TEORIA DA AMOSTRAGEM

Ben Dêivide

13 de abril de 2015

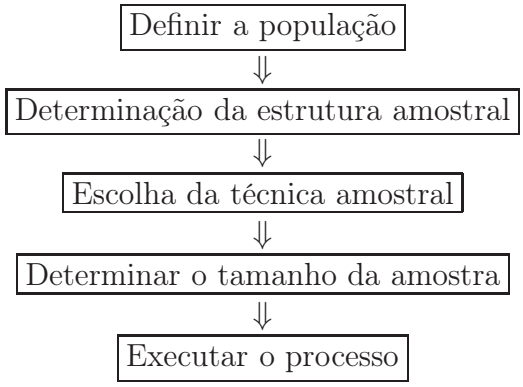
1 Definições iniciais

- **População:** É a soma de todos os elementos que compartilham algum conjunto comum de característica, conforme o universo propósito do problema.
- **Amostra:** É um sub-conjunto da população.
- **Amostragem:** O processo de colher amostras de uma população.

O levantamento por amostragem têm a finalidade de reproduzir a realidade estudada. Esses levantamentos se aplicam ao conjunto real composto de elementos, denominado população de estudo. Os elementos podem ser seres humanos, árvores, domicílios, animais, áreas ou objetos. Os dados são coletados em amostras da população de estudo e as medidas calculadas (estimativas) passam ser as informações disponíveis para os valores populacionais desconhecidos (parâmetros).

Na fase inicial dos levantamentos amostrais é necessário formular o problema e aventar hipóteses sobre o objeto de estudo ou expectativas sobre os possíveis resultados. Ainda nessa fase inicial, o investigador deve definir a população de estudo, parte identificável e acessível da população objeto, os objetivos e as variáveis observadas. Numa segunda etapa é realizado o planejamento, elaborando o plano de amostragem ou determinando o caminho a ser percorrido para atingir os objetivos propostos.

O processo de planejamento de uma amostragem pode ser mencionado cinco estágios. Eles estão interrelacionados, desde a problematização até a apresentação dos resultados.



i) **Definir a população:** consiste em estabelecer uma população alvo. Este é o primeiro passo, e tem que ser feito com precisão. A escolha errada de uma população alvo resulta sempre em uma pesquisa equivocada, que uma catástrofe.

ii) **Determinação da estrutura amostral:** consiste em representar os elementos que compõe a população alvo. Um exemplo é a lista telefônica.

iii) **Escolha da técnica amostral:** o pesquisador deverá decidir qual a técnica amostral que será utilizada: amostragem probabilística ou não-probabilística. Será discutido mais a frente.

iv) **Determinação do tamanho amostral:** o pesquisador deverá calcular qual o tamanho da amostra, em que este tamanho represente de forma fidedigna a característica da população em estudo.

v) **executar o processo:** os dados são coletados, conferidos e processados. Análises estatísticas, então realizadas nessa fase e os resultados interpretados retornando-se ao plano das hipóteses e das expectativas, a fim de que os objetivos sejam efetivamente cumpridos e que sejam obtidas as respostas para as questões estudo.

2 Técnicas amostrais

As técnicas amostrais ou técnicas de amostragem pode ser classificadas como amostragem probabilística e amostragem não-probabilística.

A amostragem probabilística se caracteriza por garantir, a priori, que todo elemento pertencente ao universo de estudo possua probabilidade conhecida e diferente de zero, de pertencer a amostra sorteada. A identificação direta ou indireta dos elementos e o uso do sorteio fundamentam as probabilidades matemáticas desse tipo de processo. Se por alguma razão, alguns elementos da população não puderem pertencer à amostra sorteada, a amostra é dita não probabilística. O organograma na Figura 1 define bem as técnicas amostrais.

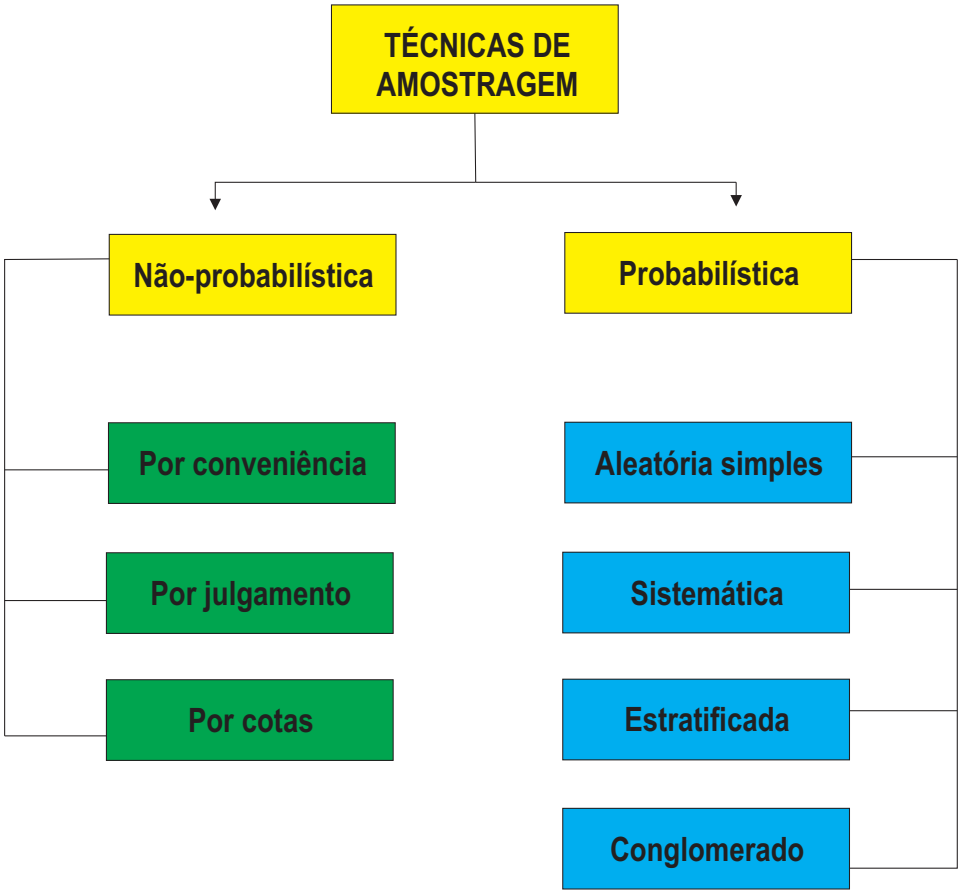


Figura 1: Organograma das técnicas amostrais.

2.1 Tipos de amostragem não-probabilística

Alguns tipos de amostragem não-probabilística podem ser empregado quando a população de estudo não é totalmente acessível, quando a amostra é realizada a esmo, ou seja, sem sorteio.

2.1.1 Amostragem não-probabilística por conveniência

Obtemos uma amostra de elementos por conveniência, que estão a seu dispor. Os elementos estão no lugar exato e no momento certo. Exemplos: questionário em shopping, questionários aplicados à lista de clientes de determinada loja, etc..

2.1.2 Amostragem não-probabilística por julgamento

Não deixa de ser uma variável da amostra não-probabilística por julgamento por conveniência, só que neste caso a escolha dos selecionados é feita com base no julgamento do pesquisador. Exemplos: a rede McDonald lançará um novo tipo de Mc lanches. Foi selecionado o estabelecimento da Av. Bartolomeu de Gusmão, em Santos/SP, com base em critérios de julgamento do pesquisador contratado.

2.1.3 Amostragem não-probabilística por cotas

O pesquisador procura uma amostra que se identifique em alguns aspectos com o universo. Esta identificação pode estar ligada ao sexo, idade, etc.. Exemplo: Um determinado pesquisador fará uma pesquisa de opinião sobre o carro “Honda Accord", para pessoas da classe A, de faixas etárias variáveis de 35 a 60 anos.

2.2 Tipos de amostragem probabilísticas

2.2.1 Amostragem probabilística casual simples

A amostragem casual simples é o processo de amostragem no qual, qualquer combinação dos n elementos da amostra, retirada dos N elementos populacionais que compõe a população, tem igual probabilidade de vir a ser sorteada. O número possível de amostras de tamanho n que podem ser retiradas de uma população de tamanho N é dado por:

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

Nesse tipo de amostragem cada uma dessas combinações tem a chance de $1/C_n^N$ de ser retirada. Na prática, cada elemento é amostrado por uma processo aleatório que confere igual chance de ser sorteado a cada elemento da população. Sorteia-se um elemento e repete o processo para se selecionar o próximo elemento, dando sempre chances iguais para todos aqueles remanescentes na população.

A partir do momento em que os elementos amostrados da população são removidos para as sucessivas retiradas subsequentes, esse método é denominado de amostragem estratificada "sem reposição". Se por outro lado, os elementos populacionais são repostos após uma retirada, a amostragem é dita "com reposição"e é perfeitamente possível de ser feita. À primeira vista, é intuitivo, podemos deduzir que não é muito vantajoso que um mesmo elemento apareça duas, três ou mais vezes na amostra.

Se a população é muito grande os dois tipos de amostragem, com ou sem reposição, não apresentam grandes diferenças. Isso porque a chance de um elemento ser sorteado duas ou mais vezes para a amostra é muito pequena.

O processo de sorteio de uma amostra aleatória simples, pode ser feito por meio de tabelas de números aleatórios, sorteio por funções de geradores de números aleatórios em programas de computadores, por uso de bolas enumerados em urnas ou papéis enumerados em algum tipo de recipiente. As tabelas de números aleatórios podem ser considerados objetos nos tempos atuais, por causa da difusão dos computadores. O uso de papéis ou de bolas enumeradas em urnas ou sacos não é operacionalmente satisfatório, principalmente se a população for muito grande. A população a esse tipo de amostragem é em geral, finita, cujos elementos possam ser identificados em uma listagem enumerada.

2.2.2 Amostragem probabilística estratificada

Uma outra característica exigida para que haja sucesso na amostragem, ou seja, para que estimativas fidedignas parâmetros populacionais possam ser obtidas, refere-se a uma homogeneidade entre os elementos dessa população. Essa homogeneidade interna da população é um tanto quanto difícil de ser caracterizada nas situações práticas com que o investigador se depara. Em hipótese alguma ela se refere a uma ausência de variabilidade da população.

Então, qual seria o limite dessa variabilidade uniforme? Essa questão é muito difícil de ser respondida na prática. Quando o investigador, maior conhecedor da população de estudo, perceber que os elementos da população podem ser agrupados usando-se os níveis (atributos) de uma variável auxiliar, cuja influência de variação da variável, objeto de estudo, está bem caracterizada ou evidente, deve-se preocupar com a adoção da amostragem, simples ao acaso (ASA).

Nessa situação, uma nova metodologia de amostragem deve ser usada. Para ilustrar esse caso, podemos pensar em um exemplo em que se pretendia estudar a opinião de alunos de uma determinada universidade para a necessidade de mudança de grade curricular com o acréscimo de uma disciplina de física avançada. Obviamente, a população de estudo deve

ser estratificada em seus cursos. Essa estratificação, é necessária, uma vez que os alunos de cursos podem ter opiniões distintas sobre a importância da nova disciplina. A variável "curso"poderia e deveria ser usada para se estratificar (ou subdividir) a população alvo.

O sistema de obtenção de amostras em que a população de N elementos é previamente dividida em grupos mutuamente exclusivos, denominados estratos, e dentro dos quais são sorteados amostras simples de tamanho n_h, chama-se de amostragem estratificada aleatória, ou simplesmente amostragem estratificada.

As subpopulações ou estratos são subdivididos previamente em grupos de tamanhos $N_1, N_2, \ldots, N_L$ unidades, mutuamente exclusivas, de tal sorte que $N = \sum_{h=1}^{L} N_h$. Após os estratos terem sido identificados, uma amostra causal simples é retirada de cada estrato, cujos tamanhos são $n_1, n_2, \ldots, n_L$, considerando $n = \sum_{h=1}^{L} n_h$.

Uma das principais razões para se usar a estratificação fundamenta-se na premissa de que esse processo leva a um ganho de precisão na estimação de parâmetros da população. Isso realmente ocorre, pois é possível dividir uma população heterogênea em subpopulações inteiramente homogêneas. Dessa forma, um estrato é considerado homogêneo inteiramente se, de elemento para elemento, há apenas uma pequena variação. Por essa razão, uma estimativa precisa para um parâmetro de estrato populacional pode ser obtida com apenas uma pequena amostra desse estrato.

Essas estimativas dos diferentes estratos são combinados para a obtenção de uma estimativa de um determinado parâmetro da população como um todo.

As notações que estão sendo empregadas até o momento usam o índice h para identificar um estrato e o índice i para definir um elemento dentro de um estrato. Assim, N_h e n_h são os tamanhos do estrato h populacional e amostral, respectivamente; X_{hi} é o valor da observação i no estrato h. Será considerado, também que $f_h = n_h/N_h$ representa a fração amostral. Os dados paramétricos são definidos para o estrato h e para toda a população. Assim, a equação (1) refere-se à média populacional, e a equação (2) à variância populacional do estrato h.

$$\mu_h = \frac{\sum_{i=1}^{N_h} X_{hi}}{N_h} \tag{1}$$

$$\sigma_h^2 = \frac{1}{N_h}\left[\sum_{i=1}^{N_h} X_{hi}^2 - \frac{\left(\sum_{i=1}^{N_h} X_{hi}\right)^2}{N_h}\right] \tag{2}$$

Os estimadores da média e da variância do estrato h são apresentados nas equações (3) e (4).

$$\overline{X}_h = \frac{\sum_{i=1}^{n_h} X_{hi}}{n_h} \tag{3}$$

$$S_h^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n_h} X_{hi}^2 - \frac{\left(\sum_{i=1}^{n_h} X_{hi}\right)^2}{n_h}\right] \tag{4}$$

Finalmente, é possível apresentar o estimador da média populacional global. Esse estimador pode não ser único. O primeiro, o mais geral, estimador da média populacional é a média ponderada das médias dos L estudos populacionais. Esse estimador está apresentado na equação (5).

$$\overline{X}_{est} = \frac{\sum_{h=1}^{L} N_h \overline{X}_h}{N} \tag{5}$$

O segundo estimador é praticamente igual ao primeiro, diferenciando apenas nos pesos utilizados, que agora são os tamanhos dos estratos amostrais. O estimador da média populacional é dado pela equação (6).

$$\overline{X} = \frac{\sum_{h=1}^{L} n_h \overline{X}_h}{n} \tag{6}$$

Os estimadores (5) e (6) são equivalentes quando a fração amostral de cada estrato é equivalente a fração populacional de cada estrato, ou seja, quando

$$\frac{n_h}{n} = \frac{N_h}{N} \quad \text{ou} \quad \frac{n_h}{N_h} = \frac{n}{N}.$$

2.2.3 Amostragem probabilística sistemática

A amostragem sistemática é um tipo de amostragem que o plano de amostragem é obtido por um critério pelo qual intervalos regulares do mesmo tamanho entre unidades da amostras são tomados até se compor uma amostra de tamanho n e toda a extensão da localização física da população-alvo. Para implementar esse sorteio os N elementos populacionais são tomados a cada $k = N/n$ elementos. O primeiro elemento deve ser sorteado entre os k primeiros. Se, por exemplo, uma população de $N = 10.000$ elementos é considerada e deseja-se extrair uma amostra de tamanho $n = 500$, então k será $10.000/500 = 20$. Assim, se o elemento 11 for primeiro a ser sorteado entre os 20 primeiros, a amostra fica determinada da seguinte forma: 11, 31, 51 e assim por diante.

Este tipo de amostragem é fácil de ser executada e provavelmente é mais precisa que a amostra casual simples. A razão disso, é a subdivisão da população em k estratos e a obtenção de um elemento por estrato.

A diferença dessa amostragem para a amostragem é fácil de ser executada e provavelmente é mais precisa que a amostra casual simples. A razão disso, é a subdivisão da

população em k estratos e a obtenção de um elemento por estrato. A diferença dessa amostragem para a amostragem estratificada original é que o elemento sorteado está na mesma posição relativa dos estratos. Por outro lado, devido a esse tipo de amostragem cobrir de forma mais regular a população em toda a sua extensão que a amostragem estratificada aleatória, essa é considerada mais precisa.

2.2.4 Amostragem probabilística por conglomerados

Quando os elementos da população são reunidos em grupos que são sorteados para compor a amostra, o processo é denominado de amostragem por conglomerado. A razão de se usar um tipo de amostragem como esse é principalmente motivada por critérios de ordem prática. Dentre esses critérios destaca-se a ausência de uma listagem de todos os elementos populacionais.

Em geral o sorteio é feito em estágios sucessivos. Assim, por exemplo, se for considerado o sorteio de uma amostra de 500 propriedades rurais em um dos Estados da Federação, poder-se-ia considerar o sorteio de 50 municípios e 10 propriedades de cada, ou de 25 municípios e 20 propriedades em cada, e assim por diante. A economia nesse tipo de amostragem é evidente, pois o método dispensa a listagem de referência ou de cadastro de toda a população. Somente os conglomerados sorteados são identificados e listados. Se, por exemplo, o Estado de Federação a ser amostrado constituído de 700 municípios, cada um com média igual a 1000 propriedades rurais, a população toda teve um total de 700.000 elementos e somente cerca de 50.000 ou 25.000 deverão ser listados. Um outro aspecto é o custo para a locomoção e acesso aos elementos populacionais para a população na realização do estado proposto. O custo total da pesquisa é reduzido, uma vez que ao sortear um município, várias de suas propriedades são amostradas, o que não ocorre com o procedimento adotado em uma amostragem aleatória simples.

2.3 Dimensionamentos de amostras

Um aspecto que aparece frequentemente no planejamento de experimentos ou de um plano amostral é: "qual deve ser o tamanho da amostra para se ter determinada precisão na estimação da média populacional?" A resposta para essa questão pode ser dada a partir do intervalo de confiança. Da teoria de estimação é possível perceber que o tamanho da amostra melhora a precisão da estimativa e diminui o comprimento do intervalo de confiança. O intervalo de confiança dado por:

$$\overline{X} \pm e, \tag{7}$$

em que $e = t_{\alpha/2}\sqrt{n}$ é a semi-amplitude do intervalo de confiança e é o estimador do erro de estimação.

É evidente que uma regra de dimensionamento de amostras, baseada na equação em (7), supõe distribuição normal ou aproximadamente normal para a população amostrada.

O tamanho da amostra especificando depende:

a) da variabilidade da população amostrada, a qual é estimada por S^2 - grande variabilidade exige maiores tamanhos amostrais;

b) do coeficiente de confiança adotado - quanto maior for o coeficiente de confiança maior deve ser o tamanho da amostra requerido;

c) do erro de estimação pretendido - quanto menor o erro de estimação (intervalos mais estreitos) maior deve ser o tamanho da amostra.

Se o pesquisador fixar um erro de estimação "e" e possuir uma alternativa da variância populacional, então é possível estimar o tamanho amostral adequado, considerando, ainda, um coeficiente de confiança $(1 - \alpha)$, também fixado a priori. Na equação (8) o valor de n é apresentado, fixados o erro de estimação (e) e o coeficiente de confiança.

$$n = \frac{S^2 t^2_{\alpha/2;\nu}}{e^2} \tag{8}$$

É conveniente mencionar algumas dificuldades que são encontradas na solução de (8), das quais temos:

i) necessidade de conhecer os graus de liberdade de t, os quais dependem do tamanho da amostra $(\nu = n - 1)$, que é o que se pretende dimensionar. Assim, a única forma de solucionar o problema é resolver a equação (8) de forma iterativa a partir de um valor arbitrário inicial para n. Com esse valor, busca-se na tabela apropriada do t de Student e resolve para n a equação n. Com o novo valor de n, repete-se o processo até que o valor encontrado seja igual ao valor usado na iteração imediatamente anterior. O processo possui convergência rápida;

ii) necessidade de conhecer uma estimativa de σ^2. Para contorná-la é necessário obter estimativas na literatura especializada em trabalhos semelhantes àquele que se está projetando. As informações assim obtidas devem ser utilizadas com cautela, observando principalmente se as condições em que a pesquisa foi realizada são semelhantes a que se pretende executar. Uma outra forma de se obterem estimativas de σ^2 é realizando amostras pilotos. Essas amostras possuem tamanhos que são frações do valor de n que se pretende estimar.

O valor de n é uma estimativa e como tal depende da acurácia de S^2. Por essa razão, a amostra piloto não deve ser muito pequena, nem pode ser muito grande por causa do custo que isso pode acarretar a pesquisa.