Tarefa I: Pesquisa de survey

Esta tarefa deve ser realizada usando o template em quarto disponibilizado no website da disciplina. A entrega deve ser feita via Google Classroom em um arquivo em PDF. Não é necessário anexar script ou dados ao Classroom.

Tarefa

Junto destas instruções, há um arquivo chamado `macae.rds` que contém uma base de dados com informações sobre a população do município de Macaé (RJ) obtidas pelo Censo de 2010. A base tem mais de 200 mil linhas, cada uma representando uma pessoa no município. Além disso, ela contém as seguintes variáveis:

- `pessoa_id`: identificador único de cada pessoa
- `codigo_setor`: código do setor censitário (quadra ou bloco onde mora cada pessoa)
- `distrito`: nome do distrito municipal onde mora cada pessoa
- `subdistrito`: nome do subdistrito municipal onde mora cada pessoa
- `bairro`: nome do bairro onde mora cada pessoa
- `alfabetizado`: variável binária que indica se a pessoa é alfabetizada (1) ou não (0)

Assumindo que essa base mensura corretamente o *status* de alfabetização da população de Macaé, e assumindo que estes são dados magicamente revelados da população do município, você deverá especificar três desenhos amostrais diferentes para extrair amostras de $n = 1200$ desta população. Os desenhos que você deverá produzir são:

1. **Um desenho AAS**, no qual cada pessoa tem a mesma probabilidade de ser selecionada para a amostra.

2. **Um desenho AAS com estratificação**. Você deverá escolher ou modificar as variáveis disponíveis (`distrito`, `subdistrito`, `bairro`) para definir os estratos com base em critérios que discutimos na aula, tais como homogeneidade interna em relação à variável de interesse (`alfabetizado`). Ficará a seu critério definir o número de estratos, h, o tamanho da população em cada estrato, N_h e a forma de alocação do número de pessoas a serem incluídas na amostr em cada estrato, n_h. Justifique brevemente suas escolhas.

3. **Um desenho por conglomerados**. Aqui os conglomerados a serem usados serão os setores censitários. Você deverá definir o número de setores (conglomerados), n, a serem incluídos na amostra e o número de entrevistas, n_i, a serem realizadas em cada setor. O número mínimo de entrevistas em cada setor deverá ser fixo e ser de, no mínimo, 10 e, no máximo, de 20. A seleção de pessoas a serem incluídas na amostra em cada setor censitário deverá ser aleatória. Justifique brevemente suas decisões.

Com os três planos amostrais, simule 1000 amostras de cada um deles e calcule a média/proporção da variável `alfabetizado` em cada uma delas. Lembre-se: todas os desenhos devem produzir amostras com exatamente $n = 1200$.

Faça um único gráfico que compare a distribuição das médias/proporções da variável `alfabetizado` para cada um dos desenhos. Para os desenhos estratificado e por conglomerado, calcule também

o *design effect* usando a variância amostral das estatísticas amostrais calculadas. Interprete seus resultados: por que o desenho estratificado e/ou por conglomerados gera maior ou menor variância do que o desenho AAS? Quais características dos desenhos explicam estes resultados?