# Análise conjunta e análise de agrupamentos (clustering)

Você vai compreender a análise conjunta e sua aplicação na avaliação de objetos, além da análise de agrupamentos, seus usos gerenciais, diferenças entre métodos e como aplicar essas técnicas utilizando comandos no software R.

Profa. Manoela Gonçalves Cabo

## Objetivos

- Analisar a análise conjunta enquanto técnica de dependência, explorando seu conceito, aplicações gerenciais e etapas de planejamento experimental.
- Interpretar a análise de agrupamentos, considerando seu conceito, funcionamento e o processo de decisão envolvido na formação de clusters.

## Introdução

Neste conteúdo, será explorada a aplicação de duas importantes técnicas estatísticas multivariadas: a análise conjunta e a análise de agrupamentos, também conhecida como clustering. Ambas desempenham um papel fundamental na tomada de decisão baseada em dados, sendo amplamente utilizadas em contextos gerenciais, de marketing e pesquisa de mercado.

A análise conjunta é uma técnica de dependência voltada à avaliação de objetos, produtos ou serviços com base nas preferências ou julgamentos dos indivíduos. Por meio dela, é possível estimar o valor atribuído a diferentes atributos que compõem um objeto, permitindo compreender o que realmente influencia as escolhas dos consumidores. Neste estudo, serão definidos os principais conceitos dessa técnica, discutidas suas aplicações práticas no ambiente organizacional e apresentada uma comparação com outros métodos multivariados. Também será abordado o planejamento de um experimento de análise conjunta, etapa essencial para a obtenção de resultados válidos e úteis.

Em seguida, será abordada a análise de agrupamentos, uma técnica de interdependência que tem como objetivo identificar grupos homogêneos dentro de um conjunto de dados. O clustering permite segmentar objetos, pessoas ou eventos com base em suas semelhanças, sendo essencial para estratégias como segmentação de mercado, desenvolvimento de produtos e personalização de serviços. Serão explicados seu funcionamento, as etapas do processo de decisão envolvido e, para ilustrar sua aplicação prática, serão apresentados exemplos com a utilização de comandos no software R para a formação dos clusters.

Este conteúdo proporcionará uma visão integrada e aplicada dessas técnicas, permitindo que você desenvolva competências analíticas voltadas à resolução de problemas reais em contextos organizacionais e de pesquisa.

# Conceito da análise conjunta


A seguir, assista ao vídeo e conheça a técnica de análise conjunta.

> **Conteúdo interativo**
>
> Acesse a versão digital para assistir ao vídeo.

A análise conjunta é a técnica de dependência utilizada na avaliação de objetos. Assim, com o uso de variáveis independentes não métricas, a análise conjunta faz lembrar da análise de variância (ANOVA), que tem uma fundamentação nas análises de experimentos. Como tal, a análise conjunta é fortemente correlacionada com a experimentação tradicional.

> A análise conjunta é uma das técnicas multivariadas usadas especificamente para entender como o respondente desenvolve as preferências por alguns tipos de objetos (produtos, serviços ou ideias).

É fundamentada nas premissas simples de que os clientes avaliam os valores de um objeto (real ou hipotético) combinados às quantias separadas de valores fornecidas por cada atributo.

Além disso, consumidores podem dar suas estimativas de preferência avaliando objetos formados por combinações de atributos. Ou seja, é um método que mostra de forma realista as decisões de clientes ou consumidores, como trocas entre produtos ou serviços de múltiplo atributo.

A análise conjunta pode ser expressa como:

$$Y1 = X1 + X2 + \ldots + Xn$$

(Não métrica ou Métrica) (Não métricas)

Atenção

A análise conjunta é, em suma, um conjunto de técnicas e métodos especificamente desenvolvidos para abranger preferência individual, e que compartilham uma fundamentação teórica com base nos modelos de integração de informação e medição funcional.

Ela é mais adequada para compreender a reação de clientes/consumidores e avaliar combinações determinadas previamente de atributos que representam produtos ou serviços em potencial. A flexibilidade e a característica da análise conjunta surgem a partir do que se segue:

a) Habilidades em acomodar tanto uma variável dependente métrica quanto não métricas.

b) Os usos de variáveis preditoras categóricas.

c) As suposições muito gerais sobre as relações de variáveis independentes com a dependente.

> A análise conjunta tradicional é uma das metodologias que empregam os princípios clássicos da análise conjunta na tarefa conjunta, usando modelos aditivos da preferência de consumidor e métodos de apresentação de comparação pareada ou de perfil completo.

Essa análise conjunta é única entre os métodos multivariados, no sentido de que primeiramente construímos um conjunto de produtos ou serviços reais ou hipotéticos, combinando níveis escolhidos de cada atributo.

Ao criarmos essas combinações, temos um planejamento, que é o conjunto de estímulos apresentados ao respondente. Essa combinação ou estímulo são então apresentados a respondentes, os quais fornecem apenas suas avaliações gerais, em um processo chamado de tarefa conjunta.

Assim, pedimos ao respondente para alcançar uma tarefa realista, que é escolher um conjunto de objetos. Os respondentes nada mais têm a dizer, como o quão importante é um atributo individual para eles ou como o objeto funciona em relação a um atributo específico.

Já que iremos construir os objetos hipotéticos de uma maneira específica, a influência de cada atributo e de cada um dos valores de cada um dos atributos sobre o julgamento de um respondente quanto a sua utilidade pode ser determinada a partir das avaliações gerais.

## Usos gerenciais da análise conjunta

A seguir, saiba como gerenciar a análise conjunta.

Conteúdo interativo

Acesse a versão digital para assistir ao vídeo.

As análises conjuntas consideram que quaisquer conjuntos de objetos (como, por exemplo, as marcas e companhias), ou conceitos (como, por exemplo, posicionamento, benefícios e imagem), são avaliados como uma coleção de atributos.

> A flexibilidade das análises conjuntas viabiliza o bom emprego em várias áreas nas quais as decisões são estudadas.

Após produzir a contribuição de cada fator à avaliação geral do consumidor, podemos então proceder da seguinte forma:

- Definir o objeto ou conceito com a combinação ótima de características.
- Mostrar as contribuições relativas de cada atributo e cada nível para a avaliação geral do objeto.
- Usar as estimativas de julgamentos de comprador ou cliente para prever preferências entre objetos com diferentes conjuntos de características (outros elementos mantidos constantes).
- Isolar grupos de clientes potenciais que atribuem diferente importância às características para definir segmentos com potenciais altos e baixos.
- Identificar oportunidades de marketing explorando o potencial de mercado para combinações de características indisponíveis no momento.

O conhecimento das estruturas de preferências para os indivíduos permite flexibilidade quase ilimitada para examinar reações agregadas e individuais em grande número de assuntos ligados aos produtos ou serviços.

> As análises conjuntas, conhecidas como modelos de decomposição, são diferentes no sentido de que precisamos conhecer apenas a preferência geral dos respondentes para um dos estímulos.

Os valores de cada atributo (variáveis independentes) já estava especificado quando o estímulo foi criado. Desse modo, a análise conjunta pode determinar (decompor) o valor de cada um dos atributos usando apenas a medida de preferência geral.

Esse método aplica uma variável estatística muito parecida em forma com aquela usada em outras técnicas multivariadas. A variável estatística conjunta é uma combinação linear de efeitos das variáveis independentes (fatores) sobre uma variável dependente. A diferença importante é que na variável estatística conjunta especificamos as variáveis independentes (fatores) e seus valores (níveis).

A única informação fornecida pelo respondente é a medida dependente. Os níveis especificados pelo pesquisador são então usados pela análise conjunta para decompor a resposta do respondente em efeitos para cada nível, muito parecido com o que é feito na análise de regressão para cada variável independente.

Sendo assim, a análise conjunta representa um tipo híbrido de técnica multivariada para estimar relações de dependência. Em um sentido, ela combina métodos tradicionais (ou seja, regressão e ANOVA), fornecendo muito da flexibilidade mostrada na regressão aliada com a tradição da experimentação de ANOVA.

No entanto, ela é única no sentido de que é decomposicional por natureza, e resultados podem ser estimados para cada respondente em separado. Como tal, a análise conjunta oferece uma ferramenta especializada de análise especificamente para compreender decisões de clientes e suas estruturas de preferência.

Ao mesmo tempo que demanda considerável trabalho de frente no planejamento da análise em si, fornece um poderoso e esclarecedor método para análise de preferências e tomadas de decisões por parte de clientes.

Os objetivos da análise conjunta são:

Ser única

A análise conjunta é única em relação a outras técnicas multivariadas, pois:

- É uma forma de modelo decomposicional que tem muitos elementos de um experimento;
- Clientes fornecem apenas uma avaliação geral de preferências para objetos (estímulos) criados;
- Estímulos são criados por combinação de um nível (valor) de cada fator (atributo);
- Cada respondente avalia estímulos o suficiente de forma que resultados conjuntos são estimados para cada indivíduo.

Ser "bem-sucedida"

Uma análise conjunta "bem-sucedida" requer que:

- Defina precisamente todos os atributos (fatores) que têm impactos positivos e negativos sobre preferência;
- Aplique o modelo apropriado sobre como os clientes combinam os valores de atributos individuais em avaliações gerais de um objeto.

Ter resultados

Os resultados de análise conjunta podem ser usados para:

- Fornecer estimativas da "utilidade" de cada nível dentro de cada atributo;
- Definir a utilidade total de quaisquer estímulos de forma que possam ser comparados com outros para se prever escolhas de clientes (p.ex., participação de mercado).

## Planejamento de um experimento de análise conjunta

A seguir, assista ao vídeo e compreenda como planejar a análise conjunta.

Conteúdo interativo

Acesse a versão digital para assistir ao vídeo.

O processo de decisão começa com uma especificação dos objetivos da análise conjunta. Como a análise conjunta é semelhante a um experimento, a conceituação da pesquisa é crítica para seu sucesso. Depois que os objetivos são definidos, as questões relacionadas ao verdadeiro plano de pesquisa são abordadas e, as suposições, avaliadas.

A discussão foca então em como o processo de decisão considera a estimação real dos resultados conjuntos, a interpretação dos resultados e os métodos usados para validá-los. A discussão termina com um exame do uso de resultados de análise conjunta em análises posteriores, como segmentação de mercado e simuladores de escolha.

Ao aplicarmos a análise conjunta, devemos tomar várias decisões ao planejar o experimento e analisar seus resultados. Os estágios de 1 a 7 listados a seguir mostram os passos gerais seguidos no delineamento e execução de um experimento de análise conjunta:

### Estágio 1

No estágio 1, objetivos da análise conjunta, como ocorre com qualquer análise estatística, o ponto de partida é a questão de pesquisa. Nesse estágio, selecionam-se objetivos; determina-se a contribuição de variáveis independentes; estabelece-se o modelo de julgamentos de consumidor e os elementos de utilidade total, identificando os critérios chave de decisão.

### Estágio 2

No estágio 2, projeto de uma análise conjunta, tendo resolvido as questões relativas aos objetivos da pesquisa, desviamos a atenção para as questões da escolha de uma metodologia conjunta: Quantos atributos devem ser usados?

Se forem seis atributos ou menos, deve-se utilizar a análise conjunta baseada em escolha; se forem menos de dez atributos, deve-se utilizar a análise conjunta tradicional; se forem dez ou mais, deve-se utilizar a escolha adaptativa.

### Estágio 3

No estágio 3 são feitas as suposições da análise conjunta, adequação da forma do modelo e representatividade da amostra. A análise conjunta tem o menor conjunto restritivo de suposições associadas com a estimação do modelo. O delineamento experimental estruturado e a natureza generalizada do modelo tornam desnecessária a maioria dos testes realizados em outros métodos de dependência.

Portanto, os testes estatísticos de normalidade, homocedasticidade e independência que foram executados para outras técnicas de dependência não são necessários para a análise conjunta. O emprego de delineamentos de estímulos baseados em estatísticas também garante que a estimação não seja confusa, e que os resultados sejam interpretáveis sob a regra de composição assumida.

Estágio 4

No estágio 4, estimação do modelo conjunto e avaliação do ajuste geral, as opções disponíveis ao pesquisador em termos de técnicas de estimação aumentaram dramaticamente nos últimos anos.

Além disso, o desenvolvimento de técnicas em conjunção com métodos especializados de apresentação de estímulos (por exemplo, a análise conjunta adaptativa ou baseada em escolhas) é apenas um melhoramento desse tipo. O pesquisador, ao obter os resultados de um estudo de análise conjunta, tem inúmeras opções disponíveis quando seleciona o método de estimação e avalia os resultados.

Estágio 5, 6 e 7

No estágio 5, é feita a interpretação dos resultados, resultados agregados versus desagregados e a importância relativa de atributos. No estágio 6, a validação dos resultados, com a validade interna e externa. Já no estágio 7 é realizada a aplicação dos resultados conjuntos, com a segmentação, análise de lucratividade e simulador de escolhas.

# Verificando o aprendizado

Questão 1

Descreva o conceito de análise conjunta e expresse a sua formulação.

Chave de resposta

Análise conjunta é uma técnica multivariada usada especificamente para entender como os respondentes desenvolvem preferências por quaisquer tipos de objetos (produtos, serviços ou ideias). É baseada na premissa simples de que os consumidores avaliam o valor de um objeto (real ou hipotético) combinando as quantias separadas de valor fornecidas por cada atributo.

Além disso, clientes podem fornecer melhor suas estimativas de preferência julgando objetos formados por combinações de atributos. Ou seja, é um método que retrata de forma realista as decisões de consumidores, como trocas entre produtos ou serviços de múltiplos atributos. A análise conjunta pode ser expressa como:

$$Y_1 = X_1 + X_2 + \ldots + X_n$$

(Não métrica ou Métrica) (Não métricas)

Quetsão 2

Descrever os objetivos da análise conjunta:

Chave de resposta

A análise conjunta é única em relação a outras técnicas multivariadas, pois:

- É uma forma de modelo decomposicional que tem muitos elementos de um experimento;
- Clientes fornecem apenas uma avaliação geral de preferências para objetos (estímulos) criados;
- Estímulos são criados por combinação de um nível (valor) de cada fator (atributo);
- Cada respondente avalia estímulos o suficiente de forma que resultados conjuntos são estimados para cada indivíduo.

Uma análise conjunta "bem-sucedida" requer que:

- Defina precisamente todos os atributos (fatores) que têm impactos positivos e negativos sobre preferência;
- Aplique o modelo apropriado sobre como os clientes combinam os valores de atributos individuais em avaliações gerais de um objeto.

Os resultados de análise conjunta podem ser usados para:

- Fornecer estimativas da "utilidade" de cada nível dentro de cada atributo;
- Definir a utilidade total de quaisquer estímulos de forma que possam ser comparados com outros para prever escolhas de clientes (p.ex., participação de mercado).

Questão 3

Os estágios de 1 a 7 mostram os passos gerais seguidos no delineamento e execução de um experimento de análise conjunta. Descreva os três primeiros estágios de um experimento de análise conjunta:

## Chave de resposta

No estágio 1, objetivos da análise conjunta, como ocorre com qualquer análise estatística, o ponto de partida é a questão de pesquisa. Nesse estágio, selecionam-se objetivos; determina-se a contribuição de variáveis independentes; estabelece-se o modelo de julgamentos de consumidor e os elementos de utilidade total, identificando os critérios chave de decisão.

No estágio 2, projeto de uma análise conjunta, tendo resolvido as questões relativas aos objetivos da pesquisa, desviamos a atenção para as questões da escolha de uma metodologia conjunta: Quantos atributos devem ser usados?

Se forem seis atributos ou menos, deve-se utilizar a análise conjunta baseada em escolha; se forem menos de dez atributos, deve-se utilizar a análise conjunta tradicional; se forem dez ou mais, deve-se utilizar a escolha adaptativa.

No estágio 3, são feitas as suposições da análise conjunta, adequação da forma do modelo e representatividade da amostra. A análise conjunta tem o menor conjunto restritivo de suposições associadas com a estimação do modelo. O delineamento experimental estruturado e a natureza

generalizada do modelo tornam desnecessária a maioria dos testes realizados em outros métodos de dependência.

Portanto, os testes estatísticos de normalidade, homocedasticidade e independência que foram executados para outras técnicas de dependência não são necessários para a análise conjunta. O emprego de delineamentos de estímulos baseados em estatísticas também garante que a estimação não seja confusa, e que os resultados sejam interpretáveis sob a regra de composição assumida.

# Conceito de análise de agrupamentos ou clustering

Assista ao vídeo e conheça o conceito de análise de atrapamento ou c*lustering*.

Conteúdo interativo

Acesse a versão digital para assistir ao vídeo.

A análise de agrupamentos, também conhecida como análise de conglomerados, classificação ou cluster, tem como objetivo dividir os elementos da amostra, ou população, em grupos, de forma que os elementos pertencentes a um mesmo grupo sejam similares entre si com respeito às variáveis (caraterística) que neles formam medidas, e os elementos em grupos diferentes sejam heterogêneos em relação a estas mesmas características.

> A análise de agrupamento é uma técnica de interdependência, porém, está concentrada somente na definição de estrutura, avaliando a interdependência sem quaisquer relações de dependência associadas, como visto nas técnicas apresentadas nas aulas anteriores.

Nenhuma das técnicas de interdependência definirá a estrutura para otimizar ou maximizar uma relação de dependência. É nossa tarefa primeiramente utilizar esses métodos na identificação de estrutura e então empregá-la onde for apropriado.

Comentário

Os objetivos de relações de dependência não são “incorporados” nesses métodos de interdependência – eles avaliam a estrutura para seus próprios objetivos, e nenhum outro

Quando realizamos um estudo ou uma pesquisa, frequentemente encontramos situações mais bem resolvidas pela definição de grupos de objetos homogêneos, sejam eles indivíduos, empresas, produtos ou mesmo comportamentos. A técnica mais comumente usada para essa finalidade é a análise de agrupamentos.

A análise de agrupamentos reúne indivíduos ou objetos em grupos tais que os objetos no mesmo grupo são mais parecidos uns com os outros do que com os objetos de outros grupos. A ideia é maximizar a homogeneidade de objetos dentro de grupos, ao mesmo tempo em que se maximiza a heterogeneidade entre os grupos.

Análise de agrupamentos é um grupo de técnicas multivariadas cuja finalidade principal é agregar objetos com base nas características que eles possuem. Ela tem sido chamada de análise Q, construção de tipologia, análise de classificação e taxonomia numérica.

Comentário

Essa variedade de nomes se deve ao uso de métodos de agrupamento nas mais diversas áreas, como psicologia, biologia, sociologia, economia, engenharia e administração.

A análise de agrupamentos, também conhecida como análise de conglomerados, classificação ou cluster, tem como objetivo dividir os elementos da amostra, ou população, em grupos, de forma que os elementos pertencentes a um mesmo grupo sejam similares entre si com respeito às variáveis (caraterística) que neles formam medidas, e os elementos em grupos diferentes sejam heterogêneos em relação a estas mesmas características.

A análise de agrupamentos classifica objetos (p.ex., respondente, produtos ou outras entidades) de modo que cada objeto é semelhante aos outros no agrupamento com base em um conjunto de características escolhidas. Os agrupamentos resultantes de objetos devem então exibir elevada homogeneidade interna (dentro dos agrupamentos) e elevada heterogeneidade externa (entre agrupamentos).

> Assim, se a classificação for bem-sucedida, os objetos dentro dos agrupamentos estarão próximos quando representados graficamente, e diferentes agrupamentos estarão distantes.

Atenção

Uma questão importante refere-se ao critério a ser utilizado para se decidir até que ponto dois elementos do conjunto de dados podem ser considerados semelhantes ou não. Para responder a essa questão é necessário considerar medidas que descrevam a similaridade entre elementos amostrais de acordo com as características que neles foram medidas.

O conceito da variável estatística é uma questão central. A variável estatística de agrupamento é o conjunto de variáveis que representam as características usadas para comparar objetos na análise de agrupamentos. Como a variável estatística de agrupamentos inclui apenas as variáveis usadas para comparar objetos, ela determina o caráter dos objetos.

A variável estatística em análise de agrupamentos é determinada de maneira muito diferente do que ocorre em outras técnicas multivariadas. A análise de agrupamentos é a única técnica multivariada que não estima a variável estatística empiricamente, mas, em vez disso, usa a variável estatística como especificada por nós.

> O foco da análise de agrupamentos é a comparação de objetos com base na variável estatística, não na estimação da variável estatística em si.

O agrupamento de objetos é, na verdade, um meio para um fim em termos de uma meta conceitualmente definida. Os papéis mais comuns que a análise de agrupamentos podem desempenhar em desenvolvimento conceitual incluem a redução de dados e geração de hipóteses:

### Redução de dados

Um pesquisador que tenha coletado dados por meio de um questionário pode se deparar com muitas observações sem significado, a não ser que sejam classificadas em grupos com os quais se possa lidar. A análise de agrupamentos pode realizar esse procedimento de redução de dados objetivamente pela redução da informação de uma população inteira ou de uma amostra para a informação sobre subgrupos específicos e menores.

### Geração de hipóteses

A análise de agrupamentos também é útil quando um pesquisador deseja desenvolver hipóteses relativas à natureza dos dados ou examinar hipóteses previamente estabelecidas.

Os objetivos da análise de agrupamentos são:

| | |
|---|---|
| **Descrição taxonômica** | Identificar grupos naturais dentro dos dados |
| **Simplificação de dados** | A habilidade de analisar grupos de observações semelhantes em vez de todas as observações individuais |
| **Identificação de relação** | A estrutura simplificada da análise de agrupamentos retrata relações não reveladas de outra forma; Considerações teóricas, conceituais e práticas devem ser levadas em conta |
| **Quando se selecionam variáveis de agrupamento para análise** | Somente variáveis que se relacionam especificamente com os objetivos da análise de agrupamentos são incluídas; variáveis irrelevantes não podem ser excluídas da análise uma vez que ela comece; Variáveis selecionadas caracterizam os indivíduos (objetos) sendo agrupado. |

Manoela Gonçalves Cabo

## O processo de decisão em análise de agrupamentos

No vídeo a seguir, conheça o processo de decisão em análise de agrupamentos.

### Conteúdo interativo

Acesse a versão digital para assistir ao vídeo.

Para demonstrar como a análise de agrupamentos opera, o processo da análise, examinamos um exemplo simples que ilustra algumas das questões-chave: Medir similaridade, formar agrupamentos e decidir sobre o número de agrupamentos que melhor representam uma estrutura. Também discutimos brevemente o equilíbrio de considerações objetivas e subjetivas que devem ser tratadas por nós.

O objetivo principal da análise de agrupamentos é definir a estrutura dos dados colocando as observações mais parecidas em grupos e, as menos parecidas, distantes. Para conseguir isso, devemos tratar três questões básicas:

Medir a similaridade

---

Necessitamos de um método de comparação simultânea de observações sobre as variáveis de agrupamentos. Diversos métodos são possíveis, incluindo a correlação entre objetos ou talvez uma medida de sua proximidade em um espaço bidimensional tal que a distância entre observações indique similaridade.

Suponha que tenha disponível um conjunto de dados constituído de elementos amostrais, tendo-se medido p-variáveis aleatórias em cada um deles. O objetivo é agrupar esses elementos em g grupos. Para cada elemento amostral j, tem-se, portanto, o vetor de medidas $X_j$ definido por:

$$X_j = [X_{1j}, X_{2j} \dots X_{pk}]', j = 1, 2, \dots, n$$

em que $X_{ij}$ representa o valor observado da variável i medida no elemento j.

Para que se possa proceder ao agrupamento de elementos, é necessário que se decida a priori a medida de similaridade ou dissimilaridade que será utilizada. Existem várias medidas diferentes e cada uma delas produz um determinado tipo de agrupamento.

Algumas medidas mais comuns, apropriadas para variáveis quantitativas, são: Distância Euclidiana, Distância generalizada ou ponderada, Distância de Minkowisky, Coeficiente de concordância simples, Coeficiente de concordância positiva, Coeficiente de concordância de Jaccard, Distância Euclidiana média, entre outras.

A distância euclidiana entre dois elementos $X_l$ e $X_k$, l ≠ k, é definida por:

$$d(X_l, X_k) = \left[(X_l - X_k)'(X_l - X_k)\right]^{\frac{1}{2}} = \left[\sum_{i=1}^{p}(X_{il} - X_{ik})^2\right]^{\frac{1}{2}}, j = 1, 2, \dots, n$$

Ou seja, os dois elementos amostrais são comparados em cada variável pertencente ao vetor de observações.

Formação dos agrupamentos

Não importa como a similaridade é medida, o procedimento deve agregar as observações mais similares em um agrupamento. Esse procedimento deve determinar a pertinência ao grupo de cada observação para cada conjunto de agrupamentos formados.

Com medidas de similaridade já calculadas, agora vamos para a formação de agrupamentos com base na medida de similaridade de cada par de observação. Geralmente formamos um número de soluções de agrupamentos (uma solução de dois agrupamentos, três etc.).

Uma vez que os agrupamentos são formados, escolhemos então a solução final a partir do conjunto de soluções possíveis. Primeiro, discutimos como os agrupamentos são formados e, em seguida, examinamos o processo para seleção de uma solução final.

Existem várias técnicas para formação dos agrupamentos, conglomerados ou clusters. Essas técnicas são classificadas em hierárquicas e não hierárquicas.

Técnica Hierárquica Aglomerativa: Identifica as duas observações mais semelhantes (mais próximas) que ainda não estão no mesmo agrupamento e combina seus agrupamentos.

Aplicamos essa regra repetidamente para gerar várias soluções, começando com cada observação em seu próprio “agrupamento” e então combinando dois agrupamentos por vez até que todas as observações estejam em um único agrupamento.

Esse processo é o chamado procedimento hierárquico, porque opera no estilo stepwise para formar um intervalo inteiro de soluções de agrupamentos. É também um método aglomerativo, porque os agrupamentos são formados pela combinação de outros já existentes.

Decisão de quantos grupos formamos

A tarefa final é selecionar um conjunto de agrupamentos como a solução final. Fazendo isso, nos deparamos com uma ponderação a ser feita: Menos agrupamentos e menos homogeneidade dentro dos agregados versus grande número de agrupamentos e maior homogeneidade interna.

A estrutura simples, com vistas à parcimônia, é refletida internamente com o menor número de agrupamentos possível. No entanto, quando o número de agrupamentos diminui, a heterogeneidade dentro dos grupos necessariamente aumenta.

Assim, deve haver um equilíbrio entre definir a estrutura mais básica (menos agrupamentos) e ainda conseguir o nível necessário de similaridade (ou distância euclidiana entre as observações) dentro dos agrupamentos.

Um método hierárquico resulta em diversas soluções de agrupamentos, nesse caso, começando com uma solução de sete agrupamentos e terminando com um. Permanece a dúvida de qual solução devemos escolher.

Atenção

Sabemos que, quando nos afastamos de agrupamentos unitários na solução de sete agrupamentos, a heterogeneidade aumenta. Portanto, por que não ficarmos com o maior número de agrupamentos possível, a opção mais homogênea possível?

O problema é que não definimos nenhuma estrutura com sete agrupamentos. Assim, devemos verificar cada solução quanto à sua descrição da estrutura versus a heterogeneidade dos agrupamentos. Primeiro,

discutimos um método simples para definir heterogeneidade de cada solução de agrupamento e então avaliamos as soluções para chegarmos a uma solução final.

Qualquer medida de heterogeneidade de uma solução de agrupamento deve representar a diversidade geral entre observações em todos os agrupamentos. Na solução inicial de uma abordagem aglomerativa, em que todas as observações estão em agrupamentos separados, a heterogeneidade é minimizada.

À medida que observações são combinadas para formarem agrupamentos, a heterogeneidade aumenta. Assim, a medida de heterogeneidade deve começar com um valor nulo e aumentar para mostrar o nível de heterogeneidade quando agrupamentos são combinados.

> Lembre-se que estamos tentando obter a estrutura mais simples possível que ainda represente agrupamentos homogêneos. Se monitoramos a medida de heterogeneidade conforme o número de agrupamentos diminui, grandes aumentos na heterogeneidade indicam que dois agrupamentos um tanto dissimilares foram unidos naquele estágio.

## Exemplo de como calcular os clusters

No vídeo a seguir, confira na prática como calcular os *custers.*

Conteúdo interativo

Acesse a versão digital para assistir ao vídeo.

Uma vez que temos os procedimentos para tratar de cada questão, podemos executar a análise. Ilustramos os princípios inerentes a cada uma dessas questões por meio de nosso exemplo simples.´

Suponha que um pesquisador de economia queira determinar segmentos de mercado em um grupo de pessoas com base em padrões de renda e idade dos indivíduos. Uma pequena amostra de seis respondentes é selecionada como um teste piloto de como a análise de agrupamentos é aplicada.

Duas medidas de segmentação – V1 (renda) e V2 (idade) – foram selecionadas para cada respondente. Os valores para cada um dos seis respondentes são mostrados na tabela a seguir:

| Indivíduo | Renda | Idade |
|---|---|---|
| A | 9 | 28 |
| B | 8 | 31 |
| C | 2 | 42 |
| D | 18,2 | 38 |
| E | 3,9 | 25 |
| F | 6,4 | 41 |

Renda e idade de seis indivíduos
Manoela Gonçalves Cabo

O método aglomerativo segue um processo simples e repetitivo:

- Começar com todas as observações formando seus próprios agrupamentos (ou seja, cada observação forma um agrupamento unitário), de forma que o número de agrupamentos seja igual ao de observações.
- Usando a medida de similaridade, combinar os dois agrupamentos mais parecidos em um novo (agora contendo duas observações), reduzindo assim a quantia de agrupamentos em uma unidade.
- Repetir o processo novamente, usando medida de similaridade para combinar os dois agrupamentos mais parecidos em um novo.
- Continuar este processo, combinando em cada passo os dois agrupamentos mais semelhantes em um novo. Repetir o processo em um total de n – 1 vezes até que todas as observações estejam contidas em um só agrupamento.

Utilizando a distância euclidiana entre dois elementos $X_I$ e $X_k, |\neq k$ , para calcular a matriz de similaridade, temos:

$$d\left(X_l, X_k\right)=\left[\left(X_l-X_k\right)'\left(X_l-X_k\right)\right]^{\frac{1}{2}}=\left[\sum_{i=1}^{p}\left(X_{il}-X_{ik}\right)^2\right]^{\frac{1}{2}},$$
$$d(A,B)=\left[(9-8)^2+(28-31)^2\right]=\left[(1)^2+(-3)^2\right]=10$$

Fazendo a distância para todos os indivíduos, termos a primeira matriz de similaridade:

| A | B | C | D | E | F |
|---|---|---|---|---|---|
| A | | | | | |
| B | 10,00 | | | | |
| C | 245,00 | 157,00 | | | |
| D | 184,64 | 153,04 | 278,44 | | |
| E | 35,01 | 52,81 | 292,61 | 373,49 | |
| F | 175,76 | 102,56 | 20,36 | 148,24 | 262,25 |

Manoela Gonçalves Cabo

Combinar os dois agrupamentos A e B mais parecidos em um novo (agora contendo duas observações), reduzindo assim a quantia de agrupamentos:

| | AB | C | D | E | F |
|---|---|---|---|---|---|
| **AB** | | | | | |
| **C** | 198,50 | | | | |
| **D** | 166,34 | 278,44 | | | |

| | AB | C | D | E | F |
|---|---|---|---|---|---|
| E | 41,41 | 292,61 | 373,49 | | |
| F | 136,66 | 20,36 | 148,24 | 262,25 | |

Manoela Gonçalves Cabo

Combinar os dois agrupamentos C e F mais parecidos em um novo, reduzindo assim a quantia de agrupamentos:

| | AB | CF | D | E |
|---|---|---|---|---|
| AB | | | | |
| CF | 162,49 | | | |
| D | 166,34 | 208,25 | | |
| E | 41,41 | 272,34 | 373,49 | |

Manoela Gonçalves Cabo

Combinar os dois agrupamentos AB e E mais parecidos em um novo:

| | ABE | D | CF |
|---|---|---|---|
| ABE | | | |
| D | 226,19 | | |
| CF | 189,90 | 208,25 | |

Manoela Gonçalves Cabo

Continuar este processo, combinando em cada passo os dois agrupamentos mais semelhantes em um novo. Repetir o processo até que todas as observações estejam contidas em um só agrupamento:

| | ABECF | D |
|---|---|---|
| ABECF | | |
| D | 173,44 | |

Manoela Gonçalves Cabo

Faz-se então o histórico de agrupamento e calcula-se a distância elevando a soma ao quadrado

$$\left[\sum_{i=1}^{p}\left(X_{il}-X_{ik}\right)^2\right]^{\frac{1}{2}}$$

Continuar este processo, combinando em cada passo os dois agrupamentos mais semelhantes em um novo. Repetir o processo até que todas as observações estejam contidas em um só agrupamento:

| Passo | No. Grupos | Fusão | Distância |
|---|---|---|---|
| 1 | 5 | {A} e {B} | 3,16 |
| 2 | 4 | {C} e {F} | 4,51 |
| 3 | 3 | {AB} e {E} | 6,44 |
| 4 | 2 | {ABE} e {CF} | 13,78 |
| 5 | 1 | {ABECF} e {D} | 13,17 |

Manoela Gonçalves Cabo

A escolha da quantidade de agrupamentos ou clusters depende do pesquisador; se escolhermos 3 grupos, teremos [ABE, CF, D].

## Verificando o aprendizado

Questão 1

Descreva o conceito de análise de agrupamento:

### Chave de resposta

A análise de agrupamentos, também conhecida como análise de conglomerados, classificação ou cluster, tem como objetivo dividir os elementos da amostra, ou população, em grupos, de forma que os elementos pertencentes a um mesmo grupo sejam similares entre si com respeito às variáveis (caraterística) que neles foram medidas, e os elementos em grupos diferentes sejam heterogêneos em relação a estas mesmas características.

A análise de agrupamento é uma técnica de interdependência, porém, está concentrada somente na definição de estrutura, avaliando a interdependência sem quaisquer relações de dependência associadas, como visto nas técnicas apresentadas nas aulas anteriores. Nenhuma das técnicas de interdependência definirá a estrutura para otimizar ou maximizar uma relação de dependência.

É nossa tarefa primeiramente utilizar esses métodos na identificação de estrutura e então empregá-la onde for apropriado. Os objetivos de relações de dependência não são “incorporados” nesses métodos de interdependência – eles avaliam a estrutura para seus próprios objetivos, e nenhum outro.

Questão 2

Suponha que tenha disponível um conjunto de dados constituído de elementos amostrais, tendo-se medido p-variáveis aleatórias em cada um deles. Para cada elemento amostral j, tem-se, portanto, o vetor de medidas $X_j$ definido por:

$$< br > X_j = [X_{1j}, X_{2j} \dots X_{pj}]' , j = 1, 2, \dots, n < br >$$

Quais as medidas de similaridade mais utilizadas na análise de agrupamentos? Descreva como se calcula uma delas.

Chave de resposta

Existem várias medidas diferentes e cada uma delas produz um determinado tipo de agrupamento.

Algumas medidas mais comuns, apropriadas para variáveis quantitativas, são: Distância Euclidiana, Distância generalizada ou ponderada, Distância de Minkowisky, Coeficiente de concordância simples, Coeficiente de concordância positiva, Coeficiente de concordância de Jaccard, Distância Euclidiana média, entre outras.

A distância euclidiana entre dois elementos $X_l$ e $X_k$, l ≠ k, é definida por:

$$d(X_l, X_k) = \left[(X_l - X_k)'(X_l - X_k)\right]^{\frac{1}{2}} = \left[\sum_{i=1}^{p}(X_{il} - X_{ik})^2\right]^{\frac{1}{2}}, j = 1, 2, \ldots, n$$

## Questão 3

Suponha que um pesquisador queira determinar segmentos de alunos de uma universidade em um grupo de pessoas com base em padrões da média de suas notas e idade dos alunos. Uma pequena amostra de seis alunos é selecionada como um teste piloto de como a análise de agrupamentos é aplicada.

Duas medidas de segmentação – V1 (média de notas) e V2 (idade) – foram selecionadas para cada respondente. Os valores para cada um dos seis respondentes são mostrados na tabela a seguir:

| Indivíduo | Renda | Idade |
|---|---|---|
| A | 9 | 28 |
| B | 8 | 31 |
| C | 2 | 42 |
| D | 10 | 38 |
| E | 3 | 25 |
| F | 6 | 41 |

Manoela Gonçalves Cabo

Calcule a primeira matriz de similaridade, utilizado a distância euclidiana.

Chave de resposta

Utilizando a distância euclidiana entre dois elementos $X_l$ e $X_k$, l ≠ k, para calcular a matriz de similaridade, temos:

$$d\left(X_l, X_k\right) = \left[\left(X_l - X_k\right)' \left(X_l - X_k\right)\right]^{\frac{1}{2}} = \left[\sum_{i=1}^{p} \left(X_{il} - X_{ik}\right)^2\right]^{\frac{1}{2}},$$

$$d(A, B) = \left[(9-8)^2 + (28-31)^2\right] = \left[\left[(1)^2 + (-3)^2\right] = 10\right.$$

$$d(A, B) = \left[(9-2)^2 + (28-42)^2\right] = \left[\left[(7)^2 + (-14)^2\right] = 245\right.$$

Fazendo a distância para todos os indivíduos, teremos a primeira matriz de similaridade:

| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| **A** | | | | | | |
| **B** | 10,00 | | | | | |
| **C** | 245,00 | 157,00 | | | | |
| **D** | 101,00 | 53,00 | 80,00 | | | |
| **E** | 45,00 | 61,00 | 290,00 | 218,00 | | |
| **F** | 178,00 | 104,00 | 17,00 | 25,00 | 265,00 | |

Os alunos A e B deveriam ser agrupados, pois são mais parecidos.

# Considerações finais

## O que você aprendeu neste conteúdo?

- A definição da análise conjunta como técnica de dependência voltada à avaliação de objetos.
- A identificação dos usos gerenciais da análise conjunta em contextos organizacionais e de mercado.
- A comparação da análise conjunta com outros métodos multivariados.
- O planejamento de um experimento de análise conjunta como etapa fundamental da aplicação prática.
- A definição da análise de agrupamentos (clustering) como técnica de interdependência.
- A explicação do funcionamento da análise de agrupamentos e de seu processo de decisão.
- A apresentação de exemplos práticos de clustering com comandos no software R.
- A integração das técnicas de análise conjunta e clustering na tomada de decisão baseada em dados.

## Explore +

- Pesquise na internet, sites, vídeos e artigos relacionados ao conteúdo visto.
- Em caso de dúvidas, converse com seu professor online por meio dos recursos disponíveis no ambiente de aprendizagem.
- No site da UFMG, acesse o Manual de introdução ao R com exemplos práticos de aplicação.

## Referência

HAIR JR., J.F. *et al*. **Análise Multivariada de Dados**, 6.ed., Porto Alegre, Bookman, 2009.

MINGOTI, S. **Análise de dados através de métodos de estatística multivariada**. Belo Horizonte: Editora UFMG, 2013.