# Lista III

Estatística, amostragem e causalidade

Bruno Marques Schaefer

26/05/2025

**Abstract**

Estes exercícios devem ser realizados usando funções do `tidyverse` e outros pacotes. 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. Os objetivos são consolidar conceitos e conteúdos trabalhados até aqui no curso

## Definições

- Leia com calma os enunciados, escreva os códigos e exporte os resultados.
- Apague o enunciado da versão final (mantenha só o número da questão)

## Amostragem

Nesta primeira parte da 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, de preferência com exposição gráfica (visualização) *E/OU* em tabela.

**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)`.
- Para visualizações comparadas considere o uso de *gridExtra*

**Resposta**

## 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.

**Resposta**

## 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.

**Resposta**

### 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 |

**Resposta**

### 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 da diferença de renda entre os dois grupos. O procedimento que você deverá implementar é o seguinte:

- tire uma amostra de 2000 pessoas da população de Belford Roxo.
- nessa amostra, calcule a renda mensal média de brancos e de pretos
- calcule a diferença entre a renda média de brancos e de pretos
- repita esse processo 100 vezes usando um for loop. O resultado do for loop deve ser um objeto chamado diferencas_renda, com 100 diferenças da renda média de brancos e pretos, calculadas a partir de 100 amostras diferentes.

Agora faça um histograma com a distribuição das 100 diferenças de médias da renda mensal que você salvou no objeto diferencas_renda. Interprete essa distribuição, tentando identificar se ela indica se as diferenças são sistemáticas ou aleatórias.

**Resposta**

## Causalidade e desenho de pesquisa

Na segunda parte da lista, vamos trabalhar com conceitos de desenho de pesquisa, descrição e inferência causal. Ou seja, vamos deixar de lado inferência e probabilidade e olhar para outros aspectos da pesquisa social quantitativa.

Para isso, trabalhe com o banco de dados *base_municipios_brasileiros.RDa*. Ele é bastante amplo e toca em diversas temáticas que são objeto de estudos de vocês.

Dicionário de dados disponível aqui: https://osf.io/3yka9/files/osfstorage

Lembre-se de selecionar (*select*) somente as variáveis que lhe interessam e filtre (*filter*) as unidades de análise que serão utilizadas (anos, regiões e estados específicos).

### 6. Desenho de pesquisa

O banco de dados *base_municipios_brasileiros.RDa* possui informações sobre municípios brasileiros em diferentes dimensões (meio-ambiente, orçamento, sociedade, políticas públicas, eleições, entre outras). A partir do banco, pense em um desenho de pesquisa e preencha os componentes de um desenho de pesquisa: pergunta, variável dependente, variável independente e hipótese principal. Ou seja, você deve preencher os componentes com vistas à realização de testes empíricos.

**Resposta**

- a) Pergunta de pesquisa:
- b) Variável dependente:
- c) Variável independente:
- d) Hipótese principal:

### 7. Formalização

A partir do pacote *ggdag* formalize a relação entre a variável dependente e independente, postulando A partir do pacote *ggdag* formalize a relação entre a variável dependente e independente, postulando possíveis variáveis que confundam a relação entre ambas, e/ou que precisam ser incluídas como controles.

Mais informações sobre o *ggdag* em: https://r-causal.github.io/ggdag/

**Resposta**

### 8. Descrição

A partir do desenho de pesquisa, faça a análise descritiva (gráfica e/ou tabular) de sua variável dependente (1), variável independente (2) e da relação entre ambas (2) (utilizando ferramentas de visualização apropriadas de acordo com a operacionalização da variável - https://rkabacoff.github.io/datavis/)

**Resposta**

### 9. Teste

Utilizando regressão linear, estime a relação entre sua variável independente e variável dependente. O importante aqui é rodar o modelo, fazer a exposição em uma tabela (pacotes que podem ajudar: *gt*, *texreg*, *summ*, entre outros), e interpretar os resultados: somente o coeficiente de $\beta_1$. Foque na discussão presente no capítulo 5 de *Data Analysis for Social Science: A Friendly and Practical Introduction*.

**Resposta**

## Entrega

Envie um único arquivo PDF, gerado usando o nosso *template*, 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` no início do *template*:

```
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.

Não se esqueça de incluir seu nome em *author*