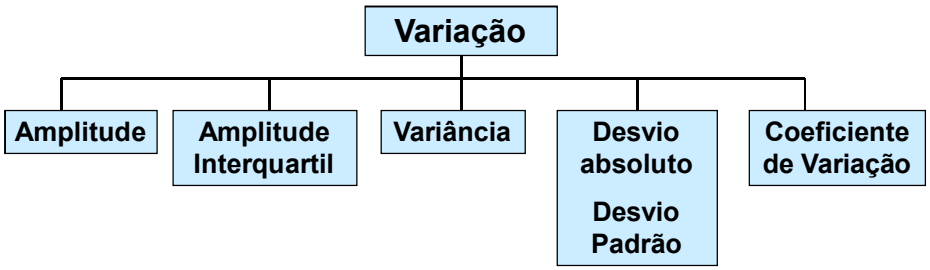
Medidas Numéricas Descritivas:

Medidas de dispersão

Medidas de Variação

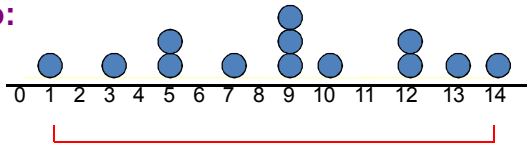


Amplitude

- Medida de variação mais simples
- Diferença entre o maior e o menor valor de um conjunto de dados:

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

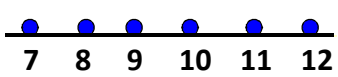
Exemplo:



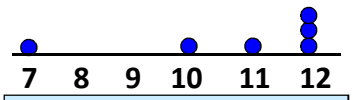
Amplitude = 14 - 1 = 13

Desvantagens da Amplitude

Ignora a forma como os dados são distribuídos



Amplitude = 12 - 7 = 5



Amplitude = 12 - 7 = 5

Sensitiva a valores extremos

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

Amplitude = 5 - 1 = 4

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

Amplitude = 120 - 1 = 119

Amplitude Interquartil (AI)

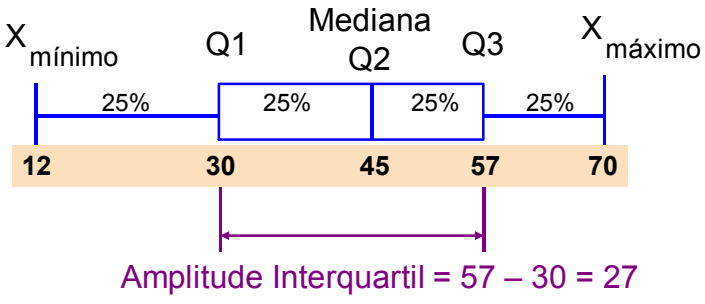
- Pode-se eliminar os problemas com valores extremos com a **amplitude interquartil**

AI = 3º quartil – 1º quartil = $Q_3 - Q_1$

Esta medida de variação não é afetada se uma fração pequena dos valores é muito pequena ou muito grande

Amplitude Interquartil

Exemplo:



Vamos introduzir o desvio de um valor x_i em relação a média:

Desvio $= x_i - \bar{x}$

Poderiamos tentar medir a dispersão em torno dos dados como a soma dos desvios. Entretanto

$$\sum_{i=1}^{n}(x_i - \bar{x}) = 0$$

Exemplo: 10 12 14 15 17 18 18 24

$n = 8 \qquad \bar{X} = 16$

$$\sum_{i=1}^{n}(x_i - \bar{x}) = (10-16)+(12-16)+(14-16)+\Lambda+(24-16) = 0$$

Resolve-se o problema usando o módulo!

Desvio médio absoluto (DM)

Usa a soma dos módulos dos desvios.

$$DM = \frac{\sum_{i=1}^{n}\left|x_i - \bar{x}\right|}{n}$$

Entretanto, o valor absoluto dos desvios é mais difícil de tratamento matemático (no uso de derivadas, por exemplo)

Exemplo

Amostra (X_i) : 10 12 14 15 17 18 18 24

$n = 8 \qquad \overline{X} = 16$

$$DM = \frac{\sum_{i=1}^{n} |x_i - \overline{x}|}{n} = \frac{|10-16| + |12-16| + |14-16| + \Lambda + |24-16|}{10}$$

$$DM = \frac{26}{10} = 2{,}6$$

Variância amostral

- Mede a dispersão em torno da média

Variância da amostra:

$$S^2 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n-1}$$

onde $\overline{X}$ = média

n = tamanho da amostra

X_i = i-ésimo valor de X

Note que se a variável x tem uma unidade de medida, por exemplo quilograma, metro, reais,…, a variância tem o inconveniente de ter a unidade da variável ao quadrado.

A fórmula para a variância pode ser reescrita como:

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

Utiliza apenas os valores dos dados

Desvio-Padrão amostral

- Medida de variação mais utilizada
- Mostra a variação em torno da média
- Raiz quadrada da variância
- Tem a **mesma unidade dos dados originais**

Desvio-Padrão da amostra:

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

Exemplo

Amostra (X_i) : 10 12 14 15 17 18 18 24

$n = 8 \qquad \bar{X} = 16$

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

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

$$= \sqrt{\frac{130}{7}} = 4{,}3095$$

Variância para dados agrupados em classes

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

k : número de classes

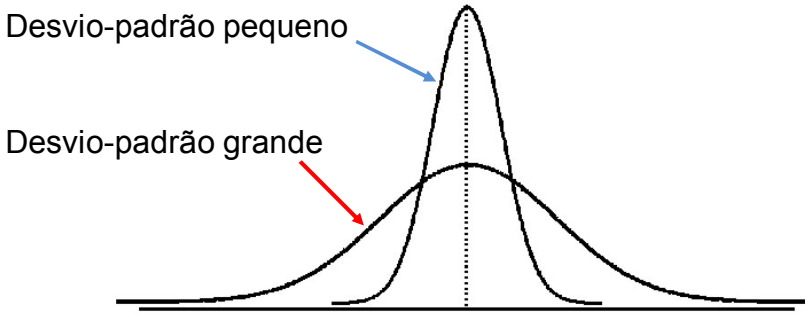
PM_i : ponto médio da classe i

f_i : frequência da classe i

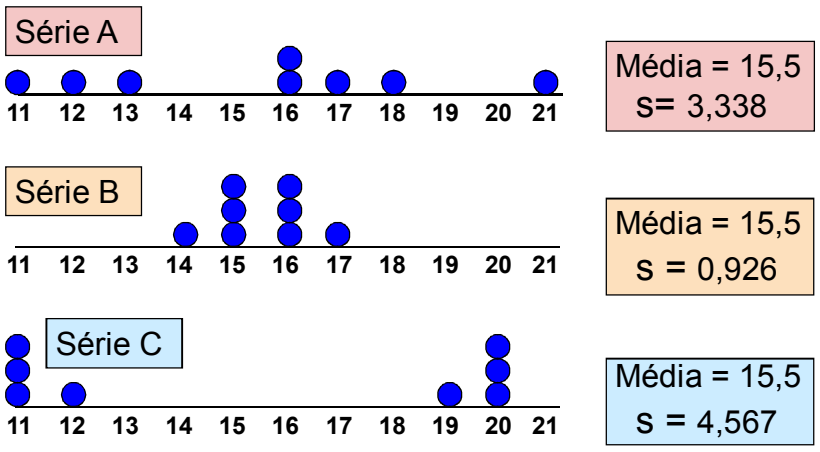
$$\bar{\bar{x}} = \frac{\sum_{i=1}^{k} f_i PM_i}{n} \quad \text{(média dos pontos médios)}$$

Desvio padrão : $s = \sqrt{s^2}$

Mensurando a Variação



Comparando Desvios-Padrão: os conjuntos de dados abaixo tem todos a mesma média. A diferença está entre eles é a dispersão.



Vantagens da Variância e do Desvio Padrão

- Todos os valores do conjunto de dados são utilizados no cálculo
- Valores muito distantes da média recebem peso diferenciado

Coeficiente de Variação

- Mede **a variação relativa à média**
- Costuma-se ser dado em porcentagem (%)
- Pode ser utilizada para comparar duas ou mais séries de dados em unidades diferentes

$$CV = \left(\frac{S}{\overline{X}} \right) \cdot 100\%$$

Comparando Coeficientes de Variação

Considere o preço e a variação de duas ações do mercado

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

$$CV_A = \left(\frac{S}{\overline{X}} \right) \cdot 100\% = \frac{\$5}{\$50} \cdot 100\% = 10\%$$

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

$$CV_B = \left(\frac{S}{\overline{X}} \right) \cdot 100\% = \frac{\$5}{\$100} \cdot 100\% = 5\%$$

As duas ações possuem o mesmo desvio-padrão, mas o preço da ação B teve uma variação relativa menor.

Escore Z

- Diferença entre um valor e a média, dividida pelo desvio-padrão

$$Z = \frac{X - \overline{X}}{S}$$

- Medida de distância da média (por exemplo, um escore Z igual a 2 significa que o valor está a 2 desvios da média)
- Um escore Z acima de 3 ou abaixo de -3 é considerado um valor extremo (outlier). Esse é um dos critérios para encontrar outlier, mas não significa que o valor seja um erro ou que não deveria fazer parte dos dados. Significa que deve ser examinado. Usar medidas baseadas em ordenamento podem ser mais adequadas para dados com muitos valores extremos.

Escore Z

Exemplo:

- Se a média é 14 e o desvio-padrão é 3, qual é o escore Z para o valor 18,5?

$$Z = \frac{x - \bar{x}}{s} = \frac{18,5 - 14}{3} = 1,5$$

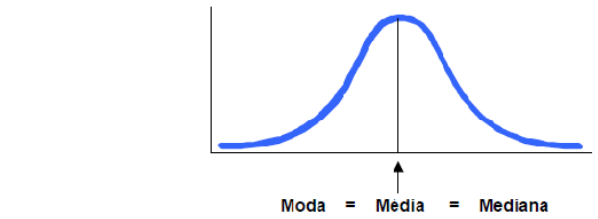
- O valor 18,5 está a 1,5 desvio-padrão acima da média
- Um escore Z negativo significa que o valor é menor que a média

Forma da Distribuição

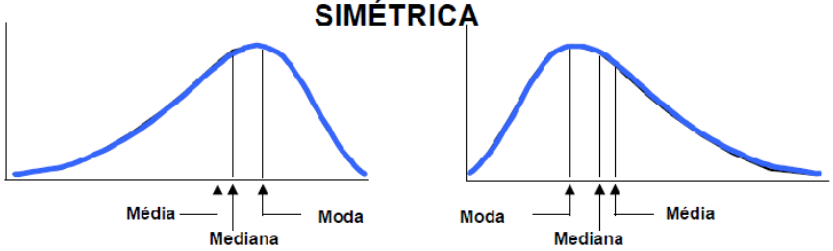
- O histograma mostra como os dados são distribuidos
- E podem apresentar formas Simétricas ou Assimétricas

Simétrica: Dados são simétricos se a metade esquerda de seu histograma é aproximadamente a imagem-espelho da metade direita.

Assimétrica: Uma distribuição de dados é assimétrica quando não é simétrica.



SIMÉTRICA



Assimétrica a esquerda: Média e mediana a esquerda da moda	Assimétrica a direita: Média e mediana a direita da moda

Medidas Numéricas Descritivas para uma População

- Medidas numéricas para uma população são chamadas **parâmetros**
- A média da população é a soma dos valores que compõem a população, dividida pelo tamanho da população (**N**)

$$\mu = \frac{\sum_{i=1}^{N} X_i}{N} = \frac{X_1 + X_2 + \Lambda + X_N}{N}$$

onde

μ = média da população

N = tamanho da população

X_i = i-ésimo valor de X

Variância da População

- Mede a dispersão em torno da média

Variância da População:

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

onde μ = média da população

N = tamanho da população

X_i = i-ésimo valor de X

Desvio-Padrão da População

- Medida de variação mais utilizada
- Mostra a variação em torno da média
- Raiz quadrada da variância
- Tem a a mesma unidade dos dados originais

  – Desvio-Padrão da população

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

regra de Chebyshev

-aplica-se a todos os conjuntos de dados.

-a proporção (ou fração) de qualquer conjunto de dados a menos de K desvios-padrão a contar da média é sempre pelo menos

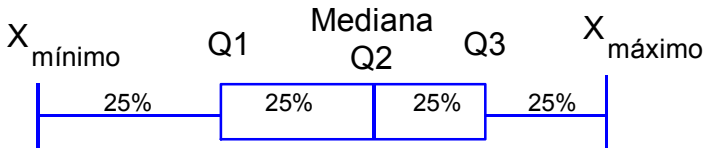
$1-\frac{1}{k^2}$, k é um número positivo maior do que 1

-pelo menos 3/4 (75%) de todos os valores estão no intervalo que vai de 2 desvios-padrão abaixo da média a 2 desvios-padrão acima da média (**75%** **em** $\bar{x} - 2s < x < \bar{x} + 2s$)

-pelo menos 8/9 (89%) de todos os valores estão no intervalo que vai de 3 desvios-padrão abaixo da média até 3 desvios-padrão acima da média. (**89%** **em** $\bar{x} - 3s < x < \bar{x} + 3s$)

Box-Plot (gráfico de caixa): representação gráfica dos dados com base no resumo de 5 números:

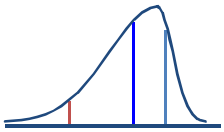
Mínimo -- Q1 -- Mediana -- Q3 -- Máximo

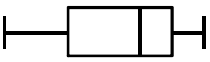


- o retângulo do *boxplot* correponde aos 50% valores centrais da distribuição.
- Um box-plot pode ser representado de forma vertical ou horizontal. Ele da uma idéia da dispersão, assimetria e da distribuiçoes de dados. São úteis na comparação de conjuntos de dados se desenhados na mesma escala.

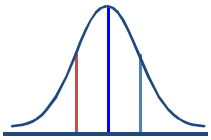
Relação entre Forma de Distribuição e Box-Plot

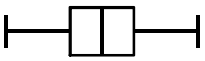
assimétrica à Esquerda



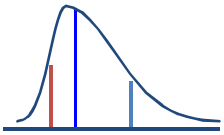


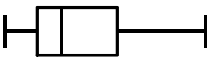
Simétrica





assimétrica à Direita





Exemplo de Box-Pot

Min	Q1		Q2		Q3		Max
0	2	2	3	4	5	10	27

- box-plot :

A distribuição do valores é assimétrica a direita.

Ajudam na escolha de medidas numéricas descritivas apropriadas:

(a) o propósito para o qual o resumo descritivo dos dados é realizado.

(b) Facilidade de interpretação.

(c) grau de sensibilidade a valores extremos.

(d) Potencial para uso em inferência estatística.

Algumas conclusões gerais sobre a descrição de dados

(a) Um gráfico mostrando a forma que os dados são distribuídos é importante para um resumo estatístico descritivo (histograma, box-plot,...)

(b) Quando várias medidas são compatíveis com o propósito do resumo descritivo é interessante apresentar todos (média, mediana, amplitude, desvio padrão,...)

(c) A forma da distribuição é importante na interpretação das medidas descritivas.

(d) Valores discrepantes (observações não usuais) devem ser relatados (por exemplo, em nota de rodapé).

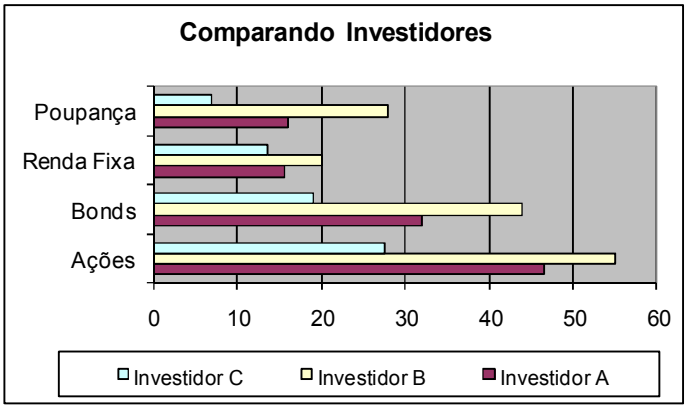
Dados Categóricos Multivariados

Tabela de Contingência para opções de investimentos

Categoria	Investidor A	Investidor B	Investidor C	Total
Ações	46,5	55	27,5	**129**
Bonds	32,0	44	19,0	**95**
RF	15,5	20	13,5	**49**
Poupança	16,0	28	7,0	**51**
Total	**110,0**	**147**	**67,0**	**324**

(Os valores individuais poderiam ser expressos em %)

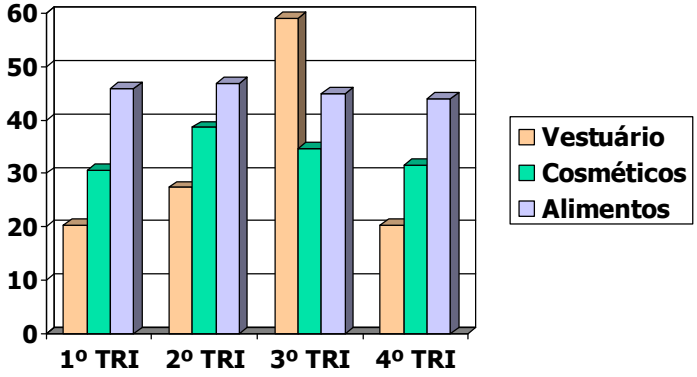
Podemos representar os dados em um Gráfico de Barras Paralelas



Poderiamos usar barras verticais.

Exemplo: vendas por trimestre

	1º TRI	2º TRI	3º TRI	4º TRI
Vestuário	20,4	27,4	59	20,4
Cosméticos	30,6	38,6	34,6	31,6
Alimentos	45,9	46,9	45	43,9



Correlação e Regressão

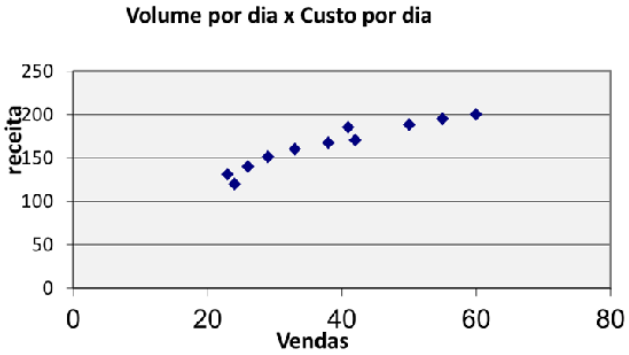
- **Dados numéricos bivariados:**
- **$(x_1, y_1), (x_2,y_2), (x_3,y_3)...(x_n,y_n)$**

- **Gráfico de Dispersão:** utilizado para identificar possíveis relações entre duas variáveis numéricas

- **O gráfico:** Uma variável é representada no eixo vertical (y) e a outra no eixo horizontal (x)

Exemplo

receita	vendas
131	23
120	24
140	26
151	29
160	33
167	38
185	41
170	42
188	50
195	55
200	60



No caso podemos identificar uma relação aproximadamente linear entre custo e volume de compras

Intensidade da relação linear

r: Coeficiente de correlação amostral

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

onde :

$$s_x = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} \qquad s_y = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}}$$

são os desvios - padrão da variável x e y, respectivamente.

se definirmos : $S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 = (n-1)\, s_x^{\,2}$

$$S_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2 = (n-1)\, s_y^{\,2}$$

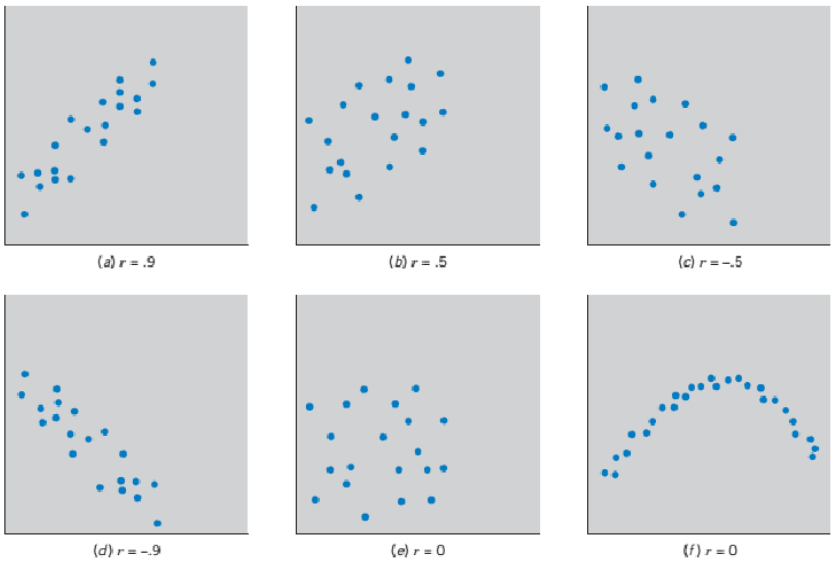
$$S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

Podemos escrever r como:

$$r = \frac{S_{xy}}{\sqrt{S_{xx}}\,\sqrt{S_{yy}}}$$

propriedade : $-1 \le r \le 1$

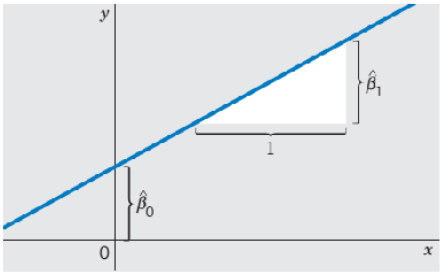
Mede o quanto o diagrama de dispersão é próximo de uma reta.



Correlação não implica causalidade: que uma variável é a causa da outra

Correlação espúria (falsa): correlação estatística existente entre duas variáveis em que não existe relação de causa e efeito entre elas. Podem ocorrer por mera coincidência ou devido a uma terceira variável que afeta as variáveis em estudo, mas não foi incluída.

Mas se existe correlação significativa e razões para usar uma variável para prever a outra, qual a equação da reta que melhor se ajusta ao gráfico de dispersão???



$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x.$$

Ou seja, como obter os coeficientes $\hat{\beta}_0$ e $\hat{\beta}_1$

a partir dos valores amostrais?

Melhor ajuste: método dos mínimos quadrados:

minimiza – se a soma das distâncias vertical entre um ponto amostral e a reta procurada : $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$

Solução:

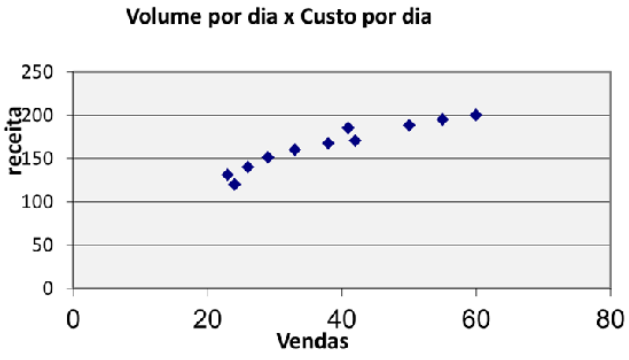
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (x - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

De volta ao exemplo

receita	vendas
131	23
120	24
140	26
151	29
160	33
167	38
185	41
170	42
188	50
195	55
200	60



No caso podemos identificar uma relação aproximadamente linear entre custo e volume de compras

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{421}{11} = 38.2727$$

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{1807}{11} = 164.2727$$

$$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 = 3229.182$$

$$S_{yy} = \sum_{i=1}^{n} (y_i - y)^2 = 7104.182$$

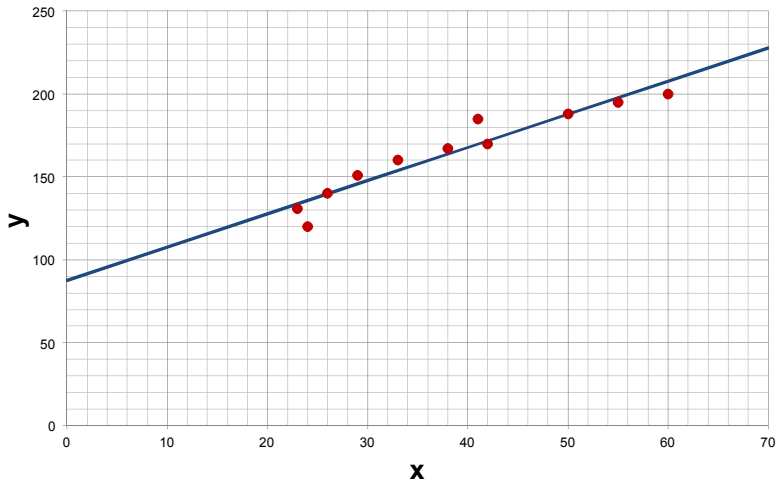
$$S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = 1612.182$$

Temos :

$$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} = \frac{1612.182}{\sqrt{3229.182}\sqrt{7104.182}} = 0.9542$$

$$\beta_1 = \frac{S_{xy}}{S_{xx}} = \frac{1612.182}{3229.182} = 2.003$$

$$\beta_0 = \bar{y} - \beta_1 \bar{x} = 87.613$$



y = 2.003x + 87.613