Iniciação à Estatística

Prof. Fernando de Souza Bastos
fernando.bastos@ufv.br

Departamento de Estatística
Universidade Federal de Viçosa
Campus UFV - Viçosa

Sumário

O resumo de um conjunto de dados por uma única medida representativa de posição central esconde toda a informação sobre a variabilidade do conjunto de observações. Por exemplo, suponhamos que cinco grupos de alunos submeteram-se a um teste, obtendo-se as seguintes notas:

- Grupo A (Variável X): 3,4,5,6,7
- Grupo B (Variável Y): 1,3,5,7,9
- Grupo C (Variável Z): 5,5,5,5,5
- Grupo D (Variável W): 3,5,5,7
- Grupo E (Variável V): 3,5,5,6,6

Vemos que $\bar{x} = \bar{y} = \bar{z} = \bar{w} = \bar{v} = 5$. A identificação de cada uma destas séries por sua média (5, em todos os casos) nada informa sobre suas diferentes variabilidades. Notamos, então, a conveniência de serem criadas medidas que sumarizem a variabilidade de um conjunto de observações e que nos permita, por exemplo, comparar conjuntos diferentes de valores, como os dados acima, segundo algum critério estabelecido.

Um critério frequentemente usado para tal fim é aquele que mede a dispersão dos dados em torno de sua média. Duas medidas são as mais usadas: desvio médio e variância.

$$Dm(X) = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n} \qquad Var(X) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}$$

O princípio básico é analisar os desvios das observações em relação à média dessas observações. Desvio é interpretado como o afastamento de uma observação em relação a uma determinada medida de posição.

Variância Amostral

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

Variância Amostral

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

Medidas foram tomadas de 3 grupos, os resultados foram:

- Grupo 1 (em Kg): 1,3,5,7,9
- Grupo 2 (em metros): 1.8,3.8,5.8,7.8,9.8
- Grupo 3 (em R$): 1002,1004,1006,1008,1010

Calcule a variância destes 3 grupos!

Sendo a variância uma medida de dimensão igual ao quadrado da dimensão dos dados (por exemplo, se os dados são expressos em cm, a variância será expressa em cm^2), ela pode causar problemas de interpretação. Costuma-se usar, então, o desvio padrão, que é definido como a raiz quadrada positiva da variância.

$$dp(X) = S(X) = \sqrt{Var(X)} \tag{1}$$

Ambas as medidas de dispersão (Dm e dp) indicam em média qual será o "erro" (desvio) cometido ao tentar substituir cada observação pela medida resumo do conjunto de dados (no caso, a média). Resolvido, portanto, o problema da unidade de medida.

Coeficiente de Variação

$$CV(X) = \frac{dp(X)}{\bar{X}} \cdot 100 = \frac{S(X)}{\bar{X}} \cdot 100 \tag{2}$$

Mesmo o DP pode induzir à conclusões errôneas com relação à variabilidade. Suponha dois conjuntos de dados $D_1 = \{10, 20, 30\}$ e $D_2 = \{10000, 10010, 10020\}$. Note que nestes casos $\bar{x}_1 = 20, dp(x) = 10, \bar{x}_2 = 10010$ e $dp(x_2) = 10$. Porém, em termos percentuais, o primeiro conjunto de dados é mais heterogênio.

Coeficiente de Variação

$$CV(X) = \frac{dp(X)}{\bar{X}} \cdot 100 = \frac{S(X)}{\bar{X}} \cdot 100 \qquad (2)$$

Mesmo o DP pode induzir à conclusões errôneas com relação à variabilidade. Suponha dois conjuntos de dados $D_1 = \{10, 20, 30\}$ e $D_2 = \{10000, 10010, 10020\}$. Note que nestes casos $\bar{x}_1 = 20, dp(x) = 10, \bar{x}_2 = 10010$ e $dp(x_2) = 10$. Porém, em termos percentuais, o primeiro conjunto de dados é mais heterogênio.

Obs.: O C.V. é utilizado para avaliar qual o percentual da média que o desvio-padrão representa. Isso é chamado de homogeneidade. 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.

Erro Padrão da Média

$$S(\overline{X}) = \sqrt{\frac{S^2(X)}{n}}$$

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

Erro Padrão da Média

$$S(\overline{X}) = \sqrt{\frac{S^2(X)}{n}}$$

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

Exemplo

Considere duas amostras de tamanhos $n = 6$, em que $S_A^2 = 5,6$ e $S_B^2 = 28$. Temos que:

$$S(\bar{X}_A) = \sqrt{\frac{5,6}{6}} = 0,966; \quad S(\bar{X}_B) = \sqrt{\frac{28}{6}} = 2,1602$$

e, portanto, a amostra A forneceu uma estimativa de média associada à uma maior precisão.

Erro Padrão da Média

Notem que o erro padrão da média é:

- Inversamente proporcional ao tamanho da amostra;
- Diretamente proporcional à variância da amostra.

Amplitude Total

A amplitude total (AT) é dada pela diferença entre o maior e o menor valor de uma amostra ou de um conjunto de dados. Se $X_1, X_2, X_3, \cdots, X_n$ é uma amostra de valores da variável X, então:

$$AT_X = X_{(n)} - X_{(1)}$$

Recorde que a notação $X_{(i)}$ indica estatísticas de ordem da amostra, isto é: $X_{(1)} \leq X_{(2)} \leq X_{(3)} \leq \cdots \leq X_{(n)}$. Portanto, a amplitude total indica que o desvio entre duas observações quaisquer é no máximo igual a AT.

Coeficiente de Correlação Amostral

Tabela: Anos de Serviço (X) versus Nº de Clientes (Y)

Agente	X	Y
A	2	48
B	4	56
C	5	64
D	6	60
E	6	65
F	6	63
G	7	67
H	8	70
I	8	71
J	10	72

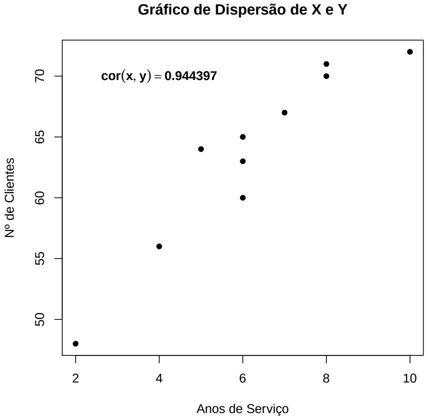
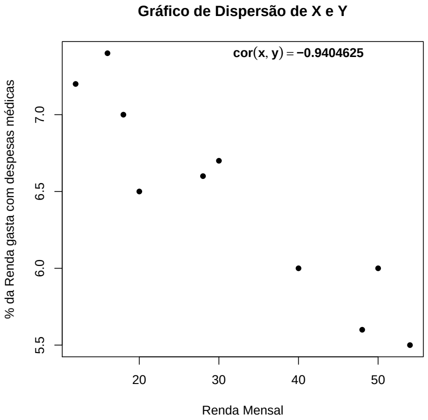


Tabela: Renda bruta mensal (X) e porcentagem da renda gasta em saúde (Y).

Família	X	Y
A	12	7.2
B	16	7.4
C	18	7.0
D	20	6.5
E	28	6.6
F	30	6.7
G	40	6.0
H	48	5.6
I	50	6.0
J	54	5.5



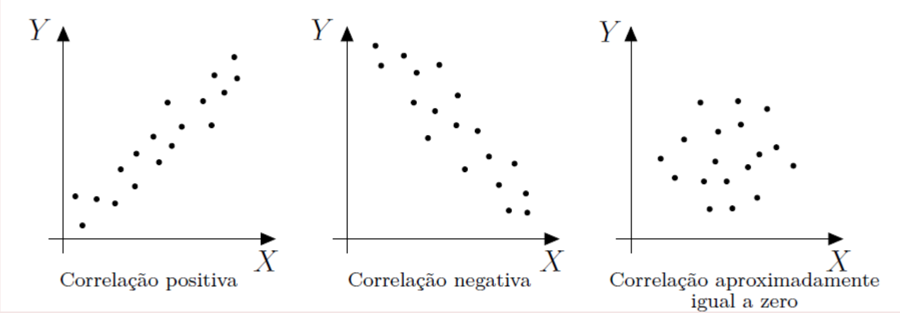
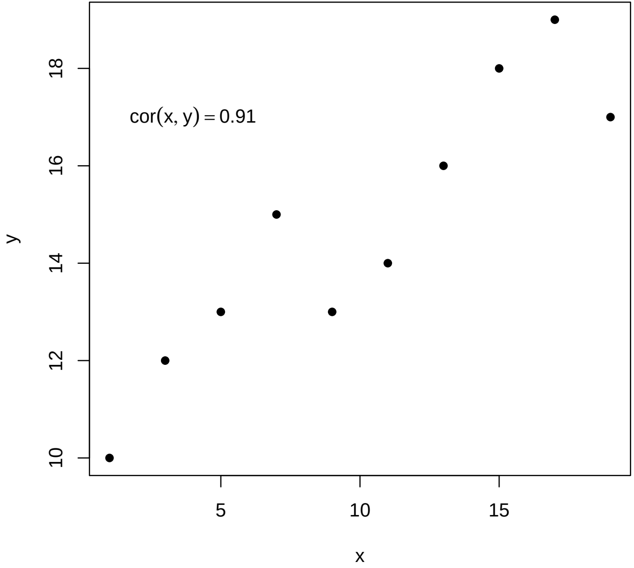


Figura: Representação gráfica de diversos coeficientes de correlação.

Gráfico de Dispersão



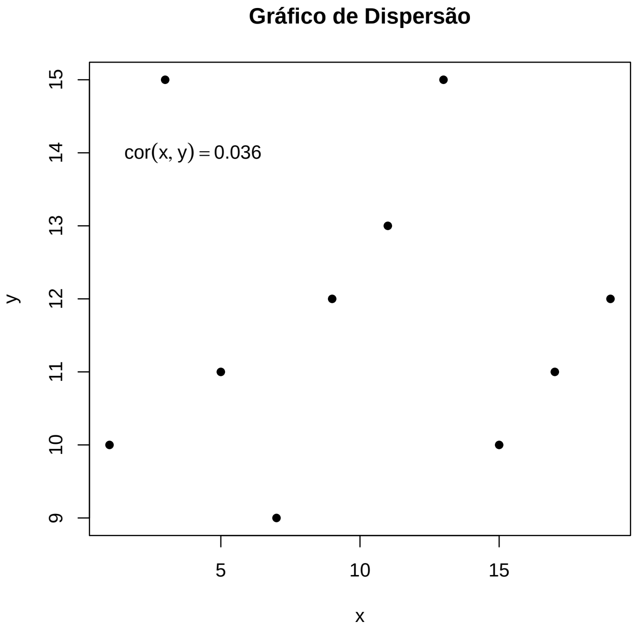


Gráfico de Dispersão
cor(x, y) = 0.036
15
14
13
12
11
10
9
y
5
10
15
x

Gráfico de Dispersão

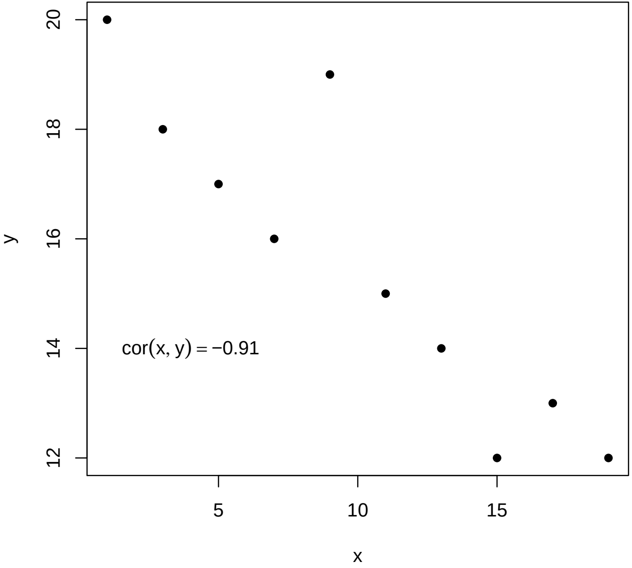



Gráfico de Dispersão
50
40
30
20
10
y
cor(x, y) = 0.015
5
10
15
x

Coeficiente de Correlação Amostral de Pearson

O coeficiente de correlação (r ou $\hat{\rho}$) mede o grau de associação linear entre duas variáveis aleatórias X e Y. Considere,

X_i	X_1	X_2	$\cdots$	X_n
Y_i	Y_1	Y_2	$\cdots$	Y_n

Assim, o coeficiente de correlação entre X e Y é dado por:

$$r_{XY} = \frac{S_{XY}}{\sqrt{S_X^2 \cdot S_Y^2}} = \frac{\dfrac{SPD_{XY}}{n-1}}{\sqrt{\dfrac{SQD_X}{n-1} \cdot \dfrac{SQD_Y}{n-1}}} = \frac{SPD_{XY}}{\sqrt{SQD_X \cdot SQD_Y}}$$

Temos,

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

Temos,

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

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

Temos,

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

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

Não é difícil provar que o coeficiente de correlação satisfaz:

$$-1 \leq cor(X, Y) \leq 1$$

DEF: Dados n pares de valores $(x_1, y_1), \cdots, (x_n, y_n)$, chamaremos de covariância entre as duas variáveis X e Y a igualdade:

$$S_{XY} = cov(X, Y) = \sum_{i=1}^{n} \frac{(x_i - \bar{x})(y_i - \bar{y})}{n}$$

DEF: Dados n pares de valores $(x_1, y_1), \cdots, (x_n, y_n)$, chamaremos de covariância entre as duas variáveis X e Y a igualdade:

$$S_{XY} = cov(X, Y) = \sum_{i=1}^{n} \frac{(x_i - \bar{x})(y_i - \bar{y})}{n}$$

Com a definição acima, o coeficiente de correlação pode ser escrito como:

$$r_{XY} = cor(X, Y) = \frac{cov(X, Y)}{dp(X)dp(Y)}$$

A covariância mede a relação linear entre duas variáveis. A covariância é semelhante à correlação entre duas variáveis, no entanto, elas diferem nas seguintes maneiras:

A covariância mede a relação linear entre duas variáveis. A covariância é semelhante à correlação entre duas variáveis, no entanto, elas diferem nas seguintes maneiras:

- Os coeficientes de correlação são padronizados. Assim, um relacionamento linear perfeito resulta em um coeficiente de correlação 1. A correlação mede tanto a força como a direção da relação linear entre duas variáveis.

A covariância mede a relação linear entre duas variáveis. A covariância é semelhante à correlação entre duas variáveis, no entanto, elas diferem nas seguintes maneiras:

- Os coeficientes de correlação são padronizados. Assim, um relacionamento linear perfeito resulta em um coeficiente de correlação 1. A correlação mede tanto a força como a direção da relação linear entre duas variáveis.

- Os valores de covariância não são padronizados. Como os dados não são padronizadas, é difícil determinar a força da relação entre as variáveis.

Referências

REGAZZI, Adair José et al. **Roteiro de Aulas da Disciplina Iniciação à Estatística**. [S.l.: s.n.], 2022.