FUNDAÇÃO EDSON QUEIROZ
UNIVERSIDADE DE FORTALEZA
ENSINANDO E APRENDENDO

Pré-Modelagem em Ciência de Dados

Prof. Rilder S. Pires

MBA em Ciência de Dados

Pré-Modelagem em Ciência de Dados

Encontros:

- Módulo 1: 09, 10 e 11 de dezembro de 2021
- Módulo 2: 13, 14 e 15 de janeiro de 2022
- Módulo 3: 27, 28 e 29 de janeiro de 2022

Projeto Final:

- Análise de Dados Sócio-Econômicos das Mesoregiões Cearenses

Pergunta Norteadora:

- Quão diferente são as Mesoregiões Cearenses?

Observações:

- Dados da Plataforma SIDRA-IBGE
- Produção Agrícola Municipal (https://sidra.ibge.gov.br/tabela/5457)
- Produto Interno Bruto dos Municípios (https://sidra.ibge.gov.br/tabela/5938)
- Estimativas de População: (https://sidra.ibge.gov.br/tabela/6579)
- Entregar os **notebooks com códigos e explicações**.

Na aula passada...

Revisão:

- Distribuições Discretas e Contínuas

Parte Teórica:

- Valor Esperado de uma Variável Aleatória
- Variância
- Inferência Estatística
  - Modelos Paramétricos e Não-Paramétricos
  - Conceitos Fundamentais
    - Estimativa Pontual
    - Regressão Linear Simples

Parte Prática:

- Exemplos
- Exploração dos Dados

Inferência Estatística

Conceitos Fundamentais:

- Muitos problemas de inferência podem ser identificados como sendo um dos seguintes tipos: **estimação, conjuntos de confiança ou teste de hipótese.**

Conjuntos de Confiança:

- Um **intervalo de confiaça** de $(1-\alpha)$ para um parâmetro θ é um intervalo $C_n = (a, b)$ onde

$$a = a(X_1, \ldots, X_n) \text{ and } b = b(X_1, \ldots, X_n)$$

são funções dos dados tais que:

$$P_\theta(\theta \in C_n) \geq 1 - \alpha, \text{ for all } \theta \in \Theta.$$

- Ou seja (a, b) engloba θ com probabilidade $(1-\alpha)$.
- Nós chamamos $(1-\alpha)$ a **cobertura** do intervalo de confiança.
- Se θ é um vetor, então usamos um **conjunto de confiança** em vez de um intervalo.

Conjuntos de Confiança:

- **Atenção!** C_n é aleatório e θ é fixo.
- Normalmente se utiliza um intervalos de confiaça de 95%, que correspondem a escolha de $\alpha = 0.05$.
- **Atenção!** Um intervalo de confiaça não é uma afirmação probabilística sobre θ já que θ é uma quantidade fixa.
- **Interpretação** Se repetimos o experimento várias vezes, o intervalo irá conter o parâmetro θ 95% das vezes.

Teste de Hipotese

Conceitos Fundamentais:

- Num **teste de hipótese**, iniciamos com alguma teoria padrão H_0 chamada **hipótese nula**.
- Então verificamos se os dados fornecem evidências suficientes para rejeitar a teoria.
- Se não, aceitamos H_0.
- Se conseguimos rejeitar H_0, então aceitamos a **hipótese alternativa** H_1.
- Para decidirmos se rejeitamos ou aceitamos H_0, realizamos um **teste estatístico**.

Tipos de Erros:

- **Erro tipo I:** Ocorre quando **rejeitamos** H_0 quando H_0 é verdadeira.
- **Erro tipo II:** Ocorre quando **aceitamos** H_0 quando H_1 é verdadeira.

Teste de Hipotese

Teste Estatístico:

- São testes que permitem decidir entre H_0 e H_1.
- O **nível de significância** α é definido como a probabilidade de se rejeitar a hipótese nula quando ela é verdadeira.
- O teste presume que aceitemos H_1 assumindo o nível de significância pre-estabelecido.

Exemplos:

- **Teste t:** É um teste onde se verifica se a variável t segue uma distribuição "t de Student".
- **Test Z:** É um teste onde se verifica se a variável Z segue uma distribuição normal.

p-value:

- O valor-p é a probabilidade de se obter um resultado mais extremo que o observado, assumindo-se que a hipótese nula é correta.
- **Cuidado!** O p-value NÃO é a probabilidade da hipótese nula ser verdadeira.

Projeto Final:

Projeto Final

Projeto Final:

Perguntas

Projeto Final

Projeto Final:

Perguntas

1. Qual a distribuição da “diversidade” dos municípios da sua região?

Projeto Final

Projeto Final:

Perguntas

1. Qual a distribuição da "diversidade" dos municípios da sua região?

2. Qual a distribuição dos valores de produção agrícola dos municípios da sua região?

Projeto Final

Projeto Final:

Perguntas

1. Qual a distribuição da "diversidade" dos municípios da sua região?

2. Qual a distribuição dos valores de produção agrícola dos municípios da sua região?

3. Qual a distribuição dos valores de produção do principal produto para municípios da sua região?

Projeto Final

Projeto Final:

Perguntas

1. Qual a distribuição da "diversidade" dos municípios da sua região?

2. Qual a distribuição dos valores de produção agrícola dos municípios da sua região?

3. Qual a distribuição dos valores de produção do principal produto para municípios da sua região?

4. e para o Ceará?

Projeto Final

Projeto Final:

Perguntas

1. Qual a distribuição da "diversidade" dos municípios da sua região?

2. Qual a distribuição dos valores de produção agrícola dos municípios da sua região?

3. Qual a distribuição dos valores de produção do principal produto para municípios da sua região?

4. e para o Ceará?

5. Quais outras variáveis podemos considerar?

Fim

Obrigado pela atenção!