# Lista III: Estatística e amostragem

---

Estes exercícios devem ser realizados usando funções do `tidyverse`. Será necessário importar dados, usar funções para calcular estatísticas descritivas e sortear elementos, além de `loops` para repetir códigos. Leia as instruções com atenção antes de começar a escrever o seu código.

---

**Descrição**

Nesta lista, seu objetivo será gerar e analisar amostras da população de Belford Roxo (RJ). Para tanto, usaremos uma base de microdados de toda a população do município extraída do Censo Demográfico de 2010. A base está salva no arquivo `belford_roxo.Rda` e tem 469313 observações (linhas) e 6 variáveis (colunas). São estas:

- `id`: identificador único de cada pessoa
- `zona_domicilio`: zona de domicílio da pessoa (urbana ou rural)
- `sexo`: sexo da pessoa (conforme registrado no Censo)
- `idade`: idade da pessoa (em anos)
- `renda_mensal`: renda mensal da pessoa (em reais de 2010)
- `cor_raca`: cor ou raça da pessoa (conforme registrado no Censo)

**Tarefas**

*1. Amostragem aleatória simples*

Com a base de microdados de Belford Roxo carregada, extraia uma amostra aleatória simples de 800 pessoas (sem repetição). Calcule a média e o desvio padrão da renda mensal da amostra. Compare esses valores com os da população do município.

*Dicas:*

- Uma forma simples de sortear uma amostra de linhas de um `data.frame` é por meio da função `slice_sample` do `tidyverse`.
- Cada sorteio feito no R gerará resultados diferentes. Para garantir que seus resultados sejam reprodutíveis – isto é, que a cada vez que você compilar seu documento você obterá os mesmos resultados –, podemos usar a função `set.seed` logo no início do código, e.g., `set.seed(123)`.

*2. Tamanho da amostra*

Refaça o exercício anterior, agora com duas novas amostras: uma com 1200 pessoas; e, a outra, com 2400 pessoas. Com esses novos resultados, reporte em uma tabela a comparação dos resultados das suas três amostras; indique o tamanho da amostra nas colunas e a média da renda mensal nas linhas. Sua tabela deve ficar mais ou menos assim:

| Renda | Amostra_800 | Amostra_1200 | Amostra_2400 | Base_completa |
|---|---|---|---|---|
| Média | 1234.56 | 1234.56 | 1234.56 | 1234.56 |

*Dicas:*

- Para criar uma tabela no R é possível usar a função `kable`, do pacote `knitr`, ou `gt` do pacote de mesmo nome.

*3. Distribuição amostral da média*

Agora simule a distribuição amostral da média da renda mensal de Belford Roxo. Para isso, repita o sorteio de 800 pessoas 100 vezes e calcule a média da renda mensal de cada amostra. Dizendo de outra forma, você precisará sortear 100 amostras diferentes, cada uma com 800 pessoas sorteadas sem repetição, e calcular a média da renda em cada uma delas. Ao final do processo, o resultado será 800 medidas da renda média de Belford Roxo. Por fim, faça um histograma com a distribuição das médias da renda mensal. Adicione uma linha vertical no histograma para indicar a média da renda mensal da população de Belford Roxo calculada a partir da base completa.

*Dicas:*

- Para repetir um código várias vezes, é possível usar um `loop` do R. Lembre-se de criar um objeto vazio para armazenar os resultados de cada iteração.

*4. Comparação de renda*

Agora crie uma nova amostra de pessoas de Belford Roxo, dessa vez com $n = 2000$, e a salve em um objeto chamado `amostra_br`. Com essa amostra, compare a média da renda mensal entre pessoas autodeclaradas brancas e pretas (desconsidere as demais cores/raças). Reporte a i) a média da renda mensal e o ii) número de pessoas em cada grupo em uma tabela como a seguinte:

| cor_raca | renda_media | n |
|---|---|---|
| Branca | 1234.56 | 1234 |
| Preta | 1234.56 | 1234 |

*5. Inferência randomizada*

Será que a diferença de renda entre pessoas autodeclaradas brancas e pretas no exercício anterior é sistemática ou, ao contrário, fruto da amostragem? Para responder a essa pergunta, use inferência randomizada para simular a distribuição nula da diferença de renda entre os dois grupos. O procedimento que você deverá implementar é o seguinte:

- Usando a base da amostra com 2000 pessoas salva no objeto `amostra_br`, mantenha apenas pessoas com cor/raça branca ou preta e embaralhe os valores da variável `cor_raca`, isto é, troque aleatoriamente a cor/raça de cada pessoa na amostra (mas mantenha o tamanho da amostra para cada cor/raça);
- Calcule a média da renda mensal de cada grupo (branco e preto) na amostra embaralhada;

- Calcule a diferença entre as médias da renda mensal dos dois grupos na amostra embaralhada (i.e., branca - preta);
- Repita esse procedimento 1000 vezes e armazene as diferenças de renda em um objeto chamado `diferencas_renda`;

Faça um histograma com a distribuição das diferenças e adicione uma linha vertical para indicar a diferença observada entre as médias da renda mensal dos dois grupos na população de Belford Roxo. Calcule e interprete o P-valor da diferença observada (escreva a sua resposta fora do código, como texto).

*Dicas:*

- Para embaralhar os valores de uma variável, você pode usar a função `sample` do R com o argumento `replace = FALSE`. Isso irá, na prática, embaralhar os valores da variável, simulando um processo no qual a cor/raça de cada pessoa é trocada aleatoriamente.

**Entrega**

Envie um único arquivo PDF, gerado usando o nosso *template* em quarto, no *Google Classroom*, isto é, *não envie scripts ou outros arquivos auxiliares*. Certifique-se também de que o código no seu PDF esteja visível trocando `echo = FALSE` por `echo = TRUE` na seguinte linha do *template*, que está perto do início do documento:

```
knitr::opts_chunk$set(echo = FALSE, message = FALSE, warning = FALSE)
```

Cada seção do seu documento (sub-títulos antecedidos por `##`) deve conter o código que você escreveu para responder o item correspondente da tarefa. Fique à vontade para escrever texto adicional para explicar o que você fez em cada seção.