

UMA ANÁLISE ESTATÍSTICA DO 2º TURNO DAS ELEIÇÕES DE 2022 PARA PRESIDENTE DO BRASIL

Resumo

Está estatisticamente provado que o *tipo de urna* (*urnas antigas* e *urnas 2020*) teve influência no resultado das eleições. Nos piores casos, em $\frac{1}{3}$ dos estados brasileiros, é mais correto afirmar “houve fraude” do que “tabagismo causa câncer de pulmão”.

Introdução

Em estatística, o *Teste Cochran-Mantel-Haenszel* (CMH) ^{1,2} é um teste usado na análise de dados categóricos estratificados. Ele permite que um investigador teste a associação entre uma variável binária, que por hipótese não tem correlação com certo resultado e o resultado em si, que deve ser também binário, levando em consideração a estratificação (repetições)³. Ao contrário do teste de McNemar, que pode lidar apenas com pares, o teste *Cochran-Mantel-Haenszel*⁴ lida com tamanhos de estratos arbitrários: há versões do Teste Cochran-Mantel-Haenszel para qualquer número de linhas e colunas do *teste de independência*, mas são raramente usados. Há três variáveis: as duas variáveis do teste de independência 2×2 (uma variável que se supõe independente ou não influente no resultado e o resultado em si), e a terceira variável que identifica as repetições (tais como horários diferentes, diferentes locais, ou diferentes estudos).

Desenvolvimento

Consideremos uma variável de resultado binária, como o resultado de uma eleição (por exemplo, as eleições de 2022 no Brasil para presidente no 2º turno) e um preditor binário, como o tipo de urna (por exemplo, urnas antigas e urnas 2020). As observações são agrupadas em estratos. Os dados estratificados são resumidos em uma série de *tabelas de independência* 2×2 , uma para cada estrato, sendo que os extratos são as localidades em que os resultados estão agrupados (zona eleitoral, cidade, grupos de cidades, estado ou região do Brasil). O teste configura-se com o *tipo de urna* como sendo a hipótese nula (o tipo de urna não tem influência no resultado ou $H_0: R = 1$). A i -ésima tabela de *independência* é:

		Candidato 1	Candidato 2	Total linhas
Fator de influência, que por hipótese não há ($H_0: R = 1$)	Urnas Antigas	A_i	B_i	N_{1i}
	Urnas 2020	C_i	D_i	N_{2i}
	Total colunas	M_{1i}	M_{2i}	T_i

A *razão de chances comum* (*common odds-ratio*, R) do total de tabelas de *independência* (dado por K) é definida como:

¹ COCHRAN, William G. *Some methods for strengthening the common χ^2 tests*. Biometrics, v.10, n.4, p.417-451, 1954.

² MANTEL, Nathan; HAENSZEL, William. Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the national cancer institute*, v.22, n.4, p.719-748, 1959.

³ AGRESTI, Alan. *An introduction to categorical data analysis*. John Wiley & Sons, 2019. 3ª ed.

⁴ MANTEL, Nathan. Chi-square tests with one degree of freedom; extensions of the Mantel-Haenszel procedure. *Journal of the American Statistical Association*, v.58, n.303, p.690-700, 1963.

$$R = \frac{\sum_{i=1}^K \frac{A_i D_i}{T_i}}{\sum_{i=1}^K \frac{B_i C_i}{T_i}}$$

A hipótese nula é que não há associação entre o tipo de urna e o resultado. Mais precisamente, a hipótese nula é $H_0: R = 1$ e a hipótese alternativa é $H_0: R \neq 1$. A estatística de teste é:

$$CMH = \frac{\left[\sum_{i=1}^K \left(A_i - \frac{N_{1i} M_{1i}}{T_i} \right) \right]^2}{\sum_{i=1}^K \frac{N_{1i} N_{2i} M_{1i} M_{2i}}{T_i^2 (T_i - 1)}}$$

e ela converge para uma distribuição χ^2 com 1 grau de liberdade sob a hipótese nula.

Implementação

Foi escolhido o software-livre *R* de estatística e a fonte primária dos dados os próprios resultados disponibilizados pelo TSE. Os arquivos contendo os dados de entrada das análises realizadas, o roteiro de cada análise, o arquivo de resultados de cada análise e o mapa das análises realizadas estão disponíveis [aqui](#).

Resultados

Os resultados gerais podem ser vistos nas Tabelas de 1 a 3 abaixo.

Tabela 1 – Resultados gerais

Extratos	R	CMH	p-value
Considerando todas as cidades			
Por zona	1,055	4.180	$< 2.2 \times 10^{-6}$
Por cidade	1,047	3.401	$< 2.2 \times 10^{-6}$
Por estado	1,174	176.735	$< 2.2 \times 10^{-6}$
Por região	1,174	176.895	$< 2.2 \times 10^{-6}$
Considerando somente as cidades menores que 100mil habitantes			
Por cidade	1,106	3.171	$< 2.2 \times 10^{-6}$
Por estado	1,119	28.499	$< 2.2 \times 10^{-6}$
Por região	1,105	23.734	$< 2.2 \times 10^{-6}$

Tabela 2 – Resultados por estado

Região	Extratos	R	σ_R	p-value
Considerando todas as cidades				
CO	DF	0.973	0.003	$< 2.2 \times 10^{-6}$
	GO	1.277	0.003	$< 2.2 \times 10^{-6}$
	MT	1.150	0.001	$< 2.2 \times 10^{-6}$
	MS	1.216	0.004	$< 2.2 \times 10^{-6}$
N	AC	1.226	0.008	$< 2.2 \times 10^{-6}$
	AP	2.632	0.008	$< 2.2 \times 10^{-6}$
	AM	1.323	0.009	$< 2.2 \times 10^{-6}$
	PA	0.884	0.002	$< 2.2 \times 10^{-6}$
	RO	2.325	0.026	$< 2.2 \times 10^{-6}$
	RR	0.869	0.001	$< 2.2 \times 10^{-6}$
	TO	1.118	0.005	$< 2.2 \times 10^{-6}$

Região	Extratos	R	σ_R	p -value
Considerando todas as cidades				
NE	AL	2.418	0.008	$< 2.2 \times 10^{-6}$
	BA	1.189	0.002	$< 2.2 \times 10^{-6}$
	CE	2.164	0.004	$< 2.2 \times 10^{-6}$
	MA	1.702	0.004	$< 2.2 \times 10^{-6}$
	PB	2.464	0.007	$< 2.2 \times 10^{-6}$
	PE	1.442	0.005	$< 2.2 \times 10^{-6}$
	PI	1.020	0.002	$< 2.2 \times 10^{-6}$
	RN	1.770	0.005	$< 2.2 \times 10^{-6}$
S	SE	1.725	0.007	$< 2.2 \times 10^{-6}$
	PR	1.934	0.004	$< 2.2 \times 10^{-6}$
	RS	0.840	0.004	$< 2.2 \times 10^{-6}$
SE	SC	0.908	0.002	$< 2.2 \times 10^{-6}$
	ES	1.212	0.003	$< 2.2 \times 10^{-6}$
	MG	0.710	0.002	$< 2.2 \times 10^{-6}$
	RJ	0.839	0.001	$< 2.2 \times 10^{-6}$
	SP	1.054	0.001	$< 2.2 \times 10^{-6}$

Tabela 3 – Resultados por estado para os piores casos de $R > 1$

Extratos	R	σ_R	p -value	$R_{\text{geral e CMH}}$
Considerando todas as cidades				
RO	2.325	0.026	$< 2.2 \times 10^{-6}$	R_{geral} 2,054 CMH 677.018
AP	2.632	0.008	$< 2.2 \times 10^{-6}$	
MA	1.702	0.004	$< 2.2 \times 10^{-6}$	
PB	2.464	0.007	$< 2.2 \times 10^{-6}$	
PR	1.934	0.004	$< 2.2 \times 10^{-6}$	
AL	2.418	0.008	$< 2.2 \times 10^{-6}$	
RN	1.77	0.005	$< 2.2 \times 10^{-6}$	
SE	1.725	0.007	$< 2.2 \times 10^{-6}$	
CE	2.164	0.004	$< 2.2 \times 10^{-6}$	

Conclusão

A magnitude do teste geral de $CMH = 677.018$ e $R_{\text{geral}} = 2,054$ para um terço dos estados brasileiros, bem como os demais resultados das análises apresentadas, permitem afirmar que, no caso apresentado das eleições 2022 de 2º turno para presidente do Brasil, a hipótese nula deve ser realmente rejeitada. Isto significa que **o tipo de urna influenciou fortemente o resultado das eleições**.

Um resultado similar de uso deste teste é a análise da correlação entre tabagismo e câncer de pulmão em cidades diferentes. Neste caso, um valor de CMH igual a 279,38 e razão de chances maior que 2,0 já foi fortemente conclusivo em desfavor da hipótese nula, que é a independência entre uma coisa e outra, entre ser fumante e o câncer de pulmão⁵.

“O valor de CMH é 279,38, que leva a uma forte rejeição da hipótese nula de independência condicional entre indicador de ser fumante e indicador de câncer de pulmão”.

⁵ CASTRO, M. de. *Teste de Cochran-Mantel-Haenszel*. Fonte: <http://wiki.icmc.usp.br/images/b/bb/Mantel2018.pdf>. Acessado em: 06/12/2022.

A hipótese de que o tipo de urna não influencia o resultado de eleição é o que deveria acontecer em eleições limpas e isentas de fraude, como uma moeda no cara-ou-coroa, ou um dado, no jogo de dados. Porém, quando o resultado, após um grande número de repetições, apresenta viés, isto é, pende mais para um lado, recai desconfiança sobre a isenção do instrumento usado, moeda, dado ou Urna. Exatamente isto que o teste estatístico de CMH se dispõe a provar quantitativamente, quando o fator que por hipótese deveria ser neutro ou isento é inextrincável de outro fator, como a localização geográfica, que carrega em si todo o perfil de preferências do eleitor.

Esta análise apresenta, então, uma medida estatística das anomalias constatadas nos resultados das eleições brasileiras, anomalias estas objeto de ampla divulgação na audiência pública realizada pelo Senado Federal em 30/11/2022, mais precisamente pelo analista técnico Fernando Cerimedo, que afirmou ser um indicativo de fraude. Segundo o site do próprio Senado Federal⁶:

Para Cerimedo, os dados levantados indicam uma “provável fraude”. Ele reafirmou que as urnas eletrônicas anteriores a 2020 não possuem memória interna e podem ter sido alvo de manipulação para favorecer o candidato Luiz Inácio Lula da Silva.

— O impedimento do fortalecimento da transparência no processo eleitoral por meio da imposição da ideia de lisura das urnas e das eleições, com a condenação das contestações, é um claro processo de inversão de conceitos — argumentou.

Já o presidente do Instituto Voto Legal, Carlos Rocha, apresentou o relatório que essa entidade fez a pedido do PL. Rocha disse que o estudo identificou “indício de mau funcionamento da urna” devido a supostos erros encontrados a partir de dados constantes na documentação pública e no “Log de Urna”, fornecidos pelo TSE. Por outro lado, ele afirmou que o próprio TSE pode realizar uma análise, a partir de colaboração com a Polícia Federal, para verificar se esses erros são apenas erros de programação.

Considerando ainda que as exatas 280.710 urnas antigas apresentaram um código espúrio onde deveria constar o código da urna, e isso repetido *em todas as linhas de todos os logs* dessas mesmas 280.710 urnas (ainda que se consiga por outros meios redundantes a identificação entre o Log e a urna), aliado ao fato de que essas urnas apresentaram anomalias inexplicáveis nos resultados em comparação com as urnas 2020, **agora quantificado estatisticamente que não é possível dizer que o tipo de urna é isento, é urgente e necessário que se proceda uma extensa auditoria nos mesmos Logs**, para que o resultado das eleições seja aceito por ambas as partes, vencedores e derrotados, sejam eles quais forem.

Referências complementares

AGRESTI, Alan; KATERI, Maria. *Foundations of statistics for data scientists: with R and Python*. Chapman and Hall/CRC, 2021.

AGRESTI, Alan; FRANKLIN, Christine. *The art and science of learning from data*. Upper Saddle River, New Jersey, v. 88, 2022. 5ª ed.

MCDONALD, John H. *Handbook of biological statistics*. Baltimore, MD: Sparky House Publishing, 2009.

FORSBERG, Ole J. *Understanding Elections through Statistics: Polling, Prediction, and Testing*. Chapman and Hall/CRC, 2020.

⁶ Fonte: <https://www12.senado.leg.br/noticias/materias/2022/11/30/participantes-de-debate-criticam-processo-eleitoral-e-pedem-que-tse-analise-denuncias>. Acessado em 06/12/2022.

Interlândia-GO, em 12 de dezembro de 2022, dia de Nossa Senhora de Guadalupe.

Flávio Mamede Pereira Gomes

CPF: 589.614.531-49