_n)^{1/n}$$

- Taxa de retorno da média geométrica
  - Mede o status de um investimento ao longo do tempo

$$\bar{r}_g = (x_1 \times x_2 \times \ldots \times x_n)^{1/n} - 1$$

  - Snedo x_i a taxa de retorno do período de tempo i

Exemplo

Um investimento de $100,000 aumenta para $150,000 ao final de um ano e aumenta para $180,000 ao final de dois anos:

$$X_1 = \$100{,}000 \quad X_2 = \$150{,}000 \quad X_3 = \$180{,}000$$

50% aumento

20% aumento

Qual é o retorno percentual médio ao longo do tempo?

Exemplo

(cont)

Use o retorno de cada ano para computer a as medias aritméticas e geométricas:

Taxa média aritmética de retorno:

$$\bar{X} = \frac{(50\%) + (20\%)}{2} = 35\%$$

Resultado enganoso

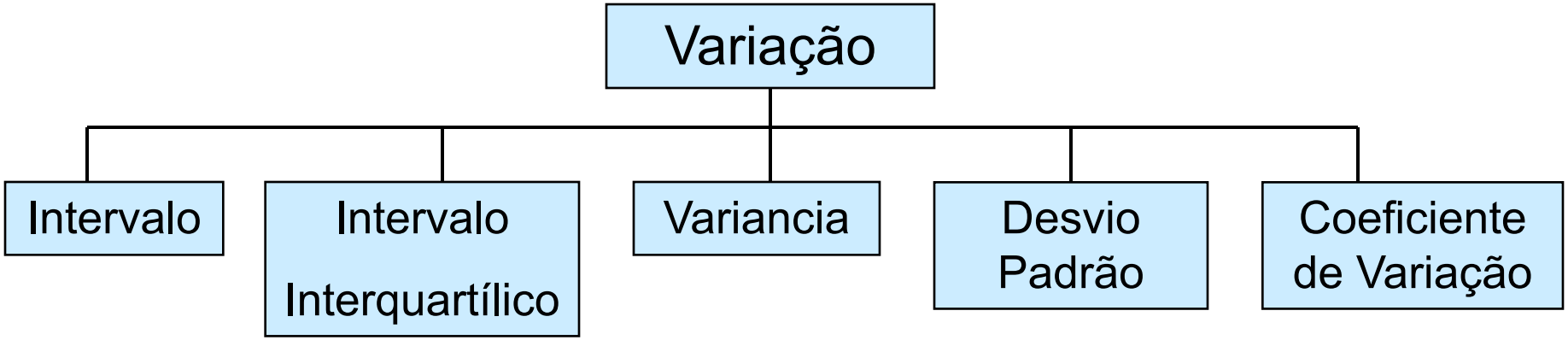
Taxa de retorno média geométrica:

$$\bar{r}_g = (x_1 \times x_2)^{1/n} - 1$$
$$= [(50) \times (20)]^{1/2} - 1$$
$$= (1000)^{1/2} - 1 = 31.623 - 1 = 30.623\%$$

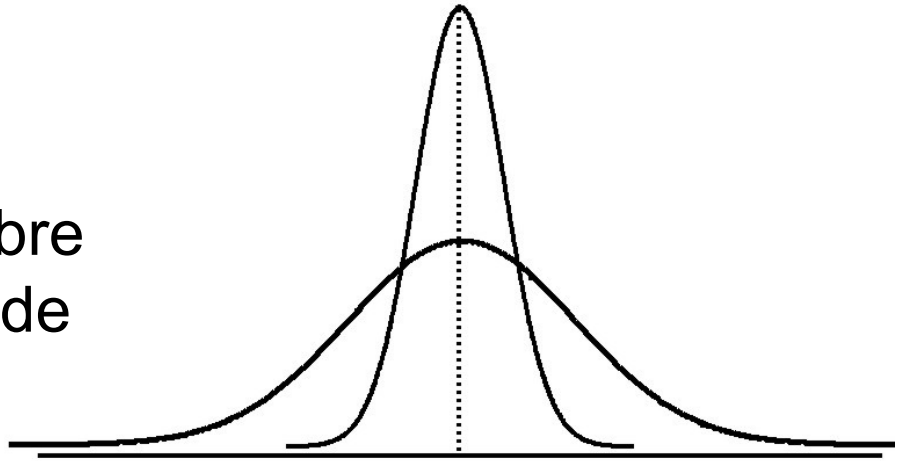
Resultado mais preciso

2.2

Medidas de Variabilidade



- As medidas de variação fornecem informações sobre a dispersão ou variabilidade dos valores dos dados.



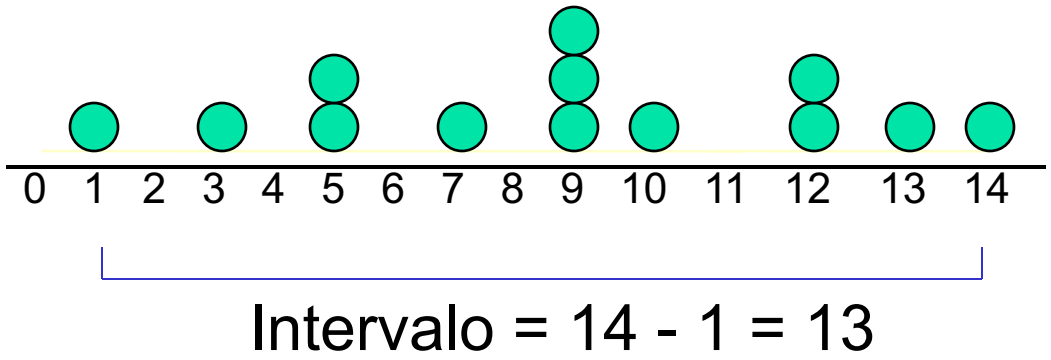
Mesma centralidade, diferentes variações

Intervalo

- Medida mais simples de variação
- Diferença entre a maior e a menor observação:

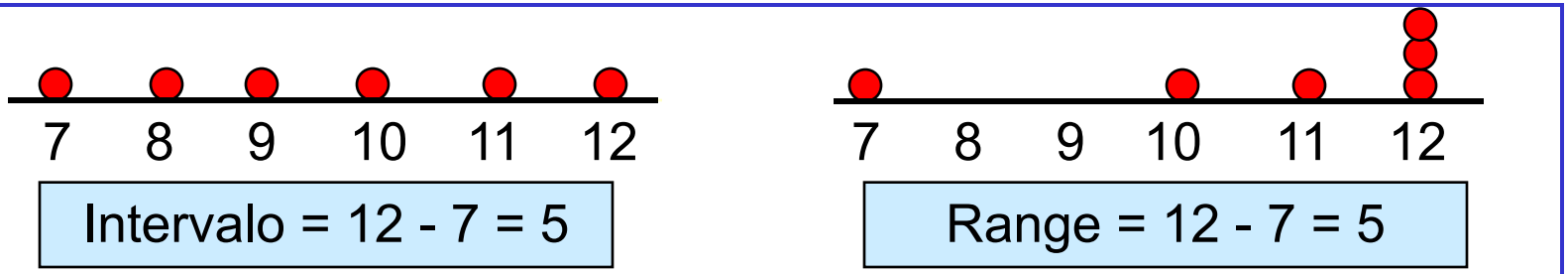
$$\text{Intervalo} = X_{\text{maior}} - X_{\text{menor}}$$

Exemplo:



Desvantagens do Intervalo

- Ignora a forma como os dados são distribuídos



- Sensível a valores discrepantes (outliers)

1,1,1,1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,2,3,3,3,3,4,5

Intervalo = 5 - 1 = 4

1,1,1,1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,2,3,3,3,3,4,120

Intervalo = 120 - 1 = 119

Intervalo Interquartílico

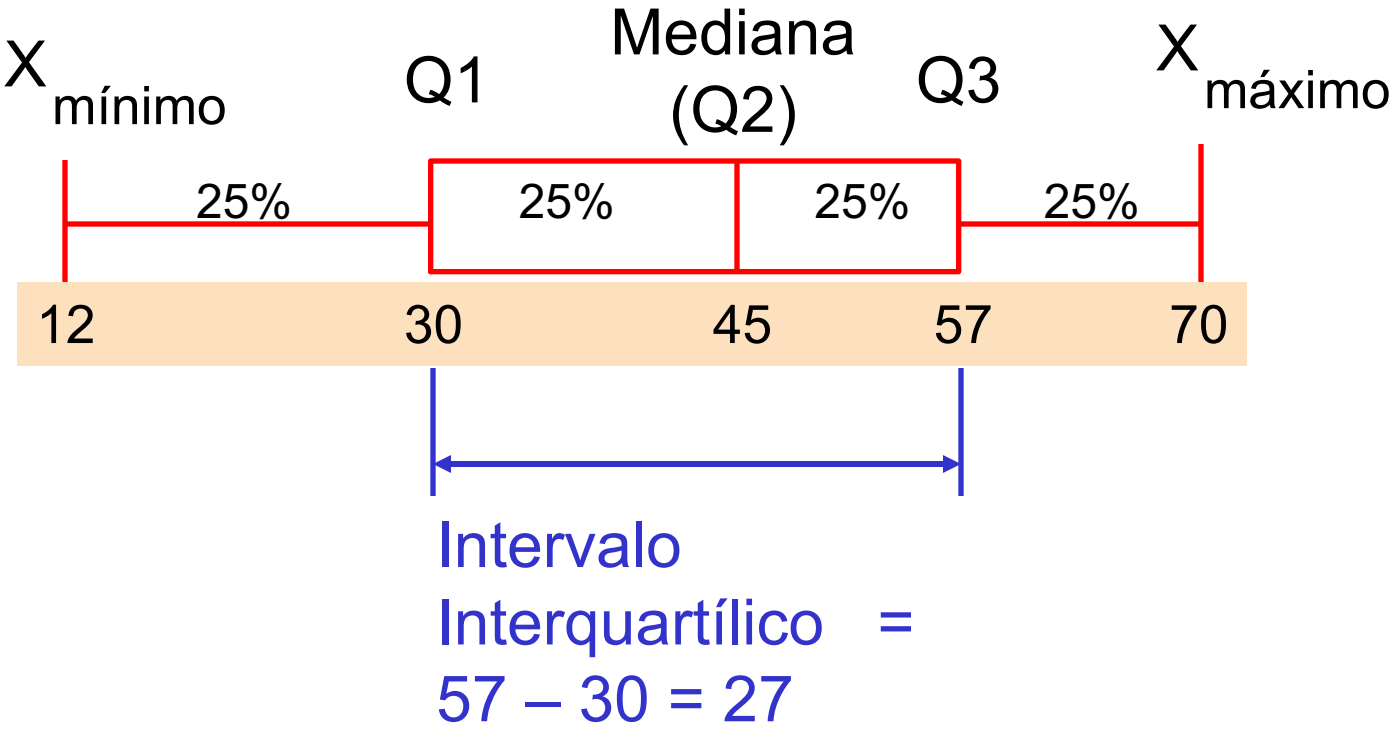
- Pode eliminar alguns problemas discrepantes usando o intervalo interquartílico

- Elimine observações de alto e baixo valor e calcule o alcance dos 50% intermediários dos dados

- Intervalo interquartílico = 3º quartil – 1º quartil
- IQR = Q3 - Q1

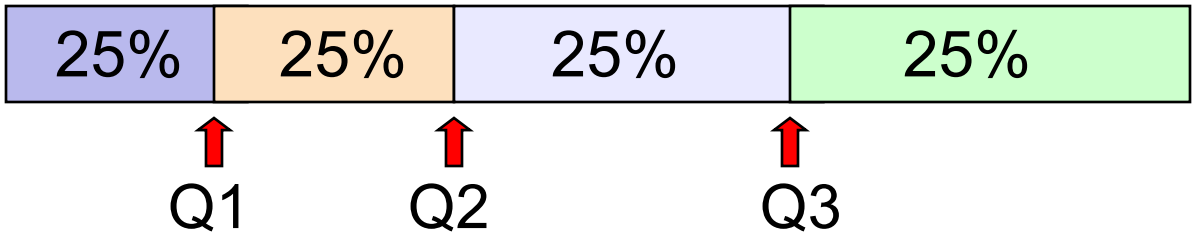
Intervalo Interquartílico

Exemplo:



Quartis

- Quartis dividem os dados classificados em 4 segmentos com um número igual de valores por segmento



- O primeiro quartil, Q1, é o valor para o qual 25% das observações são menores e 75% são maiores
- Q2 é o mesmo que a mediana (50% são menores, 50% são maiores)
- Apenas 25% das observações são maiores que o terceiro quartil

Fórmulas de quartil

Encontre um quartil determinando o valor na posição apropriada nos dados classificados, onde

Posição do primeiro quartil: Q1 = 0,25(n+1)

Posição do segundo quartil: Q2 = 0,50(n+1)
(a posição mediana)

Posição do terceiro quartil: Q3 = 0,75(n+1)

Sendo n é o número de valores observados

Quartis

- Exemplo: encontrar o primeiro quartil

Exemplo dados ordenados: 11 12 13 16 16 17 18 21 22

(n = 9)

Q_1 = é o 0.25(9+1) = 2.5 posição dos dados ordenados

portanto, use o valor intermediário entre o 2º e o 3º valores,

Então $Q_1 = 12.5$

Variância populacional

- Média dos desvios quadrados dos valores da média
- variação da população

$$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$$

sendo μ = media populacional

N = Tamanho da população

x_i = i^{th} iésimo valor da população x

Variância Amostral

- Média (aproximada) dos desvios quadrados dos valores da media
  - Variância da amostra:

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

sendo $\bar{X}$ = média aritmética

n = Tamanho da amostra

X_i = i^{th} iésimo valor da população X

Desvio padrão da população

- Medida de variação mais comumente usada
- Mostra a variação sobre a média
- Tem as mesmas unidades que os dados originais

Desvio padrão da população:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}}$$

Desvio padrão da amostra

- Medida de variação mais comumente usada
- Mostra a variação dos dados em relação à média
- Tem as mesmas unidades que os dados originais

- Desvio padrão da amostra

$$S = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$$

Exemplo de Cálculo: Desvio Padrão da Amostra

Amostra de dados (x_i) :

10	12	14	15	17	18	18	24

n = 8 Média = $\bar{x}$ = 16

$$s = \sqrt{\frac{(10-\bar{X})^2 + (12-\bar{x})^2 + (14-\bar{x})^2 + \cdots + (24-\bar{x})^2}{n-1}}$$

$$= \sqrt{\frac{(10-16)^2 + (12-16)^2 + (14-16)^2 + \cdots + (24-16)^2}{8-1}}$$

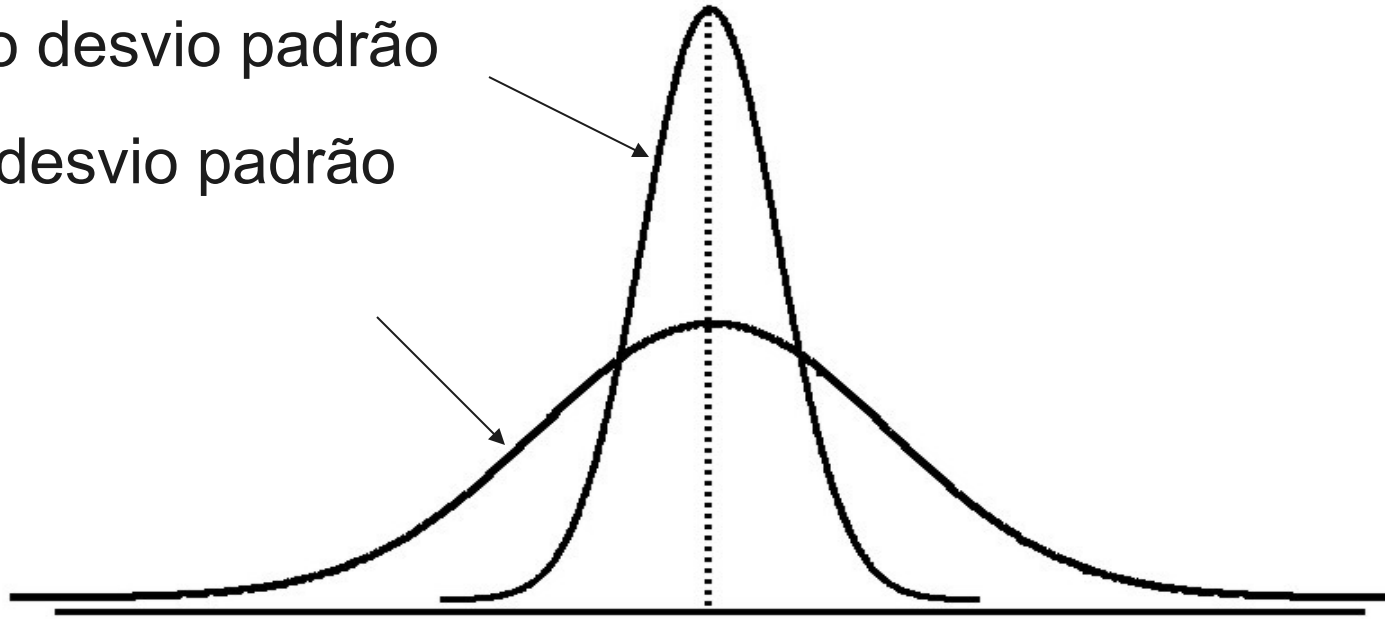
$$= \sqrt{\frac{126}{7}} = 4.2426$$

Uma medida da dispersão "média" em torno da média

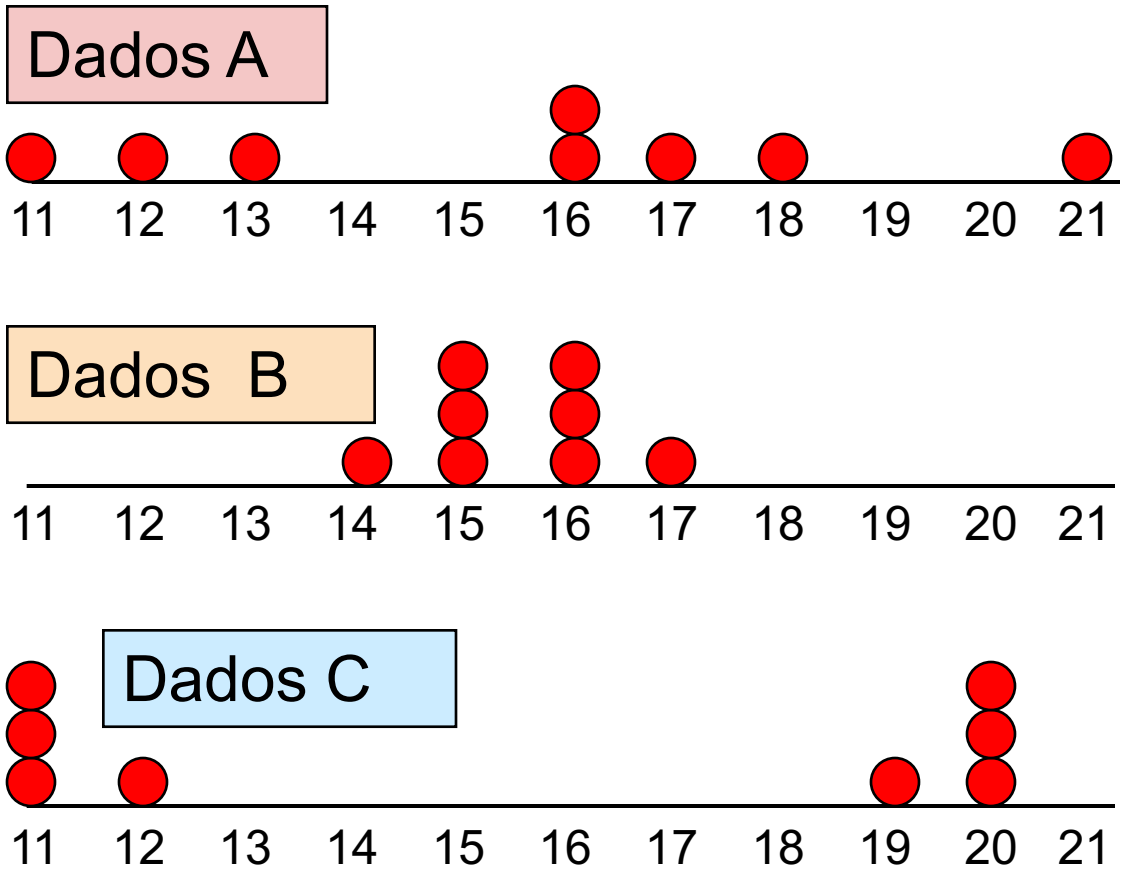
Medindo a variação

Pequeno desvio padrão

Grande desvio padrão



Comparando Desvio-Pradrões



Média = 15.5
s = 3.338

Média = 15.5
s = 0.926

Média = 15.5
s = 4.570

Vantagens da variância e do desvio padrão

- Cada valor no conjunto de dados é usado no cálculo
- Valores distantes da média recebem peso extra (porque os desvios da média são elevados ao quadrado)

Coeficiente de variação

- Mede a variação relativa
- Sempre em porcentagem (%)
- Mostra a variação em relação à média
- Pode ser usado para comparar dois ou mais conjuntos de dados medidos em unidades diferentes

$$CV = \left(\frac{s}{\bar{x}} \right) \cdot 100\%$$

Comparação do Coeficiente de Variação

- Ação A:
  - Preço médio no ano passado = $50
  - Desvio padrão = $5

$$CV_A = \left(\frac{s}{\bar{x}} \right) \cdot 100\% = \frac{\$5}{\$50} \cdot 100\% = 10\%$$

- Ação B:
  - Preço médio no ano passado = $100
  - Desvio padrão = $5

$$CV_B = \left(\frac{s}{\bar{x}} \right) \cdot 100\% = \frac{\$5}{\$100} \cdot 100\% = 5\%$$

Ambas as ações têm o mesmo desvio padrão, mas a ação B é menos variável em relação ao seu preço

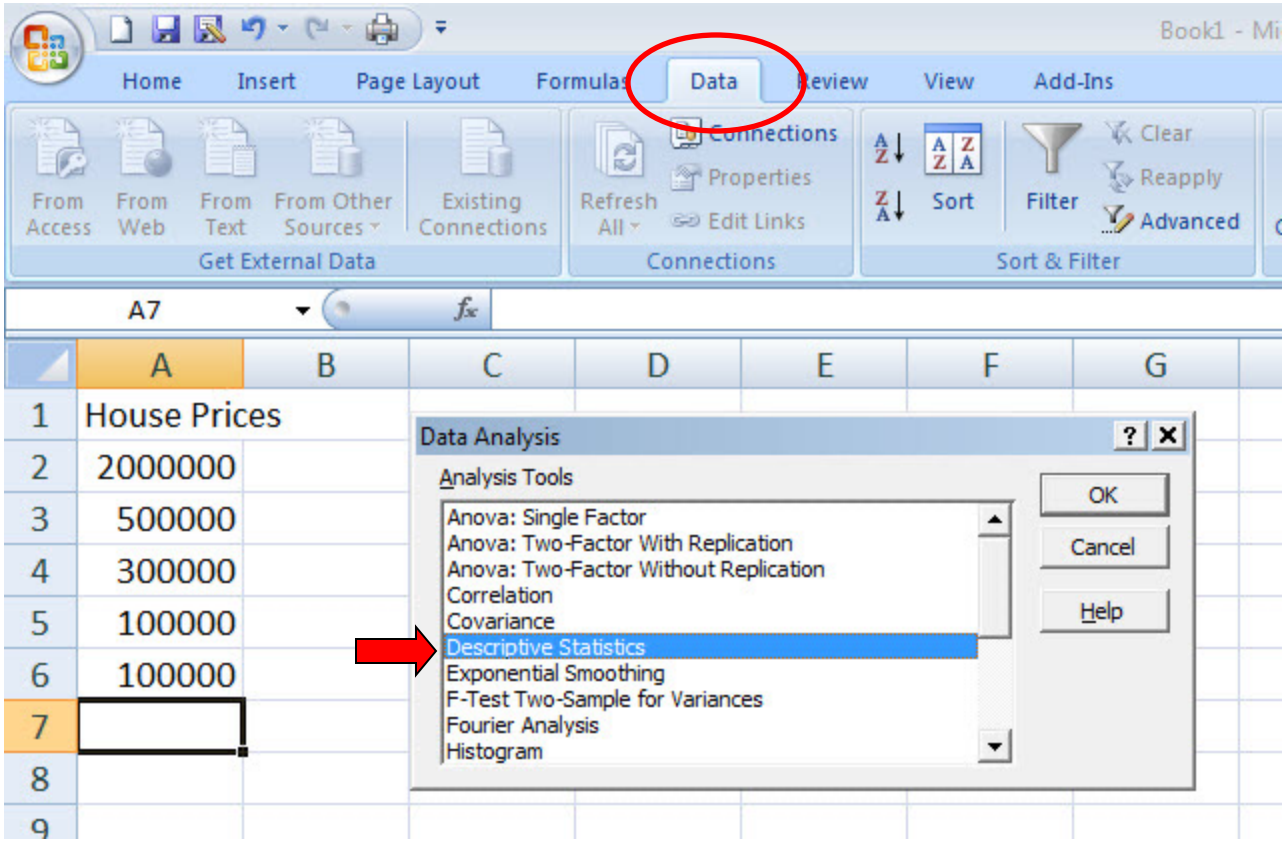
Usando Microsoft Excel

- Descriptive Statistics can be obtained from Microsoft® Excel
  - Select:

    data / data analysis / descriptive statistics
  - Enter details in dialog box

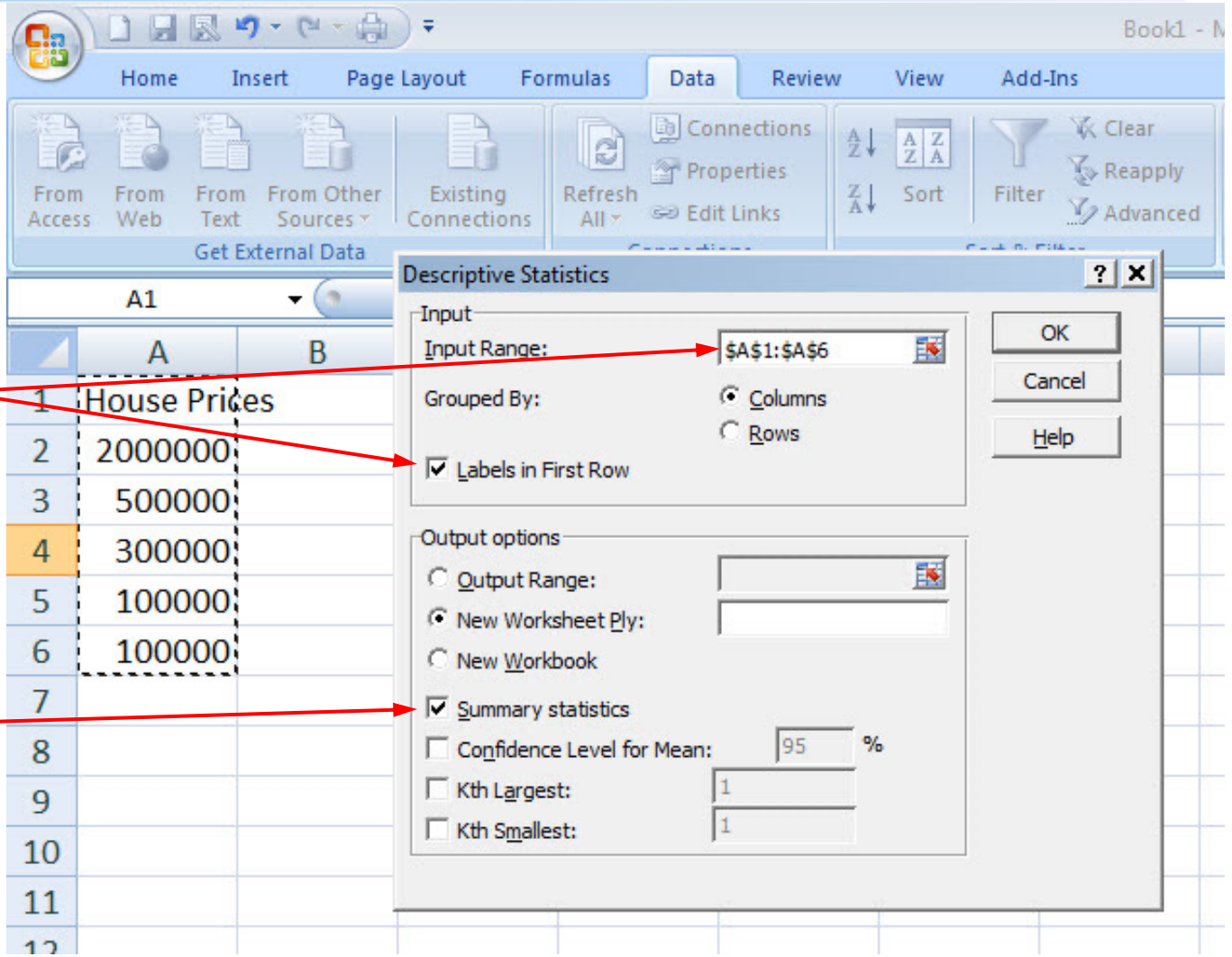
Using Excel

- Select data / data analysis / descriptive statistics



Usando Excel

- Enter input range details
- Check box for summary statistics
- Click OK



Excel output

Microsoft Excel descriptive statistics output, using the house price data:

House Prices:

$2,000,000
500,000
300,000
100,000
100,000

	A	B
1	*House Prices*	
2		
3	Mean	600000
4	Standard Error	357770.8764
5	Median	300000
6	Mode	100000
7	Standard Deviation	800000
8	Sample Variance	6.4E+11
9	Kurtosis	4.130126953
10	Skewness	2.006835938
11	Range	1900000
12	Minimum	100000
13	Maximum	2000000
14	Sum	3000000
15	Count	5

Teorema de Chebyshev

- Para qualquer população com média μ e desvio padrão σ , e k > 1 , a porcentagem de observações que caem dentro do intervalo

$$[\mu + k\sigma]$$

É *pelo menos*

$$100[1-(1/k^2)]\%$$

Teorema de Chebyshev

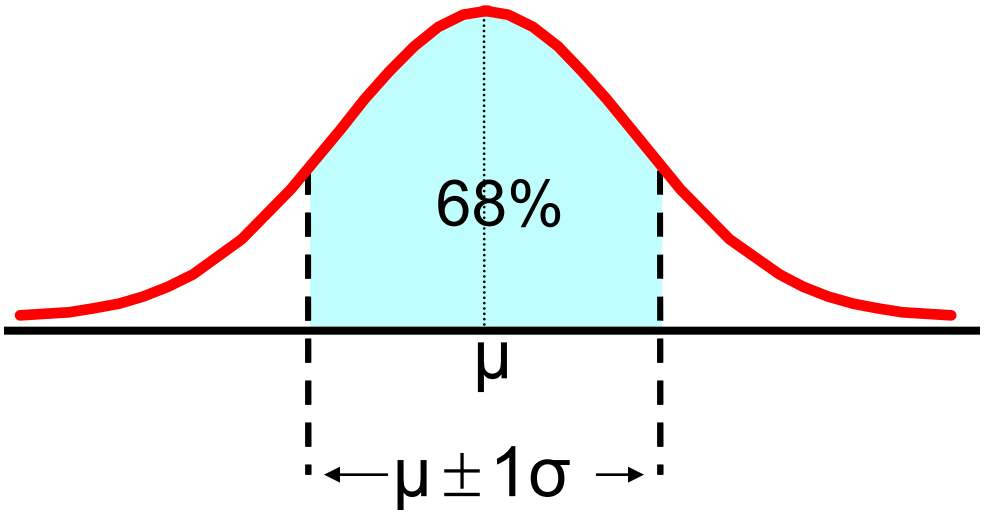
(cont.)

- Independentemente de como os dados são distribuídos, pelo menos $(1 - 1/k^2)$ dos valores cairão dentro de k desvios padrão da média (for $k > 1$)
  - Exemplos:

Pelo menos		dentro
$(1 - 1/1.5^2) = 55.6\%$	..... $k = 1.5$	$(\mu \pm 1.5\sigma)$
$(1 - 1/2^2) = 75\%$	..... $k = 2$	$(\mu \pm 2\sigma)$
$(1 - 1/3^2) = 89\%$	..... $k = 3$	$(\mu \pm 3\sigma)$

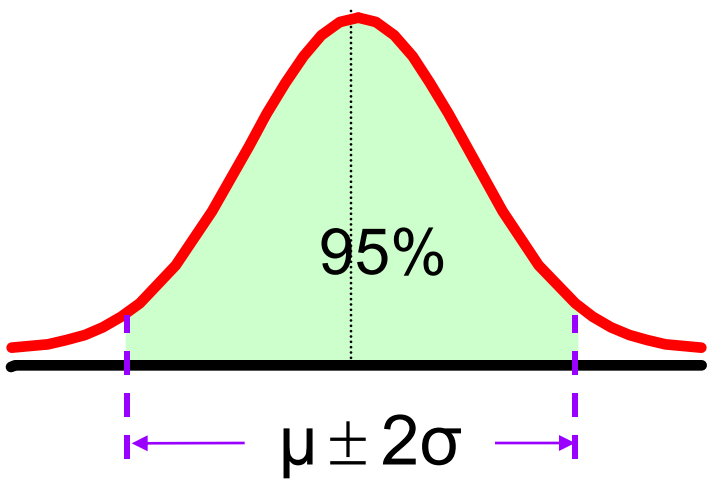
A regra empírica

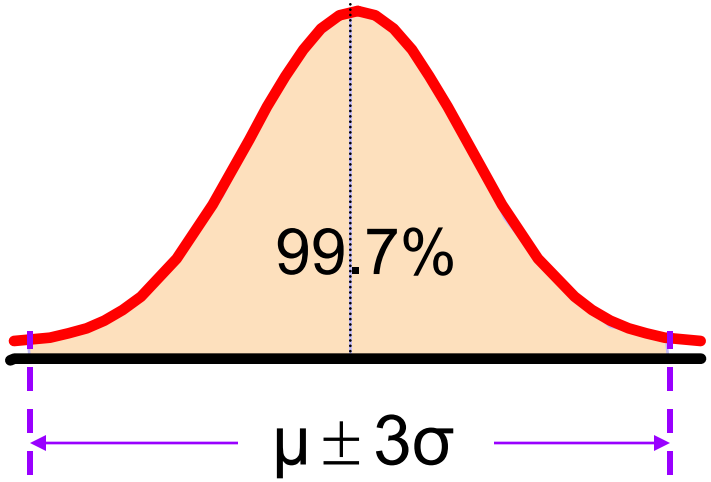
- Se a distribuição de dados for em forma de sino, então o intervalo :
- $\mu \pm 1\sigma$ contém cerca de 68% dos valores na população ou na amostra



A regra empírica

- $\mu \pm 2\sigma$ contém cerca de 95% dos valores na população ou na amostra
- $\mu \pm 3\sigma$ contém quase toda (cerca de 99.7%) dos valores na população ou na amostra





2.3

Weighted Mean

- A média ponderada de um conjunto de dados é

$$\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{n} = \frac{w_1 x_1 + w_2 x_2 + \cdots + w_n x_n}{n}$$

  - Sendo w_i o peso da i^{th} observação e

$$n = \sum w_i$$

- Use quando os dados já estiverem agrupados em n classes, com valores w_i na i^{th} classe

Aproximações para dados agrupados

Suponha que os dados estejam agrupados em K classes, com frequências $f_1, f_2, \ldots f_K$, e os pontos médioas das classes sejam $m_1, m_2, \ldots, m_K$

- Para uma classe de n observações, a media é

$$\bar{x} = \frac{\sum_{i=1}^{K} f_i m_i}{n}$$

where $n = \sum_{i=1}^{K} f_i$

Aproximações para dados agrupados

Suponha que os dados estejam agrupados em K classes, com frequências $f_1, f_2, \ldots f_K$, e os pontos médioas das classes sejam $m_1, m_2, \ldots, m_K$

- Para uma amostra de n observações, a variância é

$$s^2 = \frac{\sum_{i=1}^{K} f_i (m_i - \bar{x})^2}{n-1}$$

2.4

A amostra de covariância

- A covariância mede a força da relação linear entre duas variáveis
- A covariância da população:

$$\text{Cov}(x,y) = \sigma_{xy} = \frac{\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y)}{N}$$

- A covariância da amostra:

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

- Foca apenas a força do relacionamento
- Nenhum efeito causal está implícito

Interpretando a Covariância

- **Covariância entre duas variáveis:**

Cov(x,y) > 0. → x e y tendem a se mover na mesma direção

Cov(x,y) < 0. → x e y tendem a se mover em direções opostas

Cov(x,y) = 0 → x e y são independentes

Coeficiente de correlação

- Mede a força relativa da relação linear entre duas variáveis

- Coeficiente de correlação populacional:

$$\rho = \frac{Cov(x,y)}{\sigma_X \sigma_Y}$$

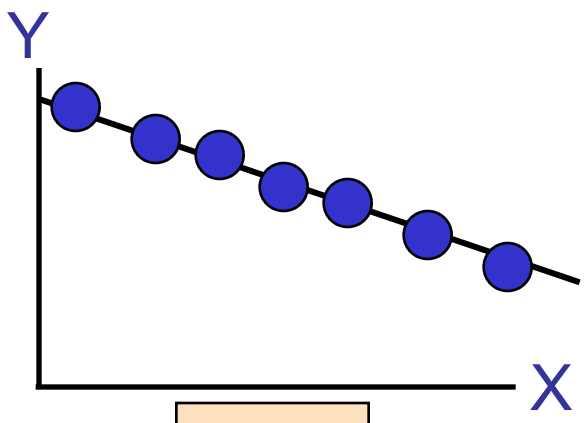
- Coeficiente de correlação da amostra:

$$r = \frac{Cov(x,y)}{s_X s_Y}$$

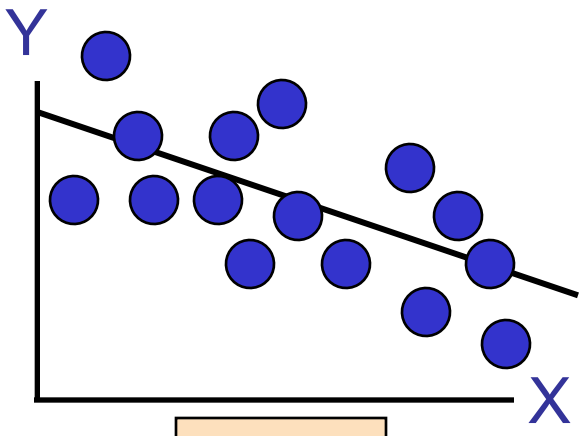
Características do Coeficiente de Correlação, r

- Unidade livre de medida
- Varia entre -1 e 1
- Quanto mais próximo de -1, mais forte é a relação linear negativa
- Quanto mais próximo de 1, mais forte é a relação linear positiva
- Quanto mais próximo de 0, mais fraca qualquer relação linear positive
- Se 0 não elimina haver relação não linear

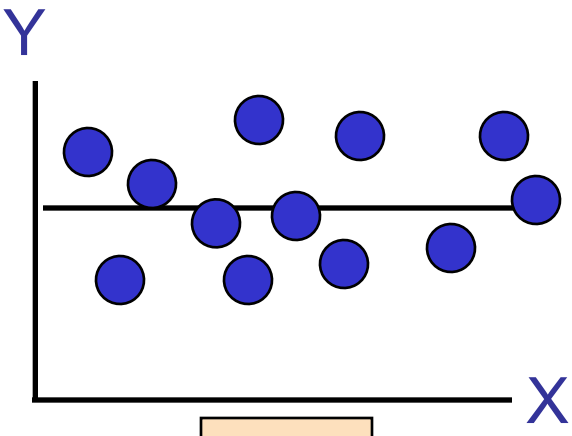
Gráficos de dispersão de dados com vários coeficientes de correlação



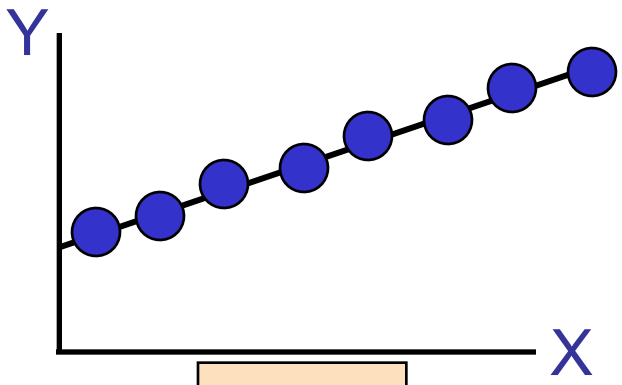
r = -1



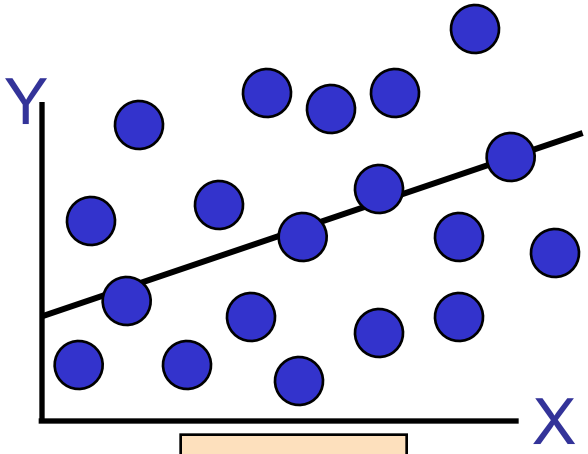
r = -.6



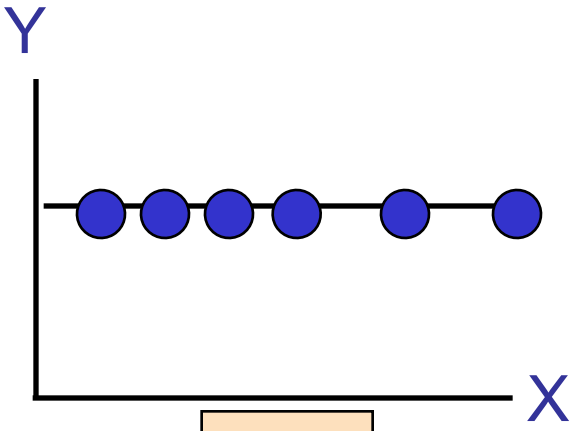
r = 0



r = +1



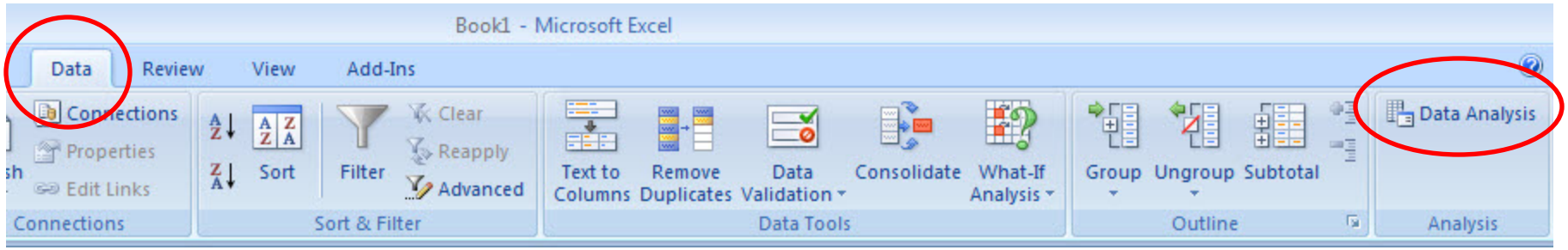
r = +.3



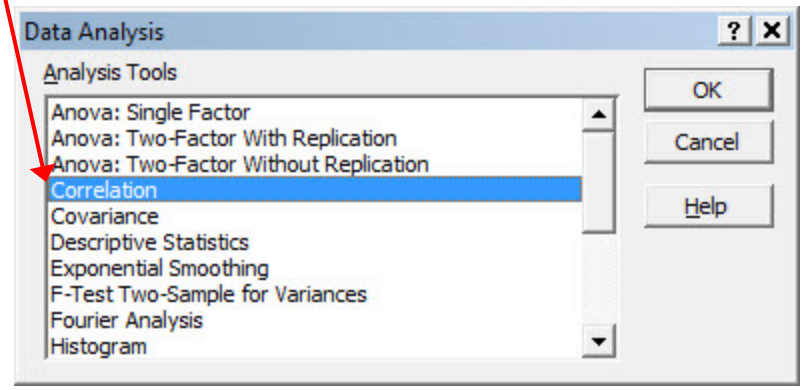
r = 0

Usando Excel para calcular Coeficiente de correlação

- Select Data / Data Analysis

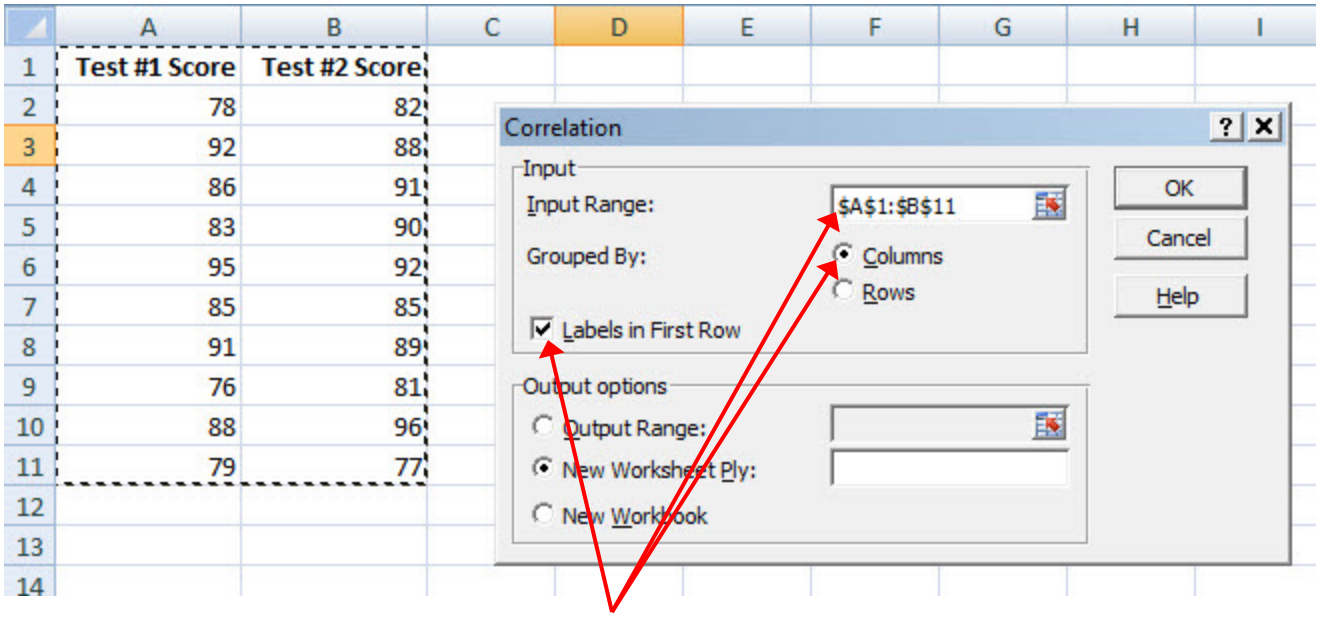


- Choose Correlation from the selection menu
- Click OK . . .



Usando Excel para calcular Coeficiente de correlação

(cont)



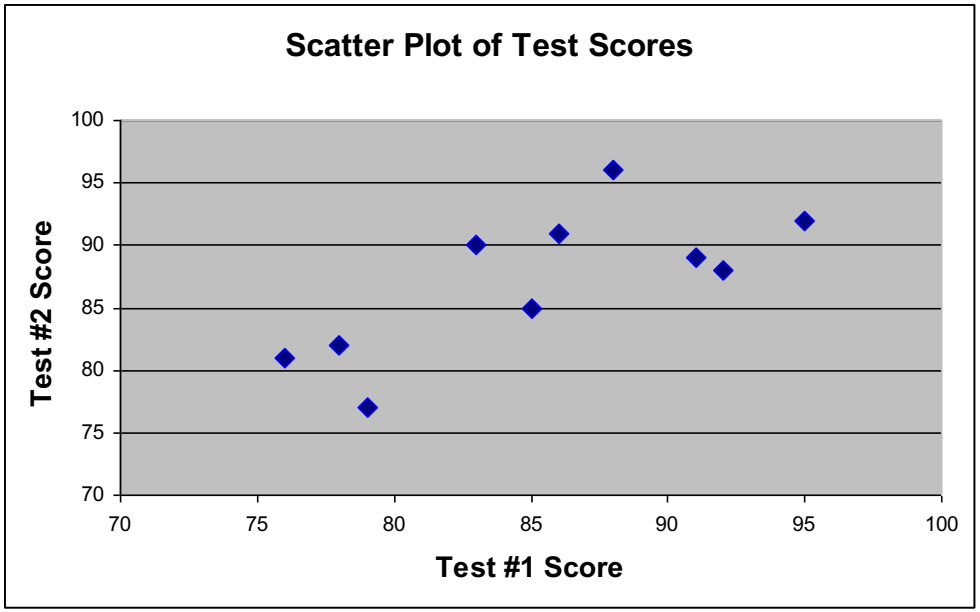
- Input data range and select appropriate options
- Click OK to get output

	A	B	C
1		*Test #1 Score*	*Test #2 Score*
2	Test #1 Score	1	
3	Test #2 Score	0.733243705	1
4			

Interpretando o Resulto

- r = .733

- existe uma relação linear positiva relativamente forte entre test score #1 e test score #2



- Alunos com notas altas no primeiro teste tendem a ter notas altas no segundo teste