DATA SCIENCE & ANALYTICS

MBA

ANACOR

0000100

Table of Contents

0. QUALI X QUANTI

Técnicas de análise de dados diferem significativamente entre variáveis categóricas (quali) e numéricas (quanti), portanto diferenciar entre elas é crucial para escolher os métodos estatísticos apropriados para análise.

0.1. VARIÁVEIS QUALITATIVAS/CATEGÓRICAS

As variáveis qualitativas, também conhecidas como variáveis categóricas, são aquelas que descrevem características ou atributos que não podem ser medidos numericamente. Elas são usadas para categorizar ou classificar indivíduos ou itens em diferentes grupos ou categorias.

0.1.1. EXEMPLOS DE VARIÁVEIS QUALITATIVAS

- Classificação em Faixas de Idade: Quando a idade é agrupada em categorias como 0-18, 19-35, 36-50, etc., ela se torna uma variável qualitativa. Cada faixa etária representa uma categoria distinta.

- Respostas Sim ou Não em Questionários: Respostas binárias como "sim" ou "não" são variáveis qualitativas, pois representam duas categorias distintas.

- Bairro do Imóvel: O bairro em que um imóvel está localizado é uma variável qualitativa, pois é uma categoria que descreve a localização sem envolver uma medida numérica.

- Países onde se localizam as empresas: Representa diferentes países como Brasil, Estados Unidos, Japão, etc. Cada país é uma categoria distinta.

- Grau de Escolaridade das Pessoas: Níveis de escolaridade como ensino fundamental, ensino médio, ensino superior, etc. Cada nível é uma categoria distinta.

- Marca do Veículo Seminovo Destinado à Venda: Marcas de veículos como Ford, Chevrolet, Toyota, etc. Cada marca é uma categoria distinta.

0.1.2. Tipos de Variáveis Categóricas

- Nominais: As categorias não têm uma ordem intrínseca. Exemplo: países, marcas de carros.

- Ordinais: As categorias têm uma ordem intrínseca. Exemplo: níveis de escolaridade (fundamental, médio, superior).

0.1.3. Análise de Variáveis Qualitativas

As variáveis qualitativas são frequentemente analisadas usando tabelas de contingência, gráficos de barras e testes estatísticos como o qui-quadrado.

0.2. Variáveis Quantitativas

As variáveis quantitativas são aquelas que podem ser medidas numericamente. Elas expressam a quantidade, magnitude ou tamanho de algo e têm uma escala contínua ou discreta.

0.2.1. Exemplo de Variável Quantitativa

- Preço em Reais de um Produto: O preço de um produto é uma variável quantitativa, pois é uma medida numérica que representa o valor monetário do produto.

0.2.2. Análise de Variáveis Quantitativas

As variáveis quantitativas são analisadas usando métodos como média, mediana, desvio padrão, gráficos de dispersão e testes como o t-test e a análise de regressão.

Exemplos de Análise

- Pesquisa de Mercado: Variáveis qualitativas como preferência de marca (qualitativa) e variáveis quantitativas como quantidade de gasto (quantitativa) são analisadas para entender o comportamento do consumidor.

- Saúde Pública: Estudos epidemiológicos podem usar variáveis qualitativas como status de fumo (sim/não) e variáveis quantitativas como número de cigarros fumados por dia.

- Educação: Avaliações educacionais podem incluir variáveis qualitativas como nível de satisfação com o ensino (categórica) e variáveis quantitativas como pontuação em testes (numérica).

Em resumo, compreender a diferença entre variáveis qualitativas e quantitativas é fundamental para a análise de dados. As variáveis qualitativas categorizam e descrevem características, enquanto as variáveis quantitativas medem e quantificam. A correta identificação e análise dessas variáveis permitem insights precisos e informados em diversas áreas de pesquisa e aplicação.

1. INTRODUÇÃO ANACOR

A análise de correspondência é uma técnica estatística utilizada para explorar e visualizar associações entre variáveis categóricas. Dentro do aprendizado de máquina não supervisionado, esta técnica é especialmente valiosa para dados qualitativos, permitindo a criação de mapas perceptuais que ilustram as relações entre categorias.

UNSUPERVISED MACHINE LEARNING

Análise de Correspondência Simples e Múltipla

A análise de correspondência é uma técnica estatística utilizada para explorar e analisar dados categóricos, revelando associações e estruturas subjacentes entre categorias de variáveis. Este método é particularmente útil em cenários onde não se possui um conhecimento prévio sobre as relações entre variáveis, enquadrando-se no âmbito do aprendizado não supervisionado.

A análise de correspondência, tanto simples quanto múltipla, é uma técnica estatística dentro do campo do aprendizado de máquina não supervisionado, que se destaca pela sua aplicabilidade na análise de variáveis categóricas. Essa técnica é especialmente útil quando se deseja verificar a existência de associações estatisticamente significativas entre variáveis e suas respectivas categorias.

1.1. Características Principais

- Técnica Exploratória:

A análise de correspondência não requer a especificação de variáveis dependentes e independentes. É usada para explorar dados e identificar padrões ou associações sem fazer suposições prévias sobre as relações entre variáveis.

- Associações Entre Categorias:

Analisar se há associações estatisticamente significativas entre as categorias das variáveis. Os resultados são representados em mapas perceptuais, onde a proximidade entre categorias indica a força da associação.

- Refazendo a Análise quando novos dados:

Como é uma técnica exploratória, é importante refazer a análise quando novas observações são adicionadas aos dados. Isso garante que as associações identificadas sejam atualizadas de acordo com os novos dados.

- Visualização Intuitiva (Mapas Perceptuais):

Os mapas perceptuais gerados pela análise de correspondência oferecem uma visualização intuitiva das associações entre categorias, facilitando a interpretação dos dados e a identificação de padrões.

- Objetivo:

Identificar associações estatisticamente significativas entre categorias de variáveis e representar essas associações visualmente.

- Tipo de Dados:

Aplica-se a dados categóricos (variáveis qualitativas). Se houver variáveis quantitativas, estas devem ser categorizadas previamente.

1.2. Processo ANACOR

1. Construção da Tabela de Contingência:
    - Coleta de dados categóricos e construção de uma tabela que mostra as frequências de cada combinação de categorias.

2. Cálculo das Frequências Esperadas:
    - Cálculo das frequências esperadas sob a hipótese de independência.

3. Cálculo dos Resíduos:
    - Determinação das diferenças entre frequências observadas e esperadas, padronizadas para avaliar a significância.

4. Geração do Mapa Perceptual:
    - Uso de métodos como a decomposição de valores singulares (SVD) para projetar os dados em um espaço de menor dimensão.

1.3. Quando Aplicar a Análise de Correspondência?

A análise de correspondência deve ser aplicada em contextos onde se busca explorar as relações entre variáveis categóricas (dados qualitativos), pois permite a criação de mapas perceptuais que visualizam essas associações de maneira clara e intuitiva.

A decisão de aplicar a análise de correspondência depende do tipo de dados e das perguntas de pesquisa específicas.

1. Exploração de Dados Categóricos: Quando os dados disponíveis são categóricos e se deseja explorar as associações entre categorias.

2. Simplificação de Dados Complexos: Útil para simplificar e visualizar dados complexos de múltiplas variáveis em um espaço de menor dimensão.

3. Revelação de Estruturas Subjacentes: Adequada para identificar padrões e relações que não são imediatamente aparentes em uma tabela de contingência.

4. Análise de Preferências ou Percepções: Comumente usada em pesquisas de mercado para analisar preferências e percepções de consumidores sobre produtos ou serviços.

5. Dados de Pesquisas e Questionários: Frequente em ciências sociais, onde dados de pesquisas e questionários são categorizados e analisados para identificar correlações.

1.4. Técnica Exploratória

A análise de correspondência é uma técnica exploratória, ou seja, não supervisionada, focada em avaliar a interdependência entre variáveis. Diferente de modelos de regressão, não se baseia em fórmulas predefinidas.

Por ser exploratória, não é adequada para inferências, e qualquer nova observação adicionada ao banco de dados requer uma nova execução da análise para atualizar os resultados.

1.5. Exemplos de Aplicação

Esta técnica pode ser aplicada em diversos cenários, tais como:

- Faixa de renda e status na aprovação de crédito: Avaliar como diferentes faixas de renda influenciam a probabilidade de aprovação de crédito.

- Nível de escolaridade e cargo ocupado em empresas: Explorar a relação entre o nível educacional dos funcionários e os cargos que ocupam.

- Tipo de solo e cultura implementada: Analisar a correspondência entre tipos de solo e as culturas agrícolas que neles prosperam.

- Gravidade dos sintomas de uma doença e comorbidades: Estudar a associação entre a gravidade dos sintomas de uma doença e a presença de comorbidades.

1.6. Análise de Variáveis Geradas por Escala Likert

A análise de correspondência é também eficaz para variáveis geradas por escalas Likert, frequentemente utilizadas em pesquisas de opinião.

Exemplos de pontos de uma escala Likert incluem:

- "concordo plenamente";
- "concordo parcialmente";
- "não concordo nem discordo";
- "discordo parcialmente";
- "discordo plenamente".

Cada um desses pontos se torna uma categoria na análise de correspondência simples ou múltipla, evitando assim o problema de ponderação arbitrária que pode surgir em outras técnicas de análise.

1.7. Categorização de Variáveis Quantitativas

- Definição de Intervalos ou Categorias: O primeiro passo é definir intervalos ou categorias que a variável quantitativa será transformada. Por exemplo, uma variável de idade pode ser categorizada em faixas etárias como 0-20, 21-40, 41-60, e 61-80 anos.

- Atribuição de Categorias: Cada valor da variável quantitativa é atribuído a uma das categorias definidas. Por exemplo, uma pessoa de 25 anos seria categorizada na faixa etária de 21-40 anos.

- Inclusão na Análise de Correspondência: A variável agora categorizada pode ser incluída na análise de correspondência, permitindo que suas categorias sejam analisadas em relação às categorias de outras variáveis qualitativas.

1.7.1. Vantagens da Categorização:

- Uniformização dos Dados: Transforma todos os dados para um formato categórico, adequado para análise de correspondência.

- Facilidade de Interpretação: Facilita a interpretação dos resultados, pois os dados são representados em categorias discretas.

2. APLICAÇÃO ANACOR

2.1. TABELA DE CONTINGÊNCIA (CLASSIFICAÇÃO CRUZADA)

A tabela de contingência, também conhecida como tabela de classificação cruzada, é uma ferramenta essencial para a análise exploratória de dados categóricos, oferecendo uma representação clara das frequências observadas e servindo como base para a investigação de associações entre variáveis.

A tabela de contingência resume a relação entre duas ou mais variáveis categóricas, apresentando as frequências absolutas observadas das combinações de categorias das variáveis, proporcionando uma visão clara e estruturada dos dados.

2.1.1. ESTRUTURA DA TABELA DE CONTINGÊNCIA

1. Linhas e Colunas:
    - As categorias de uma variável são listadas nas linhas, enquanto as categorias da outra variável são listadas nas colunas. As células da tabela contêm as frequências absolutas observadas para cada combinação de categorias.

2. Frequências Absolutas Observadas:
    - A tabela de contingência mostra quantas vezes cada combinação de categorias ocorre nos dados. Por exemplo, se estamos analisando a relação entre gênero (masculino, feminino) e preferência de produto (A, B, C), a tabela mostrará quantos homens preferem o produto A, quantos preferem o produto B, e assim por diante.

2.1.2. CARACTERÍSTICAS DA TABELA DE CONTINGÊNCIA:

- Análise de Associação:

A tabela de contingência é a base para várias análises estatísticas que investigam a associação entre variáveis categóricas, como o teste qui-quadrado de independência. Este teste avalia se as frequências observadas diferem significativamente das frequências esperadas sob a hipótese de independência.

- Resíduos e Valores Qui-Quadrado:

Embora a tabela de contingência mostre apenas as frequências observadas, os resíduos (diferenças entre frequências observadas e esperadas) e os valores qui-quadrado podem ser calculados posteriormente para avaliar a significância e o ajuste do modelo.

2.1.3. Vantagens da Tabela de Contingência:

- Simplicidade e Clareza: A tabela de contingência é fácil de interpretar e fornece uma visão clara das relações entre variáveis categóricas.
- Base para Análises Avançadas: Serve como ponto de partida para análises estatísticas mais complexas, como modelos log-lineares e análise de correspondência.

2.2. Estatística Qui-Quadrado e P-Valor em Tabelas de Contingência

A estatística qui-quadrado é uma medida usada para avaliar a independência entre duas variáveis categóricas em uma tabela de contingência. O p-valor é resultado do teste qui-quadrado.

O teste qui-quadrado é uma ferramenta estatística essencial para testar a independência entre duas variáveis categóricas em uma tabela de contingência. O p-valor obtido no teste qui-quadrado ajuda a determinar se as diferenças entre as frequências observadas e as frequências esperadas nas células da tabela são suficientemente grandes para rejeitar a hipótese nula de independência entre as variáveis.

2.2.1. Cálculo da Estatística Qui-Quadrado

- Frequências Observadas (O): As contagens reais de cada combinação de categorias das duas variáveis;
- Frequências Esperadas (E): As contagens esperadas para cada combinação de categorias, calculadas com base nas proporções marginais, assumindo que as variáveis são independentes;

- Cálculo da Estatística: A estatística qui-quadrado é calculada somando os quadrados das diferenças entre as frequências observadas e esperadas, divididos pelas frequências esperadas.

2.2.2. Interpretação do P-Valor

- P-Valor: O p-valor associado à estatística qui-quadrado indica a probabilidade de observar uma associação tão extrema quanto a presente nos dados, assumindo que a hipótese nula (de independência) é verdadeira.

- Baixo P-Valor (p < 0.05): Rejeita-se a hipótese nula, indicando que existe uma associação estatisticamente significativa entre as variáveis.

- Alto P-Valor (p >= 0.05): Não se rejeita a hipótese nula, indicando que não há evidência suficiente para afirmar que as variáveis estão associadas.

2.2.3. Vantagens do Teste Qui-Quadrado

- Simplicidade: Fácil de calcular e interpretar, sendo amplamente utilizado em análises exploratórias.

- Flexibilidade: Aplicável a qualquer tabela de contingência com variáveis categóricas.

2.2.4. Limitações

- Tamanho da Amostra: Requer um tamanho de amostra suficientemente grande para que as frequências esperadas em todas as células sejam adequadas (geralmente, pelo menos 5).

- Natureza das Variáveis: Adequado apenas para variáveis categóricas, não sendo diretamente aplicável a dados contínuos sem categorização.

2.3. Análise dos Resíduos Padronizados Ajustados

A análise dos resíduos padronizados ajustados é uma técnica poderosa para identificar associações significativas entre pares de categorias em dados categóricos. Esta técnica enriquece a análise de correspondência ao revelar detalhadamente quais combinações de categorias contribuem

significativamente para as associações observadas e quais apresentam frequências observadas que diferem de modo significativo das frequências esperadas, assumindo a hipótese de independência.

Esses insights permitem uma interpretação mais profunda e bem informada dos dados.

2.3.1. Cálculo dos Resíduos Padronizados Ajustados

1) Frequências Observadas (O): Contagens reais de cada combinação de categorias das duas variáveis;

2) Frequências Esperadas (E): Contagens esperadas para cada combinação de categorias, calculadas com base nas proporções marginais, assumindo independência;

3) Resíduos Brutos: Diferença entre as frequências observadas e esperadas ($O - E$);

4) Padronização: O resíduo bruto é dividido pelo desvio padrão das frequências esperadas para obter o resíduo padronizado ajustado.

2.3.2. Interpretação dos Resíduos Padronizados Ajustados

- Valor Absoluto Maior que 1,96: Indica uma associação significativa entre o par de categorias ao nível de significância de 5%;

- Valor Absoluto Menor que 1,96: Indica que não há associação significativa entre aquele par de categorias ao nível de significância de 5%;

- Valor Absoluto Alto: Um valor absoluto alto do resíduo padronizado ajustado (geralmente maior que 2 ou 3) indica que a frequência observada é significativamente diferente da frequência esperada, sugerindo uma associação significativa entre as categorias;

- Valor Próximo de Zero: Um valor próximo de zero indica que a frequência observada está próxima da esperada, sugerindo pouca ou nenhuma associação significativa.

2.3.3. Utilidade na Análise de Correspondência

- Identificação de Associações Significativas: Permite identificar quais pares de categorias têm associações significativas, destacando padrões que podem não ser aparentes apenas pela inspeção da tabela de contingência;

- Aperfeiçoamento de Modelos: Ao identificar associações significativas, os analistas podem ajustar seus modelos ou focar em relações específicas para obter insights mais detalhados;

- Detalhamento Fino: Fornece um nível de detalhamento que complementa a análise de correspondência, permitindo uma compreensão mais profunda das associações entre categorias;

- Foco em Áreas Críticas: Ajuda a focar em áreas críticas ou significativas dentro dos dados, facilitando a tomada de decisões informadas.

2.4. Mapa Perceptual

A análise de correspondência é uma técnica estatística utilizada para explorar e visualizar associações entre variáveis categóricas. O resultado desta análise é frequentemente representado em um mapa perceptual, que é uma representação bidimensional das relações entre categorias das variáveis analisadas.

2.4.1. Interpretação do Mapa Perceptual

1. Proximidade dos Pontos:
    - Associação Forte: Quando as categorias das variáveis estão próximas umas das outras no mapa, isso indica uma associação forte entre essas categorias.
    - Associação Fraca: Quando as categorias estão distantes no mapa, isso sugere uma associação fraca ou inexistente entre elas.

2. Eixos do Mapa:
    - Os eixos do mapa perceptual são dimensões que representam a maior variabilidade nos dados. As posições dos pontos ao longo desses eixos ajudam a entender as principais fontes de variação nas associações entre categorias.

3. Interpretação Visual:

- O mapa perceptual facilita a interpretação visual das associações, permitindo que os analistas identifiquem rapidamente quais categorias estão mais associadas. Isso é particularmente útil em estudos exploratórios, onde o objetivo é descobrir padrões ocultos nos dados.

2.4.2. Passos na Criação do Mapa Perceptual

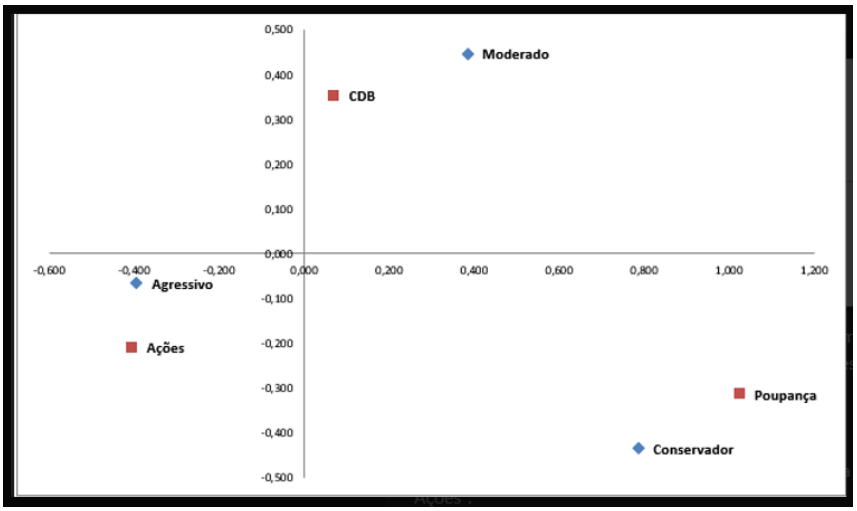
1. Construção da Tabela de Contingência:
   - Reúne os dados categóricos e constrói uma tabela de contingência com as frequências observadas das combinações de categorias.

2. Cálculo das Frequências Esperadas:
   - Calcula as frequências esperadas com base na hipótese de independência entre as variáveis.

3. Análise de Correspondência:
   - Realiza a análise de correspondência para decompor a matriz de dados e calcular as coordenadas dos pontos no mapa perceptual.

4. Geração do Mapa:
   - Plota as categorias das variáveis no mapa bidimensional, onde a proximidade dos pontos indica a força das associações.

2.4.1. Exemplo de Mapa Perceptual



- A categoria 'Agressivo' apresenta uma forte associação com a categoria 'Ações': A proximidade entre os pontos "Agressivo" e "Ações" no mapa perceptual indica uma forte associação entre essas categorias.

3. ANÁLISE DE CORRESPONDÊNCIA SIMPLES (ACS)

A Análise de Correspondência Simples (ACS) é aplicada a tabelas de contingência que contêm duas variáveis categóricas. O objetivo principal da ACS é transformar uma tabela de contingência em um mapa de pontos em um espaço de menor dimensão, tipicamente bidimensional, onde as proximidades entre os pontos refletem as associações entre as categorias das variáveis.

3.1. Passos da ACS:

1. Construção da Tabela de Contingência: A tabela de contingência é montada com as frequências observadas para cada combinação de categorias das duas variáveis;

2. Cálculo dos Perfis das Linhas e Colunas: Os perfis são vetores de proporções que representam a distribuição das frequências das categorias;

3. Cálculo das Inércias: A inércia total representa a variabilidade dos dados que pode ser explicada pela ACS;

4. Singular Value Decomposition (SVD): A decomposição de valores singulares é usada para projetar os dados em um espaço de menor dimensão;

5. Interpretação dos Mapas de Correspondência: Os mapas resultantes mostram a relação entre as categorias de uma forma visualmente intuitiva.

4. ANÁLISE DE CORRESPONDÊNCIA MÚLTIPLA (ACM)

A Análise de Correspondência Múltipla (ACM) é uma extensão da ACS que permite a análise de tabelas de contingência com mais de duas variáveis categóricas. A ACM transforma a tabela multidimensional em um espaço de menor dimensão, permitindo a visualização das associações entre todas as categorias envolvidas.

4.1. PASSOS DA ACM:

1. Construção da Tabela de Dados: A tabela contém frequências ou contagens para todas as combinações possíveis das categorias das variáveis;

2. Normalização e Cálculo das Inércias: Similar à ACS, mas aplicado a um contexto multidimensional;

3. Decomposição e Projeção: Usa métodos como a SVD para reduzir a dimensionalidade e projetar os dados;

4. Interpretação dos Resultados: Os gráficos resultantes ajudam a identificar padrões e relações complexas entre múltiplas variáveis.

Em resumo, a análise de correspondência, tanto simples quanto múltipla, são ferramentas poderosas para a análise exploratória de dados categóricos, revelando insights significativos sobre as relações entre variáveis sem a necessidade de suposições prévias. É particularmente valiosa em cenários onde a visualização das associações entre categorias facilita a interpretação e a tomada de decisões.