Detecção de fraude em urnas eletrônicas: uma estratégia puramente estatística

Daniel Chicayban Bastos[1]
Luis Antonio Brasil Kowada[1]
Maria Cristina Bessa Moreira[2]

[1]Universidade Federal Fluminense
Instituto de Computação
Av. Gal. Milton Tavares de Souza
São Domingos, Niterói, RJ, 24210-310

[2]Universidade Federal Fluminense
Instituto de Matemática e Estatística
Rua Professor Marcos Waldemar de Freitas Reis
São Domingos, Niterói, RJ, 24210-201

10 de dezembro de 2018

Autores em ordem alfabética.

Introdução

Dois aspectos de segurança são:

- Proteger um objeto contra violações.
- Detectar a violação que não foi possível evitar
  - Estimar o tamanho da violação.

Precisamos provar aos perdedores da corrida eleitoral que eles de fato perderam. (Jeroen van de Graaf, 2017.)

A votação paralela

Tabela 1: Número de urnas auditadas pela votação paralela

unidade federativa	urnas audidatas
até 15.000 seções	3
de 15.001 a 30.000	4
acima de 30.000	5

Fonte:

- TSE (2013). Resolução n. 23.397, 17 de dezembro de 2013. Diário da Justiça Eletrônico do Tribunal Superior Eleitoral, 248:2–10.
- TSE (2018). Resolução n. 23.574, 29 de maio de 2018. Diário da Justiça Eletrônico do Tribunal Superior Eleitoral, 122:111–114.

Detecção de fraude

Uma auditoria tem um procedimento A de teste de urnas eletrônicas. O teste pode responder F, significando fraude, ou L, significando urna legítima. Similarmente, uma urna U pode ser F de fraudulenta ou L de legítima. Se a auditoria é perfeita, então

$$P(A{=}L \mid U{=}L) = 1 \quad \text{e} \quad P(A{=}F \mid U{=}L) = 0.$$

Se não é perfeita, então

$$P(A{=}F \mid U{=}F) = t \quad \text{e} \quad P(A{=}L \mid U{=}F) = 1 - t.$$

Um exemplo

Suponha 5% de fraude e sensibilidade da auditoria de 70%. Qual a probabilidade de que uma urna seja classificada como fraudulenta? Dado que $P(A{=}F|U{=}F) = 0{,}70$, o teorema das probabilidades totais nos garante que

$$\begin{aligned} P(A{=}F) &= P(A{=}F|U{=}F)P(U{=}F) + P(A{=}F|U{=}L)P(U{=}L) \\ &= 0{,}70 \times 0{,}05 + 0 \times 0{,}95 = 0{,}035. \end{aligned}$$

Então a probabilidade de se detectá-la como fraudulenta é 0,035 num universo de 5% de urnas fraudulentas na população de urnas. (Como os eventos $A{=}F$ e $A{=}L$ são complementares, obtemos a probabilidade de 0,965 da auditoria detectar uma urna como legítima.)

Extrações em série

Defina X como uma variável aleatória que represente o número de urnas fraudulentas extraídas da população de urnas. Considere, como exemplo, três extrações em série. Então $X = \{0, 1, 2, 3\}$. Assuma que a extração de uma urna não altere a proporção de urnas fraudulentas na população de urnas, ou seja, assuma que a extração é com reposição. Então

$$X \sim \text{Binomial}(n = 3, p = 0{,}035).$$

Logo, a probabilidade de detecção de fraude é

$$\begin{aligned} P(X \geq 1) &= 1 - P(X < 1) = 1 - P(X{=}0) \\ &= 1 - P(A{=}L)^3 \\ &= 1 - 0{,}965^3 \\ &\approx 0{,}10. \end{aligned}$$

Extrações em série

Se desejamos, entretanto, uma probabilidade maior de revelar a fraude, como 0,99, precisamos procurar pelo número real n que satisfaça

$$\begin{aligned} 1 - P(A{=}L)^n &= 0{,}99 \\ 1 - 0{,}965^n &= 0{,}99 \\ 0{,}965^n &= 0{,}01 \\ n &= \ln 0{,}01 / \ln 0{,}965 \approx 129{,}26. \end{aligned}$$

O número de urnas portanto precisa ser $\lceil 129{,}26 \rceil = 130$.

Generalizando o exemplo anterior...

Seja:

- γ a probabilidade desejada de detecção
- $t\theta$ a probabilidade de se encontrar, ao acaso, uma urna fraudulenta, sendo t a sensibilidade do teste e θ, a proporção de urnas fraudulentas na população de urnas

Assumindo uma extração com reposição, o tamanho n da amostra de urnas é

$$n = \left\lceil \frac{\ln(1-\gamma)}{\ln(1-t\theta)} \right\rceil .$$

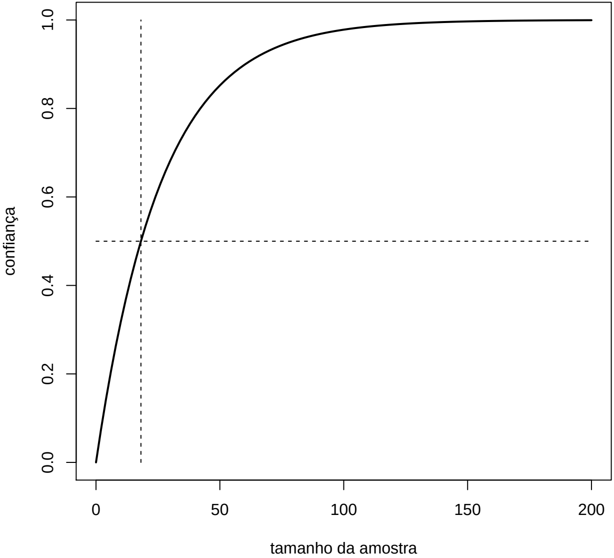
Se n for grande...

Use o que tiver — em detrimento da probabilidade γ de detecção.

Tabela 2: Decaimento da probabilidade γ de detecção com respeito à diminuição do tamanho da amostra n e da sensibilidade t dado $\theta = 0{,}05$

n	200	100	50	25	12
$t = 0{,}5$	0,99	0,92	0,72	0,47	0,26
$t = 0{,}75$	0,99	0,98	0,85	0,62	0,37
$t = 1$	0,99	0,99	0,92	0,72	0,46

Aumento da confiança em função do tamanho da amostra



Uma estimativa da proporção de fraude

Constrói-se o estimador

$$\hat{\theta} := \frac{\sum_{i=1}^{n} A_i}{nt},$$

sendo

$$A_i(U) = \begin{cases} 1 & \text{se auditoria acusa fraude em } U \\ 0 & \text{senão} \end{cases}$$

Precisaremos de uma amostra de tamanho n tal que

$$P(|\hat{\theta} - \theta| < \epsilon) \geq \gamma,$$

que é

$$n \geq \frac{z_0^2 \theta_0 (1 - t\theta_0)}{t\epsilon^2},$$

sendo z_0 o valor tal que $P(Z > z_0) = (1 - \gamma)/2$, $Z \sim \text{Normal}(0,1)$.

Exemplo de estimativa da proporção de fraude

Seja:

- $t = 0{,}75$ a sensibilidade do teste de auditoria
- $\epsilon = 0{,}01$ o erro máximo tolerável para a estimativa de θ
- $\gamma = 0{,}99$ a confiança desejada
- uma proporção inicial $\theta_0 = 0{,}05$ arbitrada

Então:

$$n \geq \frac{2{,}5758^2 \times 0{,}05 \times (1 - 0{,}75 \times 0{,}05)}{0{,}75 \times 0{,}01^2},$$

- precisamos auditar pelo menos $n = 4.258$ urnas.

Correspondente à confiança de 0,99, $z_0 \approx 2{,}5758$ na tabela da distribuição Normal padrão.

Número de urnas em função da sensibilidade e confiança

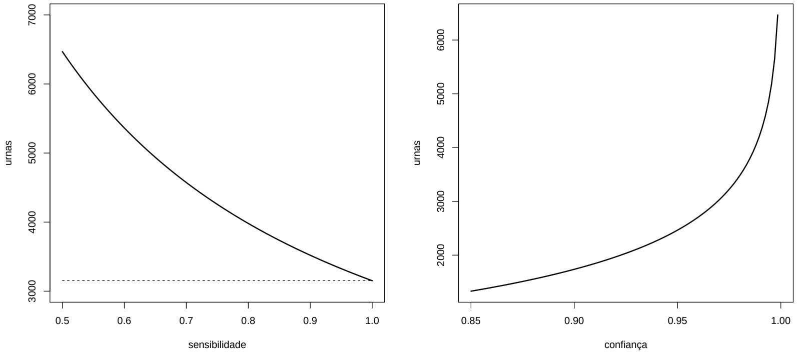


Figura 1: Impacto no tamanho da amostra usando $\epsilon = 0{,}01$ e $\theta_0 = 0{,}05$. A linha tracejada mostra o número de urnas necessárias, 3.152, se o teste da auditoria for perfeito.

Número de urnas em função de θ_0

Número de urnas auditadas em função de θ_0 e sensibilidade.

θ_0	0,01	0,02	0,03	0,04	0,05	0,10
$t = 0{,}5$	1.321	2.628	3.922	5.202	6.470	12.607
$t = 0{,}75$	879	1.743	2.595	3.433	4.258	8.184
$t = 1$	657	1.301	1.931	2.548	3.152	5.972

Sobre a extração com reposição

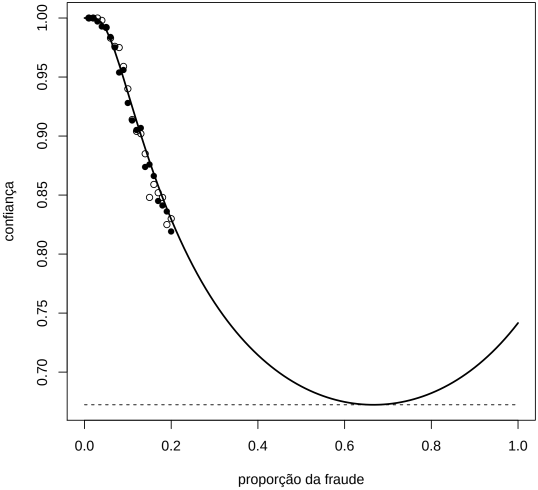


Figura 2: Simulações da confiança em amostragem com e sem reposição.

Sobre a extração com reposição

- Pontos pretos representam a confiança em função da proporção de fraude na extração sem reposição; os brancos, com reposição.
- A curva representa a relação analítica entre a confiança e a proporção de fraude considerando a amostragem com reposição.
- A linha tracejada indica a confiança mínima de 0,6723 que é atingida com $\theta = 2/3$.
- As simulações seguem os valores do exemplo: $t = 0{,}75$, $n = 4.258$, $\epsilon = 0{,}01$.
- Simulação feita com 600 mil urnas.