Universidade Federal de São João del-Rei - UFSJ

Campus Alto Paraopeba - CAP

Cláudia Mara Cardoso Reis Coluccini

A Matemática das Pesquisas por Amostragem: um olhar sobre as pesquisas de intenção de votos

Dissertação apresentada ao Departamento de Física e Matemática da Universidade Federal de São João del-Rei como parte dos requisitos exigidos para a obtenção do título de Mestre pelo Programa de Mestrado Profissional em rede Nacional, PROFMAT.

Orientador: Telles Timóteo da Silva

Ouro Branco

2017

Dissertação de Mestrado defendida em 03 de fevereiro de 2017 e aprovada pela Banca Examinadora composta pelos Professores.

Telles Timóteo da Silva

Prof. Telles Timóteo da Silva
Universidade Federal de São João del-Rei

Eduardo Sarquis Soares

Prof. Eduardo Sarquis Soares
Universidade Federal de São João del-Rei

Helvécio G. Fargnoli Filho

Prof:Helvécio Geovani Fargnoli Filho
Universidade Federal de Lavras

Intenções de Voto e Margem de Erro

Cláudia Mara Cardoso Reis Coluccini [1]
Telles Timóteo da Silva[2]

Resumo

Neste trabalho vamos apresentar parte da matemática aplicada na formatação de pesquisas por amostragem, com foco nas pesquisas eleitorais de intenção de votos, de forma a elucidar os procedimentos, os limites de alcance e a maneira como devem ser interpretadas. Vamos abordar os conceitos de margem de erro e nível de confiança no intuito de torná-los claros para professores de matemática e alunos.

Palavras-chave: Probabilidade,amostragem, erro amostral, pesquisas eleitorais.

[1]Aluna de Mestrado Profissional em Matemática, Turma 2014
Instituição: Universidade Federal de São João del-Rei - UFSJ / Campus Alto Paraopeba - CAP
E-mail:cmcardoso@fiemg.com.br
[2]Telles Timóteo da Silva
Departamento de Física e Matemática - Defim, UFSJ/CAP
E-mail: timoteo@ufsj.edu.br

1 Introdução

A cada dois anos acontecem eleições no Brasil. Durante este processo, diversas empresas são contratadas para realizar pesquisas de intenção de votos e a mídia é chamada a divulgar os resultados obtidos.

As pesquisas de intenção de votos costumam ser o ponto alto das campanhas eleitorais. Tais pesquisas retratam a intenção de votos como se a eleição fosse naquele dia em que a pesquisa foi realizada. Além de indicar a provável performance junto ao eleitorado dos candidatos, estes ainda usam-nas de forma estratégica em suas campanhas.

Neste trabalho vamos apresentar parte da matemática aplicada na formatação de pesquisas por amostragem, com foco nas pesquisas eleitorais de intenção de votos, de forma a elucidar os procedimentos, os limites de alcance e a maneira como devem ser interpretadas. Vamos abordar os conceitos de margem de erro e nível de confiança no intuito de torná-los claros para professores de matemática e alunos.

A seguir, na Seção 2 apresentaremos os conceitos básicos de probabilidade. As técnicas de amostragem simples e estratificada são abordadas na Seção 3. A forma como os erros amostrais são tratados e utilizados na estimação de tamanhos amostrais é o tema da Seção 4. A Seção 5 apresenta uma análise do discurso que a mídia usa para divulgar pesquisas de intenção de voto e a Seção 6 indica possíveis aplicações para a sala de aula.

1.1 Símbolos utilizados no texto

Para melhor entendimento de todos os conceitos que serão mostrados nesse trababalho, os seguintes símbolos serão utilizados com seus respectivos significados:
ε : erro amostral;
μ: parâmetro populacional que indica o valor médio;
$\hat{\mu}$: estimativa de μ a partir de uma amostra;
σ : parâmetro populacional que indica o desvio padrão;
Ω : espaço amostral;
ω: elemento de Ω;
f: fração de amostragem;
$\mathcal{F}$: espaço de eventos;
n: número de indivíduos na amostra da população;
N: número de indivíduos de uma determinada população;
$\mathbb{P}$: medida de probabilidade;
$\mathcal{P}(\Omega)$: conjunto das partes de Ω, i.e., todos os subconjuntos de Ω.

2 Conceitos Básicos sobre a Teoria de Probabilidade

Os primeiros estudos matemáticos sobre probabilidades ou seja, o estudo das chances, foram feitos pelos italianos Cardano (1501 - 1576) e Galileu Galiei (1564 -1642) e tratavam de jogos de dados.

Existem na natureza dois tipos de fenômenos: determinísticos e aleatórios. Os fenômenos determinísticos são aqueles em que os resultados são sempre os mesmos, desde que as condições para a realização e observação do fenômeno sejam as mesmas. Os fenômenos aleatórios apresentam resultados que não são previsíveis, mesmo havendo um grande número de repetições.

O conceito de conjunto é fundamental ao estudo de probabilidades. Para o desenvolvimento das próximas sessões será suficiente considerar um conjunto como uma coleção de elementos. O leitor interessado em aprofundar-se no tema pode consultar a referência [9].

2.1 Espaço amostral

Definição 2.1 *Espaço amostral é o conjunto de todos os resultados possíveis para um experimento. Cada elemento do espaço amostral que corresponde a um resultado é denominado ponto amostral ou amostra.*

Observação: Vale ressaltar que consideraremos aqui apenas experimentos cujos espaços amostrais são finitos.

Vejamos o seguinte exemplo de espaço amostral.

Exemplo 2.1 *Seja N um número inteiro positivo e considere uma população com N indivíduos. Fixe um inteiro positivo $n < N$. Podemos definir o seguinte experimento: escolher n indivíduos distintos dessa população. Sabemos que podemos formar $\binom{N}{n}$ subconjuntos distintos contendo n elementos, onde*

$$\binom{N}{n} = \frac{N!}{(N-n)!n!}. \tag{1}$$

O espaço amostral para este experimento é portanto o conjunto Ω formado por todos os subconjuntos com n indivíduos:

$$\Omega = \big\{\{a_1, a_2, ..., a_n\}; \{a_1, a_3, a_4, ..., a_{n+1}\}; ...; \{a_{N-n+1}, a_{N-n+2}, ..., a_N\}\big\}$$

Cada elemento $\omega \in \Omega$ é chamado de amostra.

2.2 Evento

Definição 2.2 *Evento é qualquer subconjunto contido no espaço amostral* Ω.

Quando o número amostras de um espaço amostral finito Ω é k , o número de eventos de Ω é 2^k, isso porque, considerando um número k de amostras de um espaço amostral Ω, temos $1 = \binom{k}{0}$ evento que é o evento impossível, em que nenhuma amostra é selecionada;$k = \binom{k}{1}$ eventos, que são os eventos que contêm apenas uma amostra , até a combinação $\binom{k}{k} = 1$, um evento com k amostras.

O somatório dessas combinações é dado por

$$n(\Omega) = \binom{k}{0} + \binom{k}{1} + \binom{k}{2} + ... + \binom{k}{k} = 2^k$$

Ou seja, para um conjunto com k elementos, o número total de eventos é 2^k.

Exemplo 2.2 *Considere um experimento em que três pessoas tenham que votar entre o candidato A ou B para uma determinada eleição. Os oito possíveis resultados que compõem o espaço amostral são*

$$\Omega = \{(A, A, A), (A, A, B), (A, B, A), (B, A, A), (A, B, B), (B, A, B), (B, B, A), (B, B, B)\}.$$

Considere os eventos E_A e E_B que representam, respectivamente, A ganhar a eleição e B ganhar a eleição.

Logo,

$$E_A = \{(A, A, A), (A, A, B), (A, B, A), (B, A, A)\}$$

$$E_B = \{(A, B, B), (B, A, B), (B, B, A), (B, B, B)\}.$$

Os dados do próximo exemplo serão retomados diversas vezes ao longo do texto, à medida que avançarmos na teoria.

Exemplo 2.3 *Considere uma eleição onde* 15 *pessoas deveriam votar entre os candidatos A e B, desconsiderando-se a hipótese de votos brancos ou nulos, tal que o resultado seja o seguinte:*

Indivíduo	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*
Voto	*A*	*A*	*B*	*A*	*A*	*B*	*B*	*B*	*A*	*B*	*B*	*B*	*A*	*B*	*B*

Planilha de votação dos eleitores para os candidatos A e B.

Desses indivíduos sorteia-se 5 aleatoriamente e sem reposição. O espaço amostral é

$$\Omega = \{\{1,2,3,4,5\};\{1,2,3,4,6\};\ldots;\{11,12,13,14,15\}\}\,.$$

O número total de amostras (número de elementos de Ω *) é* $\binom{15}{5} = 3003$*. O número total de eventos é* 2^{3003} *que é, aproximadamente,* 10^{903} *eventos.*

Podemos elencar alguns eventos específicos.

O evento E_0*, que contém as amostras em que ninguém vota em A, possui* $\binom{9}{5} = 126$ *elementos; visto que há* 9 *indivíduos que não votam em A e deseja-se escolher* 5*pessoas entre essas* 9 *.*

O evento E_1*, que contém as amostras em que* 1 *e apenas* 1 *indivíduo vota em A, possui* $\binom{9}{4}\binom{6}{1} = 756$ *elementos.*

O evento E_2*, que contém as amostras em que* 2 *e apenas* 2 *indivíduos votam em A, possui* $\binom{9}{3}\binom{6}{2} = 1260$ *elementos.*

O evento E_3*, que contém as amostras em que* 3 *e apenas* 3 *indivíduos votam em A, possui* $\binom{9}{2}\binom{6}{3} = 720$ *elementos.*

O evento E_4*, que contém as amostras em que* 4 *e apenas* 4 *indivíduos votam em A, possui* $\binom{9}{1}\binom{6}{4} = 135$ *elementos.*

O evento E_5*, que contém as amostras em que os* 5 *indivíduos votam em A, possui* $\binom{6}{5} = 6$ *elementos.*

2.3 Medida de probabilidade

Definição 2.3 *Seja* Ω *um espaço amostral finito e seja* $\mathcal{F} = \mathcal{P}(\Omega)$ *o espaço de eventos.* ***Uma medida de probabilidade*** *é uma função* $\mathbb{P}$ *sobre* $\mathcal{F}$ *tal que*

$$\mathbb{P} : \mathcal{F} \to [0,1]$$

satisfazendo os seguintes axiomas:
(i) $\mathbb{P}(\Omega) = 1$;
(ii) Se A e B são subconjuntos de Ω *disjuntos, então* $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$.

Exemplo 2.4 *Retomando o Exemplo* 2.3*, se cada amostra* $\omega \in \Omega$ *tiver a mesma chance de ser selecionada, então*

$$\mathbb{P}(\{\omega\}) = \frac{1}{\binom{15}{5}} = \frac{1}{3003}.$$

2.4 Variáveis aleatórias

O espaço amostral associado a uma experiência é definido como o conjunto de todos os possíveis resultados da experiência. Os elementos desse espaço não são necessariamente números, mas é sempre de interesse que se tenha uma representação numérica. Desta maneira, associar cada elemento $\omega \in \Omega$ a um número real corresponde a introduzir uma função definida em Ω e tomando valores do conjunto $\mathbb{R}$ dos números reais.

Uma função X que mapeia todos os pontos de Ω em $\mathbb{R}$ é denominada uma **variável aleatória**. Portanto,

$$X : \Omega \longrightarrow \mathbb{R}$$

$$\omega \longmapsto X(\omega). \tag{2}$$

Exemplo 2.5 *Continuando o exemplo* 2.3*, seja* $X : \Omega \to \mathbb{R}$ *uma função que, para cada* $\omega \in \Omega$ *conta o número de eleitores que votam em* A*.*

Para a amostra $\omega_1 = \{1, 2, 3, 4, 6\}$ *temos*

$$X(\{1, 2, 3, 4, 6\}) = 3,$$

desta forma, nesta amostra a proporção de eleitores que votam em A *é* $\frac{3}{5}$*.*

Para a amostra $\omega_2 = \{1, 7, 10, 14, 15\}$ *temos*

$$X(\{1, 7, 10, 14, 15\}) = 1,$$

desta forma, nesta amostra a proporção de eleitores que votam em A *é* $\frac{1}{5}$*.*

Para a amostra $\omega_3 = \{2, 3, 5, 7, 11\}$ *temos*

$$X(\{2, 3, 5, 7, 11\}) = 2,$$

desta forma, nesta amostra a proporção de eleitores que votam em A *é* $\frac{2}{5}$*.*

Vemos pela tabela do Exemplo 2.3 que a proporção de eleitores que votam em A *é* $\frac{2}{5}$*. Assim, para a obtenção da informação sobre a proporção de eleitores que votam em* A*, notamos que a amostra* ω_3 *é mais representativa do estado da população do que as amostras* ω_1 *ou* ω_2*.*

Exemplo 2.6 *Considere agora os dados dos Exemplos 2.3, 2.4 e 2.5. Em termos de probabilidade, os valores de X se distribuem da seguinte forma:*

$$\mathbb{P}([X=0]) = \frac{\binom{9}{5}}{\binom{15}{5}} = \frac{126}{3003} = 0,04196$$

i.e., a probabilidade de se selecionar uma amostra em que ninguém vote em A é aproximadamente $0,04196$.

$$\mathbb{P}([X=1]) = \frac{\binom{9}{4}\binom{6}{1}}{\binom{15}{5}} = \frac{126.6}{3003} = \frac{135}{3003} = 0,25175,$$

i.e., a probabilidade de se selecionar uma amostra em que 1 *e apenas* 1 *indivíduo vote em A é aproximadamente* $0,25175$.

$$\mathbb{P}([X=2]) = \frac{\binom{9}{3}\binom{6}{2}}{\binom{15}{5}} = \frac{84.15}{3003} = \frac{1260}{3003} = 0,41958,$$

i.e., a probabilidade de se selecionar uma amostra em que 2 *e apenas* 2 *indivíduos votem em A é aproximadamente* $0,41958$.

$$\mathbb{P}([X=3]) = \frac{\binom{9}{2}\binom{6}{3}}{\binom{15}{5}} = \frac{36.20}{3003} = \frac{720}{3003} = 0,23976,$$

i.e., a probabilidade de se selecionar uma amostra em que 3 *e apenas* 3 *indivíduos votem em A é aproximadamente* $0,23976$.

$$\mathbb{P}([X=4]) = \frac{\binom{9}{1}\binom{6}{4}}{\binom{15}{5}} = \frac{9.15}{3003} = \frac{135}{3003} = 0,04495,$$

i.e., a probabilidade de se selecionar uma amostra em que 4 *e apenas* 4 *indivíduos votem em A é aproximadamente* $0,04495$.

$$\mathbb{P}([X=5]) = \frac{\binom{6}{5}}{\binom{15}{5}} = \frac{6}{3003} = 0,00199,$$

i.e., a probabilidade de se selecionar uma amostra em que os 5 *indivíduos votem em A é aproximadamente* $0,00199$.

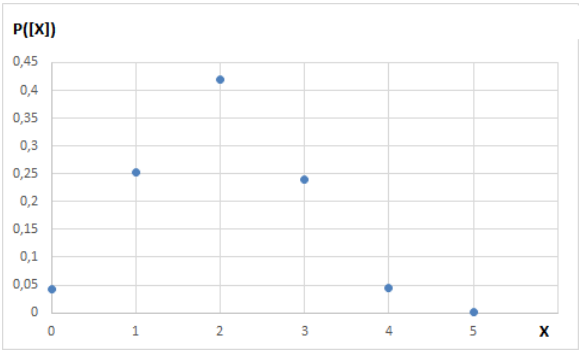

Gráfico referente à distribuição de probabilidades da variável aleatória X.

Neste exemplo vemos que é grande a possibilidade de se selecionar uma amostra que seja "representativa" do estado da população, o que no exemplo anterior está relacionado ao evento $[X = 2]$; entretanto, existe a possibilidade de se selecionar outros tipos de amostra que *não reflitam muito bem* o estado da população, visto que a probabilidade selecioná-las não é nula. Esta é a origem da necessidade de se lidar com margens de erro quando se faz amostragens estatísticas.

2.4.1 Esperança matemática

Considere uma variável aleatória discreta que assume k valores $x_1, x_2, ..., x_k$ e seja $p(x)$ a função de probabilidade de X, isto é, $p(x_i) = \mathbb{P}([X = x_i])$. Então, o valor esperado de X, também chamado de Esperança Matemática de X, é dado por

$$E[X] = \sum_{i=1}^{k} x_i.p(x_i) \tag{3}$$

A esperança matemática pode ser entendida como uma média dos valores x_i ponderados pelos pesos $p(x_i)$.

Exemplo 2.7 *Retomando o Exemplo* 2.6*, para o cálculo de* $E(X)$ *monta-se a tabela* 1.

x_i	$\mathbb{P}([X = x_i])$	$x_i.\mathbb{P}([X = x_i])$
0	$0,04196$	$0,0$
1	$0,25175$	$0,25175$
2	$0,41958$	$0,83916$
3	$0,23976$	$0,71928$
4	$0,04496$	$0,17984$
5	$0,0020$	$0,010$
Total	1	$2,0$

Tabela de valores esperados para a variável aleatória que conta o número de votos no candidato A.

O valor da última linha e da última coluna representa portanto a expressão:

$$E[X] = 0.\mathbb{P}([X = 0])+1.\mathbb{P}([X = 1])+2.\mathbb{P}([X = 2])+3.\mathbb{P}([X = 3])+4.\mathbb{P}([X = 4])+5.\mathbb{P}([X = 5]).$$

O resultado $E[X] = 2$ *representa o valor esperado do número de eleitores numa amostra que votam em A.*

Em outras palavras, numa amostragem aleatória, obteríamos que em média 2 *em cada* 5 *dos eleitores votam em A, o que corresponde à proporção do eleitorado na população de* 15 *indivíduos que vota em A.*

2.4.2 Variância e desvio padrão

Seja X uma variável aleatória. A *variância* de X é uma medida da dispersão dos valores da variável aleatória X em torno de $E[X]$; se os valores tendem a concentrar próximos à média, a variância é pequena, caso os valores tendem a afastar-se da média a variância é grande. Definimos a variância por

$$Var(X) = E\{[X - E(X)]^2\}. \tag{4}$$

Daí obtemos

$$Var(X) = E(X^2) - [E(X)]^2. \tag{5}$$

Se X assume os valores $x_1, x_2, \ldots, x_k$ então a variância de X também pode ser escrita por

$$Var(X) = \sum_{i=1}^{k} (x_i - E[X])^2.p(x_i). \tag{6}$$

O *desvio padrão* de X também é uma medida de dispersão dos valores de X em torno de $E[X]$. No entanto, ao contrário da variância, ele tem a mesma dimensão de $E[X]$. O valor mínimo do desvio padrão é 0, indicando que não há variabilidade, ou seja, que todos os valores são iguais à média. Por definição, o desvio-padrão é dado por

$$\sigma = \sqrt{E\{[X - E(X)]^2\}}. \tag{7}$$

Quando X assume os valores $x_1, x_2, \ldots, x_k$, o desvio padrão de X também pode ser escrito por

$$\sigma = \sqrt{\sum_{i=1}^{k} (x_i - E[X])^2 . p(x_i)}. \tag{8}$$

Temos então que

$$\sigma = \sqrt{Var(X)}.$$

Exemplo 2.8 *Retomando o Exemplo* 2.7*, podemos calcular a variância e o desvio-padrão de X montando a seguinte tabela:*

x_i	$\mathbb{P}([X = x_i])$	$x_i\mathbb{P}([X = x_i])$	$(x_i - E[X])$	$(x_i - E[X])^2$	$(x_i - E[X])^2\mathbb{P}([X = x_i])$
0	$0,04196$	0	$-2,0$	$4,0$	$0,16784$
1	$0,25175$	$0,25175$	$-1,0$	$1,0$	$0,25175$
2	$0,41958$	$0,83916$	$0,0$	$0,0$	$0,0$
3	$0,23976$	$0,71928$	$1,0$	$1,0$	$0,23976$
4	$0,04496$	$0,17984$	$2,0$	$4,0$	$0,17984$
5	$0,00199$	$0,010$	$3,0$	$9,0$	$0,17910$
Total	1	2	3	19	$1,01829$

Tabela de cálculo de variância do número médio de votos no candidato A na amostra.

Para esse caso, a variância é $Var(X) = 1,01829$; que indica a variação da variável aleatória em relação à média, em uma dimensão superior à média e o desvio padrão é $\sigma = \sqrt{1,01829} = 1,00910$ indicando uma variação em torno da média, e que tem a mesma dimensão que a média.

3 Técnicas de amostragem

Quando se deseja tirar determinadas conclusões sobre um grupo de pessoas, ao invés de fazer entrevistas com todo o grupo, o que demanda muito tempo e custo, pode-se estudar apenas uma amostra da população, fazendo-se inferências estatísticas. A amostragem determina estimativas de um ou mais parâmetros da população total. As referências [2, 3, 4] apresentam diversas técnicas de amostragem utilizadas na prática, além daquelas apresentadas nesta seção.

3.1 Parâmetros e estimadores

Suponha que numa população de N indivíduos, a cada indivíduo se atribua um valor numérico $y_1, y_2, \ldots, y_N$. Um parâmetro de interesse é a **média** (μ) desse atributo:

$$\mu = \frac{1}{N} \sum_{i=1}^{N} y_i. \tag{9}$$

Outro parâmetro de interesse é o **desvio padrão** desse atributo na população, denotado por σ, e dado por

$$\sigma = \sqrt{\frac{1}{N-1} \sum_{i=1}^{N} (y_i - \mu)^2}. \tag{10}$$

Cálculo de **proporções** de características da população, que são casos especiais da média, bem como de **totais** também costumam ser de interesse, por exemplo ao se considerar pesquisas eleitorais.

Com a amostragem deseja-se em geral estimar o valor de um ou mais destes parâmetros numéricos representativos da população.

O termo **estimativa** se refere ao cálculo feito a partir de uma amostra para estimar o valor de um parâmetro. Já o termo **estimador** se refere à variável aleatória que, para cada amostra no espaço amostral, retorna o valor estimado para o parâmetro. Nas amostragens aleatórias simples e estratificada apresentadas a seguir, é possível considerar a distribuição de probabilidade dos estimadores e assim tecer ponderações sobre o erro cometido ao se utilizar a estimativa do parâmetro ao invés do próprio parâmetro.

Se a é um parâmetro populacional e X é um estimador para este parâmetro, dizemos que o estimador X é **não tendencioso** se

$$E[X] = a.$$

Em outras palavras, um método de estimativa é não tendencioso se o valor médio do estimador é exatamente igual ao valor verdadeiro da população (ver Cochran [4]).

Para o que se segue, fixe os valores populacionais $y_1, y_2, \ldots, y_N$, bem como os parâmetros: média populacional

$$\mu = \frac{1}{N}\sum_{i=1}^{N} y_i \tag{11}$$

e variância populacional

$$\sigma^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \mu)^2. \tag{12}$$

Notamos que a expressão (12) pode ser desenvolvida da seguinte forma:

$$\begin{aligned}
\sigma^2 &= \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \mu)^2 \\
&= \frac{1}{N-1}\sum_{i=1}^{N}(y_i^2 - 2y_i\mu + \mu^2) \\
&= \frac{1}{N-1}\left\{\sum_{i=1}^{N}[y_i^2] - 2N\mu^2 + N\mu^2\right\} \\
&= \frac{1}{N-1}\left\{\sum_{i=1}^{N}(y_i^2) - N\mu^2\right\}
\end{aligned}$$

então,

$$\frac{1}{N-1}\sum_{i=1}^{N}(y_i^2) = \sigma^2 + \frac{N}{N-1}\mu^2 \tag{13}$$

Lema 3.1 *O somatório dos quadrados dos valores dos elementos da população depende da variância populacional e da média populacional pela seguinte expressão:*

$$\frac{1}{N-1}\sum_{i=1}^{N}(y_i^2) = \sigma^2 + \frac{N}{N-1}\mu^2$$

3.2 Amostragem aleatória simples

A amostragem aleatória simples, também chamada casual, elementar, randômica, etc., é equivalente a um sorteio lotérico. Nela, todos os elementos da população têm igual probabilidade de pertencer à amostra ou, equivalentemente, todas as possíveis amostras têm também igual probabilidade de serem sorteadas.

Em geral, considera-se a amostragem sem reposição, assim o número de amostras diferentes de tamanho n que podem ser retiradas das N unidades é dado pela seguinte fórmula combinatória

$$\binom{N}{n} = \frac{N!}{n!(N-n)!}. \tag{14}$$

Sendo N o número de elementos da população e n o número de elementos da amostra, cada elemento da população tem probabilidade

$$f = \frac{\binom{N-1}{n-1}}{\binom{N}{n}} = \frac{n}{N} \tag{15}$$

de pertencer à amostra. A essa relação $\frac{n}{N}$ denomina-se fração de amostragem.

Observe que o numerador da fração de amostragem $\binom{N-1}{n-1}$ conta o número amostras em que um elemento fixado aparece.

Considere o estimador da média μ, denominado **média amostral** e definido por

$$\bar{X}_\omega = \frac{1}{n}\sum_{k=1}^{n} y_{i_k}, \tag{16}$$

para cada amostra $\omega = \{y_{i_1}, y_{i_2}, \ldots, y_{i_n}\}$.

Teorema 3.1 *A variável aleatória média amostral $\bar{X}$ é um estimador sem tendência de μ.*

Demonstração: Por definição temos que

$$E[\bar{X}] = \sum_{\omega\in\Omega} \bar{X}(\omega).\mathbb{P}(\{\omega\}) \tag{17}$$

Observa-se que

$$\mathbb{P}(\{\omega\}) = \frac{1}{\binom{N}{n}} \tag{18}$$

Logo

$$\begin{aligned} E[\bar{X}] &= \sum_{\omega\in\Omega} .\bar{X}(\omega).\frac{1}{\binom{N}{n}} \\ &= \frac{1}{\binom{N}{n}} \sum_{\{y_{i_1},\ldots,y_{i_n}\}} \left(\frac{1}{n}\sum_{k=1}^{N} y_{i_k}\right) \\ &= \frac{1}{n}.\frac{1}{\binom{N}{n}} \sum_{\{y_{i_1},\ldots,y_{i_n}\}} \left(\sum_{k=1}^{N} y_{i_k}\right) \\ &= \frac{1}{n}.\frac{1}{\binom{N}{n}} \sum_{\{i_1,\ldots,i_n\}} \left(\sum_{k=1}^{N} y_{i_k}\right) \end{aligned} \tag{19}$$

Cada y_{i_k} nessas somas aparece $\binom{N-1}{n-1}$ vezes, logo

$$\begin{aligned} E[\bar{X}] &= \frac{1}{n}.\frac{1}{\binom{N}{n}}\binom{N-1}{n-1}\sum_{i=1}^{N} y_i \\ &= \frac{1}{n.\dfrac{N!}{n!(N-n)!}}.\frac{(N-1)!}{(n-1)!(N-n)!}\sum_{i=1}^{N} y_i \\ &= \frac{1}{N}\sum_{i=1}^{N} y_i = \mu \end{aligned} \tag{20}$$

□

Lema 3.2 *O valor esperado de $\bar{X}^2$ pode ser calculado por*

$$E[\bar{X}^2] = \left[\frac{1}{n}\left(1-\frac{n}{N}\right)\sigma^2 + \mu^2\right]$$

Demonstração:

$$\begin{aligned}
E[\bar{X}^2] &= \frac{1}{\binom{N}{n}} \sum_{\{y_{i_1},\ldots,y_{i_n}\}} \left(\frac{1}{n}\sum_{k=1}^{n} y_{i_k}\right).\left(\frac{1}{n}\sum_{r=1}^{n} y_{i_r}\right) \\
&= \frac{1}{n^2\binom{N}{n}} \sum_{\{y_{i_1},\ldots,y_{i_n}\}} \sum_{k=1}^{n} (y_{i_k} y_{i_r}) \\
&= \frac{1}{n^2\binom{N}{n}} \left\{ \sum_{\{y_{i_1},\ldots,y_{i_n}\}} \sum_{k=1}^{n} (y_{i_k}) + \sum_{\{y_{i_1},\ldots,y_{i_n}\}} \sum_{k\neq r} y_{i_k} y_{i_r} \right\}
\end{aligned}$$

Sabemos que em $\sum_{\{y_{i_1},\ldots,y_{i_n}\}} \sum_{k\neq r} y_{i_k} y_{i_r}$ cada par y_{i_k}, y_{i_r} aparece $\binom{N-2}{n-2}$ vezes então,

$$\begin{aligned}
E[\bar{X}^2] &= \frac{1}{n^2\binom{N}{n}} \left\{ \binom{N-1}{n-1} \sum_{i=1}^{N} (y_i)^2 + \binom{N-2}{n-2} \sum_{i\neq j} y_i y_j \right\} \\
&= \frac{1}{n^2\binom{N}{n}} \left\{ \binom{N-1}{n-1} \sum_{i=1}^{N} (y_i)^2 + \binom{N-2}{n-2} \left[\sum_{i,j=1}^{N} y_i y_j - \sum_{i=1}^{N} {y_i}^2 \right] \right\} \\
&= \frac{1}{n^2\binom{N}{n}} \left\{ \left[\binom{N-1}{n-1} - \binom{N-2}{n-2} \right] \sum_{i=1}^{N} y_i^2 + \binom{N-2}{n-2} \sum_{i,j=1}^{N} y_i y_j \right\} \\
&= \frac{1}{n^2\binom{N}{n}} \left[\binom{N-1}{n-1} - \binom{N-2}{n-2} \right] \sum_{i=1}^{N} y_i^2 + \frac{1}{n^2\binom{N}{n}} \binom{N-2}{n-2} \left(\sum_{i}^{N} y_i\right).\left(\sum_{j=1}^{N} y_j\right) \\
&= \frac{N-n}{nN(N-1)} \sum_{i=1}^{2} y_i^2 + \frac{(n-1)}{nN(N-1)} (N\mu)(N\mu)
\end{aligned}$$

logo,

$$E[\bar{X}^2] = \frac{N-n}{nN(N-1)} \sum_{i=1}^{N} y_i^2 + \frac{(n-1)N}{n(N-1)} \mu^2.$$

Substituindo a equação 13 na equação acima temos que

$$\begin{aligned} E[\bar{X}^2] &= \frac{N-n}{nN}\left\{\sigma^2 + \frac{N}{N-1}\mu^2\right\} + \frac{(n-1)N}{n(N-1)}\mu^2 \\ &= \frac{N-n}{nN}\sigma^2 + \frac{N}{n(N-1)}\left[\frac{N-n}{N} + n - 1\right]\mu^2 \\ &= \frac{N-n}{nN}\sigma^2 + \frac{N}{n(N-1)}\left(\frac{N+nN-N}{N}\right)\mu^2 \\ &= \frac{1}{n}\left(1 - \frac{n}{N}\right)\sigma^2 + \mu^2 \end{aligned} \tag{21}$$

□

Em uma amostragem aleatória simples, o estimador **variância amostral** S^2 é dado por

$$S^2 = \frac{\sum_{k=1}^{n}(y_{i_k} - \bar{X})^2}{n-1}. \tag{22}$$

Teorema 3.2 *O estimador variância amostral S^2 é um estimador não tendencioso da variância populacional*

$$\sigma^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \mu)^2. \tag{23}$$

Demonstração:

$$\begin{aligned}
S^2(\omega) &= \frac{1}{n-1}\sum_{k=1}^{n}[y_{i_k} - \bar{X}(\omega)]^2 \\
&= \frac{1}{n-1}\sum_{k=1}^{n}\{(y_{i_k})^2 - 2y_{i_k}\bar{X}(\omega) + [\bar{X}(\omega)]^2\} \\
&= \frac{1}{n-1}\{\sum_{k=1}^{n}(y_{i_k})^2 - 2\bar{X}.\sum_{k=1}^{n}(y_{i_k}) + n[\bar{X}(\omega)]^2\} \\
&= \frac{1}{n-1}\{\sum_{k=1}^{n}(y_{i_k})^2 - 2[\bar{X}]^2.n + n[\bar{X}(\omega)]^2\} \\
&= \frac{1}{n-1}\{\sum_{k=1}^{n}(y_{i_k})^2 - n[\bar{X}(\omega)]^2\}
\end{aligned}$$

pois $\sum y_{i_k} = n\bar{X}(\omega)$

$$\begin{aligned}
E[S^2] &= \frac{1}{\binom{N}{n}}\sum_{\{y_{i_k},\ldots,y_{i_n}} S^2(\omega) \\
&= \frac{1}{\binom{N}{n}}\sum_{\{y_{i_k},\ldots,y_{i_n}\}}\{\frac{1}{n-1}\sum_{k=1}^{n}(y_{i_k})^2 - n[\bar{X}(\omega)]^2\} \\
&= \frac{1}{\binom{N}{n}}\left(\frac{1}{n-1}\right).\binom{N-1}{n-1}\sum_{i=1}^{n}(y_1)^2 - \frac{n}{(n-1)\binom{N}{n}}\sum_{\{y_{i_k},\ldots,y_{i_n}\}}\bar{X}[(\omega)]^2 \\
&= \frac{n}{(n-1)}\sum_{i=1}^{n}(y_i)^2 - \frac{n}{n-1}\frac{1}{\binom{N}{n}}\sum_{\{y_{i_1},\ldots,y_{i_N}\}}\left[\frac{1}{n}\sum_{k=1}^{n}y_{i_k}\right]^2 \\
&= \frac{n}{N(n-1)}\sum_{i=1}^{N}(y_i)^2 - \frac{n}{n-1}E[\bar{X}^2]
\end{aligned}$$

Substituindo as equações 13 e 21 na equação acima, temos que:

$$
\begin{aligned}
E[S^2] &= \frac{n}{N(n-1)}\left[(N-1)\sigma^2 + N\mu^2\right] - \frac{n}{n-1}\left[\frac{1}{n}\left(1-\frac{n}{N}\right)\sigma^2 + \mu^2\right] \\
&= \left[\frac{n(N-1)}{n(N-1)} - \frac{1}{n-1}\left(1-\frac{n}{N}\right)\right]\sigma^2 \\
&= \left(\frac{nN-n-N+n}{N(n-1)}\right)\sigma^2 \\
&= \sigma^2 \qquad (24)
\end{aligned}
$$

□

Teorema 3.3 *A variância do estimador $\bar{X}$ sob amostragem aleatória simples é dada pela fórmula*

$$
\begin{aligned}
Var(\bar{X}) &= E[(\bar{X}-\mu)^2] \\
Var(\bar{X}) &= \frac{\sigma^2}{n}.\frac{(N-n)}{N} \\
Var(\bar{X}) &= \frac{\sigma^2}{n}(1-f) \qquad (25)
\end{aligned}
$$

Demonstração:

$$
\begin{aligned}
Var(\bar{X}) &= E[(\bar{X}-\mu)^2] \\
&= E\left\{\bar{X}^2 - 2E[\bar{X}]\mu + \mu^2\right\} \\
&= E(\bar{X}^2) - \mu^2 \qquad (26)
\end{aligned}
$$

Utilizando a equação 21, e como $f = \frac{n}{N}$, segue o resultado

$$
\begin{aligned}
Var(\bar{X}) &= E(\bar{X}^2 - \mu^2) \\
&= \frac{1}{n}(1-f)\sigma^2 \qquad (27)
\end{aligned}
$$

□

Temos que as estimativas para os parâmetros populacionais obedecem a seguinte correspondência:

Para $\mu = E[\bar{X}]$:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} y_i. \tag{28}$$

Para S^2:

$$s^2 = \hat{S}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \hat{\mu})^2. \tag{29}$$

Para $Var(\bar{X}) = \frac{\sigma^2}{n}(1-f)$:

$$\widehat{Var}(\bar{X}) = \frac{s^2}{n}(1-f). \tag{30}$$

Exemplo 3.1 *Vamos retomar os dados do Exemplo 2.3, estendendo a tabela para ter mais uma coluna onde associamos o voto no candidato A a* 1*, e o voto no candidato B a* 0*. Vamos verificar como a proporção de votos no candidato A varia segundo a amostra selecionada.*

Indivíduo	*Voto*	*Valor*
1	*A*	1
2	*A*	1
3	*B*	0
4	*A*	1
5	*A*	1
6	*B*	0
7	*B*	0
8	*B*	0
9	*A*	1
10	*B*	0
11	*B*	0
12	*B*	0
13	*A*	1
14	*B*	0
15	*B*	0

Tabela que atribui valor 1 *para A e* 0 *para B.*

Os parâmetros populacionais são:

$$\mu = 0,4$$
$$\sigma^2 = 0,2571$$

sendo o desvio padrão $\sigma = 0,507$.

Para a amostra $\omega_1 = \{1, 2, 3, 4, 6\}$ *temos as estimativas*

$$\hat{\mu} = 0,6$$
$$s^2 = 0,3$$
$$s = 0,547.$$

Para a amostra $\omega_2 = \{1, 7, 10, 14, 15\}$ *temos as estimativas*

$$\hat{\mu} = 0,2$$
$$s^2 = 0,4$$
$$s = 0,632.$$

Para a amostra $\omega_3 = \{2, 3, 5, 7, 11\}$ *temos as estimativas*

$$\hat{\mu} = 0,4$$
$$s^2 = 0,35$$
$$s = 0,5916.$$

3.3 Amostragem estratificada

A amostragem estratificada é uma técnica que consiste em dividir a população, supostamente heterogênea, em subgrupos que chamamos de *estratos*. Esses subgrupos deverão ser mais homogêneos que a população em relação à variável de estudo. As estratificações mais comuns são por classe social, idade, gênero, profissão ou qualquer outro atributo que revele os estratos dentro da população. Mas, a amostragem estratificada é um processo recomendável quando se utiliza da estratificação *natural* de uma cidade em bairros ou distritos, ou de um país em estados.

Os N indivíduos da população são divididos em L subpopulações (estratos) de tamanhos $N_1, N_2, \ldots, N_L$ que não se sobrepõem e, juntas, abrangem a totalidade da população:

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

Em cada estrato h faz-se a amostragem de n_h elementos; desta forma cada estrato tem uma fração amostral dada por

$$f_h = \frac{n_h}{N_h}. \tag{31}$$

Notação utilizada

Seja L o número de estratos. Para cada estrato h, anotaremos:

N_h: número total de unidades no estrato;

n_h: número de unidades da amostra no estrato;

y_{hi}: valor obtido para a unidade de ordem i;

$W_h = \dfrac{N_h}{N}$: peso do estrato;

$f_h = \dfrac{n_h}{N_h}$: fração amostral do estrato;

$\mu_h = \dfrac{\sum_{i=1}^{N_h} y_{hi}}{N_h}$: parâmetro que mede o valor médio de $y_{h1}, y_{h2}, \ldots, y_{hN_h}$;

$\bar{X}_h = \dfrac{\sum_{i=1}^{n_h} y_{hi}}{n_h}$: estimador do valor médio da amostra;

$\sigma_h^2 = \dfrac{\sum_{i=1}^{N_h}(y_{hi} - \mu_h)^2}{N_h - 1}$: parâmetro que mede a variância de μ_h;

$S_h^2 = \dfrac{\sum_{i=1}^{n_h}(y_{hi} - \bar{X}_h)^2}{n_h - 1}$: estimador do parâmetro σ_h.

O estimador do parâmetro μ quando a amostragem é estratificada é representado por $\bar{X}_{st}$ e dado por

$$\bar{X}_{st} = \frac{\sum_{h=1}^{L} N_h \bar{X}_h}{N} \tag{32}$$

onde X_h é o estimador do valor médio no estrato h.

O estimador $\bar{X}_{st}$ em geral é diferente do estimador $\bar{X}$ calculado pela equação

$$\bar{X} = \frac{\sum_{h=1}^{L} n_h \bar{X}_h}{n}. \tag{33}$$

No estimador $\bar{X}_{st}$ os estratos recebem seus pesos corretos:

$$W_h = \frac{N_h}{N}. \tag{34}$$

O valor de $\bar{X}$ coincide com $\bar{X}_{st}$ apenas se em todos os estratos tivermos

$$\frac{n_h}{n} = \frac{N_h}{N} \text{ ou } \frac{n_h}{N_h} = \frac{n}{N} \text{ ou } f_h = f,$$

ou seja, se no estrato h o tamanho da amostra for proporcional ao total de unidades na amostra n em relação ao tamanho populacional N. Se este for o caso, a amostragem estratificada é dita com alocação proporcional: o número de elementos sorteados em cada estrato é proporcional ao número de elementos existentes no estrato.

Outros tipos de amostragem estratificada são: a uniforme, em que o tamanho da amostra em qualquer estrato é fixo; e a alocação ótima, em que o tamanho da amostra em cada estrato é definido de forma a minimizar alguma determinada função-custo.

As principais propriedades da estimativa $\bar{X}_{st}$ estão indicadas nos teoremas seguintes.

Teorema 3.4 *Se, em todos os estratos, o estimador da média amostral $\bar{X}_h$, for sem tendência, então $\bar{X}_{st}$ é um estimador sem tendência do valor médio da população μ.*

Demonstração: Tomemos

$$\begin{aligned} E(\bar{X}_{st}) &= E\left[\frac{\sum_{h=1}^{L} N_h \bar{X}_h}{N}\right] \\ &= \frac{\sum_{h=1}^{L} N_h \mu_h}{N} \end{aligned} \tag{35}$$

uma vez que as estimativas são sem tendência nos estratos individuais. Entretanto, o valor médio da população μ pode ser dado pela fórmula

$$\mu = \frac{\sum_{h=1}^{L} \sum_{i=1}^{N_h} y_{hi}}{N}$$

Temos que

$$\sum_{i=1}^{N_h} y_{h_i} = N_h \mu_h$$

Logo

$$\mu = \frac{\sum_{h=1}^{L} N_h \mu_h}{N}$$

Coincidindo com a equação 35. Portanto o resultado segue. □

Para a amostragem estratificada, a variância de $\bar{X}_{st}$ é dada pela equação

$$\begin{aligned} Var(\bar{X}_{st}) &= \frac{\sum_{h=1}^{L} (N_h)^2 Var(\bar{X}_h)}{N^2} \\ &= \sum_{h=1}^{L} (W_h)^2 Var(\bar{X}_h). \end{aligned} \tag{36}$$

onde

$$V(\bar{X}_h) = E[(\bar{X}_h - \mu_h)^2] \tag{37}$$

Teorema 3.5 *Se a amostragem aleatória for estratificada e se em cada estrato se fizer uma amostragem aleatória simples, então a variância do estimador $\bar{X}_{st}$ é*

$$Var(\bar{X}_{st}) = \frac{1}{N^2} \sum_{h=1}^{L} N_h (N_h - n_h) \frac{\sigma_h^2}{n_h} = \sum_{h=1}^{L} W_{h^2} \frac{\sigma_h^2}{n_h} (1 - f_h) \tag{38}$$

Demonstração: Uma vez que $\bar{X}_h$ é um estimador sem tendência de μ_h, podemos aplicar o teorema anterior, além disso temos que

$$Var(\bar{X}_h) = \frac{\sigma_h^2}{n_h}\frac{N_h - n_h}{N_h}$$

que, substituindo na expressão (47), obtemos

$$\begin{aligned} Var(\bar{X}_{st}) &= \frac{1}{N^2}\sum_{h=1}^{L} N_h^2 Var(\bar{X}_h) \\ &= \frac{1}{N^2}\sum_{h=1}^{L} N_h(N_h - n_h)\frac{\sigma_h^2}{n_h} \\ &= \sum W_h^2 \frac{\sigma_h^2}{n_h}(1 - f_h) \end{aligned}$$

□

Alguns casos particulares dessa equação são dados nos corolários abaixo.

Corolário 3.1 *Se as frações de amostragem $\frac{n_h}{N_h}$ forem desprezíveis em todos os estratos, então*

$$Var(\bar{X}_{st}) = \frac{1}{N^2}\sum \frac{N_h^2\sigma_h^2}{n_h} = \sum \frac{W_h^2\sigma_h^2}{n_h} \tag{39}$$

Em outras palavras, quando se pode desprezar as correções populacionais finitas usa-se a equação anterior.

Corolário 3.2 *No caso em que a repartição é proporcional, pode-se substituir n_h por seu valor $\frac{nN_h}{N}$ na equação* (38)*, desse modo reduzindo o cálculo da variância para*

$$Var(\bar{X}_{st}) = \sum \frac{N_h}{N}\frac{\sigma_h^2}{n}\left(\frac{N-n}{N}\right) = \frac{1-f}{n}\sum W_h\sigma_h^2 \tag{40}$$

Corolário 3.3 *Se a amostragem for proporcional e as variâncias de todos os estratos tiverem o mesmo valor σ^2, obtem-se a equação simplificada para a variância*

$$Var(\bar{X}_{st}) = \frac{\sigma^2}{n}\left(\frac{N-n}{N}\right) \tag{41}$$

4 Erro amostral

Nesta seção vamos tratar do erro amostral ε que é a máxima diferença que o investigador admite suportar entre μ e $\bar{X}$.

Na teoria dos erros, o erro amostral ε é considerado uma quantidade desconhecida da qual se supõe conhecer apenas a distribuição de probabilidades em função da grandeza que está sendo mensurada.

O erro total em geral pode ser escrito como uma soma de i erros elementares ε_1, ε_2, ...,ε_i

$$\varepsilon = \varepsilon_1 + \varepsilon_2 + ... + \varepsilon_i. \tag{42}$$

Estes erros elementares ε_i podem ter diferentes distribuições de probabilidades, como retangular, triangular, gaussiana, entre outras. Entretanto, quando o erro total ε resulta em uma superposição de vários erros elementares independentes, a distribuição de probabilidades para ε tende a se tornar gaussiana.

No caso de pesquisas de intenções de votos, o tamanho dos erros tolerados é estipulado de acordo com o tempo restante para a ocorrência das eleições. Observa-se que a magnitude dos erros diminui à medida em que a data das eleições se aproxima, comumentemente decrescendo de 5% a 3% e 2%.

4.1 Distribuição normal

A distribuição normal constitui a base teórica de toda inferência estatística, por descrever muitas distribuições de frequências de erros de observações e mensurações, como dito na subseção anterior. Também conhecida como distribuição gaussiana, ela tem aplicação prática na descrição de fenômenos naturais e sociais.

Uma variável aleatória X tem distribuição normal com média μ e variância σ^2 se sua função densidade de probabilidade for dada por:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\left[\frac{(x-\mu)^2}{2\sigma^2}\right]} \tag{43}$$

onde $-\infty < x < \infty$, $\pi = 3,1416...$, $e = 2,7182....$

O gráfico de uma distribuição normal assemelha-se muito a um sino e seu formato dependerá dos parâmetros μ e σ.

O gráfico a seguir mostra a curva para a densidade de uma variável aleatória normal de média 0 e variância 1.

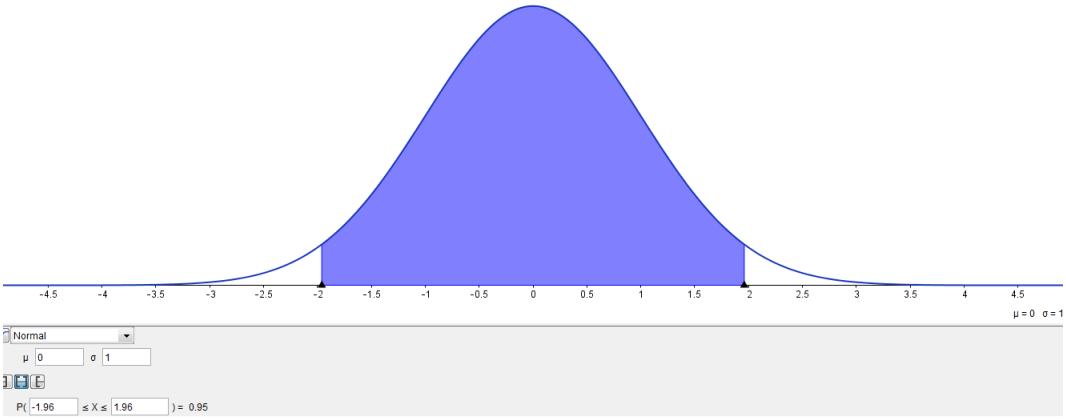

Gráfico da densidade de uma v.a. normal de média 0 *e variância* 1. *A área sobreada equivale a* 95% *de toda a área sob a curva.*

Uma propriedade útil é que se X tiver distribuição normal com média μ e variância σ^2, então a v.a. Z dada pela transformação linear

$$Z = \frac{X - \mu}{\sigma} \tag{44}$$

tem distribuição normal com média 0 e variância 1.

Assim

$$\mathbb{P}([-a < Z < a]) = \mathbb{P}([\mu - a\sigma < X < \mu + a\sigma])$$

O principal teorema aplicado na inferência estatística é o teorema central do limite. Segue uma versão deste teorema.

Teorema 4.1 *Sejam* $X_1, X_2, \ldots$ *variáveis aleatórias independentes e identicamente distribuídas com* $E[X_n] = \mu$ *e* $Var(X_n) = \sigma^2$, *onde* $0 < \sigma^2 < \infty$. *Defina* $S_n = X_1 + \cdots + X_n$. *Então a seqüência de variáveis* Z_n *definidas por*

$$Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}} \tag{45}$$

converge em distribuição para uma variável normal com média 0 *e variância* 1 *quando* $n \to \infty$.

Para demonstração ver James [9].

Na amostragem estatística em populações finitas, não é possível tomar o tamanho da amostra n tendendo a infinito. Considere então $\bar{X}$, estimador da média amostral, com média μ

e variância $\frac{\sigma^2}{n}(1-f)$. Para uma amostra razoavelmente grande, $\bar{X}$ tenderá aproximadamente a convergir para uma distribuição normal com média μ e variância $\frac{\sigma^2}{n}(1-f)$.

Neste sentido é possível definir intervalos de confiança para a estimativa $\hat{\mu}$. De fato, definindo

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}\sqrt{1-f}} \tag{46}$$

podemos considerar que Z tem distribuição normal com média 0 e variância 1. Então

$$\mathbb{P}([-a < Z < a]) = \mathbb{P}\left(\left[\mu - a\frac{\sigma}{\sqrt{n}}\sqrt{1-f} < \bar{X} < \mu + a\frac{\sigma}{\sqrt{n}}\sqrt{1-f}\right]\right). \tag{47}$$

Se escolhermos, em particular, $a = 1,96$, obtemos

$$\mathbb{P}\left(\left[\mu - 1,96\frac{\sigma}{\sqrt{n}}\sqrt{1-f} < \bar{X} < \mu + 1,96\frac{\sigma}{\sqrt{n}}\sqrt{1-f}\right]\right) = 95\%$$

Assim, se $\hat{\mu}$ é uma estimativa de μ, podemos dizer que, com 95% de probabilidade (i.e. ao nível de confiança de 95%) μ pertence ao intervalo

$$\left(\hat{\mu} - 1,96\frac{\sigma}{\sqrt{n}}\sqrt{1-f}\ ,\ \hat{\mu} + 1,96\frac{\sigma}{\sqrt{n}}\sqrt{1-f}\right).$$

De forma mais geral, se

$$\mathbb{P}\left(\left[\mu - a\frac{\sigma}{\sqrt{n}}\sqrt{1-f} < \bar{X} < \mu + a\frac{\sigma}{\sqrt{n}}\sqrt{1-f}\right]\right) = p_a,$$

então ao nível de confiança p_a, μ pertence ao intervalo

$$\left(\hat{\mu} - a\frac{\sigma}{\sqrt{n}}\sqrt{1-f}\ ,\ \hat{\mu} + a\frac{\sigma}{\sqrt{n}}\sqrt{1-f}\right). \tag{48}$$

4.2 Tamanho da amostra

Para determinação do tamanho da amostra, o pesquisador precisa especificar o **erro amostral tolerável**, ou seja, o quanto ele adminte errar na avaliação dos parâmetros de interesse. Por exemplo, na divulgação de pesquisas eleitorais, é comum encontrarmos no relatório algo como: *a presente pesquisa tolera um erro de* 2%. Isso quer dizer que, quando a pesquisa aponta determinado candidato com 20% de preferência do eleitorado, está afirmando, na verdade, que a preferência por este candidato é um valor do intervalo de 18% a 22% (ou seja, $20\% \pm 2\%$).Geralmente essas pesquisas são feitas com um intervalo de confiança de 95%.

Suponha que o erro amostral seja

$$\varepsilon = |\mu - \hat{\mu}|. \tag{49}$$

Como μ deve pertencer ao intervalo

$$\left(\hat{\mu} - a\frac{\sigma}{\sqrt{n}}\sqrt{1-f}\ ,\ \hat{\mu} + a\frac{\sigma}{\sqrt{n}}\sqrt{1-f}\right) \tag{50}$$

com probabilidade p_a, então

$$|\mu - \hat{\mu}| \approx a\frac{\sigma}{\sqrt{n}}\sqrt{1-f}. \tag{51}$$

Daí obtemos que

$$n \approx \frac{a^2\sigma^2}{\varepsilon^2}(1-f). \tag{52}$$

Caso $f \approx 0$, definimos

$$n_0 = \frac{a^2\sigma^2}{\varepsilon^2}. \tag{53}$$

Mas se quisermos corrigir este valor quando f não for desprezível, então substituindo (53) em (52) teremos

$$n = n_0(1-f) = n_0\left(1 - \frac{n}{N}\right), \tag{54}$$

donde, explicitando n, obtemos

$$n = \frac{n_0}{1 + \frac{n_0}{N}}. \tag{55}$$

Exemplo 4.1 *Deseja-se fazer um levantamento por amostragem de* $N = 2000$ *funcionários de uma determinada empresa em relação a melhorias no setor onde se encontram. Tais características são especialmente do tipo percentual, como percentual de pessoas que já sofreram acidentes de trabalho, pessoas que obtiveram promoções por meio de desempenho no trabalho, entre outras. Qual deverá ser o tamanho mínimo de uma amostra aleatória simples, tal que se possa admitir com erro amostral que não ultrapasse* $2\%(\varepsilon = 0,02)$*?*

Para uma primeira aproximação em que $a = 1,96$ *ou seja,* 95% *de confiança e* $\sigma = 0,5$ *temos que*

$$n_0 = \frac{(1,96)^2.(0,5)^2}{(0,02)^2}$$

Pode-se considerar que $(1,96)^2 \approx 4$ *e tendo que* $(0,5)^2 = \frac{1}{4}$ *então,*

$$n_0 = \frac{4.\frac{1}{4}}{(0,02)^2},$$

donde temos

$$n_0 = \frac{1}{(0,02)^2} = 2500$$

ou seja, 2500 *funcionários. Fazendo a correção do tamanho* N *do número de funcionários temos:*

$$n = \frac{(2000).(2500)}{2000 + 2500} = \frac{500000}{2750} = 1111 \text{ funcionários.}$$

Há no total $\binom{2000}{1111}$ *amostras para tal pesquisa com o erro amostral estipulado.*

Em pesquisas de inferência estatística os peritos estão interessados em construir uma família de conjuntos que contenham o valor do verdadeiro parâmetro com uma probabilidade alta especificada. Uma alternativa para apresentar um único valor sensato para representar o parâmetro que está sendo estimado é calcular e relatar um intervalo completo de valores plausíveis, sendo então essa uma estimativa de intervalo de confiança.

Os níveis de confiança usados com mais frequência são 90%, 95% e 99%. Quanto maior o nível de confiança, mais fortemente acredita-se que o valor do parâmetro que está sendo estimado está dentro do intervalo. No caso do exemplo anterior o tamanho da amostra está muito próximo da população, o que não é interessante em uma pesquisa de intenções de voto.

Vamos, a seguir, ilustrar como calcular o tamanho de uma amostra em uma pesquisa eleitoral de acordo com um erro esperado informado. Em primeiro lugar os institutos de pesquisa procuram saber qual é a população eleitoral de determinada região.

Por dados obtidos do Tribunal Regional Eleitoral (TRE), temos que atualmente a população votante no estado de Minas Gerais é de $N = 15.697.993$, ver [14]. Assumindo-se o erro de 2% determinamos o tamanho da amostra por

$$n_0 = \frac{1}{(0,02)^2} = 2500.$$

A seguir, corrigindo para uma população finita, temos que

$$n = \frac{(15697933) \cdot (2500)}{15697933 + 2500} = 2499,60$$

Portanto, usa-se o valor de $n = 2500$ eleitores a serem entrevistados.

No ano de 2016 houve eleições para prefeito e foram observadas as pesquisas ao longo da campanha eleitoral, citadas no próximo tópico, na cidade de Belo Horizonte. Com uma população eleitoral de 1.927.460 eleitores, estima-se que a amostra da população seja de 400 pessoas com margem de erro de 5%, de 1110 pessoas com margem de erro de 3% e de 2497 eleitores com margem de erro de 2%.

Exemplo 4.2 *Considere uma população com* $N = 100$*, desejamos a partir dessa, retirar uma amostra, com e* $n = 30$ *e* 95% *de certeza. Segue uma tabela com os supostos eleitores e seus votos.*

Eleitor	*Voto*	*Eleitor*	*Voto*	*Eleitor*	*Voto*	*Eleitor*	*Voto*	*Eleitor*	*Voto*
1	*B*	2*1*	*A*	41	*B*	61	*A*	81	*B*
2	*A*	22	*B*	42	*A*	62	*B*	82	*A*
3	*A*	23	*A*	43	*B*	63	*A*	83	*A*
4	*B*	24	*A*	44	*A*	64	*A*	84	*A*
5	*B*	25	*A*	45	*B*	65	*B*	85	*A*
6	*B*	26	*A*	46	*B*	66	*B*	86	*A*
7	*A*	27	*B*	47	*A*	67	*B*	87	*B*
8	*A*	28	*B*	48	*A*	68	*B*	88	*B*
9	*B*	29	*A*	49	*B*	69	*A*	89	*B*
10	*B*	30	*B*	50	*A*	70	*A*	90	*A*
11	*B*	31	*B*	51	*B*	71	*A*	91	*B*
12	*A*	32	*B*	52	*B*	72	*B*	92	*B*
13	*B*	33	*A*	53	*B*	73	*A*	92	*B*
14	*B*	34	*B*	54	*A*	74	*B*	94	*B*
15	*B*	35	*B*	55	*A*	75	*A*	95	*B*
16	*B*	36	*B*	56	*B*	76	*A*	96	*B*
17	*A*	37	*B*	57	*B*	77	*A*	97	*B*
18	*B*	38	*A*	58	*B*	78	*B*	98	*A*
19	*A*	39	*A*	59	*A*	79	*B*	99	*B*
20	*A*	40	*B*	60	*A*	80	*A*	100	*B*

Planilha de votação para 100 *eleitores nos candidatos A e B.*

Observa-se então, nessa tabela, que o valor n_0 *para se obter uma amostra com* 30 *eleitores é*

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

$$\Downarrow$$

$$30 = \frac{100.n_0}{100 + n_0}$$

$$\Downarrow$$

$$30(100 + n_0) = 100.n_0$$

$$\Downarrow$$

$$n_0 \cong 42,85$$

Por meio da expressão 73*, podemos obter que*

$$\varepsilon^2 = \frac{a^2 \cdot \sigma^2}{n_0}, \tag{56}$$

onde $a^2 \cdot \sigma^2 = 1$ *devido ao fato de ser atribuído* 95% *de confiança e* $\sigma^2 = 0,25$.

Para a equação $n_0 = \dfrac{1}{\varepsilon^2}$ *temos então que*

$$42,85 = \frac{1}{\varepsilon^2} \Rightarrow\Rightarrow \varepsilon \cong 0,15$$

ou seja, a margem de erro é de, aproximadamente 15%.

O número de amostras que se pode obter com 30 *eleitores é* $\binom{100}{30}$.

Os parâmetros populacionais referentes aos dados da tabela são:

$$\mu = 0,5600\text{; } \sigma^2 = 0,00196\text{; } \sigma = 0,04422$$

A variância do estimador $\bar{X}$ *é* $Var(\bar{X}) = 0,000028$.

Escolhendo-se então uma amostra $\omega_1 = \{1, 3, 5, 12, 14, 16, 21, 22, 27, 30, 32, 33, 39, 41, 44, 50, 56, 68, 69, 71,\ 72, 79, 83, 87, 88, 89, 90, 91, 99, 100\}$*, temos que há 13 eleitores votando em A e* 17 *eleitores votando em B.*

Nessa amostra,

$$\hat{\mu} = 0,56667;\ s^2 = 0,00648 \text{ e } s = 0,08047.$$

O intervalo de confiança para a amostra ω_1 *para* $a = 95\%$ *é*

$$(0,54261 < \bar{X} < 0,59079)$$

Os valores dos limites do intervalo de confiança indicam que, para a amostra, os votos em relação ao candidato A, estão entre 54,26% *e* 59,07% *aproximadamente.*

Na amostra $\omega_2 = \{1, 3, 5, 7, 9, 11, 13, 15, 17, 19, 21, 23, 25, 27, 29, 31, 33, 35, 37, 39, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50\}$ *obtivemos que* 15 *eleitores votaram em A.*

Temos então que, para a amostra ω_2 *:*

$$\hat{\mu} = 0,5;\ s^2 = 0,00862 \text{ e } s = 0,09284$$

O intervalo de confiança para a amostra ω_2 *para* $a = 95\%$ *é*

$$(0,47220 < \bar{X} < 0,52780)$$

Os valores dos limites do intervalo de confiança indicam que, para a amostra, os votos em relação ao candidato A, estão entre $47,22\%$ *e* $52,78\%$ *aproximadamente.*

Na amosta $\omega_3 = \{54, 55, 56, 57, 61, 62, 63, 64, 65, 66, 67, 68,\ 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 87, 88, 89, 96, 97, 98\}$*, temos que* 16 *eleitores votaram em A.*

Temos então que, para a amostra ω_3 *:*

$$\hat{\mu} = 0,5\bar{3};\ s^2 = 0,00981e\ s = 0,09904$$

O intervalo de confiança para a amostra ω_3 *para* $a = 95\%$ *é*

$$(0,50365 < \bar{X} < 0,56295)$$

Os valores dos limites do intervalo de confiança indicam que, para a amostra, os votos em relação ao candidato A, estão entre $50,36\%$ *e* $56,29\%$ *aproximadamente.*

5 Análise de procedimentos e divulgação de pesquisas de intenção de votos

Analisando as pesquisas referentes às eleições do ano de 2014 observam-se divergências entre institutos de pesquisa. Embora os níveis de confiança e margens de erro sejam semelhantes, o número de eleitores entrevistados varia entre os institutos de pesquisa, o que supõe métodos diferentes de amostragem. Observa-se também que alguns institutos consideram como parâmetro o número de votos válidos e outros, o total de votos.

As pesquisas feitas nos estágios iniciais das campanhas eleitorais apresentam margem de erro maior, contudo, à medida em que a data da pesquisa se aproxima da data da eleição, reduz-se a margem de erro amostral.

Todas as pesquisas eleitorais feitas, embora algumas com número de entrevistados ou tempo de realização diferentes, são registradas pelo Tribunal Superior Eleitoral (TSE). A página do TSE na internet exibe a numeração dos protocolos de todas as pesquisas realizadas.

À medida em que as eleições se aproximam, a mídia mostra a evolução das intenções de votos, ou seja, se o percentual de votos em relação aos candidatos aumentou ou diminuiu ao longo das campanhas eleitorais.

Na análise da campanha eleitoral do ano de 2014 os institutos de pesquisa IBOPE e Data Folha realizaram pesquisas no início da campanha eleitoral em que a margem de erro era de 5%. No mês de setembro do mesmo ano as duas empresas de pesquisa fizeram pesquisas com margem de erro de 3%. Já no período de 1 a 4 de outubro o IBOPE pesquisou a intenção de voto de 2002 eleitores. No período de 3 e 4 de outubro, o instituto Datafolha fez sua pesquisa com 2325 eleitores.

Outro ponto a se destacar são os muitos comentários indevidos sobre margem de erro. A imprensa e os demais meios de comunicação usam termos que não são adequados ao que está sendo mostrado. Uma reportagem de 27/10/2014 do jornal *"O Tempo"* [10] mostrou uma comparação das margens de erros de alguns institutos de pesquisa, dizendo que apenas um instituto conseguiu calcular com exatidão o resultado das eleições presidenciais de 2014. A idéia que se pretendeu passar foi a de que os institutos erraram o prognóstico. Porém, como visto nas seções anteriores, o resultado de uma pesquisa por amostragem apresenta um valor provável dentro de um limite de confiança. O fato da estimativa não bater com o da eleição não significa que o instituto errou, mas sim que a amostra aleatória extraída para realizar a pesquisa era pouco representativa do estado da população.

Em relação ao que se informa sobre o nível de confiança, por exemplo, de 95%, segundo o IBOPE [10], se 100 amostras forem tiradas da população e se um candidato teve 30% de intenção de votos, então, em pelo menos 95 amostras, o índice deste candidato deverá variar

entre 27% e 33% (o erro amostral sendo de 3%), mas em 5 amostras os resultados estarão fora deste intervalo. Esta é uma explicação rebuscada, mas equivocada. Já muitos jornais escrevem, de forma errônea, que se a pesquisa fosse feita 100 vezes, em 95 delas o resultado seria o mesmo. Na verdade o nível de confiança garante apenas que a probabilidade do parâmetro estimado estar dentro do intervalo de confiança é de 95%.

6 Proposta para aplicação em sala de aula

Proposta 1

Podemos fazer uma exemplificação do que acontece em uma pesquisa de intenção de votos propondo aos alunos uma atividade prática.

Considere a população N do total de alunos de um turno. Desejamos fazer uma pesquisa a respeito das chapas para o grêmio estudantil. Como demanda muito tempo a pesquisa feita com todos os alunos do turno, vamos então fazer uma pesquisa por amostragem.

1. Pesquise na secretaria escolar o número de alunos de cada uma das turmas do turno da manhã.

2. Calcule o valor **n** da amostra de alunos que devem ser entrevistados para que a pesquisa tenha erro de 5%.

3. Calcule a proporção de alunos por sala a serem entrevistados e realize uma pesquisa de acordo com a proporção de alunos por sala.

4. Calcule as estimativas de proporção de intenção de votos bem como a estimativa da variância.

5. Construa os gráficos referentes ao resultado da pesquisa e divulgue o resultado.

Proposta 2

Pesquisa simulada com auxílio de uma planilha eletrônica.

1. Monte uma tabela de 100 eleitores de uma determinada região e disponha aleatoriamente as letras A e B para descreverem os candidatos.

2. Calcule os parâmetros populacionais de proporção de votos em A.

3. Determine o tamanho da amostra para o erro 15%.

4. Faça uma escolha aleatória de 5 amostras.

5. Calcule as estimativas em cada amostra da proporção de votos em A juntamente com a estimativa da variância.

6. Analise os resultados.

7 Conclusão

Neste trabalho apresentamos os conceitos principais de probabilidade e de amostragem necessários para se entender como é feita e interpretar o resultado de uma pesquisa amostral. Vimos que o fato da amostra ser selecionada aleatoriamente é crucial para se poder estimar o erro cometido quando se compara a estimativa com o valor de um parâmetro populacional.

Quando se analisa casos reais, verifica-se que o discurso utilizado na divulgação de pesquisas destoa do que é demonstrado matematicamente na teoria. Esperamos que este trabalho contribua para a elucidação de conceitos junto à população em geral e aprimoramento da educação auxiliando professores e alunos. Em especial é importante levar os alunos a terem uma postura crítica frente aos resultados de pesquisas, ensinando-os a questionar as informações recebidas.

Agradecimentos

Agradeço em primeiro lugar a Deus, por ter me mostrado que sou capaz de vencer obstáculos para alcançar meus objetivos, aos professores da UFSJ/CAP, em especial ao professor Telles Timóteo pela paciência em minha orientação, às professoras Gilcélia e Mariana, que ajudaram nos momentos finais do curso, à minha mãe, pelo incentivo ao estudo, à minha sogra e minha cunhada, tão importantes por me auxiliar nos momentos em que estive ausente e ajudar com a Isabel, e a essa, agradeço imensamente a paciência, a compreensão de ver que a mãe esteve ausente em muitos momentos da sua vida, ao meu marido e aos demais familiares por entender minha ausência. Aos amigos pela disposição em me ajudar como a Ana Carolina Gonçalves, nas traduções em inglês, Carla Renata e Ronaldo Figueiredo pelo incentivo acadêmico, aos colegas e direção das escolas onde trabalho por me incentivar.

Agradeço aos grandes amigos feitos nesse curso, pelo apoio em não deixar que o cansaço e as dificuldades me deixassem desistir, em especial à Eliane e Luis Gustavo, dois anjos que Deus colocou na minha vida.

Agradeço também à CAPES pelo apoio financeiro.

Referências

[1] Albuquerque, José Paulo de Almeida, *et al. Probabilidade, variáveis aleatórias e processos estocásticos.* Editora Interciência, Rio de Janeiro, 2008.

[2] Barbetta, Pedro Alberto. *Estatística aplicada às ciências sociais.* Editora da UFSC, Florianópolis, 2004.

[3] Bolfarine, Heleno e Bussab, Wilton O. *Elementos de amostragem.* Editora Blücher, São Paulo, 2005.

[4] Cochran, William G. *Técnicas de Amostragem.* Editora Fundo de Cultura, Rio de Janeiro, primeira edição brasileira 1965.

[5] Devore, Jay L. *Probabilidade e Estatística: para Engenharia e Ciências.* Editora Thomson [Tradução Joaquim Pinheiro Nunes da Silva], São Paulo, 2006.

[6] Fonseca, Jaime. *Estatística Matemática - Vol 1.*, Sílabo, Lisboa, 2001.

[7] Fonseca, Jaime. *Estatística Matemática - Vol 2.*, Sílabo, Lisboa, 2001.

[8] IBGE. *http://www.eleicoes2014.com.br/pesquisa-eleitoral-minas-gerais/* , acesso em 19 de julho de 2016. Acesso em 28 fev. 16

[9] James, Barry R.*Probabilidade: um curso em nível intermediário.*, IMPA, Rio de Janeiro, 2011.

[10] Jornal O Tempo. *http://www.otempo.com.br/blogs/pol%C3%ADtica−19.298822/datafolha-acerta-em-cheio-ibope-na-margem-de-erro-sensus-e-verit%C3%A1−fracassam−19.328953.* Acesso em 22 jun.16.

[11] Loesch, Cláudio. *Probabilidade e Estatística.* LTC, Rio de Janeiro, 2014.

[12] Ross, Sheldon. *Probabilidade, um curso moderno com aplicações.* Bookman, Porto Alegre, 2010.

[13] Spiegel, Murray R. *Probabilidade e estatística.* Editora da Pearson, São Paulo, 2004.

[14] TRE.*http://www.tse.jus.br/eleitor/estatisticas-de-eleitorado/consulta-quantitativo*, acesso em 26 de agosto de 2016.