



Como saber se nossas mudanças de produto realmente funcionam?

Por Reyso Teixeira

Em empresas digitais, decisões sobre layout, botão de compra ou sugestão de produtos são feitas todos os dias.

Para responder, mergulhei em um conjunto de dados real de experimentos A/B realizados por uma empresa de e-commerce, onde múltiplos testes foram conduzidos em paralelo para medir efeitos em métricas de negócio.

Como medimos isto?

Usamos testes A/B, comparando grupos de controle e tratamento. Analisando a taxa de conversão de cada grupo.

Para cada experimento:

- O grupo controle representa a experiência original do site;
- O grupo tratamento recebeu alguma alteração (ex: novo layout, novo botão, personalização etc.)



O que exatamente estamos medindo?

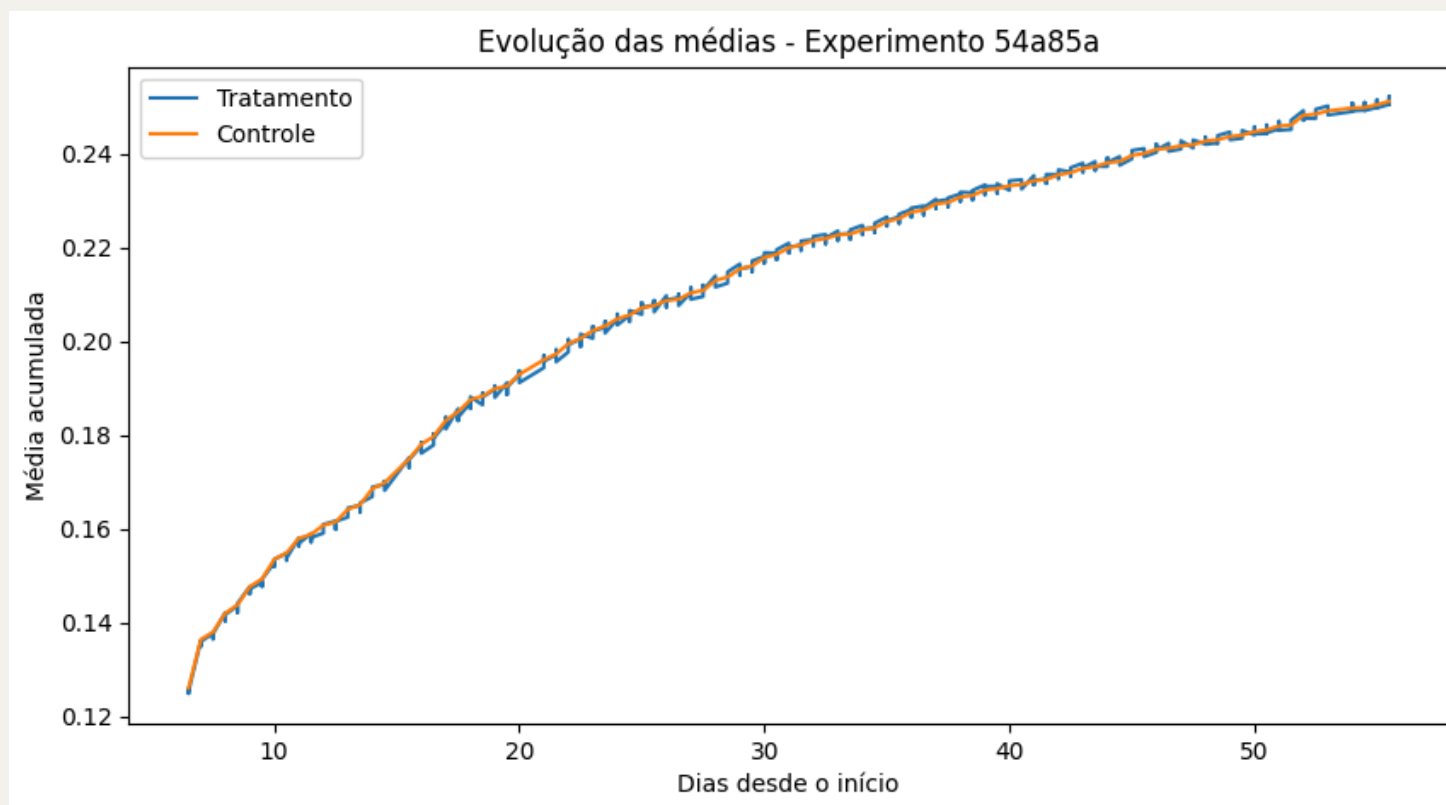


Gráfico de médias acumuladas por tempo de um determinado experimento

Estamos avaliando se há diferença estatisticamente significativa na taxa média de conversão entre usuários do grupo controle e do grupo tratamento.

$$\text{*efeito} = \mu_t - \mu_c$$

Segui nesta abordagem pois não temos acesso a uma maior granularidade dos dados, apenas resumos estatísticos por grupo no dataset.

Dataset: experimentos reais, métricas anonimizadas


	experiment_id	variant_id	metric_id	time_since_start	count_c	count_t	mean_c	mean_t	variance_c	variance_t
0	036afc	2	1	1.5	188065.0	186686.0	0.107808	0.107828	0.096186	0.096201
1	036afc	2	1	2.0	245041.0	243694.0	0.131790	0.131435	0.114422	0.114160
2	036afc	2	1	2.5	277237.0	275949.0	0.143065	0.142711	0.122598	0.122345
3	036afc	2	1	3.0	315689.0	314676.0	0.161789	0.160997	0.135613	0.135077
4	036afc	2	1	3.5	338631.0	337715.0	0.172474	0.171067	0.142727	0.141803
...
24148	fdaf62	1	4	28.0	2182559.0	2180705.0	38.270483	38.858464	8478.894092	8708.839083
24149	fdaf62	1	4	29.0	2259937.0	2257899.0	38.485748	39.065107	8627.264583	8821.982824
24150	fdaf62	1	4	30.0	2341537.0	2339309.0	38.691410	39.211843	8916.435308	8848.559719
24151	fdaf62	1	4	31.0	2422152.0	2419745.0	38.705264	39.263485	8836.078773	8884.955277
24152	fdaf62	1	4	32.0	2478580.0	2475916.0	38.962364	39.554518	9077.322251	9215.231273

24153 rows × 10 columns

O dataset representa resultados agregados de 78 experimentos A/B de uma grande empresa de e-commerce

Cada linha contém estatísticas cumulativas por experimento, variante e métrica, durante todo o processo.

Para este projeto, assumimos que a Métrica 1 representa taxa de conversão (binária) – uma métrica comum para negócios online, por exemplo:

- “Usuário comprou?”
 - “Usuário clicou no botão?”
 - “Usuário concluiu o cadastro?”
- 

O resultado final é a média dessas conversões em cada grupo.

Os experimentos: Evolução da métrica ao longo do tempo

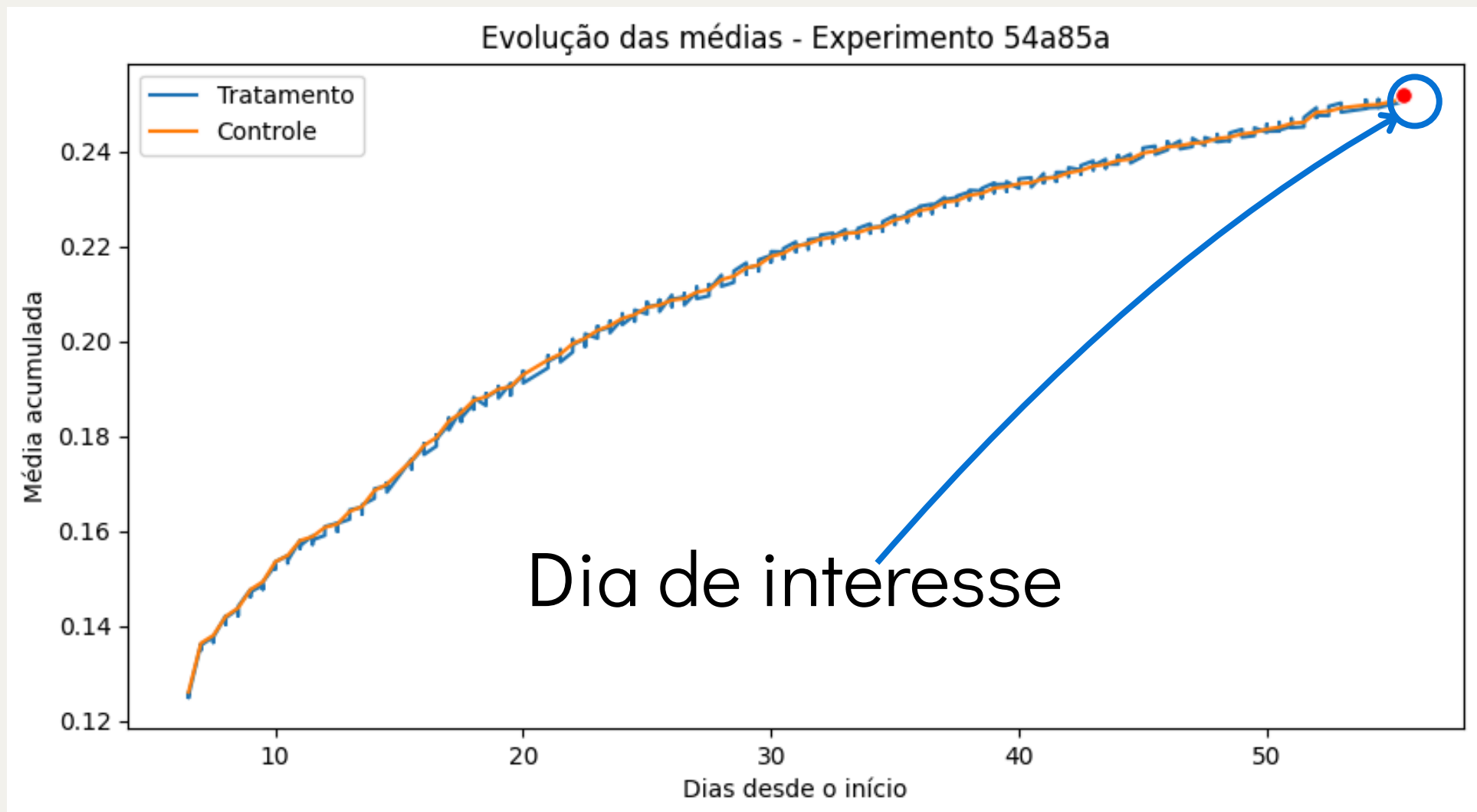


Gráfico de médias acumulada por tempo de um determinado experimento

Nossa análise se concentra no último dia disponível para cada experimento, com base na suposição de que o experimento foi encerrado naquele ponto e decisões seriam tomadas com esses dados

Hipóteses estatísticas

Dessa forma, podemos formalizar nossas hipóteses da seguinte forma:

Hipótese nula (H_0): Não há diferença entre o grupo de tratamento e o de controle.

Hipótese alternativa (H_1): Há uma diferença estatisticamente significativa entre os grupos.

Em outras palavras:

- H_0 : “A nova versão não mudou nada. O comportamento dos usuários é o mesmo com ou sem ela.”
- H_1 : “A nova versão provocou uma mudança real no comportamento dos usuários – para melhor ou pior.”

queremos nega-la

Utilizamos o teste Z para diferença entre médias (com variâncias conhecidas) e construímos intervalos de confiança de 95% para cada experimento.

Estatísticas aplicadas a um experimento

Exemplificando com um experimento isolado

```
new_df.iloc[0] # ultimo timestamp, mais tempo desde o inicio, mais dados
✓ 0.0s

experiment_id      54a85a
variant_id         1
metric_id          1
time_since_start   55.5
count_c            3320536.0
count_t            3316730.0
mean_c             0.251191
mean_t             0.252204
variance_c         0.188094
variance_t         0.188597
Name: 8459, dtype: object
```

Para aplicar o teste estatístico, antes precisamos do Standard Error (erro padrão)

Com os dados em mãos, aplicamos:

$$SE = \sqrt{\frac{s_c^2}{n_c} + \frac{s_t^2}{c_t}}$$

```
SE = np.sqrt((variance_c/count_c)+(variance_t/count_t))
print(f'Standard Error: {SE}')
✓ 0.1s

Standard Error: 0.0003369097700431301
```


Estatísticas aplicadas a um experimento

Aplicamos o Zscore

$$Z_{score} = \frac{\bar{x}_t - \bar{x}_c}{SE}$$

```
Z = (mean_t - mean_c)/SE
print(f'Z score: {Z}')
✓ 0.0s
Z score: 3.006906995504404
```

O Z-score quantifica o quão diferente foi o tratamento do controle. Indicando que a diferença está por volta de 3 desvios padrão da média esperado sob a H_0

Estatísticas aplicadas a um experimento

Por fim, o p-valor:

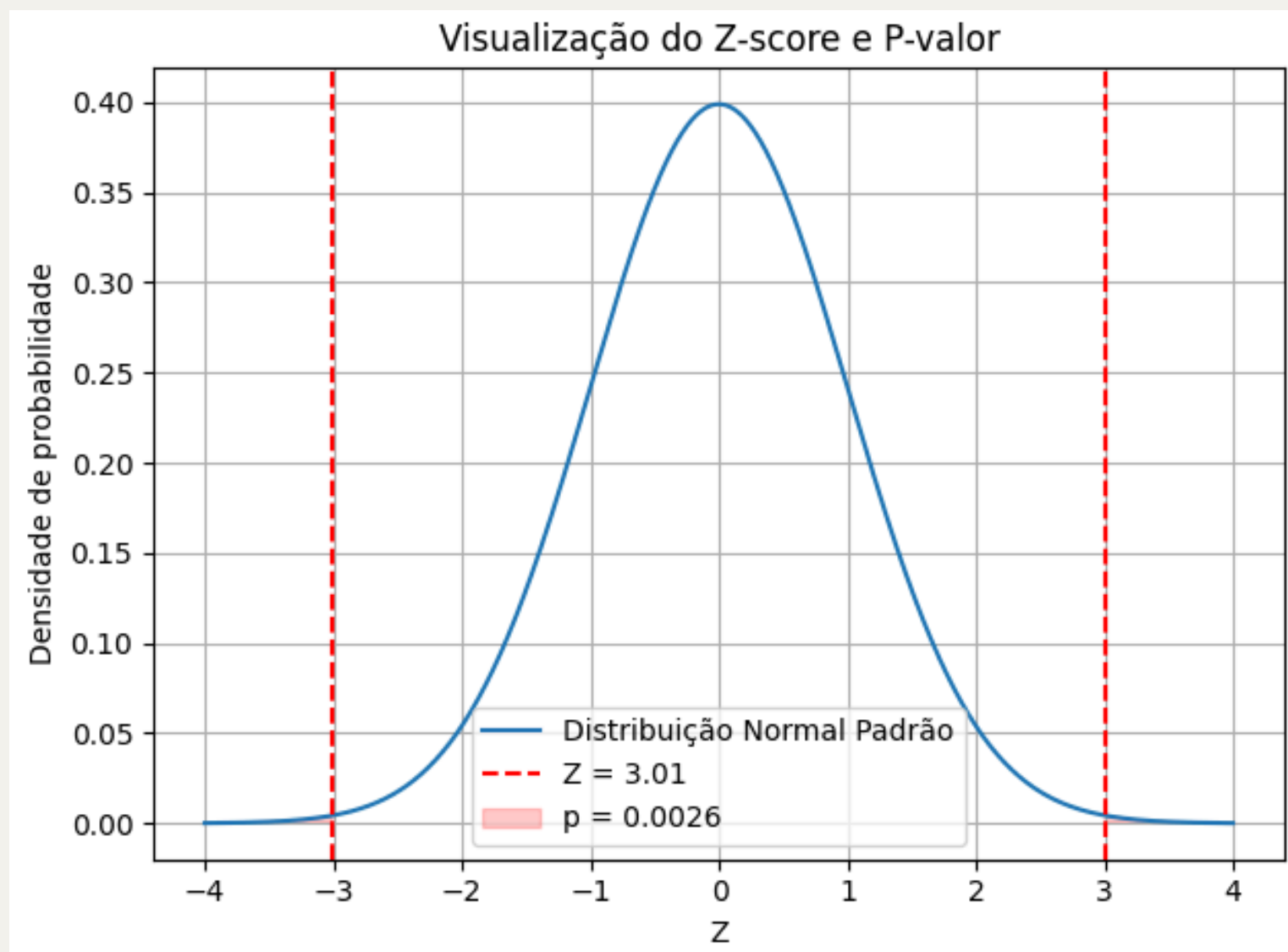
```
p_valor = 2 * (1 - norm.cdf(abs(Z)))  
print(f'P-valor: {p_valor}')
```

✓ 0.05

P-valor: 0.0026392049550985064

~~H_0~~

H_a ✓



O p-valor de 0.0026 é muito menor que 0.05, então podemos rejeitar a hipótese nula

Logo, a nova versão provocou uma alteração no comportamento dos usuários

Avaliando todos os experimentos

```
report_df.sample(20).sort_values('P-Valor', ascending= True)
```

✓ 0.0s

Python

	Experiment ID	Métrica	Tempo Final (dias)	Taxa Controle	Taxa Tratamento	Dif. Absoluta	Dif. Relativa (%)	IC 95% (Inf)	IC 95% (Sup)	Z-Score	P-Valor	Significativo	interpretação
19	4509ec	1	28.5	0.0569	0.0559	-0.0009	-1.67%	-0.0011	-0.0008	-15.5230	0.000000	Yes	Redução significativa na métrica.
9	2c8a04	1	131.0	0.0630	0.0642	0.0012	1.94%	0.0011	0.0013	28.5835	0.000000	Yes	Aumento significativo na métrica.
69	e4c4a1	1	36.0	0.1856	0.1841	-0.0016	-0.84%	-0.0022	-0.0009	-4.8412	0.000001	Yes	Redução significativa na métrica.
25	54a85a	1	55.5	0.2512	0.2522	0.0010	0.40%	0.0004	0.0017	3.0069	0.002639	Yes	Aumento significativo na métrica.
55	c3d89d	1	39.0	0.0493	0.0490	-0.0002	-0.50%	-0.0005	0.0000	-1.8820	0.059838	No	Inconclusivo: sem diferença estatística detect...
40	8b436e	1	45.5	0.0468	0.0469	0.0001	0.16%	-0.0000	0.0002	1.4739	0.140514	No	Inconclusivo: sem diferença estatística detect...
1	058875	1	21.5	0.0456	0.0457	0.0001	0.22%	-0.0000	0.0002	1.4601	0.144249	No	Inconclusivo: sem diferença estatística detect...
60	d0b910	1	32.5	0.2058	0.2054	-0.0003	-0.16%	-0.0008	0.0002	-1.3102	0.190113	No	Inconclusivo: sem diferença estatística detect...

Assumindo a que todos seguem as mesmas premissas do exemplo anterior.

Modularizei os testes criando uma função em python, que gera, além das métricas padrões do dataset, um reporte mais detalhado. Incluindo:

- Diferença (efeito) absoluto e relativo
- Taxa (media) de controle e tratamento
- Limites dos intervalos de confiança
- Z-score
- P-valor
- avaliação do p-valor
- Uma breve interpretação

Avaliando todos os experimentos

```
report_df.loc[report_df['Experiment ID'] == '4db6c7']
```

✓

0.0s

Python

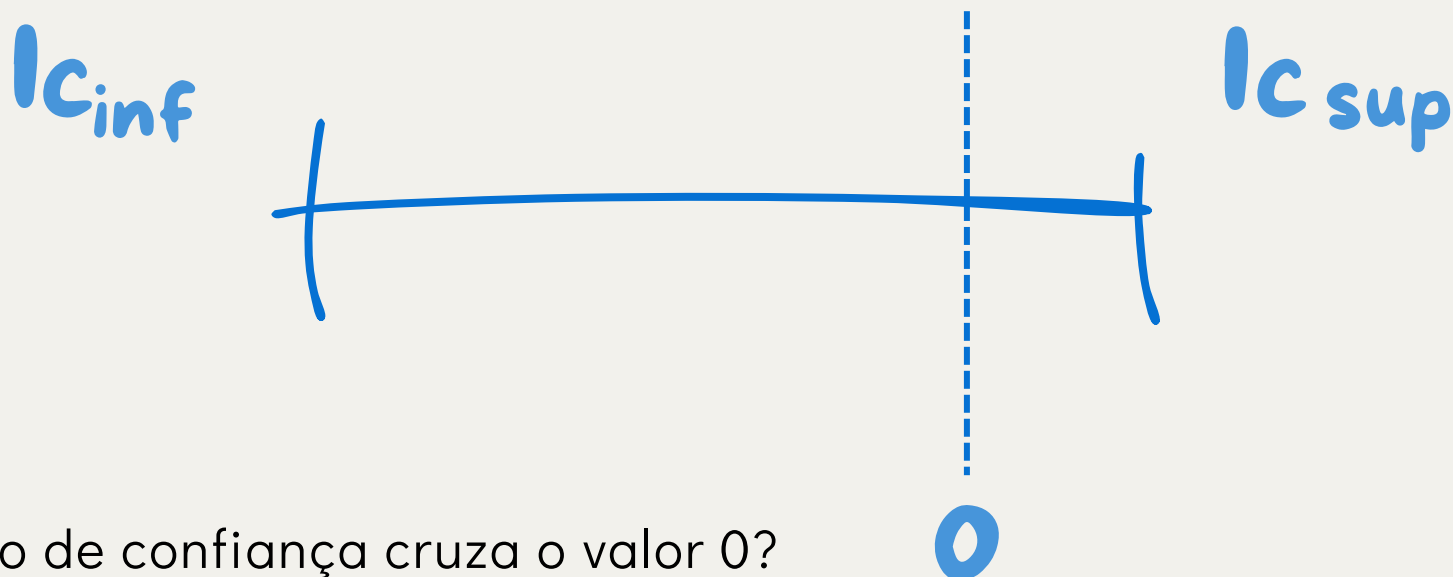
	Experiment ID	Métrica	Tempo Final (dias)	Taxa Controle	Taxa Tratamento	Dif. Absoluta	Dif. Relativa (%)	IC 95% (Inf)	IC 95% (Sup)	Z-Score	P-Valor	Significativo	interpretação
22	4db6c7	1	32.0	0.495	0.4935	-0.0016	-0.31%	-0.0032	0.0001	-1.8732	0.061037	No	Inconclusivo: sem diferença estatística detect...

Exemplo, no experimento '4db6c7', foi utilizado:

- Métrica 1 (assumimos como conversão)
- Experimento esteve no ar por 32 dias
- A diferença (efeito) relativo foi de -0.31%
- Zscore igual a -1.87 e p-valor de 0.0610
- Resultado inconclusivo, sem diferença estatística.

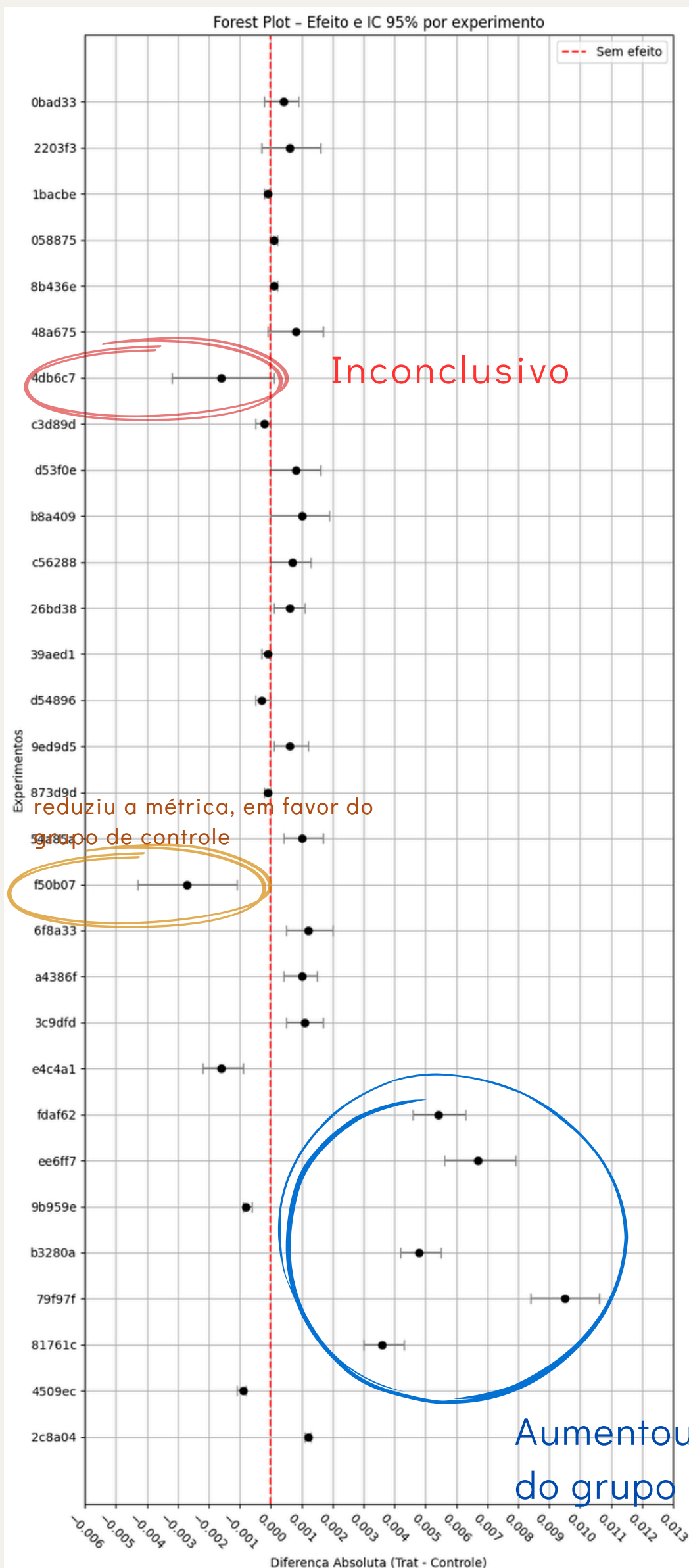
Logo não podemos rejeitar a hipótese nula

Inconclusivo, pois o p-valor > 0.05 e os intervalos de confiança cruzam o valor 0



Intervalo de confiança cruza o valor 0?

Avaliando todos os experimentos



Se esse intervalo inclui o valor 0, então:

- Pode ser que o tratamento aumente a métrica
- Pode ser que o tratamento reduza a métrica
- Pode ser que não haja efeito algum

Ou seja: o resultado é inconclusivo do ponto de vista estatístico.

Conclusão

Nem todo teste gera valor visível. E tudo bem. Saber aceitar a insignificância estatística de forma madura é parte essencial de uma cultura de experimentação saudável.

Um p-valor > 0.05 não significa que o experimento “deu errado” – significa que não houve evidência suficiente para afirmar uma mudança real.

Um leve aumento ou redução na conversão pode ser aceitável se houver ganhos em outros aspectos – como custo, escalabilidade, ou tempo de manutenção

Um efeito de $+0.01\%$ com p-valor < 0.001 pode ser estatisticamente relevante, mas irrelevante do ponto de vista de negócio. É necessário contextualizar os resultados com impacto real (número de usuários, faturamento, esforço técnico).

Dúvidas, sugestões ou críticas?

Comente!



Reyso Teixeira
Engenharia da Computação/UFPA



[linkedin.com/in/reyso-teixeira/](https://www.linkedin.com/in/reyso-teixeira/)



<https://github.com/reyso>



https://reyso.github.io/portifolio_projetos/