MBA USP ESALQ

# UNSUPERVISED MACHINE LEARNING: Análise de Correspondência Simples e Múltipla

Prof. Dr. Wilson Tarantin Junior

*A responsabilidade pela idoneidade, originalidade e licitude dos conteúdos didáticos apresentados é do professor. **Proibida a reprodução,** total ou parcial, sem autorização. Lei nº 9610/98

# Contextualização

- Quando aplicar a análise de correspondência?
  - Técnica adequada para a análise de variáveis categóricas (**qualitativas**)
  - O objetivo é verificar se existe **associação** estatisticamente significativa entre as variáveis e suas categorias, criando o **mapa perceptual** para visualizar as associações
  - Caso exista uma variável quantitativa, é necessário que ela passe por um processo de categorização previamente
    - Por exemplo: a idade é uma variável quantitativa (25, 42, 73, 81 anos) e poderia ser categorizada como: 0-30 anos é a categoria 1, 31-60 anos é categoria 2, 61-90 anos é categoria 3...

# Contextualização

- Trata-se de técnica exploratória (não supervisionada)
  - Para avaliar a relação conjunta entre as variáveis
  - Não há modelos do tipo "$y_i = x_{1i} + x_{2i} + ... + u_i$"
  - Não são adequadas para fins de inferência
  - Se novas observações forem adicionadas ao banco de dados, é adequado refazer a análise

# Contextualização

- Exemplos de aplicação
  - Faixa de renda e status na aprovação de crédito
  - Nível de escolaridade e cargo ocupado em empresas
  - Tipo de solo e cultura implementada
  - Gravidade dos sintomas da doença e comorbidades
  - Outros...

# Contextualização

- Análise de variáveis geradas por escala Likert
    - Exemplos: concordo plenamente; concordo parcialmente; não concordo, nem discordo; discordo parcialmente; discordo plenamente
    - Evita o problema da ponderação arbitrária
    - Cada ponto da escala Likert torna-se uma categoria da variável na análise de correspondência simples ou múltipla

# Implementação
# Análise de Correspondência Simples

# Análise de Correspondência Simples

- Também conhecida como Anacor
  - Quando o objetivo é estudar a **associação entre duas variáveis e suas categorias**
  - É possível separar a Anacor em duas partes:

    **1.** Análise da significância estatística da associação entre as variáveis e suas categorias por meio do teste qui-quadrado ($\chi^2$)

    **2.** Elaboração e interpretação do mapa perceptual

# 1. Análise da significância estatística (teste qui-quadrado)

# Análise de Correspondência Simples

## 1. Tabela de contingência

- Contém as frequências absolutas observadas para cada par de categorias das variáveis
- Trata-se de uma tabela de classificação cruzada (*cross-tabulation*)

| | | Variável B | | | | | |
|---|---|---|---|---|---|---|---|
| | | **Categoria 1** | **Categoria 2** | **Categoria 3** | **...** | **Categoria J** | **Total** |
| **Variável A** | **Categoria 1** | $n_{11}$ | $n_{12}$ | $n_{13}$ | ... | $n_{1J}$ | $\Sigma_{L1}$ |
| | **Categoria 2** | $n_{21}$ | $n_{22}$ | $n_{23}$ | ... | $n_{2J}$ | $\Sigma_{L2}$ |
| | **Categoria 3** | $n_{31}$ | $n_{32}$ | $n_{33}$ | ... | $n_{3J}$ | $\Sigma_{L3}$ |
| | **...** | ... | ... | ... | ... | ... | ... |
| | **Categoria I** | $n_{I1}$ | $n_{I2}$ | $n_{I3}$ | ... | $n_{IJ}$ | $\Sigma_{LI}$ |
| | **Total** | $\Sigma_{C1}$ | $\Sigma_{C2}$ | $\Sigma_{C3}$ | ... | $\Sigma_{CJ}$ | $N$ |

# Análise de Correspondência Simples

## 2. Tabela de frequências absolutas esperadas

- Para a célula referente às categorias 1 das duas variáveis, a frequência absoluta esperada é:

$$\textit{Freq. absoluta esperada} = \frac{(\Sigma_{L1} \times \Sigma_{C1})}{N}$$

- Este mesmo cálculo deve ser realizado para cada par de categorias das variáveis, alterando-se apenas o numerador

# Análise de Correspondência Simples

## 3. Tabela de resíduos

- Para a célula referente às categorias 1 das duas variáveis, o valor do resíduo é:

$$Resíduo = n_{11} - \frac{(\Sigma_{L1} \times \Sigma_{C1})}{N}$$

- **Ou seja, resíduo = frequência absoluta observada – frequência absoluta esperada**
- O mesmo cálculo é realizado para cada par de categorias

# Análise de Correspondência Simples

4. Tabela com os valores $\chi^2$

- Para a célula referente às categorias 1 das duas variáveis, o valor da estatística $\chi 2$ é:

$$\chi 2 = \frac{(\ resíduo_{11}\ )^2}{(\ freq.\ absoluta\ esperada_{11}\ )}$$

- O mesmo cálculo é realizado para cada par de categorias e, em seguida, os valores de todas as células são somados

# Análise de Correspondência Simples

## 4. Tabela com os valores $\chi^2$

- O objetivo é verificar se há associação estatisticamente significante entre as variáveis (utilizando a soma do $\chi 2$)
  - $H_0$: as variáveis se associam de forma aleatória.
  - $H_1$: a associação entre as variáveis não se dá de forma aleatória.
- Dados o nível de significância e os graus de liberdade, se o valor da estatística $\chi^2$ for maior do que seu valor crítico, há associação significante entre as duas variáveis ($H_1$)
  - Graus de liberdade = (I – 1) x (J – 1)

# Análise de Correspondência Simples

5. Tabela de resíduos padronizados (e padronizados ajustados)

- Enquanto a análise do $\chi^2$ permite verificar se há ou não a dependência entre as duas variáveis, a análise de resíduos padronizados ajustados permite aprofundar a análise com foco nas categorias das variáveis

- Como as categorias de uma variável se relacionam com as categorias da outra variável?

  - Para tanto, observa-se o excesso ou falta de ocorrência de casos nas categorias das duas variáveis

# Análise de Correspondência Simples

## 5. Tabela de resíduos padronizados

- Para a célula referente às categorias 1 das duas variáveis, o valor do resíduo padronizado é:

$$Resíduo_{padronizado} = \frac{resíduo_{11}}{\sqrt{(freq.\ absoluta\ esperada_{11})}}$$

- O mesmo cálculo é realizado para cada par de categorias

# Análise de Correspondência Simples

5. Tabela de resíduos padronizados ajustados

- Para a célula referente às categorias 1 das duas variáveis, o valor do resíduo padronizado ajustado é:

$$Resíduo_{padronizado\ ajustado} = \frac{resíduo\ padronizado_{11}}{\sqrt{\left[\left(1-\frac{\Sigma_{C1}}{N}\right) x \left(1-\frac{\Sigma_{L1}}{N}\right)\right]}}$$

  - O mesmo cálculo é realizado para cada par de categorias

# Análise de Correspondência Simples

5. Tabela de resíduos padronizados ajustados

- Se o valor do resíduo padronizado ajustado em certa célula for maior do que **1,96**, interpreta-se que existe associação significativa, ao nível de significância de **5%**, entre as duas categorias que interagem na célula; se for menor do que 1,96, não há associação estatisticamente significativa
    - A referência de 1,96 é o valor crítico da normal padrão para o nível de significância de 5%

# 2. Elaboração e interpretação do mapa perceptual

# Análise de Correspondência Simples

1. Determinar os autovalores ($\lambda^2$)

- A quantidade ($m$) de autovalores depende da quantidade de categorias nas variáveis: $m = \text{mín}(I - 1, J - 1)$

- Na Anacor, os autovalores referem-se às inércias principais parciais e são base para determinar a inércia principal total e o percentual da inércia principal total em cada dimensão do mapa perceptual

# Análise de Correspondência Simples

1. Determinar os autovalores ($\lambda^2$)

- Como base para o cálculo dos autovalores, inicialmente, define-se uma matriz A
  - Um modo de obter a matiz A, baseando-se nas etapas anteriores, é fazer para cada célula da matriz de resíduos padronizados o seguinte cálculo:

$$\frac{(\ Resíduo_{padronizado}\ )}{\sqrt{N}}$$

- Com base na matriz A, obtém-se a matriz W: **W = A' . A**

# Análise de Correspondência Simples

1. Determinar os autovalores ($\lambda^2$)

- Identificando **W**, os autovalores são obtidos pela solução da seguinte expressão: $\det(\lambda^2 \,.\, \mathbf{I} - \mathbf{W}) = 0$

$$\begin{vmatrix} \lambda^2 - w_{11} & - w_{12} & - w_{13} \\ - w_{21} & \lambda^2 - w_{22} & - w_{23} \\ - w_{31} & - w_{32} & \lambda^2 - w_{33} \end{vmatrix} = 0$$

  - **I** é a matriz identidade

# Análise de Correspondência Simples

1. Determinar os autovalores ($\lambda^2$)

- Com base nos autovalores ($\lambda^2$), encontra-se o percentual da inércia principal total de cada dimensão

$$\textit{\% da Inércia Principal Total (para cada dimensão)} = \frac{\lambda^2_{\textit{dimensão}}}{\lambda^2_{\textit{total}}}$$

- Quanto maior a inércia principal total (e o $\chi^2$), mais forte será a associação entre as variáveis em análise

$$\textit{Inércia Principal Total} = \frac{\chi^2}{N}$$

# Análise de Correspondência Simples

2. Determinar as massas em linha e coluna

- As **massas** representam a influência que cada categoria exerce sobre as demais categorias de sua variável, seja na coluna (*column profiles*) ou linha (*row profiles*)

- Com base nos "totais" da tabela de contingência, para a categoria 1 das variáveis, obtém-se as massas médias:

$$\textit{Massa na coluna} = \frac{\Sigma_{L1}}{N} \qquad \textit{Massa na linha} = \frac{\Sigma_{C1}}{N}$$

  - O mesmo cálculo é realizado para as demais categorias

# Análise de Correspondência Simples

## 3. Determinar os autovetores

- Para a matriz **W**, é possível encontrar os autovetores a partir dos autovalores ($\lambda^2$) já calculados
  - Substituindo os autovalores de cada dimensão na matriz definida como **$\det(\lambda^2 \cdot I - W) = 0$** e resolvendo o sistema de equações que parte dela, é possível encontrar os autovetores da coluna (V) e, com base neles, encontrar os autovetores da linha (U)

$$u'_k = \underbrace{[\, D_l^{-1/2} \cdot (P - Ic') \cdot D_c^{-1/2} \,]} \cdot v'_k \cdot \lambda_k^{-1}$$

**Trata-se da matriz A**

# Análise de Correspondência Simples

4. Determinar as coordenadas das categorias

- Variável **em linha** na tabela de contingência
  - Coordenadas das abscissas (X)

$$X_l = \sqrt{\lambda_1} \cdot D_l^{-1/2} \cdot u_1$$

  - Coordenadas das ordenadas (Y)

$$Y_l = \sqrt{\lambda_2} \cdot D_l^{-1/2} \cdot u_2$$

  - Coordenadas da K-ésima dimensão (k = quantidade de $\lambda$)

$$Z_l = \sqrt{\lambda_k} \cdot D_l^{-1/2} \cdot u_k$$

# Análise de Correspondência Simples

4. Determinar as coordenadas das categorias

- Variável **em coluna** na tabela de contingência
  - Coordenadas das abscissas (X)

$$X_c = \sqrt{\lambda_1} \cdot D_c^{-1/2} \cdot v_1$$

  - Coordenadas das ordenadas (Y)

$$Y_c = \sqrt{\lambda_2} \cdot D_c^{-1/2} \cdot v_2$$

  - Coordenadas da K-ésima dimensão (k = quantidade de $\lambda$)

$$Z_c = \sqrt{\lambda_k} \cdot D_c^{-1/2} \cdot v_k$$

# Análise de Correspondência Múltipla (MCA)

# Análise de Correspondência Múltipla

- Associação entre mais de duas variáveis
  - Só participam da ACM as variáveis que apresentam associação estatisticamente significativa com pelo menos uma outra variável contida na análise
  - Antes de elaborar a ACM, é importante realizar um teste $\chi^2$ para cada par de variáveis
  - Se alguma delas não apresentar associação com outras, não é incluída na análise de correspondência

# Análise de Correspondência Múltipla

1º método: utilizando a **Matriz Binária**

- A matriz binária é obtida pela transformação das variáveis qualitativas em variáveis binárias, ou seja, valores 0 ou 1
- Com base na matriz binária (Z), pode ser obtida a inércia principal total na ACM
- Supondo que a matriz binária (Z) seja semelhante a uma tabela de contingência da Anacor, é possível obter a inércia principal parcial das dimensões, autovalores, autovetores e coordenadas dessa matriz

# Análise de Correspondência Múltipla

## 1º método: utilizando a **Matriz Binária**

- Um exemplo de matriz binária (Z) é:

| Obs. | Variável A | | Variável B | | | Variável C | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Categ. 1 | Categ. 2 | Categ. 1 | Categ. 2 | Categ. 3 | Categ. 1 | Categ. 2 | Categ. 3 | Categ. 4 |
| **1** | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 |
| **2** | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 0 |
| **3** | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 0 |
| **N** | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 |

- **Quantidade de dimensões ($\lambda^2$) = J – Q = (9 – 3) = 6**, em que "**J**" é a quantidade total de categorias e **"Q"** a quantidade de variáveis

# Análise de Correspondência Múltipla

2º método: utilizando a **Matriz de Burt**

- A matriz de Burt é definida como: **B = Z' . Z**
  - É possível combinar em uma única matriz as tabelas de contingência com o cruzamento de todos os pares variáveis
  - Ao considerar a matriz de Burt uma tabela de contingência, é possível realizar uma Anacor e obter as coordenadas das categorias

# Referência

Fávero, Luiz Paulo; Belfiore, Patrícia. (2017). Manual de análise de dados: estatística e modelagem multivariada com Excel®, SPSS® e Stata®. Rio de Janeiro: Elsevier