Trabalho Parte 1

Trabalho Parte 1: Análise dos dados e distribuições amostrais

Dicionário dos dados:

Variável	Tipo	Nome da variável	Descritivo e unidades
Idade	Objetivo	***age***	int (days)
Altura	Objetivo	***height***	int (cm)
Peso	Objetivo	***weight***	float (kg)
Gênero	Objetivo	***gender***	1: Feminino 2: Masculino
Pressão sistólica (contração e saída de sangue)	Exame	***ap_hi***	Int (mmHg)
Pressão diastólica (relaxamento e entrada de sangue)	Exame	***ap_lo***	Int (mmHg)
Colesterol	Exame	***cholesterol***	1: normal, 2: above normal, 3: well above normal
Glicose	Exame	***gluc***	1: normal, 2: above normal, 3: well above normal
Se é fumante	Subjetivo	***smoke***	Binary 0: Não fumante 1: Fumante
Se ingere bebida alcoólica	Subjetivo	***alco***	binary
Se pratica atividade física	Subjetivo	***active***	binary
Presença ou ausência de doença cardiovascular	Target	***cardio***	binary

Definições:

População: Considere todos os elementos (70k indivíduos) como a sua população.

Amostra: uma amostra aleatória dessa População

Obs: utilize algum programa (R/Python/Matlab/Scilab/Octave/..) ou software (Excel/Calc/...) para realizar o processo de amostragem aleatória simples.

Faça um trabalho contendo o seguinte (.pdf):

- **Capa**: Trabalho – Parte 1, Nome completo, matrícula e turma

- **Sumário**

Tópicos:

1. Análise exploratória simples

Esse tópico é aberto, ou seja, o aluno poderá explorar a base de dados de diferentes maneiras, visando extrair informação dos dados.

Algumas poucas sugestões são dadas a seguir:

- Verifique se a base de dados possui valores inconsistentes (***weight*** muito inferior, ***height*** muito acima,...)

- Faça transformações nos dados:

    - Transforme idade (days) para idade (years)

    - Calcule o IMC

    - ... outras que achar necessário

- Construa diferentes gráficos para explicar relações ou obter informação a respeito dessa população.

Exemplos:

- Histograma do IMC (calcule) de toda a população

*- Histograma do IMC de toda a população por **genre***

*- Histograma do IMC de toda a população por **cardio***

- Gráfico de dispersão entre as pressões diastólica e sistólica, por grupos (alco, cardio, ...)

- Existência de correlação linear entre variáveis? Quais? explique.. (pode montar uma matriz de correlação)

2. Distribuições amostrais

Nesse tópico o aluno deverá efetuar diferentes procedimentos e responder algumas questões ligadas à distribuição amostral. O aluno deverá realizar amostragens aleatórias simples na população e construir as distribuições amostrais além de responder e concluir os resultados encontrados.

2.1 Parâmetros e distribuições amostrais

2.1.1 Calcule os parâmetros populacionais de cada variável, discreta ou contínua.

Calcule o Valor Esperado (μ), Variância (σ^2), Desvio Padrão (σ), e se a variável segue algum modelo que você conheça, no caso de não seguir um modelo específico informe que segue um modelo empírico. Para variáveis contínuas construa um histograma (e densidades) e para as discretas uma distribuição massa de probabilidade (Bernoulli).

Variáveis: *age, height, weight, smoke, alco, active, cardio, ap_hi, ap_lo*

Exemplo:

Variável: Massa/peso/weight [kg]

Tipo de variável: Contínua

Gráficos: Histograma

Parâmetros: E[], V(), DP(), ...

Depois defina algumas subpopulações de interesse e calcule os parâmetros, o Valor Esperado (μ), Variância (σ^2), Desvio Padrão (σ), e se a variável segue algum modelo que

você conheça, no caso de não seguir um modelo específico informe que segue um modelo empírico e acrescente mais medidas, como os quartis.

Exemplo:

Subpopulação: Homens, fumantes, não praticantes de atividade física.

Variável: Massa/peso [kg]

Tipo de variável: Contínua

Gráficos: Histograma (apresente)

Parâmetros: E[], V(), DP(), ...

2.1.2 Tome 100000 amostras aleatórias de tamanho n = 5, com reposição e calcule as médias amostrais (ou proporções amostrais quando for o caso), construa um histograma dos valores e calcule o valor esperado, variância e desvio padrão.

Variáveis: *age, height, weight, smoke, alco, active, cardio, ap_hi, ap_lo*

2.1.3 Tome 100000 amostras aleatórias de tamanho n = 35, com reposição e calcule as médias amostrais (ou proporções amostrais quando for o caso), construa um histograma dos valores e calcule o valor esperado, variância e desvio padrão.

Variáveis: *age, height, weight, smoke, alco, active, cardio, ap_hi, ap_lo*