

A/B Testing en el mundo digital – Día 1

SKYSCANNER

Jose Parreño García

Curso 2022

Fecha 31/03/2022

AGENDA DÍA 1

- Romper el hielo: Me presento ☺
- Parte 1. A/B testing en el mundo digital.
- Parte 2. Tomando decisiones en vuestro primer experimento.
- Parte 3. Diseño de ejecución de un experimento y toma de decisiones
 - Repaso de la parte 3
- Parte 4. Consideraciones prácticas de implementación
 - Repaso de la parte 4



Me presento 😊

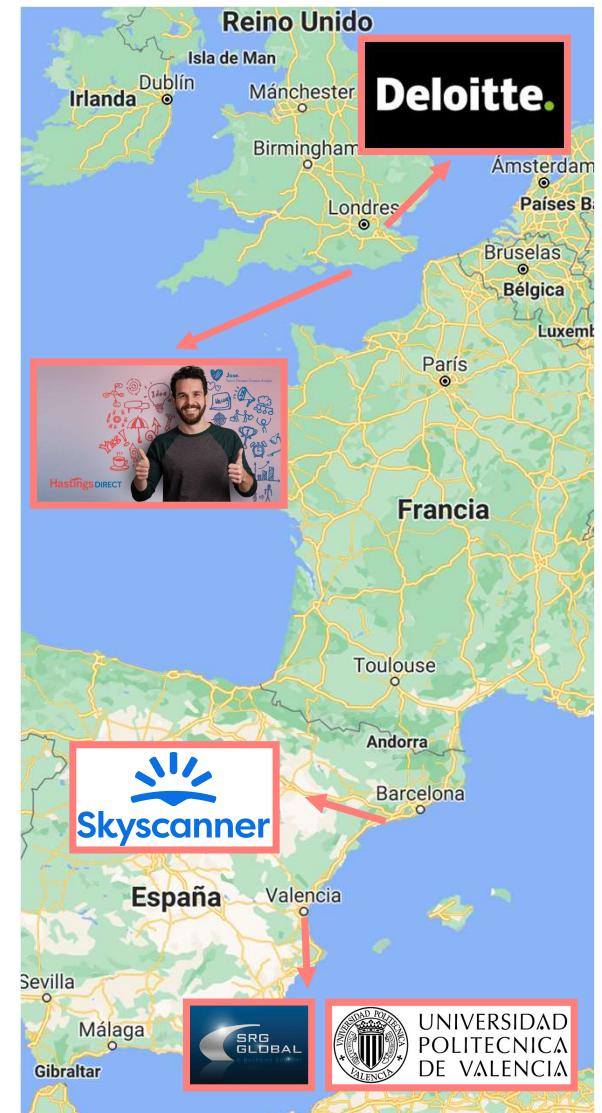
JOSE PARREÑO GARCIA

- Ingeniería Industrial – especialidad en robótica
- Prácticas en SRG – análisis estadístico de defectos en piezas y optimización logística.
- Deloitte UK – analista de datos en proyectos de crimen financiero
 - DieselGate de VW y blanqueo de capitales en HSBC México
- Hastings Direct – de junior a senior manager del equipo de Data Science.
 - Crecimiento de equipo de 3 a 10 DS + montar un equipo nuevo de DE
 - Optimización de precios para la venta en comparadoras de seguros
- Skyscanner – Data Science Manager
 - (1) Equipo de personalización y recomendaciones.
 - Proyecto de optimización de 'BEST' ranking para diferentes verticales.
 - (2) Equipo de marketing

En Hastings Direct y en Skyscanner la aplicación de experimentación es clave.



www.linkedin.com/in/joseparrenogarcia





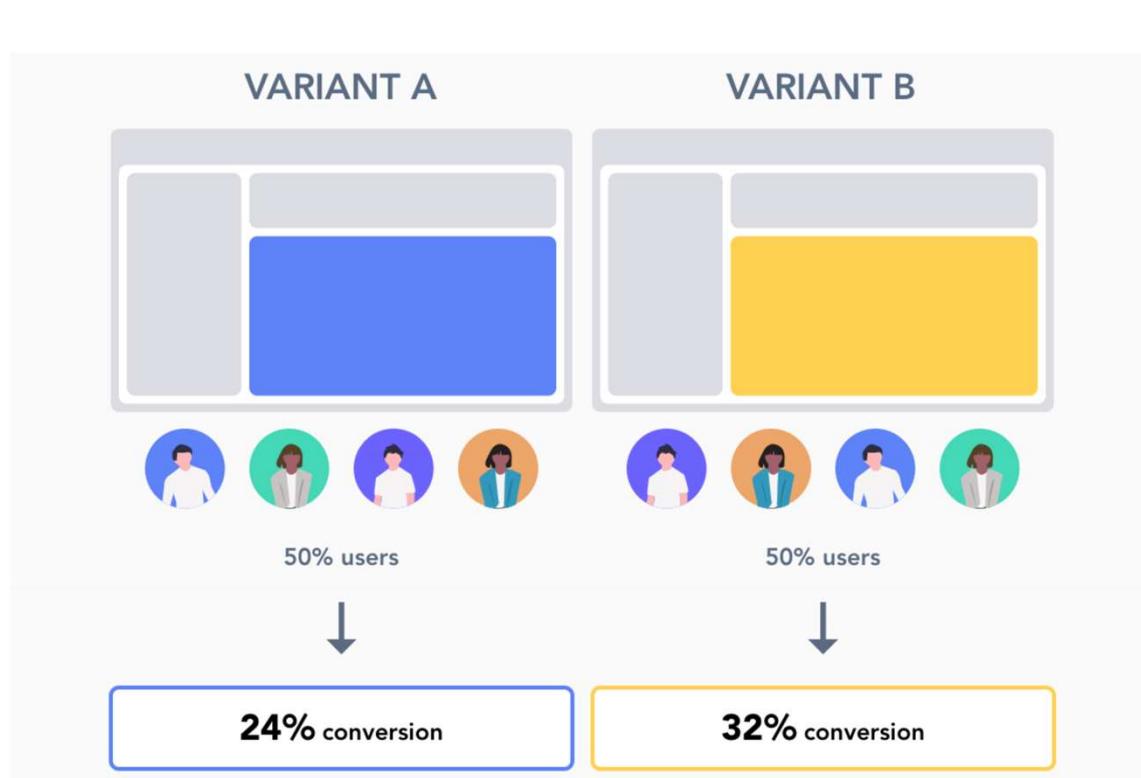
Parte 1. A/B testing en el mundo digital

PARTE 1. A/B TESTING EN EL MUNDO DIGITAL.

Objetivos:

- *Entender qué es el A/B Testing*
- *¿Por qué es importante para aquellos que quieran dedicarse al mundo del dato?*

¿QUÉ ES A/B TESTING EN EL MUNDO DIGITAL?



¿Mejora mi nueva idea respecto a la versión actual?

Se busca medir variaciones en nuestras métricas cuando realizamos cambios en el producto.

Control sobre la cantidad de cambios

Control sobre la exposición del experimento a nuestro usuarios.

Análisis estadístico y robusto de los resultados.

¿CÓMO DE COMÚN ES?

Las grandes empresas lo consideran una parte clave de su ecosistema



- +**7000** experimentos en 2011
- Se comenta que en 2020, casi todo el tráfico de Google era parte de un A/B test...



- +79% en conversión de donaciones
- +\$75 millones



- Experimentos lanzados:
- En menos de **una hora**
 - En **75 países**
 - **43 idiomas**



- Equipo de **300 personas** dedicadas a crear un motor de experimentación
- Coste: **150 millones**
- Beneficio: **500 millones**

99.9% QUE TENDRÉIS QUE HACERLO EN ROLES DE DATA

Experimentation is a major focus of Data Science across Netflix

 Netflix Technology Blog [Follow](#) 
Jan 11 · 18 min read



 Michael Barber
Jan 28, 2018 · 20 min read ★

Data science you need to know! A/B testing

This is part 2 of a 5-part series of posts aiming to quickly introduce some core concepts in data science and data analysis, with a specific focus on areas that I feel are overlooked or treated briefly in other materials.

This post outlines A/B testing, and the steps necessary to plan and build your own robust A/B test.

This post will suit data scientists working in product development, and product managers hoping to communicate better with their data scientists.

<https://www.upwork.com> > ab-testing ▾ [Traducir esta página](#)

A/B Testing Jobs | Upwork™

Browse 74 open jobs and land a remote A/B Testing job today. See detailed job requirements, compensation, duration, employer history, & apply today.

<https://www.workingnomads.com> > ... ▾ [Traducir esta página](#)

Remote Ab Testing Jobs - Working Nomads

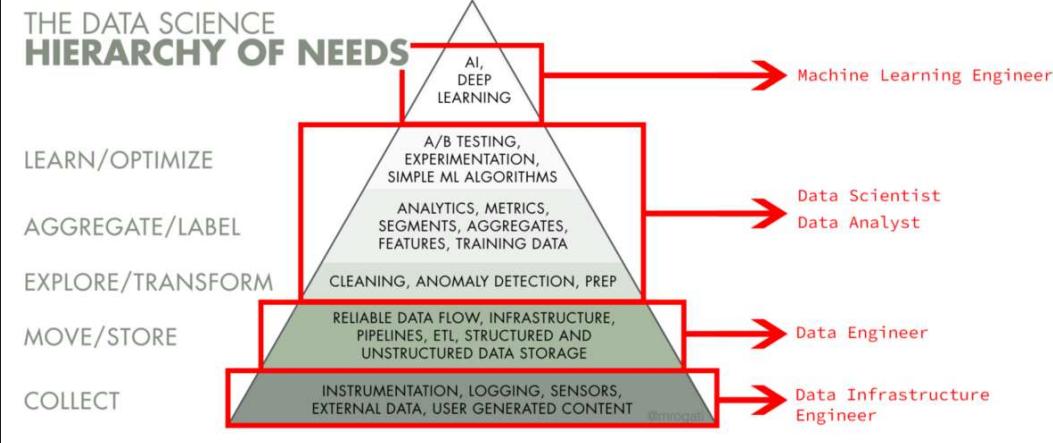
Remote Ab Testing Jobs · SEO Specialist · Digital & Automated Marketing Consultant · Senior Demand Generation Marketing Manager · Head of Demand Generation · Sr.

<https://productschool.com> > blog ▾ [Traducir esta página](#)

Product Management Skills: A/B Testing

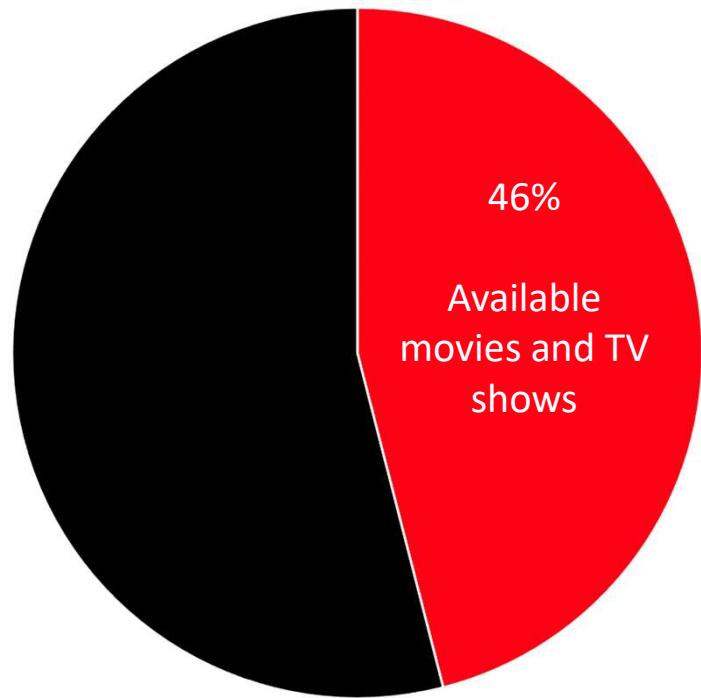
Here's everything product managers need to know about A/B testing, including how to plan, execute, and prioritize your test backlog.

18 mar 2021 · Subido por Product School



EJEMPLO DE NETFLIX

Netflix quería mejorar el ratio de conversión para nuevos usuarios



What **one thing** would you like to know more about before signing for Netflix?

EJEMPLO DE NETFLIX

Se pusieron a trabajar y crearon múltiples variantes de la landing page.

Landing page sin ninguna indicación de contenido



Landing page donde se añade el contenido más popular



EJEMPLO DE NETFLIX

Después de multiples experimentos...

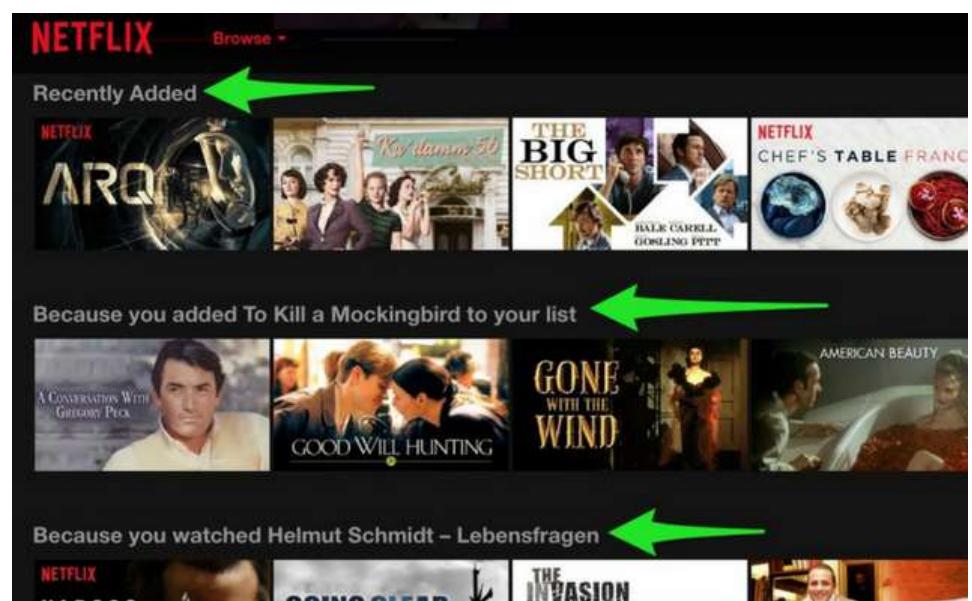
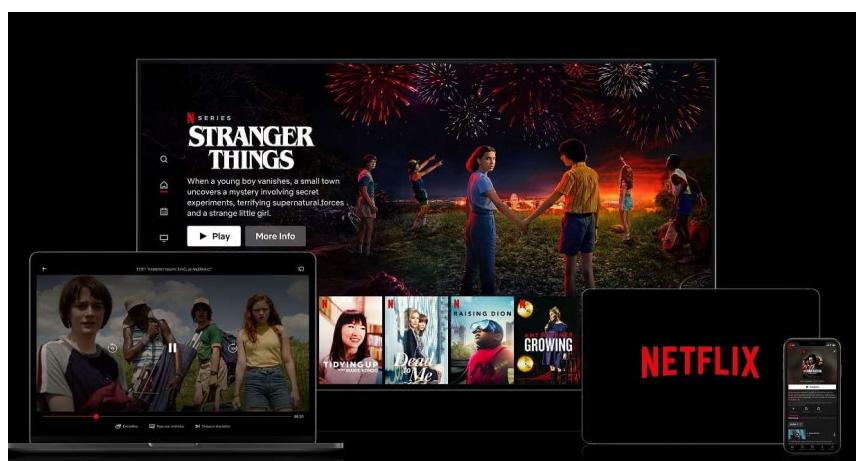
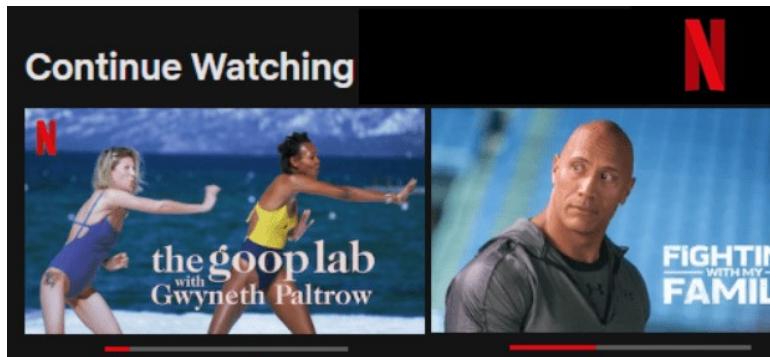
Landing page sin ninguna indicación de contenido



Landing page donde se añade el contenido más popular

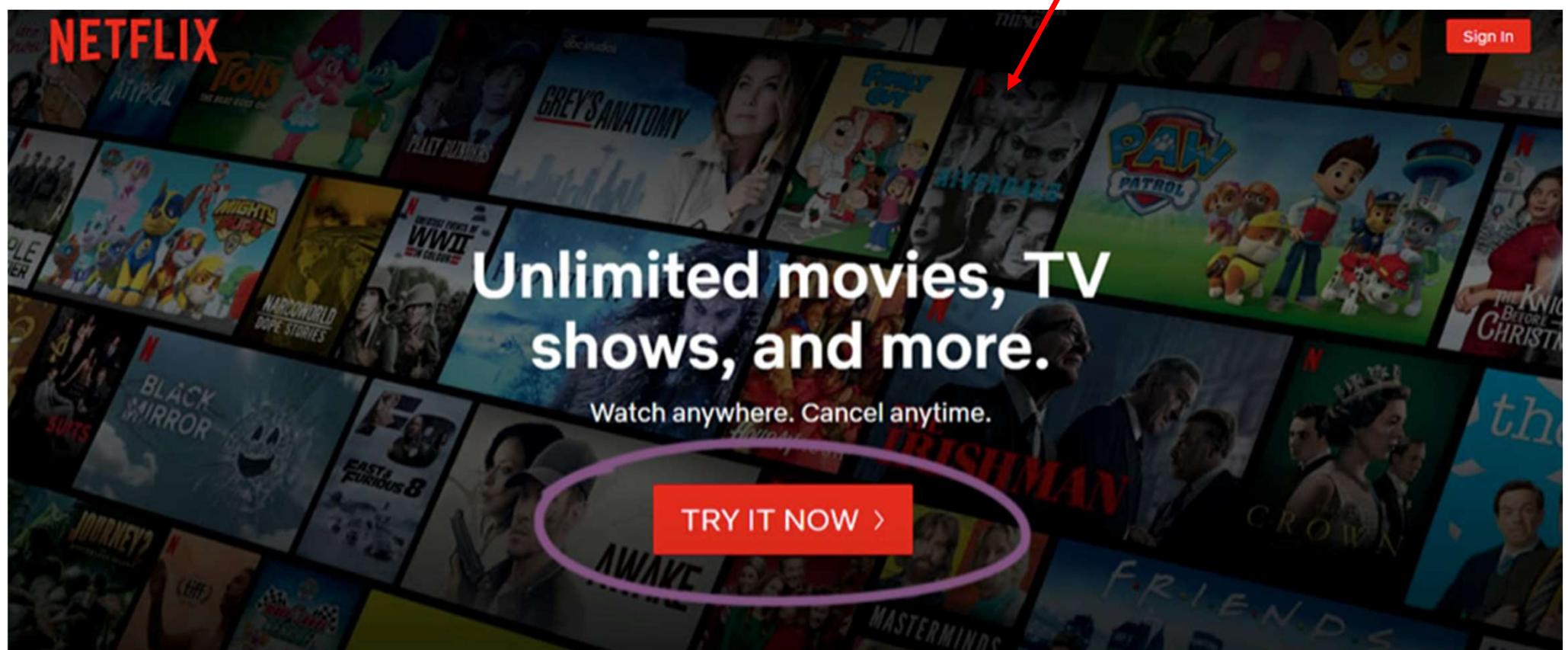


EJEMPLO DE NETFLIX



EJEMPLO DE NETFLIX

Página estática con muestra de contenido popular en la región del usuario.



A minimalist, modern concrete staircase with wide treads and a dark grey wall to its left. The stairs lead upwards towards a white ceiling. The overall aesthetic is clean and architectural.

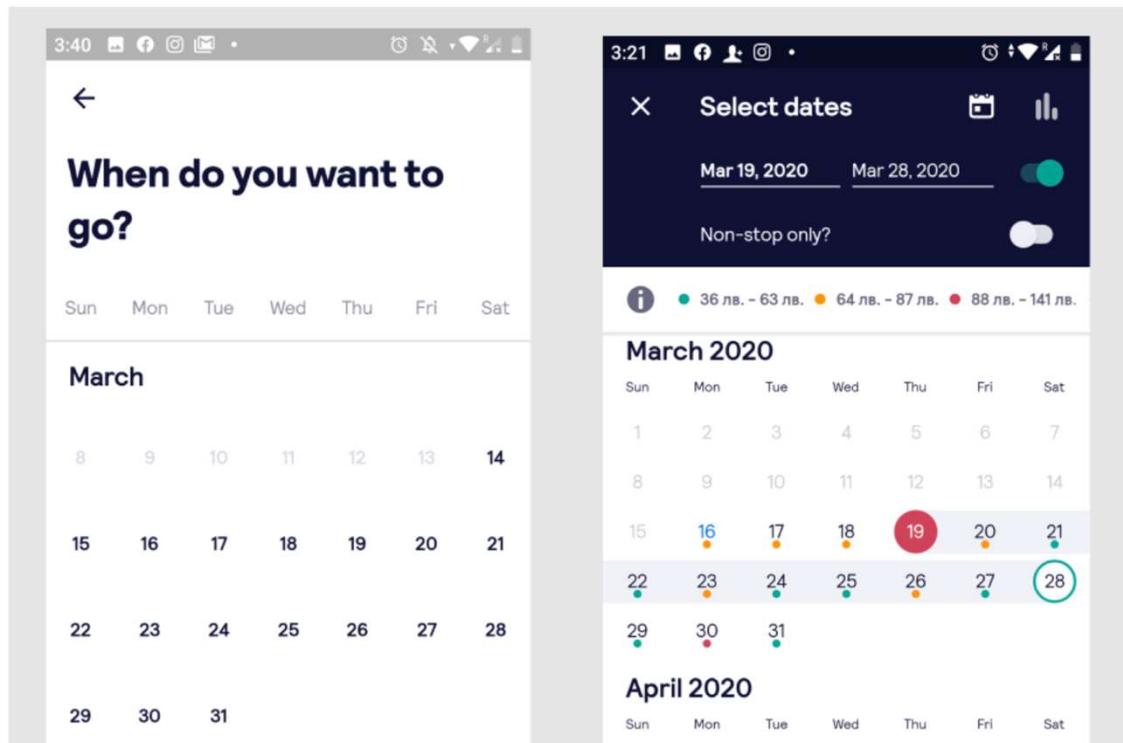
Parte 2. Tomando
decisiones en vuestra
primer experimento.

PARTE 2. TOMANDO DECISIONES EN VUESTRO PRIMER EXPERIMENTO.

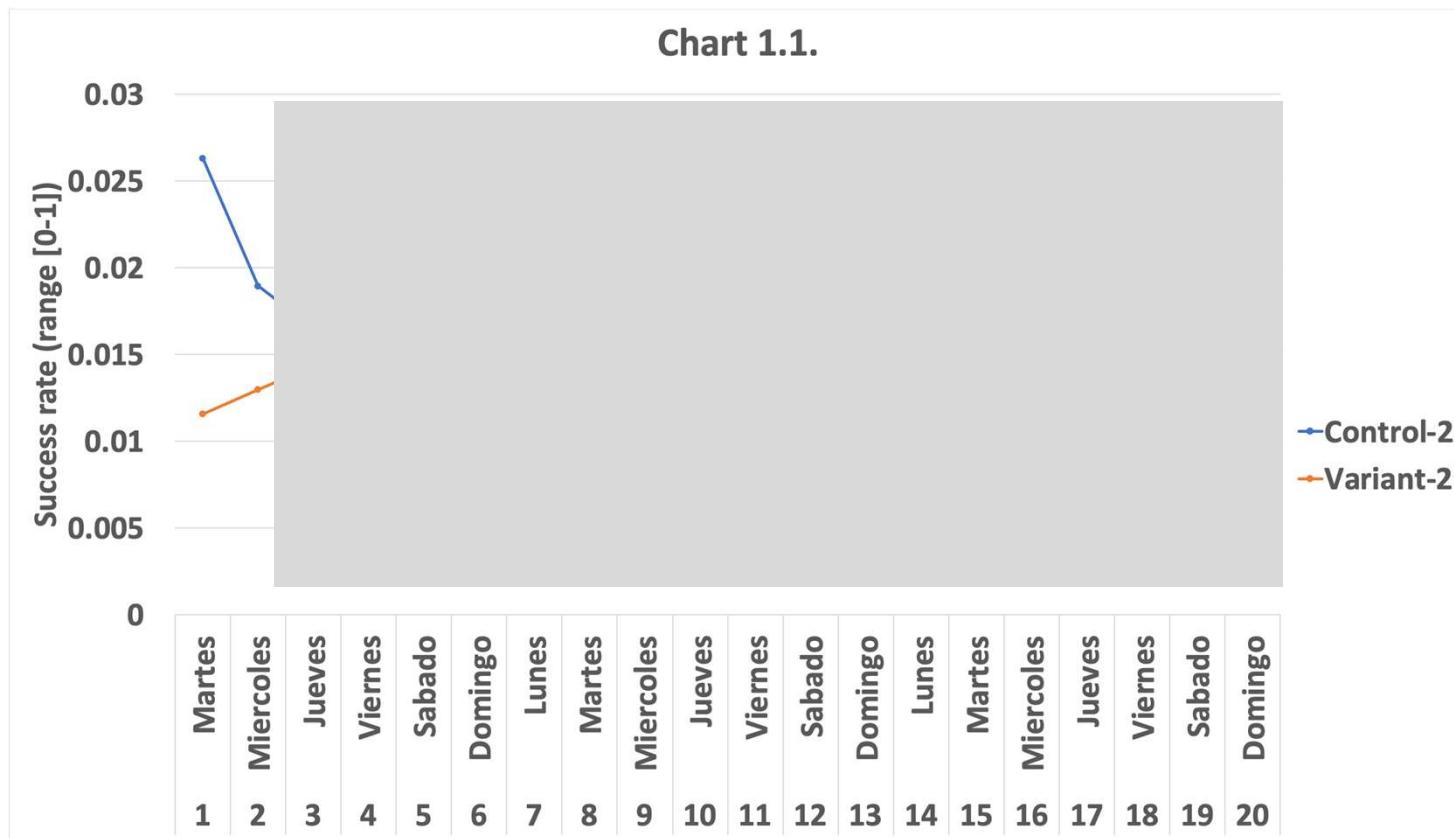
Objetivos:

- *Poneros a prueba con los resultados de un experimento ☺*

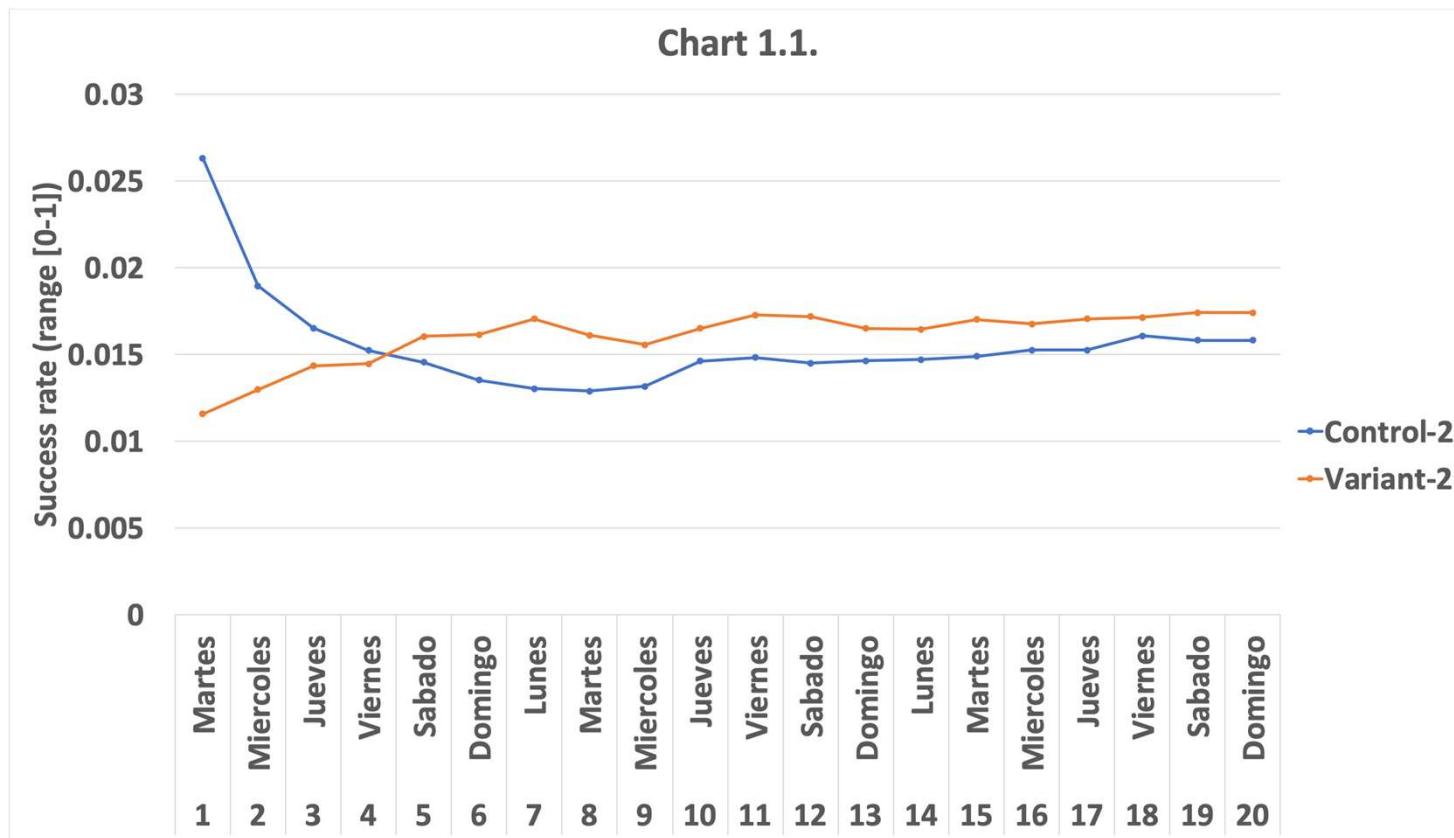
EJEMPLO DE SKYSCANNER



¿QUÉ DECISIÓN TOMARÍAIS?

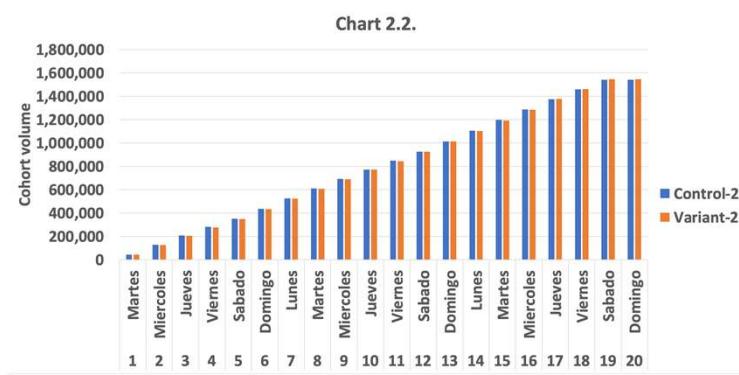
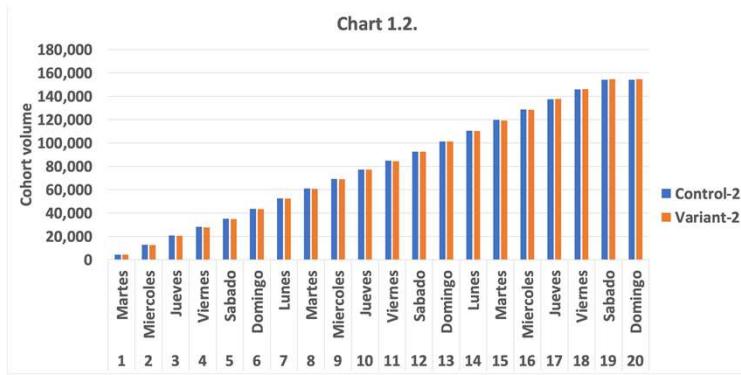
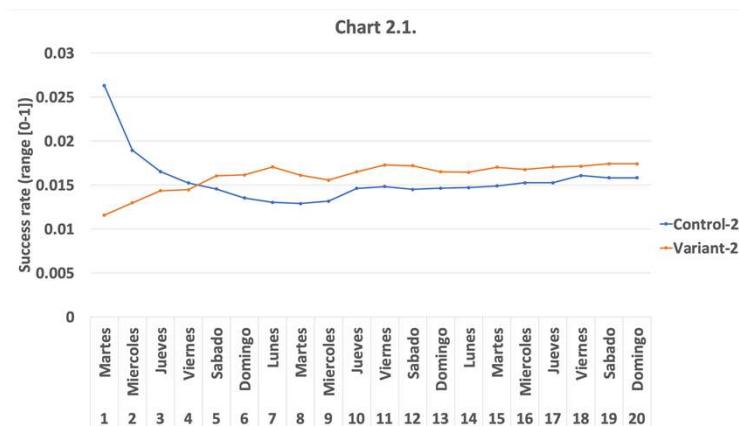
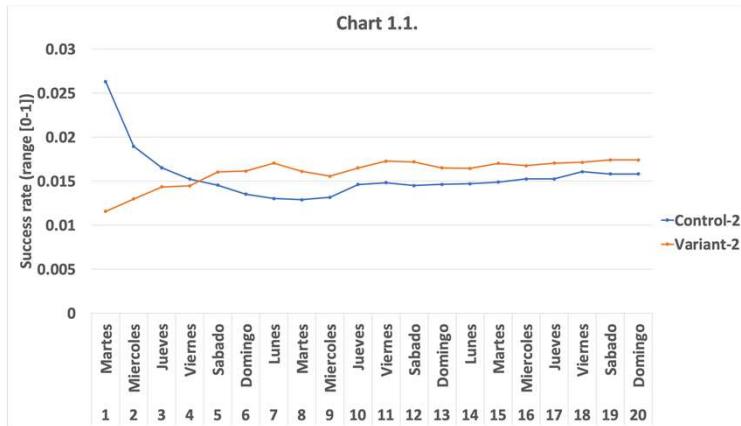


¿QUÉ DECISIÓN TOMARÍAIS?



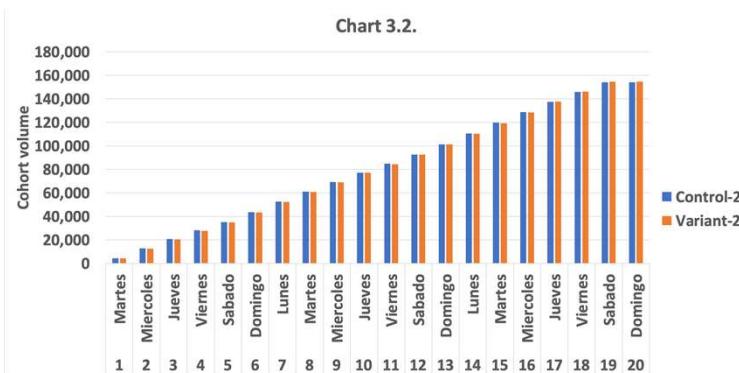
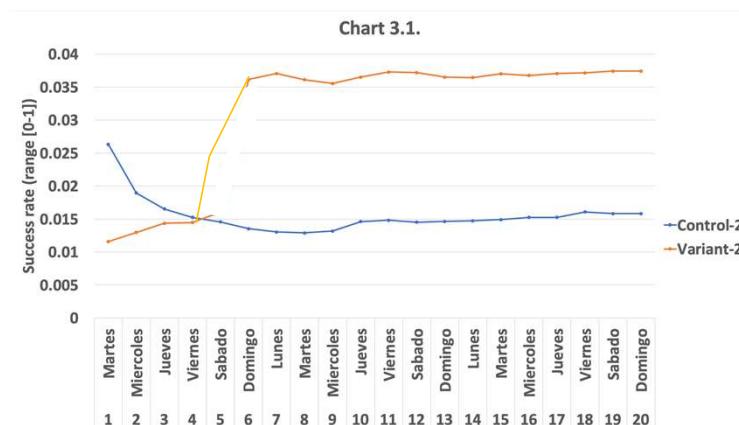
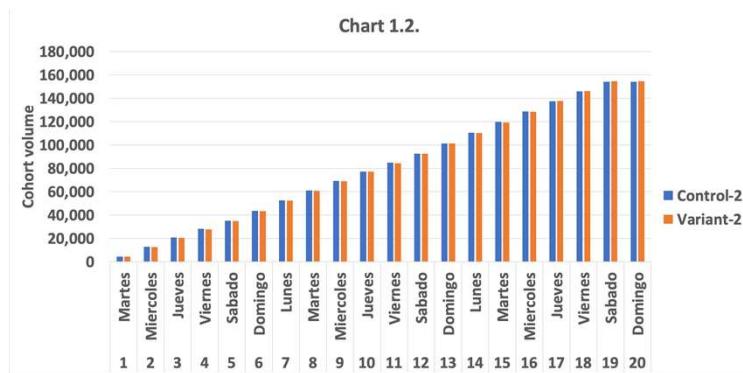
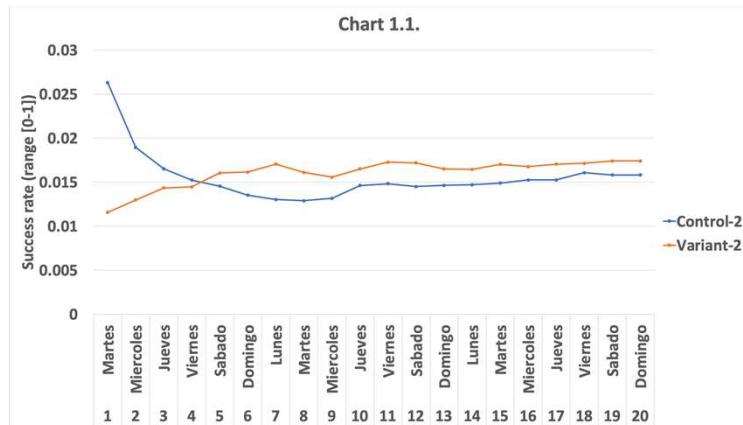
¿QUÉ DECISIÓN TOMARÁIS?

Concepto clave: tamaño de muestra



¿QUÉ DECISIÓN TOMARÁIS?

Concepto clave: magnitud del efecto a detectar



RESUMEN

- $(\text{Ratio de conversión de A}) - (\text{Ratio de conversión de B})$ no sirve
- Factores importantes:
 - tamaño de muestra => **¿Qué tamaño de muestra necesito?**
 - efecto a detectar => **¿Qué magnitud a detectar es adecuada?**
 - nivel de confianza en la toma de decisiones => **¿Qué probabilidad tengo de tomar una decisión equivocada?**

Parte 3. Diseño de ejecución de un experimento y toma de decisiones

PARTE 3. DISEÑO DE EJECUCIÓN DE UN EXPERIMENTO Y TOMA DE DECISIONES

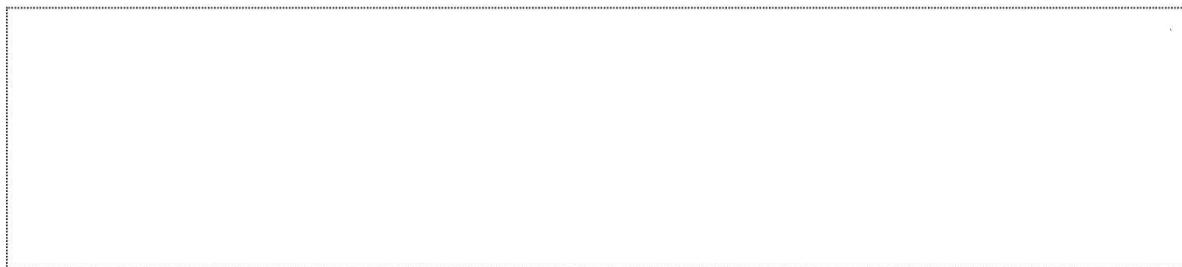
Objetivos:

- ¿Qué tamaño de muestra necesito?
- ¿Qué magnitud a detectar es adecuada?
- ¿Qué probabilidad tengo de tomar una decisión equivocada?
- Tengo los resultados de mi experimento, ¿qué decisión tomo?

- ¿Qué tamaño de muestra necesito?
- ¿Qué magnitud a detectar es adecuada?
- ¿Qué probabilidad tengo de tomar una decisión equivocada?
- Tengo los resultados de mi experimento, ¿qué decisión tomo?

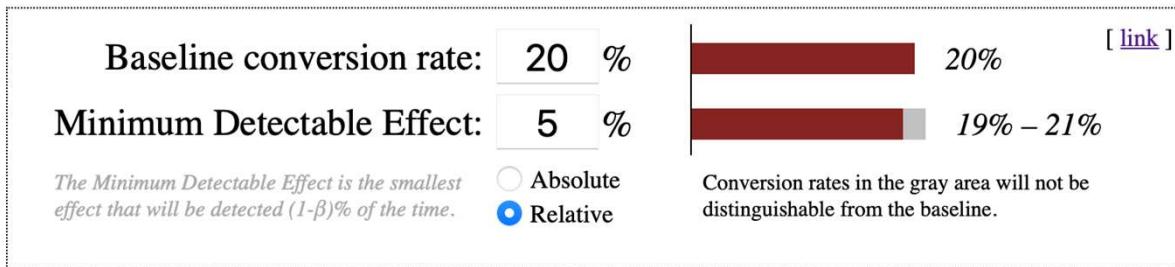
LA CALCULADORA MÁGICA

Question: How many subjects are needed for an A/B test?



¿CÓMO AFECTAN 'BCR' Y 'MDE' A N?

Question: How many subjects are needed for an A/B test?

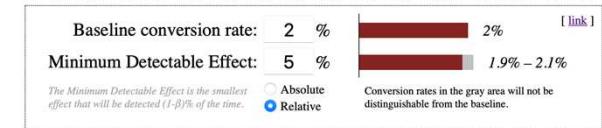


Sample size:

25,255

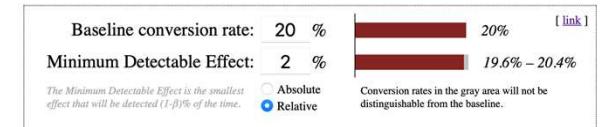
per variation

¿Qué pasa si nuestro ratio de conversión es de 2%?



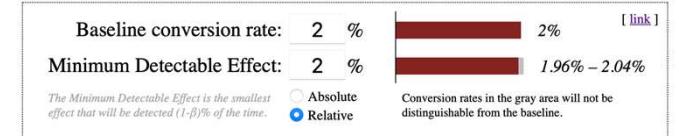
Sample size:
309,928
per variation

¿Qué pasa si queremos detectar un efecto del 2%?



Sample size:
157,328
per variation

¿Qué pasa si nuestro ratio de conversión es de 2% y queremos detectar un efecto del 2%?

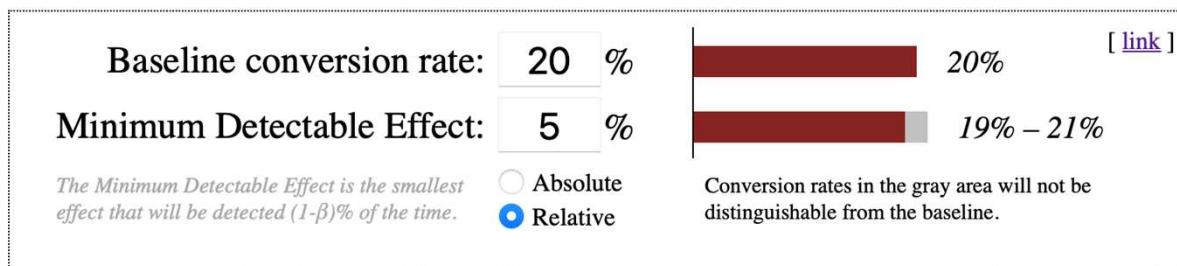


Sample size:
1,928,622
per variation

- ¿Qué tamaño de muestra necesito?
- ¿Qué magnitud a detectar es adecuada?
- **• ¿Qué probabilidad tengo de tomar una decisión equivocada?**
- Tengo los resultados de mi experimento, ¿qué decisión tomo?

ALPHA Y BETA

Question: How many subjects are needed for an A/B test?



Sample size:

25,255

per variation

Statistical power $1-\beta$:  80%

Significance level α :  5%

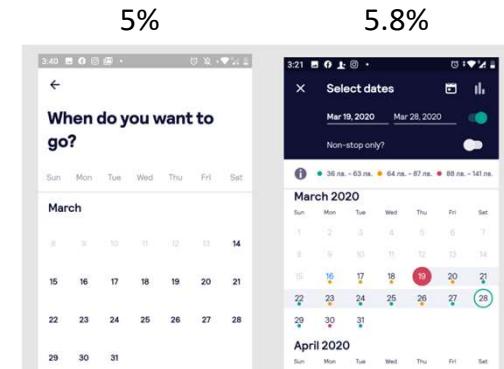
See also: [How Not To Run an A/B Test](#)



PRUEBAS DE HIPÓTESIS

Hipótesis nula vs alternativa

- Una prueba de hipótesis es una regla que especifica si se puede aceptar o rechazar una afirmación acerca de una población.
 - *Por lo general, se busca entender si hay una diferencia significativa entre A y B.*
- La hipótesis nula es el enunciado que se probará:
 - *Por lo general, asume que NO hay diferencias entre A y B*
- La hipótesis alternativa es el enunciado que se desea poder concluir que es verdadera



Queremos probar, de manera **estadísticamente robusta** que, la conversion de 5.8% de la nueva versión vs el 5% de la antigua , es un **efecto real**, y así decidir usarlo como nuestra nueva versión.

H_0 = mostrar el rango de precios en el calendario, NO tiene efecto en el ratio de conversión.

$$H_0: p_A = p_B$$

H_a = mostrar el rango de precios en el calendario, SI tiene efecto en el ratio de conversion (en cualquier dirección)

$$H_a: p_A \neq p_B$$

¿QUÉ NIVEL DE ERROR ESTAMOS DISPUESTOS A ACEPTAR?

- H_0 = mostrar el rango de precios en el calendario, NO tiene efecto en el ratio de conversion, $p_A = p_B$
- H_a = mostrar el rango de precios en el calendario, SI tiene efecto en el ratio de conversion (en cualquier dirección), $p_A \neq p_B$

Decisión que tomamos

Efecto real		
	Null hypothesis is false	Null hypothesis is true
Reject null hypothesis	Correct decision ($p = 1 - \beta$)	Type I error ($p = \alpha$)
Accept null hypothesis	Type II error ($p = \beta$)	Correct decision ($p = 1 - \alpha$)

Situación ideal: situación ideal.

- Hemos diseñado un nuestro H_0 para poder rechazarla posteriormente.

Creemos que Sí hay un efecto cuando NO lo hay.

- Evitar a toda costa
- Podríamos estar lanzando un producto que estuviera haciendo daño

Creemos que NO hay un efecto cuando Sí lo hay.

- Queremos evitarlo, pero no tan grave como el (I)
- Coste de oportunidad

EFFECTOS DE CAMBIAR ALPHA Y BETA

Question: How many subjects are needed for an A/B test?

Baseline conversion rate: 20 %



[[link](#)]

Minimum Detectable Effect: 5 %



The Minimum Detectable Effect is the smallest effect that will be detected ($1-\beta$)% of the time.

Absolute
 Relative

Conversion rates in the gray area will not be distinguishable from the baseline.

Sample size:

25,255

per variation

Statistical power $1-\beta$:

80% Percent of the time the minimum effect size will be detected, assuming it exists

Significance level α :

5% Percent of the time a difference will be detected, assuming one does NOT exist

See also: [How Not To Run an A/B Test](#)

		Null hypothesis is false	Null hypothesis is true
Reject null hypothesis	Correct decision ($p = 1 - \beta$)	Type I error ($p = \alpha$)	
	Type II error ($p = \beta$)	Correct decision ($p = 1 - \alpha$)	
Accept null hypothesis			

¿Qué pasa si queremos detectar un efecto del 5% relativo un 95% de las veces?

Sample size:
41,932

per variation

Statistical power $1-\beta$: 95% Percent of the time the minimum effect size will be detected, assuming it exists

Significance level α : 5% Percent of the time a difference will be detected, assuming one does NOT exist

¿Qué pasa si queremos si queremos fallar sólo un 1% de la veces?

Sample size:
37,542

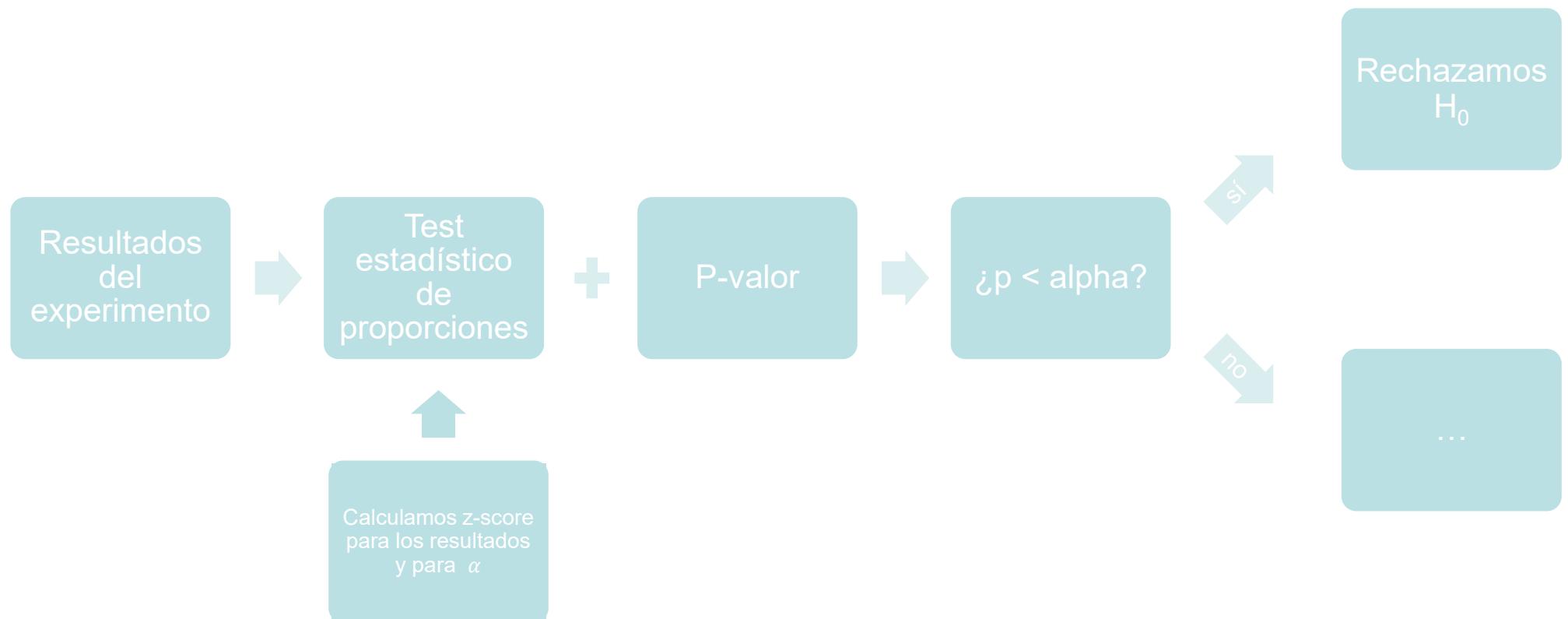
per variation

Statistical power $1-\beta$: 80% Percent of the time the minimum effect size will be detected, assuming it exists

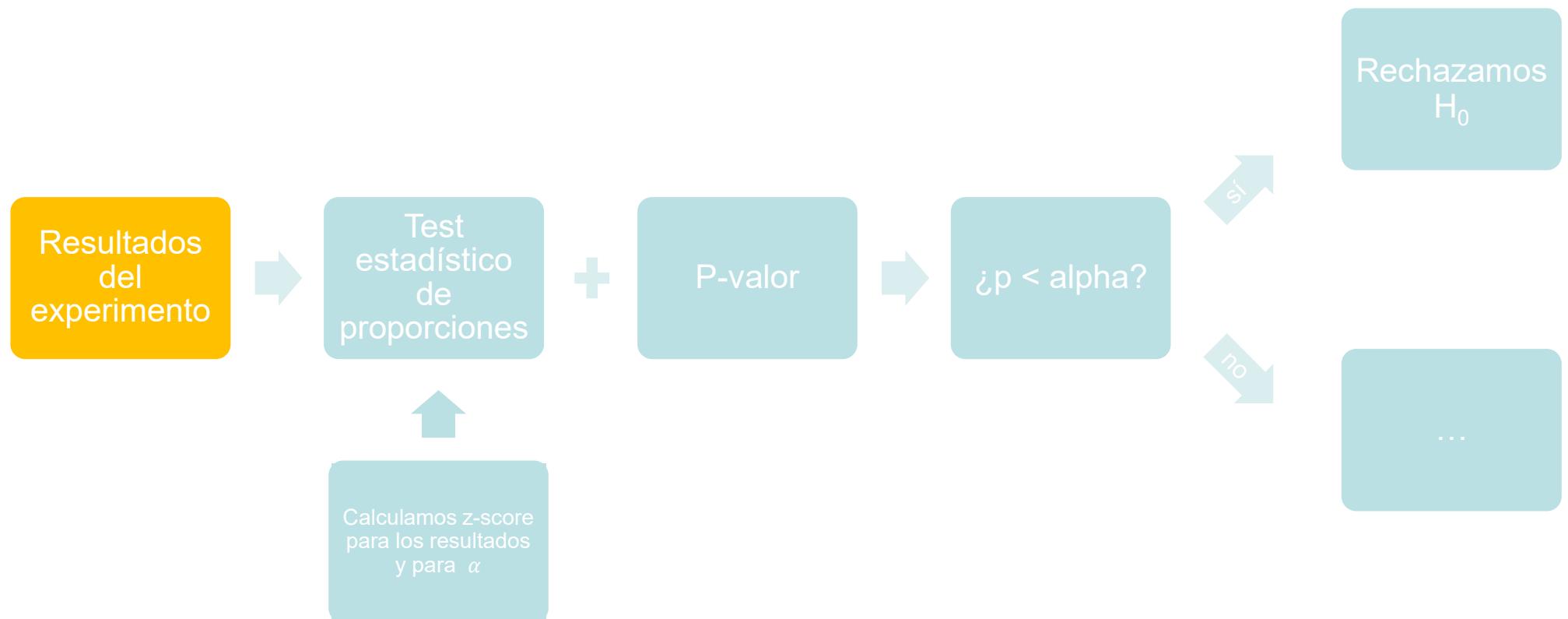
Significance level α : 1% Percent of the time a difference will be detected, assuming one does NOT exist

- ¿Qué tamaño de muestra necesito?
- ¿Qué magnitud a detectar es adecuada?
- ¿Qué probabilidad tengo de tomar una decisión equivocada?
- **Tengo los resultados de mi experimento, ¿qué decisión tomo?**

PASOS PARA LA TOMA DE DECISIONES



PASOS PARA LA TOMA DE DECISIONES



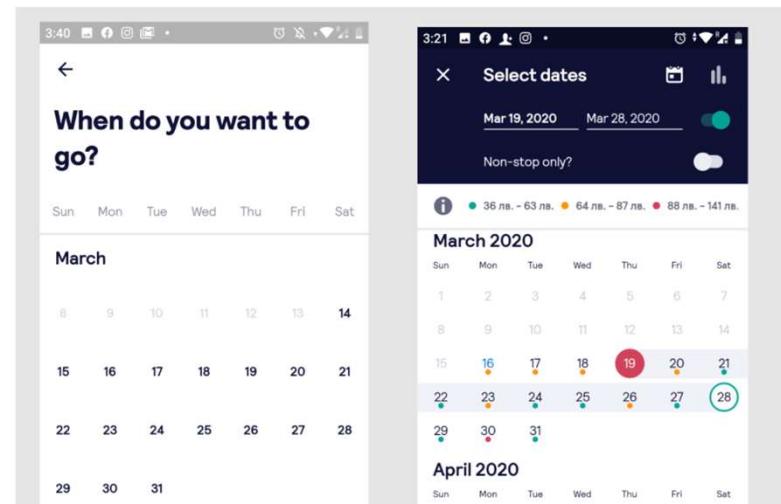
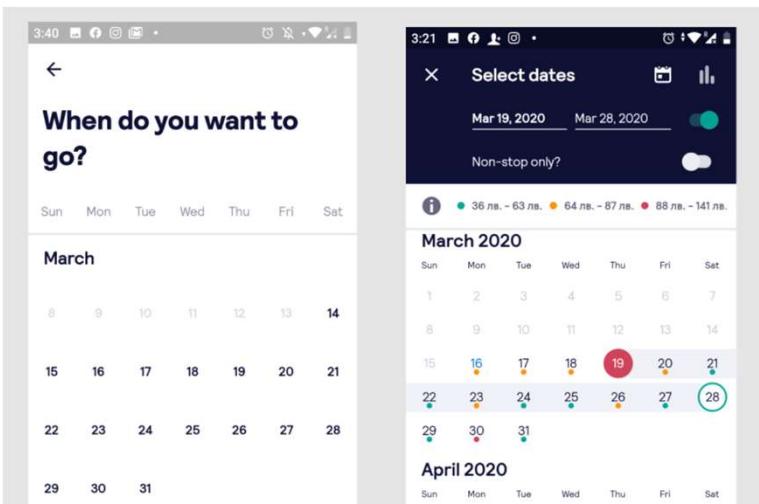
RESULTADOS DE NUESTRO EXPERIMENTO

Caso 1

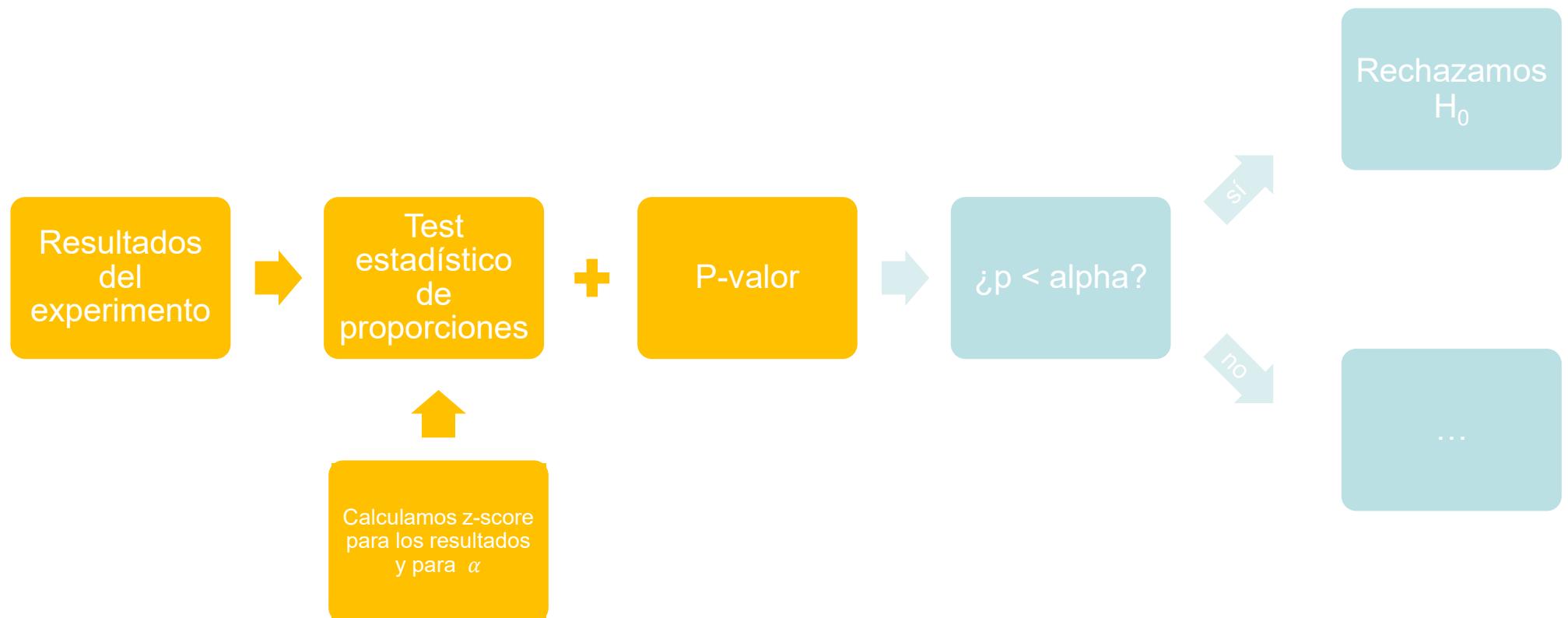
Tamaño de muestra	100,123
Conversiones	5,000
Tasa de conversión	4.99%

Caso 2

100,123	100,133
5,000	5,175
4.99%	5.17%

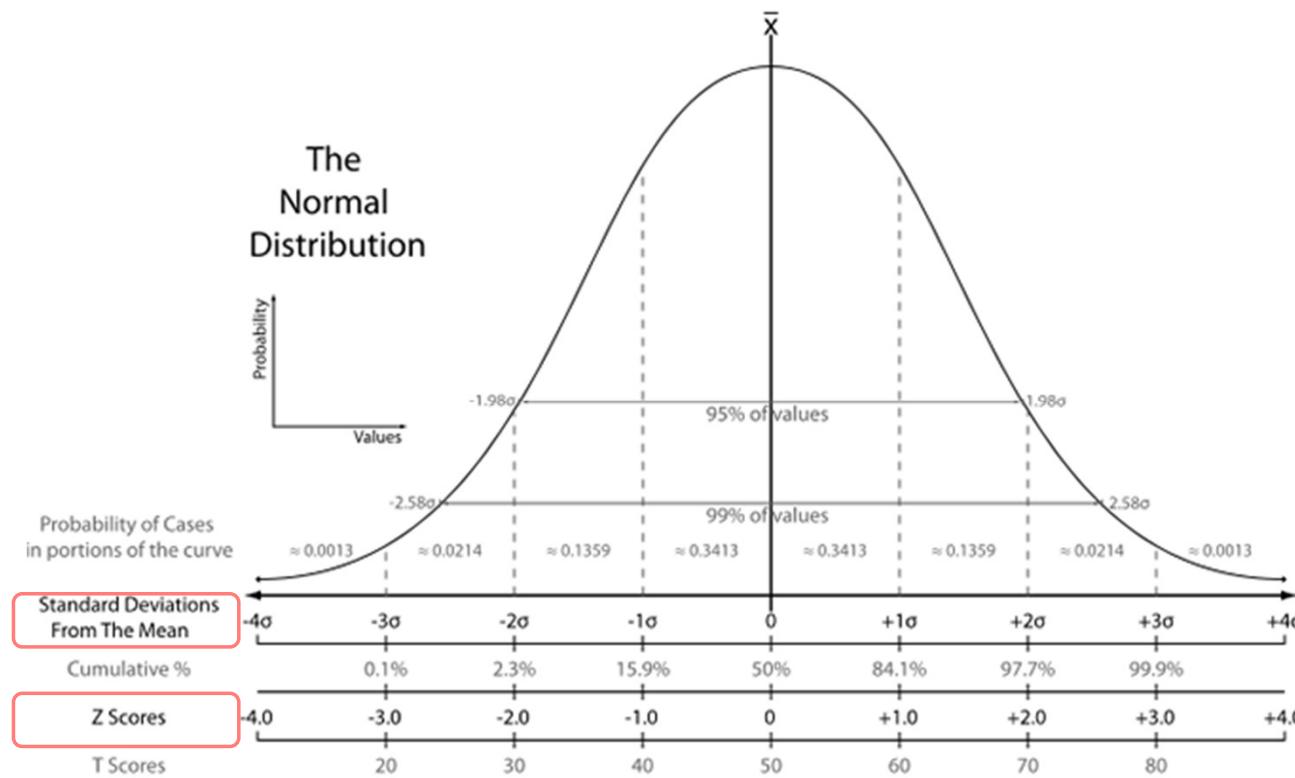


PASOS PARA LA TOMA DE DECISIONES



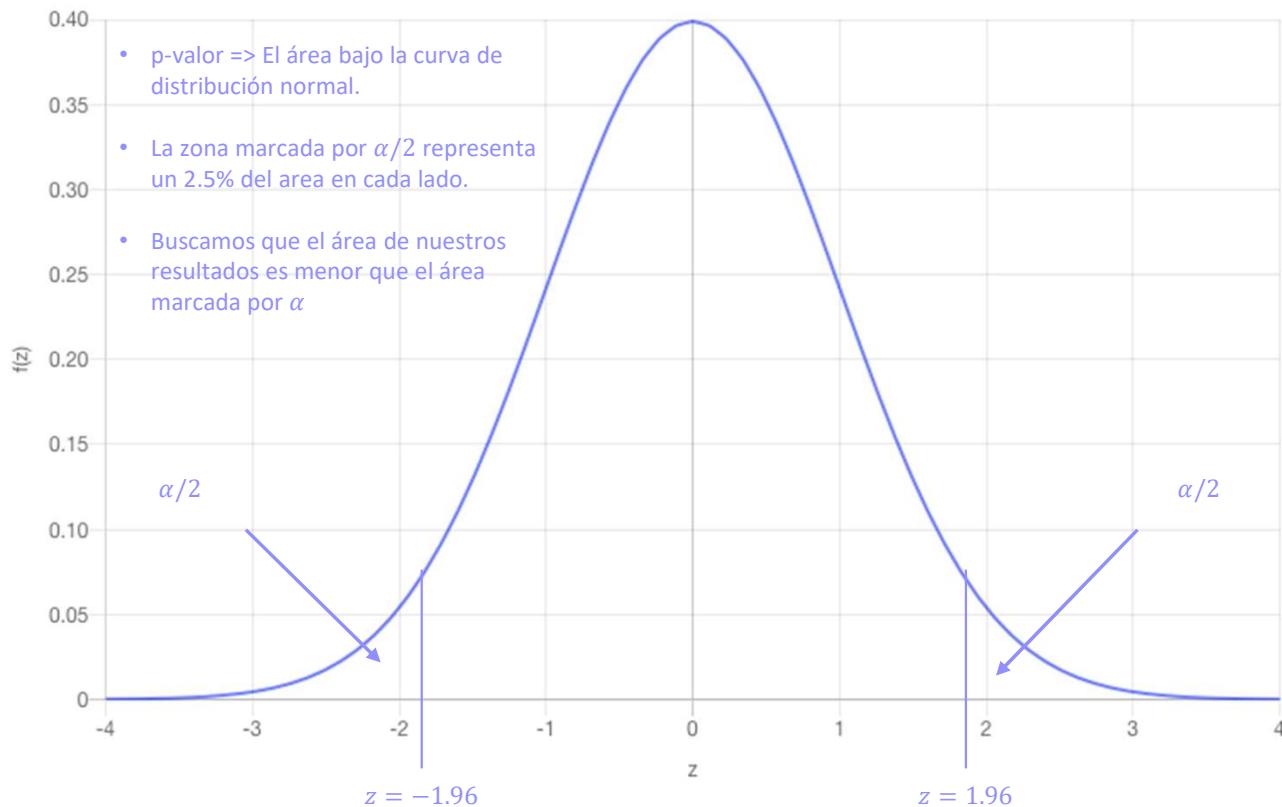
TEST ESTADÍSTICO DE PROPORCIONES (I)

z-score es equivalente a una distribución normal



TEST ESTADÍSTICO DE PROPORCIONES (II)

Para $\alpha_{0.05}$, el z-score = 1.96



$$p_A \neq p_B$$

La equivalencia entre p-valor y z-score de hace mediante la siguiente tabla

	P						
	0.1	0.05	0.025	0.01	0.005	0.001	0.0005
one-tail							
two-tails	0.2	0.1	0.05	0.02	0.01	0.002	0.001
DF							
1	3.078	6.314	12.706	31.821	63.656	318.289	636.578
2	1.886	2.92	4.303	6.965	9.925	22.328	31.6
3	1.638	2.353	3.182	4.541	5.841	10.214	12.924
4	1.533	2.132	2.776	3.747	4.604	7.173	8.61
5	1.476	2.015	2.571	3.365	4.032	5.894	6.869
6	1.44	1.943	2.447	3.143	3.707	5.208	5.959
7	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	1.397	1.86	2.306	2.896	3.355	4.501	5.041
9	1.383	1.833	2.262	2.821	3.25	4.297	4.781
10	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	1.356	1.782	2.179	2.681	3.055	3.93	4.318
13	1.35	1.771	2.16	2.65	3.012	3.852	4.221
14	1.345	1.761	2.145	2.624	2.977	3.787	4.14
15	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	1.337	1.746	2.12	2.583	2.921	3.686	4.015
17	1.333	1.74	2.11	2.567	2.898	3.646	3.965
18	1.33	1.734	2.101	2.552	2.878	3.61	3.922
19	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	1.325	1.725	2.086	2.528	2.845	3.552	3.85
21	1.323	1.721	2.08	2.518	2.831	3.527	3.819
22	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	1.319	1.714	2.069	2.5	2.807	3.485	3.768
24	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	1.316	1.708	2.06	2.485	2.787	3.45	3.725
26	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	1.314	1.703	2.052	2.473	2.771	3.421	3.689
28	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	1.311	1.699	2.045	2.462	2.756	3.396	3.66
30	1.31	1.697	2.042	2.457	2.75	3.385	3.646
60	1.296	1.671	2	2.39	2.66	3.232	3.46
120	1.289	1.658	1.98	2.358	2.617	3.16	3.373
1000	1.282	1.646	1.96	2.33	2.581	3.098	3.3
Inf	1.282	1.645	1.96	2.326	2.576	3.091	3.291

TEST ESTADÍSTICO DE PROPORCIONES (II)

z-Test con dos poblaciones

Condiciones de aplicación:

- ✓ Los datos han de ser obtenido de manera aleatoria
- ✓ Tamaño de muestra > 30
- ✓ Se necesita saber exactamente el número de casos total y afectos por cada variante.

$$p = \frac{5000 + 5250}{100123 + 100133} = 0.0512,$$

Caso 1

100,123

100,133

5,000

5,250

4.99%

5.24%

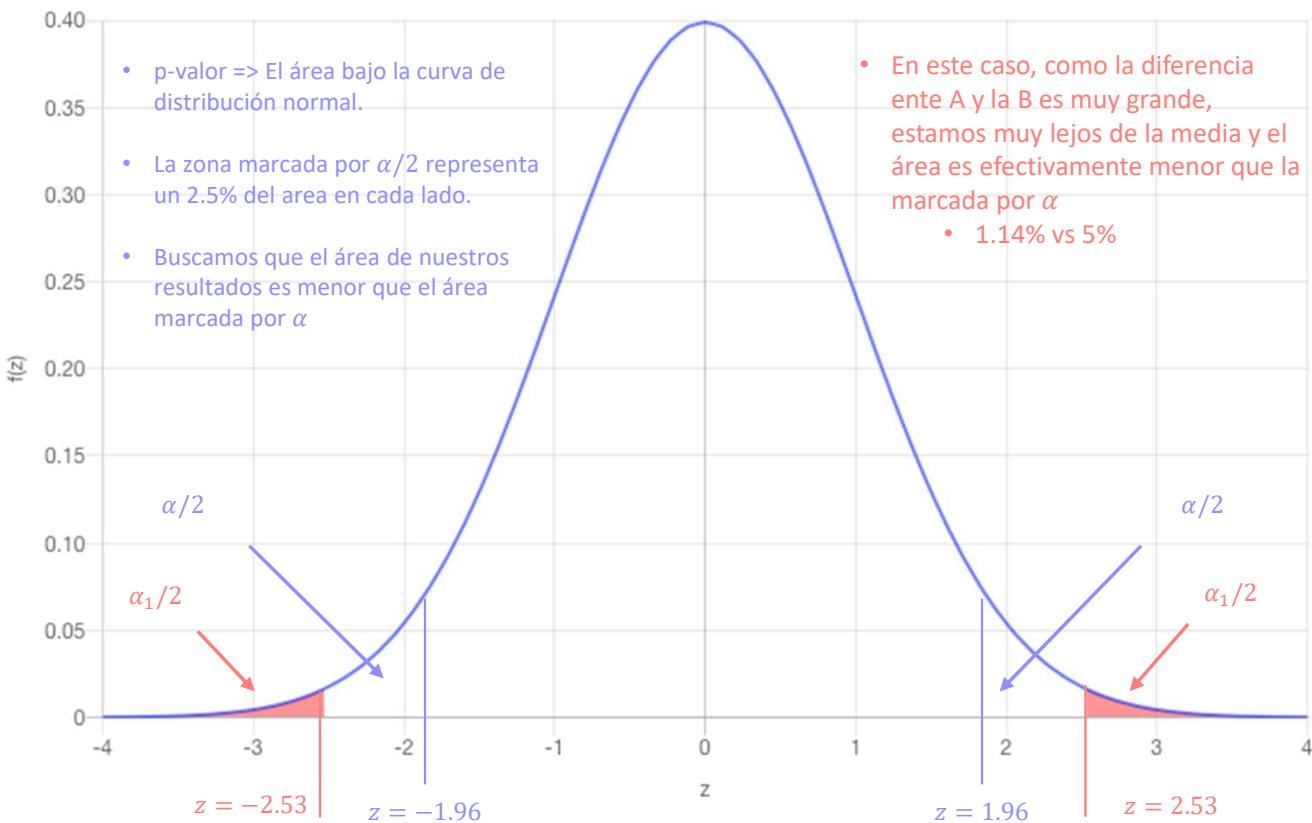
$$Z = \frac{(p_A - p_B) - 0}{\sqrt{p(1-p) \left(\frac{1}{n_A} + \frac{1}{n_B} \right)}},$$

$$\text{donde } p = \frac{y_A + y_B}{n_A + n_B}$$

$$Z = \frac{(0.0499 - 0.0524) - 0}{\sqrt{0.0512(1 - 0.0512) \left(\frac{1}{100123} + \frac{1}{100133} \right)}} = -2.53$$

TEST ESTADÍSTICO DE PROPORCIONES (III)

Caso 1: z-score = 2.53 =>p-valor = 0.0114

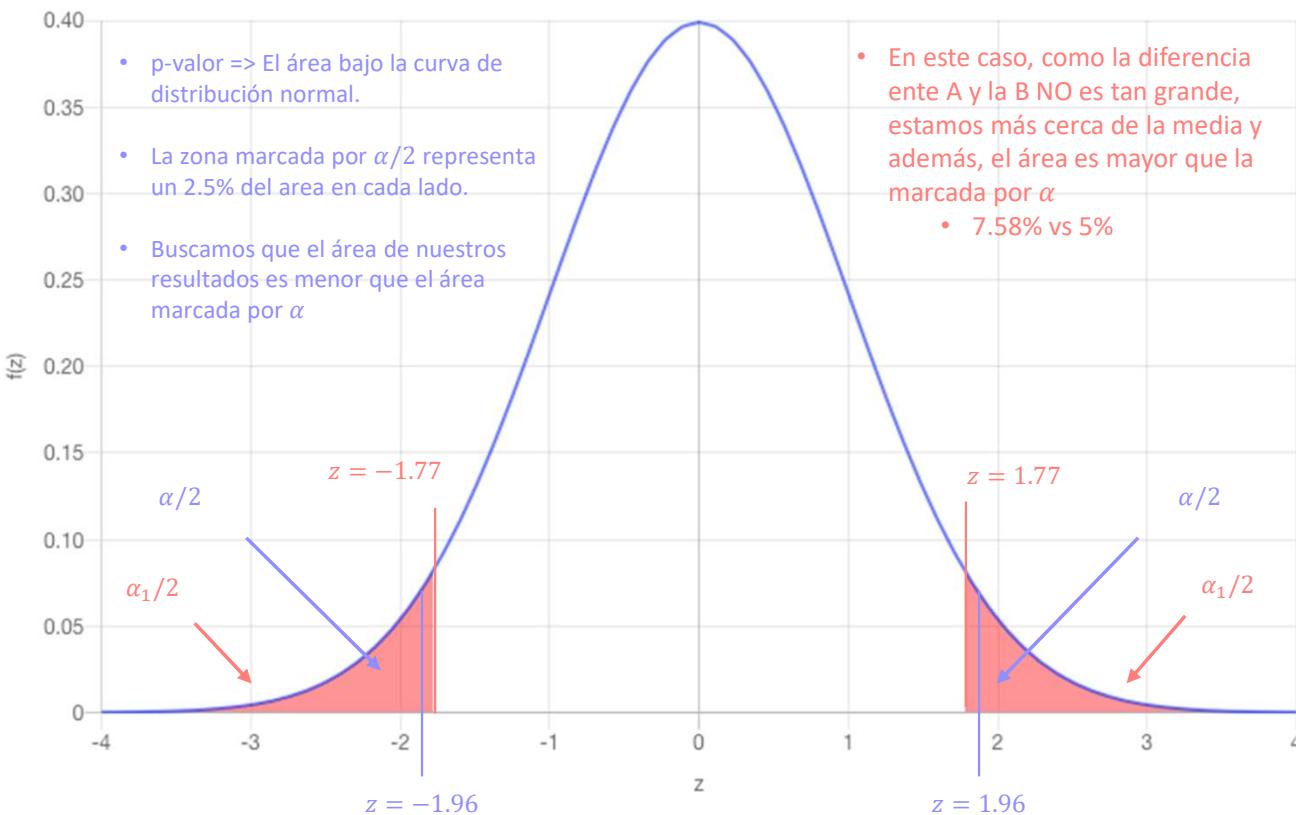


$p = 0.0114$

	P						
	0.1	0.05	0.025	0.01	0.005	0.001	0.0005
one-tail	0.2	0.1	0.05	0.02	0.01	0.002	0.001
two-tails	0.4	0.2	0.05	0.02	0.01	0.002	0.001
DF							
1	3.078	6.314	12.706	31.821	63.656	318.289	636.578
2	1.886	2.92	4.303	6.965	9.925	22.328	31.6
3	1.638	2.353	3.182	4.541	5.841	10.214	12.924
4	1.533	2.132	2.776	3.747	4.604	7.173	8.61
5	1.476	2.015	2.571	3.365	4.032	5.894	6.869
6	1.44	1.943	2.447	3.143	3.707	5.208	5.959
7	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	1.397	1.86	2.306	2.896	3.355	4.501	5.041
9	1.383	1.833	2.262	2.821	3.25	4.297	4.781
10	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	1.356	1.782	2.179	2.681	3.055	3.93	4.318
13	1.35	1.771	2.16	2.65	3.012	3.852	4.221
14	1.345	1.761	2.145	2.624	2.977	3.787	4.14
15	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	1.337	1.746	2.12	2.583	2.921	3.686	4.015
17	1.333	1.74	2.11	2.567	2.898	3.646	3.965
18	1.33	1.734	2.101	2.552	2.878	3.61	3.922
19	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	1.325	1.725	2.086	2.528	2.845	3.552	3.85
21	1.323	1.721	2.08	2.518	2.831	3.527	3.819
22	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	1.319	1.714	2.069	2.5	2.807	3.485	3.768
24	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	1.316	1.708	2.06	2.485	2.787	3.45	3.725
26	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	1.314	1.703	2.052	2.473	2.771	3.421	3.689
28	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	1.311	1.699	2.045	2.462	2.756	3.396	3.66
30	1.31	1.697	2.042	2.457	2.75	3.385	3.646
60	1.296	1.671	2	2.39	2.66	3.232	3.46
120	1.289	1.658	1.98	2.358	2.617	3.16	3.373
1000	1.282	1.646	1.962	2.33	2.581	3.098	3.3
Inf	1.282	1.645	1.96	2.326	2.576	3.091	3.291

TEST ESTADÍSTICO DE PROPORCIONES (III)

Caso 2: z-score = 1.776 =>p-valor = 0.0758



$p = 0.0758$

	P						
	0.1	0.05	0.025	0.01	0.005	0.001	0.0005
one-tail	0.2	0.1	0.05	0.02	0.01	0.002	0.001
two-tails	0.4	0.2	0.1	0.02	0.01	0.002	0.001
DF	1	3.078	6.314	12.706	31.821	63.656	318.289
	2	1.886	2.92	4.303	6.965	9.925	22.328
	3	1.638	2.353	3.182	4.541	5.841	10.214
	4	1.533	2.132	2.776	3.747	4.604	7.173
	5	1.476	2.015	2.571	3.365	4.032	5.894
	6	1.44	1.943	2.447	3.143	3.707	5.208
	7	1.415	1.895	2.365	2.998	3.499	4.785
	8	1.397	1.86	2.306	2.896	3.355	4.501
	9	1.383	1.833	2.262	2.821	3.25	4.297
	10	1.372	1.812	2.228	2.764	3.169	4.144
	11	1.363	1.796	2.201	2.718	3.106	4.025
	12	1.356	1.782	2.179	2.681	3.055	3.93
	13	1.35	1.771	2.16	2.65	3.012	3.852
	14	1.345	1.761	2.145	2.624	2.977	3.787
	15	1.341	1.753	2.131	2.602	2.947	3.733
	16	1.337	1.746	2.12	2.583	2.921	3.686
	17	1.333	1.74	2.11	2.567	2.898	3.646
	18	1.33	1.734	2.101	2.552	2.878	3.61
	19	1.328	1.729	2.093	2.539	2.861	3.579
	20	1.325	1.725	2.086	2.528	2.845	3.552
	21	1.323	1.721	2.08	2.518	2.831	3.527
	22	1.321	1.717	2.074	2.508	2.819	3.505
	23	1.319	1.714	2.069	2.5	2.807	3.485
	24	1.318	1.711	2.064	2.492	2.797	3.467
	25	1.316	1.708	2.06	2.485	2.787	3.45
	26	1.315	1.706	2.056	2.479	2.779	3.435
	27	1.314	1.703	2.052	2.473	2.771	3.421
	28	1.313	1.701	2.048	2.467	2.763	3.408
	29	1.311	1.699	2.045	2.462	2.756	3.396
	30	1.31	1.697	2.042	2.457	2.75	3.385
	60	1.296	1.671	2	2.39	2.66	3.232
	120	1.289	1.658	1.98	2.358	2.617	3.16
	1000	1.282	1.646	1.962	2.33	2.581	3.098
	Inf	1.282	1.645	1.96	2.326	2.576	3.091

$p = 0.0758$

$z = 1.776$

TEST ESTADÍSTICO DE PROPORCIONES (IV)

Cálculo con python

```
from statsmodels.stats.proportion import proportions_ztest
import numpy as np

def calculate_proportions_ztest(sample_size_a, sample_size_b, successes_a, successes_b,
                                alpha, test_type='two-sided'):

    # check our sample against H0 for Ha != H0
    successes = np.array([successes_a, successes_b])
    samples = np.array([sample_size_a, sample_size_b])

    # note, no need for a H0 value here - it's derived from the other parameters
    stat, p_value = proportions_ztest(count=successes, nobs=samples, alternative=test_type)

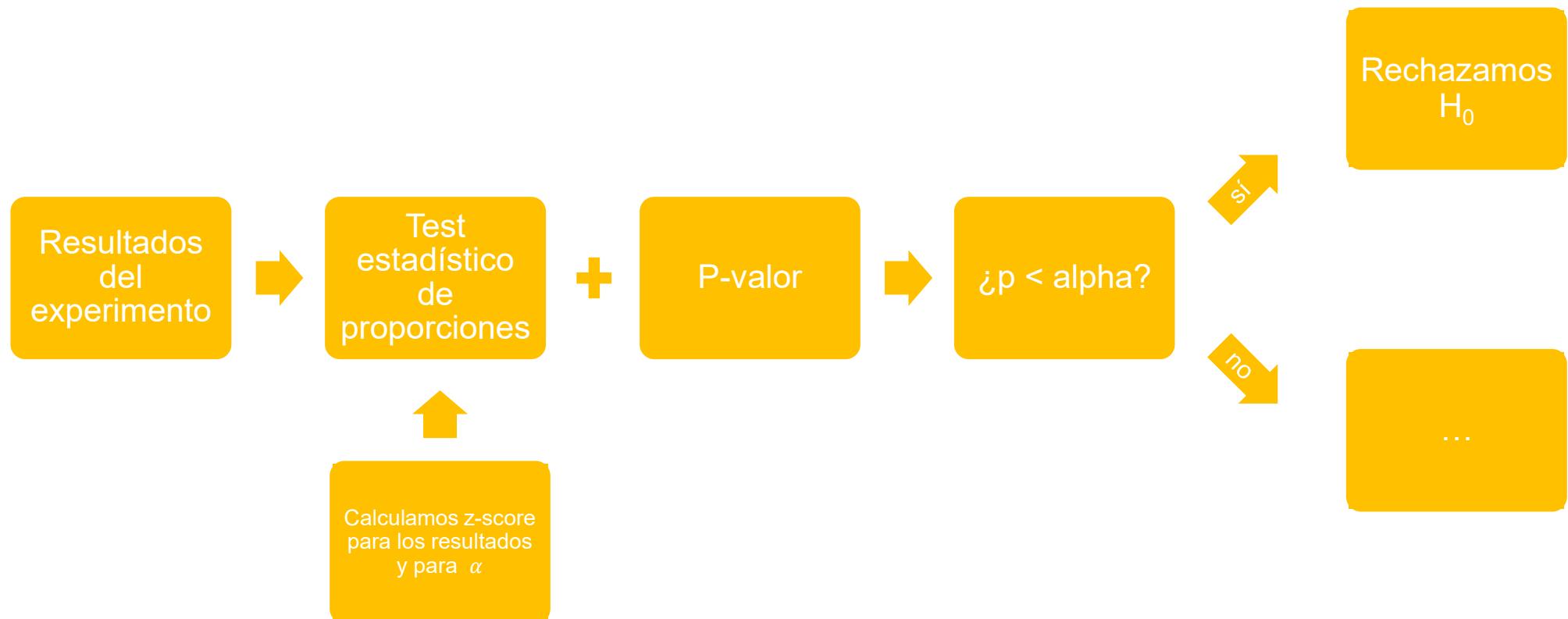
    # report
    print('z_stat: %0.3f, p_value: %0.3f' % (stat, p_value))

    if p_value > alpha:
        print("Fail to reject the null hypothesis - we have nothing else to say")
    else:
        print("Reject the null hypothesis - suggest the alternative hypothesis is true")
```

```
-----
Resultados caso 1:
-----
z_stat: -2.530, p_value: 0.011
Reject the null hypothesis - suggest the alternative hypothesis is true

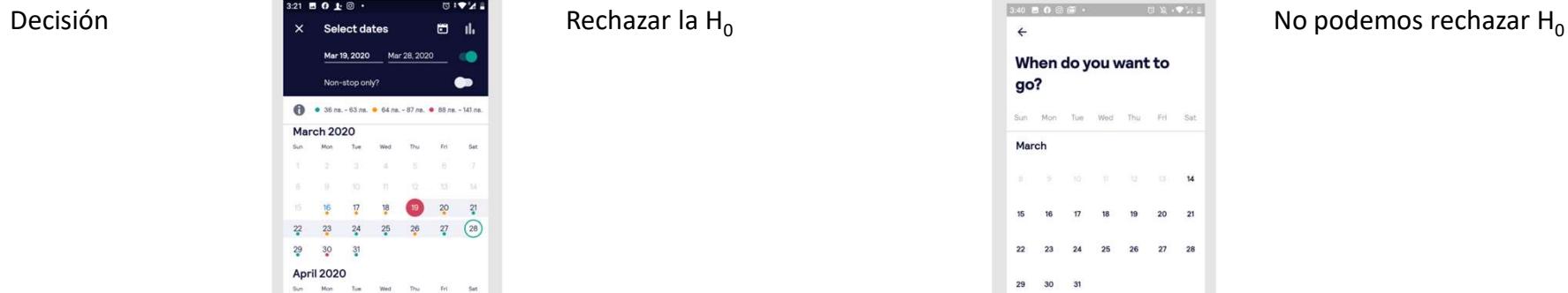
-----
Resultados caso 2:
-----
z_stat: -1.776, p_value: 0.076
Fail to reject the null hypothesis - we have nothing else to say
```

PASOS PARA LA TOMA DE DECISIONES



RESULTADOS DE NUESTRO EXPERIMENTO

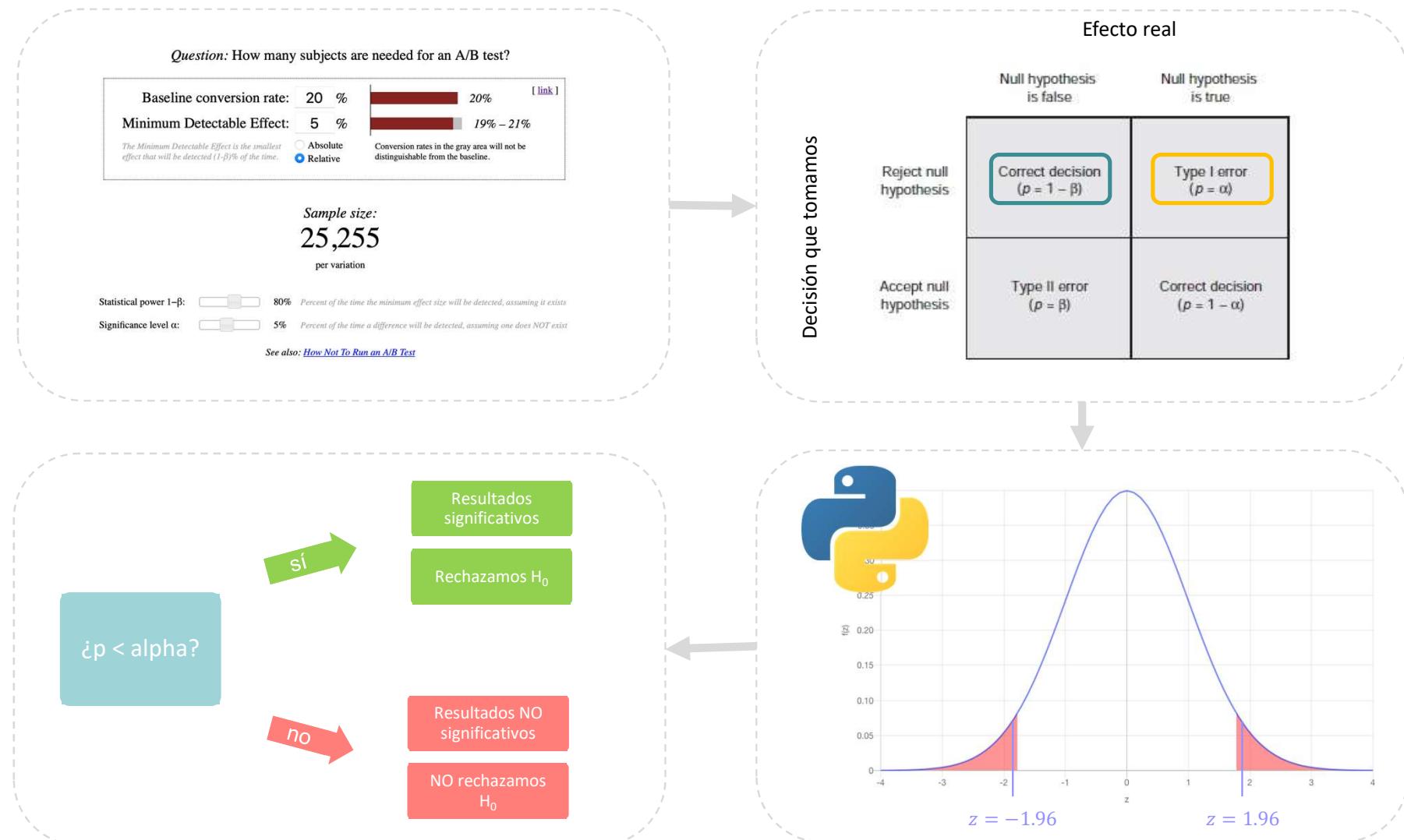
	Caso 1	Caso 2
Tamaño de muestra	100,123	100,133
Conversiones	5,000	5,250
Tasa de conversión	4.99%	5.24%
p-valor	0.0114	0.0758
alpha	0.05	0.05



- ¿Qué tamaño de muestra necesito?
- ¿Qué magnitud a detectar es adecuada?
- ¿Qué probabilidad tengo de tomar una decisión equivocada?
- Tengo los resultados de mi experimento, ¿qué decisión tomo?



Repaso de la parte 3





Parte 4. Consideraciones prácticas de implementación

PARTE 4. CONSIDERACIONES PRÁCTICAS DE IMPLEMENTACIÓN

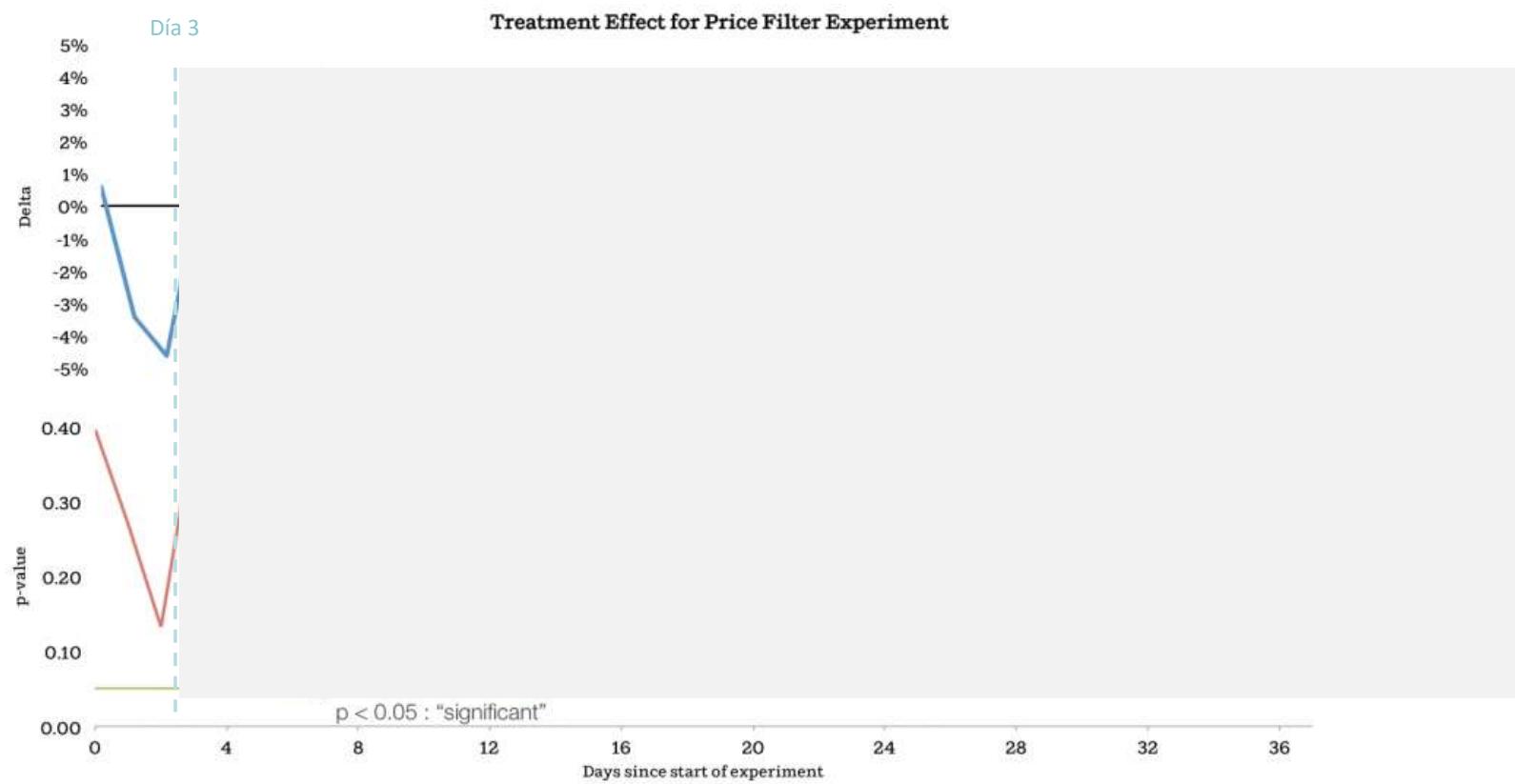
Objetivos:

- P-eaking
- Métricas
- ¿Cuántas métricas analizamos?
- ¿Cómo tratar diferentes segmentos?

- P-eaking
- Métricas
- ¿Cuántas métricas analizamos?
- ¿Cómo tratar diferentes segmentos?

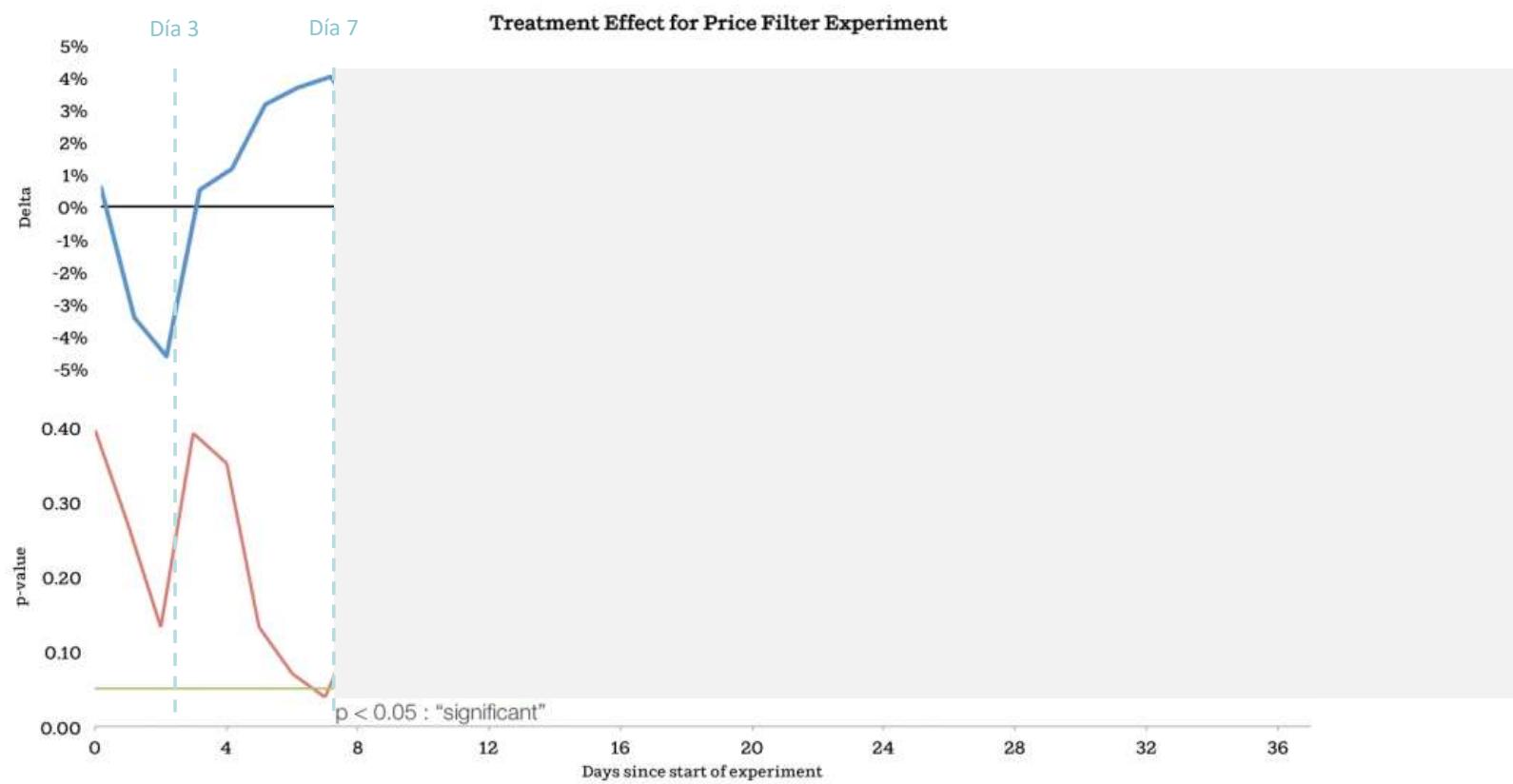
P-EAKING EN EXPERIMENTACIÓN

Ejemplo de experimento de AirBnB



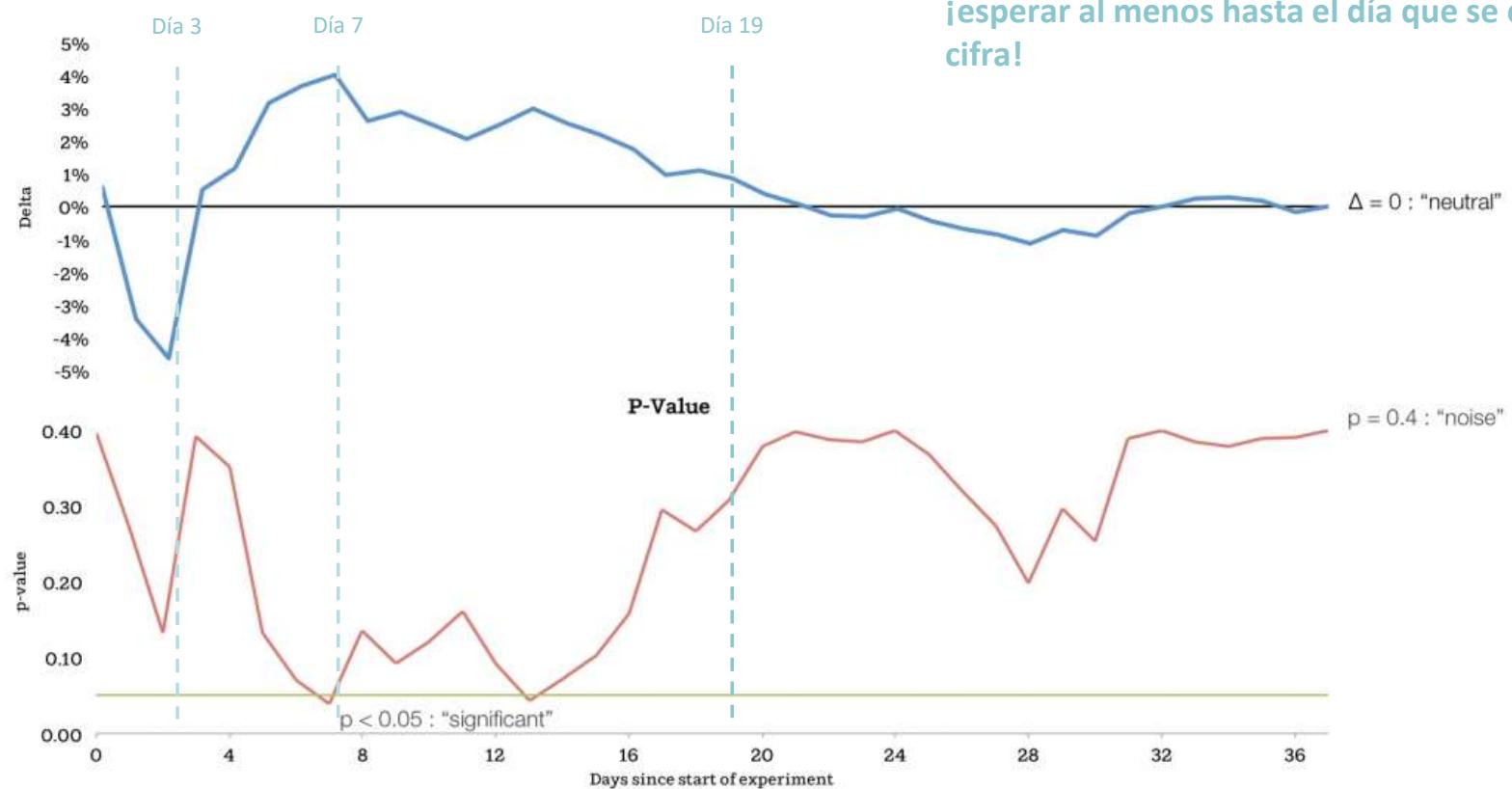
P-EAKING EN EXPERIMENTACIÓN

Ejemplo de experimento de AirBnB



P-EAKING EN EXPERIMENTACIÓN

¡Recordad la calculadora mágica!

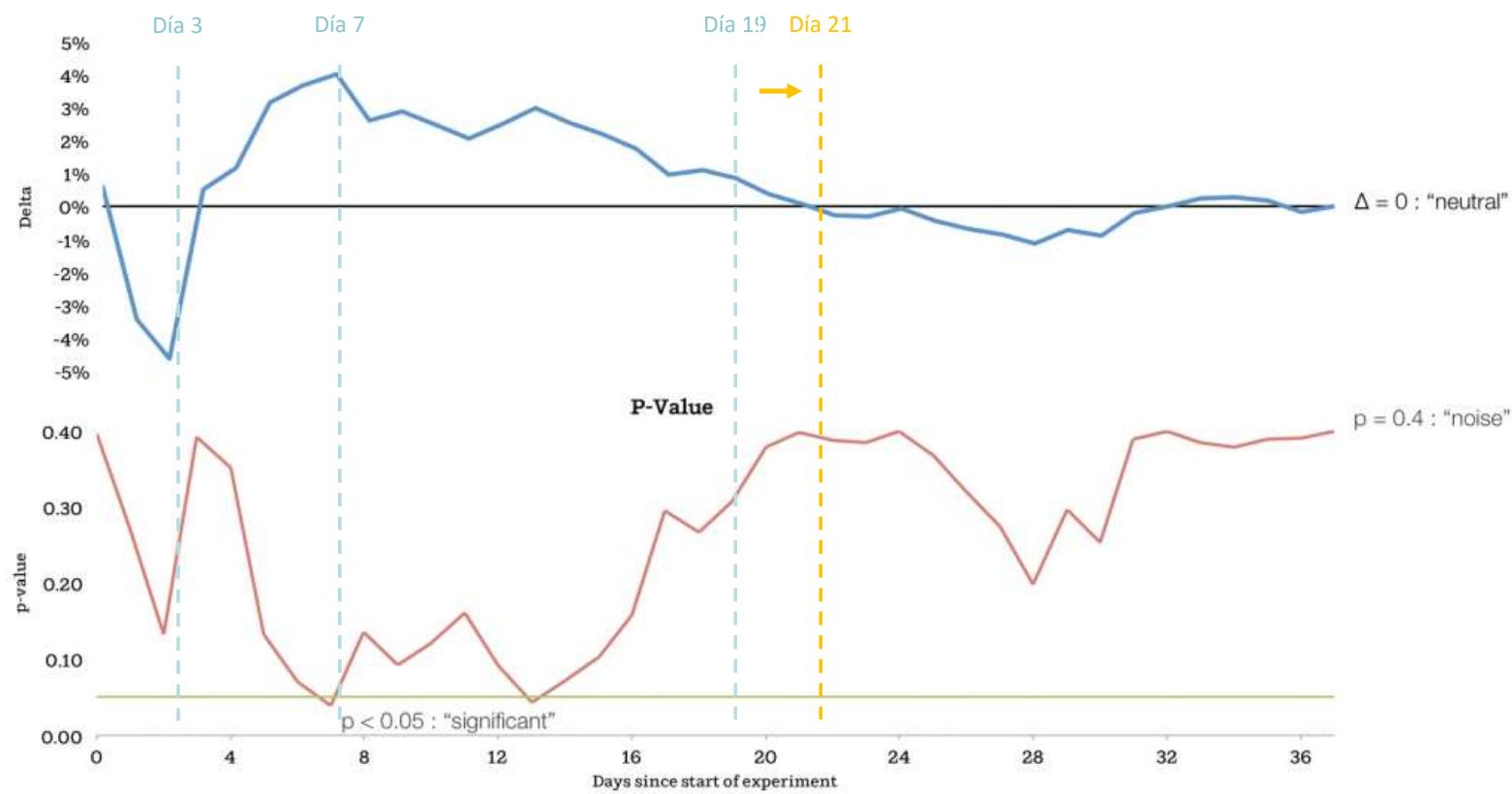


Si diseñamos nuestro experimento con un tamaño de muestra de 2 millones de usuarios...
esperar al menos hasta el día que se consiga esa cifra!

P-EAKING EN EXPERIMENTACIÓN

Recomendación: Ciclos de negocio

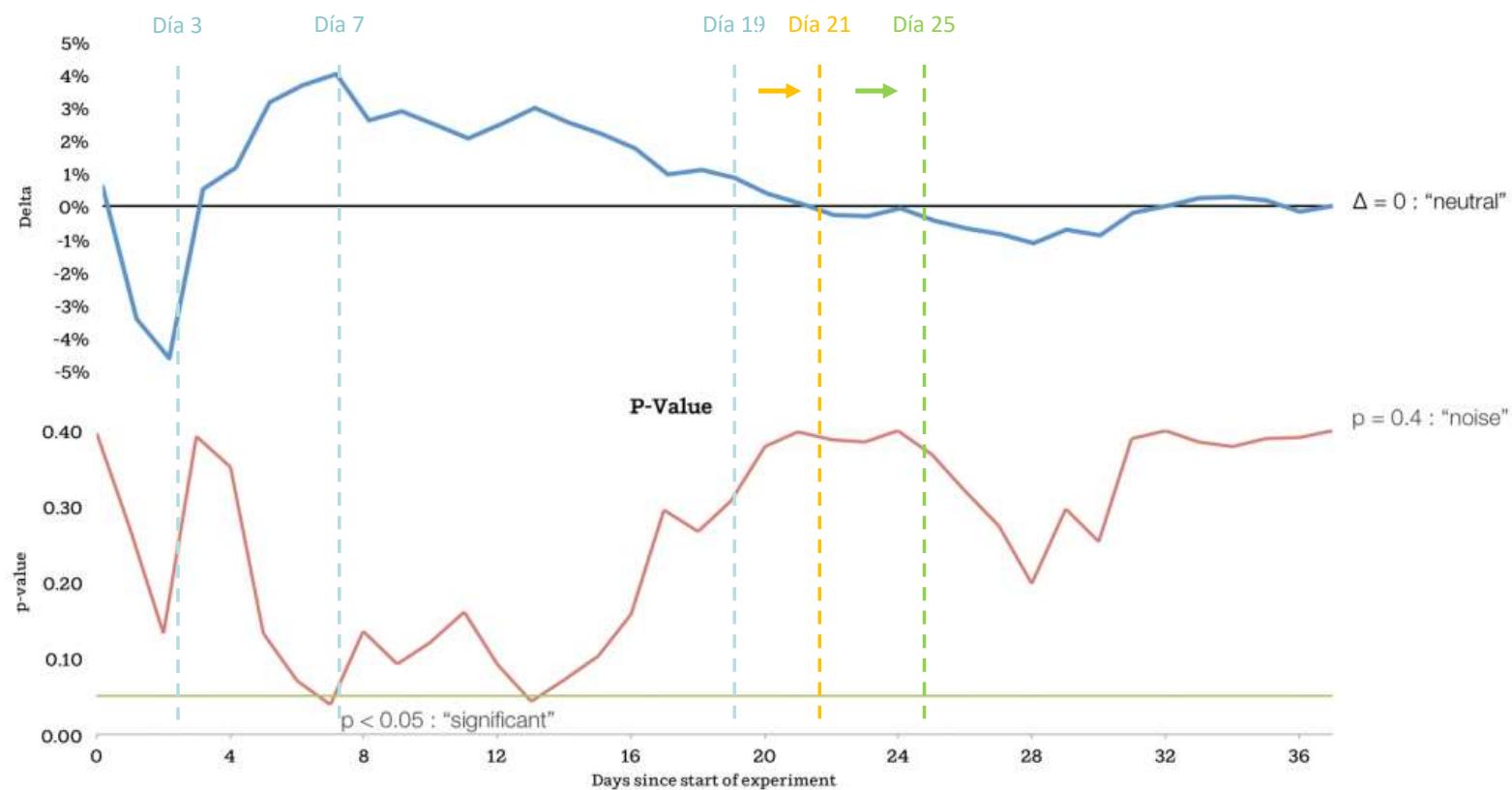
- Ciclo de negocio: semanal => Claras diferencias de comportamiento entre semana y el fin de semana.
- Recomendación: alcanzar cifra de usuarios Y completar el ciclo.



P-EAKING EN EXPERIMENTACIÓN

Recomendación: Estabilidad del p-valor

- Si estamos tratando con un experimento súper crítico, tambien puede ser interesante asegurar que p-valor no tenga fluctuaciones de significativo a no significativo.



- P-eaking
- **Diseño de Métricas**
- ¿Cuántas métricas analizamos?
- ¿Cómo tratar diferentes segmentos?

¿QUÉ CARACTERIZA UNA BUENA MÉTRICA?

Hasta ahora sólo hemos trabajado con ratios conversión...

Definición de una buena métrica

1. **Sensitiva:** la métrica debería variar con pequeños cambios en la satisfacción del usuario o valor generado.
2. **Direccional:** si la satisfacción del usuario aumenta, también debería hacerlo nuestra métrica.
Excepción: cuando un aumento de la satisfacción del usuario va en detrimento del valor generado (ejemplo, alerta de un desastre natural para que el cliente no reserve un hotel en esa zona).
3. **Entendida:** todo el mundo tiene que entender cómo está la métrica relacionada con la generación de valor y satisfacción.



Core Metrics

- Todos los experimentos han de reportar estas métricas.

Targeted Metrics

- Cualquier experimento puede definir su métrica de éxito individualmente.



Success metrics

How will you determine the success of your experiment? [Read more about how to choose your metric](#). Not finding the right metric? Get started with Metrics Catalog. Read more about MOE or how it's used.

Show only core metrics and those from my mission/tribe/squad.

Success metric *

Expected change %

Guardrail metrics

How will you measure that you haven't messed something up? [Read more about how to choose your metric](#). Not finding the right metric? Get started with Metrics Catalog. Read more about NIM.

Show only core metrics and those from my mission/tribe/squad.

Guardrail metric *

Should not %

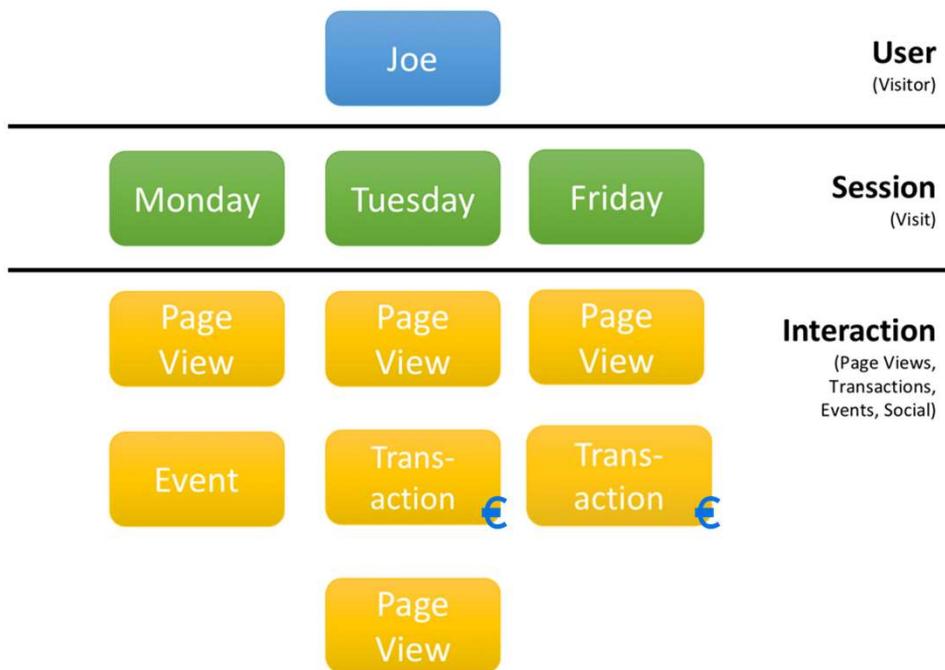
Guardrail metric *

Should not %

Guardrail metric *

Should not %

¿CÓMO DISEÑAR UNA BUENA MÉTRICA?



¿Qué métrica escogeríais para diseñar y evaluar vuestro experimento?

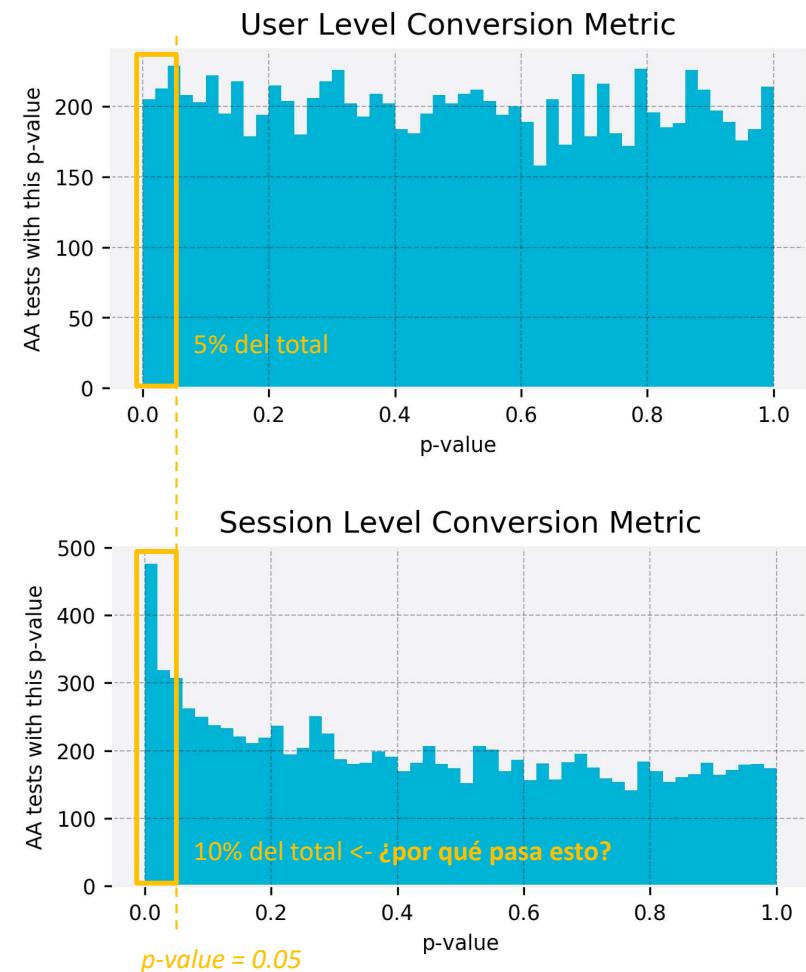
- a) Transaction/Session = 2/3
- b) Transaction/User = 2/1
- c) Unique transaction/user = 1/1

MÉTRICAS BASADAS EN USUARIO VS SESIONES

Simulaciones:

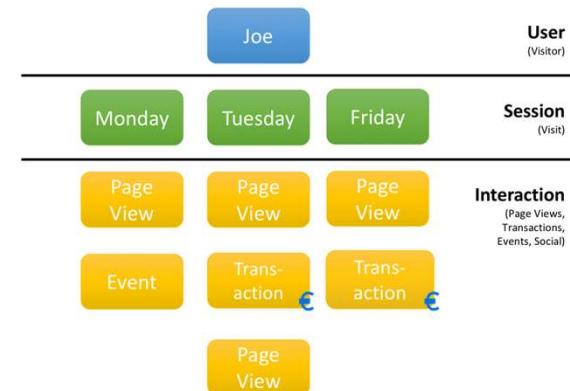
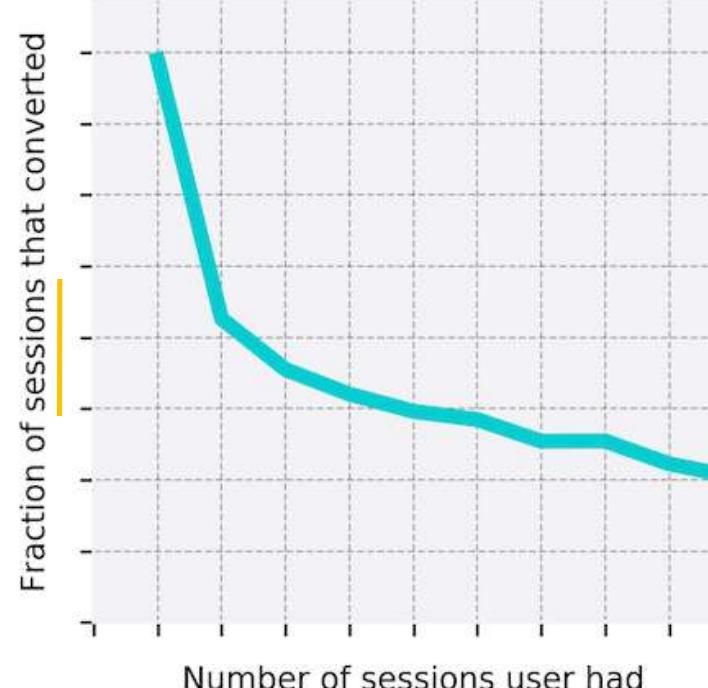
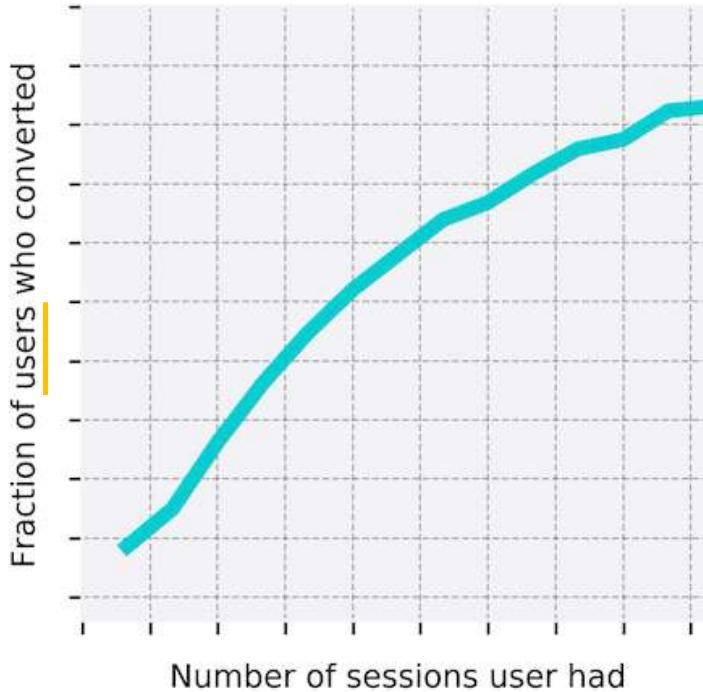
1. Cogemos 1 semana de transacciones en Skyscanner.
2. Formamos dos grupos seleccionando usuarios aleatoriamente (A/A test)
3. Calculamos el p-valor de la diferencia entre los dos ratios de conversión a nivel de sesión y de usuario
4. Repetimos (2) y (3) múltiples veces

- A/A test implica que NO debería haber diferencia entre los 2 grupos.
- Si elegimos un p-valor de 0.05, **deberíamos esperar que ~5% of los tests tuvieran un error tipo I.**
 - Si los grupos fueran estadísticamente independientes, los p-valores deberían estar distribuidos uniformemente.



MÉTRICAS BASADAS EN USUARIO VS SESIONES

Concepto clave: independencia estadística

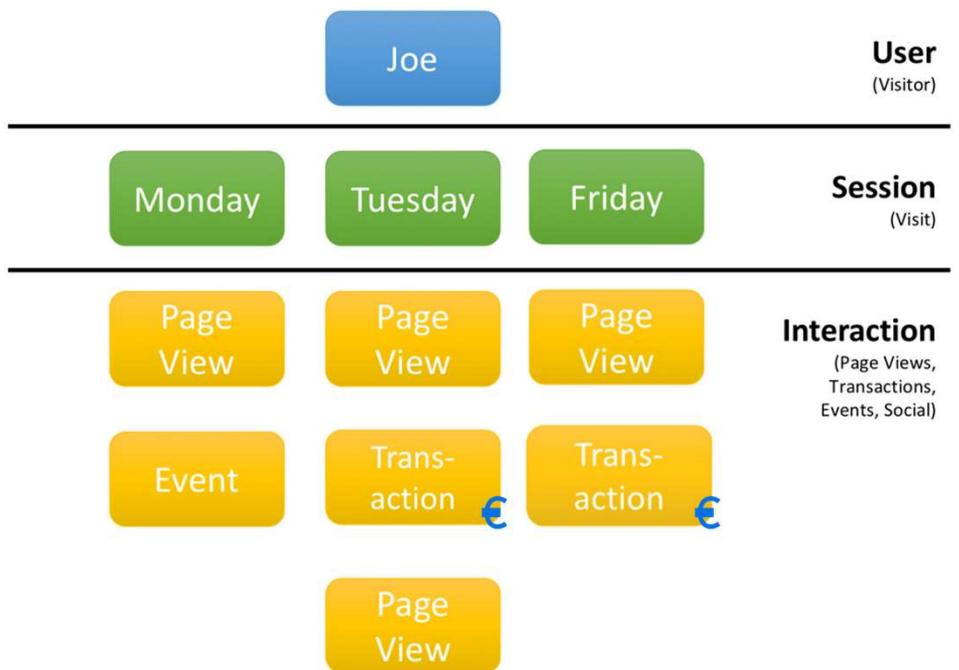


Claramente cuando un usuario empieza varias sesiones, lo que pase en una anterior afectará a una futura.

- *¿Podemos asegurar que la compra que hice Joe el viernes no se debe a su experiencia de compra el martes?*

No se cumple la independencia estadística si medimos con respecto a las sesiones.

¿CÓMO DISEÑAR UNA BUENA MÉTRICA?



¿Qué métrica escogeríais para diseñar y evaluar vuestro experimento?

- a) ~~Transaction/Session = 2/3~~
- b) Transaction/User = 2/1
- c) Unique transaction/user = 1/1

MÉTRICAS BASADAS EN TRANSACCIÓN ÚNICA VS TOTAL TRANSACCIONES

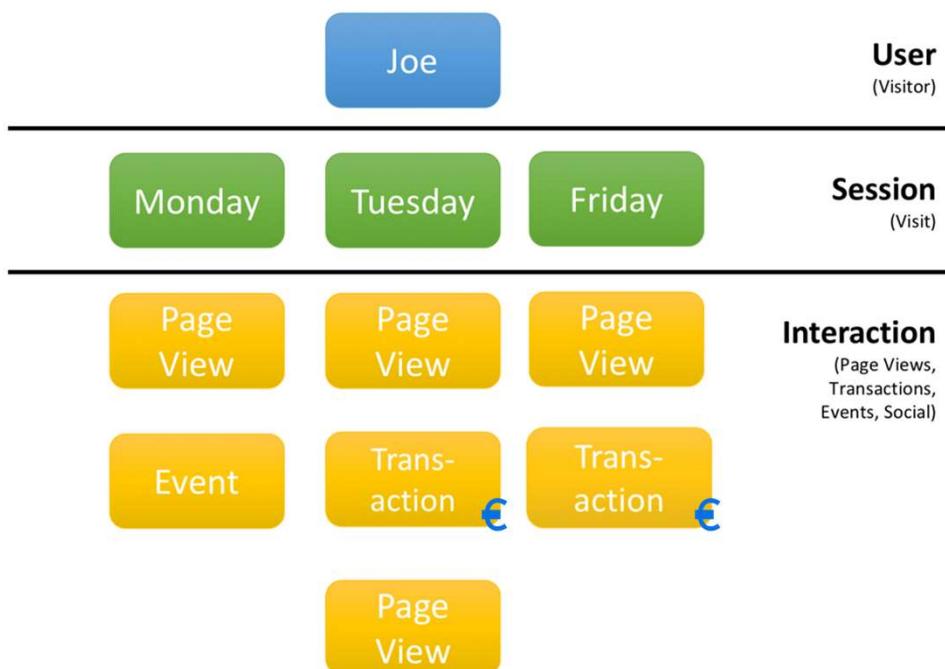
¿Qué pasa cuando tienes “power users”?

Variante	Total Users	Usuarios que hicieron al menos 1 transacción	Transacciones totales	Unique transactions		Total transactions Users
				Users		
A	10,000	100	120	100 / 10,000 = 1%		120 / 10,000 = 1.2%
B	10,000	80	150	80 / 10,000 = 0.8%		150 / 10,000 = 1.5%

Hay usuarios que hicieron más de 1 transacción

- Recordad -> concepto clave: independencia estadística
 - Total Transactions/Users viola la independencia estadística
- Recomendación:
 1. Para el test de estadístico de proporciones y la toma de decision -> usar *Unique Transactions/Users*
 2. Usar *Total Transactions/Users* como métrica de apoyo para presentar vuestros análisis.

¿CÓMO DISEÑAR UNA BUENA MÉTRICA?



¿Qué métrica escogeríais para diseñar y evaluar vuestro experimento?

a) ~~Transaction/Session = 2/3~~

b) ~~Transaction/User = 2/1~~

c) Unique transaction/user=1/1

- P-eaking
- Diseño de Métricas
- **• ¿Cuántas métricas analizamos?**
- **• ¿Cómo tratar diferentes segmentos?**

¿CUÁNTAS MÉTRICAS ANALIZAMOS?

967 results

Sort by: Best, Cheapest first, Fastest first, Outbound: Departure time, Return: Departure time.

	Flight Details	Price	Action
1	IBERIA 18:55 BCN → 20:15 LHR 06:05 LHR → 09:10 BCN	299 € 2h 13 (average)	Select →
2	IBERIA 18:55 BCN → 20:15 LHR 06:05 LHR → 09:10 BCN	299 € 2h 13 (average)	Select →
3	IBERIA 18:55 BCN → 20:15 LHR 06:05 LHR → 09:10 BCN	299 € 2h 13 (average)	Select →
4	IBERIA 11:40 BCN → 13:05 LHR Operated by British Airways	289 € 2h 26 (average)	Select →
5	IBERIA 06:05 LHR → 09:10 BCN Operated by British Airways	289 € 2h 05 (average)	Select →

Re-sorting rate

Recommended London hotels for you

Hotel Name	Rating	Reviews	Price
Travelodge London C...	4.0	177 reviews	81 €/night
Holiday Inn Express Lo...	4.5	77 reviews	91 €/night
Best Western London L...	3.5	72 reviews	46 €/night

Click-through Rate to Partner selection

Outbound Sun, 20 Feb 2022

18:55 BCN → 20:15 LHR
Operated by British Airways

Return Sun, 27 Feb 2022

06:05 LHR → 09:10 BCN
Operated by British Airways

Book your ticket

Economy class, 1 adult

Booking Site	Total Price	Action
eDreams	312 €	Select →
Opodo	317 €	Select →
Travelgenio	323 €	Select →
Trip.com	324 €	Select →

Click-through Rate to Partner site

Price Details

Passengers

Passenger Type	Price
Adult	€323.24 x 1
Fare	€263.34 x 1
Taxes & fees	€59.90 x 1

Total **€323.24**

Trip Coins ⓘ
You'll earn **92 Trip Coins** after your trip.

Book with Confidence

- Trusted by Millions
- 24/7 Customer Support
- Safe & Secure

Booking Rate

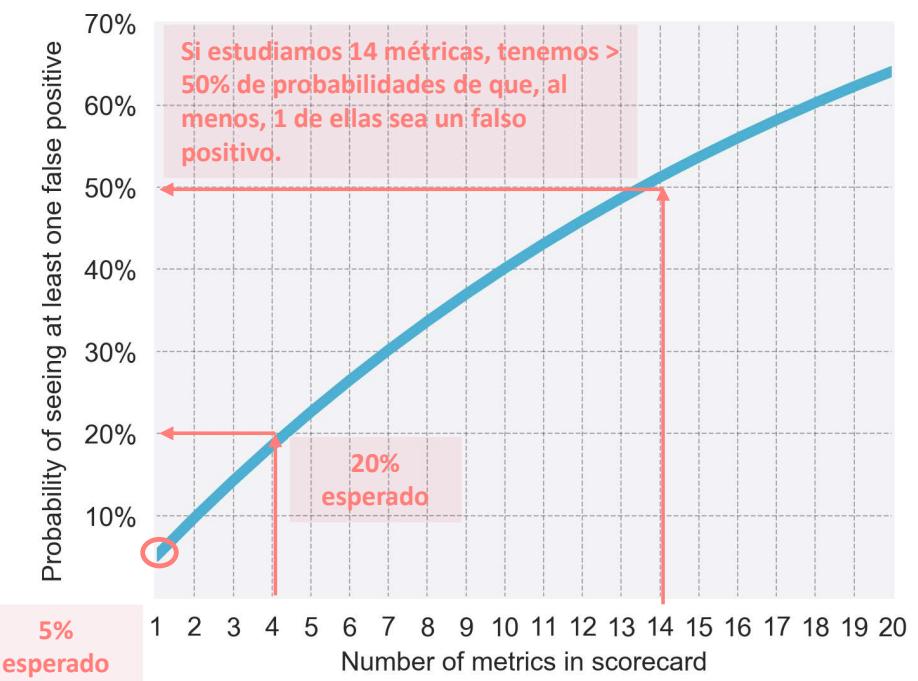
¿CUÁNTAS MÉTRICAS ANALIZAMOS?

Recomendación: Corrección de alpha

Corrección de Bonferroni: $\alpha/n_{\text{métricas}}$

- Reducir α por el número de métricas que quieras analizar.
- Por ejemplo, para mantener nuestro 95% de confianza midiendo 5 métricas, necesitaríamos un alpha de 0.01 (en lugar de 0.05 para 1 sola).
- ¿Qué consiguimos?
 - Ser más estrictos cuando comparamos el p-valor de nuestro experimento con un α más pequeño

A/A test comparando el % de falsos positivos (errores tipo I) cuando medimos más de 1 métrica a la vez.

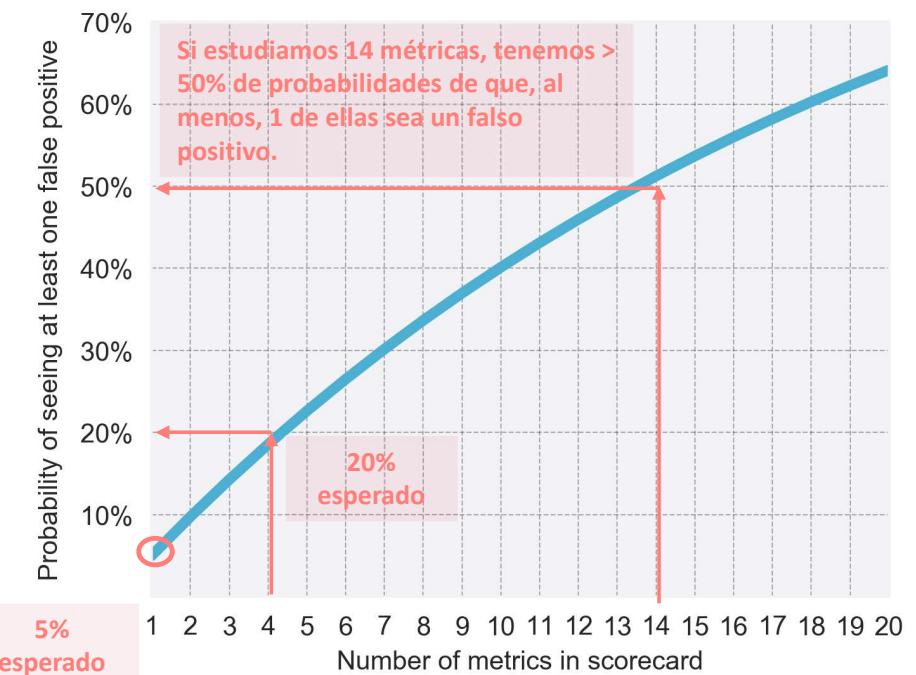


¿CUÁNTAS MÉTRICAS ANALIZAMOS?

Recomendación: Fijar una métrica principal

- Escoger 1 sola métrica como la importante.
- Diseñar el experimento en base a esta métrica
 - ¿Tasa de éxito? -> tamaño de muestra necesario
- Usar las otras métricas como apoyo extra
- Ejemplo 1. Métrica principal es positivamente significativa, pero la métrica 3 negativamente significativa.
 - No lanzar la nueva variante y diseñar un experimento con la métrica 3 como la principal para asegurar que este efecto es real.
- Ejemplo 2. Métrica principal NO es significativa, pero la métrica 3 es positivamente significativa.
 - No lanzar la nueva variante y diseñar un experimento con la métrica 3 como la principal para asegurar que este efecto es real.

A/A test comparando el % de falsos positivos (errores tipo I) cuando medimos más de 1 métrica a la vez.

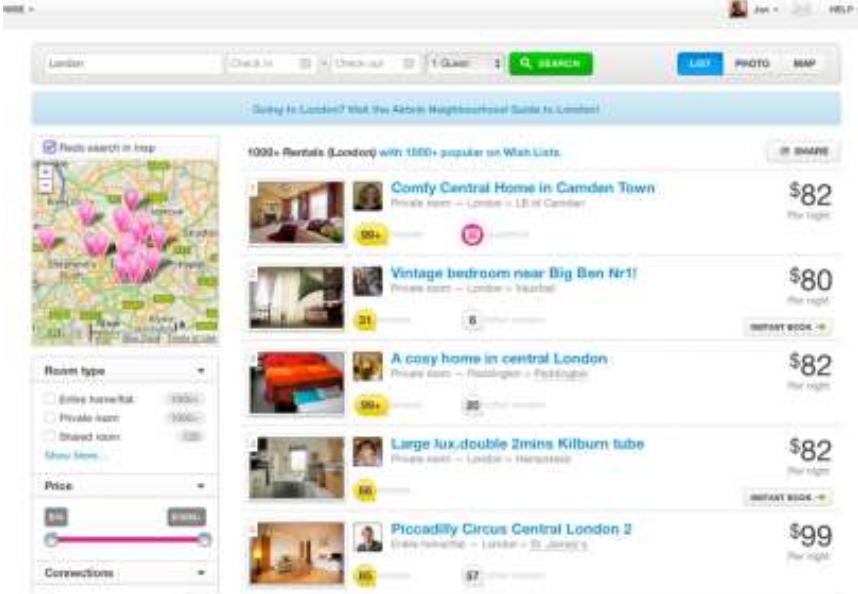


- P-eaking
- Diseño de Métricas
- ¿Cuántas métricas analizamos?
- **¿Cómo tratar diferentes segmentos?**

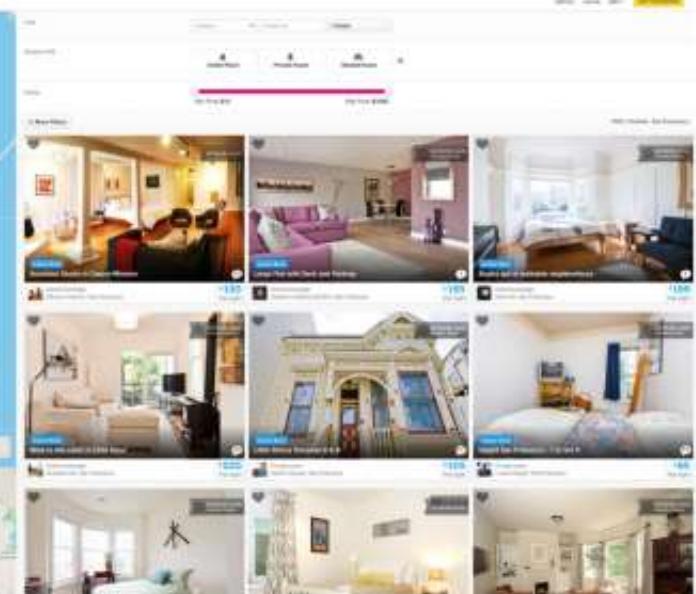
¿CÓMO TRATAR DIFERENTES SEGMENTOS?

Ejemplo de AirBnB (de hace muuchos años...)

Before



After



¿CÓMO TRATAR DIFERENTES SEGMENTOS?

¿Sirve de algo mirar los resultados en diferentes segmentos?

Before

After

Browser	Δ	p
All	-0.27%	0.29
Chrome	2.07%	0.01
Firefox	2.81%	0.00
IE	-3.66%	0.00
Safari	0.86%	0.26
Rest	-0.74%	0.33

Estos resultados no coincidían para nada con el user research que llevaron a cabo, donde la nueva versión ganaba por amplia mayoría.

En realidad, lo que vieron es que la nueva funcionalidad había roto un importante componente en Internet Explorer.

¿CÓMO TRATAR DIFERENTES SEGMENTOS?

Recomendación: segmentos independientes

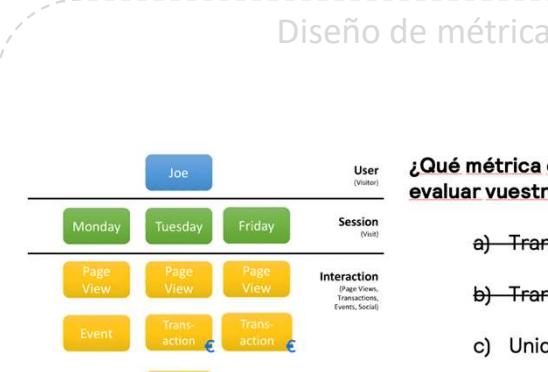
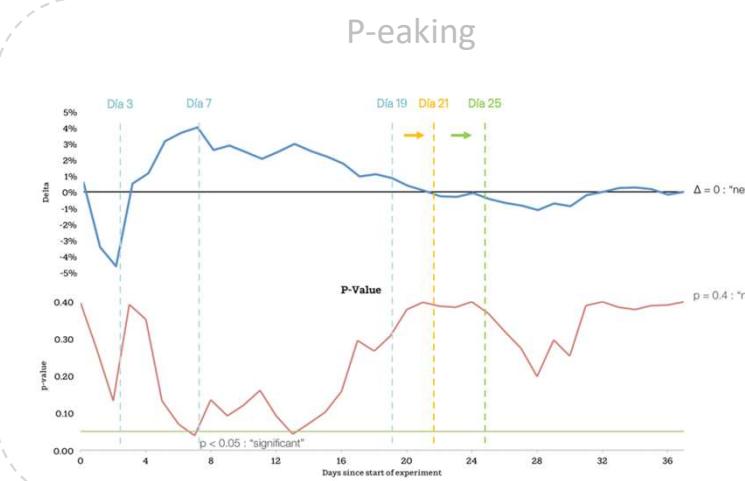
Browser	Δ	p
All	-0.27%	0.29
Chrome	2.07%	0.01
Firefox	2.81%	0.00
IE	-3.66%	0.00
Safari	0.86%	0.26
Rest	-0.74%	0.33

- Recordatorio 1 -> tamaño de muestra
- Recordatorio 2 -> al igual que cuándo aumentamos el número de métricas a analizar, si aumentamos el número de segmentos, la probabilidad de un falso positivo aumenta.
- Recordatorio 3 -> concepto clave: independencia estadística
 - Tenemos que tener muy claro que los segmentos no tienen una influencia clara entre ellos.
- Recomendación: no tomar decisiones estadísticas sobre segmentos individuales si no se ha diseñado el experimento para tal efecto.
 - Puede haber diferencias en la tasa de conversión (tamaño de muestra) o en el volumen tráfico (qué día paramos)

- P-eaking
- Diseño de Métricas
- ¿Cuántas métricas analizamos?
- ¿Cómo tratar diferentes segmentos?



Repaso de la parte 4



¿Qué métrica escogeríais para diseñar y evaluar vuestro experimento?

- a) Transaction/Session = 2/3
- b) Transaction/User = 2/1
- c) Unique transaction/user = 1/1

Segmentos

Browser	Δ	p
All	-0.27%	0.29
Chrome	2.07%	0.01
Firefox	2.81%	0.00
IE	-3.66%	0.00
Safari	0.86%	0.26
Rest	-0.74%	0.33

Múltiples métricas



AGENDA DÍA 1

- Romper el hielo: Me presento 😊
- Parte 1. A/B testing en el mundo digital.
- Parte 2. Tomando decisiones en vuestro primer experimento.
- Parte 3. Diseño de ejecución de un experimento y toma de decisiones
 - Repaso de la parte 3
- Parte 4. Consideraciones prácticas de implementación
 - Repaso de la parte 4

The background features a vibrant, multi-colored gradient from purple on the left to red and orange on the right. Several balloons of various colors (pink, yellow, blue, white) are scattered throughout the scene. In the foreground, the silhouettes of several people's hands are visible against the bright background. One person's hands are specifically highlighted, forming a heart shape. The overall atmosphere is celebratory and joyful.

Gracias