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**.

No módulo passado...

Aula 1:

- Revisão: Distribuições
- Parte Teórica: Distribuições Discretas
- Parte Prática: Exemplos, Exploração dos Dados

Aula 2:

- Parte Teórica: Distribuições Contínuas
- Parte Prática: Exemplos, Exploração dos Dados

Aula 3:

- Parte Teórica: Princípio de Pareto e Distribuições Multivariadas
- Parte Prática: Exemplos, Exploração dos Dados

Pré-Modelagem em Ciência de Dados

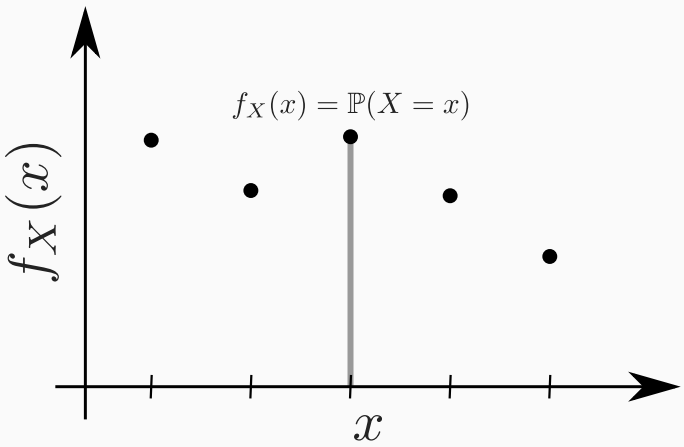
Ementa:

- Conceitos de Axiomas da Probabilidade
- Atribuições das Probabilidades
- O que é uma variável aleatória?
- Distribuição de Probabilidade Discretas:
  - Distribuição de Bernoulli,
  - Distribuição Binomial,
  - Distribuição de Poisson,
  - Distribuição Geométrica e Hipergeométrica
- Distribuições Contínuas:
  - Distribuição Uniforme,
  - Distribuição Exponencial,
  - Distribuição Normal ou Gaussiana,
  - Cálculo de Probabilidade em Distribuições Normais e Funções lineares de Distribuições Normais.
- Inferência Estatística: Noções de amostragem e estimação.

Revisão: Distribuições

Variável Discreta

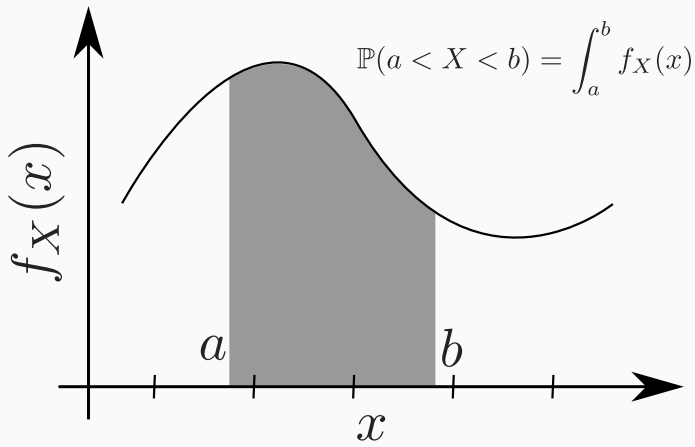
Função de Probabilidade:



$$F_X(x) = \mathbb{P}(X \leq x) = \sum_{x_i < x} f_X(x_i)$$

Variável Contínua

Função Densidade de Probabilidade:



$$F_X(x) = \int_{-\infty}^{x} f_X(t)dt.$$

Valor Esperado de uma Variável Aleatória

Definição:

- O **valor esperado**, **média** ou **primeiro momento** de uma variável aleatória X é definido como sendo:

$$\mathbb{E}(X) = \begin{cases} \sum_x x f(x) & \text{se } X \text{ é discreta} \\ \int x f(x) dx & \text{se } X \text{ é contínua} \end{cases}$$

- Assumiremos a seguinte notação para o valor esperado de X.

$$\mathbb{E}(X) = \mu_X = \mu$$

Variância

Definição:

- Seja X uma variável aleatória com média μ, a variância de X é definida como:

$$\mathbb{V}(X) = \mathbb{E}((X - \mu)^2)$$

- Assumindo que a variância existe, desvio padrão de X é:

$$\text{sd}(X) = \sigma_X = \sigma = \sqrt{\mathbb{V}(X)}$$

Inferência Estatística

Inferência Estatística

Introdução:

- Inferência Estatística é o processo de usar dados para inferir a distribuição que o gerou.
- Uma pergunta típica em Inferência Estatística é: **Dada uma amostra $X_1, \ldots, X_n$, como inferimos F?**
- Em alguns casos, nós queremos inferir apenas alguma característica de F como a sua média.

Modelos Paramétricos e Não-Paramétricos:

- Um **modelo estatístico** $\mathcal{F}$ é um conjunto de distribuições.
- Um **modelo paramétrico** é um conjunto $\mathcal{F}$ que pode ser parametrizado por um número finito de parâmetros.
- Um **modelo não-paramétrico** é um conjunto $\mathcal{F}$ que NÃO pode ser parametrizado por um número finito de parâmetros.

Modelo Paramétrico:

- Por exemplo, se nós assumirmos que os dados vêm de uma Distribuição Normal, então o modelo $\mathcal{F}$ é

$$\mathcal{F} = \left\{ f(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{ -\frac{1}{2\sigma^2}(x-\mu)^2 \right\}, \mu \in \mathbb{R}, \sigma > 0 \right\}.$$

  Esse é um exemplo de um modelo de dois parâmetros.

Modelo Não-Paramétrico:

- Por exemplo, seja $X_1, \ldots, X_n$ observações independentes de uma distribuição cumulativa F. Suponha que queremos estimar F assumindo apenas que:

$$F \in \mathcal{F}_{ALL} = \{\text{todas as distribuições cumulativas}\}.$$

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.**

Estimativa Pontual:

- Uma **estimativa pontual** trata de fornecer um único "melhor palpite" de alguma quantidade de interesse. A quantidade de interesse pode ser:
- um parâmetro em um modelo paramétrico,
- um distribuição cumulada F,
- uma função de densidade de probabilidade f,
- uma função de regressão r
- ou uma previsão para um valor futuro Y de alguma variável aleatória.

Estimativa Pontual:

- Por convenção, denotamos uma estimativa pontual de θ por $\hat{\theta}$.
- Obs: θ é uma quantidade desconhecida fixa. A estimativa $\hat{\theta}$ depende dos dados, então $\hat{\theta}$ é uma variável aleatória.
- Matematicamente, um estimador $\hat{\theta}_n$ de um parâmetro $\hat{\theta}$ é alguma função de $X_1, \ldots, X_n$:

$$\hat{\theta}_n = g(X_1, \ldots, X_n).$$

- O viés de um estimador é definido por

$$\text{viés}(\hat{\theta}_n) = \mathbb{E}_\theta(\hat{\theta}_n) - \theta.$$

  dizemos que o estimador é **não enviesado** se $\mathbb{E}_\theta(\hat{\theta}_n) = \theta$.
- Dizemos que um estimador é **consistente** se

$$\hat{\theta}_n \to \hat{\theta}$$

Estimativa Pontual:

- A qualidade de uma estimativa é as vezes obtida pelo erro quadrático médio (MSE):

$$\text{MSE} = \mathbb{E}_\theta((\hat{\theta}_n - \theta)^2)$$

- O MSE pode também ser escrito como

$$\text{MSE} = \text{viés}^2(\hat{\theta}_n) + \mathbb{V}_\theta(\hat{\theta}_n)$$

Regressão:

- Regressão é um método para estudar a relação que existe entre uma variável Y e outra X:
- A meta é estimar a função de regressão $r(x)$ a partir de dados da forma:

$$(Y_1, X_1), \ldots, (Y_n, X_n) \; F_{X,Y}.$$

Regressão Linear Simples:

- A versão mais simples de regressão ocorre quando X_i é unidimensional e $r(x)$ assume a forma linear:

$$r(x) = \beta_0 + \beta_1 x$$

Regressão Linear Simples:

- Definição:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

  onde $\mathbb{E}(\epsilon_i|X_i) = 0$ e $\mathbb{V}(\epsilon_i|X_i) = \sigma^2$
- onde os parâmetros do modelo são β_0, β_1 e σ^2
- A **reta ajustada** é

$$\hat{r}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$$

  onde os valores ajustados são $\hat{Y}_i = \hat{r}(X_i)$ e os resíduos são $\hat{\epsilon}_i = Y_i - \hat{Y}_i$
- A **soma residual dos quadrados** é definida como:

$$\text{RSS} = \sum_{i=1}^{n} \hat{\epsilon}_i^2$$

- As **estimativas de mínimos quadrados** são valores $\hat{\beta}_0$ e $\hat{\beta}_1$ que minimizam o RSS.

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!