Universidade de São Paulo - USP
Instituto de Ciências Matemáticas e de Computação - ICMC
Departamento de Matemática Aplicada e Estatística - SME

Introdução à Inferência Estatística

Trabalho 3

Fernando Henrique Cardoso - Número USP: 10365671
Gabriel Ribeiro - Número USP: 8531000
Nelson Ricardo Coelho do Nascimento - Número USP: 11218448
Stéfane Tame Monteiro Oliveira - Número USP: 10829970
Profa. Dra. Katiane Silva Conceição

1. Queremos determinar o tamanho da amostra n de modo que

$$P(|\overline{X} - \mu| \leq \varepsilon) \geq \gamma = 1 - \alpha$$

com $0 < \gamma < 1$ e ε, o erro amostral máximo que podemos suportar, ambos valores fixos.

Sabemos que $\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$, logo $\overline{X} - \mu \sim N\left(0, \frac{\sigma^2}{n}\right)$ e, portanto, pode ser escrita

$$P(-\varepsilon \leq \overline{X} - \mu \leq \varepsilon) = P\left(\frac{-\sqrt{n}\varepsilon}{\sigma} \leq Z \leq \frac{-\sqrt{n}\varepsilon}{\sigma}\right) \approx \gamma$$

com $Z = (\overline{X} - \mu)\sqrt{\frac{n}{\sigma}}$.

Dado γ, podemos obter z_γ da $N(0, 1)$, tal que

$P\left(-z_\gamma \leq Z \leq z_\gamma\right) = \gamma$, de modo que

$$\frac{\sqrt{n}\varepsilon}{\sigma} = z_\gamma$$

do que, finalmente, obtemos

$$n = \frac{\sigma^2 z_\gamma^{\,2}}{\varepsilon^2}$$

porém, não conhecemos a variância σ^2 da população.

Como nosso estudo tem distribuição de Bernoulli, com média $\mu = p$ e variância $\sigma^2 = p(1 - p)$ e são duas a duas independentes, podemos escrever

$$Y_n = n\overline{X}$$

Mas pelo TLC, $\overline{X}$ terá distribuição aproximadamente normal, com média p e variância $p(1 - p)\ /n$, ou seja

$$\overline{X} \sim N\left(p, \frac{p(1-p)}{n}\right)$$

Logo a transformação $Y_n = n\overline{X}$ terá a distribuição

$$Y_n \sim N(np, np(1 - p))$$

Observe que $\overline{X}$, na expressão acima, é a própria variável $\hat{p}$ e desse modo para n grande podemos considerar a distribuição amostral de p como aproximadamente normal. Sendo assim,

$$n = \frac{p(1-p)z_\gamma^{\,2}}{\varepsilon^2}$$

Caso não conheçamos p, a verdadeira proporção populacional, podemos usar o fato de que $p(1 - p) \leq 1/4$, para todo p, então podemos obter um valor conservador

$$n \approx \frac{z_\gamma^{\,2}}{4\varepsilon^2}$$

Para $\gamma = 1 - \alpha = 1 - 0,05 = 0,95$

Tem-se

$$P(|\overline{X} - \mu| \leq \varepsilon) \geq \gamma = P(|\overline{X} - \mu| \leq 0,02) \geq 0,95$$
$$P\left(-z_{\gamma} \leq Z \leq z_{\gamma}\right), \text{ para } Z = \overline{X} - \mu \text{ e } z_{0,95} = 1,96$$

Determina-se o tamanho da amostra da região R_i, com $i = 1, 2, 3, 4\, e\, 5$, para um grau de confiança maior do que 95% e um erro menor do que 2%, pela igualdade.

$$n = \frac{p(1-p)z_{\gamma}^{2}}{\varepsilon^{2}}$$

Considere o tamanho da amostra do colégio eleitoral igual a 10002 e, para n da tabela, o novo valor da amostra sobre o erro e o grau de confiança predeterminados será arredondado para o inteiro maior.

Tabela 1 - Distribuição de votantes da amostra do colégio eleitoral.

	C_1	C_2	C_3	O	B	N	I
Total	4015	3297	1452	268	278	169	181

Tabela 2 - Proporções e o tamanho da amostra referente ao colégio eleitoral.

	C_1	C_2	C_3	O	B	N	I
p	$\frac{4015}{10002}$	$\frac{3297}{10002}$	$\frac{1452}{10002}$	$\frac{268}{10002}$	$\frac{278}{10002}$	$\frac{169}{10002}$	$\frac{181}{10002}$
$1-p$	$\frac{5987}{10002}$	$\frac{6705}{10002}$	$\frac{8550}{10002}$	$\frac{9734}{10002}$	$\frac{9724}{10002}$	$\frac{9833}{10002}$	$\frac{9821}{10002}$
n	2308	2123	1192	251	260	160	171

2. A partir dos dados apresentados da amostra 4, temos que os intervalos com os erros percentuais de 2% para mais e para menos para as regiões 1-5, e esses seguem na Tabela 3.

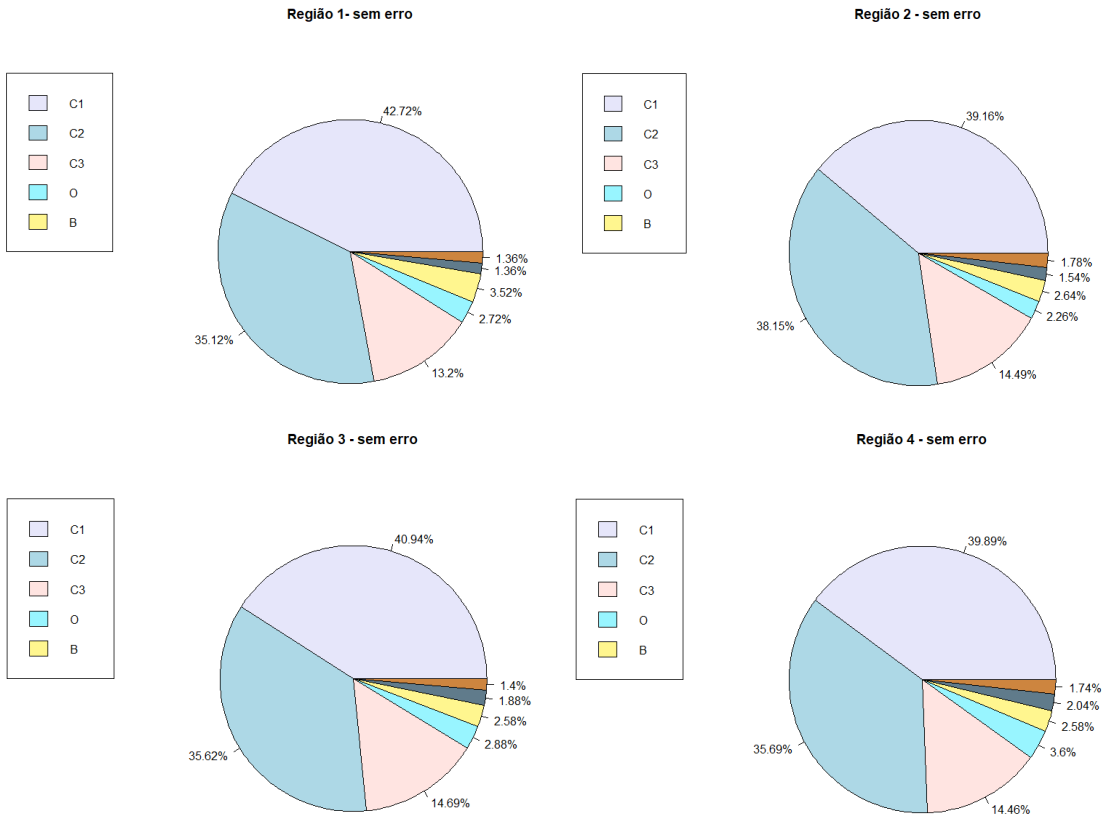
Tabela 3 – Intervalo de confiança para a proporção de votos das respectivas regiões estudadas.

	C1	C2	C3	O	B	N	I
R1							
-2%	0,4072	0,3312	0,1120	0,0072	0,0152	0,0000	0,0000
+2%	0,4472	0,3712	0,1520	0,0472	0,0552	0,0336	0,0336
R2							
-2%	0,3715	0,3614	0,1249	0,0025	0,0064	0,0000	0,0000
+2%	0,4115	0,4014	0,1649	0,0425	0,0464	0,0353	0,0377
R3							
-2%	0,3893	0,3362	0,1269	0,0088	0,0058	0,0000	0,0000
+2%	0,4293	0,3762	0,1669	0,0488	0,0458	0,0388	0,0340
R4							
-2%	0,3789	0,3369	0,1245	0,0159	0,0058	0.0004	0,0000
+2%	0,4189	0,3769	0,1645	0,0560	0,0458	0,0404	0,0374

R5							
-2%	0,3687	0,3539	0,1309	0,0048	0,0088	0,0000	0,0000
+2%	0,4087	0,3939	0,1709	0,0448	0,0488	0,0352	0,0374

A partir da análise da tabela 3, com uma confiança de 95%, podemos concluir que i) Região 1: candidato C1 ganha; ii) Região 2: Empate técnico de C1 e C2; iii) Região 3: candidato C1 ganha; iv) Região 4: candidato C1 ganha; v) Região 5: Empate técnico de C1 e C2.
Nas tabelas 5, 6 e 7, que estão adiante no trabalho, podemos ver que as regiões 1, 2 e 4 tem o erro amostral maior que 2%, logo podemos perceber que o tamanho da amostra de cada uma destas regiões não é o suficiente para atingir o erro proposto.

3. É possível visualizar melhor por meio de gráficos de pizza na Figura 1 e da Tabela 3 o desempenho de cada candidato, pela figura sem os erros percentuais, e pela tabela computando os erros percentuais.
Como se pode observar, e como dito na questão anterior, é notório que na Região 1 o C1 é favorito, na Região 2, a disputa se equilibra entre os C1 e C2, já na Região 3 e 4, os respectivos C1 lideram a disputa, e , por fim, na Região 5, há um empate técnico entre os C1 e C2.



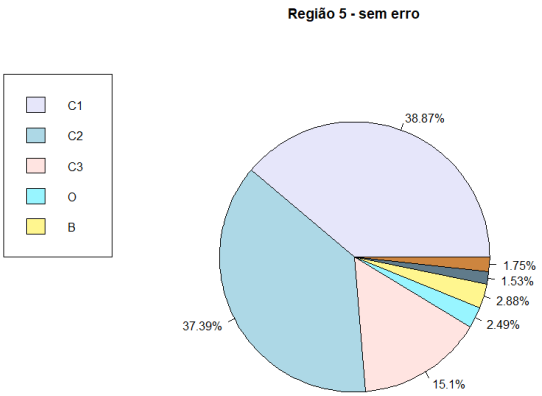


Figura 1 – Proporção de votos de cada candidato em cada região de acordo com a Amostra 4.

4. Utilizando os dados da amostra fornecida e implementando em R, foram feitos os cálculos para encontrar os erros amostrais, considerando as abordagens otimista e conservativa, e os intervalos de confiança. Sendo assim, temos as seguintes tabelas:

Tabela 4 - Erro amostral numa visão conservativa para cada região.

Região	R_1	R_2	R_3	R_4	R_5
Erro	*0,028*	*0,021*	*0,019*	*0,024*	*0,020*

Tabela 5 - Intervalos de confiança numa visão otimista das regiões R_1 e R_2.

	R_1			R_2		
	Lim. inf.	Lim. sup.	ε amostral	Lim. inf.	Lim. sup.	ε amostral
C_1	0,400	0,455	0,027	0,370	0,413	0,021
C_2	0,325	0,378	0,026	0,361	0,402	0,021
C_3	0,113	0,151	0,019	0,130	0,160	0,015
O	0,018	0,036	0,001	0,016	0,029	0,006
B	0,025	0,045	0,010	0,020	0,033	0,007
N	0,007	0,020	0,006	0,010	0,021	0,005
I	0,007	0,020	0,006	0,012	0,023	0,006

Tabela 6 - Intervalos de confiança numa visão otimista das regiões R_3 *e* R_4.

	R_3			R_4		
	Lim. inf.	Lim. sup.	ε amostral	Lim. inf.	Lim. sup.	ε amostral
C_1	0,391	0,428	0,019	0,375	0,422	0,024
C_2	0,338	0,374	0,018	0,334	0,380	0,023
C_3	0,134	0,160	0,013	0,128	0,161	0,017
O	0,022	0,035	0,006	0,027	0,045	0,009
B	0,020	0,032	0,006	0,018	0,161	0,008
N	0,014	0,024	0,007	0,014	0,027	0,007
I	0,010	0,018	0,006	0,011	0,024	0,006

Tabela 7 - Intervalos de confiança numa visão otimista da região R_5.

	Limite inferior	Limite superior	Erro amostral
C_1	0,369	0,409	0,020
C_2	0,354	0,394	0,020
C_3	0,136	0,166	0,015
O	0,018	0,031	0,006
B	0,022	0,036	0,007
N	0,010	0,020	0,005
I	0,012	0,023	0,005

Tabela 8 - Intervalos de confiança numa visão conservativa das regiões R_1, R_2 e R_3.

	R_1		R_2		R_3	
	Lim. Inf.	**Lim. Sup.**	**Lim. Inf.**	**Lim. Sup.**	**Lim. Inf.**	**Lim. Sup.**
C_1	0,399	0,455	0,370	0,413	3,905e-01	0,428
C_2	0,323	0,379	0,360	0,401	3,374e-01	0,375
C_3	0,104	0,160	0,123	0,166	1,281e-01	0,166
O	-0,001	0,055	0,001	0,044	9,964e-03	0,048
B	0,007	0,063	0,005	0,048	7,011e-03	0,045
N	-0,014	0,041	-0,006	0,037	-2,619e-06	0,038
I	-0,014	0,041	-0,004	0,039	-4,801e-03	0,033

Tabela 9 - Intervalos de confiança numa visão conservativa da região R_4 e R_5.

	R_4		R_5	
	Lim. inferior	**Lim. superior**	**Lim. inferior**	**Lim. superior**
C_1	0,375	0,423	0,368	0,409
C_2	0,333	0,381	0,353	0,394
C_3	0,125	0,169	0,130	0,171
O	0,012	0,060	0,004	0,045
B	0,002	0,050	0,008	0,049
N	0,002	0,044	-0,005	0,036
I	-0,004	0,041	-0,003	0,038

Tabela 10 - Distribuição de probabilidades.

Regiões	C_1	C_2	C_3	O	B	N	I
R_1	0,427	0,351	0,132	0,027	0,035	0,014	0,014
R_2	0,392	0,381	0,145	0,023	0,026	0,015	0,018
R_3	0,409	0,356	0,150	0,029	0,026	0,019	0,014
R_4	0,399	0,357	0,145	0,036	0,026	0,020	0,017
R_5	0,389	0,374	0,151	0,025	0,029	0,015	0,017