# 1 Aspectos da análise multivariada

## 1.1. Introdução

Nos trabalhos científicos, o problema de se inferir, a partir de dados mensurados pelo pesquisador, sobre os processos ou fenômenos físicos, biológicos ou sociais, que não se pode diretamente observar, é uma realidade constante. A pesquisa científica se constitui num processo interativo de aprendizado. Para explicação de um fenômeno, o pesquisador em geral coleta e analisa dados de acordo com uma hipótese. Por outro lado, a análise destes mesmos dados coletados de amostragem ou experimentação geralmente sugere modificações da explicação do fenômeno, além disso, devido a complexidade destes fenômenos, o pesquisador deve coletar observações de diferentes variáveis. Neste contexto, a inferência estatística é realizada de acordo com o paradigma hipotético-dedutivo (Bock, 1975).

Devido aos fenômenos serem estudados a partir de dados coletados ou mensurados em muitas variáveis, os métodos estatísticos delineados para obter informações a partir destes conjuntos de informações, são denominados de métodos de análises multivariados. A necessidade de compreensão das relações entre as diversas variáveis faz com que as análises multivariadas sejam um

assunto complexo ou até mesmo difícil. O objetivo do presente material é apresentar a utilidade das técnicas multivariada de uma forma clara, usando exemplos ilustrativos e evitando o máximo de possível de cálculo.

Sendo assim, os objetivos gerais, para os quais a análise multivariada conduz são:

a. redução de dados ou simplificação estrutural: o fenômeno sob estudo é representado da maneira mais simples possível, sem sacrificar informações valiosas e tornando as interpretações mais simples
b. ordenação e agrupamento: agrupamento de objetos (tratamentos) ou variáveis similares, baseados em dados amostrais ou experimentais.
c. investigação da dependência entre variáveis: estudos das relações estruturais entre variáveis muitas vezes é de interesse do pesquisador
d. predição: relações entre variáveis devem ser determinadas para o propósito de predição de uma ou mais variável com base na observação de outras variáveis.
e. construção e teste de hipóteses.

Os modelos multivariados possuem em geral, um propósito através do qual o pesquisador pode testar ou inferir a respeito de uma hipótese sobre um determinado fenômeno. No entanto a sua utilização adequada depende do bom conhecimento das técnicas e das suas limitações. A frase utilizada por Marriott (1974) descreve bem este fato: “Não há mágica com os métodos numéricos, e que apesar de serem uma importante ferramenta para análise e interpretação de dados, não devem ser utilizados como máquinas automáticas de encher lingüiça, transformando massas numéricas em pacotes de fatos científicos”.

## 1.2. Aplicação de técnicas multivariadas

As técnicas estatísticas constituem se uma parte integral da pesquisa científica e em particular as técnicas multivariadas tem sido regularmente aplicada em várias investigações científicas nas áreas de biologia, física, sociologia e ciências médicas. Parece, neste instante, ser apropriado descrever as situações em que as técnicas multivariadas têm um grande valor.

### Medicina

Nos estudos onde as reações de pacientes a um determinado tratamento são mensuradas em algumas variáveis e possuem difícil diagnóstico, as técnicas multivariadas podem ser usadas para construir uma medida de resposta simples ao tratamento, na qual é preservada a maior parte da informação da amostra e das múltiplas variáveis respostas. Em outras situações as técnicas multivariadas podem ser usadas também quando a classificação de um paciente, baseada nos sintomas medidos em algumas variáveis, é difícil de ser realizada. Neste caso, uma técnica multivariada de classificação, em que se cria uma função que pode ser usada para separar as pessoas doentes das não doentes, pode ser implementada.

## Sociologia

Em alguns estudos o inter-relacionamento e o agrupamento de indivíduos, cidades ou estados em grupos homogêneos em relação à mobilidade, número de estrangeiros nascidos e de segunda geração em determinado país é necessária em alguns estudos sociológicos. As técnicas de análise multivariada, conhecidas como análise de agrupamento (Cluster analysis), pode ser empregada com esta finalidade.

## Biologia

No melhoramento de plantas é necessário, após o final de uma geração, selecionar aquelas plantas que serão os genitores da próxima geração. a seleção deve ser realizada de maneira que a próxima geração seja melhorada em relação à resposta média de uma série de características da geração anterior. O objetivo do melhorista consiste em maximizar o ganho genético em um espaço mínimo de tempo. As análises multivariadas podem ser usadas para converter uma série de características para um índice, na qual a seleção e escolha dos pais possam ser feitas.

Em algumas situações se deseja a separação de algumas espécies, e as técnicas multivariadas tem sido utilizadas com esta finalidade. Uma função é construída, e os seus valores são usados para esta separação.

## 1.3. Organização de dados

Através deste material pretende-se tratar das análises realizadas em muitas características ou variáveis. Essas medidas, muitas vezes chamadas de dados, devem ser organizados e apresentados em várias formas. Por exemplo, a utilização de gráficos e arranjos tabulares são importantes auxiliares nas análises de dados. Por outro lado, números que resumem, ou seja, que descrevem quantitativamente certas características, são essenciais para a interpretação de os dados amostrais ou experimentais.

## Arranjos

Os dados multivariados são provenientes de uma pesquisa em determinada área em que são selecionadas $p \geq 1$ variáveis ou características para serem mensuradas. As medidas são tomadas em cada unidade da amostra ou do experimento. A representação destes dados é feita com a notação $x_{jk}$ para indicar um valor particular da j-ésima unidade amostral ou experimental e da k-ésima

variável mensurada. Conseqüente, estas medidas de p variáveis em n unidades amostrais ou experimentais, podem ser representadas conforme o arranjo apresentado na Tabela 1.1.

Tabela 1.1. Representação de dados através da notação $x_{jk}$ para indicar um valor particular da k-ésima variável mensurada na j-ésima unidade amostral ou experimental.

| | Variáveis | | | |
|---|---|---|---|---|
| Unidades amostrais ou experimentais | 1 | 2 ... | k ... | p |
| 1 | $X_{11}$ | $X_{12}$... | $X_{1k}$... | $X_{1p}$ |
| 2 | $X_{21}$ | $X_{22}$... | $X_{2k}$... | $X_{2p}$ |
| . | . | . | . | . |
| . | . | . | . | . |
| . | . | . | . | . |
| j | $X_{j1}$ | $X_{j2}$... | $X_{jk}$... | $X_{jp}$ |
| . | . | . | . | . |
| . | . | . | . | . |
| . | . | . | . | . |
| n | $X_{n1}$ | $X_{n2}$... | $X_{nk}$... | $X_{np}$ |

Estes valores, apresentados na Tabela 1.1, podem ser representados em um arranjo retangular, denominado de **X**, com n linhas e p colunas, da seguinte forma:

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1k} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2k} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{j1} & x_{j2} & \cdots & x_{jk} & \cdots & x_{jp} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nk} & \cdots & x_{np} \end{bmatrix}$$

**Exemplo 1.1**

Uma seleção de 4 firmas de ração de Minas Gerais foi obtida para avaliar a venda de rações. Cada observação bivariada forneceu a quantidade de sacos de ração vendidos e a quantidade de reais de cada venda. Os dados obtidos na forma tabular são:

| | | | | |
|---|---|---|---|---|
| Variável 1 (Reais/venda) | 80 | 120 | 90 | 110 |
| Variável 2 (número de sacos de ração vendidos) | 10 | 12 | 6 | 8 |

Usando a notação proposta anteriormente, tem-se:

$X_{11}=80$ $X_{21}=120$ $X_{31}=90$ $X_{41}=110$ $X_{12}=10$ $X_{22}=12$ $X_{32}=6$ $X_{42}=8$

E a matriz X dos dados é:

$$X = \begin{bmatrix} 80 & 10 \\ 120 & 12 \\ 90 & 6 \\ 110 & 8 \end{bmatrix}$$

A organização dos dados em arranjos facilita a exposição e permite que os cálculos sejam efetuados de uma forma ordenada e eficiente. Os ganhos na eficiência são: (1) descrição dos cálculos como operações com matrizes e vetores; e (2) sua fácil implementação em computadores.

## ESTATÍSTICAS DESCRITIVAS

Grandes conjuntos de dados possuem um sério obstáculo para qualquer tentativa de extração de informações visuais pertinentes aos mesmos. muitas das informações contidas nos dados podem ser obtidas por cálculo de certos números, conhecidos como estatísticas descritivas. Por exemplo, a média aritmética ou média amostral, é uma estatística descritiva que fornece informação de posição, isto é, representa um valor central para o conjunto de dados. Como um outro exemplo, a média das distâncias ao quadrados de cada dado em relação à média, fornece uma medida de dispersão, ou variabilidade.

Às estatísticas descritivas que mensuram posição, variação e associação linear são enfatizadas. As descrições formais destas medidas estão apresentadas a seguir.

A média amostral, simbolizada por $\bar{x}$, é dada por:

$$\bar{X}_k = \frac{1}{n}\sum_{j=1}^{n} X_{jk} \qquad \textbf{k=1, 2, ..., p} \qquad (1.1)$$

Uma medida de variação é fornecida pela variância amostral, definida para as n observações de i-ésima variável por:

$$S_k^2 = S_{kk} = \frac{1}{n-1}\sum_{j=1}^{n}\left(X_{jk} - \bar{X}_k\right)^2 \quad \textbf{k = 1, 2, ..., p} \tag{1.2}$$

A raiz quadrada da variância amostral, $\sqrt{S_{kk}}$, é conhecido como desvio padrão amostral. Esta medida de variação está na mesma unidade de medida das observações.

Uma medida de associação entre as observações de duas variáveis, variáveis k e k', é dada pela covariância amostral:

$$S_{kk'} = \frac{1}{n-1}\sum_{j=1}^{n}\left(X_{jk} - \bar{X}_k\right)\left(X_{jk'} - \bar{X}_{k'}\right) \quad \textbf{k, k'=1,2, ..., p} \tag{1.3}$$

Se grandes valores de uma variável são observados em conjunto com grandes valores da outra variável, e os pequenos valores também ocorrem juntos, $S_{kk'}$ será positiva. Se grandes valores de uma variável ocorrem com pequenos valores da outra, $S_{kk'}$ será negativa. Se não há associação entre os valores das duas variáveis, $S_{kk'}$ será aproximadamente zero. Quando k=k', a covariância reduz-se a variância amostral. Além disso, $S_{kk'} = S_{k'k}$, para todo k e k'.

A última estatística descritiva a ser considerada aqui é o coeficiente de correlação amostral. Esta medida de associação linear entre duas variáveis

não depende da unidade de mensuração. O coeficiente de correlação amostral para k-ésima e k'-ésima variável, é definido por:

$$r_{kk'} = \frac{S_{kk'}}{\sqrt{S_{kk}}\sqrt{S_{k'k'}}} = \frac{\sum_{j=1}^{n}\left(X_{jk} - \bar{X}_k\right)\left(X_{jk'} - \bar{X}_{k'}\right)}{\sqrt{\sum_{j=1}^{n}\left(X_{jk} - \bar{X}_k\right)^2}\sqrt{\sum_{j=1}^{n}\left(X_{jk'} - \bar{X}_{k'}\right)^2}} \qquad (1.4)$$

Verifica-se que $r_{kk'}=r_{k'k}$ para todo k e k'. O coeficiente de correlação amostral é a versão estandardizada da covariância amostral, onde o produto das raízes das variâncias das amostras fornece a estandardização.

O coeficiente de correlação amostral pode ser considerado como uma covariância amostral. Suponha que os valores $X_{jk}$ e $X_{jk'}$ sejam substituídos pelos valores padronizados, $(X_{jk}-\bar{X}_k)/\sqrt{S_{kk}}$ e $(X_{jk'}-\bar{X}_{k'})/\sqrt{S_{k'k'}}$. Esses valores padronizados são expressos sem escalas de medidas (adimensionais), pois são centrados em zero e expressos em unidades de desvio padrão. O coeficiente de correlação amostral é justamente a covariância amostral das observações estandardizadas.

A correlação amostral (r), em resumo, tem as seguintes propriedades:

1. Os valores de r devem ficar compreendidos entre -1 e 1;
2. Se r = 0, implica em inexistência de associação linear entre as variáveis. Por outro lado, o sinal de r, indica a direção da associação: se r < 0 há uma tendência de um dos valores do par ser maior que sua média, quando o outro for menor do que a sua média, e r > 0 indica que quando um valor do par for

grande o outro também o será, além de ambos valores tender a serem pequenos juntos;

3. Os valores de $r_{kk'}$ não se alteram com a alteração da escala de uma das variáveis.

As estatísticas $S_{kk'}$ e $r_{kk'}$, em geral, não necessariamente refletem todo o conhecimento de associação entre duas variáveis. Associações não lineares existem, as quais, não podem ser reveladas por estas estatísticas descritivas. Por outro lado, estas estatísticas são muito sensíveis a observações discrepantes (outliers).

Além destas, outras estatísticas como a soma de quadrados de desvios em relação à média ($W_{kk}$) e a soma de produtos de desvios ($W_{kk'}$), são muitas vezes de interesse. Essas estão apresentadas a seguir:

$$W_{kk} = \sum_{j=1}^{n} (X_{jk} - \bar{X}_k)^2$$

$$W_{kk'} = \sum_{j=1}^{n} (X_{jk} - \bar{X}_k)(X_{jk'} - \bar{X}_{k'})$$

As estatísticas descritivas multivariadas calculadas de n observações em p variáveis podem ser organizadas em arranjos.

## Médias da amostra

$$\underset{\sim}{\overline{X}} = \begin{bmatrix} \overline{X}_1 \\ \overline{X}_2 \\ \vdots \\ \overline{X}_p \end{bmatrix}$$

## Matriz de covariância amostral

$$\boldsymbol{S} = \begin{bmatrix} \boldsymbol{s}_{11} & \boldsymbol{s}_{12} & \cdots & \boldsymbol{s}_{1p} \\ \boldsymbol{s}_{21} & \boldsymbol{s}_{22} & \cdots & \boldsymbol{s}_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{s}_{p1} & \boldsymbol{s}_{p2} & \cdots & \boldsymbol{s}_{pp} \end{bmatrix}$$

## Matriz de correlações amostral

$$\boldsymbol{R} = \begin{bmatrix} \mathbf{1} & \boldsymbol{r}_{12} & \cdots & \boldsymbol{r}_{1p} \\ \boldsymbol{r}_{21} & \mathbf{1} & \cdots & \boldsymbol{r}_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{r}_{p1} & \boldsymbol{r}_{p2} & \cdots & \mathbf{1} \end{bmatrix}$$

### Exemplo 1.2

Considerando os dados introduzidos no exemplo 1.1, encontrar as o vetor de médias $\underset{\sim}{\overline{X}}$ e as matrizes S e R. Neste exemplo, cada firma de ração, representa uma das observações multivariadas, com p = 2 variáveis (valor da venda em reais e número de sacos de rações vendidas).

As médias amostral são:

$$\bar{X}_1 = \frac{1}{4}\sum_{j=1}^{4} X_{j1} = \frac{1}{4}(80+120+90+110) = 100$$

$$\bar{X}_2 = \frac{1}{4}\sum_{j=1}^{4} X_{j2} = \frac{1}{4}(10+12+6+8) = 9$$

$$\underset{\sim}{\bar{X}} = \begin{bmatrix} \bar{X}_1 \\ \bar{X}_2 \end{bmatrix} = \begin{bmatrix} 100 \\ 9 \end{bmatrix}$$

A matriz de covariância amostral é:

$S_{11}=[(80-100)^2+(120-100)^2+(90-100)^2+(110-100)^2]/3 = 333,333$

$S_{22}=[(10-9)^2+(12-9)^2+(6-9)^2+(8-9)^2]/3 = 6,667$

$S_{12}=[(80-100)(10-9)+(120-100)(12-9)+(90-100)\ (6-9)+(110-100)(8-9)]/3 = 20,000$

$S_{21}=S_{12}=20,000$, e

$$S = \begin{bmatrix} 333,333 & 20,000 \\ 20,000 & 6,667 \end{bmatrix}$$

A correlação amostral é:

$$r_{12} = \frac{20}{\sqrt{33,333}\sqrt{6,667}} = 0,4243$$

$r_{21}=r_{12}=0,4243$

Portanto,

$$R = \begin{bmatrix} 1,0000 & 0,4243 \\ 0,4243 & 1,0000 \end{bmatrix}$$

## 1.4. Distâncias

A maioria das técnicas multivariadas é baseada no simples conceito de distância, por mais formidável que isso possa parecer. O conceito de distância euclidiana deve ser familiar para a maioria dos estudantes. Se for considerado um ponto $P=(x_1, x_2)$ no plano cartesiano, a distância deste ponto P da origem $O=(0, 0)$, definida por d(O,P), é dada pelo teorema de Pitágoras por:

$$\boldsymbol{d}(\boldsymbol{O},\boldsymbol{P}) = \sqrt{\boldsymbol{x}_1^2 + \boldsymbol{x}_2^2} \qquad (1.5)$$

Esta situação é ilustrada na Figura 1.1. Em geral, se o ponto P tem p coordenadas, de tal forma que P=($x_1$, $x_2$, ... $x_p$), a distância de P da origem O=(0, 0, ..., 0), pode ser generalizada por:

$$\boldsymbol{d(O,P) = \sqrt{x_1^2 + x_2^2 + ... + x_p^2}} \qquad (1.6)$$

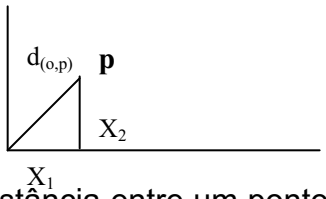


Figura 1.1. Distância entre um ponto P=($x_1$, $x_2$) e a origem O=(0, 0), fornecida pelo teorema de Pitágoras.

Todos os pontos ($x_1$, $x_2$, .., $x_p$) que contém uma distância ao quadrado, denominada $c^2$, da origem, satisfaz a equação:

$$\boldsymbol{d^2(O,P) = x_1^2 + x_2^2 + ... + x_p^2 = c^2} \qquad (1.7)$$

A expressão em (1.7) representa a equação de uma hiperesfera (um círculo se p=2), e os pontos eqüidistantes da origem pertencem a mesma. A distância de um ponto P a um ponto arbitrário Q, com coordenadas P=($x_1$, $x_2$, ... $x_p$) e Q=($y_1$, $y_2$, ... $y_p$) é dada por:

$$\boldsymbol{d(P,Q) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + ... + (x_p - y_p)^2}} \qquad (1.8)$$

A distância euclidiana é insatisfatória para muitas situações estatísticas. Isso ocorre, devido a contribuição de cada coordenada ter o mesmo peso para o cálculo da distância. Quando estas coordenadas representam medidas que estão sujeitas a flutuações aleatórias de diferentes magnitudes, é muitas vezes desejável ponderar as coordenadas com grande variabilidade por menores pesos do que aquelas com baixa variabilidade. Isto sugere o uso de uma nova medida de distância.

Será apresentada a seguir uma distância que considere as diferenças de variação e a presença de correlação. Devido a esta escolha de a distância depender das variâncias e das covariâncias amostrais, a partir deste instante, será utilizado o termo “distância estatística” para distinguí-la da distância euclidiana.

Para iniciar, será considerada a construção de uma distância entre um ponto P, com p coordenadas, da origem. O argumento que pode ser usado refere-se ao fato de que as coordenadas do ponto P podem variar produzindo diferentes posições para os pontos. Para ilustrar, suponha que se tenha n pares de medidas em duas variáveis ($x_1$ e $x_2$) e que as medidas de $x_1$ variam independentemente das mensurações em $x_2$. O significado de independente neste ponto pode ser dado pelo fato de que os valores de $x_1$ não podem ser preditos com nenhuma acurácia a partir dos valores de $x_2$ e vice-versa. Em adição, é assumido que as observações de $x_1$ possuem maior variabilidade que as de $x_2$. Uma ilustração desta situação está apresentada na Figura 1.2.

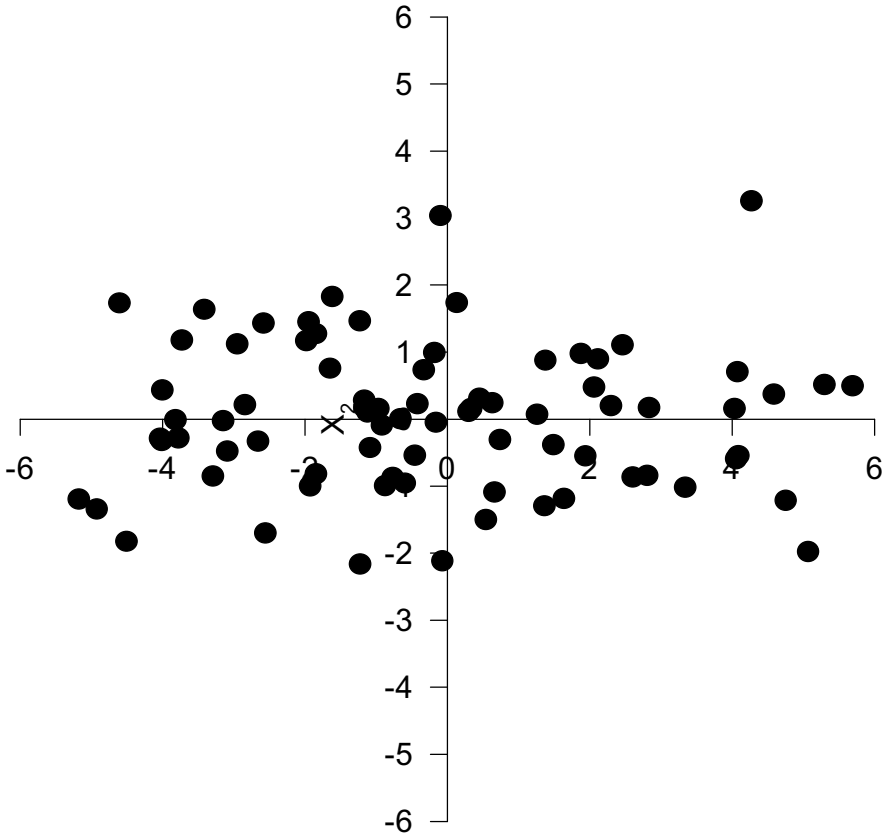


Figura 1.2. Diagrama de dispersão, mostrando a maior variabilidade na direção de $x_1$ do que na direção de $x_2$.

Observando a Figura 1.2, verifica-se que não é surpreendente encontrar desvios na direção de $x_1$ que se afastem da origem consideravelmente, o que não ocorre na direção de $x_2$. Parece ser razoável, então, ponderar $x_2$ com mais peso do que $x_1$ para um mesmo valor, quando as distâncias da origem forem calculadas.

Um modo de fazer isso é dividir cada coordenada pelo desvio padrão amostral. Após a divisão, tem-se as coordenadas estandardizadas $x_1^* = x_1 / \sqrt{s_{11}}$ e $x_2^* = x_2 / \sqrt{s_{22}}$. Após eliminar as diferenças de variabilidade das variáveis (coordenadas), determina-se à distância usando a fórmula euclidiana padrão:

$$d(O,P) = \sqrt{(x_1^*)^2 + (x_2^*)^2} = \sqrt{\frac{x_1^2}{s_{11}} + \frac{x_2^2}{s_{22}}} \qquad ]= \qquad (1.9)$$

Usando a equação (1.9) todos os pontos tendo como coordenadas ($x_1$, $x_2$) e com distância quadrada ($c^2$) da origem devem satisfazer:

$$\frac{x_1^2}{s_{11}} + \frac{x_2^2}{s_{22}} = c^2 \qquad (1.10)$$

A expressão (1.10) é a equação de uma elipse, cujos maiores e menores eixos coincidem com os eixos das coordenadas. A Figura 1.3 mostra o caso geral para p=2 coordenadas.

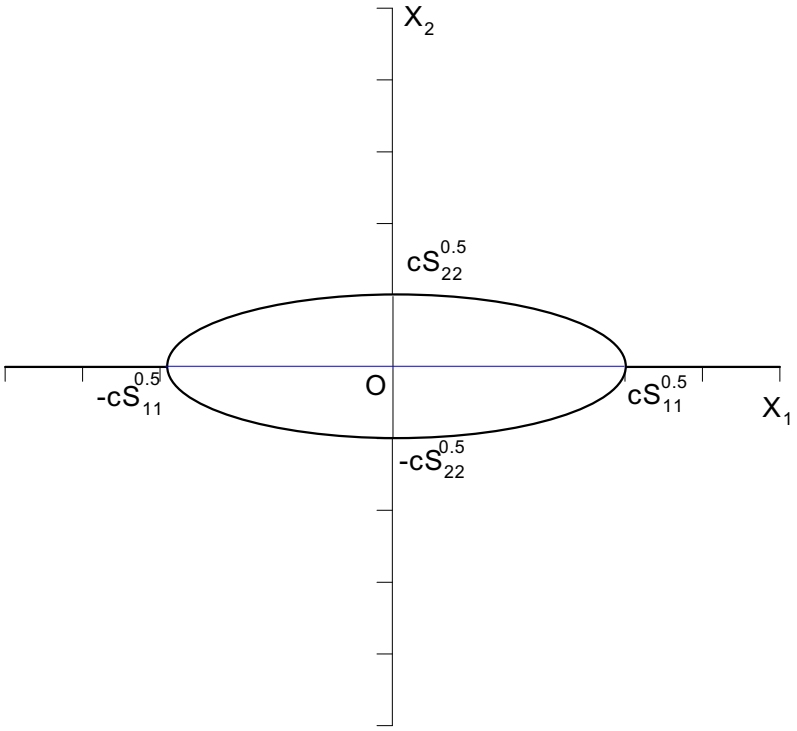


Figura 1.3. Elipse de uma distância estatística quadrática $d^2(O,P)= \frac{x_1^2}{S_{11}} + \frac{x_2^2}{S_{22}} = c^2$.

## Exemplo 1.2

Um conjunto de pares ($x_1$, $x_2$) de duas variáveis forneceu $\overline{X}_1 = \overline{X}_2 = 1$, $S_{11}=9$ e $S_{22}=1$. Supõem-se que as observações de $x_1$ são independentes de $x_2$. A distância quadrática de um ponto arbitrário (P) da origem, uma vez que as variâncias da amostra não são iguais, é dada por:

$$d^2(O,P) = \frac{x_1^2}{9} + \frac{x_2^2}{1}$$

Todos os pontos ($x_1$, $x_2$) que possuem distâncias quadrada da origem igual a 1, satisfaz a equação:

$$\frac{x_1^2}{9} + \frac{x_2^2}{1} = 1 \qquad (1.11)$$

As coordenadas de alguns pontos com distância quadrática unitária da origem foram apresentadas na Tabela 1.2.

Tabela 1.2. Coordenadas de alguns pontos com distância quadrática unitária da origem.

| Coordenadas ($x_1$, $x_2$) | Distância ao quadrado |
|---|---|
| ( 0, 1) | $\frac{0^2}{9} + \frac{1^2}{1} = 1$ |
| ( 0,-1) | $\frac{0^2}{9} + \frac{(-1)^2}{1} = 1$ |
| ( 3, 0) | $\frac{3^2}{9} + \frac{0^2}{1} = 1$ |
| (-3, 0) | $\frac{(-3)^2}{9} + \frac{0^2}{1} = 1$ |

O gráfico da equação (1.11) é uma elipse centrada na origem (0,0), cujo maior eixo é o da direção de $x_1$ e o menor da direção de $x_2$. A metade do maior eixo (semi-eixo maior) é $c\sqrt{s_{11}} = 3$ e do menor $c\sqrt{s_{22}} = 1$. A elipse de distância quadrática unitária foi plotada na Figura 1.4.

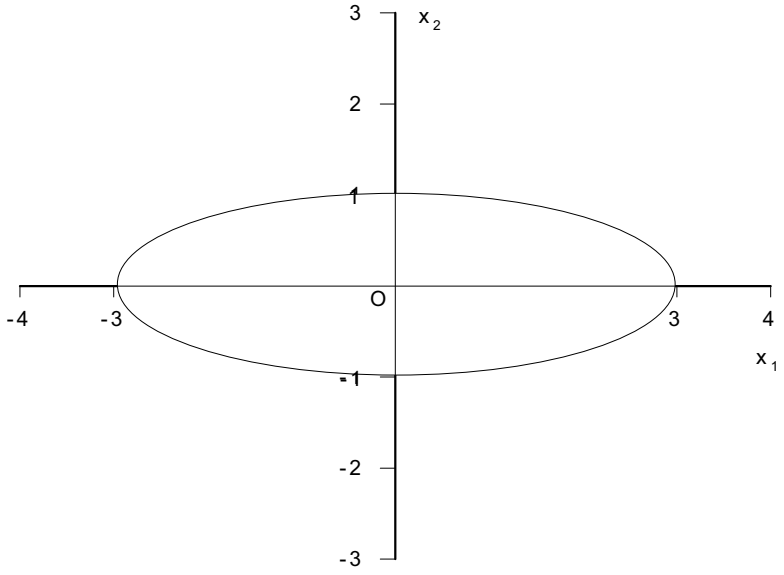


Figura 1.4. Elipse de distância unitária quadrática da origem obtida a partir da equação 1.11.

A expressão (1.9) pode ser generalizada para o cálculo da distância entre pontos P e Q, cujas coordenadas variam, mutuamente independentemente uma da outra. O caso mais geral, em que a hipótese de independência não é satisfeita, será abordado futuramente.

$$\boldsymbol{d}(\boldsymbol{P},\boldsymbol{Q}) = \sqrt{\frac{(\boldsymbol{x}_1 - \boldsymbol{y}_1)^2}{\boldsymbol{S}_{11}} + \frac{(\boldsymbol{x}_2 - \boldsymbol{y}_2)^2}{\boldsymbol{S}_{22}} + \cdots + \frac{(\boldsymbol{x}_p - \boldsymbol{y}_p)^2}{\boldsymbol{S}_{pp}}} \qquad (1.12)$$

Todos os pontos (P) situados a uma distância quadrática constante de Q, pertencem a uma hiperelipsóide centrada em Q, cujos maiores e menores eixos são paralelos aos eixos das coordenadas.

O programa SAS, apresentado a seguir, contem os códigos necessários para a obtenção das principais estatísticas descritivas multivariadas apresentadas nesse capítulo. O programa contém códigos matriciais e será abordado com mais detalhe nos próximos capítulos. Os dados do exemplo 1.1 são utilizados para a ilustração.

```
Proc IML;
  X={  80   10,
      120   12,
       90    6,
      110    8};
  Print X;
  n=nrow(X);p=ncol(X);
  Xbar=x`*j(n,1,1)/n;
  Print Xbar;
  q=i(n)-(1/n)*j(n,n,1);
  print q;
  S=(1/(n-1))*X`*q*X;
  W=(n-1)*S;
  print S W;
  V=diag(S);
  Vroot=half(V);
  IVroot=inv(Vroot);
  R=Ivroot*S*Ivroot;
  Print V Vroot IVroot;
  Print R;
Quit;
```

Foi motivado nesse capítulo o estudo das análises multivariadas e tentou-se fornecer alguns rudimentares, mas importantes, métodos de organizar e resumir os dados. Em adição, o conceito geral de distância foi apresentado, e será abordado e generalizado nos próximos capítulos.

## 1.5. Exercícios

■ Considere as amostras com 8 observações e 3 variáveis apresentadas a seguir:

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| $x_1$ | 3 | 5 | 6 | 4 | 8 | 9 | 6 | 7 |
| $x_2$ | 6 | 11 | 11 | 9 | 15 | 16 | 10 | 12 |
| $x_3$ | 14 | 9 | 9 | 13 | 2 | 2 | 9 | 5 |

a) Construa o gráfico de dispersão dos pontos das variáveis $x_1$ e $x_2$, $x_1$ e $x_3$, $x_2$ e $x_3$. Comente sobre sua aparência.

b) Calcule: $\overline{\boldsymbol{x}}$, S e R e interprete os valores em R.

c) Calcule a distância euclidiana dada em (1.8) de um ponto P=( $x_1$, $x_2$, $x_3$)=(5, 12, 8) em relação a origem e em relação a $\overline{\boldsymbol{x}}$.

d) Calcule as mesmas distâncias do item c, usando (1.12).

## 1.6. Referências

BOCK, R.D. **Multivariate statistical methods in behavioral research.** McGrawn Hill, 1975.

CLEVELAND, W.S.; RELLES, D.A. Clustering by identification with special application to two way tables of counts. **Journal of American Statistical Association. v.70, n.351, 1975**. 626-630p.

JOHNSON, R.A.; WICHERN, D.W. **Applied multivariate statistical analysis.** 4th edition. Prentice Hall, New Jersey, 1998. 816p.

MARRIOTT, F.H.C. **The interpretation of multiple observations.** London, Academic Press, 1974.