

Seis critérios, uma confirmada: o que o 1º turno de **2026** diz sobre o método

Versão técnica, para quem quer auditar. Resultado: snapshot do TSE de 4 de outubro de 2026, 22h57, com 99,97% das seções. Pré-registro: cópia congelada em 1º de outubro. Arquivos e scripts estão nomeados para quem for refazer a conta.

O método mostrou que a incerteza de especificação supera a amostral e acertou a ordem de grandeza da correção. Falhou no desenho dos próprios critérios de teste. Das seis previsões, uma foi confirmada, por uma adjudicação posterior que a seção 7 expõe, e duas morreram por evidência. As outras três não respondem, por motivos diferentes.

1. O desenho, e o que ele compra e custa

Fixei o método antes de olhar o resultado. Método escolhido depois de ver o dado tende a achar o que se quer achar. A Metodologia_Decomposicao.md tem data de 30 de setembro e diz, logo na abertura, que foi escrita antes dos dados chegarem.

O pré-registro tem seis previsões, cada uma com a regra que a confirma e a que a mata. Nem todo gatilho é numérico: P2 tem um corte (3,5 pontos), P3 dois limiares de Spearman (0,5 e 0,2), P1 compara dois erros, e P4, P5 e P6 são regras de sinal ou de comparação entre trechos. Isso pesa adiante.

O que compra. Impede que o critério seja afinado depois da resposta. Dá uma régua ao leitor: a auditoria de 4 de outubro achou correspondência integral com a cópia congelada. E deixa o erro de desenho aparecer como erro de desenho, não como surpresa dos dados.

O que custa. O texto congelado não se conserta. Quando um critério sai ruim, fica ruim. Três saíram assim (P4, P5, P6). O congelamento só vale se o defeito for declarado. Antes de congelar, houve revisão adversarial (18 achados) e uma comparação cega em outra ferramenta, que achou o erro da seção 3.

2. A decomposição do erro

Para um candidato c , o erro é $E_c = P_c - A_c$, pesquisa menos resultado. Com g um grupo do eleitorado, w seu peso (P na amostra, T entre quem compareceu) e p o voto de c no grupo:

$$\begin{aligned} E_c &= \text{SOMA}_g (w_g^P - w_g^T) * p_c | g^P && (1) \text{ composição} \\ &+ \text{SOMA}_g w_g^T * (p_c | g^P - p_c | g^T) && (2) \text{ preferência} \\ &+ (3) \text{ alocação} \end{aligned}$$

Os dois primeiros termos fecham sem termo cruzado. O terceiro entra porque pesquisa e urna não usam a mesma base. Como (2) é o resíduo, a identidade fecha por construção.

Composição é o único termo calculável sem inferência: o TSE publica quem compareceu por sexo, idade, escolaridade e UF. A abstenção de 2022 foi de 20,95%, mas de 52,08% entre analfabetos e 11,85% entre quem tem superior. O limite é a amostra realizada, que muitos institutos não publicam. Reponderando os cruzamentos de setembro de 2026 pela abstenção de 2022, a margem se move -1,0 com a BTG/Nexus e +1,6 com a AtlasIntel, porque os dois discordam em 31 pontos no gradiente de escolaridade.

Preferência sai por resíduo e mistura não resposta, mudança tardia e o que sobra de comparecimento. Por grupo, não se conclui nada: voto é secreto, e renda não existe nem no TSE. "A pesquisa errou X entre mulheres" compara pesquisa com pesquisa, não com a verdade.

Alocação. Pesquisa publica percentual do total, a urna conta válidos. Em 2022 os válidos foram 95,59% do comparecimento. No snapshot de 2026, 95,23%, com brancos em 1,84% e nulos em 2,93%. Instituto que supôs outra proporção errou antes da preferência. E o viés histórico já contém o comparecimento: somar os dois conta duas vezes.

O que foi executado. A identidade é o arcabouço, e não rodou por inteiro. A composição foi calculada só para 2026. A alocação não foi calculada. Os de 2022 vêm de uma comparação mais simples, pesquisa final contra resultado.

3. A armadilha de base, e como ela inverteu 2018

Há duas formas de errar a ponte entre total e válidos: comparar sem converter, ou converter pela metade, normalizando sobre os nomeados e comparando com o percentual oficial, calculado sobre todos. Se os nomeados não somam todos os válidos, toda estimativa sai inflada, no mesmo sentido.

O episódio. Em 2018 os quatro nomeados somaram 92,54% dos válidos. Cada estimativa minha saiu inflada em cerca de 8%, ou 3,71 pontos no primeiro colocado. Eu reportava o antipetista superestimado em 1,49. Os válidos publicados pelos próprios institutos dizem o contrário:

Instituto	Válido publicado	Erro com o publicado	Erro pela minha conversão
Datafolha	40,0	-6,03	+0,1
DataPoder360	33,0	-13,03	-7,1
CNT/MDA	42,6	-3,43	+2,0
Veritá	51,2	+5,17	+10,9
Média		-4,33	+1,49

Erro é pesquisa menos o oficial, 46,03. O sinal inverte, e "dois ciclos subestimam e um superestima" vira "três de três", ainda três observações. O erro não parecia erro: a versão errada dizia que o sinal de 2018 dependia de um instituto só, o Veritá. História coerente, sobre conta torta.

Como foi achado. Dei a outra ferramenta só a pergunta, sem as minhas conclusões (PROMPT_para_outra_ferramenta.md). Ela preservou os válidos publicados, e nos quatro institutos em comum a diferença apareceu. O GABARITO_comparacao.md tem dez armadilhas escritas por mim, e a primeira é a mistura de bases. Caí nela. Saber qual é a armadilha não impede de cair. O que impede é alguém de fora refazer a conta. A grade e o componente comum de +1,96 vêm do painel do Nexa, que inclui a coluna "Outros", e não foram afetados.

Para quem vai refazer. Prefira o válido publicado à reconversão sua. Converta só com a lista completa. Confira a soma dos nomeados antes de comparar: foi ela, 92,54%, que

entregou o erro. O mesmo tipo de erro voltou depois do resultado: uma versão anterior do `compara_previsao_vs_urna.py` tratou "falta para 50%" (9,9) como total dos demais (16,4).

4. A correlação que não se move

Estimei o patamar de cada casa com as rodadas anteriores, deixando de fora a final, e testei contra o erro da rodada final (`casa_vs_erro_v2.py`). Treze casas de 2022. Resultado: r de +0,96, inclinação de 0,97. Parecia dizer que o erro de cada casa é previsível. Tratei como achado.

Não é. O erro é pesquisa menos resultado, e o resultado é uma constante. Para qualquer x :

$$\text{corr}(x, \text{pesquisa} - R) = \text{corr}(x, \text{pesquisa}) \quad \text{para qualquer constante } R$$

A correlação com o "erro" é a correlação com o nível bruto da pesquisa. Não depende do resultado, logo não prevê erro. O teste: refazer a conta trocando o resultado real por 0, por 30, por -20 e por 999. O r ficou em 0,9557 e a inclinação em 0,9658 nos quatro casos, até a quarta decimal. Estatística que não se move quando se troca o resultado da eleição não prevê erro eleitoral.

A v1 tinha dois defeitos reais, base errada e a rodada final nos dois lados. A v2 consertou os dois, e o r continuou invariante: consertar a mecânica não conserta a lógica.

Rebaixei o 0,96 a erro metodológico, e o `LEIA-ME.txt` da cópia congelada registra a correção, feita antes de qualquer circulação. O que sobra é persistência: cada casa repete o próprio patamar de uma rodada para a seguinte, o que num mês estável vale quase por construção. O `casa_vs_erro_v2.py` ainda imprime esse r e uma "redução de incerteza de 70%", e as duas linhas medem persistência. O efeito de casa tem base sólida em outra coisa, a distância entre as casas e a consistência interna de cada uma, ambas sem tocar o resultado. Elas dizem que as casas medem populações diferentes, não qual está mais perto da verdade.

5. A grade de 24 estimativas

Publicar um número só, com banda, parece rigoroso. É falsa precisão. A grade sai de duas perguntas sem resposta única.

O que as pesquisas dizem hoje. São 39 pesquisas de 16 casas, com margens de -3,37 a +11,83. Há quatro formas defensáveis de resumi-las: média ou tendência, peso por pesquisa ou por casa. A Quaest fez cinco rodadas e nove casas fizeram uma.

O quanto erraram da última vez. Rodei oito recortes de quais casas de 2022 entram e publiquei seis, com viés de margem de 1,96 (as 13 casas com três rodadas ou mais) a 8,07 (as quatro de maior circulação). O outlier é o Brasmarket, com -21,0.

Quatro consensos vezes seis vieses, margem de Lula em votos válidos (positivo é vantagem dele):

Consenso de 2026	13 casas	sem outlier	mediana	mín. 6 rod.	6 maiores	4 maiores
Média simples (+3,35)	+1,39	-0,53	-0,55	-1,69	-3,63	-4,72
Tendência por pesquisa (+1,41)	-0,55	-2,47	-2,49	-3,63	-5,57	-6,66
Tendência, peso por casa (+4,28)	+2,32	+0,40	+0,38	-0,76	-2,70	-3,79
Média, peso por casa (+3,65)	+1,69	-0,23	-0,25	-1,39	-3,33	-4,42

Amplitude de **+2,32** a **-6,66**, quase **9** pontos. Mediana **-1,54**. Em **19** das **24** células o antipetista fica à frente. As duas dimensões são escolhas de método, não fatias do eleitorado.

É este o argumento central. A margem declarada de uma pesquisa, uns 2 pontos, descreve incerteza amostral. Escolhas defensáveis de especificação movem a estimativa em 9. A incerteza de especificação domina a amostral. Quem publica um ponto com banda estreita escolheu uma célula e escondeu 23.

Duas ressalvas. A tendência de setembro não está estabelecida: a margem de Lula cai 0,112 por dia com peso por pesquisa (t de 1,68) e sobe 0,036 com peso por casa (t de -0,20), e essas duas linhas produzem os extremos. E o viés de 2022 foi medido da última pesquisa à urna, enquanto dois consensos são médias de setembro: calibrando no mesmo período, a mediana vai de -1,54 a +0,70. A direção também depende de especificação.

O que a urna disse. A margem oficial é -1,89 na convenção de Lula. Cabe na amplitude: cinco células a menos de 1 ponto, e a mediana errou por 0,35. Fora do placar, porque a grade nunca foi previsão pontual.

6. Voto útil e voto envergonhado, sem boca de urna

Em 2022 nenhum instituto fez boca de urna, pela primeira vez desde a redemocratização. O método separou os dois fenômenos pelo lugar onde deixam marca. **Voto útil** é migração antes da urna e aparece na série de pesquisas: menores encolhem, grandes crescem. **Voto envergonhado** é eleitor que não declara e só aparece entre a última medição e o resultado. Mede-se o trecho visível (início de setembro à véspera) e o invisível (véspera à urna).

O `voto_util_vs_envergonhado.py` roda com 61 pesquisas de 15 institutos. A versão restrita, que prefiro, usa só as casas presentes nas duas pontas (janelas de 25/08 a 10/09 e de 26/09 à urna): oito casas, 28 pesquisas, com uma restrição adicional no mesmo script. Sem ela, parte do "movimento" é troca de quem media.

2022, votos válidos	Início	Véspera	Urna	Visível	Invisível
Lula, restrita	46,4	48,8	48,4	+2,4	-0,4
Bolsonaro, restrita	38,4	37,3	43,2	-1,1	+5,9
Menores, restrita	15,2	13,9	8,4	-1,3	-5,5

"Bolsonaro caía nas pesquisas enquanto subia para o resultado" existe na especificação restrita e não na outra: sem restrição, ele sobe **0,9** no visível e **3,9** no invisível. As duas concordam no essencial, o invisível é maior que o visível. O voto útil visível existiu, foi pequeno e foi para o lado de Lula.

O teste da data de fechamento de campo. Se o invisível fosse mudança tardia, quem fechou mais perto da urna erraria menos. Se fosse não resposta, fechar um dia antes ou cinco não mudaria nada. Catorze institutos fecharam entre 20/09 e 1º/10 de 2022 (teste_data_de_campo.py). Só três mediram na véspera: Datafolha, Ipec e Quaest, com erro médio de -6,4 sobre o antipetista. Quem fechou quatro dias antes ou mais teve -3,0 (5,5 em valor absoluto, contra 6,4). A correlação entre antecedência e erro absoluto é +0,18, no sentido que a mudança tardia prevê, fraca demais para sustentar algo.

O grupo de quatro dias ou mais inclui o Brasmarket, que errou 7,4 para o outro lado. Sem ele esse grupo erra 5,1 contra 6,4 da véspera, o que aponta contra a mudança tardia. Tirá-lo favorece a minha leitura, e por isso não o tirei do cálculo principal. Na grade ele também pesa: o viés de 2022 é 1,96 com o Brasmarket e 3,88 sem ele. Isso não refuta a mudança tardia: é ausência de evidência onde a hipótese previa evidência forte. Importa porque não resposta se transporta de uma eleição para outra, e mudança tardia depende da campanha. São 14 casas e uma eleição. Repetir com 2026 está previsto e não consta nos arquivos desta auditoria.

7. O placar dos seis critérios

#	Regra congelada	Resultado	Classe
P1	Média das 39 pesquisas erra menos, em margem, que a pesquisa mediana	5,2403 contra 5,1851	Morreu
P2	Erro do consenso abaixo de 3,5 pontos	5,2403	Morreu
P3	Spearman do nível de setembro contra o erro real acima de 0,5	Não calculada	Pendente de cálculo
P4	O viés comum de 2026 tem o sinal de 2022	Depende da grandeza	Confirmada, por adjudicação posterior
P5	Gradiente municipal de renda bate com uma só das duas medições	Não adjudicável	Sem resposta, por desenho
P6	Menores caem mais entre 28/09 e a urna que entre 25/08 e 15/09	Não testável	Sem resposta, por desenho

Uma confirmada, duas mortas, três sem resposta. O que separa as três sem resposta é o que mais ensina.

P1 e P2: morreram por evidência

O erro absoluto da margem do consenso foi 5,2403. O da pesquisa mediana, 5,1851. Diferença de 0,0552, contra o consenso. O critério exigia que ele ganhasse, e "igual ou maior" matava. Esses 0,0552 são a média das 39 margens (3,3519) menos a mediana (3,2967), e não dependem de onde o resultado cai, desde que caia abaixo das duas. Aqui P1 mediu a assimetria da distribuição, o que é mais estreito que "o agregado erra menos que a pesquisa individual". P2 não tem esse problema: o corte era 3,5.

A classificação é preliminar porque o snapshot é parcial, e a seção 8 mostra que não muda. P1 e P2 mortas juntas significam, pelo pré-registro, que o agregado também não serve. Eu atribuíra a isso de 10 a 20%, palpite meu, sem conta por trás.

P4: confirmada, com uma adjudicação que precisa estar à vista

O critério dizia que o viés comum de 2026 tem o mesmo sinal de 2022, confirmava se o consenso superestimasse o PT e morria se o sinal fosse contrário. Não dizia sobre qual grandeza. O resultado admite duas leituras, e elas apontam para lados opostos:

- **Em margem, confirma.** A média das 39 pesquisas dava Lula à frente por 3,4. A urna deu Flávio à frente por 1,89. O consenso superestimou a posição relativa do PT em 5,29 pontos de margem.
- **Em nível, refuta.** Lula apareceu com 43,5 e teve 45,15. Foi subestimado em 1,65, não superestimado.

A escolha foi minha, em 4 de outubro, com o resultado à vista. Escrevi o critério e escolhi a leitura. Vale a de margem, e P4 fica confirmada.

Escolher entre duas leituras depois do resultado é exatamente o que o pré-registro existe para impedir. A decisão não é ilegítima, porque a ambiguidade era real e estava no texto congelado, mas só é honesta se declarada como adjudicação posterior. O registro, com data, está em DECISA0_P4_margem.md. Quem lê precisa poder descontar: lida em nível, P4 morre, e o placar vira zero confirmadas, três mortas (P1, P2, P4) e três sem resposta. A escolha não autoriza reinterpretar P1, P2, P5 ou P6. (O 5,29 usa 43,5 e 40,1 arredondados. Sem arredondar, é 5,24, o mesmo de P1 e P2.)

P3: pendente de cálculo

A auditoria de 4 de outubro tinha escopo que excluía ranking de institutos, e P3 pede um. Não foi calculada, e não é falha nem confirmação. Com os dados que existem, é calculável.

P5: sem resposta por desenho

O critério perguntava se o gradiente municipal de renda, do resultado regredido sobre a renda do Censo, tem o sinal de **exatamente uma** das duas medições de pesquisa. As duas têm sinais opostos: +19,7 na AtlasIntel de 2026, -10,0 na BTG/Nexus. Logo qualquer gradiente municipal diferente de zero casa com exatamente uma delas. Só um gradiente nulo falharia. A regra não discrimina nada. É verdadeira quase por construção.

Foi escrita para não tomar partido de nenhuma medição, e com sinais opostos isso a esvaziou. Deixou em aberto o que devia ter fixado: qual das duas medidas de renda do Censo 2022, transformação linear ou logarítmica, peso igual ou proporcional a eleitores, e como harmonizar faixas em reais (AtlasIntel) e em salários mínimos (BTG/Nexus). A junção existe (5.570 de 5.571 municípios, por código IBGE, 59 ainda com seções não totalizadas). Serve como sensibilidade, nunca como teste confirmatório, e seria associação ecológica: não diz como indivíduos votaram.

P6: sem resposta por desenho, de outro jeito

O critério comparava a queda dos menores entre 28/09 e a urna com a queda entre 25/08 e 15/09, nas casas presentes nas duas janelas. A série tem duas, AtlasIntel/Bloomberg e Palver, cada uma com uma observação na janela inicial. Sem duas ondas dentro da janela, a queda inicial não se estima. Os números descritivos (resíduo de 18,12 na janela inicial, 10,64 na tardia, 7,81 no resultado) comparam médias de janelas, e não são a comparação registrada. A primeira versão do critério era mais simples, véspera contra setembro. A revisão apontou falta de regra de medição, e a correção, "só nas casas presentes nas duas janelas", reduziu a série a duas.

Erro de desenho do critério, não do mundo. P5 e P6 não eram calculáveis desde que foram escritas. P3 só não foi calculada.

8. Robustez e limites declarados

Snapshot parcial. O TSE registrou 499.076 de 499.248 seções, 99,97%. Faltavam 172, com 53.248 eleitores aptos. A sensibilidade joga todo esse eleitorado, no limite, para um polo e depois para o outro, o que leva a margem de -1,8884 a algo entre -1,9321 e -1,8429. Em toda a faixa, P1 e P2 ficam na mesma classe. Não é previsão de comparecimento: só testa se a classificação depende de tratar o snapshot como final, e não depende.

Denominador. A base congelada não tem série uniforme de indecisos, brancos e nulos: os nomeados somam em média 91,3% do total. Normalizar sobre os nomes disponíveis redistribui esse resíduo em proporção, a armadilha da seção 3 em versão atenuada. Em setembro, cinco denominadores alternativos moveram a estimativa entre -0,44 e -0,57. A auditoria pós-eleição não refez essa sensibilidade.

Nomes, fonte e reprodução. O painel mistura cenários com e sem Pablo Marçal, que não estava na urna. A fonte é uma compilação secundária (AFOS). Não há intervalo amostral por pesquisa nem propagação de incerteza para a média. O `erro_historico.py` não executa com os diretórios atuais (espera 2014_base na raiz, a pasta está em 03_dados), então os históricos não foram rerodados.

9. O que o método acertou

A direção e a ordem de grandeza. O método dizia que a pesquisa subestima o antipetista e que a correção é grande, de 3 a 6 pontos de margem. O erro de margem da pesquisa crua foi 5,29, e Flávio foi subestimado em 6,94.

Cenário	Lula	Flávio	Margem (Flávio menos Lula)	Erro de margem
Pesquisa crua	43,5	40,1	-3,4	5,29
Correção mediana, 3,5 p.p.	42,5	43,6	+1,1	0,79
Correção extrema, 6,3 p.p.	41,7	46,4	+4,7	-2,81
Urna	45,15	47,04	+1,89	

(compara_previsao_vs_urna.py.) A correção mediana levou o erro de margem de 5,29 para 0,79. A extrema passou do ponto e errou 2,81 para o outro lado. **A mediana ganhou da ponta.** É informação de calibração: o meio da faixa de vieses ficou mais perto da urna que a ponta, que vem das casas de maior circulação.

10. O que o método errou

O nível de Lula. Previ Lula superestimado, como em 2022. Foi subestimado em 1,65. A margem acertou porque o erro do antipetista (6,94) foi bem maior que o de Lula, e a correção empurrou a margem para o lado certo mesmo levando Lula a 42,5, a 2,65 do que teve. Quem lê só a margem vê acerto. Quem lê o nível vê erro de direção.

Acertar a razão pelo motivo errado. A proporção de 2022, 78 contra 22, quase se repetiu: 81 contra 19 em 2026 (Flávio 6,94 de 8,59, Lula 1,65). Parece confirmação. O mecanismo foi outro.

Em 2022, parte do ganho do antipetista vinha de tirar do petista: Bolsonaro apareceu 5,0 abaixo do que teve, Lula 1,4 acima. Em 2026 nada veio do petista. Os dois subiram, e a soma das duas subestimações, 8,59, é exatamente o que os demais perderam. Tudo saiu do estoque dos menores, que caiu 8,59 dos 16,4, ou 52%. Em 2022 foram 40%. Acertar a razão por mecanismo errado é acerto frágil: a regra "repartir 80 para 20" não se transporta se o mecanismo muda. Isto é problema de método, não detalhe.

E o 52 contra 40 compara bases diferentes. O 40% de 2022 parte da véspera (13,9 a 8,4). O 52% de 2026 parte da média de setembro inteiro (16,4). Pela base de início de setembro, 2022 dá 6,8 de 15,2, ou 45%. Pela janela tardia das duas casas pareadas de 2026, o resíduo foi de 10,64 a 7,81, ou 27% (descritivo). Alinhada a base, a diferença encolhe ou troca de sinal, e não sei dizer que o estoque de 2026 se moveu mais.

11. O que muda no desenho para o 2º turno

O segundo turno é em 25 de outubro. Quatro mudanças, na ordem em que teriam evitado os erros deste.

1. Todo critério nomeia a grandeza e o denominador. Margem ou nível, sobre qual base. P4 não disse, e margem e nível apontaram para lados opostos. Se um critério não puder ser avaliado sem uma escolha posterior, não está pré-registrado de verdade. P1 também pedia: "pesquisa mediana" podia ser a das margens ou a dos erros. Aqui dão o mesmo número, mas o critério devia ter dito. E toda adjudicação posterior vai para um arquivo com data, fora da peça.

2. Antes de registrar, conferir se o dado permite testar. Foi o que matou P6: o critério pedia duas ondas por instituto pareado, e a série tinha uma. A checagem é barata: rodar o critério em modo seco contra a série que já existe.

3. Um critério tem de ser falsificável antes de ser registrado. Para cada um, perguntar qual resultado plausível o mataria. Se não houver resposta clara, reescrever. Se existe um resultado plausível que o critério não rejeita, ele não é um teste, é um enfeite. P5 falhava assim. Rodar essa pergunta em cada critério do segundo turno antes de congelar. Ela vale para P3: se o erro de cada instituto é pesquisa menos resultado e o resultado é um só, a ordem por erro com sinal é a ordem por nível, e a invariância da seção 4 se aplica. A primeira P3 (r acima de 0,8) já tinha sido reescrita por ser autoconfirmada. Antes de calcular a atual, convém conferir se ela mede algo além de persistência. Não calculei: é a checagem da seção 4 aplicada ao critério, não uma conclusão.

4. Hipótese a testar: no segundo turno não há mais estoque de menores. No primeiro, 8,59 pontos saíram do bloco dos demais. No segundo restam dois nomes. Se o viés de subestimar o antipetista persistir, ele não pode vir do encolhimento dos menores, e isola melhor a hipótese de não resposta. É hipótese, não resultado. Não separa por completo: continua sem boca de urna, e o movimento entre indecisos, brancos e nulos continua possível.

O método de 2026 errou onde nenhuma das 24 células apontava: nos próprios critérios. Para o segundo turno, a regra é escrever cada critério como se um estranho fosse aplicá-lo sem saber o que eu queria dizer.

Felipe Feldens

Fundador da MudAção

4 de outubro de 2026