

TÉCNICAS MULTIVARIADAS EXPLORATÓRIAS

TEORIAS E APLICAÇÕES NO *SOFTWARE* ***STATISTICA®***

editoraufsm

LORENA VICINI
ADRIANO MENDONÇA SOUZA
FIDEL ERNESTO CASTRO MORALES
FRANCISCA MENDONÇA SOUZA



TÉCNICAS MULTIVARIADAS EXPLORATÓRIAS

TEORIAS E APLICAÇÕES NO *SOFTWARE* ***STATISTICA®***

LORENA VICINI
ADRIANO MENDONÇA SOUZA
FIDEL ERNESTO CASTRO MORALES
FRANCISCA MENDONÇA SOUZA

editoraufsm
Santa Maria, RS - 2018

UNIVERSIDADE FEDERAL DE SANTA MARIA

Reitor	Paulo Afonso Burmann
Vice-Reitor	Luciano Schuch
Diretor da Editora	Daniel Arruda Coronel
Conselho editorial	Adão Pilar Damasceno
	Adriano Mendonça Souza
	Alisson Vicente Zarnot
	Carlise Scalamato Duarte
	Daniel Arruda Coronel (Presidente)
	Gil Roberto Costa Negreiros
	Glauber Rodrigues de Quadros
	José Renato Ferraz da Silveira
	Lucas Reinehr
	Marcelo Battesini
	Marilda Oliveira de Oliveira
	Melina da Silva Mota
	Rosmari Horner
Revisão de texto	Maicon Antonio Paim
	Tagiane Mai
	Matheus Von Ende Schwertner (Bolsista)
Capa	Douglas Mastela Dal Forno (Bolsista)
Diagramação	Gilberto de Moraes Jr.

© 2018, Lorena Vicini, Adriano Mendonça Souza,
Fidel Ernesto Castro Morales e Francisca Mendonça Souza

T255 Técnicas multivariadas exploratórias : teorias e aplicações no Software Statistica / Lorena Vicini ... [et al.]. – Santa Maria, RS : Ed. UFSM, 2018.
240 p. : il. ; 23 cm

Inclui bibliografia

1. Estatística 2. Software Statistica I. Vicini, Lorena

CDU 004.4
519.22:004.4 ISBN: 978-85-7391-323-1

Ficha catalográfica elaborada por Alenir Goularte CRB – 10/990
Biblioteca Central – UFSM

Editora Associada à
Associação Brasileira das Editoras Universitárias

editoraufsm

Direitos reservados à:

Editora da Universidade Federal de Santa Maria
Prédio da Reitoria – Campus Universitário
Camobi – CEP 97105.900 – Santa Maria – RS
Fone: (55) 3220.8610/8115
editufsm@gmail.com | www.ufsm.br/editora

Aos nossos Pais.

SUMÁRIO

PREFÁCIO **7**

APRESENTAÇÃO **10**

1 INTRODUÇÃO **13**

2 CONCEITOS BÁSICOS **17**
2.1 Variável estatística 18
2.2 Escalas de medidas 19
2.3 Validade e confiabilidade 21

3 ANÁLISE DE AGRUPAMENTOS – AA **27**
3.1 Alguns coeficientes de medida de distância 33
3.2 Alguns métodos de encadeamento 39
3.3 Qual o melhor método de agrupamento? 53
3.4 Como interpretar o dendrograma 60
3.5 Aplicações em análise de agrupamentos 62

4 ANÁLISE DE COMPONENTES PRINCIPAIS – ACP **75**
4.1 Geração das componentes principais 78
4.2 Seleção e interpretação das componentes principais 84
4.3 Matriz de variância-covariância e matriz de correlação 87
4.4 Alguns conceitos úteis para relembrar 134

5 ANÁLISE FATORIAL – AF **136**
5.1 Medidas de adequação ao modelo fatorial 139
5.2 O modelo fatorial 142
5.3 Tipos de estimação dos fatores 144
5.4 Melhor visualização das variáveis – rotação de eixos 146
5.5 Como proceder à seleção do número de fatores e à validação do modelo . 147
5.6 Aplicação no *software* – AF 148

6 ANÁLISE FATORIAL DE CORRESPONDÊNCIA – AFC 167
6.1 Exemplificando a técnica 177
6.2 Como comparar perfis 188

7 ANÁLISE DISCRIMINANTE – AD 198
7.1 A função discriminante 201
7.2 Escore de corte 203
7.3 Aplicação 206

BIBLIOGRAFIA 216

PREFÁCIO

A disponibilidade de grandes volumes de informação estatística constitui um desafio central das sociedades modernas, não só no que diz respeito à forma e à velocidade de divulgação dessa informação, mas também no que diz respeito ao modo como os dados são tratados e analisados.

O rigor na recolha e no tratamento da informação estatística exige, cada vez mais, a utilização de recursos humanos e técnicos especializados que permitam analisar os dados de forma isenta e tecnicamente rigorosa. Para isso, muito têm contribuído os avanços no conhecimento das diversas técnicas e instrumentos estatísticos para recolha, tratamento e análise de dados estatísticos.

Ora, é precisamente no âmbito das técnicas multivariadas exploratórias que a presente obra se adscreve, mantendo, por um lado, o rigor teórico da apresentação e discussão dos modelos analisados e, por outro, fornecendo vários exemplos práticos com utilização do *software* estatístico *Statistica*. A obra debruça-se sobre cinco modelos temáticos, cada um deles apresentado num capítulo diferente: análise de agrupamentos, análise de componentes principais, análise fatorial, análise fatorial de correspondência e análise discriminante. Os conceitos são apresentados de forma simples, mas mantendo o rigor estatístico sempre que necessário, tendo em vista facilitar a compreensão dos seus conteúdos também por leitores menos familiarizados com a "dureza" da linguagem estatística.

Os autores da presente obra possuem larga experiência no ensino e na investigação de modelos e técnicas estatísticas em diversos níveis de ensino e dirigidos a públicos-alvo diferenciados no que concerne à aplicação empírica dessas técnicas. Reforçando o que acima se referiu, trata-se de uma obra muito bem fundamentada e descrita de forma didática, sem perder o rigor do tratamento estatístico sempre que, e somente quando, necessário. Os exemplos de aplicação com *software* são muito bem apresentados com um passo a passo claro e produtivo para o leitor.

A temática abordada neste livro é, sem dúvida, abrangente e atual, despertando, certamente, o interesse de professores, pesquisadores e estudantes que pretendam realizar uma análise conjunta de dados. Esta obra poderá ser adotada por alunos de graduação e pós-graduação nos cursos de Administração, Ciências Contábeis, Finanças e Economia, bem assim como em diversas outras áreas das Ciências Sociais.

Importa saudar os autores por mais esta iniciativa num mercado no qual, de algum modo, escasseiam as obras publicadas em língua portuguesa. A obra contribui, dessa forma, para a disseminação dessas técnicas no seio da comunidade científica dos países que integram a Comunidade dos Países de Língua Portuguesa.

Rui Menezes[1]

1 Professor catedrático do Departamento de Métodos Quantitativos do Instituto Superior de Ciências do Trabalho e da Empresa – Instituto Universitário de Lisboa – e investigador associado da Unidade de Investigação em Desenvolvimento Empresarial (UNIDE-IUL).

APRESENTAÇÃO

A estatística mostra-se, cada vez mais, como uma poderosa ferramenta para a análise e avaliação de dados, em várias áreas do conhecimento, sendo muitas vezes um tanto difícil para os profissionais trabalharem conceitos e elaborarem exemplos práticos, devido à limitação de materiais didáticos que expressem, com simplicidade e clareza, métodos e procedimentos da aplicação de certas técnicas multivariadas, que só passaram a ser utilizadas, em larga escala, a partir do advento dos computadores.

Embora a estatística multivariada tenha surgido por volta de 1901, apenas nos dias de hoje tornou-se possível desenvolver e aplicar essa técnica com uma maior facilidade, pois sem o auxílio de programas computacionais não seria possível desenvolver certos gráficos apresentados na análise, bem como realizar tão rápido e com a devida clareza a interpretação dos gráficos que possibilitam estudar o inter-relacionamento das *n* variáveis.

Verifica-se, no decorrer da pesquisa, que as técnicas de análise de agrupamentos e análise de componentes principais são técnicas matemáticas, com grande fundamentação na álgebra e na geometria, o que muitas vezes faz com que os estatísticos não as considerem como técnicas estatísticas. Por outro lado, figuram, quase sempre, em congressos nacionais e revistas especializadas que tratam de assuntos sobre estatística.

Já a análise fatorial pressupõe a existência de um modelo e permite que se façam inferências e que se cumpram algumas pressuposições básicas sobre as variáveis em análise, como a multinormalidade dos dados. Por fim, a técnica de análise discriminante é apresentada e possibilita que um indivíduo seja classificado com certo grau de certeza em determinado grupo.

Ao aplicar as técnicas multivariadas, há a necessidade de que exista uma estrutura de correlação entre as variáveis, pois, se estas não estiverem ligadas entre si, tem-se que utilizar uma análise univariada, uma vez que essa, se bem aplicada, é capaz de fornecer um nível muito bom de informação.

O uso de um *software* é indispensável para a execução e análise multivariada e existe uma diversidade deles, inclusive *softwares* livres. Esta obra traz, passo a passo, o desenvolvimento das técnicas de análise de agrupamentos, análise de componentes principais, análise fatorial e análise fatorial de correspondência aplicadas ao *software Statistica*.

Sugere-se ao leitor não familiarizado com álgebra matricial, conceitos de autovalores e autovetores que busque um material auxiliar para melhor entender esses conceitos.

Salientamos que este material é de responsabilidade dos autores e que quaisquer dúvidas ou sugestões devem ser encaminhadas para eles.

Os autores

1
INTRODUÇÃO

A estatística multivariada surgiu por volta de 1901, sendo que apenas com o auxílio de programas computacionais tornou-se possível desenvolver e aplicar essa técnica com maior facilidade.

Os pesquisadores devem ter cautela ao trabalhar com as técnicas multivariadas, pois a arte do seu uso está na escolha das opções mais apropriadas para detectar os padrões esperados nos dados. Para escolher as opções corretas, recomenda-se que os leitores exercitem-nas, com cautela, durante o tempo necessário para apreender as limitações de cada técnica antes de tentar explorar suas grandes potencialidades (MAGNUSSON; MOURÃO, 2003).

Sabe-se que os métodos estatísticos, para analisar variáveis, estão dispostos em dois grandes grupos: um que analisa as variáveis de forma isolada – a estatística univariada – e outro de forma conjunta – a estatística multivariada.

Tanto no caso univariado quanto no multivariado, a variável deve conter as informações, de maneira a manter a integridade dos dados, mediante tabelas ou gráficos e medidas estatísticas que irão resumir e representar a informação.

Muitas vezes, quando se realiza uma análise com várias variáveis, tem-se a ideia de que já se está trabalhando em um caso multivariado, desde que se tenha um conjunto formado por diversos indivíduos, mas deve-se observar que um conjunto de variáveis será considerado multivariado se várias medidas forem realizadas do mesmo indivíduo, apresentando, assim, uma ligação entre as variáveis determinada por meio da correlação. A partir daí, dois procedimentos podem ser adotados.

O primeiro é utilizar métodos estatísticos univariados e fazer a análise de cada variável, tomando-se, assim, uma decisão isolada, sem levar em consideração os possíveis inter-relacionamentos entre as demais variáveis. O segundo é realizar a análise considerando todas as variáveis simultaneamente, levando em conta as inter-relações existentes entre essas variáveis mediante métodos da análise multivariada (AM).

Por exemplo, ao se considerar uma pessoa, pode-se observar diversas variáveis, como altura, peso, idade, renda mensal, grau de instrução, preferência por determinado estilo musical e assim por diante. Nesse caso, tem-se uma pessoa e várias medidas que, de uma forma ou de outra, poderão estar ligadas entre si. Para que seja possível analisar

diversas variáveis simultaneamente sobre um conjunto de indivíduos, será necessário um método de estimação conjunta, o qual caracteriza uma análise multivariada.

Também é importante saber que, ao se utilizar técnicas estatísticas multivariadas, dois dos principais objetivos são a classificação e a redução da dimensionalidade, e nesse processo sempre ocorrerá alguma perda de informação. Essa perda é mensurável, e cabe ao pesquisador verificar o limite aceitável para que o seu estudo seja validado e confiável. Dessa forma, surge um *trade off*, em que ele deve decidir se é preferível ter alguma informação sobre o estudo ou ignorar os efeitos de correlação existente e realizar uma análise univariada.

Para Pereira (2001), a AM envolve uma grande multiplicidade de conceitos estatísticos matemáticos, sendo, a rigor, qualquer abordagem analítica que considere o comportamento de muitas variáveis simultaneamente. Essa análise permite ainda ao pesquisador avaliar erros de interpretação dos dados e revelar informações que não são imediatamente percebidas, quando se utiliza uma análise mais simples.

Vale relembrar que duas condições são importantes para a realização dos estudos multivariados: os valores das diferentes variáveis devem ser obtidos sobre os mesmos indivíduos e/ou casos; e as mensurações das variáveis devem ser interdependentes e consideradas simultaneamente (KENDALL, 1980).

Entre os métodos multivariados, existe uma diversidade de procedimentos, passando pela inferência estatística multivariada que aborda testes de hipóteses sobre o vetor média, comparações de diversas médias, modelos de regressões multivariados e métodos centrados na estrutura da matriz de variância-covariância e nas distâncias de vetores de médias.

Os métodos multivariados que utilizam a decomposição da matriz de variância-covariância e os métodos de distância são, em maioria, métodos exploratórios, já que não usam nenhum teste específico para comprovar os resultados encontrados, mas são úteis para descrever o conjunto de dados na presença de correlação e servem para levantar hipóteses sobre o estudo realizado (HAIR et al., 2005).

A obra apresenta os principais métodos estatísticos multivariados, acompanhados pela literatura, por comentários, descrições e interpretações de cada janela utilizada para a realização desses métodos no *software Statistica*.

Para a leitura deste livro, não é necessário um conhecimento aprofundado de estatística, pois, da maneira como o material foi organizado, tendo o leitor um conhecimento básico, é possível obter o máximo de proveito.

2
CONCEITOS BÁSICOS

Para medir, avaliar ou quantificar informações, o pesquisador utiliza-se das mais variadas formas de escalas de medida, sendo essas muito importantes para o estudo, pois, dependendo da característica da variável mensurada, será definido o tipo de método multivariado a ser aplicado.

Definindo-se a escala a ser utilizada, o pesquisador depara-se com os critérios de significância e a precisão do instrumento de medida, que são a validade e a confiabilidade. Essas medidas devem ser testadas de forma a validarem o estudo e a possibilitarem conclusões futuras.

2.1 Variável estatística

É uma característica em estudo que assume diferentes valores para diferentes elementos. Variável é tudo aquilo que se pode medir e se divide em variáveis quantitativas e qualitativas. Essa classificação definirá o tipo de análise a ser realizada.

2.1.1 Variável quantitativa

É aquela que assume um valor em um conjunto predeterminado denominado domínio da variável. A variável assume um número real, uma unidade constante e comum de mensuração, podendo ser contínua ou discreta, dependendo do tipo de dados coletados na pesquisa e do tipo de medida que foi utilizada.

Quando se utiliza a variável contínua, é admissível utilizar as estatísticas paramétricas que pressupõem o conhecimento de alguns parâmetros, como média, desvio padrão, entre outras medidas. Também é possível trabalhar com estatísticas descritivas e testes paramétricos, como o teste "*t*" de *Student*, o teste "f" de *Snedecor*, entre outros.

É importante observar que uma variável contínua geralmente resulta de algum tipo de medição, pois, dependendo da precisão do instrumento de medida utilizado, pode-se obter vários dígitos após a vírgula, caracterizando, dessa maneira, que o dado a ser medido assume um ponto dentro de um intervalo, como é o caso das balanças eletrônicas de precisão, medidas feitas com paquímetros e outros instrumentos de precisão.

A variável discreta, ao contrário das variáveis contínuas, assume um único valor, o qual, em geral, é o de números inteiros (conjunto dos números inteiros). Essa variável, geralmente, resulta de enumerações e contagens.

2.1.2 Variável qualitativa

A variável qualitativa tem uma conotação com a abstração, pois depende do conceito de cada indivíduo. Ela está intimamente ligada ao conceito de atributo, que, ao contrário de uma variável, é algo que não é fácil de quantificar numericamente. Um atributo pode ter um significado para uma pessoa, e para outra um significado completamente diferente, pois não existe um valor numérico para determiná-lo nem um padrão de comparação.

Um exemplo claro de atributo é quando uma pessoa analisa uma obra de arte e a considera como de extrema beleza, ao passo que a segunda pessoa pode, simplesmente, desprezá-la, afirmando que aquela obra não é tão estarrecedora quanto parece. Um atributo depende de pessoa para pessoa, como conceito de beleza, preferência por uma cor e assim por diante.

2.2 Escalas de medidas

O registro das ocorrências de um estudo científico necessita de um meio de se representar os acontecimentos e fenômenos adequadamente, o que é realizado através das chamadas **escalas de mensuração** ou **escalas numéricas**. Basicamente, trata-se de modos de expressar qualidade e/ou quantidade.

Para que possam cumprir adequadamente a sua tarefa, as escalas precisam apresentar duas propriedades:

- Exaustividade: abrangência que permita representar todos os fatos e ocorrências possíveis.
- Exclusividade: coerência para que qualquer fato ou acontecimento só possa ser representado de uma única maneira.

Em suma, as escalas precisam ser capazes de exaurir todas as possibilidades e, ao mesmo tempo, ser mutuamente excludentes (MARTINS; SANTOS, 2006).

Uma escala de medida obedece a níveis de mensuração, o que é uma forma de hierarquizar a escala, sendo distribuídas em quatro níveis: nominal, ordinal, intervalar e de razões.

2.2.1 Escala nominal ou classificatória

É uma das formas mais simples de mensuração que existe, pois se utiliza de nomes ou outros símbolos, para classificar objetos, pessoas ou outras características, estabelecendo relações de igualdade ou diferenças. Nesses casos, não é permitido que se utilize uma medida estatística para representar o conjunto de dados além da moda. Em geral, o teste estatístico univariado recomendado para se realizar é o qui-quadrado, portanto a análise multivariada recomendada, nesses casos, seria a análise de correspondência, a qual também trata de classificações.

2.2.2 Escala ordinal ou por postos

Neste tipo de escala, os elementos podem ter algum tipo de ordenação, categorias ou simplesmente uma relação de ordem de maior, menor, superior e assim por diante. O importante é que as escalas transmitam a **relação** entre as variáveis. A escala nominal está incluída nesta categoria também. Portanto, a relação de igualdade e diferença também é válida. O que se deve observar é que os números escolhidos para codificar uma variável não refletem um significado real e, dessa forma, não é possível utilizar outras técnicas estatísticas.

Nas escalas ordinais, aplicam-se somente as medidas estatísticas, como a mediana, o percentil, o quartil, a correlação por postos e os testes de Spearman, Kendall e Tau e o de significância de Mann-Whitney, e na área multivariada são aplicadas as técnicas de agrupamentos. Muitos autores revelam que, ao se utilizar uma escala de **Likert**, é possível aplicar outras técnicas estatísticas.

2.2.3 Escala intervalar

A escala intervalar é mais forte que a ordinal e possibilita a ordenação e o cálculo de distâncias. Quando a escala tem todas as características de uma escala ordinal, e quando, além disso, conhecem-se as distâncias entre dois números quaisquer da escala, então se consegue uma mensuração consideravelmente mais forte que a ordinal. Atribui-se à variável um número real, uma unidade constante e comum de mensuração, e um ponto zero, o qual é arbitrário.

Uma propriedade básica da escala intervalar, que é a primeira escala verdadeiramente quantitativa, derivada do modo como ela é definida, é o fato de que duas variações iguais em termos de medidas intervalares necessariamente correspondem a variações iguais em termos do valor do que está sendo medido.

As estatísticas utilizadas são a média aritmética, o desvio padrão, o desvio médio, a correlação e o teste de "*t*". É possível utilizar quase todos os procedimentos multivariados.

2.2.4 Escala de razões

A escala de razões é a mais completa e sofisticada das escalas numéricas. Ela é uma quantificação produzida a partir da identificação de um ponto zero que é fixo e absoluto, representando, de fato, um ponto de nulidade, ausência e/ou mínimo. Nela, uma unidade de medida é definida em termos da diferença entre o ponto zero e uma intensidade conhecida. A partir disso, cada observação é aferida segundo a sua distância do ponto zero, expressa em unidades da medida que foi definida. Quando esse tipo de escala é utilizado, todos os procedimentos estatísticos, uni e multivariados, podem ser utilizados.

2.3 Validade e confiabilidade

A validade é um conceito importante, pois nos reporta à capacidade que o instrumento tem de medir o que realmente se deseja saber. A validade significa que, se um instrumento de coleta for replicado, ele deverá

apresentar sempre os mesmos resultados em termos estatísticos, isto é, replicáveis e consistentes, apresentando apenas uma variação aleatória.

Validade de constructo é uma avaliação do quanto um instrumento mede o constructo teórico que o investigador deseja medir. Ao mesmo tempo, é uma avaliação desse próprio constructo e da teoria que o sustenta. Ao contrário da validade aparente, a validade de constructo requer mais do que a opinião de especialistas: requer uma demonstração de que o constructo em questão existe, de que é distinto de outros constructos e de que o instrumento mede este constructo em particular e não outros (SELLTIZ et al., 1987).

A propósito, a palavra "constructo" vem do latim e, em termos de psicologia, significa a percepção ou o pensamento formado a partir de combinações de lembranças com acontecimentos atuais. Relativamente à matemática, a palavra também é uma combinação, só que expressa em termos de combinações lineares, que vão desde a combinação linear das variáveis originais até a combinação linear escrita em termos de fatores.

Likert (1967) propõe, como forma de avaliar a validade de constructo, a análise do item, que consiste no cálculo da correlação de cada item com o total de itens. Uma correlação negativa implica que os extremos da escala não foram propriamente especificados, devendo ser invertidos. Se o coeficiente de correlação for zero ou muito baixo, significa que o item falhou em medir o que o resto dos itens está medindo. Tais itens são chamados de não diferenciáveis e não têm qualquer contribuição para o grupo de itens; portanto, poderiam ser excluídos da análise.

Assim, na prática, a validade de constructo pode ser realizada pela correlação entre os escores obtidos em cada item, para cada observação, e o obtido para uma variável auxiliar, como, por exemplo, a soma de todos os escores obtidos nos outros itens.

Considerando que os cálculos dos coeficientes de correlação eram muito trabalhosos, Likert (1967) sugeriu que fosse utilizado o critério da consistência interna. Nesse critério, normalmente na área da psicologia, comparam-se as atitudes dos grupos em extremos opostos, ou seja, comparam-se aqueles que fortemente aprovam com os que fortemente desaprovam um item. Isto é feito calculando-se a soma total do instrumento para todos os observadores, ordenando-os em ordem