

MAT02025 - Amostragem 1

Amostragem aleatória simples

Rodrigo Citton P. dos Reis
citton.padilha@ufrgs.br

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE MATEMÁTICA E ESTATÍSTICA
DEPARTAMENTO DE ESTATÍSTICA

Porto Alegre, 2023

Amostragem aleatória simples

Amostragem aleatória simples

- ▶ Os **levantamentos por amostragem** dizem respeito a amostras retiradas de populações que contêm um número finito, N , de unidades.
- ▶ Se todas essas unidades podem ser distinguidas umas das outras, o número de amostras distintas de tamanho n , que podem ser retiradas das N unidades, é dado pela seguinte fórmula combinatória:

$$\binom{N}{n} = {}_N C_n = \frac{N!}{n!(N-n)!}.$$

Amostragem aleatória simples

- ▶ Por exemplo, se a população contiver cinco unidades designadas por A, B, C, D e E , há dez amostras distintas de tamanho 3:

$ABC \quad ABD \quad ABE \quad ACD \quad ACE$
 $ADE \quad BCD \quad BCE \quad BDE \quad CDE$

- ▶ Note que a mesma letra não se repete dentro da mesma amostra.
 - ▶ Não se leva em consideração a ordem em que as letras aparecem na amostra, de modo que as seis amostras ABC, ACB, BAC, BCA, CAB e CBA são consideradas como uma única.

Amostragem aleatória simples

A **amostragem aleatória simples**¹ (AAS) é um processo para selecionar n unidades de N de modo que cada uma das ${}_NC_n$ amostras distintas tenha uma **chance igual de ser extraída**.

¹Também conhecida como **amostragem casual simples** ou **amostragem acidental irrestrita**

Amostragem aleatória simples

- ▶ Na prática, uma amostra aleatória simples é sorteada **unidade por unidade**.
- ▶ As unidades da população são numeradas de **1 a N** .
- ▶ Uma série de **números aleatórios** entre **1 e N** é então sorteada, por meio de uma **tabela de números aleatórios** ou por meio de um **programa de computador** que produz tal tabela.
- ▶ Em qualquer sorteio, o processo usado deve dar uma chance igual de seleção a qualquer número na população **ainda não sorteado**.
- ▶ As unidades que contêm esses **n números** **constituem a amostra** (selecionada).

Amostragem aleatória simples

- ▶ É facilmente verificado que todas as ${}_NC_n$ amostras distintas têm chances iguais de serem selecionadas por este processo.
- ▶ Considere uma amostra distinta, ou seja, um conjunto de n unidades especificadas.
- ▶ No primeiro sorteio (**o sorteio da primeira unidade da amostra**), a probabilidade de que alguma das n unidades especificadas seja selecionada é n/N .
- ▶ No segundo sorteio (**o sorteio da segunda unidade da amostra**), a probabilidade de que alguma das unidades restantes $(n - 1)$ especificadas seja retirada é $(n - 1)/(N - 1)$ e assim por diante.

Amostragem aleatória simples

- Portanto, a probabilidade de que todas as n unidades especificadas sejam selecionadas em n sorteios é

$$\frac{n}{N} \cdot \frac{(n-1)}{(N-1)} \cdot \frac{(n-2)}{(N-2)} \cdots \frac{1}{(N-n+1)} = \frac{n!(N-n)!}{N!} = \frac{1}{{}_N C_n}.$$

Amostragem aleatória simples

- ▶ Como um número sorteado é removido da população em todos os sorteios subsequentes, esse método também é chamado de amostragem aleatória **sem reposição**.
- ▶ A amostragem aleatória **com reposição** é inteiramente viável: em qualquer sorteio, todos os N membros da população têm a mesma chance de serem sorteados, não importa quantas vezes eles já tenham sido sorteados.
- ▶ As fórmulas para as variâncias e variâncias estimadas das estimativas feitas a partir da amostra são frequentemente mais simples quando a amostragem é “com reposição” do que quando é “sem reposição”.
 - ▶ Por esta razão, a amostragem com reposição é às vezes usada nos planos de amostragem mais complexos, embora à primeira vista pareça fazer pouco sentido em ter a mesma unidade duas ou mais vezes na amostra.

Seleção de uma amostra aleatória simples

Seleção de uma AAS

- ▶ As tabelas de números aleatórios são tabelas de dígitos $0, 1, 2, \dots, 9$, cada dígito tendo uma chance igual de seleção em qualquer sorteio.
- ▶ Entre as maiores tabelas estão as publicadas pela **Rand Corporation (1955)**².
- ▶ Diversas tabelas estão disponíveis, muitas em livros de estatística.
- ▶ A tabela a seguir apresenta 1000 dígitos aleatórios para ilustração, de **Snedecor e Cochran (1967)**³.

²Rand Corporation (1955). **A Million Random Digits**. Free Press, Glencoe, Ill.

³Snedecor, G. W., and Cochran, W. G. (1967). **Statistical Methods**. Iowa State University Press, Ames, Iowa, sixth edition.

Seleção de uma AAS

	00-04	05-09	10-14	15-19	20-24	25-29	30-34	35-39	40-44	45-49
00	54463	22662	65905	70639	79365	67382	29085	69831	47058	08186
01	15389	85205	18850	39226	42249	90669	96325	23248	60933	26927
02	85941	40756	82414	02015	13858	78030	16269	65978	01385	15345
03	61149	69440	11286	88218	58925	03638	52862	62733	33451	77455
04	05219	81619	10651	67079	92511	59888	84502	72095	83463	75577
05	41417	98326	87719	92294	46614	50948	64886	20002	97365	30976
06	28357	94070	20652	35774	16249	75019	21145	05217	47286	76305
07	17783	00015	10806	83091	91530	36466	39981	62481	49177	75779
08	40950	84820	29881	85966	62800	70326	84740	62660	77379	90279
09	82995	64157	66164	41180	10089	41757	78258	96488	88629	37231
10	96754	17676	55659	44105	47361	34833	86679	23930	53249	27083
11	34357	88040	53364	71726	45690	66334	60332	22554	90600	71113
12	06318	37403	49927	57715	50423	67372	63116	48888	21505	80182
13	62111	52820	07243	79931	89292	84767	85693	73947	22278	11551
14	47534	09243	67879	00544	23410	12740	02540	54440	32949	13491
15	98614	75993	84460	62846	59844	14922	48730	73443	48167	34770
16	24856	03648	44898	09351	98795	18644	39765	71058	90368	44104
17	96887	12479	80621	66223	86085	78285	02432	53342	42846	94771
18	90801	21472	42815	77408	37390	76766	52615	32141	30268	18106
19	55165	77312	83666	36028	28420	70219	81369	41943	47366	41067

Seleção de uma AAS

- ▶ Ao usar essas tabelas para selecionar uma amostra aleatória simples, o primeiro passo é numerar as unidades na população de 1 a N .
- ▶ Se o primeiro dígito de N for um número entre 5 e 9, o seguinte método de seleção é adequado.

- ▶ Suponha que $N = 528$, e queremos $n = 10$.
 - ▶ Selecione três colunas da tabela de números aleatórios, digamos as colunas 25 a 27.
 - ▶ Desça pelas três colunas, selecionando os primeiros 10 números distintos entre 001 e 528.
 - ▶ Estes são 36, 509, 364, 417, 348, 127, 149, 186, 290 e 162.
 - ▶ Para os dois últimos números, saltamos para as colunas 30 a 32.

Seleção de uma AAS

- ▶ Com $N = 128$, por exemplo, um segundo método que envolve menos rejeição e é facilmente aplicado é o seguinte.
- ▶ Em uma série de números de três dígitos, subtraia 200 de todos os números entre 201 e 400, 400 de todos os números entre 401 e 600, 600 de todos os números entre 601 e 800, 800 de todos os números entre 801 e 999 e, é claro, 000 de todos os números entre 000 e 200.
- ▶ Todos os restantes maiores que 129 e os números 000, 200, e assim por diante, são rejeitados.
- ▶ Usando as colunas 05 a 07 na tabela de números aleatórios, obtemos 26, 52, 7, 94, 16, 48, 41, 80, 128 e 92, o sorteio exigindo 15 números de três dígitos para $n = 10$.

Seleção de uma AAS

- Uma alternativa seria utilizar uma urna.



Seleção de uma AAS

- Hoje em dia possuímos “urnas eletrônicas”!

```
sample(x = 1:528, size = 10, replace = FALSE)
```

```
## [1] 354 36 334 218 247 522 189 496 28 276
```

```
sample(x = 1:128, size = 10, replace = FALSE)
```

```
## [1] 39 52 90 110 93 27 55 32 113 43
```


Definições e notação

Definições e notação

- ▶ Nos levantamentos por amostragem, decidimos sobre certas propriedades, que tentamos medir e registrar, para cada unidade incluída na amostra.
- ▶ Essas propriedades das unidades são chamadas de **características**.

Definições e notação

- ▶ Os valores obtidos para qualquer característica específica⁴ nas N unidades que compõem a população são denotados por Y_1, Y_2, \dots, Y_N .
- ▶ Os valores correspondentes para as unidades na amostra são denotados por Y_1, Y_2, \dots, Y_n ou, se desejarmos nos referir a um membro típico da amostra, por Y_i ($i = 1, 2, \dots, n$).
 - ▶ Observe que a amostra **não consistirá** nas primeiras n unidades da população, exceto no caso, geralmente raro, em que essas unidades foram sorteadas.

⁴Note que em um levantamento típico estaremos interessadas em diversas características das unidades (sexo, idade, altura, peso, renda, escolaridade, IMC, etc.). Para cada uma destas características empregaremos um letra (Y, X, Z, W, V, R, S , etc.), ou podemos utilizar a mesma letra se estamos a representar diversas variáveis, uma de cada vez.

Definições e notação

- ▶ As letras **maiúsculas** referem-se às características da população e as **minúsculas** às da amostra.
 - ▶ Para totais e médias, temos as seguintes definições.

Na população

Total

$$Y_T = \sum_{i=1}^N Y_i = Y_1 + Y_2 + \dots + Y_N.$$

Média

$$\bar{Y} = \frac{\sum_{i=1}^N Y_i}{N} = \frac{Y_1 + Y_2 + \dots + Y_N}{N}.$$

Definições e notação

Na amostra

Total

$$y_T = \sum_{i=1}^n Y_i = Y_1 + Y_2 + \dots + Y_n.$$

Média

$$\bar{y} = \frac{\sum_{i=1}^n Y_i}{n} = \frac{Y_1 + Y_2 + \dots + Y_n}{n}.$$

Definições e notação

Embora se possa realizar uma amostragem com muitas finalidades, na maioria dos casos o interesse se concentra em **quatro características da população**:

1. A **média**: \bar{Y} (por exemplo, o número médio de crianças por escola).
2. O **total** Y_T (por exemplo, o número total de acres de trigo em uma região).
3. A relação (**razão, ou índice**) entre dois totais ou médias $R = Y_T/X_T = \bar{Y}/\bar{X}$ (por exemplo, relação entre custo semanal em alimentos e tamanho da família).
4. A **proporção** de unidades que pertencem a uma determinada classe/categoria P (por exemplo, proporção de pessoas com diabetes).

Definições e notação

- ▶ A estimativa das três primeiras quantidades será discutida na primeira parte desta área.
- ▶ O símbolo “ $\hat{\cdot}$ ” denota uma estimativa de uma característica da população feita a partir de uma amostra.

Definições e notação

- De acordo com a relação acima, temos:

	Estimador
Média da população \bar{Y}	$\hat{\bar{Y}} = \bar{y} = \sum_{i=1}^n Y_i/n$
Total da população Y_T	$\hat{Y}_T = N\bar{y} = N \sum_{i=1}^n Y_i/n$
Índice da população R	$\hat{R} = \sum_{i=1}^n Y_i / \sum_{i=1}^n X_i$

Definições e notação

- ▶ Em \hat{Y}_T , o **fator N/n** , pelo qual o valor total da amostra é multiplicado, é às vezes chamado de **fator de expansão**, ou **fator de crescimento**, ou ainda de **fator de inflação**.
- ▶ Seu inverso n/N , a relação entre o tamanho da amostra e o da população, é chamado de **fração de amostragem** e é denotado pela letra f .

Para casa

- ▶ Realize um sorteio de uma amostra aleatória simples como o apresentado no exemplo de aula utilizando a tabela de números aleatórios.
- ▶ Repita o sorteio utilizando a função `sample` do R, ou algum outro recurso computacional.

Próxima aula

- ▶ Propriedades e variâncias dos estimadores.

Por hoje é só!

Bons estudos!

