PROJETO 2

MODELOS PROBABILÍSTICOS E DADOS

IDENTIFICAÇÃO DE DISTRIBUIÇÕES

OBJETIVO:

O objetivo deste projeto é identificar quais distribuições (funções de densidade de probabilidade - no caso contínuo, ou funções de probabilidade - no caso discreto) descrevem melhor variáveis quantitativas extraídas de um *dataset*.

O resultado final esperado é um relatório que identifique, com bons argumentos, a escolha de um ou mais modelos probabilísticos para ajuste de uma variável quantitativa extraída de um *dataset*.

Este projeto é **estritamente individual**.

O QUE DEVE SER FEITO:

Estude a expectativa de vida de países de três particulares anos, seguindo as seguintes etapas:

1. **Seleção de três anos para analisar:**
   - Você deve escolher três anos diferentes do *dataset* `Life.xlsx`, o qual apresenta, em todas as colunas, a expectativa de vida de quase todos os país do mundo entre 1800 e 2013. Necessariamente, a escolha de cada ano deve ser tal que as expectativas de vida de cada um dos três anos tenham formatos de distribuição dos dados diferentes entre si.
   - Escolhido os anos, limpe e prepare os dados para processamento (tratando valores `NaN` ou `N/A`, por exemplo).
2. **Análise Descritiva:**
   - Calcule algumas medidas resumo úteis para entender o comportamento das expectativas de vida dos três anos escolhidos por você. Analise.
   - Construa um histograma e compare as distribuições da expectativa de vida nos três anos. O que aconteceu com o formato da distribuição da expectativa de vida ao longo desses três anos. Explique.

- Construa o QQ-Plot considerando quantil amostral da expectativa de vida de um ano vs quantil amostral da expectativa de vida de um outro ano. Repita esse gráfico com todas as combinações dos três anos. Esse gráfico de quantis amostrais auxilia a compreender possíveis mudanças na expectativa de vida ao longo dos três anos escolhidos? Justifique claramente.

  Dica: veja construção desse gráfico no Magalhães e Lima (7ª edição) - pág. 27 e 29.

3. Aderência de um modelo probabilístico normal aos dados:

Assuma que X:expectativa de vida de um específico ano tenha o parâmetro μ estimado pela média amostral e tenha o parâmetro σ^2 estimado pela variância amostral, ambas de um mesmo ano escolhido na sua análise.

Verifique, visualmente/graficamente, se: **A expectativa de vida em um respectivo ano é bem modelada por uma distribuição normal?** Para isso, construa e interprete os seguintes gráficos:

- Para cada ano, construa um histograma da variável junto com a função densidade de probabilidade da distribuição normal e analise.

  Dica: Para calcular cada valor `f(x)` da distribuição normal, use o comando `stats.norm.pdf`, do pacote Stats da biblioteca SciPy do Python.

- Construa um gráfico dos valores observados ordenados vs frequência relativa acumulada empírica (a partir dos dados) e vs função de distribuição acumulada e analise.

  Dica: veja Exemplo 6.8 do Magalhães e Lima (7ª. edição).

  Considere a notação para a variável quantitativa expectativa de vida:

  Valores observados: $x_1, x_2, x_3 \dots, x_n$.

  Valores ordenados: $x_{(1)} \leq x_{(2)} \leq x_{(3)} \leq \cdots \leq x_{(n)}$.

  Para calcular a **frequência relativa acumulada empírica**, assuma que cada observação terá um peso igual a $1/n$. Logo, para calcular o quanto se tem de frequência relativa acumulada empírica até a observação $x_{(i)}$, faça:

$$f_{ae(i)} = \frac{i - 0{,}5}{n}, \text{com } i = 1, \dots, n.$$

  Para calcular a **probabilidade acumulada até a observação $x_{(i)}$** a partir da distribuição normal, ou seja, a **função de distribuição acumulada $P(X \leq x_{(i)})$**, use o comando `stats.norm.cdf`, do pacote Stats da biblioteca SciPy do Python.

- Construa o Gráfico de probabilidade considerando o quantil amostral vs o quantil teórico da distribuição normal e analise.

  Dica: veja Exemplo 6.10 do Magalhães e Lima (7ª. edição).

  O **quantil amostral** é dado pelos valores $x_{(i)}$, com $i = 1, \dots, n$. Ainda, vimos que a frequência relativa acumulada empírica até a observação $x_{(i)}$ é dada por $f_{ae(i)}$.

  O **quantil teórico** da distribuição normal é dado pelo valor $x_{T(i)}$ tal que $(X \leq x_{T(i)}) = f_{ac(i)}$. Para obter cada valor $x_{T(i)}$, com $i = 1, \dots, n$, use o comando `stats.norm.ppf(p, loc=mu, scale=sigma)`, sendo `p` a frequência relativa acumulada empírica até a observação $x_{(i)}$ dada por $f_{ae(i)}$, `mu` estimada pela média amostral e `sigma` estimada pelo desvio padrão amostral.

4. Aderência de um modelo probabilístico não normal aos dados:

Apenas para os anos escolhidos cuja expectativa de vida tem formato assimétrico, procure identificar uma função densidade de probabilidade adequada que descreva as probabilidades de ocorrência dos valores que essa variável pode assumir. Para esses anos, faça:

- Para cada ano, construa um histograma da variável junto com a função densidade de probabilidade da distribuição de probabilidade escolhida e analise.
- Construa um gráfico dos valores observados ordenados vs frequência relativa acumulada (a partir dos dados) e vs função de distribuição acumulada da distribuição escolhida e analise.
- Construa o Gráfico de probabilidade considerando o quantil amostral vs o quantil teórico da distribuição escolhida e analise.

5. Cálculo de probabilidades a partir da distribuição normal:

Independente da sua resposta do item anterior, assuma que X: expectativa de vida de um ano segue uma distribuição normal considerando as estimativas amostrais como valores dos parâmetros de cada bairro. Nesse caso, para cada ano escolhido, calcule:

- probabilidade de um país ter expectativa de vida superior a 70 anos.
- sabendo que um páis tem expectiva de vida superior a 70 anos, qual a probabilidade desse possuir expectativa inferior a 75 anos?
- qual a maior expectativa de vida dos 10% de países com menores expectativas de vida?
- qual a menor expectativa de vida dos 10% de países com maiores expectativas de vida?

ENTREGÁVEIS ESPERADOS E DATAS:

Item	Data	Descrição
Desenvolvimento do projeto em sala de aula	**30/03/2017 Até às 15h30**	Publicar na pasta "Projeto2" no github, com nome Projeto2-EntregaAula
Projeto final	**04/02/2017 Até às 23h59**	Publicar na pasta "Projeto2" no github, com nome Projeto2-EntregaFinal

RUBRICS DE AVALIAÇÃO DO OBJETIVO DE APRENDIZADO

Objetivo de aprendizado	Insatisfatório (I)	Em desenvolvimento (D)	Essencial (C)	Proficiente (B)	Avançado (A)
Especificar as distribuições de probabilidades adequadas para as variáveis	Apresentou entregas insuficientes ou atrasadas	Para os três anos escolhidos: - Selecionou três anos de expectativa de vida com formatos de distribuição diferentes entre si.	Para os três anos escolhidos: - Fazer adequadamente a análise descritiva. - Fazer adequadamente as análises para o ajuste da distribuição normal.	Realizou os comportamentos de C de maneira excelente e para os anos com assimetria nos dados: - Fazer adequadamente as análises para um ajuste de uma distribuição teórica com comportamento de assimetria semelhante ao apresentado nos dados. - Fazer corretamente as contas das probabilidades da normal.	Realizou os comportamentos de B e C de maneira excelente e: - Análises feitas em todos os gráficos com excelentes argumentos e interpretações.