

7 - AMOSTRAGEM

No Capítulo 6 apresentamos as principais etapas do planejamento de uma pesquisa estatística. Na Seção 6.2.2 - Coleta de Dados, vimos que há duas formas de conduzir um estudo: por censo ou por amostragem, sendo que o censo consiste em simplesmente estudar todos os elementos da população, e a amostragem pesquisa apenas uma pequena parte dela, suposta representativa do todo. Avaliamos quando era aconselhável ou não usar amostragem. Para realizar um estudo por amostragem, de maneira que seus resultados sejam válidos e possam generalizados para a população algumas técnicas precisam ser empregadas. Estas técnicas constituem o conteúdo deste Capítulo. A essência deste processo é mostrada na figura 1:

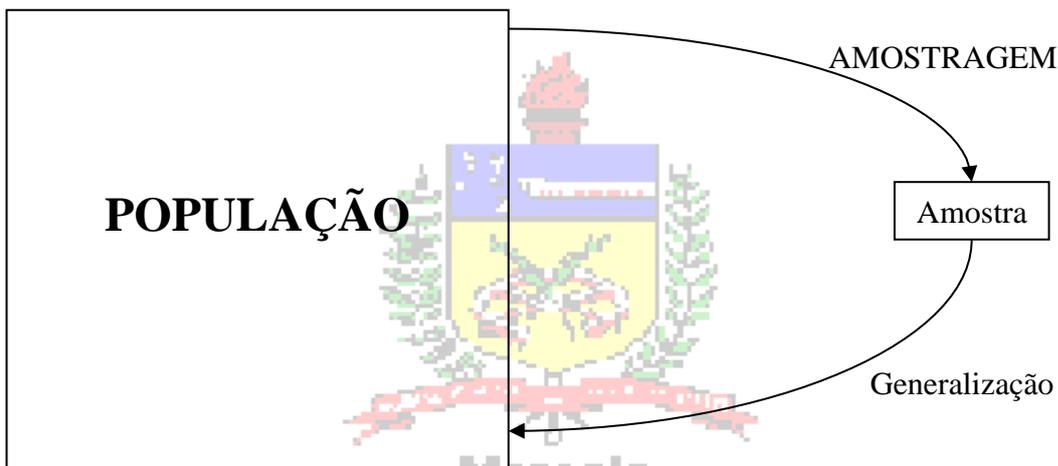


Figura 1 - Processo de amostragem e generalização

7.1 - Plano de Amostragem

Uma vez tendo decidido realizar a pesquisa selecionando uma amostra da população é preciso elaborar o plano de amostragem. O plano de amostragem consiste em definir as unidades amostrais, modo como a amostra será retirada (o tipo de amostragem), e o próprio tamanho da amostra.

As unidades amostrais são as unidades selecionadas para se chegar aos elementos da própria população. Podem ser os próprios elementos da população, quando há acesso direto a eles, ou qualquer outra unidade que possibilite chegar até eles: selecionar os domicílios como unidades de amostragem, para chegar até as famílias (que são os elementos da população); selecionar as turmas como unidades de amostragem, para chegar até os alunos (que são os elementos da população).

A definição de como a amostra será retirada consiste em definir o tipo de amostragem, e podemos ver um resumo dos tipos na Figura 2:



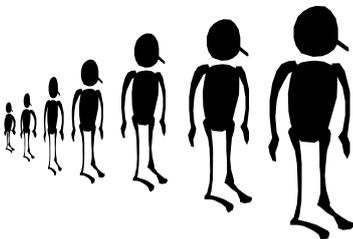
Figura 2 - Tipos de amostragem

Os tipos de amostragem probabilística serão estudados na Seção 7.2, a amostragem não probabilística na Seção 7.3, e o cálculo do tamanho de amostra será visto na Seção 7.4.

7.2 - Amostragem Probabilística

A amostragem probabilística também é chamada de aleatória ou casual. 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 probabilidade¹ 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, e as técnicas de inferência estatística² porventura utilizadas terão validade questionável.

A condição primordial para uso da amostragem probabilística é que "todos os elementos da população tenham uma probabilidade maior do que zero de pertencerem à amostra". Tal condição é materializada se:



Há acesso a toda a população. Ou seja, não há teoricamente problema em selecionar nenhum dos elementos, todos poderiam ser pesquisados.

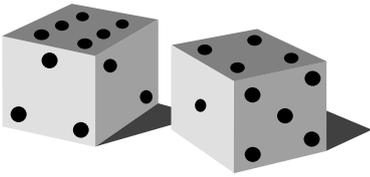


Há possibilidade de obter uma listagem dos elementos da população, concretizando então o acesso a todos os elementos. Se pensarmos em uma pesquisa de opinião, seria uma listagem com todos os possíveis respondentes.

Os elementos da amostra são selecionados através de alguma forma de **sorteio não viciado**: tabelas de números aleatórios, números pseudo-aleatórios gerados por computador. Com a utilização de sorteio elimina-se a ingerência do pesquisador na obtenção da amostra, e garante-se que todos os integrantes da população têm probabilidade de pertencer à amostra.

¹ Os conceitos de Probabilidade serão vistos no Capítulo 8.

² Os conceitos de Inferência Estatística serão vistos no Capítulo 9.



Vamos ver então os diversos tipos de amostragem probabilística.

7.2.1 - Amostragem aleatória (casual) simples

Este tipo de amostragem probabilística somente é recomendável se a população for **homogênea** em relação à variável de interesse. Há uma listagem dos elementos da população, atribuem-se números a eles, e através de alguma espécie de sorteio não viciado (por meio de tabelas de números aleatórios³ ou números pseudo-aleatórios gerados por computador) os integrantes da amostra são selecionados. Neste tipo de amostragem probabilística todos os elementos da população têm a **mesma** probabilidade de pertencer à amostra.

Exemplo 7.1 - Queremos realizar uma pesquisa de opinião sobre a qualidade de um curso universitário, que tem cerca de 1000 alunos, perguntando aspectos relativos ao encadeamento das disciplinas no currículo. Decidimos utilizar amostragem aleatória simples para selecionar os respondentes. Este método de amostragem é o mais apropriado?

A amostragem aleatória simples pode ser utilizada quando houver homogeneidade na população em relação à variável de interesse. No presente caso estamos interessados na opinião dos alunos sobre o currículo. É razoável imaginar que um aluno da oitava fase tenha um conhecimento diferente do currículo do que outro da segunda fase, acarretando diferentes opiniões, que deveriam ser registradas. Se utilizarmos uma amostragem aleatória simples, totalmente por acaso, apenas alunos das fases mais adiantadas, ou das fases iniciais, comprometendo o resultado da pesquisa. Assim, como não há homogeneidade na população acerca da variável de interesse, a amostragem aleatória simples não é apropriada para este caso.

Se a amostra for retirada com o auxílio de um computador o processo da amostragem aleatória simples pode tomar muito tempo, pois para uma grande população, sendo os números selecionados ao acaso, o programa poderia escolher o número 235634 e logo em seguida o número 7. Esta "flutuação" pode fazer com que a retirada da amostra leve algum tempo. O mesmo se aplica se for utilizada uma tabela de números aleatórios.

7.2.2 - Amostragem sistemática

Quando a lista de respondentes for muito grande a utilização de amostragem aleatória simples pode ser um processo moroso. Utiliza-se então uma variação, a amostragem sistemática, que também supõe que a população é homogênea em relação à variável de interesse, mas que consistem em retirar elementos da população a intervalos regulares, até compor o total da amostra. A amostragem sistemática somente pode ser retirada se a ordenação da lista não tiver relação com a variável de interesse: imagine que queremos obter uma amostra de idades de uma listagem justamente ordenada desta forma, neste caso a amostragem sistemática não seria apropriada (a não ser que reordenássemos a lista).

O procedimento para a amostragem sistemática:

- obtém-se o tamanho da população (N);

³ Ver BARBETTA, P. A. Estatística Aplicada às Ciências Sociais. Ed. da UFSC, 9 ed. Florianópolis, 2014.

- calcula-se o tamanho da amostra (n) - espere pela Seção 7.5;
- encontra-se o intervalo de retirada $k = N/n$
 - # se k for fracionário, deve-se aumentar n até tornar o resultado inteiro;
 - # se N for um número primo, excluem-se *por sorteio* alguns elementos da população para tornar k inteiro.
- sorteia-se o ponto de partida (um dos k números do primeiro intervalo), usando uma tabela de números aleatórios, ou qualquer outro dispositivo (isso precisa ser feito para garantir que todos os elementos da população terão chance de pertencer à amostra).
- a cada k elementos da população retira-se um para a amostra, até completar o valor de n.

Veja a Figura 3:

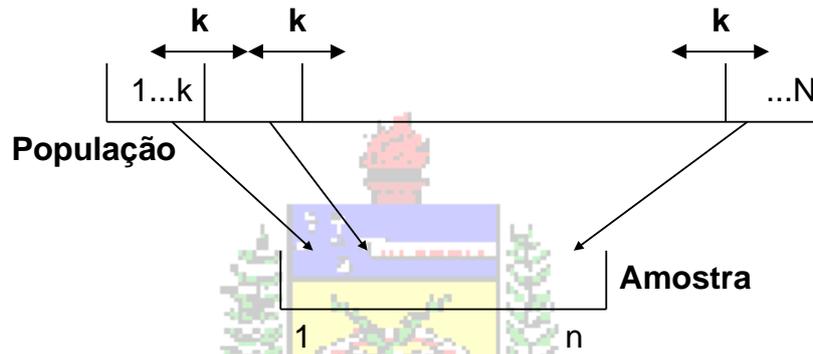


Figura 3 - Amostragem sistemática

Exemplo 7.2 - Uma operadora telefônica pretende saber a opinião de seus assinantes comerciais sobre seus serviços na cidade de Florianópolis. Supondo que há 25037 assinantes comerciais, e a amostra precisa ter no mínimo 800 elementos, mostre como seria organizada uma amostragem sistemática para selecionar os respondentes.

A operadora dispõe de uma lista ordenada alfabeticamente com todos os seus assinantes, o intervalo de retirada será:

$$k = N/n = 25037/800 = 31,2965$$

Como o valor de k é fracionário algo precisa ser feito. Aumentar o tamanho da amostra não resolverá o problema, porque 25037 é um número primo. Como não podemos reduzir o tamanho de amostra, devendo permanecer igual a 800, se excluirmos por sorteio 237 elementos da população, e refizermos a lista teremos:

$$k = N/n = 24800/800 = 31$$

A cada 31 assinantes um é retirado para fazer parte da amostra. Devemos sortear o ponto de partida: um número de 1 a 31 (do 1º ao 31º assinante). Imagine que o sorteio resultasse em 5, então amostra seria (número de assinantes):

{5, 36, 67, 98, ..., 24774}

7.2.3 - Amostragem Estratificada

É bastante comum que a população alvo de uma pesquisa seja heterogênea em relação à variável de interesse. No caso de uma pesquisa eleitoral para governador, por exemplo, podemos esperar que a opinião deva ser diferente dependendo da idade, classe social e mesmo profissão dos entrevistados. Contudo podemos supor que haja certa homogeneidade de opinião dentro de cada grupo. Então se supõe que haja heterogeneidade entre os estratos, mas homogeneidade dentro dos estratos, e que eles sejam mutuamente exclusivos (cada elemento da população pode pertencer a apenas um estrato). Para garantir que a amostra seja *representativa* da população precisamos garantir que os diferentes estratos sejam nela representados.



Figura 4 - Amostragem estratificada

A seleção dos elementos de cada estrato pode ser feita usando amostragem aleatória simples ou sistemática.

A amostragem estratificada pode ser:

- proporcional, quando o número de elementos selecionados de cada estrato é proporcional ao seu tamanho na população (por exemplo, se o estrato representa 15% da população, 15% da amostra deverá ser retirada dele);
- uniforme, quando os mesmos números de elementos são selecionados de cada estrato.

A amostragem estratificada proporcional possibilita resultados melhores, mas exige um grande conhecimento da população (para saber quantos são e quais são os tamanhos dos estratos). A amostragem estratificada uniforme é mais usada em estudos comparativos.

7.2.3.1 – Cálculo de média e proporção amostrais em uma amostra probabilística estratificada proporcional

Quando a amostra é estratificada proporcional é preciso tomar alguns cuidados ao calcular as principais medidas de síntese. Calculam-se normalmente as medidas dentro de cada estrato amostral (parte da amostra retirada de um estrato da população) como visto em Análise Exploratória de Dados (Estatística I). Mas para obter o valor total da amostra é preciso ponderar as médias com os pesos dos estratos na população⁴. Vejam as equações:

$$\bar{x}_{Estratificada} = \sum_{i=1}^E W_i \times \bar{x}_i \quad p_{Estratificada} = \sum_{i=1}^E W_i \times p_i$$

Onde:

- E é o número de estratos identificados na população;
- \bar{x}_i é a média de uma variável quantitativa na parte da amostra retirada de um estrato i qualquer;
- p_i é a proporção do valor de interesse de uma variável qualitativa na parte da amostra retirada de um estrato i qualquer;
- W_i é a proporção que o estrato i têm na população.

Exemplo 7.3 – Uma pesquisa de opinião busca conhecer melhor o comportamento da renda familiar e da opinião de eleitores de uma cidade em relação a um projeto de lei. Para os objetivos da pesquisa é apropriado dividir a população em estratos definidos por região, e as proporções de cada um em relação ao total de habitantes, de acordo com a tabela abaixo:

Região	Número de habitantes (Ni)	Wi (peso do estrato) = Ni/N
Nordeste	35000	35000/350000 = 0,1
Noroeste	70000	70000/350000 = 0,2
Sudoeste	140000	140000/350000 = 0,4
Sudeste	17500	17500/350000 = 0,05
Central	87500	87500/350000 = 0,25
Total (N)	3500000	-

⁴ Isso precisa ser feito para garantir que as medidas amostrais sejam estimadores não viciados dos valores populacionais. Maiores detalhes a respeito serão vistos no Capítulo 9, e podem ser consultados nas referências: COCHRAN, W.G. Técnicas de Amostragem, Rio de Janeiro: Fundo de Cultura, 1965 e BOLFARINE, H., BUSSAB, W.O. Elementos de Amostragem, São Paulo: Edgar Blücher, 2005.

Retirou-se uma amostra probabilística estratificada proporcional de 700 elementos, e foram registradas as rendas familiares e as opiniões dos eleitores sobre o projeto de lei, sendo calculadas as médias amostrais de renda (em salários mínimos) e as proporções amostrais de eleitores favoráveis ao projeto de lei, registradas na tabela a seguir:

Região	Número de elementos na amostra (n_i)	Média amostral de renda por estrato (\bar{x}_i) em salários mínimos	Proporção amostral de favoráveis por estrato (p_i)
Nordeste	70	3,35	0,20
Noroeste	140	2,76	0,25
Sudoeste	280	1,89	0,10
Sudeste	35	4,78	0,42
Central	175	3,05	0,52
Total (N)	700	-	-

Obtenha a média amostral ponderada de renda e a proporção amostral ponderada de favoráveis ao projeto de lei.

Conhecendo os pesos de cada estrato é possível obter os valores ponderados das medidas amostrais em cada estrato e depois os totais, conforme visto na tabela a seguir:

Região	W_i	\bar{x}_i	$W_i \times \bar{x}_i$ (salários mínimos)	p_i (favoráveis)	$W_i \times p_i$
Nordeste	0,1	3,35	$0,1 \times 3,35 = 0,335$	0,20	$0,1 \times 0,20 = 0,02$
Noroeste	0,2	2,76	$0,2 \times 2,76 = 0,952$	0,25	$0,2 \times 0,25 = 0,05$
Sudoeste	0,4	1,89	$0,4 \times 1,89 = 0,756$	0,10	$0,4 \times 0,10 = 0,04$
Sudeste	0,05	4,78	$0,05 \times 4,78 = 0,239$	0,42	$0,05 \times 0,42 = 0,021$
Central	0,25	3,05	$0,25 \times 3,05 = 0,7625$	0,52	$0,25 \times 0,52 = 0,13$
Total	1,0	-	3,0445	1,0	0,261 (26,1%)

Então a média amostral de renda é de 3,0445 salários mínimos, e a proporção amostral de favoráveis ao projeto de lei é de 26,1%.

7.2.4 - Amostragem por conglomerados

Teoricamente a amostragem estratificada proporcional apresenta os melhores resultados possíveis. A grande dificuldade em implementá-la deve-se ao grau de conhecimento necessário sobre a população, que geralmente não existe ou é impraticável de ser obtido. Uma forma alternativa de amostragem consiste no uso de conglomerados.

Os conglomerados também são grupos mutuamente exclusivos de elementos da população, mas são definidos de forma mais arbitrária do que os estratos: é bastante comum definir os conglomerados geograficamente. Por exemplo, os bairros de uma cidade, que constituiriam conglomerados de domicílios.

O procedimento para a amostragem por conglomerados:

- divide-se a população em conglomerados;
- sorteiam-se os conglomerados (usando tabela de números aleatórios ou qualquer outro método não viciado);
- pesquisam-se todos os elementos dos conglomerados sorteados, ou sorteiam-se elementos deles.

Veja a Figura 5:



Figura 5 - Amostragem por conglomerados

Exemplo 7.4⁵ - Pesquisa Nacional por Amostra de Domicílios (PNAD) do IBGE. Coleta informações demográficas e sócio-econômicas sobre a população brasileira. Utiliza amostragem por conglomerados.

Primeiro estágio: amostras de municípios (conglomerados) para cada uma das regiões geográficas do Brasil;

Segundo estágio: setores censitários sorteados em cada município (conglomerado sorteado);

Terceiro estágio: domicílios sorteados em cada setor censitário.

A utilização de amostragem por conglomerados permite uma redução substancial nos custos de obtenção da amostra, sem comprometer em demasia a precisão, e em alguns casos é a única alternativa possível.

7.3 - Amostragem Não Probabilística

A obtenção de uma amostra probabilística exige que se obtenha uma listagem com os elementos da população. Em suma, exige acesso a todos os elementos da população, que a população acessível seja igual à população alvo. Nem sempre é possível obter tal listagem na prática, o que teoricamente inviabilizaria a retirada de uma amostra aleatória. Então deve-se recorrer à amostragem não probabilística.

Ao usar a amostragem não probabilística o pesquisador não sabe qual é a probabilidade de 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. Se as características da população acessível forem semelhantes às da população alvo os resultados podem ser equivalentes aos de uma amostragem probabilística, mas não podemos garantir a sua confiabilidade.

A amostragem não probabilística pode ser usada

Alguns dos usos habituais da amostragem não probabilística são os seguintes:

- a) Como etapa preliminar em projetos de pesquisa;
- b) em projetos de pesquisa qualitativa;
- c) em casos onde a população de trabalho não pode ser enumerada.

E os vários tipos de amostragem não probabilística.

⁵ MOREIRA, Daniel Augusto. Levantamentos Amostrais - Mestrado em Administração de Empresas - FECAP- SP.

Amostragem a esmo⁶. Neste caso o pesquisador procura ser o mais aleatório possível, mas sem fazer um sorteio formal. Imagine um lote de 10000 parafusos, do qual queremos tirar uma amostra de 100, se fôssemos realizar uma amostragem aleatória simples o processo seria muito trabalhoso. Então simplesmente retiramos os elementos a esmo. Este tipo de amostragem também pode ser utilizado quando a população for formada por material contínuo (gases, líquidos, minérios), bastando homogeneizar o material e proceder a retirada da amostra.

Amostragem por julgamento (intencional). O pesquisador deliberadamente escolhe alguns elementos para fazer parte da amostra, com base no seu julgamento de aqueles seriam representativos da população. Este tipo de amostragem é bastante usado em estudos qualitativos. Obviamente o risco de obter uma amostra viciada é grande, pois se baseia totalmente nas preferências do pesquisador, que pode se enganar (involuntária ou "voluntariamente").

Amostragem por cotas. Parece semelhante a uma amostragem estratificada proporcional, da qual se diferencia por não empregar sorteio na seleção dos elementos. A população é dividida em vários subgrupos, na realidade é comum dividir em um grande número para compensar a falta de aleatoriedade, e seleciona-se uma cota de cada subgrupo, proporcional ao seu tamanho. Por exemplo, em uma pesquisa de opinião eleitoral poderíamos dividir a população de eleitores por sexo, nível de instrução, faixas de renda entre outros aspectos, e obter cotas proporcionais ao tamanho dos grupos (que poderia ser obtido através das informações do IBGE). Na amostragem por cotas os elementos da amostra são escolhidos pelos entrevistadores (de acordo com os critérios...), geralmente em pontos de grande movimento, o que sempre acarreta certa subjetividade (e impede que qualquer um que não esteja passando pelo local no exato momento da pesquisa possa ser selecionado). Na prática muitas pesquisas são realizadas utilizando amostragem por cotas.

Exemplo 7.5⁷ - Imagine que queremos saber a opinião dos eleitores do bairro Goiaba sobre o governo municipal. Supõe-se que as principais variáveis que condicionariam as respostas seriam sexo, idade e classe social. O bairro apresenta a seguinte composição demográfica para as variáveis:

Sexo	Idade (faixa etária)	Classe social	% populacional
Masculino	18 -- 35	A	1%
Masculino	18 -- 35	B	4%
Masculino	18 -- 35	C	10%
Feminino	18 -- 35	A	1%
Feminino	18 -- 35	B	2%
Feminino	18 -- 35	C	9%
Masculino	35 -- 60	A	5%
Masculino	35 -- 60	B	8%
Masculino	35 -- 60	C	12%
Feminino	35 -- 60	A	4%
Feminino	35 -- 60	B	8%
Feminino	35 -- 60	C	10%
Masculino	Mais de 60	A	1%
Masculino	Mais de 60	B	9%
Masculino	Mais de 60	C	3%
Feminino	Mais de 60	A	3%
Feminino	Mais de 60	B	7%
Feminino	Mais de 60	C	3%

Se, por exemplo, o tamanho de nossa amostra fosse igual a 200 (200 pessoas serão entrevistadas), o número de pessoas deveria ser dividido de forma proporcional: 1% do sexo masculino, com idade entre 18 e 25 anos, da classe A, totalizando 2 pessoas; 4% do sexo masculino, com idade entre 18 e 25 anos, da classe B, totalizando 8 pessoas, e assim por diante. Os entrevistadores receberiam suas cotas, e deveriam escolher pessoas, em pontos de movimento do referido bairro, que se aproximem

⁶ COSTA NETO, P.L. da O. Estatística. São Paulo. Edgard Blücher, 1977.

⁷ Adaptado de MARCONI, Marina de A., LAKATOS, Eva Maria. Técnicas de Pesquisa - 5ª ed. São Paulo: Atlas, 2003.

dos critérios e entrevistá-las, recolhendo suas opiniões sobre o governo municipal. Usualmente os resultados são generalizados estatisticamente para a população, empregando as técnicas que serão vistas no Capítulo 9, mas rigorosamente os resultados da amostragem por cotas **NÃO TÊM VALIDADE ESTATÍSTICA**, visto que não contemplam o princípio de aleatoriedade na seleção da amostra.

Amostragem "bola de neve"⁸. "Particularmente importante quando é difícil identificar respondentes em potencial. A cada novo respondente que é identificado e entrevistado, pede-se que identifique outros que possam ser qualificados como respondentes".

Há ainda a amostragem por conveniência (selecionam-se elementos supostos semelhantes à população, e que estão disponíveis) e a de passantes eventuais (selecionam-se aqueles que estejam passando pelos pesquisadores no momento; é comum conjugar este tipo com a amostragem por cotas).

7.4 - Tamanho de uma amostra aleatória

A determinação do tamanho de amostra é um dos aspectos mais controversos da técnica de amostragem, e envolve uma série de conceitos (probabilidade, inferência estatística e a própria teoria da amostragem). Neste tópico apresentaremos uma visão simplificada para obter o tamanho mínimo de uma amostra aleatória simples que atenda aos seguintes requisitos:

- estamos interessados no parâmetro (ver Seção 1.2) **proporção** de uma determinada característica na população (os casos em que há interesse na média de uma variável quantitativa serão vistos no Capítulo 9);
- a confiabilidade dos resultados da amostra deve ser igual a aproximadamente 95% (ou seja, há 95% de probabilidade de que a proporção populacional esteja no intervalo definido pelos resultados da amostra - maiores detalhes no Capítulo 9);
- estamos fazendo uma estimativa exagerada do tamanho de amostra (no Capítulo 9 aprenderemos como trabalhar com outros tipos de estimativa);
- não vamos nos preocupar com aspectos financeiros relacionados ao tamanho da amostra (embora obviamente seja uma consideração importante).

O primeiro passo para calcular o tamanho da amostra é definir o erro amostral tolerável, que será chamado de e_0 . Este erro é o valor máximo que o pesquisador admite errar na estimativa do parâmetro. Lembrem-se das pesquisas de opinião eleitoral: "o candidato Fulano está com 18% de intenção de voto, a precisão da pesquisa é de 3% para mais ou para menos". O 3% é o valor do erro amostral tolerável, então o percentual de pessoas declarando o voto no candidato Fulano é igual a $18\% \pm 3\%$. Além disso, há uma **probabilidade** de que este intervalo não contenha o valor real do parâmetro (ou seja o percentual de eleitores que declaram o voto no candidato), pelo fato de que estamos usando uma amostra. Os detalhes a respeito serão apresentados no Capítulo 9.

É razoável imaginar que quanto menor o erro amostral tolerável escolhido maior será o tamanho da amostra necessário para obtê-lo. Isso fica mais claro ao ver a fórmula para obtenção da primeira estimativa do tamanho de amostra:

$$n_0 = \frac{1}{e_0^2}$$

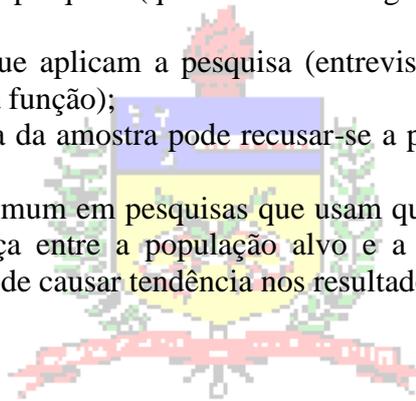
⁸ MOREIRA, Daniel Augusto. Levantamentos Amostrais - Mestrado em Administração de Empresas - FECAP- SP.

Uma vez calculado o tamanho mínimo da amostra, basta retirar os elementos utilizando alguma espécie de sorteio não viciado. Se a população dividir-se em subgrupos, o procedimento deve ser realizado para cada um deles, e o tamanho da amostra total será a soma deles.

7.5 - Fontes de erro em pesquisas por amostragem

O erro amostral tolerável (mencionado na Seção 7.4) considera que a amostra foi retirada seguindo rigorosamente o plano de amostragem, e que não houve viesamento algum por parte do pesquisador. Caso contrário, ou seja, se ocorrem *erros não amostrais*, o erro amostral não pode mais ser garantido. Estes erros poderiam ser:

- problemas no instrumento de pesquisa (questionário ambíguo, opções não conseguem medir as respostas do respondente);
- problemas com as pessoas que aplicam a pesquisa (entrevistadores mal treinados, cansados ou simplesmente inadequados para função);
- falta de resposta (uma parcela da amostra pode recusar-se a participar da pesquisa, ao menos em um primeiro momento).
- erro de cobertura (bastante comum em pesquisas que usam questionários por correio, drop-off ou on-line, em suma é a diferença entre a população alvo e a acessível, somente os interessados respondem à pesquisa, o que pode causar tendência nos resultados).



**Marcelo
Menezes
Reis**