

MAT02036 - Amostragem 2

Aula 19 - Amostragem Sistemática - Introdução, Parâmetros e Estimação

Markus Stein

Departamento de Estatística, IME/UFRGS

2022/2

Housekeeping

- Aproveitem o momento presencial para tirar dúvidas
- Se estivéssemos no ensino remoto ou à distância
 - vocês poderiam estar somente ouvindo, sem interação
 - ou assistindo vídeos e material em outro momento
- Depois das aulas, rever material da aula passada
 - fazer exercícios
 - se preparar para a próxima aula

Aula passada

Avaliação Parcial 2

Questão 1

Considere uma população com $N = 8$ indivíduos, onde

$$Y = (9, 10, 11, 17, 20, 31, 32, 30).$$

a. Seja a divisão A desta população:

$$U_A = (C_1, C_2) = ((9, 10, 11, 17), (20, 31, 32, 30)).$$

Calcule o coeficiente de correlação intraclasse e o interprete. Qual é o menor valor que o coeficiente pode assumir nesse caso?

b. Considere agora a divisão B :

$$U_B = (C_1, C_2) = ((10, 20, 30, 11), (32, 9, 17, 31)).$$

Calcule o coeficiente de correlação intraclasse. Compare os resultados das duas divisões.

c. Na divisão A você recomendaria utilizar um plano AC1S ou AAS2 E na

Aula passada

Avaliação Parcial 2

Questão 2

Uma empresa de táxis deseja estudar a situação dos pneus dos veículos da sua frota, que é composta por 175 táxis. Para tanto, uma amostra de 10 táxis foi selecionada com reposição e, para cada um, se avaliou o número de pneus (dentre os 4 pneus em uso) que estavam fora de condições de segurança. Os resultados obtidos foram:

2, 4, 0, 1, 2, 0, 4, 1, 3, 1

- Estime a proporção de pneus da frota fora de condições pontualmente e por *IC* 95%.
- Usando esses resultados como um estudo piloto, qual seria o número de táxis necessário para obter uma estimativa da proporção de pneus fora das condições, com um erro absoluto de 2,5% e 95% de confiança? Considere **AC1s** com reposição.

Aula passada

Avaliação Parcial 2

Questão 3

Considere o banco de dados `agpop` do pacote `SDaA` do R. Após instalar o pacote, ao executar os comandos abaixo o banco de dados será carregado e poderá ser utilizado. Considere que os dados se referem a população de distritos dos EUA.

```
library(SDaA)
data(agpop)
```

Responda:

- Usando o seu cartão UFRGS como semente aleatória (`set.seed(XXXXXXX)`, onde `XXXXXXX` é o número do seu cartão), sorteie uma **AC1S** de 15 estados (variável `state`) sem reposição.
- A partir da amostra sorteada, obtenha e apresente a estimativa pontual e por *IC* 95% da média da variável `largef92`. Interprete os resultados.
- Produza dois gráficos que descrevam a variável `largef92`: um na

Amostragem Sistemática

Amostragem Sistemática

Introdução

- A *Amostragem Sistemática Simples* - **AS** seleciona com **equiprobabilidade** unidades de uma população
 - Alternativa à **AAS** bastante utilizada na prática.
- **AS** pode ser aplicada mesmo **quando não existe cadastro prévio** da população de pesquisa,
 - o cadastro ser construído ao mesmo tempo em que é feita a seleção da amostra.
- *O método*: selecionar cada K -ésima unidade da população;
 - seja N o tamanho da população e n o tamanho da amostra, selecione um **valor inicial** r ao acaso entre os números inteiros de 1 a K ;
 - A amostra formada pelas unidades $U_r, U_{r+K}, U_{r+2K}, \dots, U_{r+(n-1)K}$ é **sem reposição** e todas as unidades da população têm a **mesma chance**, $1/K$, de serem selecionadas.

Amostragem Sistemática Simples

Introdução

Seja $U = \{1, 2, \dots, N\}$ a população de pesquisa, de tamanho $N = nK + c$, com $0 \leq c < K$.

- O número inteiro K define o chamado **intervalo de seleção**,
- o tamanho amostral $n = \lfloor N/K \rfloor$ é a parte inteira de N/K , e c é o resto dessa divisão.

Exemplo (Bussab e Bolfarine, apostila pg. 37)

Seja $N = 24$ e $n = 4$. O intervalo de amostragem é $k = 6$.

- A primeira unidade será selecionada por **AAS** entre as unidades 1 e 6 da lista de *unidades elementares*.
- Supondo que selecionamos $r = 5$, então as outras unidades serão:

$$2a. \text{ unidade: } r + k = 5 + 6 = 11$$

$$3a. \text{ unidade: } r + 2k = 5 + 12 = 17$$

$$4a. \text{ unidade: } r + 3k = 5 + 18 = 23$$

Amostragem Sistemática Simples

Introdução

Exemplo 2 (Livro "Amostragem: Teoria e Prática usando o R", exemplo 8.1)

Suponha que se deseja aplicar um questionário a uma amostra dos espectadores de uma peça teatral sendo encenada em determinado teatro, num determinado dia. Nesse caso não estaria disponível uma lista das pessoas que irão ao teatro naquela data. Pode-se selecionar uma **AS** utilizando os passos indicados a seguir.

1. Definir o valor de K . Por exemplo, seja $K = 10$, significa que a amostra será composta por aproximadamente um de cada dez, ou 10% dos espectadores da noite.
2. Selecionar o *valor inicial* entre 1 e 10. Suponha que seja $r = 3$.
3. Iniciar o processo de **cadastramento sequencial** dos espectadores de acordo com a ordem de chegada (ou de saída) ao teatro, numerando cada espectador cadastrado.
4. Entrevistar os espectadores selecionados por **AS**, começando pelo de número 3 e seguindo com os de números de chegada (ou saída) iguais a 13, 23, 33, etc.

Amostragem Sistemática Simples

Introdução

Duas características funcionam como um apelo para a adoção da **AS**:

- simplicidade, pois ao selecionar a primeira unidade a ser incluída na amostra, todas as demais estarão automaticamente escolhidas;
- possibilidade de aplicação mesmo quando não se tem disponível um cadastro prévio da população.
- Menos sujeita a erros do entrevistador do que outros planos.
- Na maioria dos casos a sua eficiência é próxima da **AAS**, principalmente se o sistema de referência, isto é, a lista, está numa ordem aleatória.
- **AS** pode fornecer uma amostra estratificada proporcional se a população estiver arranjada numa ordem em função da variável a ser estudada.

Desvantagens da AS: Nos casos em que o tamanho da população não é múltiplo do intervalo de amostragem, os estimadores são viesados, porém o vício é pequeno em geral.

Amostragem Sistemática Simples

Introdução

Censo IBGE

- As características acima foram determinantes para adoção de **AS**, em cada setor censitário, dos domicílios que deveriam responder ao questionário da amostra nos **Censos Demográficos** realizados pelo **IBGE** de 1960 até 2000.
- Com esse método, cada recenseador utilizava um formulário denominado **Folha de coleta do setor** ao percorrer um setor censitário.
 - O formulário servia para cadastrar as unidades da população (domicílios) encontradas em cada setor, instrumento para a seleção da **AS**.
 - Linhas marcadas em sombreado indicavam que domicílios seria aplicado o questionário mais longo, denominado da amostra, em lugar do questionário simplificado (denominado básico), aplicado aos demais domicílios não incluídos na amostra.

Amostragem Sistemática Simples

Introdução

Censo IBGE

- Ao terminar de percorrer o setor censitário, o recenseador teria um cadastro dos domicílios de seu setor e selecionado a amostra correspondente.
 - O **cadastro de domicílios** construído era utilizado como base para o trabalho de campo das **outras pesquisas por amostragem** realizadas pelo **IBGE** ao longo da década subsequente ao Censo. Ver, por exemplo, @Albieri2015.
- Nos censos de 1960, 1970 e 1980 o intervalo de seleção usado pelo **IBGE** foi sempre com $K = 4$. Começando no Censo de 1991, o **IBGE** passou a usar valores de K que podiam variar conforme o tamanho do município em que a amostra estava sendo selecionada.
- Maiores detalhes sobre a amostragem nos Censos de 2000 e 2010 podem ser vistos em @IBGE2003 e @IBGE2016, respectivamente. Uma revisão dos aspectos de amostragem dos Censos Demográficos brasileiros desde 1960 pode ser encontrada em @Albieri2017.

Amostragem Sistemática Simples

Método de seleção da amostra

O método de seleção de uma **AS** pode ser generalizado consistindo nos seguintes passos:

1. Defina o valor de K , que determina o *intervalo de seleção* da AS.
2. Selecione a *partida aleatória* r , igual a um número inteiro sorteado entre 1 e K com probabilidades iguais a $1/K$ para todos os inteiros no intervalo; sendo assim, r tem distribuição Uniforme Discreta de parâmetro K , ou seja, $r \sim UD(K)$.
3. Inclua na amostra sistemática s_r todas as unidades que satisfazem a regra indicada abaixo.

$$s_r = \{i : i = r + lK \leq N; \ l = 0, \dots, n\}$$

Em consequência desse método, há exatamente K *amostras sistemáticas* distintas possíveis. Todas têm igual probabilidade de ser a amostra selecionada, logo:

$$p(s) = \begin{cases} 1/K, & \text{se } s = s_r \text{ para } r = 1, 2, \dots, K \\ 0, & \text{caso contrário} \end{cases}$$

Amostragem Sistemática Simples

Método de seleção da amostra

- O **tamanho efetivo** da amostra selecionada por **AS**, aqui denotado por n_r , não é fixado a priori, depende de r e pode tomar dois valores possíveis:

$$n_r = \begin{cases} n + 1, & \text{quando } r \leq c \\ n, & \text{quando } r > c \end{cases}$$

- O *tamanho efetivo* da amostra será $n + 1$ com probabilidade c/K e n com probabilidade $1 - (c/K)$.

Exemplo 4: (Livro "Amostragem: Teoria e Prática usando o R", exemplo 8.2)

Considere uma população de $N = 20$ unidades, da qual se quer selecionar uma amostra sistemática simples com intervalo de seleção definido com $K = 5$. Nesse caso, existem 5 amostras distintas, todas de tamanho $n = 4$.

Amostragem Sistemática Simples

Método de seleção da amostra

Exemplo 4: (Livro "Amostragem: Teoria e Prática usando o R", exemplo 8.2)

```
## livro 'amostragem: teoria e pratica usando o R', exemplo 8.2
## populacao
N <- 20          # tamanho da populacao
U <- 1:N         # Geração da lista das unidades da população U
K <- 5          # intervalo de selecao
##amostra
n <- floor(N/K)   # tamanho desejado da amostra
c <- N - n * K    # parte de N nao multipla de K
r <- sample(1:K, 1) # valor inicial
s_r <- subset(U, (U%%K)==r) # unidades da amostra
s_r              # AS selecionada

## [1] 3 8 13 18
```

Amostragem Sistemática Simples

Método de seleção da amostra $N \neq nK$

- Em geral, N não é múltiplo de k , portanto, diferentes amostras sistemáticas selecionadas da mesma população podem ter uma diferença de uma unidade no tamanho da amostra.

Exemplo 3: (Apostila pg. 38)

Com $N = 23$ e $k = 5$, as diferentes amostras sistemáticas são:

Amostragem Sistemática Simples

Método de seleção da amostra $N \neq nK$

- Abaixo, cada linha, r , é formada por uma das possíveis amostras, s_r .
 - Se $N \neq nK$, algumas das últimas células são vazias
 - O tamanho efetivo da amostra é n ou $n - 1$, dependendo do valor de r selecionado.

Unidades que compõem as possíveis K amostras sistemáticas, $s_1, s_2, \dots, s_r, \dots, s_K$, com partida aleatória r no intervalo $[1; K]$

Possíveis amostras		Índices	das	unidades	U_i
s_1	1	$K + 1$	$2K + 1$...	$(n - 1)K + 1$
s_2	2	$K + 2$	$2K + 2$...	$(n - 1)K + 2$
...
s_r	r	$K + r$	$2K + r$...	$(n - 1)K + r$
...
s_K	K	$2K$	$3K$...	-

Amostragem Sistemática Simples

Método de seleção da amostra $N \neq nK$

Exemplo 5:

Identifique as amostras sistemáticas simples possíveis quando a população tem $N = 19$ unidades e o tamanho desejado da amostra é de $n = 4$ unidades.

Como $N = 19 = 4 \times 4 + 3$, temos que $K = 4$ e $c = 3$. Logo, as quatro amostras sistemáticas possíveis nesse caso são:

$$s_1 = \{1; 5; 9; 13; 17\} \text{ com } n_1 = 5;$$

$$s_2 = \{2; 6; 10; 14; 18\} \text{ com } n_2 = 5;$$

$$s_3 = \{3; 7; 11; 15; 19\} \text{ com } n_3 = 5;$$

$$s_4 = \{4; 8; 12; 16\} \text{ com } n_4 = 4.$$

Amostragem Sistemática Simples

Método de seleção da amostra $N \neq nK$

Exemplo 5:

```
N=19 # Tamanho da População  
print(paste("Tamanho da população N:", N), quote=FALSE)
```

```
## [1] Tamanho da população N: 19
```

```
n=4 # Tamanho desejado da amostra  
print(paste("Tamanho desejado da amostra n:", n), quote=FALSE)
```

```
## [1] Tamanho desejado da amostra n: 4
```

```
K=trunc(N/n) # Cálculo do intervalo de seleção  
print(paste("Intervalo de seleção K:", K), quote=FALSE)
```

```
## [1] Intervalo de seleção K: 4
```

```
c=N-n*K # Cálculo da constante c  
print(paste("Constante c:", c), quote=FALSE)
```

Amostragem Sistemática Simples

Método de seleção da amostra $N \neq nK$

Exemplo 6:

Calcule o tamanho efetivo da amostra resultante da seleção sistemática em uma população com $N = 149$ unidades, quando o tamanho desejado da amostra sistemática simples é de $n = 60$ unidades. Como $N = 149 = 60 \times 2 + 29$, resulta que $K = 2$ com $n = 74$ e $c = 1$. Sendo assim as duas únicas amostras possíveis são:

$s_1 = \{\text{números ímpares até 149, inclusive}\}$ com $n_1 = 75$.

$s_2 = \{\text{números pares até 148, inclusive}\}$ com $n_2 = 74$.

Nesse caso, verifica-se que o tamanho efetivo da amostra poderá ser 74 ou 75, um pouco maiores que o tamanho desejado de 60.

Amostragem Sistemática Simples

Probabilidades de inclusão na AS

É fácil notar que a probabilidade de inclusão na amostra de uma unidade i qualquer é dada por:

$$\pi_i = \frac{1}{K}, \quad i = 1, \dots, N$$

- A primeira unidade a ser incluída na amostra é a unidade r , que é um inteiro selecionado com equiprobabilidade no intervalo $[1; K]$.
 - Como há K números inteiros nesse intervalo, segue-se que a probabilidade de sortear um qualquer desses números é $1/K$.
 - As demais unidades selecionadas são obtidas somando a r os múltiplos lK , com l variando de 1 a n , enquanto $lK + r \leq nK + c$.

A probabilidade de inclusão das unidades $i \neq j$ na amostra é dada por:

$$\pi_{ij} = \begin{cases} 1/K, & \text{se } i \neq j \in s_r \text{ para } r = 1, \dots, K \\ 0, & \text{caso contrário} \end{cases}$$

Amostragem Sistemática Simples

Variáveis aleatórias indicadoras

- Ao escolher r , $1 \leq r \leq K$, selecionamos a amostra inteira. Sejam as variáveis indicadoras associadas às amostras possíveis s_r :

$$I(r) = \begin{cases} 1, & \text{se a amostra é } s_r \text{ para } 1 \leq r \leq K \\ 0, & \text{caso contrário} \end{cases}$$

- O valor esperado de $I(r)$ é

$$E_{AS}[I(r)] = 1/K, \quad r = 1, \dots, K,$$

a variância

$$Var_{AS}[I(r)] = E_{AS}\{[I(r)]^2\} - \{E_{AS}[I(r)]\}^2 = \frac{1}{K} - \frac{1}{K^2} = \frac{1}{K} \left(1 - \frac{1}{K}\right)$$

e a covariância entre $I(r)$ e $I(q)$ quando $r \neq q$ é:

$$Cov_{AS}[I(r), I(q)] = E_{AS}[I(r)I(q)] - E_{AS}[I(r)]E_{AS}[I(q)] = 0 - \frac{1}{K^2} = -\frac{1}{K^2}$$

já que apenas uma das duas partidas r ou q pode ser selecionada.

Amostragem Sistemática Simples

Estimação de totais na AS

- O estimador tipo Horvitz-Thompson do total sob **AS**, denotamos por t_r a soma amostral dos valores observados da variável y para a amostra s_r , definida como:

$$t_r = \sum_{i \in s_r} y_i$$

- Como a probabilidade de inclusão de uma amostra sistemática simples s_r qualquer é $1/K$, o peso amostral das unidades dessa amostra é sempre igual a $d_i = 1/\pi_i = K$. Sendo assim, sob AS o estimador de Horvitz-Thompson para o total é dado por:

$$\hat{T}_{AS} = K t_r = K \sum_{i \in s_r} y_i$$

Com base nas propriedades do estimador de Horvitz-Thompson para o total, já sabemos que este estimador é não viciado para o total populacional. Mas vamos aqui demonstrar esse resultado também para o caso particular da AS, pois a prova nos ajudará com a obtenção posterior de expressão para a variância do estimador sob esse plano amostral. Note então que:


Para casa

- Fazer a lista 2 de exercícios.
- Continuar exercícios.
- Rever os slides.
- Preparação para avaliação parcial 2

Próxima aula

- Acompanhar o material no moodle.

Amostragem por Conglomerados

- Exercícios.
- Laboratório de 

Muito obrigado!



Fonte: imagem do livro *Combined Survey Sampling Inference: Weighing of Basu's Elephants*.

Referências

- Amostragem: Teoria e Prática Usando o R
- **Elementos de Amostragem**, Bolfarine e Bussab.
- Cochran(1977)

Resumo da notação