



“R” Com Data Science

Gabrielle Gomes dos Santos Ribeiro
2024



AMOSTRAGEM



INTRODUÇÃO

No cotidiano é comum nos depararmos com perguntas que aos olhos, requer muito tempo e dinheiro para serem respondidas. Como por exemplo:

- Quantos eleitores irão votar em certo candidato à presidência?
 - Da população de uma determinada cidade, quantas pessoas são idosas, quantas vivem na área rural e quantas estão desempregadas?
 - Qual o nível de poluição da água do mar?
-



AMOSTRAGEM

- Ao invés de entrevistarmos uma população inteira, existe outro processo possível, que consiste em consultar apenas um grupo de pessoas dessa população, que constituem uma AMOSTRA.
 - Se a amostra representa de fato toda a população, podemos utilizar as características dos seus elementos para estimar as características de toda população.
-



AMOSTRAGEM

- A amostragem é uma técnica ou conjunto de procedimentos necessários para descrever e selecionar amostras de uma população, e quando realizada com técnicas adequadas, é um fator responsável pela determinação da representatividade da população em questão.
 - Algumas das vantagens da utilização da amostragem são:
 - economia de tempo
 - redução de custos
 - a obtenção de resultados menos propícios ao erro.
-



OBJETIVOS

- Objetivo primário ao estabelecer um plano de amostragem → promover um levantamento de dados o mais representativo possível da área avaliada, considerando-se um custo de investigação já fixado, ou se possível que seja minimizado.
 - Segundo objetivo → a adoção de um esquema de amostragem simples e eficiente, que facilite a análise dos dados e a sua **implantação em campo**.
 - O processo pode ser:
 1. Probabilístico
 2. Não-probabilístico
-



Amostragem Não-Probabilística

- A escolha dos elementos da amostra é feita através de um procedimento de seleção, segundo critérios estabelecidos pelo pesquisador, portanto alguns elementos não têm nenhuma chance de serem escolhidos.
 - Ao usar a amostragem não probabilística o pesquisador não sabe qual é a probabilidade que um elemento da população tem de pertencer à amostra. Portanto, os resultados da amostra não podem ser estatisticamente generalizados para a população, porque não se pode estimar o erro amostral.
-

Amostragem Não-Probabilística

Amostra por conveniência

O pesquisador seleciona membros da população mais acessíveis.

Ex.: - Solicitar a pessoas que voluntariamente testem um produto e que em seguida respondam a uma entrevista.

- Colocar linhas de telefone adaptadas para que durante um programa de televisão os telespectadores possam dar suas opiniões

Amostra por julgamento

O pesquisador usa o seu julgamento para selecionar os membros da população que são boas fontes de informação precisa.

Ex.: Entrevista com os representantes de turma do curso de turismo, aplicação de questionários com os líderes da comunidade.

Amostragem Probabilística

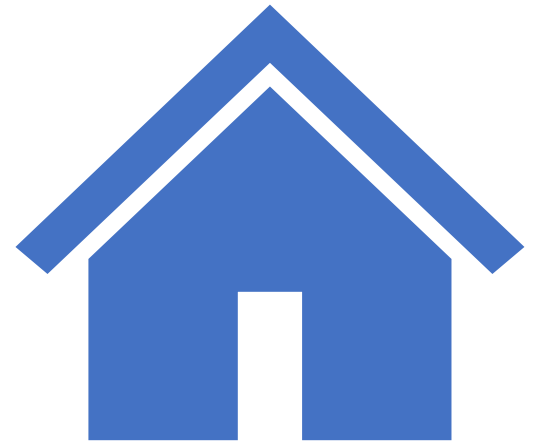
- A amostragem probabilística também é chamada de aleatória ou casual. Caracterizado pelo fato de **todos os elementos da população terem alguma chance não nula e conhecida de serem selecionados**;
- A sua importância decorre do fato de que apenas **os resultados provenientes de uma amostra probabilística podem ser generalizados estatisticamente para a população** da pesquisa. O que significa estatisticamente? Significa que podemos associar aos resultados uma probabilidade de que estejam corretos, ou seja uma medida da confiabilidade das conclusões obtidas. Se a amostra não for probabilística não há como saber se há 95% ou 0% de probabilidade de que os resultados sejam corretos.

Amostragem Probabilística

- É necessário possuir **uma listagem com os elementos da população**. Em suma, exige acesso a todos os elementos da população;

PRINCIPAIS TIPOS DE AMOSTRAGEM:

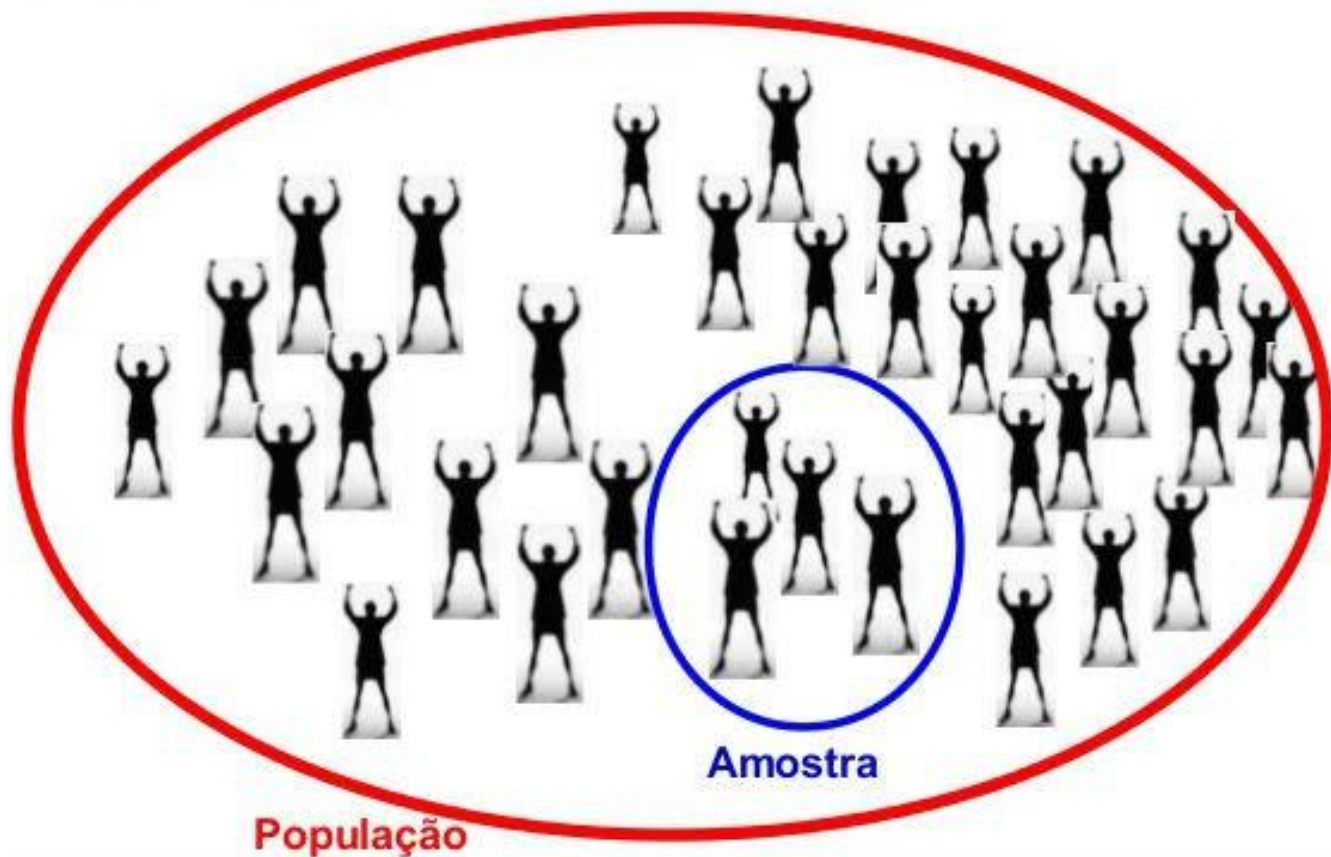
- Amostragem Aleatória Simples
- Amostragem Aleatória Estratificada
- Amostragem Sistemática



AMOSTRAGEM ALEATÓRIA SIMPLES

- É o método mais simples de amostragem probabilística.
- “De uma lista com N unidades elementares, sorteiam-se com igual probabilidade n unidades” (BOLFARINE; BUSSAB, 2005).
- Cada unidade desta população tem a mesma probabilidade igual a $1/N$ de entrar na amostra.
- Comparando-se com outros métodos → oferece o pior resultado, podendo ter áreas não amostradas e áreas com pontos agrupados.





AMOSTRAGEM ALEATÓRIA SIMPLES

- No R – Vamos usar como exemplo o conjunto “amostra.csv”

```
dados <- read.csv("amostra.csv", header=T, sep=";")
```

```
set.seed(50) #Definição da semente a ser utilizada
```

Função *runif* → *runif*(*n*, *mínimo*, *máximo*)

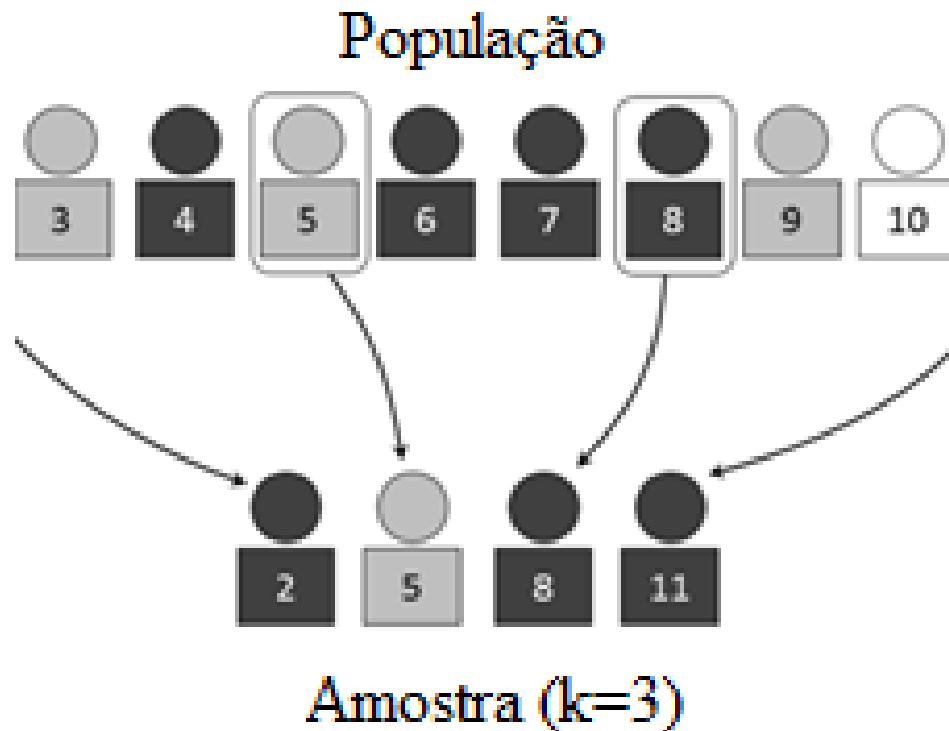
```
amostra <- dados[runif(10,1,nrow(dados)),] #Definição da amostra
```

AMOSTRAGEM SISTEMÁTICA

- É uma variação da amostragem aleatória.
- A população deve ser ordenada de modo tal que cada um de seus elementos possa ser identificado pela sua posição.
- Considere uma população com N elementos e k um número inteiro, tal que $k = N/n$.
- Seleciona-se então aleatoriamente um número, simbolizado por r , entre 1 e k , que será o 1º elemento da amostra.
- Os outros elementos seguintes são obtidos a partir dessa primeira unidade, selecionadas em intervalos de comprimento k .

$$(r, r+k, r+2k, \dots, r+(i-1)k), i=1, \dots, n.$$

Amostragem Sistemática(AS)



- Resumindo, os elementos da amostra serão selecionados obedecendo a um determinado intervalo (k) entre eles:

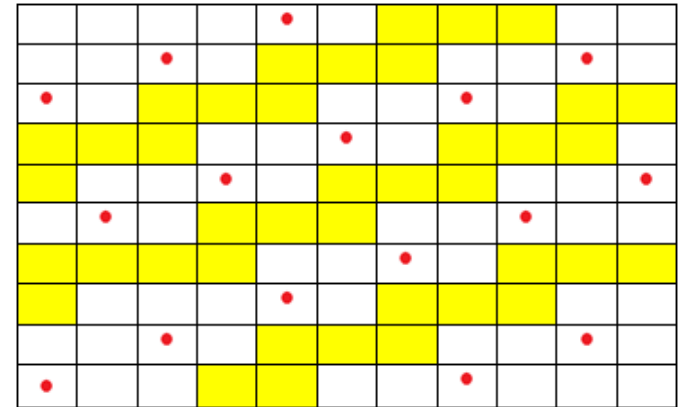
Amostragem Sistemática(AS)

Por exemplo, se o tamanho da população for $N = 1600$ e o tamanho da amostra for $n = 100$, tem-se que $k = 16$. Sorteia-se então um número entre 1 e 16 (de forma aleatória), que será o primeiro número da amostra, logo as próximas unidades amostrais serão retiradas de 16 em 16, até obter as 100 unidades. Supondo que o primeiro número sorteado foi 10, a amostra ficaria da seguinte maneira:

10, 26, 42, 58, 74, ..., 1594.

Amostragem Sistemática(AS)

- **CUIDADO** → verificar se os elementos da população não apresentam nenhum ciclo ou periodicidade, o que poderá inviabilizar o uso dessa metodologia de seleção de amostras.



	.				.				.

Amostragem Sistemática(AS)

- **No R**

```
N <- nrow(dados)
```

```
n <- 10
```

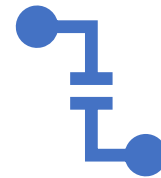
```
k <- round(N/n)
```

```
set.seed(22222) #Definição da semente
```

```
n0 <- runif(1, min=1, max=k) #Definição do 1º elemento da amostra
```

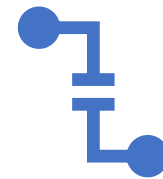
```
sequencia <- seq(n0, n0 + (n-1) * k, k) #Metodologia da amostragem
```

```
amostra2 <- dados[sequencia,]
```



Amostragem Sistemática(AS)

	ID	EstCivil	Renda	Idade
7	7	Casado	8.25	18
27	27	Casado	12.80	61
47	47	Divorciado	3.45	30
67	67	Casado	12.80	30
87	87	Casado	14.55	32
107	107	Casado	7.00	23
127	127	Casado	4.65	18
147	147	Solteiro	8.10	25
167	167	Casado	8.75	21
187	187	Divorciado	2.35	20




Amostragem Sistemática(AS)

- Outra alternativa: Usar o função *S.SY*

```
install.packages("TeachingSampling")  
library(TeachingSampling)
```

```
N<-nrow(dados)  
n<-10  
K<-round(N/n)  
amostra<-S.SY(N,k)
```



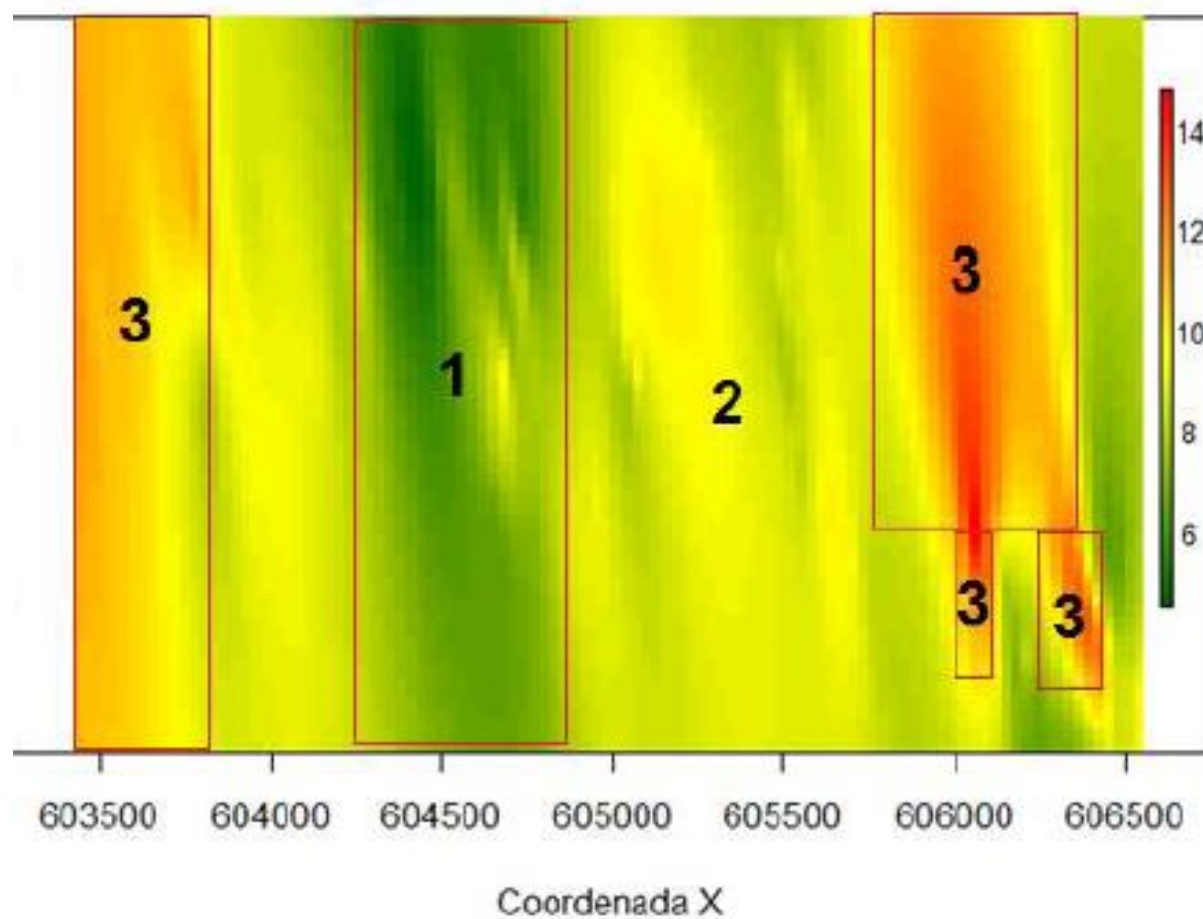
Pausa para exercício..

- Considere o conjunto Milsa.txt e selecione uma amostra de 12 funcionários mediante o processo sistemático.

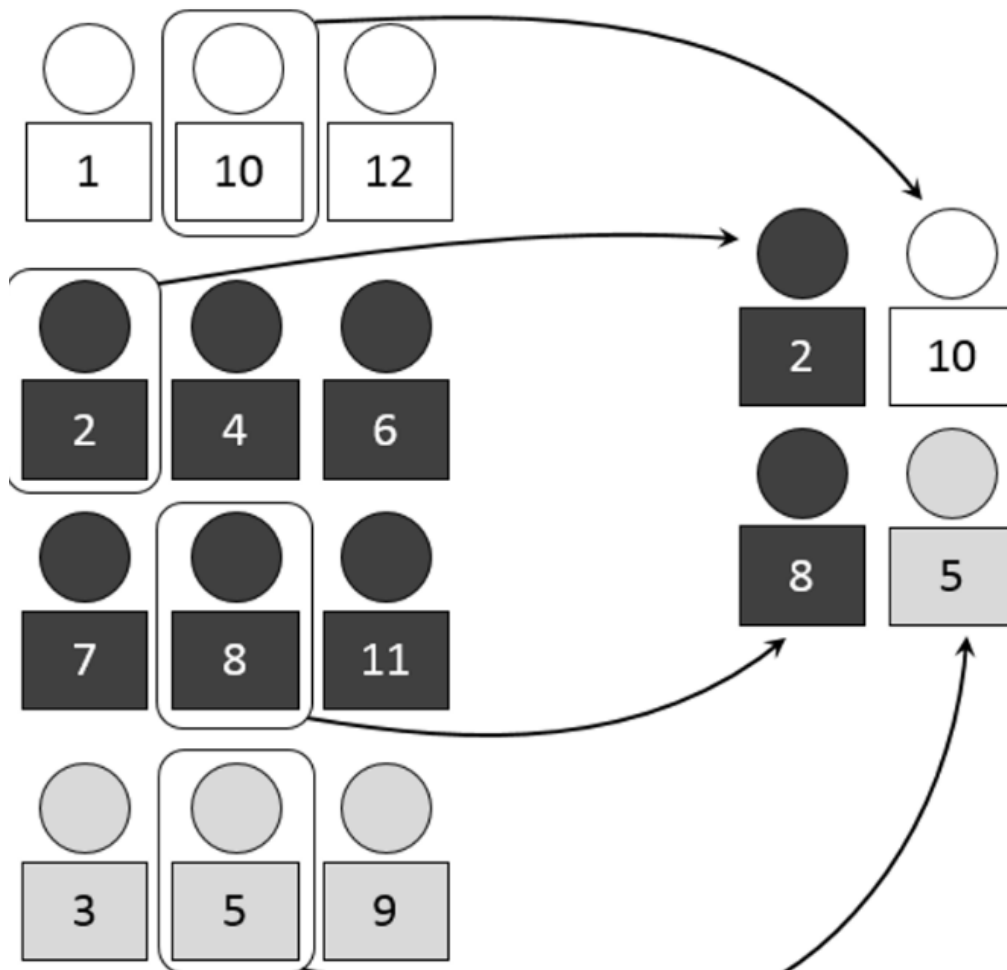
AMOSTRAGEM ALEATÓRIA ESTRATIFICADA

- Esta técnica de amostragem usa informação existente sobre a população para dividi-la em grupos bem definidos, chamados estratos (homogêneos dentro deles e heterogêneos entre si).
- De cada um desses estratos, é selecionada uma amostra mediante um processo aleatório simples.
- Tem a vantagem de fornecer resultados com menor probabilidade de erro associada.
- Esse esquema de amostragem assegura que todas as subáreas que compõe o local em estudo sejam amostradas.
- Desvantagem → nem sempre se consegue obter subpopulações bem distintas, é comum que os estratos fiquem sobrepostos.

Krigagem Ordinária da variável clorofila-a



AMOSTRAGEM
ESTRATIFICADA



Amostragem

AMOSTRAGEM
ESTRATIFICADA

Amostragem Aleatória Estratificada (AAE)

- Este método de amostragem estratificada tem a vantagem de fornecer resultados com menor probabilidade de erro associada.
- Em uma AAE a população de tamanho N é dividida em L estratos de N_1, N_2, \dots, N_L unidades, tal que:

$$N = N_1 + N_2 + \dots + N_L.$$

Quando os estratos são determinados, uma AAS é retirada de dentro de cada estrato independente. O tamanho amostral dentro de cada estrato é n_1, n_2, \dots, n_L , tal que:

$$n = n_1 + n_2 + \dots + n_L.$$

Amostragem Aleatória Estratificada (AAE)

- A distribuição das “ n ” unidades da amostra pelos estratos é feita proporcionalmente à quantidade de elementos dentro de cada estrato. O tamanho amostral de cada estrato é definido por:

$$n_L = n \cdot W_L$$

onde W_L é o peso ou a proporção do estrato L dentro da população, definido por

$$W_L = N_L / N$$

Amostragem Aleatória Estratificada (AAE)

EXEMPLO (Scheaffer, 1986): Uma agência de opinião pública tem que realizar uma pesquisa num município que tem 2 bairros e uma área rural. Os elementos de interesse na população são todos os adultos acima de 21 anos. Uma AAE pode ser obtida deste município, isto é, os dois bairros e a área rural representam 3 estratos separados dos quais podemos obter uma AASs de cada um. Sabemos que o Bairro A tem 155 domicílios, o Bairro B tem 62 domicílios e na área rural temos 93 domicílios. Com o objetivo de saber quantas horas semanais uma família gasta assistindo TV, uma agência decide selecionar uma amostra de 40 domicílios, coletadas proporcionalmente ao tamanho do estrato. Quantos domicílios serão selecionados de cada estrato?

Amostragem Aleatória Estratificada (AAE)

$$N = 155 + 62 + 93 = 310$$
$$n = 40$$

Primeiro vamos calcular o peso (w_h) e a média de cada estrato:

$$w_1 = \frac{155}{310} = 0,5 \quad ; \quad w_2 = \frac{62}{310} = 0,2 \quad ; \quad w_3 = \frac{93}{310} = 0,3$$

Agora vamos calcular a quantidade de amostras que será coletada de cada estrato:

$$n_1 = n \cdot w_1 = 40 * 0,5 = 20 \quad ; \quad n_2 = 40 * 0,2 = 8 \quad ; \quad n_3 = 40 * 0,3 = 12$$

AMOSTRAGEM ESTRATIFICADA no R

```
NL=as.numeric(table(dados$Regiao))
```

```
n=10
```

```
n1=(87/200)*n
```

```
n2=(23/200)*n
```

```
n3=(52/200)*n
```

```
n4=(38/200)*n
```

```
round(n1)
```

```
round(n2)
```

```
round(n3)
```

```
round(n4)
```

```
nL=c(4,1,3,2)
```

```
#Definição da amostra
```

```
amostra2 <- S.STSI(as.factor(dados$Regiao), NL, nL)
```

AMOSTRAGEM ESTRATIFICADA no R

```
> NL=as.numeric(table(dados$Região))
> n=10
> n1=(87/200)*n
> n2=(23/200)*n
> n3=(52/200)*n
> n4=(38/200)*n
> round(n1)
[1] 4
> round(n2)
[1] 1
> round(n3)
[1] 3
> round(n4)
[1] 2
> nL=c(4,1,3,2)
>
> amostra2 <- S.STSI(as.factor(dados$Região), NL, nL)
> amostra2
      [,1]
[1,]  114
[2,]  149
[3,]   10
[4,]  102
[5,]  126
[6,]   36
[7,]   30
[8,]   31
[9,]   19
[10,] 185
```