MAF 105 - Iniciação à Estatística

Prof. Fernando de Souza Bastos

Instituto de Ciências Exatas e Tecnológicas
Universidade Federal de Viçosa
Campus UFV - Florestal

Sumário

Introdução

Até agora vimos como organizar e resumir informações pertinentes a uma única variável (ou a um conjunto de dados), mas freqüentemente estamos interessados em analisar o comportamento conjunto de duas ou mais variáveis aleatórias.

Introdução

Os dados aparecem na forma de uma matriz, usualmente com as colunas indicando as variáveis e as linhas os indivíduos (ou elementos).

Introdução

Indivíduo	Variável					
	X_1	X_2	$\cdots$	X_j	$\cdots$	X_p
1	X_{11}	X_{12}	$\cdots$	X_{1j}	$\cdots$	X_{1p}
2	X_{21}	X_{22}	$\cdots$	X_{2j}	$\cdots$	X_{2p}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
i	X_{i1}	X_{i2}	$\cdots$	X_{ij}	$\cdots$	X_{ip}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	X_{n1}	X_{n2}	$\cdots$	X_{nj}	$\cdots$	X_{np}

Tabela: Tabela de dados MEPS 2001

Obs.	Damb	Ind	lDamb	Idade	Sexo	Educ	Blhisp	Totcr	Ins	Renda
50	0	0	0.000	4.5	1	12	0	2	0	24.480
51	996	1	6.904	6.4	1	9	0	2	0	30.000
52	2765	1	7.925	3.7	0	16	0	1	0	47.000
53	89	1	4.489	6.3	0	14	0	0	0	50.013
54	568	1	6.342	5.7	1	16	0	0	0	50.013
55	0	0	0.000	3.9	0	16	0	0	1	14.424
56	204	1	5.318	4.0	1	16	0	1	1	28.849
57	0	0	0.000	2.7	0	12	0	0	0	39.000
58	108	1	4.682	2.5	1	12	0	0	0	18.700
59	1293	1	7.165	3.0	1	12	0	0	0	5.667
60	133	1	4.890	5.0	1	14	0	0	1	1.370

Quando consideramos duas variáveis (ou dois conjuntos de dados), podemos ter três situações:

1. as duas variáveis são qualitativas;
2. as duas variáveis são quantitativas; e
3. uma variável é qualitativa e outra é quantitativa.

As técnicas de análise de dados nas três situações são diferentes.

Suponha que queiramos analisar o comportamento conjunto das variáveis X : grau de instrução e Y : região de procedência, cujas observações estão contidas na Tabela abaixo:

Tabela: Distribuição Conjunta das frequências das variáveis Y e X.

Y \ X	Ensino Fundamental	Ensino Médio	Ensino Superior	Total
Capital	4	5	2	11
Interior	3	7	2	12
Outra	5	6	2	13
Total	12	18	6	36

- 4 indivíduos procedem da capital e possuem o ensino fundamental;

- 4 indivíduos procedem da capital e possuem o ensino fundamental;
- Na última coluna, está representada a frequência absoluta da variável Y;

- 4 indivíduos procedem da capital e possuem o ensino fundamental;
- Na última coluna, está representada a frequência absoluta da variável Y;
- Na última linha está representada a frequência absoluta da variável X;

- 4 indivíduos procedem da capital e possuem o ensino fundamental;
- Na última coluna, está representada a frequência absoluta da variável Y;
- Na última linha está representada a frequência absoluta da variável X;
- As frequências absolutas (parte interna da tabela) são chamadas de frequências absolutas conjuntas entre X e Y.

Em vez de trabalharmos com as freqüências absolutas, podemos construir tabelas com as freqüências relativas (proporções), como foi feito no caso unidimensional. Mas aqui existem três possibilidades de expressarmos a proporção de cada casela:

- em relação ao total geral;

Em vez de trabalharmos com as freqüências absolutas, podemos construir tabelas com as freqüências relativas (proporções), como foi feito no caso unidimensional. Mas aqui existem três possibilidades de expressarmos a proporção de cada casela:

- em relação ao total geral;
- em relação ao total de cada linha;

Em vez de trabalharmos com as freqüências absolutas, podemos construir tabelas com as freqüências relativas (proporções), como foi feito no caso unidimensional. Mas aqui existem três possibilidades de expressarmos a proporção de cada casela:

- em relação ao total geral;
- em relação ao total de cada linha;
- ou em relação ao total de cada coluna.

De acordo com o objetivo do problema em estudo, uma delas será a mais conveniente.

Tabela: Distribuição conjunta das proporções (em porcentagem) em relação ao total geral das variáveis Y e X.

Y \ X	Ensino Fundamental	Ensino Médio	Ensino Superior	Total
Capital	11%	14%	6%	31%
Interior	8%	19%	6%	33%
Outra	14%	17%	5%	36%
Total	33%	50%	17%	100%

Podemos, então, afirmar que 11% dos empregados vêm da capital e têm o ensino fundamental. Os totais nas margens fornecem as distribuições unidimensionais de cada uma das variáveis. Por exemplo, 31% dos indivíduos vêm da capital, 33% do interior e 36% de outras regiões.

Tabela: Distribuição conjunta das proporções (em porcentagem) em relação aos totais de cada coluna.

Y \ X	Ensino Fundamental	Ensino Médio	Ensino Superior	Total
Capital	33%	28%	33%	31%
Interior	25%	39%	33%	33%
Outra	42%	33%	34%	36%
Total	100%	100%	100%	100%

Entre os empregados com ensino médio:

- 28% vêm da capital;

Entre os empregados com ensino médio:

- 28% vêm da capital;
- 39% vêm do interior;

Entre os empregados com ensino médio:

- 28% vêm da capital;
- 39% vêm do interior;
- 33% vêm de outros locais.

Esse tipo de tabela serve para comparar a distribuição da procedência dos indivíduos conforme o grau de instrução.

Tabela: Distribuição conjunta das proporções (em porcentagem) em relação aos totais de cada linha.

Y \ X	Ensino Fundamental	Ensino Médio	Ensino Superior	Total
Capital	36.4%	45.4%	18.2%	100%
Interior	25%	58.3%	16.7%	100%
Outra	38.5%	46.1%	15.4%	100%
Total	33%	50%	17%	100%

Entre os empregados do interior:

- 25% têm Ensino Fundamental;

Entre os empregados do interior:

- 25% têm Ensino Fundamental;
- 58.3% têm Ensino médio;

Entre os empregados do interior:

- 25% têm Ensino Fundamental;
- 58.3% têm Ensino médio;
- 16.7% têm ensino superior.

Esse tipo de tabela serve para comparar a distribuição do grau de instrução dos indivíduos conforme a procedência.

Um dos principais objetivos de se construir uma distribuição conjunta de duas variáveis qualitativas é descrever a associação entre elas, isto é, queremos conhecer o grau de dependência entre elas, de modo que possamos prever melhor o resultado de uma delas quando conhecermos a realização da outra.

Por exemplo, se quisermos estimar qual a renda média de uma família moradora da cidade de São Paulo, a informação adicional sobre a classe social a que ela pertence nos permite estimar com maior precisão essa renda, pois sabemos que existe uma dependência entre as duas variáveis: renda familiar e classe social.

- Suponhamos que uma pessoa seja sorteada ao acaso e devamos adivinhar o sexo dessa pessoa. Existe sexo mais provável?

- Suponhamos que uma pessoa seja sorteada ao acaso e devamos adivinhar o sexo dessa pessoa. Existe sexo mais provável?
- E se a mesma pergunta fosse feita, porém fosse dito que a pessoa sorteada trabalha na indústria siderúrgica, qual a resposta mais provável?

- Suponhamos que uma pessoa seja sorteada ao acaso e devamos adivinhar o sexo dessa pessoa. Existe sexo mais provável?
- E se a mesma pergunta fosse feita, porém fosse dito que a pessoa sorteada trabalha na indústria siderúrgica, qual a resposta mais provável?

Ou seja, há um grau de dependência grande entre as variáveis sexo e ramo de atividade.

Queremos verificar se existe ou não associação entre o sexo e a carreira escolhida por 200 alunos de Economia e Administração.

Tabela: Distribuição conjunta de alunos segundo o sexo (X) e o curso escolhido (Y).

Y \ X	Masculino	Feminino	Total
Economia	85	35	120
Administração	55	25	80
Total	140	60	200

Fonte: Morettin e Bussab (2009).

Queremos verificar se existe ou não associação entre o sexo e a carreira escolhida por 200 alunos de Economia e Administração.

Tabela: Distribuição conjunta de alunos segundo o sexo (X) e o curso escolhido (Y).

Y \ X	Masculino	Feminino	Total
Economia	85	35	120
Administração	55	25	80
Total	140	60	200

Fonte: Morettin e Bussab (2009).

Inicialmente, verificamos que fica muito difícil tirar alguma conclusão, Devemos, pois, construir as proporções segundo as linhas ou as colunas para podermos fazer comparações.

Tabela: Distribuição conjunta das proporções de alunos segundo o sexo (X) e o curso escolhido (Y).

Y \ X	Masculino	Feminino	Total
Economia	61%	58%	60%
Administração	39%	42%	40%
Total	100%	100%	100%

Devemos, pois, construir as proporções segundo as linhas ou as colunas para podermos fazer comparações.

Tabela: Distribuição conjunta das proporções de alunos segundo o sexo (X) e o curso escolhido (Y).

Y \ X	Masculino	Feminino	Total
Economia	61%	58%	60%
Administração	39%	42%	40%
Total	100%	100%	100%

observar que, independentemente do sexo, 60% preferem Economia e 40% preferem Administração. Não havendo dependência entre as variáveis, esperaríamos essas mesmas proporções para cada sexo.

Outro exemplo:

Tabela: Distribuição conjunta das proporções de alunos segundo o sexo (X) e o curso escolhido (Y).

Y \ X	Masculino	Feminino	Total
Física	100(71%)	20(33%)	120(60%)
Ciências Sociais	40(29%)	40(67%)	80(40%)
Total	140(100%)	60(100%)	200(100%)

Fonte: Morettin e Bussab (2009).

Verifique se a criação de determinado tipo de cooperativa está associada com algum fator regional:

Tabela: Cooperativas autorizadas a funcionar por tipo e estado, junho de 1974.

Estado	Tipos de Cooperativa				Total
	Consumidor	Produtor	Escola	Outras	
São Paulo	214(33%)	237(37%)	78(12%)	119(18%)	648(100%)
Paraná	51(17%)	102(34%)	126(42%)	22(7%)	301(100%)
Rio G. do Sul	111(18%)	304(51%)	139(23%)	48(8%)	602(100%)
Total	376(24%)	643(42%)	343(22%)	189(12%)	1.551(100%)

Tabela: Valores esperados ao assumir independência entre as variáveis.

Estado	Tipos de Cooperativa				Total
	Consumidor	Produtor	Escola	Outras	
São Paulo	157(24%)	269(42%)	143(22%)	79(12%)	648(100%)
Paraná	73(24%)	124(42%)	67(22%)	37(12%)	301(100%)
Rio G. do Sul	146(24%)	250(42%)	133(22%)	73(12%)	602(100%)
Total	376(24%)	643(42%)	343(22%)	189(12%)	1.551(100%)

Tabela: Anos de Serviço (X) versus
N$^{\underline{o}}$ 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

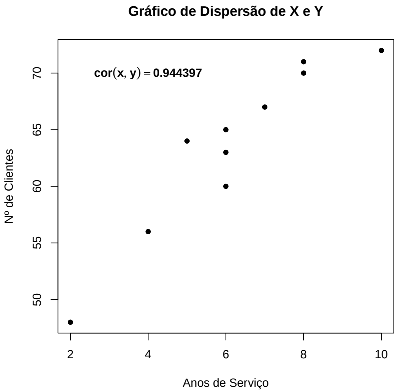
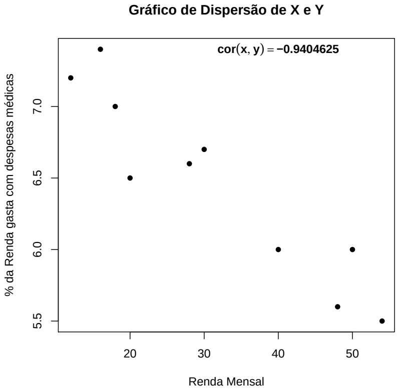


Tabela: Renda bruta mensal (X) e porcentagem da renda gasta em saúde (Y).

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



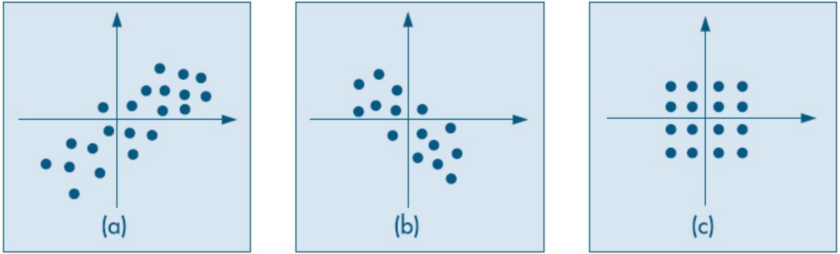


Figura: Morettin e Bussab (2009)

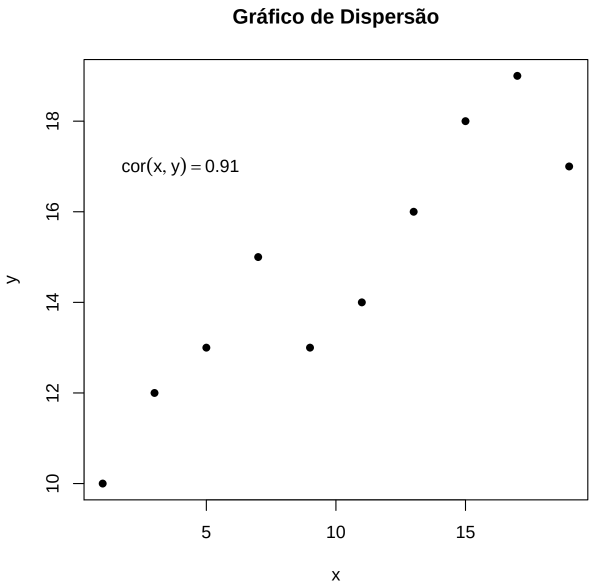


Gráfico de Dispersão
cor(x, y) = 0.91
y
x
10
12
14
16
18
5
10
15

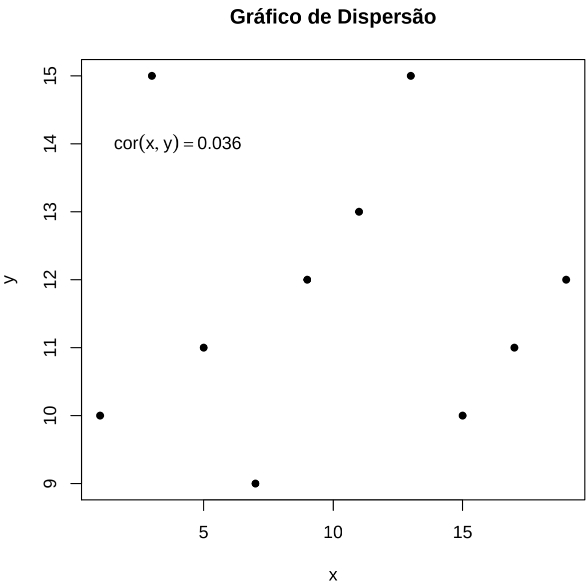


Gráfico de Dispersão
cor(x, y) = 0.036
y
15
14
13
12
11
10
9
5
10
15
x

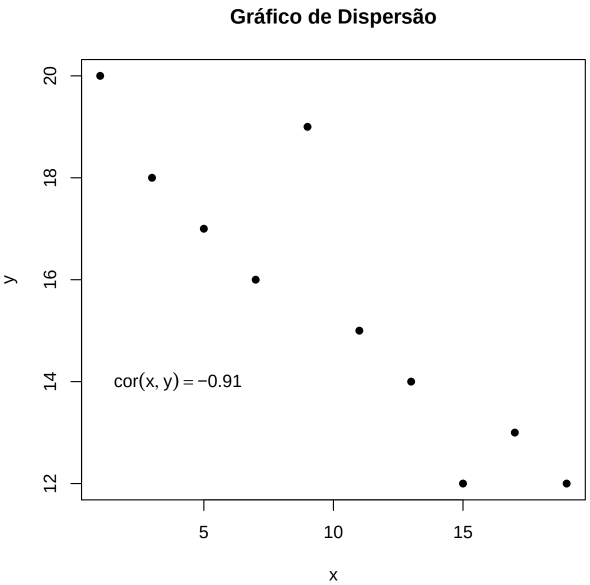


Gráfico de Dispersão
cor(x, y) = −0.91
y
x
20
18
16
14
12
5
10
15

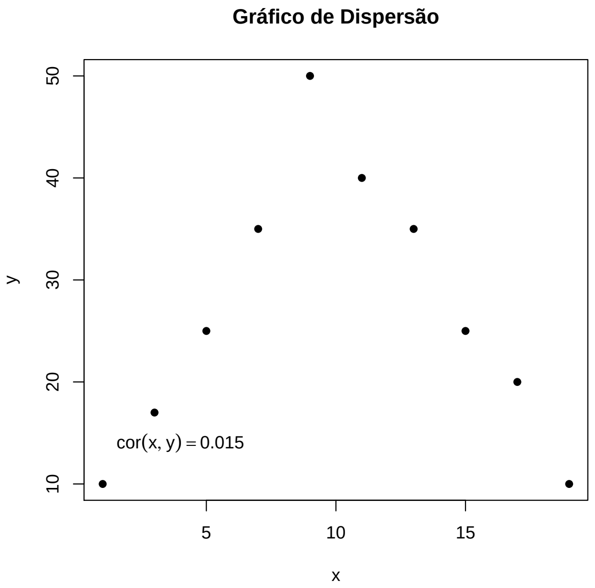


Gráfico de Dispersão
cor(x, y) = 0.015
x
y
5
10
15
10
20
30
40
50

Tabela: Cálculo do coeficiente de correlação.

Agente	Anos	Clientes	$x - \bar{x}$	$y - \bar{y}$	$\frac{x - \bar{x}}{dp(x)} = z_x$	$\frac{y - \bar{y}}{dp(y)} = z_y$	$z_x.z_y$
A	2	48	-3.7	-8.5	-1.54	-1.05	1.617
B	3	50	-2.7	-6.5	-1.12	-0.8	0.896
C	4	56	-1.7	-0.5	-0.71	-0.06	0.043
D	5	52	-0.7	-4.5	-0.29	-0.55	0.160
E	4	43	-1.7	-13.5	-0.71	-1.66	1.179
F	6	60	0.3	3.5	0.12	0.43	0.052
G	7	62	1.3	5.5	0.54	0.68	0.367
H	8	58	2.3	1.5	0.95	0.18	0.171
I	8	64	2.3	7.5	0.95	0.92	0.874
J	10	72	4.3	15.5	1.78	1.91	3.400
Total	57	565	0	0			8.759

$$Cor = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{dp(X)}\right)\left(\frac{y_i - \bar{y}}{dp(Y)}\right) = \frac{8.759}{10} = 0.8759$$

Não é difícil provar que o coeficiente de correlação satisfaz:

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

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:

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

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:

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

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

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

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:

- 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.

Associação entre Variáveis Qualitativas e Quantitativas

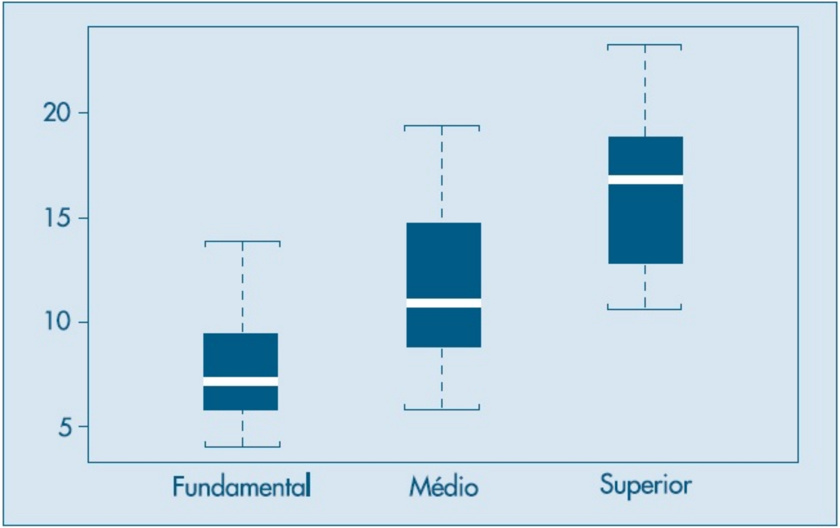


Figura: Box plots de salário segundo grau de instrução. Morettin e Bussab

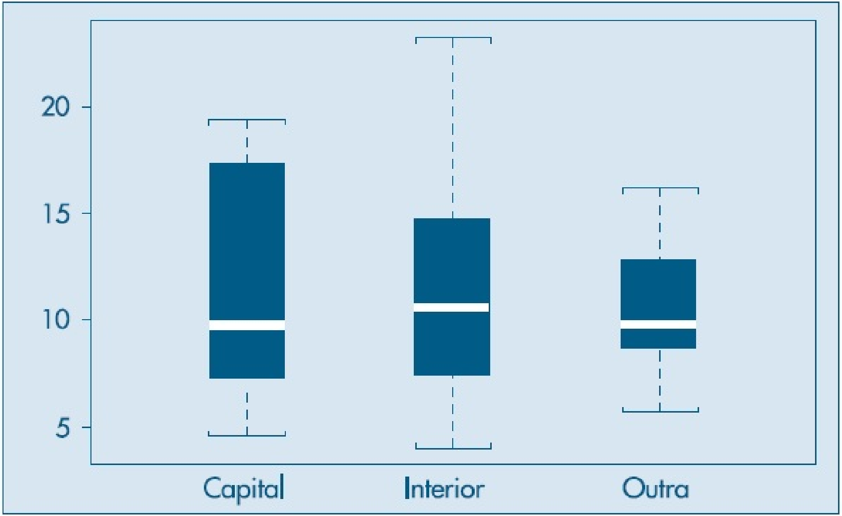


Figura: Box plots de salário segundo região de procedência. Morettin e Bussab (2009)

Referências Bibliográficas

P. Morettin e W. Bussab. *Estatística básica*. Editora Saraiva, São Paulo, 6 edition, 2009.