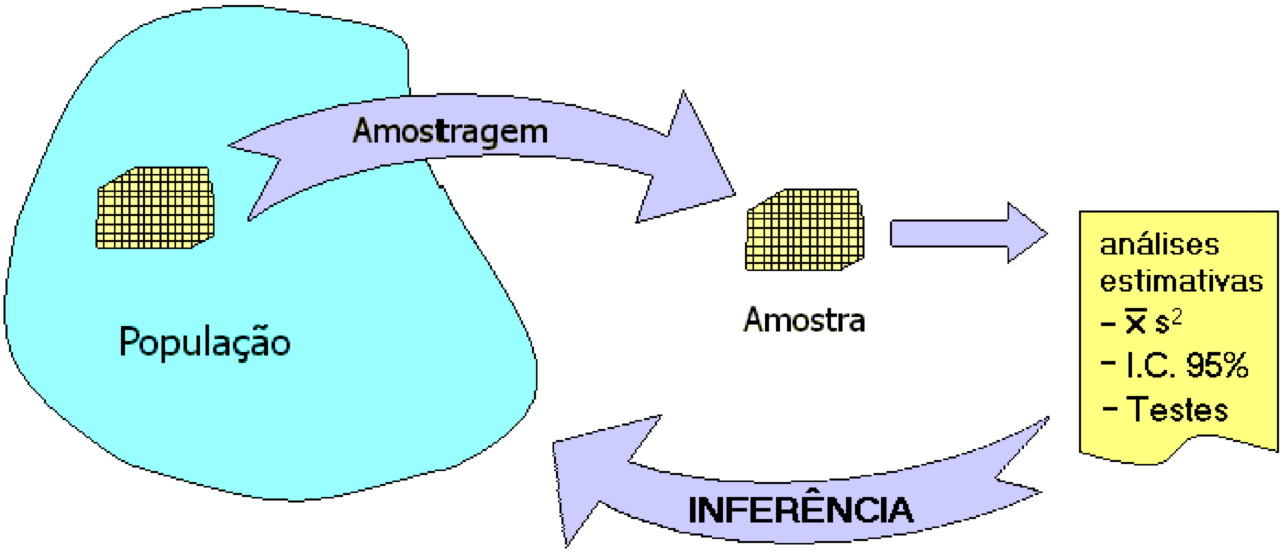
1.4- Técnicas de Amostragem

É a parte da Teoria Estatística que define os procedimentos para os planejamentos amostrais e as técnicas de estimação utilizadas.

As técnicas de amostragem, tal como o planejamento amostral, são amplamente utilizados nas pesquisas científicas e de opinião para se conhecer alguma característica da população.

Nos ***planejamentos amostrais***, a coleta dos dados deve ser realizada observando-se uma metodologia adequada para que os resultados possam ser extrapolados para a população como um todo.

Esse processo de ***extensão dos resultados para a população*** é o que, na estatística, chamamos de ***INFERÊNCIA***.



I- Conceitos

i) População e amostra: (das definições anteriores)

População objetivo: é formada pelo conjunto de indivíduos (ou elementos) que queremos abranger em nosso estudo e para os quais desejamos que as conclusões da pesquisa. Os indivíduos d a população têm ***pelo menos uma característica em comum***.

População amostral: conjunto de indivíduos da população que estão de fato acessível para serem amostrados.

ii) Amostra: é a parcela da população amostral efetivamente selecionada para a realização do estudo, segundo um processo de seleção adequado.

iii) Parâmetro: é uma ***característica*** fixa e desconhecida da ***população*** a qual se tem interesse em estudar.

Os parâmetros representam ***quantidades numéricas*** que podem ser ***interpretadas*** pelo pesquisador, como por exemplo: média; proporção; variação; taxa de crescimento; etc...

Exemplos:

- Proporção de crianças com a cobertura vacinal completa (estudo PI);
- número de latrocínios em Minas Gerais, por região administrativa;
- percentual de intenção de votos para um candidato numa pesquisa eleitoral;
- tempo até a cura de pacientes submetidos a um novo tratamento ou a uma nova droga;
- medida do desempenho escolar de crianças expostas à violência doméstica do pai contra a mãe.

iv) **Estimativa**: valor calculado a partir dos dados obtidos pela ***amostra*** para se ***estimar o valor desconhecido*** do parâmetro.

Exemplo: média amostral, proporção amostral, variância amostral, etc...

v) **Unidade amostral**: é o ***indivíduo*** (ou elemento) da população amostral sobre o qual a medida de interesse será observada.

As ***unidades amostrais*** podem ser os próprios elementos da população amostral ou podem ser formadas por ***grupos de elementos***, compondo o que será chamado de ***conglomerado***.

Conglomerados podem ser formados por:

- ⇨ quarteirões;
- ⇨ ruas (face dos quarteirões);
- ⇨ departamentos;
- ⇨ prateleiras;
- ⇨ caixas;
- ⇨ lotes de produtos;
- ⇨ etc...

vi) **Sistema de referência**: é uma ***listagem completa*** de todos os unidades da população amostral (aptas a serem selecionadas na amostra);

vii) **Amostragem probabilística**: é a pesquisa por amostragem realizada segundo critérios bem definidos da teoria estatística das probabilidades.
Na amostragem probabilística ***todas*** as unidades da população amostral devem ter a ***mesma probabilidade*** de serem selecionadas.

Por que fazer amostragem ao invés de um censo?

Vantagens da pesquisa por amostragem em relação ao censo:

a) é mais barata;
b) é mais rápida;
c) é mais fácil de ser controlada por envolver operações menores.

Desvantagens da pesquisa por amostragem em relação ao censo:

a) o censo pode ser mais vantajoso quando a população é pequena e/ou as informações são de fácil obtenção.
b) os resultados da pesquisa por amostragem carregam erro;
c) se a população for muito heterogênea o erro pode ser muito grande (e a precisão muita baixa).
   Neste caso pode ser necessária uma amostra muito grande;

II- planos de Amostragem

Para a definição do *plano amostral* devem-se ter bem definidos:

i) ***Unidade amostral***: indivíduos ou grupos de indivíduos (conglomerados);

ii) ***Sistema de referência:*** lista completa das unidades amostrais.

iii) N = tamanho da população, é definido pelo número de indivíduos da população amostral;

iv) n = tamanho da amostra, definido pelo número de indivíduos selecionados na amostra.

⇨ $n < N$

Fatores que interferem na escolha do Plano Amostral:

⇨ Tamanho da população N;
⇨ Custo;
⇨ Heterogeneidade da população;

Os elementos da amostra devem ser selecionados da população amostral segundo alguma forma de sorteio.

Os Planos de Amostragem mais comuns são:

A) Amostragem Aleatória Simples (*A.A.S.*):

Na *A.A.S.*, a amostra de tamanho n é selecionada ***ao acaso*** dentre os N elementos da população amostral.

Procedimento de sorteio:

i) Um indivíduo é selecionado ao acaso dentre os N possíveis;
ii) O segundo indivíduo é selecionado ao acaso dentre os $(N - 1)$ restantes...
iii) . . . e assim por diante, até que todos os n indivíduos sejam sorteados.

Esse procedimento tem a característica de ser "***sem reposição***", o que significa que: cada indivíduo aparece uma única vez na amostra.

Procedimentos "***com reposição***", quando o indivíduo pode aparecer mais de uma vez na amostra, não serão abordados por serem poucos comuns na prática.

Obs: Quando o tamanho da população for muito grande, os dois procedimentos de sorteio (sem e com reposição) são equivalentes.

Na *A.A.S.* a probabilidade de qualquer indivíduo, ou elemento, da população fazer parte da amostra é igual a $\dfrac{n}{N}$.

Como realizar o sorteio?

i) geração números aleatórios, pelo computador;

ii) tabela de números aleatórios;

iii) globos com bolinhas numeradas;

iv) qualquer outra forma aleatória de escolha que preserve a propriedade de que cada unidade amostral tenha a mesma chance de ser selecionada.

B) Amostragem Aleatória Estratificada (*A.A.E.*):

Quando a população é muito ***heterogênea***, ou seja, quando as características observadas variam muito de um indivíduo para outro, é aconselhável ***subdividir*** a população em ***estratos homogêneos***.

A população é dividida em k estratos sendo que, uma *A.A.S.* é aplicada em cada um dos deles.

Definições:

i) tamanhos dos estratos: $N_1, N_2, N_3, \ldots, N_k$.

$$N_1 + N_2 + N_3 + \ldots + N_k = N$$

ii) tamanhos das amostras nos estrados: $n_1, n_2, n_3, \ldots, n_k$.

$$n_1 + n_2 + n_3 + \ldots + n_k = n$$

Obs: A *A.A.E.* produz ***resultados mais precisos*** do que a *A.A.S.* com o mesmo tamanho de amostra.

É mais cara, por segmentar a população.

Pergunta:

Sabendo que o tamanho da amostra é $\boldsymbol{n}$, como ***alocar***, ou, determinar o número de indivíduos a serem selecionados em cada um dos estratos?

i) **Alocação por igual**: se se desconfia de que os estratos são todos de tamanhos parecidos, ou seja,

$$N_1 \approx N_2 \approx N_3 \approx \ldots \approx N_k$$

Então pode-se fazer: $n_1 = n_2 = n_3 = \ldots = n_k = \dfrac{n}{k}$

Exemplo: Se o tamanho de uma amostra for $n = 56$ e, o número de estratos é $k = 4$, então, $n_1 = n_2 = n_3 = n_4 = 14$.

ii) **Alocação proporcional ao tamanho do estrato**: na alocação proporcional ao tamanho, os tamanhos das amostras devem seguir a mesma relação de proporcionalidade dos tamanhos dos estratos, ou seja,

$$\frac{n_1}{n} = \frac{N_1}{N}, \qquad \frac{n_2}{n} = \frac{N_2}{N}, \qquad \ldots \qquad \frac{n_k}{n} = \frac{N_k}{N}$$

Desta forma, tem-se

$$n_1 = \frac{nN_1}{N}, \qquad n_2 = \frac{nN_2}{N}, \qquad \ldots \qquad n_k = \frac{nN_k}{N}$$

Exemplo:

Considere uma amostra de tamanho $n = 48$ a ser selecionada de uma população dividida em 3 estratos, tais que $N_1 = 40$, $N_2 = 80$ e $N_3 = 120$, então

$$\boldsymbol{N} = 20 + 60 + 180 = 240$$

$$\frac{N_1}{N} = \frac{40}{240} = \frac{1}{6} \quad \Rightarrow \quad n_1 = \frac{48}{6} = 8$$

$$\frac{N_2}{N} = \frac{80}{240} = \frac{1}{3} \quad \Rightarrow \quad n_2 = \frac{48}{3} = 16$$

$$\frac{N_3}{N} = \frac{120}{240} = \frac{1}{2} \quad \Rightarrow \quad n_3 = \frac{48}{2} = 24$$

Portanto, $n_1 = 8$, $n_2 = 16$ e $n_3 = 24$ é a alocação proporcional ao tamanho dos estratos.

Esse resultado significa que se deve selecionar 8 indivíduos do primeiro estrato, 16 do segundo estrato e 24 do terceiro.

iii) **Alocação ótima**: alocação que ***otimiza*** uma ***relação conhecida*** (***função***) e que normalmente envolve o tamanho dos estratos, as suas heterogeneidades e o custo da amostragem.

Por otimizar entende-se escolher os tamanhos de amostras em cada estratos que ***maximizam***, ou ***minimizam***, a função escolhida.

C) **Amostragem Aleatória por Conglomerados** (***A.A.C.***): na amostragem por conglomerados os elementos da população são agrupados em ***conglomerados*** ou ***clusters*** (grupos), que serão as unidades amostrais a serem selecionadas.

A divisão deve ser feita de forma que os conglomerados tenham as ***mesmas características*** da população.

Na ***A.A.C.*** uma ***A.A.S.*** é aplicada para a seleção aleatória de ***k*** conglomerados.

Uma vez selecionados os conglomerados, ***todos os seus elementos*** devem são observados.

O procedimento descrito acima é uma ***A.A.C.*** em ***um estágio***, quando se realiza uma única seleção de conglomerados.

A ***A.A.C.*** pode, ainda, ser aplicada em ***dois ou mais estágios***:

Na A.A.C. em ***dois estágios***, após a escolha dos conglomerados, aplica-se um segundo sorteio aleatório dentre os seus elementos.

Exemplos: Estudo sobre a percepção social dos problemas de quantidade, qualidade e custo dos recursos hídricos em São Carlos. Definindo-se os quarteirões como sendo os conglomerados:

***a)* A.A.C. em 1 estágio:**

Uma A.A.S. é aplicada para a seleção de uma amostra aleatória de quarteirões, e o questionário é aplicado a todos os domicílios dos quarteirões selecionados.

***b)* A.A.C. em 2 estágios:**

- *i)* no 1º. estágio: aplica-se uma $A.A.S.$ para se selecionar uma amostra de quarteirões;
- *ii)* no 2º. estágio: dentre os quarteirões selecionados no 1º. estágio, sorteia-se uma amostra aleatória de domicílios que efetivamente participarão da amostra.

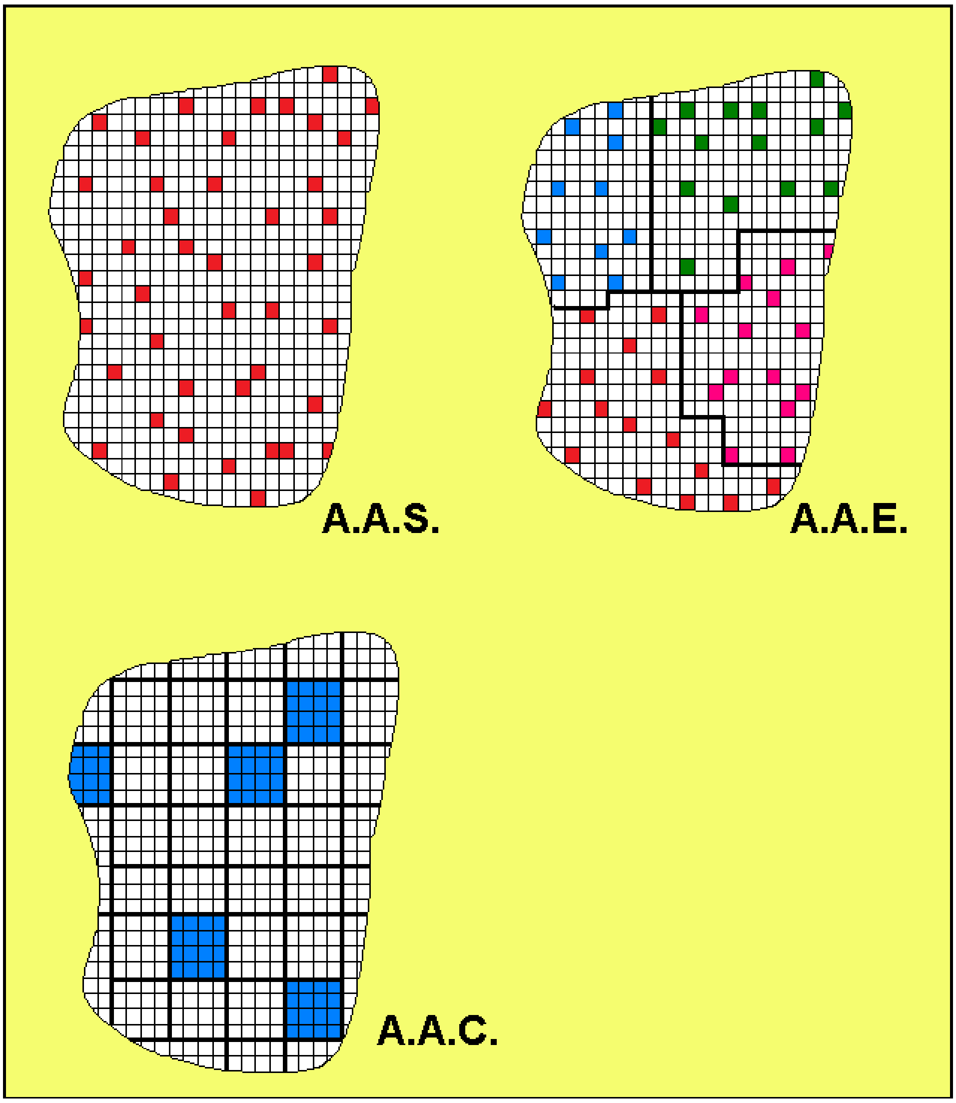
⇨ A $A.A.C.$ produz ***resultados menos precisos*** do que a $A.A.S.$ com o mesmo tamanho de amostra e, por consequência, do que a $A.A.E.$

⇨ É ***mais barata*** por agrupar os elementos da população.

⇨ Na $A.A.C.$ o tamanho da amostra $\boldsymbol{n}$ será determinado a posteriori, pelo número total de elementos observados nos conglomerados (no estágio final de amostragem).

Quadro comparativo entre os três métodos de amostragem:

A.A.E.	⇨	Mais precisa do que a $A.A.S.$, porém mais cara. • considera a heterogeneidade da população
A.A.S.	⇨	Planejamento ideal. • pode ser muito cara • não considera a heterogeneidade da população
A.A.C.	⇨	Menos precisa do que a $A.A.S.$ e $A.A.E.$, porém mais barata. • resolve o problema do custo



D) **Amostragem Sistemática:** é aplicada de forma sistemática, tendo em mão um sistema de referência de fácil acesso.

Na amostragem sistemática além da facilidade de acesso ao sistema de referência, a informação a ser coletada também é de fácil acesso.

⇨ Fichas de cadastro de assinantes (revistas, provedores de acesso à internet, serviço telefônico, etc...); cadastro de funcionários; peças numa linha de produção; mudas num canteiro; etc...

Procedimento: com o sistema de referência em mãos

a) determina-se o intervalo de seleção, que é dado por $R = \frac{N}{n}$;

b) sorteia-se um indivíduo, ou item, dentre os $\boldsymbol{R}$ primeiros da relação;

c) a partir daí, seleciona-se os indivíduos sistematicamente a cada intervalo de tamanho $\boldsymbol{R}$.

Exemplo: se a população tem tamanho $\boldsymbol{N = 84}$ e deve-se selecionar uma amostra de tamanho $\boldsymbol{n = 6}$, então, tendo-se em mão uma relação com os 84 indivíduos da população:

i) divide-se população em 6 seções de tamanho $\frac{84}{6} = 14$;

ii) seleciona-se aleatoriamente o primeiro indivíduo da amostra dentre os *14 primeiros* (por exemplo, o de número 5);

iii) o segundo indivíduo a ser selecionado é o 5 + 14 = 19, ou seja, o 19º. da relação;

iv) o terceiro é o 19 + 14 = 33, ou seja, o 33º. da relação, e assim por diante.

ordem	Indivíduo selecionado
1	5º
2	19º
3	33º
4	47º
5	61º
6	75º

Outro exemplo: $N = 79$ e $n = 7$ => $\frac{79}{7} = 11.3 \approx 11$

* O primeiro selecionado é o 3º., e, dai por diante a seleção é feita em intervalos de tamanho 11 (ver tabela).

ordem	Indivíduo selecionado
1	3º
2	14º
3	25º
4	36º
5	47º
6	58º
7	69º

Situações especiais:

Se, por acaso: $N = 68$ e $n = 7$ => $\frac{68}{7} = 9.7 \approx 10$

* O primeiro selecionado é o 9º e, a partir daí:

ordem	Indivíduo selecionado
1	9º
2	19º
3	29º
4	39º
5	49º
6	59º
7	***69º !***

Note que nesse caso, o 69º indivíduo da relação não existe, pois $N = 68$, logo, ***a amostra fica com uma unidade a menos***.

Ou ainda: $N = 80$ e $n = 7$ => $\dfrac{80}{7} = 11.4 \approx 11$

* O primeiro selecionado é o 2º.

ordem	Indivíduo selecionado
1	2º
2	13º
3	24º
4	35º
5	46º
6	57º
7	68º
8	**79º**

Já, nesse caso, o 79º indivíduo é o penúltimo da relação e deve ser incluído, logo, ***a amostra fica com uma unidade a mais***.

* A amostra pode ter uma unidade a mais ou a menos em função do arredondamento.

Amostragens não aleatórias

Muitas vezes não se tem acesso a um sistema referência para a realização do sorteio.

A ***A.A.C.*** pode resolver a maioria desses casos.

Uma outra saída é a utilização de métodos de amostragem não aleatórios.

i) **Amostragem por cotas**: a população é dividida em grupos, assemelhando-se à A.A.E., mas a seleção não é aleatória.

ii) **Amostragem por julgamento e estudos comparativos**: seleciona-se as unidades da amostra segundo um determinado perfil definido segundo os objetivos da pesquisa.

No estudo comparativo certas características são comparadas em duas, ou mais, populações através de amostras escolhidas por julgamento

Exemplos:

1) Estudo sobre a produção científica dos departamentos de ensino de uma universidade.

2) Estudo sobre a percepção do conceito de morte em crianças de diferentes períodos de desenvolvimento cognitivo (subperíodo pré-operacional, subperíodo das operações concretas, período formal).
Estudo comparativo da incidência de câncer de pulmão em grupos de Fumante e Não Fumantes.

Obs: Nos estudos comparativos, normalmente não se busca a generalidade, mas sim as diferenças entre os grupos em análise.

Nesse contexto, as amostras devem ser o mais similares possíveis, diferindo apenas em relação ao fator de comparação.

O Erro Amostral

O **erro amostral** é definido como sendo a diferença entre a estimativa obtida para um parâmetro e o seu verdadeiro valor. É decorrente da variabilidade natural das unidades amostrais (**é aleatório**).

$$\boxed{\textit{erro amostral} = \textit{estimativa} - \theta}$$

Como medir o erro?

A amostra é retirada sem erro?

O erro decorrente da coleta dos dados é chamado de erro não amostral.

Os planejamentos e a execução da pesquisa devem ser feitos com muita cautela para evitar os erros não amostrais.

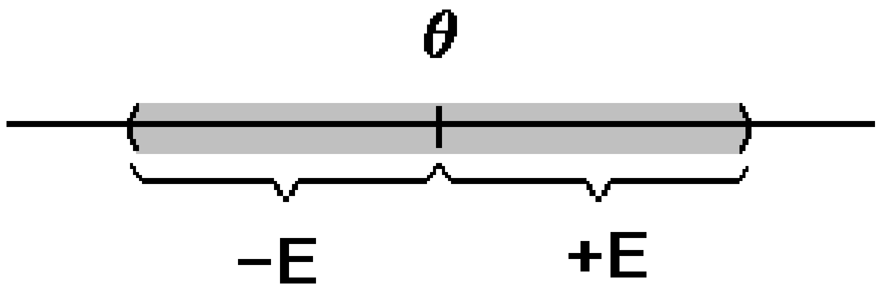
Alguns erros em amostragem:

i) População acessível diferente da população alvo;
ii) Falta de resposta;
iii) Erros de mensuração.

Determinação do tamanho da amostra

A determinação do tamanho da amostra é, talvez, o grande dilema dos pesquisadores, pois deve levar em conta um erro tolerável e a probabilidade de se cometer tal erro.

O **erro tolerável** é uma margem de erro das estimativas em relação ao parâmetro θ, para mais ou para menos, o qual o pesquisador está disposto a aceitar.



O tamanho da amostra é determinado tal que a probabilidade de que a estimativa do parâmetro θ esteja dentro da margem de erro seja alta, por exemplo, de 95%

$$P(\text{ estimativa de } \theta \text{ estar dentro da margem de erro }) = 0.95$$

Em linguagem estatística:

$$P(|\textit{estimativa} - \theta| < E) = 0.95$$

Obs: Para o cálculo acima, deve-se considerar uma distribuição de probabilidades para a estimativa, normalmente a ***Gaussiana***, ou ***normal***.

Na prática, pode-se escolher um tamanho inicial n_0 em função de um *erro relativo* tal que:

$$n_0 = \frac{1}{(Erro\ Relativo)^2}$$

Conhecendo o tamanho da população, deve-se fazer a correção:

$$n = \frac{N\,n_0}{N + n_0}$$

<u>Exemplo</u>: Se N = 780, e com um erro relativo de no máximo 5%, então

$$n_0 = \frac{1}{(0.05)^2} = 400$$

Fazendo a correção pelo tamanho da população, tem-se

$$n = \frac{400 \cdot 780}{(400 + 780)} = \frac{312000}{1180} = 264{,}4 \approx 265,$$

ou seja, a amostra ser de 265 unidades amostrais.

Obs: → $n = \dfrac{N\,n_0}{N + n_0} = \dfrac{n_o}{\dfrac{n_0}{N} + 1}$

→ se a <u>população é muito grande</u>, ou seja, N é muito grande, então, $\dfrac{n_0}{N} \approx 0$, logo $n = n_0$

Conceito:

Estatística = é uma característica obtida como função dos dados para descrever a amostra.
(soma dos valores da amostra, média amostral, proporção de uma dada resposta, etc...)

Desta forma, uma **estimativa** é um valor obtido de uma estatística.

Obs: Toda estimativa é uma estatística, mas nem toda estatística é uma estimativa.

<u>Exemplo</u>: Estudar a influência do fato do chefe da família ser analfabeto e/ou mulher no perfil sócio-econômico das famílias de UFSCarlândia.

<u>Características a serem observadas</u>:

iii) número de moradores no domicílio (morad);
iv) número de filhos estudando (estdte);
v) chefe da família é analfabeto sim/não (anfbto);
vi) chefe da família é mulher sim/não (chmul);
vii) renda familiar, em salários mínimos (1s.m.= R$ 380,00).

<u>Dados da População</u>:

viii) 250 domicílios;
ix) 1046 moradores (UFSCarlenses).

Números gerados aleatoriamente entre 1 e 250 para a seleção dos domicílios.

83	121	166	161	145	1	178	164	219	26
188	131	105	172	61	57	141	70	133	100
54	211	203	173	226	64	16	87	182	7
184	73	213	58	146	99	225	174	112	96
79	97	195	71	192	51	158	228	208	42
190	80	46	27	196	135	56	207	143	118

Números gerados aleatoriamente entre 1 e 34 para a seleção dos quarteirões.

31	7	28	29	22	3	13	4
6	20	27	19	8	30	15	12