Untitled

Gustavo Almeida Silva

02/07/2023

Resumo

Este trabalho apresenta um estudo de simulação sobre métodos de amostragem complexa para a estimação da média amostral. Os métodos analisados são baseados em amostragem conglomerada em 1, 2 e 3 estágios. O objetivo é comparar o desempenho desses métodos em termos de eficiência e precisão da estimativa.

A amostragem complexa é amplamente utilizada em pesquisas em que a população de interesse possui uma estrutura hierárquica ou está dividida em subpopulações distintas. A amostragem conglomerada é uma técnica comumente aplicada nesse contexto, em que a população é dividida em conglomerados e, em seguida, uma amostra é selecionada em cada conglomerado.

Neste estudo, são simulados diferentes cenários com base em parâmetros de amostragem realistas. São considerados os métodos de amostragem conglomerada em 1, 2 e 3 estágios, nos quais a seleção dos conglomerados e das unidades amostrais é realizada de forma sequencial.

Através das simulações, são comparados os estimadores das médias amostrais obtidos pelos diferentes métodos, levando em consideração a variância da estimativa e a eficiência em relação ao tamanho da amostra. Além disso, são avaliados possíveis vieses de estimadores e a precisão das estimativas em cada estágio da amostragem conglomerada.

Os resultados das simulações fornecem insights valiosos sobre a adequação e o desempenho dos métodos de amostragem conglomerada em diferentes estágios. Espera-se que este estudo contribua para a compreensão das complexidades da amostragem em pesquisas com estrutura hierárquica e auxilie pesquisadores na escolha do método de amostragem mais apropriado para suas necessidades.

Introdução

A amostragem desempenha um papel fundamental na estatística, permitindo aos pesquisadores obterem informações sobre uma população a partir de uma amostra representativa. Através de métodos estatísticos robustos, é possível extrapolar conclusões precisas e confiáveis sobre a população em geral. No entanto, a amostragem muitas vezes enfrenta desafios práticos, como a seleção adequada das unidades amostrais e a consideração de complexidades inerentes a certos planos de amostragem.

De forma geral, é amplamente reconhecido na teoria da amostragem que, embora o esquema de amostragem aleatória simples (AAS) seja teoricamente simples, na prática, é pouco utilizado devido às restrições orçamentárias e à busca por métodos probabilísticos que forneçam informações mais precisas. Além disso, é comum encontrar dificuldades na obtenção de cadastros adequados para o AAS, bem como lidar com situações de não resposta, o que requer considerar observações com pesos desiguais (Vierira 2005). A especificação inadequada na análise do plano amostral selecionado também pode resultar em estimativas enviesadas, destacando a importância de estudar metodologias que levem em conta o esquema de amostragem adotado.

Este artigo tem como objetivo explorar a interseção entre a amostragem em estatística e a simulação computacional, destacando como essa abordagem combinada pode contribuir para aprimorar a qualidade das

inferências estatísticas. Serão apresentados conceitos fundamentais da amostragem, incluindo diferentes métodos de seleção amostral e as respectivas propriedades, e, em seguida, será discutido como a simulação computacional pode ser aplicada para investigar essas técnicas em contextos específicos.

Ao integrar a simulação computacional à amostragem estatística, os pesquisadores podem explorar virtualmente uma ampla gama de cenários de amostragem, considerando diferentes planos amostrais, tamanhos de amostra e distribuições populacionais. Além disso, a simulação permite a avaliação de métricas de desempenho, como viés e erro padrão, fornecendo insights valiosos sobre a precisão e a eficiência dos métodos de amostragem em diferentes contextos.

Metodologia

O objetivo deste trabalho é comparar diferentes planos de amostragem em estágios complexos, como a amostragem estratificada e a amostragem conglomerada. Para realizar essa comparação, foi conduzido um estudo de simulação. O estudo tem como propósito investigar e avaliar o desempenho desses diferentes planos amostrais em termos de eficiência, precisão e viés. Através da simulação, é possível criar cenários controlados que permitem analisar o impacto de cada plano amostral em diferentes características da população. Com base nos resultados obtidos na simulação, será possível identificar quais planos de amostragem são mais adequados para determinados contextos e auxiliar na tomada de decisões estatísticas mais embasadas.

Para isso, foi utilizado o conjunto de dados: **Alunos.txt**, que se trata de dados sobre notas de alunos na prova de portugues. Os dados são populacionnais, ou seja, é um cadastro completo dos alunos da rede básica de **** lugar. Assim, os diferentes métodos de amostragem complexa foram utilizados em cima desse conjunto de dados.

O conjunto de dados possui 6 variáveis:

- Aluno
  - Se trata de um ID individual para cada observação no cadastro
- Rede
  - Se trata de um ID para cada rede de ensino no cadastro, cada rede pode possuir mais de uma escola
- Escola
  - Se trata de um ID para cada escola no cadastro
- Turma
  - Se trata de um ID para cada turma no cadastro
- Port
  - Se trata da nota no teste de portugues de cada aluno, é tambem a variável de interesse desse trabalho

Os metodos utilizados foram:

- Amostragem Estratificada
  - Foram testados estratificação por Rede e estratificação por Escola
- Amostragem Conglomerada
  - 1 estágio por Escolas

  - 1 estágio por Turmas
  - 2 estágios: UPA-Escolas, USA-Turmas
  - 3 estágios: UPA-Escolas, USA-Turmas, UTA-Alunos

- Amostragem Conglomerada com PPT Poisson
  - 1 estágio por Escolas, tamanho via número de turmas
  - 1 estágio por Escolas, tamanho via número de alunos

Estudo de Simulação

Considerou-se como variavel de interesse a média da variavel Port com transformação logaritmo natural, ou seja, a variavel estimada via diferentes metodos de amostragem complexa foi: $ln(\bar{Port})$

Para a cada plano amostral, foram replicadas 1000 vezes amostras de tamanho 500 e 1000 vezes amostras de tamanho 750, para cada *pool* foram calculadas as estimativas pontuais, o erro padrão e o intervalo de confiança de 95%

Para avaliar o desempenho de cada plano amostral. foram consideradas metricas como **Víes, Erro-padrão e Erro Quadrático Médio**.

O verdadeiro valor da variavel estimada é de:

$$\bar{Port} = 511.7484$$

E portanto:

$$ln(\bar{Port}) = ln(511.7484) = 6.237833$$

O conhecimento de talk valor é importantíssimo para o calculo do viés e consequente a decisão sobre o plano maostral mais adequado para o problema

```
set.seed(123)
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.1.3
```

```
## Warning: package 'ggplot2' was built under R version 4.1.3
```

```
## Warning: package 'tibble' was built under R version 4.1.3
```

```
## Warning: package 'tidyr' was built under R version 4.1.3
```

```
## Warning: package 'readr' was built under R version 4.1.3
```

```
## Warning: package 'purrr' was built under R version 4.1.3
```

```
## Warning: package 'dplyr' was built under R version 4.1.3
```

```
## Warning: package 'stringr' was built under R version 4.1.3
```

```
## Warning: package 'forcats' was built under R version 4.1.3

## Warning: package 'lubridate' was built under R version 4.1.3

## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.2     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.2     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.1
## -- Conflicts ---------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(survey)
```

```
## Warning: package 'survey' was built under R version 4.1.3

## Carregando pacotes exigidos: grid
## Carregando pacotes exigidos: Matrix

## Warning: package 'Matrix' was built under R version 4.1.3

##
## Attaching package: 'Matrix'
##
## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack
##
## Carregando pacotes exigidos: survival

## Warning: package 'survival' was built under R version 4.1.3

##
## Attaching package: 'survey'
##
## The following object is masked from 'package:graphics':
##
##     dotchart
```

```
library(sampling)
```

```
## Warning: package 'sampling' was built under R version 4.1.3

##
## Attaching package: 'sampling'
##
## The following objects are masked from 'package:survival':
##
##     cluster, strata
```

```
df_alunos = read.table('https://raw.githubusercontent.com/Gustavo039/survey_articles/main/Alunos.txt', h
```

Amostragem Estratificada

```
aes_unif_size=function(n_size){
  ret = ((rep(1/3,3) * n_size) |> ceiling() )
  return(ret)
}

aes_prop_size = function(n_size){
  df_alunos |>
  group_by(rede) |>
  reframe(group_size = (n()/(df_alunos |>nrow())) |>
            {\(x) x * n_size}() |>
            ceiling())
}

aes_neyman_size =  function(n_size){

  neyman_calc = function(data, n) {
  num = length(data) * sd(data)
  return(as.data.frame(num))
}

strata_size_neyman = df_alunos |>
  group_by(rede)

strata_size = strata_size_neyman |>
  group_modify(~ neyman_calc(data = .x$port,n = n_size)) |>
  ungroup() |>
  mutate(deno = sum(num)) |>
  mutate(group_size = ((num/deno)*n_size) |> ceiling())

}


AES_estimates = function(df_data = df_alunos, strat_name, rep_size, sample_size, sample_size_type)
{
  if(sample_size_type == 'uniform'){
    strata_size = aes_unif_size(sample_size)
  }
  else
    if(sample_size_type == 'proportional'){
    strata_size = aes_prop_size(sample_size) |>
      {\(x) x$group_size}()
   }
  else{
    strata_size = aes_neyman_size(sample_size)|>
      {\(x) x$group_size}()
  }
```

```
IAESs = replicate(n = rep_size,
                  expr = sampling::strata(df_alunos,
                       stratanames = strat_name,
                       size = strata_size,
                       method = 'srswor') |>
                    {\(x) data.frame(x$ID_unit, x$Prob)}()
)

AESs = sapply(1:rep_size, function(i) {
  df_alunos |>
  filter(aluno %in% IAESs[,i]$x.ID_unit) |>
  mutate(log_port = log(port), .keep = 'unused')
  }
)

fpc_calc = function(N,n){
  ret = (((N-n)/(N-1))**(1/2))
  return(ret)
}

AESs_estimates = sapply(1:rep_size,function(i){
  plan = svydesign(~1, strata=~rede, data = AESs[,i] |> as.data.frame(), probs=~IAESs[,i]$x.Prob)
  svymean(~log_port,plan) |>
  as.data.frame()
}
)


data.frame('Estimativas' = AESs_estimates[1,] |> unlist(),
           'ErroPadrão' = AESs_estimates[2,] |> unlist())

}

AES_estimates_unif = AES_estimates(df_alunos, strat_name = 'rede', rep_size = 1000, sample_size = 500, s

AES_estimates_prop = AES_estimates(df_alunos, strat_name = 'rede', rep_size = 1000, sample_size = 500, s

AES_estimates_neyman = AES_estimates(df_alunos, strat_name = 'rede', rep_size = 1000, sample_size = 500
```

Temos as seguintes estatísticas após 1000 replicações:

```
aess_table = function(data){
  ret = data |>
  modify(mean) |>
  slice(1) |>
  mutate(Vies = Estimativas - 6.237833) |>
  mutate(EQM = Vies^2 + ErroPadrão^2)

  return(ret)
}

AES_estimates_table =  list(AES_estimates_unif, AES_estimates_prop, AES_estimates_neyman) |>
  map(aess_table)
```

```
AES_estimates_table = bind_rows(AES_estimates_table[[1]], AES_estimates_table[[2]], AES_estimates_table
rownames(AES_estimates_table) = c('Uniforme', 'Proporcional', 'Neyman')
```