

Princípios básicos da Análise Exploratória de Dados UFCD: 10807

CARGA HORÁRIA: 25 horas

ALEXANDRA CAMPOS VIDAL DE SOUZA
FORMADORA



Aula 5, 6 e 7

Métodos de análise de dados para duas variáveis

- Covariância e correlação
- R de Spearman
- Medidas de concentração
- Medidas de concentração
- Números índice
- Princípios básicos sobre probabilidades
- Princípios básicos sobre amostragem e metodologia de recolha de dados
- Princípios da análise de componentes principais
- Projeto de análise exploratória de dados



Covariância e correlação



Tipos de Variáveis de Dados

Os cálculos de covariância e correlação têm diversas aplicações em estatística, ciência de dados, economia, finanças, e outras áreas. Aqui está um panorama das funcionalidades e utilidades de cada uma:



Covariância

A **covariância** é uma medida estatística que indica como duas variáveis se comportam em relação uma à outra. Ela analisa a **tendência conjunta** das variáveis, ou seja, se aumentam ou diminuem juntas, ou se uma sobe enquanto a outra desce.

Covariância positiva: Se **X aumenta**, então **Y também tende a aumentar** (exemplo: altura e peso de pessoas).

Covariância negativa: Se **X aumenta**, então **Y tende a diminuir** (exemplo: preço de um produto e sua demanda).

Covariância próxima de zero: Indica que não há uma relação linear clara entre as variáveis.



Cálculos

- O cálculo da covariância geralmente utiliza o total de linhas para pupulacao e menos 1 ($n-1$) quando se trata de uma amostra. Isso é feito para obter um estimador não tendencioso da covariância populacional.



Cenários:

1. Amostra (com $n - 1$):

- Usar $n - 1$ no denominador ajusta a variabilidade por ser uma amostra e fornece um estimador mais preciso para a população.
- Fórmula:

$$\text{Covariância (amostra)} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{n - 1}$$

2. População (com n):

- Usar n no denominador é adequado quando os dados representam toda a população.
- Fórmula:

$$\text{Covariância (população)} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{n}$$

Resumo:

- Amostra: Use $n - 1$.
- População: Use n .



Aplicações da Covariância:

1. Análise de Relações em Finanças: Avaliar como dois ativos (ações, índices) se comportam juntos:

- Um valor de covariância positiva indica que os preços tendem a subir e descer juntos.
- Uma covariância negativa indica movimentos opostos.



Aplicações da Covariância:

2. Identificação de Tendências Gerais:

- Usado para determinar se duas variáveis têm uma direção comum sem considerar a escala.

3. Construção de Portfólios de Investimentos:

- Covariância é essencial para medir o risco conjunto de ativos e para criar portfólios diversificados.



Correlação

A correlação mede a força e a direção da relação linear entre duas variáveis. É um coeficiente que varia entre -1 e +1:

1: Correlação positiva perfeita (quando uma variável aumenta, a outra também aumenta proporcionalmente).

0: Nenhuma correlação linear (não há relação linear entre as variáveis).

-1: Correlação negativa perfeita (quando uma variável aumenta, a outra diminui proporcionalmente).

Ela é útil para entender padrões em dados e verificar se duas variáveis estão relacionadas, sendo amplamente usada em estatística, análise de dados e aprendizado de máquina.



A **correlação** mede a relação entre duas variáveis numéricas, indicando se elas variam juntas. No Power BI, podemos calcular a correlação de **Pearson** usando a fórmula:

$$r = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum(X - \bar{X})^2 \times \sum(Y - \bar{Y})^2}}$$

$$r = \frac{\text{Covariância}(X, Y)}{\text{Desvio Padrão de } X \times \text{Desvio Padrão de } Y}$$



Aplicações da Correlação:

1. Estudo de Relações Estatísticas:

1. Entender quão forte é a relação entre variáveis em áreas como psicologia, sociologia, biologia, etc.
2. Exemplo: Estudar a correlação entre horas de estudo e desempenho em provas.



Aplicações da Correlação:

2. Análise de Dados em Ciência e Engenharia:

1. Identificar variáveis mais relevantes para modelos preditivos em aprendizado de máquina.
2. Estudo de séries temporais, como temperatura e precipitação ao longo do tempo.



Aplicações da Correlação:

3. Financeiro e Econômico:

1. Analisar relações entre indicadores econômicos, como PIB e desemprego.
2. Entender a relação entre diferentes mercados (ações, câmbio).

4. Controle de Qualidade e Engenharia:

1. Avaliar a relação entre diferentes parâmetros de produção para melhorar processos industriais.



Diferença entre Covariância e Correlação

Embora ambos sejam usados para medir a relação entre variáveis:

- **Covariância** dá a direção da relação, mas não é padronizada. Seu valor depende das unidades das variáveis.
- **Correlação** é padronizada (valores entre -1 e 1), permitindo a comparação direta entre diferentes conjuntos de dados.



R de Spearman



O R de Spearman é uma medida de correlação que avalia a força e a direção da associação monotônica (não necessariamente linear) entre duas variáveis. Ele é baseado nos ranks (ordens) dos dados, em vez dos valores absolutos. Isso o torna robusto contra outliers e adequado para variáveis ordinais ou relações não lineares.



Fórmula do Spearman : $R_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$

- Onde:
- **di** = Diferença entre os **ranks** das observações nas duas variáveis.
- **n** = Número total de observações.



R_s	Interpretação
1.0	Correlação positiva perfeita 
0.9 a 0.99	Correlação muito forte (+)
0.7 a 0.89	Correlação forte (+)
0.5 a 0.69	Correlação moderada (+)
0.3 a 0.49	Correlação fraca (+)
0.0 a 0.29	Correlação muito fraca ou inexistente
-0.3 a -0.49	Correlação fraca (-)
-0.5 a -0.69	Correlação moderada (-)
-0.7 a -0.89	Correlação forte (-)
-0.9 a -0.99	Correlação muito forte (-)
-1.0	Correlação negativa perfeita 



Exercicio Moodle

Sobre os métodos de análise de dados para duas variáveis, fazer em Power Bi os cálculos referentes a:

1. Medias Idade, Renda Anual e media compras
2. Desvio padrão idade e desvio padrão de renda
3. Covariância sobre Idade e Renda
4. Correlação sobre população
5. R de Spearman

Utilizar o ficheiro Excel como origem de dados para base de calculo.

Enviar em ficheiro com print com os resultados e cálculos das formulas associados aos resultados.



Para os calculos

Diferença Idade = idade – media de idade

Diferença renda anual = renda – media renda

Diferença compras mês = comprasmes-media compras mes

Covariância idade x renda população

Covariância_Idade_RendaAnual =

```
VAR Media_Idade = AVERAGE(Folha3[Idade])
```

```
VAR Media_Renda = AVERAGE(Folha3[Renda Anual])
```

```
RETURN
```

```
    AVERAGEX(
```

```
        Folha3,
```

```
        (Folha3[Idade] - Media_Idade) * (Folha3[Renda Anual] - Media_Renda)
```

```
    )
```

- Covariância idade x renda amostra **repetir**
- Correlação – para calcular precisa primeiro calcular o desvio padrao de idade e renda anual
Desvio_Padrão_Idade = STDEV.P(Folha3[Idade])

A formula para calcular a correlação é covariânciaIdadeRenda(para população e para amostra) / desvio padrão idade * desvio padrao renda



FicheiroBaseInserirModelaçãoVerOtimizarAjuda

Cortar

Copiar

Pincel de formatação

Obter dados

Livro do Excel

Hub de dados do OneLake

SQL Server

Introduzir dados

Dataverse

Origens recentes

Transformar dados

Atualizar dados

Novo visual

Caixa de texto

Mais elementos visuais

Novo cálculo visual

Nova medida

Medida rápida

Confidencialidade

Publicar

Copilot

Área de TransferênciaDadosConsultasInserirCálculosConfidencialidadePartilharCopilot

Partilhar

MedidaIdade

MedidaRendaAnual

MedidaComprasMes

35,60

66,00

10,40

Desvio_Padrão_Idade

Desvio_Padrão_Renda

10,65

18,55

Idade

Diferença Idade

Renda Anual

DiferençaRenda Anual

Compras por Mês

diferença com porasmes

Covariância_Idade_RendaAnual

23

-12,60

40

-26,00

7

-3,40

174,40

45

9,40

70

4,00

8

-2,40

174,40

25

-10,60

50

-16,00

10

-0,40

174,40

35

-0,60

80

14,00

12

1,60

174,40

50

14,40

90

24,00

15

4,60

174,40

174,40

Covariância População

218,00

Covariância Amostra

0,88

indica uma correlação positiva forte, o que significa que existe um relação linear direta: à medida que a idade aumenta, a renda anual tende a aumentar.

1,10

Visualizações

Dados

Procurar

CorrelaçãoPopulacao

Covariância_Idade_RendaAnual

CovariânciaAmostraldade_Renda

CovariânciaPopulaçãoldade_Renda

Desvio_Padrão_Idade

Desvio_Padrão_Renda

Diferença Idade

Diferença_RanKs

diferença com porasmes

DiferençaRendaAnual

Idade

Quadrado_Diferenca

RanK_Idade

RanK_Renda

Renda Anual

Soma_Quadrados

Spearman_Corr

Medidas

Column

MedidaComprasMes

MedidaIdade

MedidaRendaAnual

Vendas

Vendas

Página 1

Página 2

+



Para os calculos R Spermán

Idade	RanK_ Idade	Renda Anual	RanK_Renda	Diferenca_ RanKs	Quadrado_ Diferenca	Soma_Quadrados
23	1	40	1	0	0,00	0,00
25	2	50	2	0	0,00	0,00
35	3	80	4	-1	1,00	1,00
45	4	70	3	1	1,00	1,00
50	5	90	5	0	0,00	0,00
Total						2,00

0,90

Spearman_Corr

Spearman = 0,9: Isso significa que há uma forte tendência para que, à medida que a **Idade** aumenta, a **Renda Anual** também aumenta, mas não perfeitamente (não é uma correlação perfeita de 1).

Rank idade `RanK_Idade = RANKX(ALLSELECTED(Folha3), Folha3[Idade], , ASC, Dense)`

Rank Renda

Diferenca dos ranks

Quadrado da diferenca `Quadrado_Diferenca = [Diferenca_RanKs] ^ 2`

Soma dos quadrados `Soma_Quadrados = SUM(Folha3[Quadrado_Diferenca])`

A formula do R Spermán ficaria assim:

`Spearman_Corr =`

`1 - (6 * [Soma_Quadrados]) / (COUNTROWS(Folha3) * (COUNTROWS(Folha3)^2 - 1))`



Area de Transferencia

Dados

Consultas

Inserir

Calculos

Confidencialidade

Partilhar

Copilot

Idade	RanK_ Idade	Renda Anual	RanK_Renda	Diferenca_RanKs	Quadrado_Diferenca	Soma_Quadrados
23	1	40	1	0	0,00	0,00
25	2	50	2	0	0,00	0,00
35	3	80	4	-1	1,00	1,00
45	4	70	3	1	1,00	1,00
50	5	90	5	0	0,00	0,00
Total						2,00

0,90

Spearman_Corr

Spearman = 0,9: Isso significa que há uma forte tendência para que, à medida que a **Idade** aumenta, a **Renda Anual** também aumenta, mas não perfeitamente (não é uma correlação perfeita de 1).

Visualizações

Criar visual

Filtros

Adicionar campos de dados...

Pormenorizar

Relatório cruzado

Manter todos os filtros

Adicionar os campos de p...

Dados

Procurar

Folha3

Cliente

Compras por Mês

Correlação amostra

Correlação População

Covariância Idade_Renda Anual

Covariância Amostral Idade_Renda

Covariância População Idade_Renda

Desvio_Padrão Idade

Desvio_Padrão_Renda

Diferença Idade

Diferença_RanKs

diferença compras mes

Diferença Renda Anual

Idade

Quadrado_Diferenca

RanK_Idade

RanK_Renda

Renda Anual

Soma_Quadrados

Spearman_Corr

Medidas

Column

Vendas

Página 1

Página 2

+

INSTITUTO DO EMPREGO E FORMAÇÃO PROFISSIONAL, IP

www.iefp.pt

Medidas de Concentração

- **As medidas de concentração** são ferramentas usadas para descrever o quão distribuídos ou concentrados os valores estão em torno de uma área específica do conjunto de dados. Estas medidas fornecem uma visão sobre a distribuição e a centralização dos valores.



Medidas de Concentração

Principais Medidas de Concentração:

- Medidas de Tendência Central
- Índices de Concentração Estatística
- Curva de Lorenz



Medidas de Concentração

Medidas de Tendência Central

Representam os valores centrais de um conjunto de dados.

Média: Soma de todos os valores dividida pelo número total de observações.

Mediana: O valor central de um conjunto de dados ordenados.

Moda: O valor que ocorre com mais frequência no conjunto de dados.



Medidas de Concentração

Índices de Concentração

Avaliam como os valores estão concentrados em torno do centro.

Coeficiente de Gini: Mede a desigualdade entre os valores (muito usado em estudos econômicos).

Índice de Herfindahl-Hirschman (HHI): Mede a concentração de mercado em termos de participação.

Percentis (ou Quartis): Dividem os dados em partes iguais para entender a dispersão dos valores (ex.: 25%, 50%, 75%).



Medidas de Concentração

Curva de Lorenz

Representa graficamente a distribuição de concentração. É utilizada principalmente em estudos económicos para mostrar a distribuição da renda.



Números Índice

Os **números índice** são ferramentas estatísticas usadas para medir mudanças relativas em variáveis ao longo do tempo, permitindo comparações e análises de tendências.

- Índices economicos
- Índices educacionais
- Índices Ambientais
- Índices Sociais entre outros



Números Índice

Tipos de Números Índice:

Simples: Compara o valor de uma variável em um período com um período base.

Fórmula : $I = \text{Valor Atual} / \text{Valor Base} \times 100$

Composto:

Mede mudanças em um conjunto de variáveis.

Exemplo: Índices de Preços ao Consumidor (IPC).

Princípios Básicos sobre Probabilidades

A probabilidade é uma área da matemática que mede a chance ou a possibilidade de um evento ocorrer. É amplamente utilizada em áreas como estatística, ciência de dados, análise de riscos e tomada de decisão. Os princípios básicos fornecem o alicerce para lidar com incertezas e prever resultados.



Princípios Básicos sobre Probabilidades

Conceitos Fundamentais:

Espaço Amostral (S): Conjunto de todos os resultados possíveis de um experimento.

Evento (A): Subconjunto de S , representando os resultados de interesse.

Probabilidade Clássica:

Fórmula: $P(A) = \text{Número de casos favoráveis} / \text{Número de casos possíveis}$



Princípios Básicos sobre Probabilidades

A probabilidade de um evento A ocorrer é dada pela fórmula:

$$P(A) = \frac{\text{Número de resultados favoráveis}}{\text{Número total de resultados possíveis}}$$

$P(A)$: Probabilidade do evento A .

- valor de $P(A)$ sempre está entre 0 e 1:
- $P(A)=0$: Evento impossível.
- $P(A)=1$: Evento certo.



Princípios Básicos sobre Probabilidades

Case de Probabilidade:

Sorteio de Prêmios em uma empresa com 100 cupons, distribuídos da seguinte maneira:

10 cupons correspondem a prêmios de 100 Euros.

5 cupons correspondem a prêmios de 500 Euros.

Os demais (85 cupons) não correspondem a nenhum prêmio.



Pergunta:

Qual a probabilidade de:

1. Sortear um cupom de R\$ 100?
2. Sortear um cupom de R\$ 500?
3. Não ganhar nada?



Respostas:

Probabilidade de sortear um cupom de 100 euros:

- $P(100) = \text{Numero de cupos de 100 euros} / \text{total de cupons}$
- $P(100) = 10 / 100 = 0,01$ ou 10%

Probabilidade de sortear um cupom de 500 euros:

- $P(500) = \text{Numero de cupos de 500 euros} / \text{total de cupons}$
- $P(500) = 5 / 100 = 0,05$ ou 5%



Respostas:

Probabilidade de sortear um cupom (Não ganhar):

- $P(\text{Não ganhar}) = \text{Numero de cupos de 100 euros} / \text{total de cupons}$
- $P(\text{Não ganhar}) = 85 / 100 = 0,85 \text{ ou } 85\%$



Princípios básicos sobre amostragem e metodologia de recolha de dados

- A amostragem e metodologia de recolha de dados são fundamentais na estatística, a fim de garantir que as informações coletadas sejam representativas, confiáveis e úteis para a análise.



Princípios básicos sobre amostragem

- **A amostragem** é o processo de selecionar uma parte da população (**amostra**) para analisar, em vez de coletar dados de toda a **população**, o que pode ser inviável devido a custos ou limitações de tempo.



Tipos de Amostragem

a) Amostragem Probabilística

Cada elemento da população tem uma probabilidade conhecida e diferente de zero de ser selecionado. Exemplos:

Aleatória Simples: Todos os elementos têm a mesma probabilidade de serem escolhidos.

Sistemática: Seleciona elementos em intervalos regulares (ex.: cada 10ª pessoa).

Estratificada: Divide a população em subgrupos (estratos) homogêneos e seleciona amostras de cada estrato.

Por Conglomerados: Divide a população em grupos heterogêneos (conglomerados) e seleciona grupos inteiros.



Tipos de Amostragem

b) Amostragem Não Probabilística

A seleção é baseada em critérios não aleatórios, o que pode introduzir viés. Exemplos:

Por Conveniência: Baseia-se na acessibilidade dos participantes.

Intencional ou por Julgamento: O pesquisador escolhe elementos que considera representativos.

Bola de Neve: Participantes indicam novos participantes.



Tamanho da Amostra

O tamanho da amostra deve ser suficiente para:

Garantir representatividade.

Minimizar erros de amostragem.

Considerar o nível de confiança (ex.: 95%) e a margem de erro desejada.

A fórmula para o tamanho da amostra pode variar dependendo do método, mas para populações grandes e proporções é comum usar:



Tamanho da Amostra

A fórmula para o tamanho da amostra pode variar dependendo do método, mas para populações grandes e proporções é comum usar:

$$n = \frac{Z^2 \cdot p \cdot (1 - p)}{e^2}$$

Onde:

- n = tamanho da amostra.
- Z = valor crítico da distribuição normal (ex.: 1,96 para 95% de confiança).
- p = proporção esperada.
- e = margem de erro.



Princípios de Recolha de Dados

- A recolha de dados é o processo de obter informações da amostra de forma sistemática e controlada.



Métodos de Recolha de Dados

Os métodos podem variar dependendo do tipo de dados (quantitativos ou qualitativos):

Observação:

Observação direta do comportamento ou fenómeno.

Pode ser estruturada (sistemática) ou não estruturada.

Entrevistas:

Conversas diretas com perguntas estruturadas ou semi-estruturadas.

Útil para coletar dados qualitativos.



Métodos de Recolha de Dados

Questionários ou Inquéritos:

Conjunto de perguntas enviadas para os participantes.

Pode ser presencial, online ou por telefone.

Experimentos:

Dados coletados em um ambiente controlado, onde variáveis podem ser manipuladas.

Dados Secundários:

Informações já existentes em bancos de dados, relatórios ou arquivos.



Erros na Amostragem e na Recolha de Dados

Erros podem surgir em qualquer etapa do processo.
Alguns exemplos:

Erros na Amostragem:

Erro de Amostragem: Diferença entre a amostra e a população.

Viés de Seleção: Amostra não representativa da população.



Erros na Amostragem e na Recolha de Dados

Erros na Recolha de Dados:

Erro de Medição: Dados incorretos devido a instrumentos ou entrevistadores.

Viés do Respondente: Participantes respondem de forma não honesta ou incompleta.

Taxa de Não Resposta: Participantes selecionados não fornecem dados.



Princípios da Análise de Componentes Principais (PCA)

PCA é uma técnica estatística usada para reduzir a dimensionalidade de dados, preservando a maior variação possível. Ela transforma variáveis correlacionadas em componentes principais não correlacionados, ordenados pela quantidade de variância explicada. É útil para simplificar modelos e identificar padrões em dados complexos.



Projeto de Análise Exploratória de Dados (AED)

AED é o processo de exploração inicial de um conjunto de dados para identificar padrões, outliers e relações entre variáveis. Envolve resumo estatístico, visualização de dados (gráficos, histogramas) e detecção de tendências, preparando os dados para modelagem e tomada de decisões.



Projeto de Análise Exploratória de Dados (AED)

Objetivos

- **Entendimento do Conjunto de Dados:**
 - Explorar as características principais das variáveis, como distribuições, tipos de dados e metadados.
- **Identificação de Padrões e Relações:**
 - Verificar correlações, tendências e agrupamentos entre variáveis.
- **Detecção de Problemas nos Dados:**
 - Encontrar valores ausentes, inconsistentes, duplicados ou outliers.
- **Preparação para Modelagem:**
 - Gerar insights iniciais para definir hipóteses e estratégias de modelagem ou visualização.



Etapas de um Projeto de AED

1-Coleta e Importação dos Dados:

- Obter os dados de fontes confiáveis (bancos de dados, APIs, arquivos CSV, etc.).
- Conferir a integridade dos dados ao importar.

2-Limpeza de Dados:

- Identificar e lidar com valores ausentes, inconsistências e duplicações.
- Verificar tipos de dados e converter variáveis, se necessário.



Etapas de um Projeto de AED

3-Resumo Estatístico:

- Gerar estatísticas descritivas:
 - Média, mediana, moda.
 - Variância, desvio padrão.
 - Quartis e percentis.

4-Visualizações Básicas:

- Utilizar gráficos para entender os dados:
 - Histogramas para distribuições.
 - Gráficos de dispersão para relações entre variáveis.
 - Boxplots para detectar outliers..



Etapas de um Projeto de AED

5-Análise de Correlação e Relações:

- Calcular coeficientes de correlação para variáveis numéricas quando necessário.
- Avaliar relações lineares ou não-lineares.

6-Identificação de Outliers:

- Avaliar a relevância ou o impacto dos outliers.



Etapas de um Projeto de AED

7-Documentação e Comunicação:

- Registrar os achados principais, destacando padrões, tendências e problemas encontrados.
- Gerar relatórios ou dashboards para comunicação.



Ferramentas Utilizadas em AED

- **Excel:** Estatísticas descritivas e gráficos básicos.
- **Power BI/Tableau:** Visualizações interativas e exploração visual.
- **Python/R:** Análise estatística avançada e manipulação de dados com bibliotecas como Pandas, NumPy, Matplotlib e Seaborn.



Obrigado!

