PRO3200 – EPUSP
Professora Celma de Oliveira Ribeiro
Assistente: Pedro Gerber Machado & Monitor: Vinícius Castanho

Estudo de Caso 1: Análise de Dados

Neste estudo de caso, utilizaremos a base de dados IMDB.rds adaptada no último case.

1. Análise de Dados.

Muito bem, até aqui vimos algumas funções importantes para o ambiente de programação, funções de importação e alguns cuidados iniciais. A partir daqui, com os dados já devidamente importados, alguns itens são pedidos.

Utilize o comando help, referências e fóruns online para gerar/calcular os dados requisitados.

a. Crie uma variável para cada status e analise as principais medidas de posição central e dispersão (são elas: média, mediana, moda, variância, desvio padrão, amplitude). Faça essa análise para as variáveis “duracao”, "orcamento" e “nota_imdb”. Para criar as variáveis por status você pode utilizar o comando subset().

b. Represente, via boxplot, a comparação entre grupos para as categorias “duracao”, “orcamento” e “nota_imdb”. Comente sobre a comparação dos três diagramas.

c. Apresente o histograma de frequências da variável “nota_imdb” **em cor amarela** para filmes com status “novos”.

d. Faça agora o histograma com **a frequência relativa dos dados (função densidade de probabilidade no eixo y)**.

e. Calcule a média dos dados utilizados no histograma montado. Criaremos então uma linha vertical com o valor da média indicando graficamente sua posição no eixo x. Após isso, utilizando o comando *text()* escreva o nome “Média” próximo à linha criada.

PRO3200 – EPUSP
Professora Celma de Oliveira Ribeiro
Assistente: Pedro Gerber Machado & Monitor: Vinícius Castanho

2. Amostragem

Atenção! Vamos imaginar que o conjunto de dados disponível representa a população. Dessa forma, vamos selecionar amostras dessa população de forma a mantê-las representativas.

Parte 1 – Gerando amostras sem estratificação

Vamos gerar amostras de tamanho K dos dados do arquivo. Inicialmente considere K=200.

Para realizar essa amostragem, primeiro você deve sortear K índices, limitados ao tamanho da sua base de dados. Esses índices serão utilizados para selecionar uma amostra aleatória dos dados do arquivo original. Construa um vetor, por meio da função *sample (x, size, replace=TRUE)* contendo os índices que serão consideradas para construir a sua amostra.

Tome cuidado: a cada vez que você usa o comando sample ele gera uma nova amostra! Exemplo:

Z<- sample (1:número de observações, K, replace=TRUE)

Construa um novo data-frame contendo as linhas do data-frame original que você selecionou pelo comando sample. Exemplo:

amostra<- df[Z[1:K],]

a. Por que você deve utilizar replace = TRUE?

b. Usando a amostra selecionada, faça uma análise gráfica com as porcentagens dos tipos de classificação sorteados e compare de forma gráfica e quantitativa com as porcentagens observadas na população.

Parte 2 – Gerando amostras estratificadas

Se seu interesse é observar características classificatórias é razoável estratificar a sua população, buscando construir uma amostra que a represente. Para isso, você deve construir uma amostra que represente a estratificação das várias classificações. Vamos separar os conjuntos em quatro dataframes somente por razões didáticas.

c. Para cada classificação, construa um novo banco de dados relativo àquela classificação.

Exemplo: *mais_dezoito <- subset(df, df$classificacao=="A partir de 18 anos")*

d. Determine o número de elementos de cada classificação. Lembre-se que estamos considerando que os dados são da população. Consequentemente, cada novo subconjunto construído representa uma população.

e. Se você for retirar uma amostra de 200 elementos de sua população original que represente os vários extratos, quantos elementos de cada região devem ser retirados?

f. Construa uma amostra da população original com estas características.

3. Entrega no Moodle.

Os cases devem ser enviados no e-disciplinas em um arquivo .pdf com o script do R anexo ao final do próprio PDF, de forma a possibilitar o Ctrl c, Ctrl v do mesmo para efeitos de correção.

Lembre-se de que dissertações e conclusões acerca dos resultados são mais importantes que a própria construção do código em R. Indique todos os resultados da maneira mais expositiva possível.

O prazo de entrega é domingo, 02/04, às 23h59.