

# A/B Testing en el mundo digital – Día 2

**SKYSCANNER**

Jose Parreño García

Curso 2022

Fecha 07/04/2022

# AGENDA DÍA 2

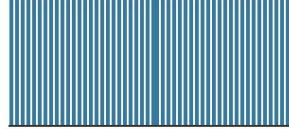
- Romper el hielo.
  - Un par de recursos extra por si os apetece seguir indagando en A/B testing
  - Más allá del test estadístico de proporciones
- Parte 1. Repaso del día 1.
- Parte 2. Debilidades del método frecuentista
- Parte 3. A/B testing con análisis Bayesiano
- Parte 4. ¿Soluciona Bayes las debilidades del método frecuentista?
- Parte 5. Bayes vs Frecuentista
- Parte 6. Test de hipótesis y Bayes en otros usos

A photograph of a renewable energy facility. In the foreground, several blue solar panels are angled upwards. In the background, a white wind turbine with three blades is visible against a clear blue sky with a few wispy clouds.

Un par de recursos extra  
por si os apetece seguir  
indagando en A/B testing

# DATASETS PARA JUGAR

- Opción 1: <https://www.kaggle.com/tklimonova/grocery-website-data-for-ab-test>

RecordID	IP Address	# LoggedInFlag	ServerID	# VisitPageFlag
identifier of the row of data	address of the user, who is visiting website	1 - when user has an account and logged in	one of the servers user was routed through	1 - when user clicked on the loyalty program page
	<b>99516</b> unique values	0 1	1 3	0 1
1	39.13.114.2	1	2	0
2	13.3.25.8	1	1	0
3	247.8.211.8	1	1	0
4	124.8.220.3	0	3	0

- IP Address: usuario
- ServerID: control o variante
- VisitPageFlag: lo que queremos medir
- LoggedInFlag: segmento

# DATASETS PARA JUGAR

- Opción 2: <https://www.kaggle.com/osuolaleemmanuel/ad-ab-testing/version/1>

experiment	date	# hour	device_ma...	platform_os	browser	# yes
exposed	2020-07-10	8	Generic Smartphone	6	Chrome Mobile	0
exposed	2020-07-07	10	Generic Smartphone	6	Chrome Mobile	0
exposed	2020-07-05	2	E5823	6	Chrome Mobile WebView	0
control	2020-07-03	15	Samsung SM-A705FN	6	Facebook	0

- experiment:** control o variante
- yes:** lo que queremos medir
- platform\_os o device\_make:** segmento

# MOOCS



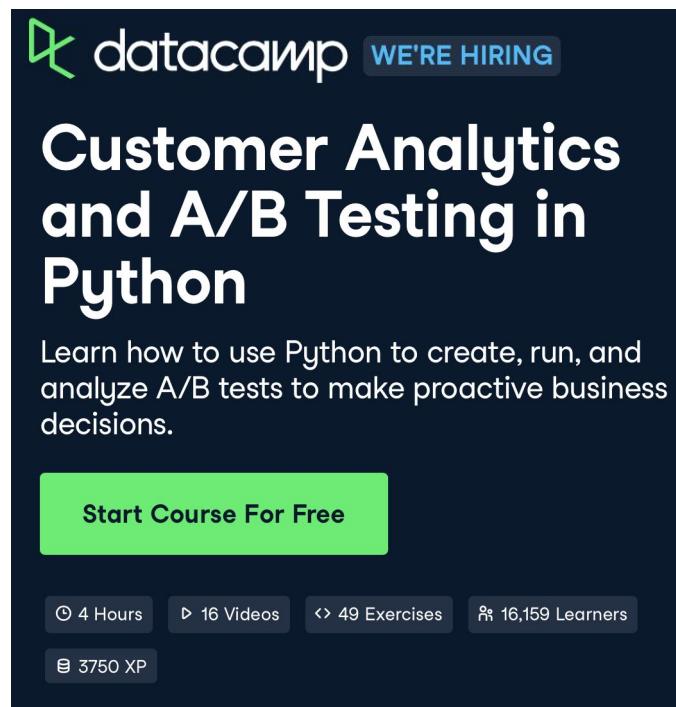
[Home](#) > [Catalog](#) > [Introduction to A/B Testing](#)

FREE COURSE

## A/B Testing by

Online Experiment Design and Analysis

[START FREE COURSE](#)



The screenshot shows a DataCamp course page titled "Customer Analytics and A/B Testing in Python". At the top right, there's a "WE'RE HIRING" button. The main title is "Customer Analytics and A/B Testing in Python" in large white font. Below it is a subtitle: "Learn how to use Python to create, run, and analyze A/B tests to make proactive business decisions." A prominent green button labeled "Start Course For Free" is centered below the subtitle. At the bottom, there are several course statistics: "4 Hours", "16 Videos", "49 Exercises", "16,159 Learners", and "3750 XP".



### Bayesian Machine Learning in Python: A/B Testing

Data Science, Machine Learning, and Data Analytics Techniques for Marketing, Digital Media, Online Advertising, and More

Lazy Programmer Inc.

**4.6** ★★★★★ (5,065)

10.5 total hours • 78 lectures • All Levels

Bestseller

### A/B Testing and Experimentation for Beginners

Learn how to improve Landing Pages, Conversion Rate and Marketing ROI with AB Testing and Data Driven Decisions.

Anil Batra

**4.2** ★★★★★ (1,095)

1 total hour • 14 lectures • Beginner

### Complete Course on Product A/B Testing with Interview Guide

AB Testing, Multivariate Testing, Multi-armed Bandit, Mock Interview Questions, R Coding, Statistics, Hypothesis Testing

Preeti Semwal

**4.6** ★★★★★ (352)

2.5 total hours • 27 lectures • All Levels



Más allá del test  
estadístico de  
proporciones

# MUCHOS MÁS TIPOS DE TESTS DE HIPÓTESIS

¿Qué estás testeando?

## Proporciones

### 1 sample z-test

Comparar la proporción en 1 muestra contra un valor esperado

### 2 sample z-test

Comparar la proporción en 1 muestra A contra otra muestra B

## Medias

### 1 sample t-test

Comparar la media de 1 muestra contra un valor esperado

### 2 sample t-test

Comparar la media de una muestra A contra la media de otra muestra B

2 sample paired t-test  
Comparar la media de 2 muestras pero de la misma población

## Frecuencias / Conteos

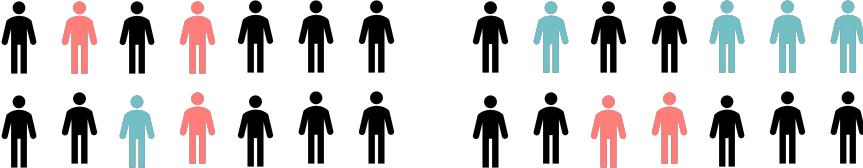
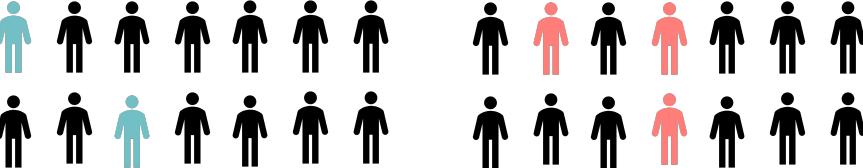
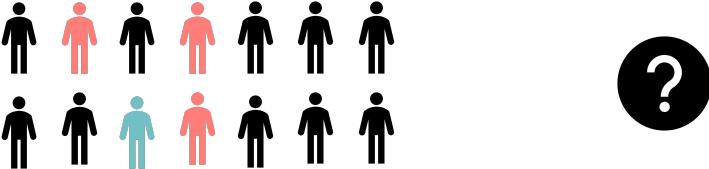
### Chi-squared goodness-of-fit

Comparar las frecuencias de ciertas variables en 1 muestra contra un distribución esperada

### Chi-squared homogeneity

Comparar las frecuencias de ciertas variables en entre dos muestras

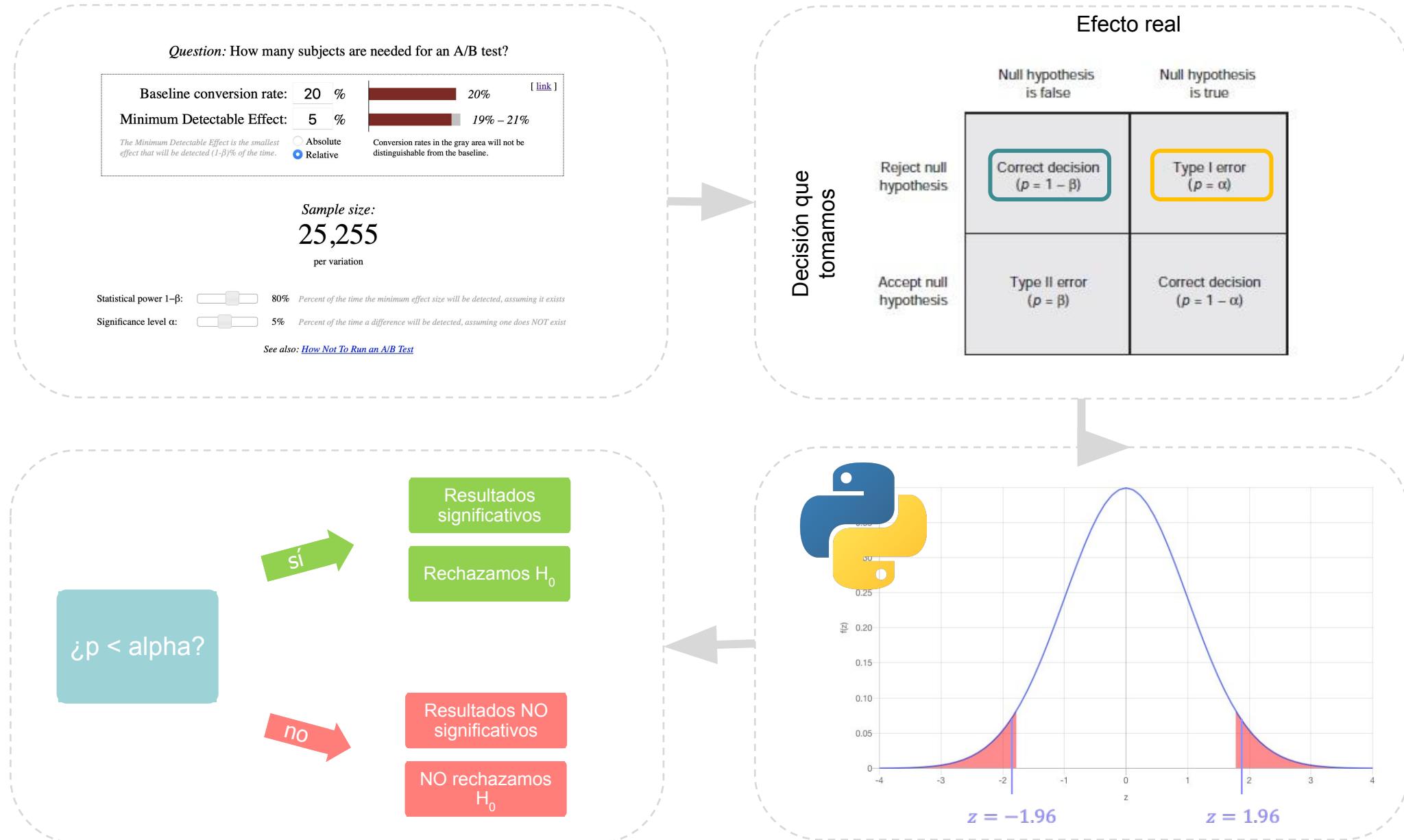
# Y EXISTEN MUCHOS TIPOS DE EXPERIMENTACIÓN

Experimento		Control y variante son idénticos <i>Biología, Física</i>
Experimento estadístico		Control y variante NO son idénticos, pero están divididos de forma aleatoria. <i>A/B testing, análisis Bayesiano</i>
Quasi-experiment		Control y variante NO son idénticos, y están divididos de forma 'natural'. <i>Differences-in-differences, regression discontinuity, matching, etc</i>
Counter-factuals		No existe un grupo de control. Para entender posibles diferencias se usa un modelo predictive para qué habría pasado sin no hubiera habido el tratamiento. <i>Causal Impact, Synthetic Differences-in-differences</i>

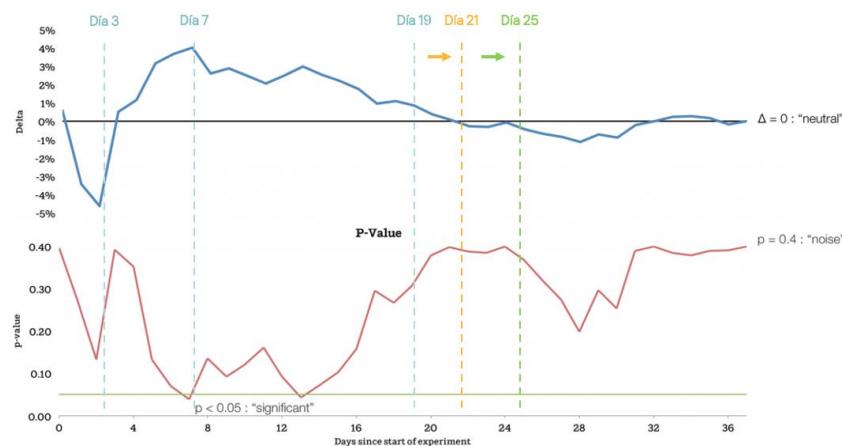


# Repaso del día 1





## P-eaking



## Diseño de métrica



¿Qué métrica escogeríais para diseñar y evaluar vuestro experimento?

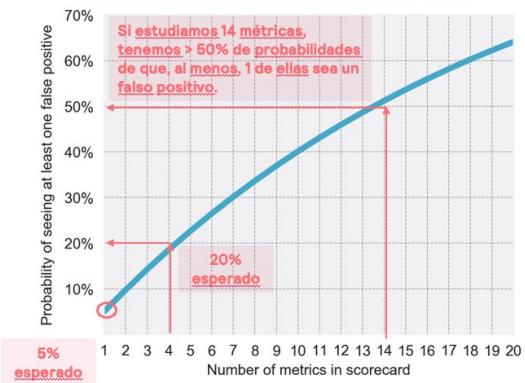
- a) Transaction/Session = 2/3
- b) Transaction/User = 2/1
- c) Unique transaction/user=1/1

## Segmentos

Browser	$\Delta$	p
All	-0.27%	0.29
Chrome	<b>2.07%</b>	0.01
Firefox	<b>2.81%</b>	0.00
IE	<b>-3.66%</b>	0.00
Safari	0.86%	0.26
Rest	-0.74%	0.33

## Múltiples métricas

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

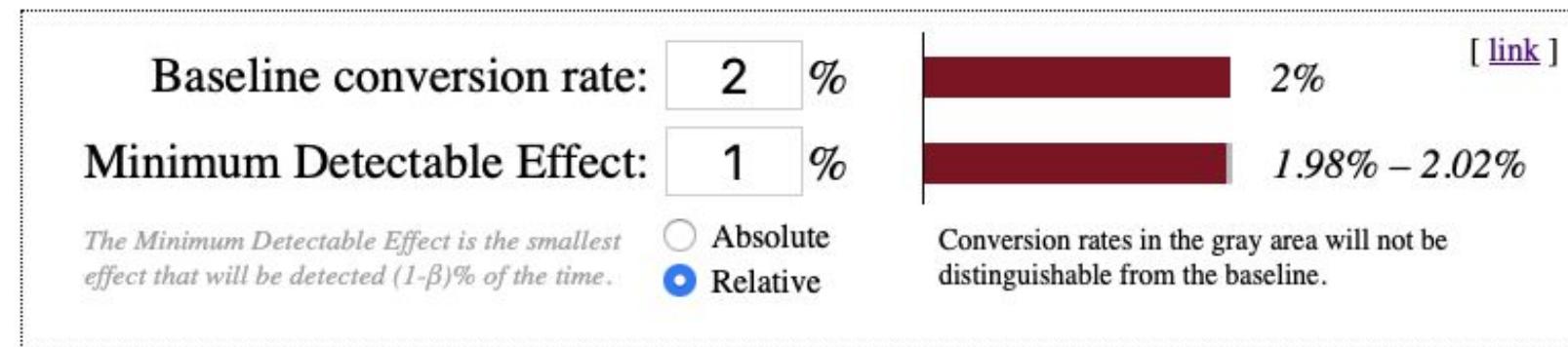


A brown teddy bear with a textured, shaggy fur sits on a light-colored wooden surface, facing towards the right. The bear's body is angled slightly, with its front paws resting on the surface. The background is a soft-focus wooden wall.

## Parte 2. Debilidades del método frecuentista

Experimentos con tasas de éxitos pequeñas pueden ser inviables por el volumen requerido.

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



*Sample size:*

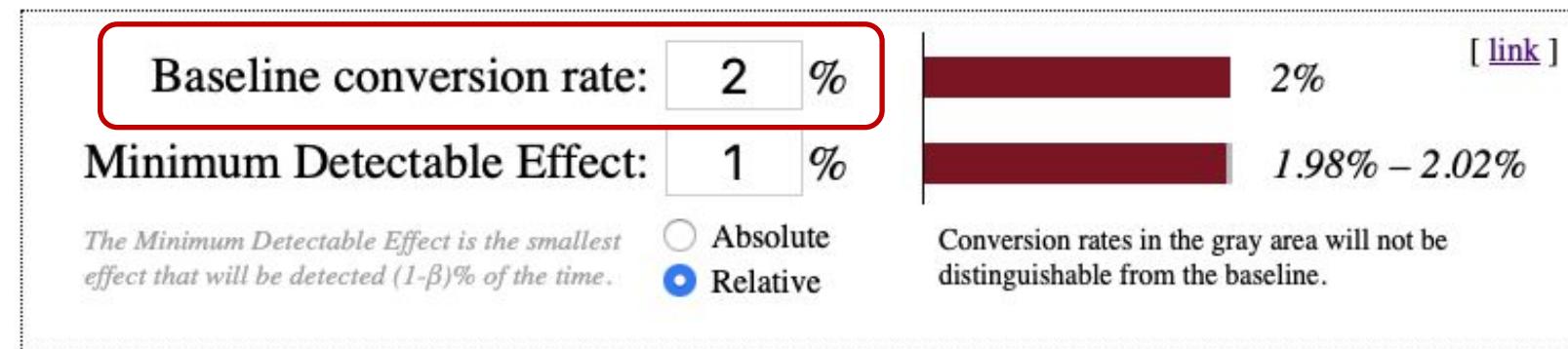
**7,703,208**

per variation

# ADAPTABILIDAD

Conocimientos previos sobre nuestro producto sólo intervienen en el cálculo del tamaño de muestra.

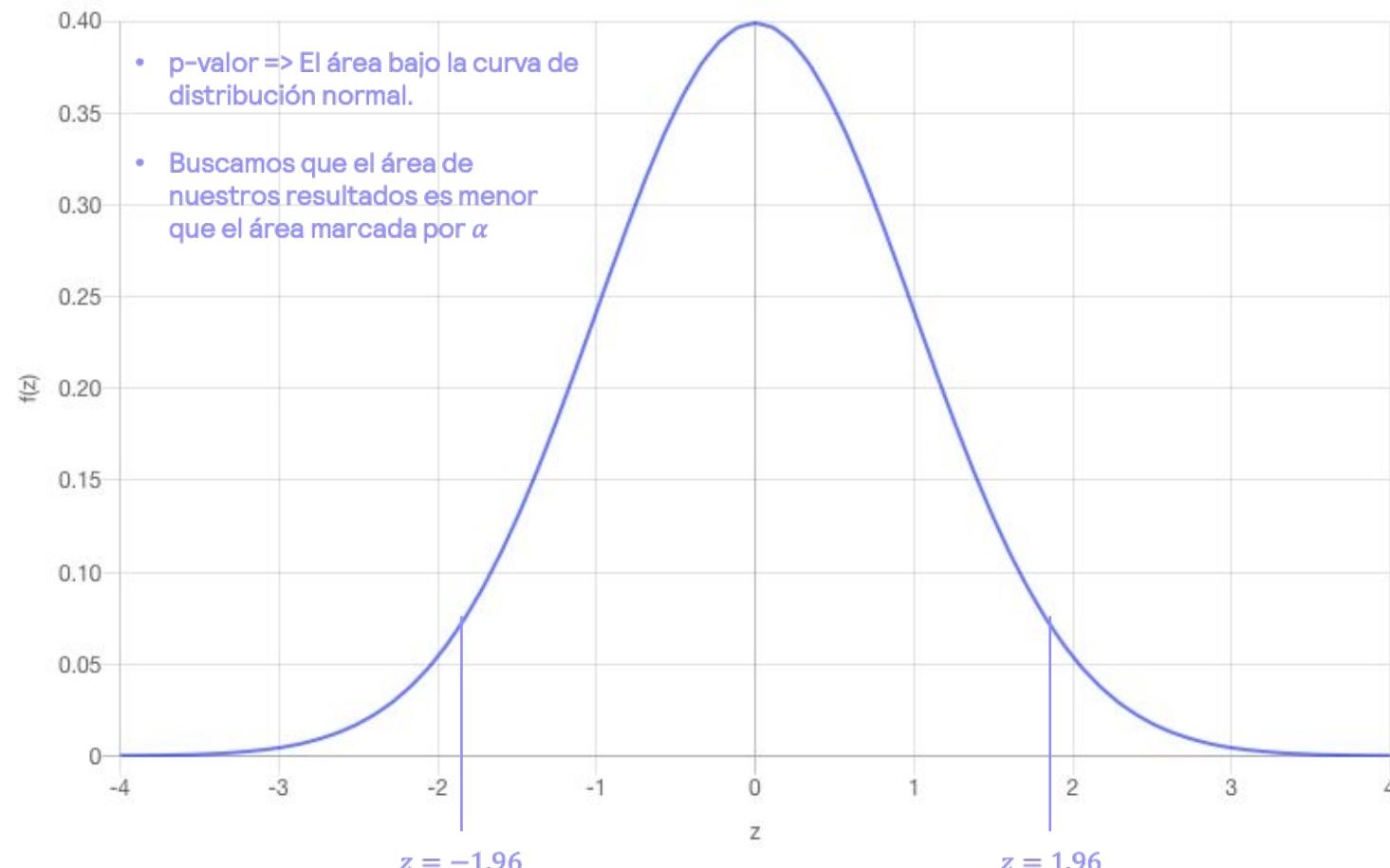
*Question:* How many subjects are needed for an A/B test?



*Sample size:*  
**7,703,208**  
per variation

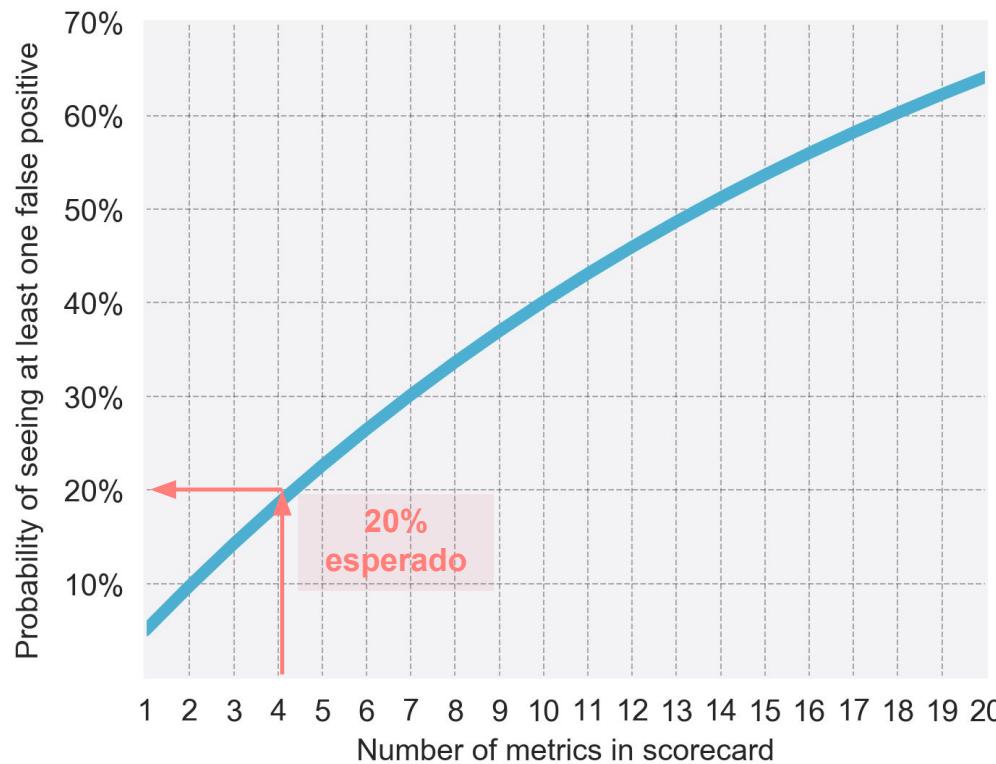
# RIGIDEZ

$\alpha = 0.05$ , es práctica estándar, pero, ¿por qué no  $\alpha = 0.049$  o  $\alpha = 0.051$ ?

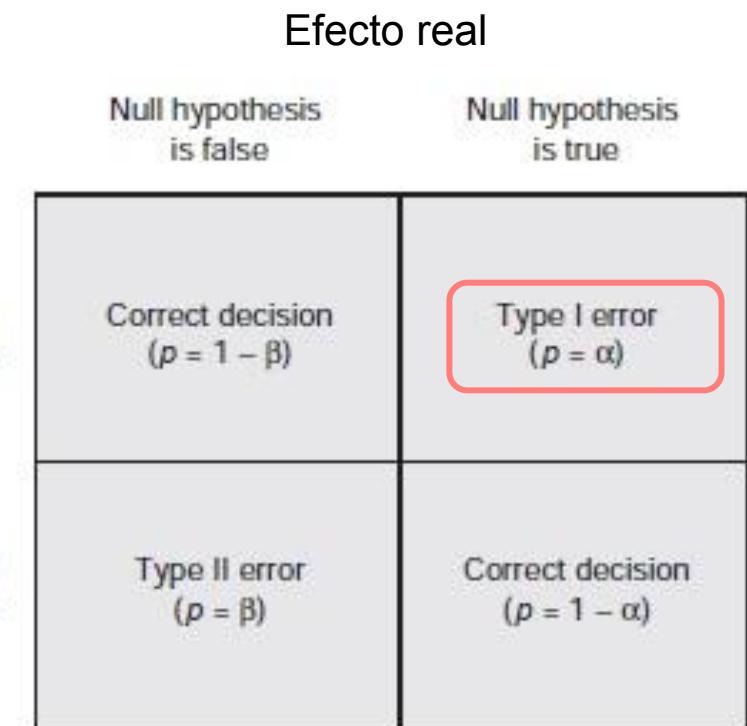


# MÚLTIPLES MÉTRICAS: CORRECCIONES

Hay que tener cuidado para no incurrir en errores de tipo I

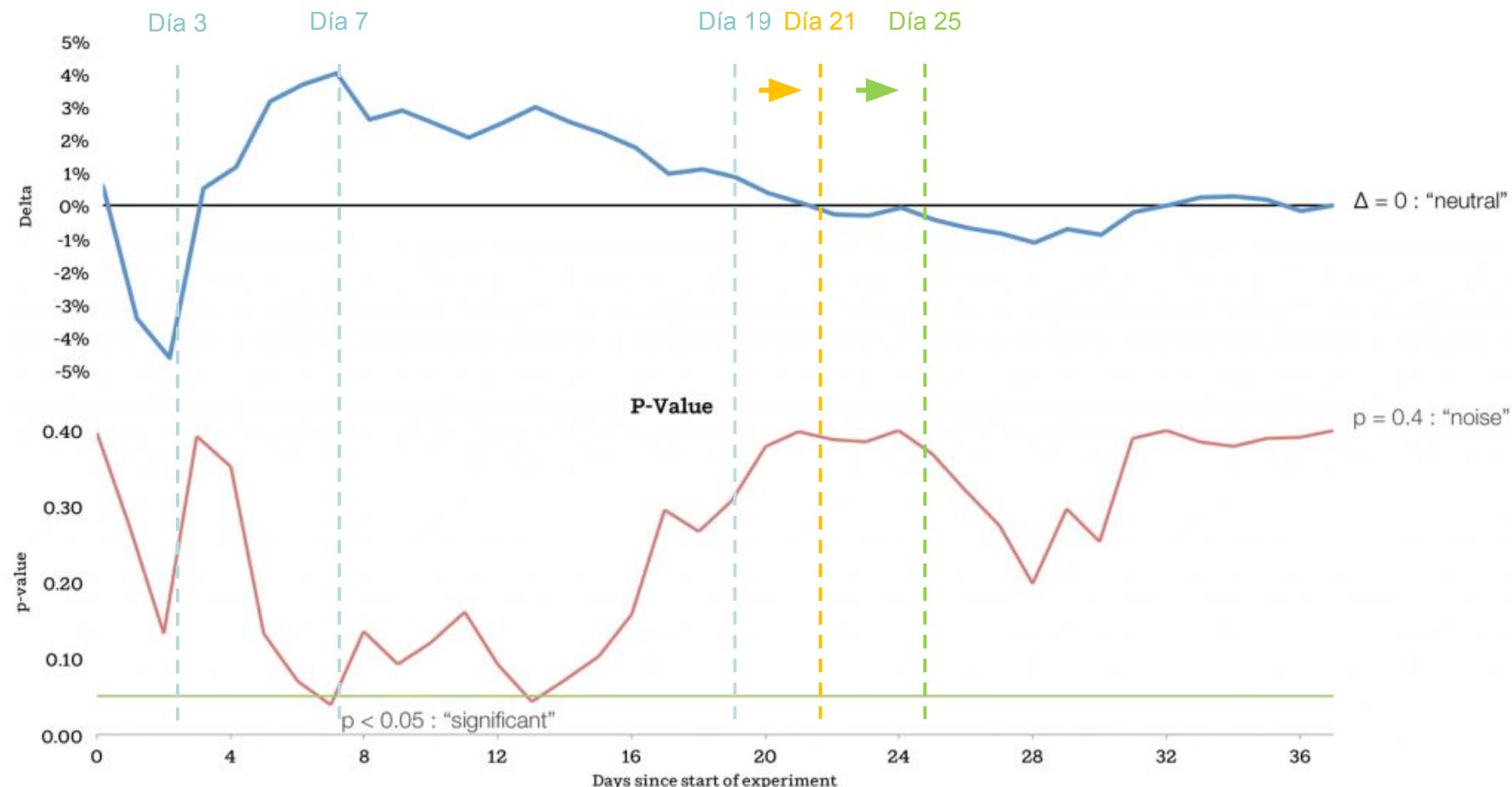


Decisión que tomamos



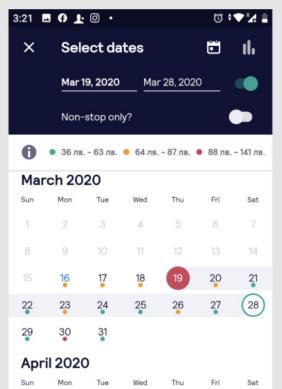
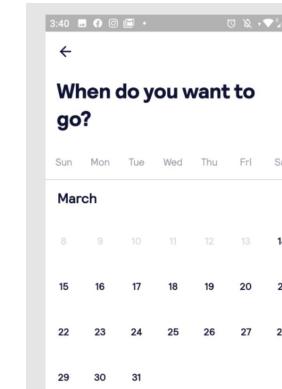
# P-EAKING

La tentación de mirar resultados antes de tiempo puede ser perjudicial



# INTERPRETABILIDAD

- Si es significativo, genial (aunque el p-valor ya es de por si, difícil de explicar)
- Si es no significativo, simplemente, no puedes sacar ninguna conclusión sobre los resultados

	Caso 1	Caso 2
Tasa de conversión	4.99%	5.24%
p-valor	0.0114	0.0758
alpha	0.05	0.05
Decisión	 <span style="border: 1px solid green; border-radius: 50%; padding: 2px;">Rechazar la <math>H_0</math></span>	 <span style="border: 2px solid red; border-radius: 50%; padding: 2px;">No podemos rechazar <math>H_0</math></span>

Por una diferencia de 0.07%, en el caso 2, nuestro nuevo calendario ya no vale... ¡ja ver cómo le explicas esto a un Product Owner!!



# Parte 3. A/B testing con análisis Bayesiano

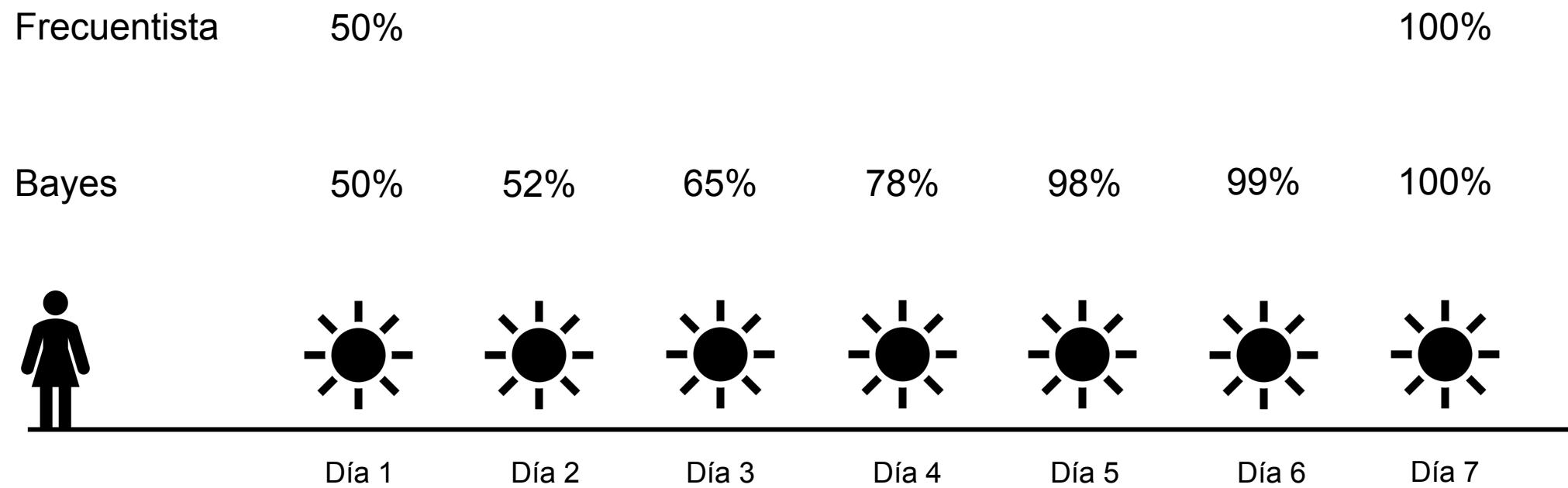
- Intuición Bayesiana
- Distribución beta
- Análisis de resultados

- Intuición Bayesiana
- Distribución beta
- Análisis de resultados

# INTUICIÓN BAYESIANA

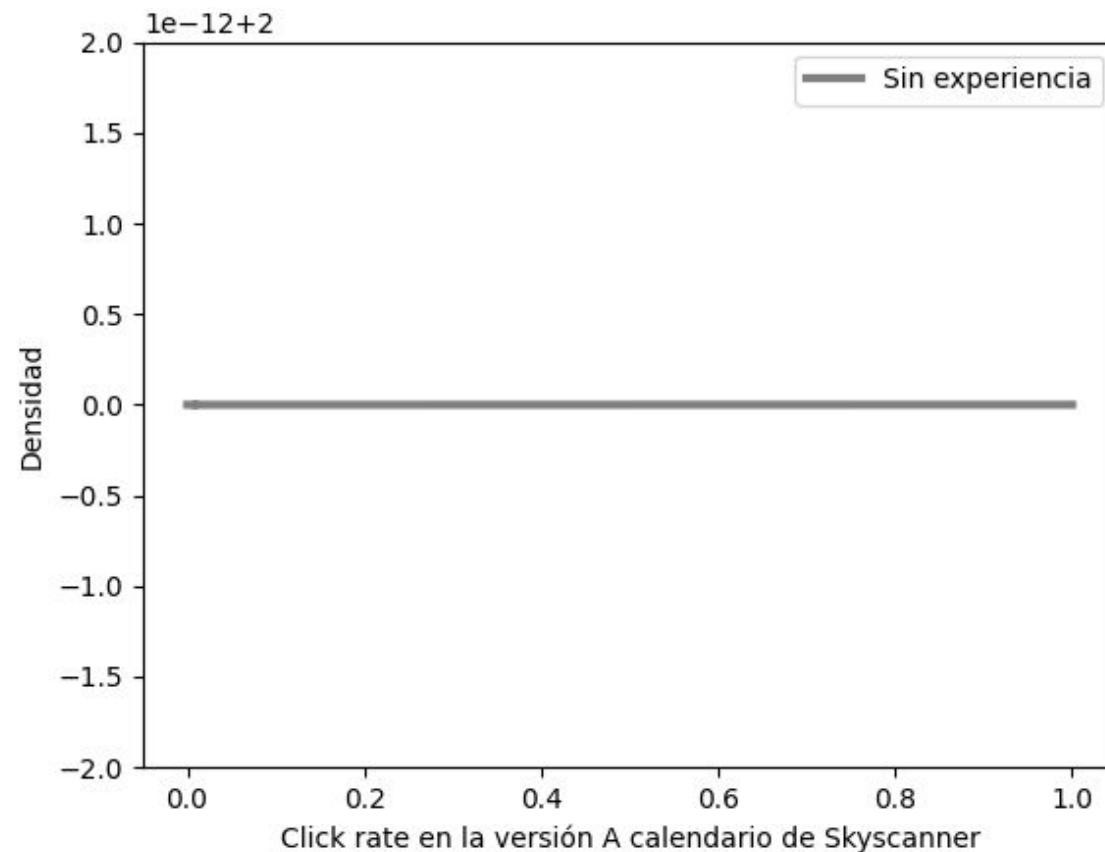
Los humanos pensamos como ‘Bayes’, no como ‘Frecuentistas’

*Aprendemos con cada nueva experiencia, no con ‘batches’ de experiencias*



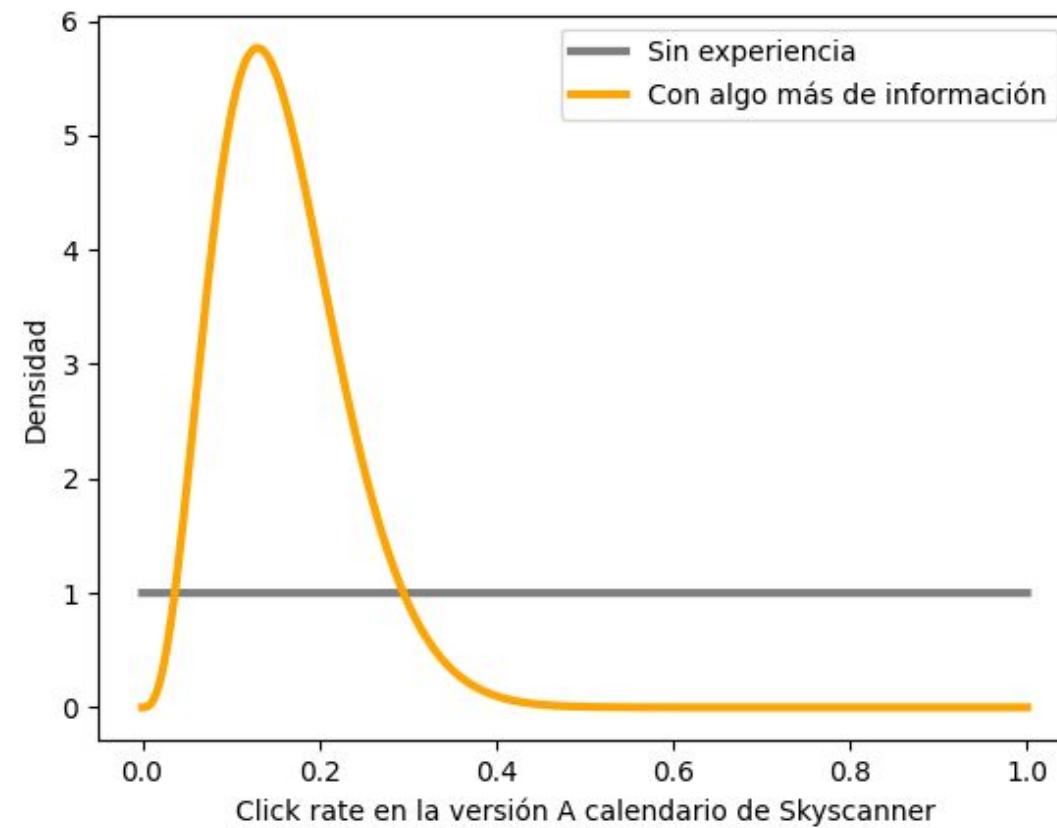
# ¿CÓMO SE MODELA UNA PROBABILIDAD?

Con una distribución



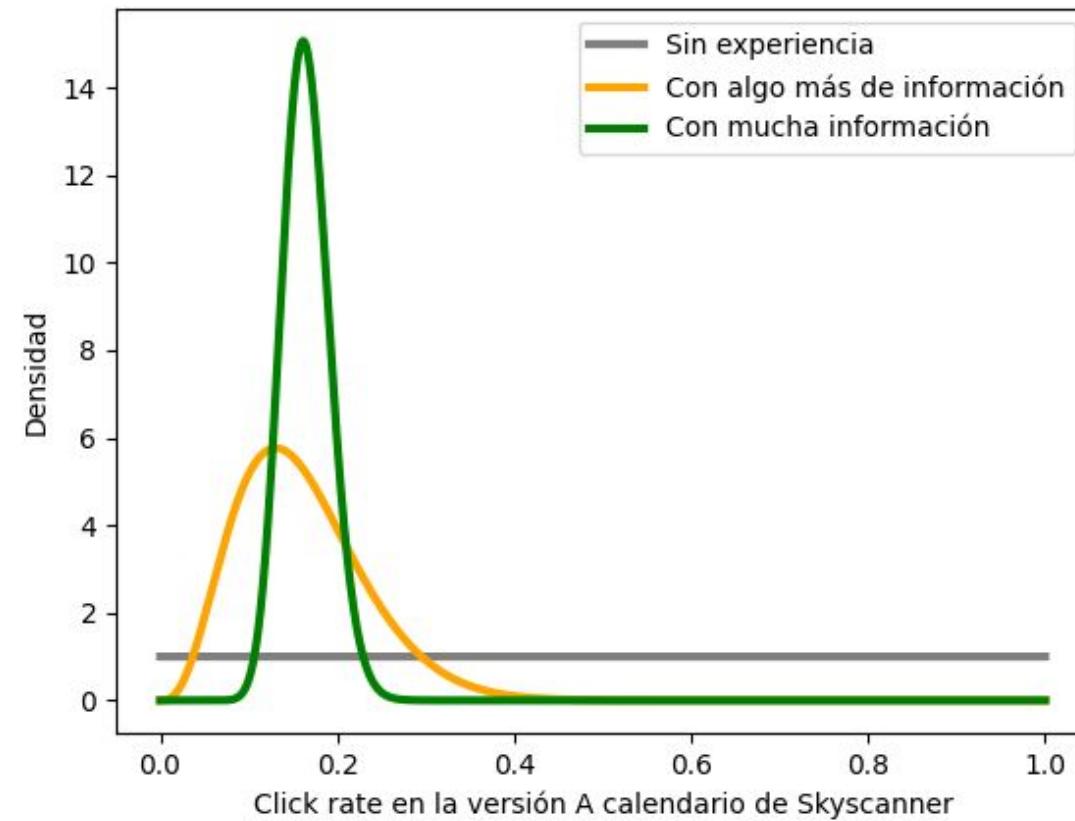
# ¿CÓMO SE MODELA UNA PROBABILIDAD?

Con una distribución

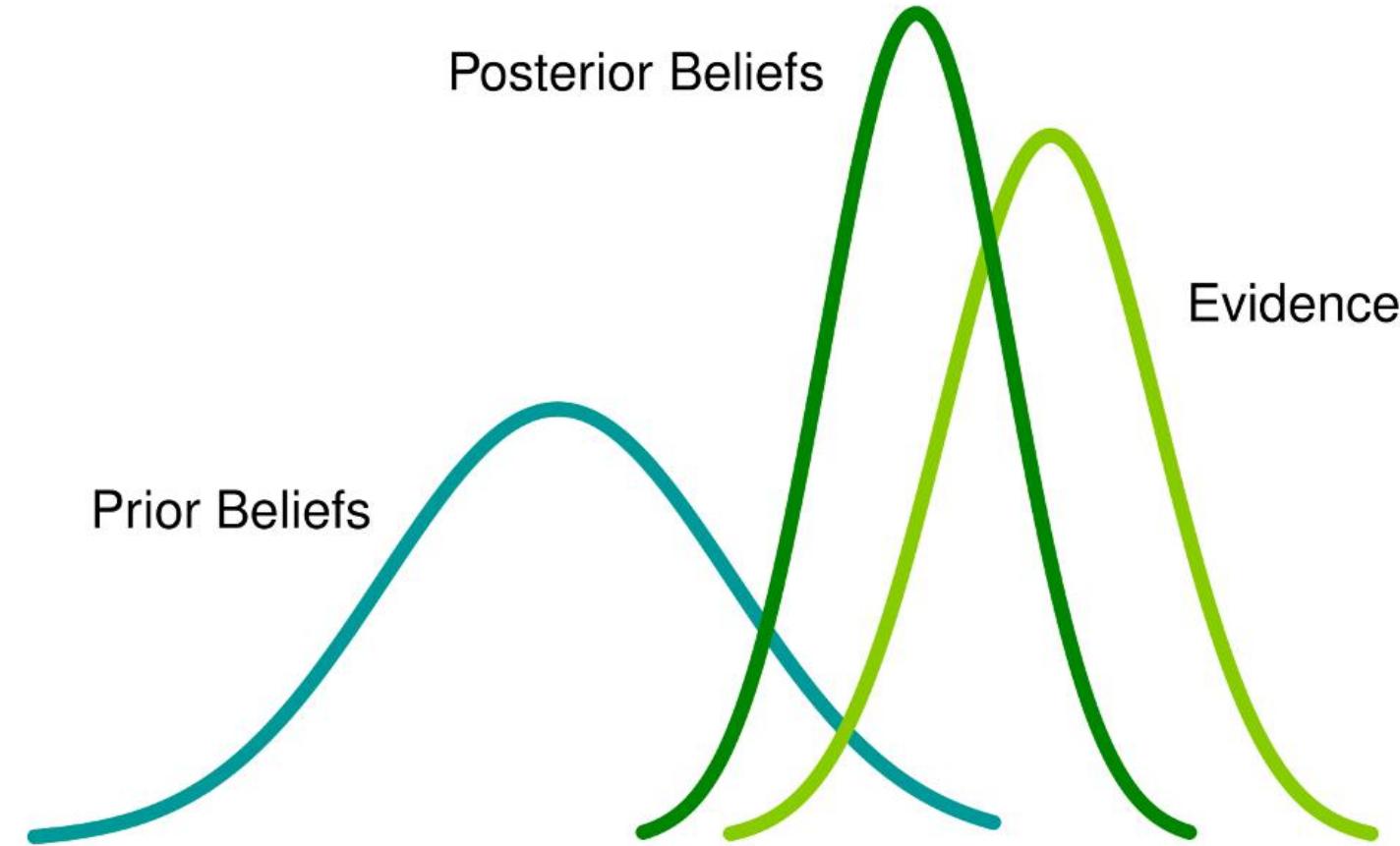


# ¿CÓMO SE MODELA UNA PROBABILIDAD?

Con una distribución



# BAYES ACTUALIZA SU CONOCIMIENTO



- Intuición Bayesiana
- **Distribución beta**
- Análisis de resultados

# ¿QUÉ ES LA DISTRIBUCIÓN BETA?

Una distribución que es capaz de modelar probabilidades, ya que trabaja en el rango de [0,1]

$$\frac{x^{(\alpha-1)}(1-x)^{(\beta-1)}}{B(\alpha, \beta)}$$

siendo:

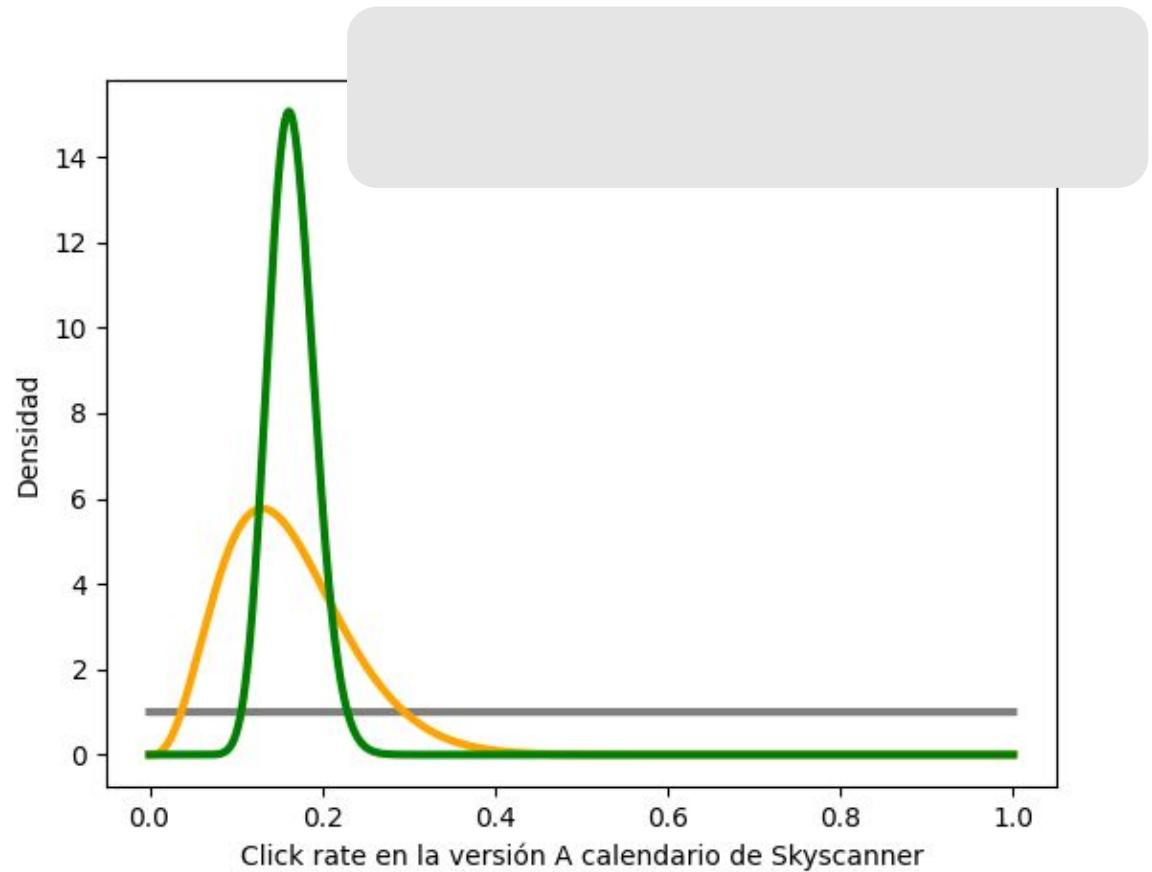
- $\alpha - 1$  → número de éxitos
- $\beta - 1$  → número de fracasos
- $B(\alpha, \beta)$  → función beta

# EFEKTOS DE $\alpha$ Y $\beta$ EN LA DISTRIBUCIÓN BETA

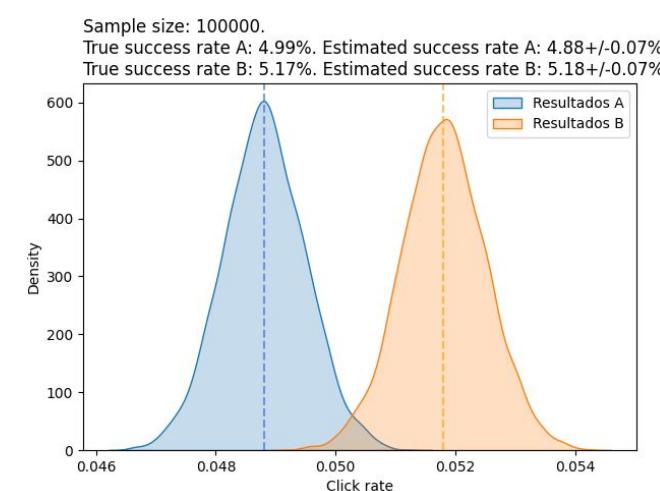
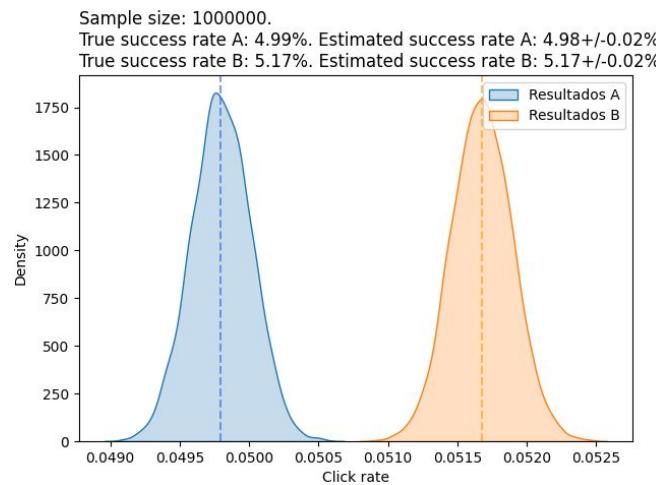
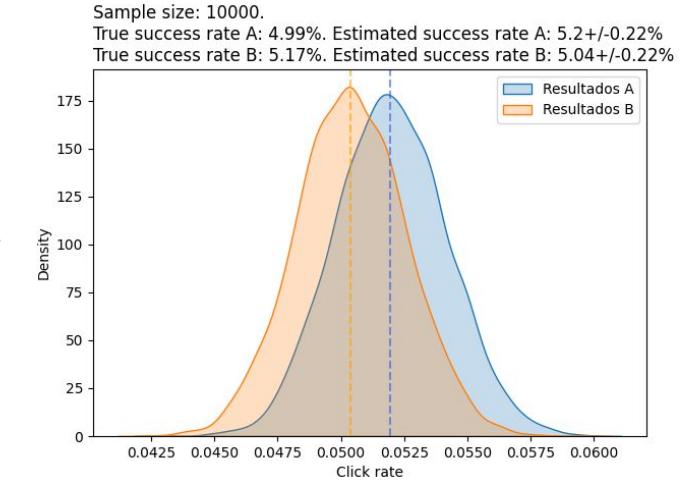
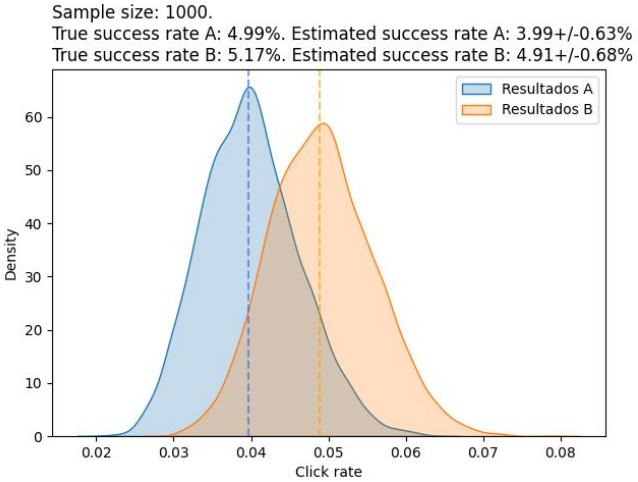
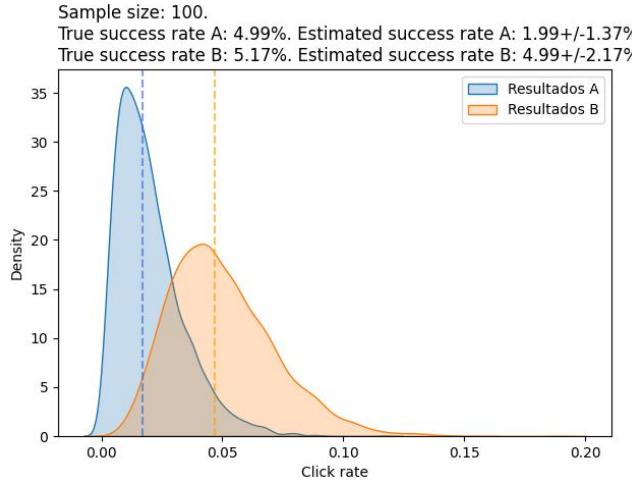
$\alpha - 1 \rightarrow$  número de éxitos,  $\beta - 1 \rightarrow$  número de fracasos

$$\frac{x^{(\alpha-1)}(1-x)^{(\beta-1)}}{B(\alpha, \beta)}$$

- Si  $\uparrow \alpha, \uparrow \beta \equiv$  la distribución se vuelve más espigada
  - Sube la confianza en el valor de % éxito
- Si  $\uparrow \alpha \equiv$  la distribución se mueve a la derecha
  - % de éxito sube + espigada
- Si  $\uparrow \beta \equiv$  la distribución se mueve a la izquierda
  - % de éxito baja + espigada



# SIMULACIÓN QUE MUESTRA EL EFECTO DE IR RECOGIENDO MÁS DATOS



- Intuición Bayesiana
- Distribución beta
- **Análisis de resultados**

# PASO 1: SIMULACIONES MONTECARLO

1. Obtienes número éxitos y fracasos (resultados del experimento).
2. Introduces ese valor para crear una distribución Beta.
3. Aleatoriamente extraemos un valor de esta distribución Beta miles de veces (simulación MonteCarlo)
4. Puedes calcular diferencias entre las muestras simuladas de A y de B

```
beta_distribution_A = beta(A_exitos+1, A_fracasos+1) ① ②
beta_distribution_B = beta(B_exitos+1, B_fracasos+1)

# Extraemos n_muestras de nuestras distribuciones
n_muestras = 50_000
A_muestras = pd.Series(beta_distribution_A.rvs() for _ in range(n_muestras))
B_muestras = pd.Series(beta_distribution_B.rvs() for _ in range(n_muestras))

# Para que trabajéis mejor y visualicéis, lo ponemos en un dataframe
resultados_df = pd.DataFrame({'resultados_A': A_muestras,
                               'resultados_B': B_muestras})
resultados_df['B_mejor_que_A'] = resultados_df['resultados_B'] - resultados_df['resultados_A']
resultados_df['B_mejor_que_A_rel'] = resultados_df['resultados_B']/resultados_df['resultados_A']
```

	resultados_A	resultados_B	B_mejor_que_A	B_mejor_que_A_rel
0	0.016141	0.023410	0.007269	1.450325
1	0.028795	0.006214	-0.022581	0.215807
2	0.020334	0.010847	-0.009487	0.533423
3	0.056110	0.036481	-0.019630	0.650160
4	0.020486	0.034953	0.014467	1.706187

# PASO 2: VISUALIZACIÓN DE RESULTADOS

Caso 1

Tamaño de muestra	100,123	100,133
Conversiones	5,000	5,250
Tasa de conversión	4.99%	5.24%

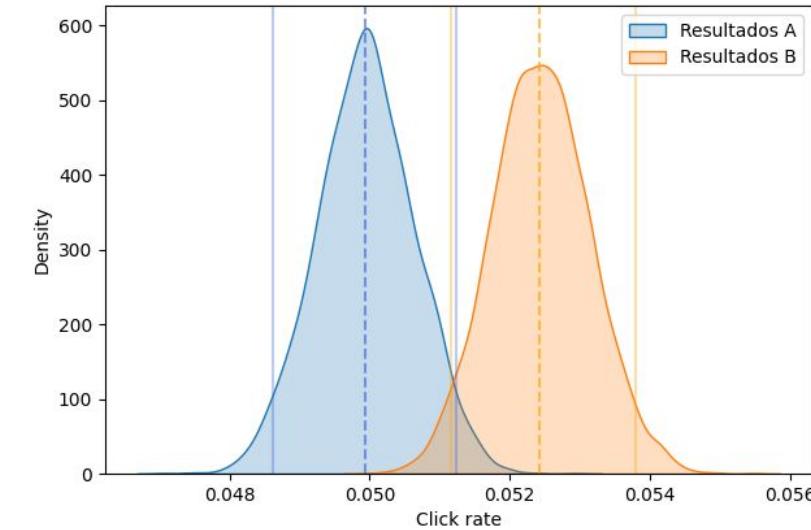
```
beta_distribution_A = beta(A_exitos+1, A_fracasos+1) ① ②
beta_distribution_B = beta(B_exitos+1, B_fracasos+1)

# Extraemos n_muestras de nuestras distribuciones
n_muestras = 50_000
A_muestras = pd.Series(beta_distribution_A.rvs() for _ in range(n_muestras)) ③
B_muestras = pd.Series(beta_distribution_B.rvs() for _ in range(n_muestras))

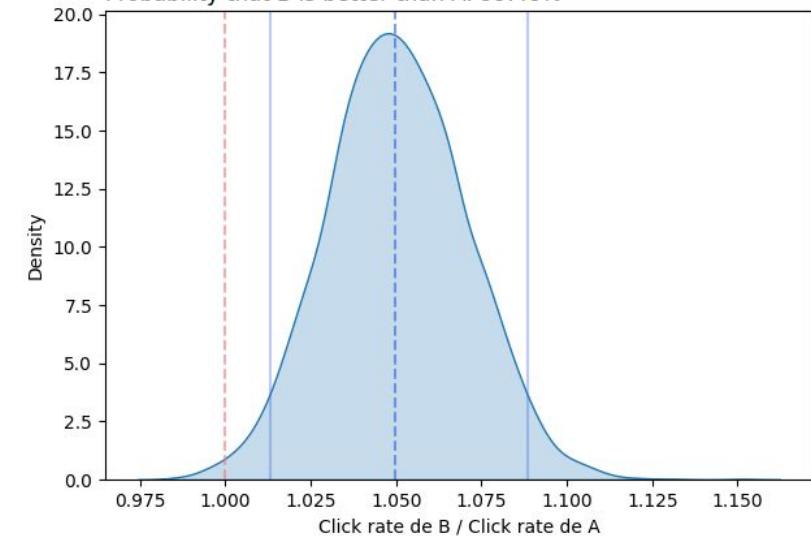
# Para que trabajéis mejor y visualicéis, lo ponemos en un datafram
resultados_df = pd.DataFrame({'resultados_A': A_muestras,
                               'resultados_B': B_muestras}) ④

resultados_df['B_mejor_que_A'] = resultados_df['resultados_B'] - resultados_df['resultados_A']
resultados_df['B_mejor_que_A_rel'] = resultados_df['resultados_B']/resultados_df['resultados_A']
```

Sample size A: 100123, Sample size B: 100133.  
Estimated success rate A: 4.99 +/- 0.07%  
Estimated success rate B: 5.24 +/- 0.07%



Sample size A: 100123, Sample size B: 100133.
Estimated difference of B/A: 1.05 +/- 0.02%
Probability that B is better than A: 99.48%



# PASO 2: VISUALIZACIÓN DE RESULTADOS

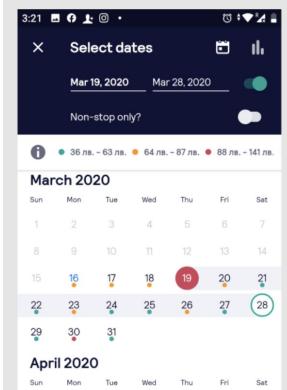
Caso 1

Tamaño de muestra	100,123	100,133
Conversiones	5,000	5,250
Tasa de conversión	4.99%	5.24%

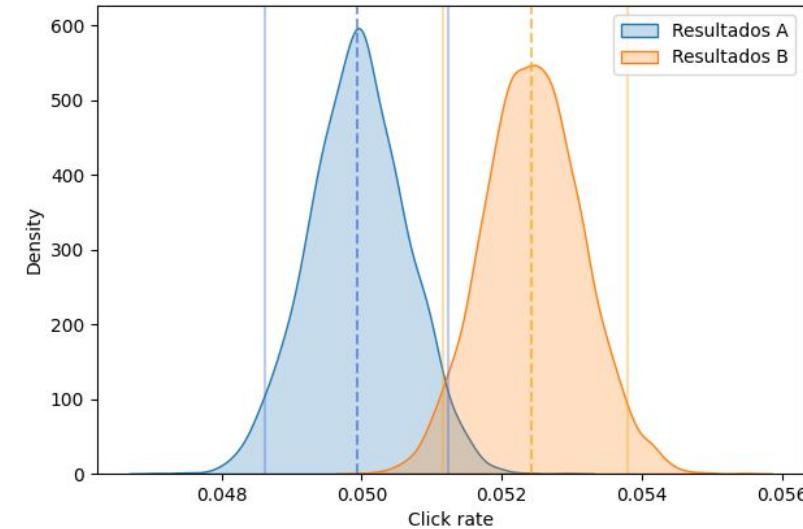
p-valor 0.0114

alpha 0.05

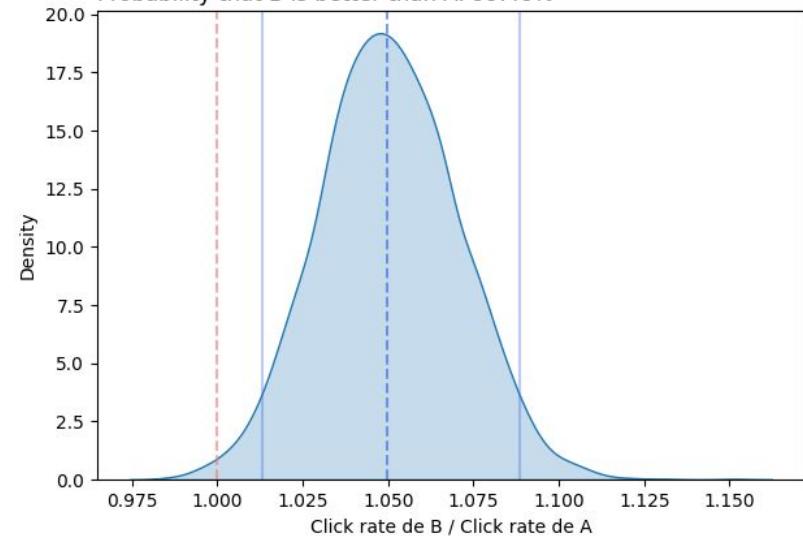
Decisión Rechazar la  $H_0$



Sample size A: 100123, Sample size B: 100133.  
Estimated success rate A: 4.99% +/- 0.07%  
Estimated success rate B: 5.24% +/- 0.07%



Sample size A: 100123, Sample size B: 100133.  
Estimated difference of B/A: 1.05% +/- 0.02%  
Probability that B is better than A: 99.48%



# PASO 2: VISUALIZACIÓN DE RESULTADOS

## Caso 2

Tamaño de muestra	100,123	100,133
Conversiones	5,000	5,175
Tasa de conversión	4.99%	5.17%

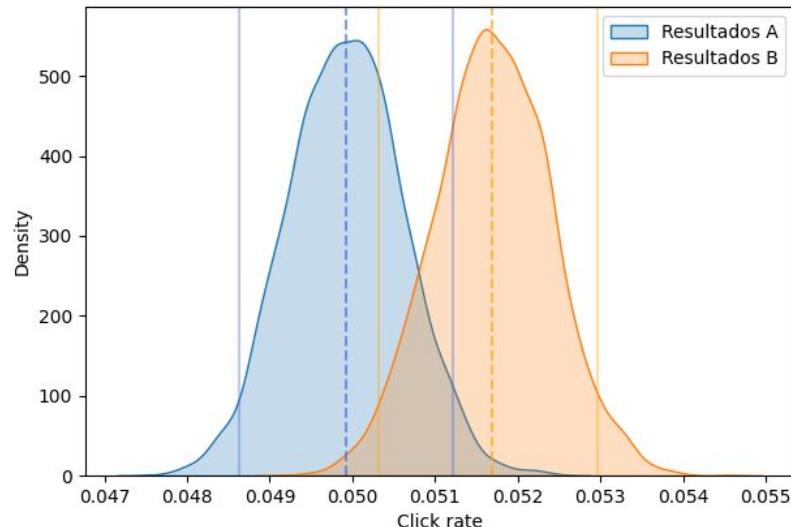
```
beta_distribution_A = beta(A_exitos+1, A_fracasos+1) ① ②
beta_distribution_B = beta(B_exitos+1, B_fracasos+1)

# Extraemos n_muestras de nuestras distribuciones
n_muestras = 50_000
A_muestras = pd.Series(beta_distribution_A.rvs() for _ in range(n_muestras)) ③
B_muestras = pd.Series(beta_distribution_B.rvs() for _ in range(n_muestras))

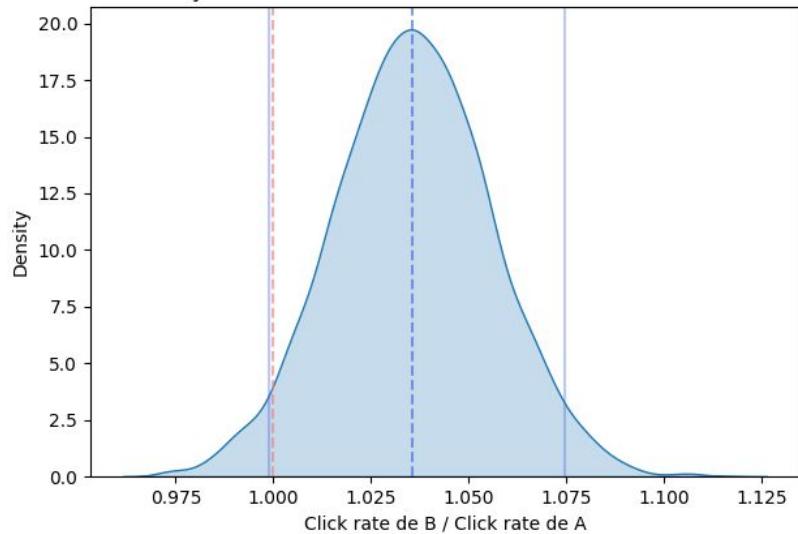
# Para que trabajéis mejor y visualicéis, lo ponemos en un datafram
resultados_df = pd.DataFrame({'resultados_A': A_muestras,
                               'resultados_B': B_muestras}) ④

resultados_df['B_mejor_que_A'] = resultados_df['resultados_B'] - resultados_df['resultados_A']
resultados_df['B_mejor_que_A_rel'] = resultados_df['resultados_B']/resultados_df['resultados_A']
```

Sample size A: 100123, Sample size B: 100133.  
Estimated success rate A: 4.99% +/- 0.07%  
Estimated success rate B: 5.17% +/- 0.07%



Sample size A: 100123, Sample size B: 100133.  
Estimated difference of B/A: 1.04% +/- 0.02%  
Probability that B is better than A: 96.22%



# PASO 2: VISUALIZACIÓN DE RESULTADOS

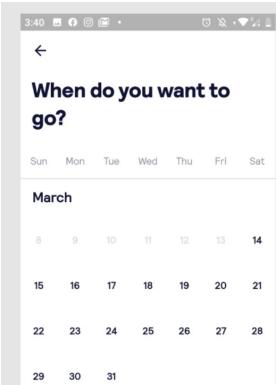
Caso 2

Tamaño de muestra	100,123	100,133
Conversiones	5,000	5,175
Tasa de conversión	4.99%	5.17%

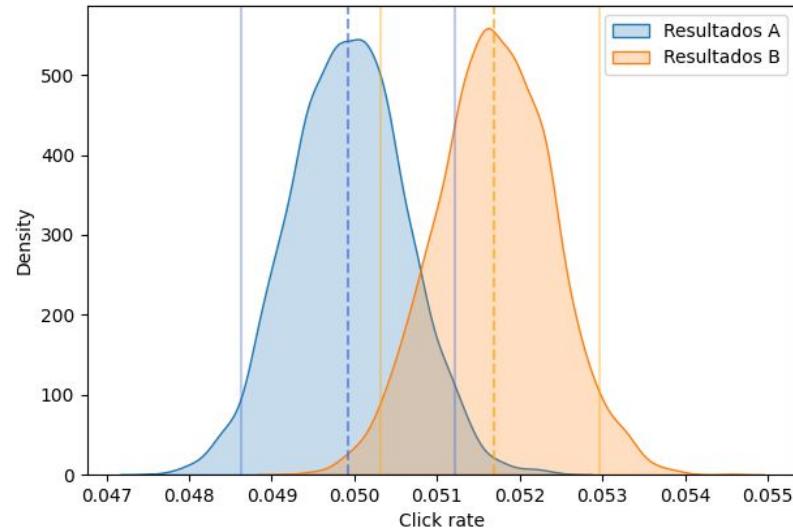
p-valor 0.0758

alpha 0.05

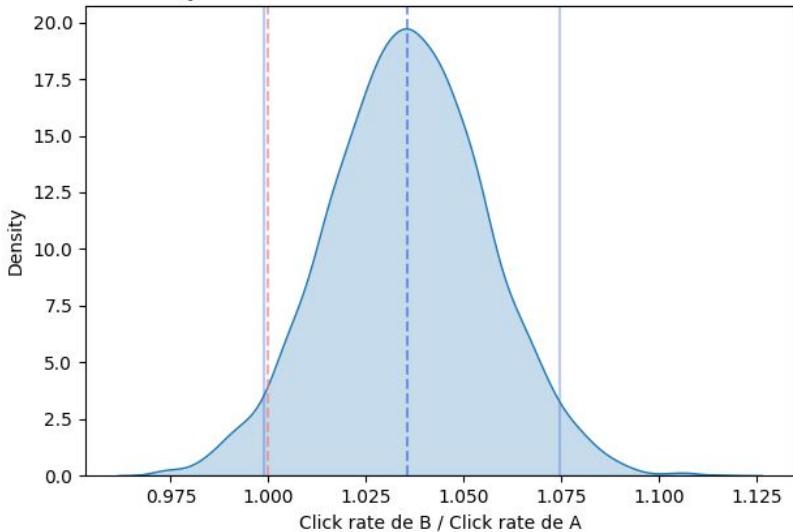
Decisión No podemos rechazar  $H_0$

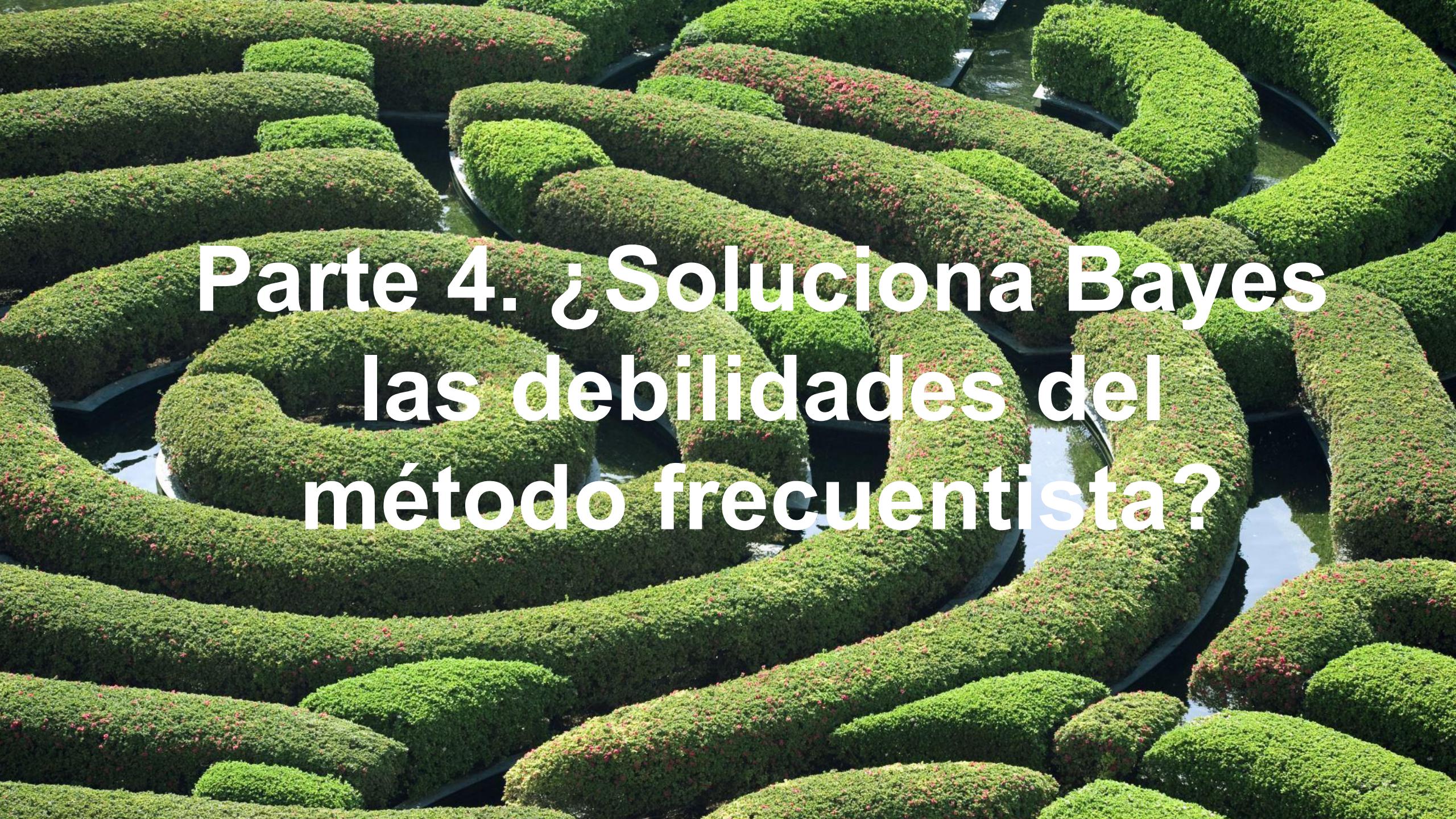


Sample size A: 100123, Sample size B: 100133.  
Estimated success rate A: 4.99% +/- 0.07%  
Estimated success rate B: 5.17% +/- 0.07%



Sample size A: 100123, Sample size B: 100133.  
Estimated difference of B/A: 1.04% +/- 0.02%  
Probability that B is better than A: 96.22%



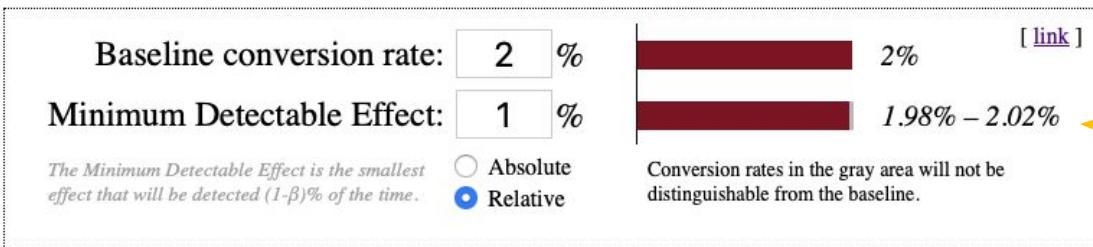


Parte 4. ¿Soluciona Bayes  
las debilidades del  
método frequentista?

# RAPIDEZ

Por lo general, Bayes converge mucho antes que el método frecuentista.

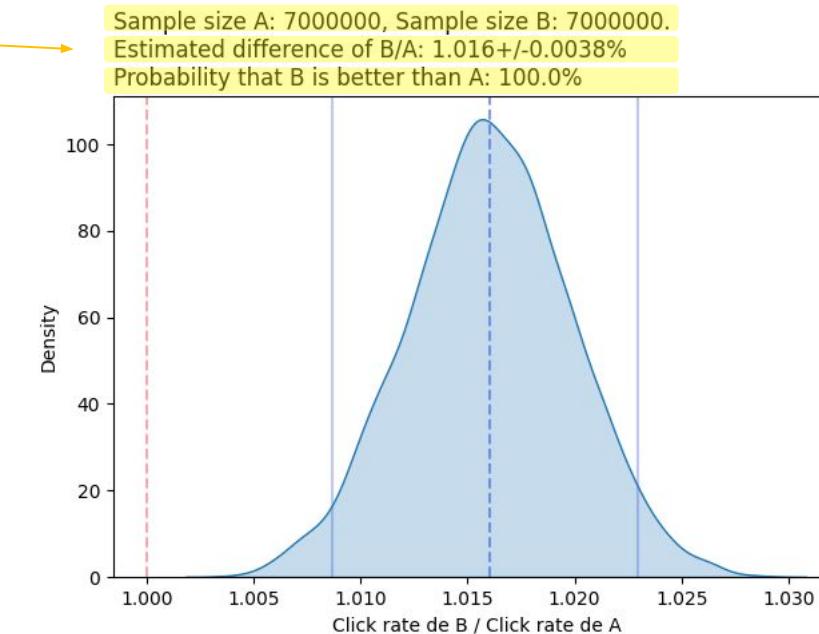
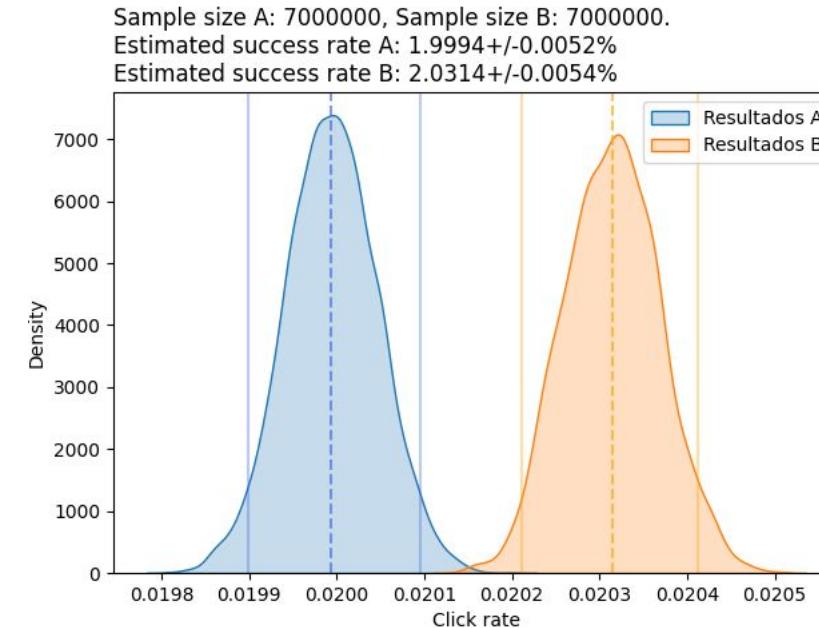
*Question:* How many subjects are needed for an A/B test?



Sample size:

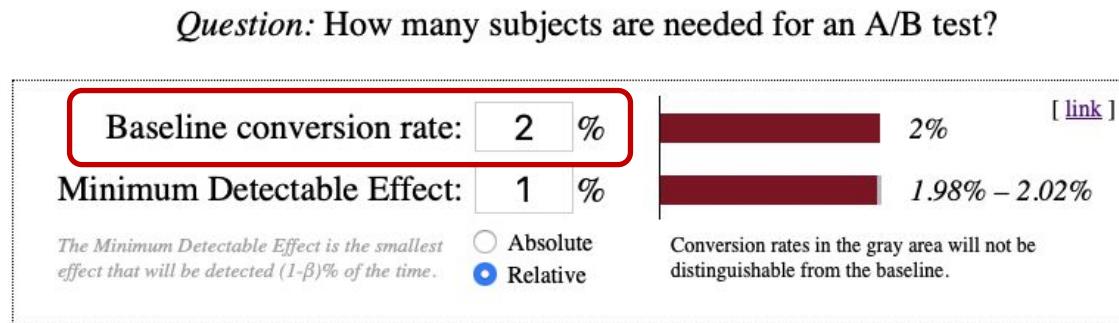
7,703,208

per variation

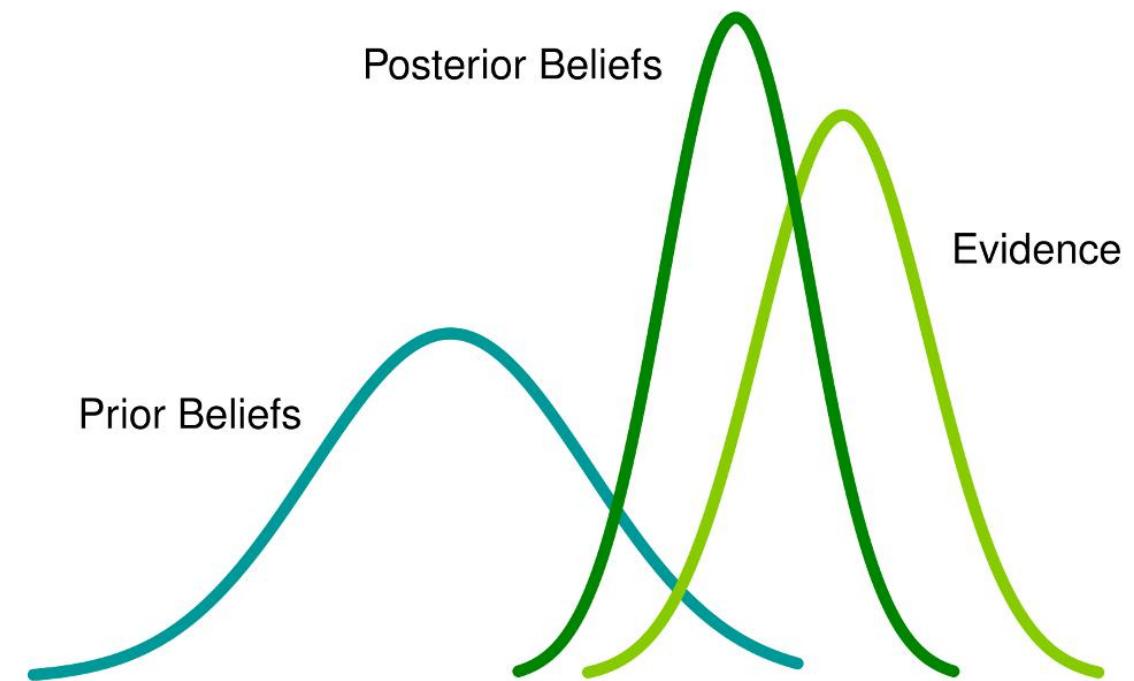


# ADAPTABILIDAD

Bayes actualiza el conocimiento con cada nueva muestra

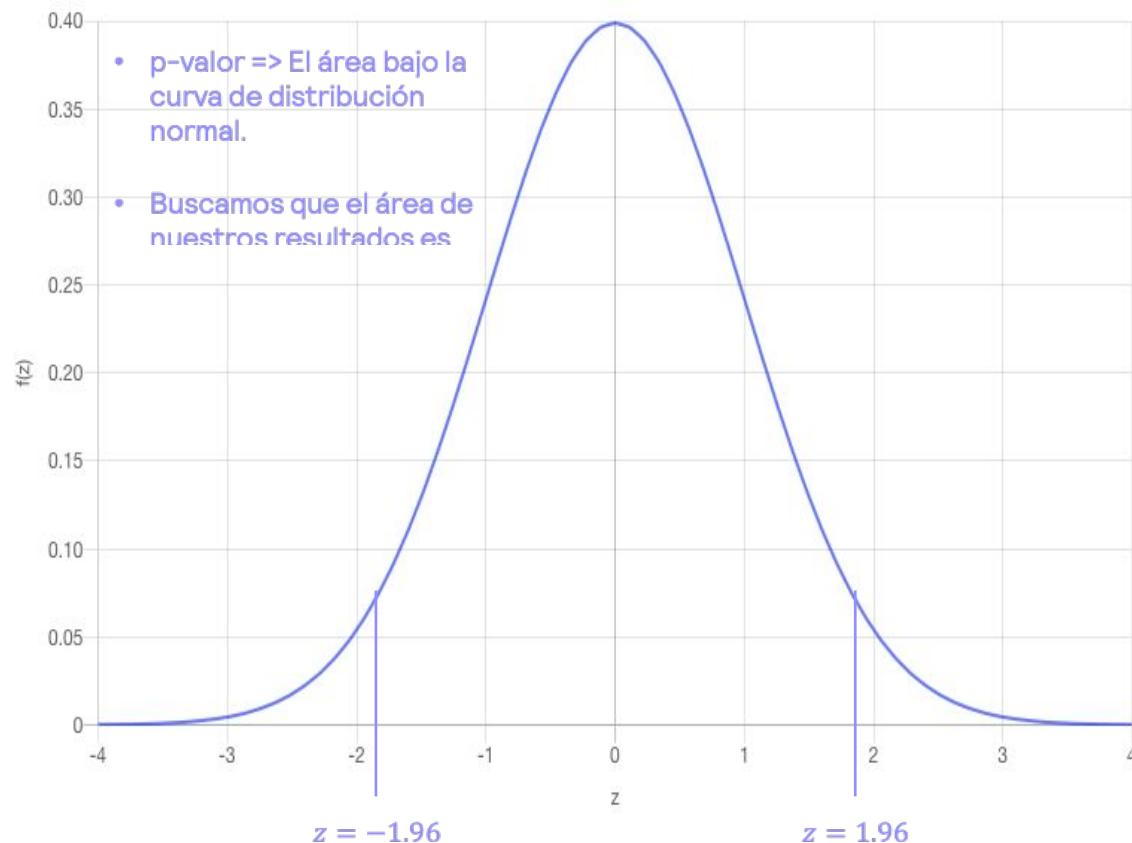


Sample size:  
7,703,208  
per variation



# RIGIDEZ

Para Bayes, no es estrictamente necesario poner límites

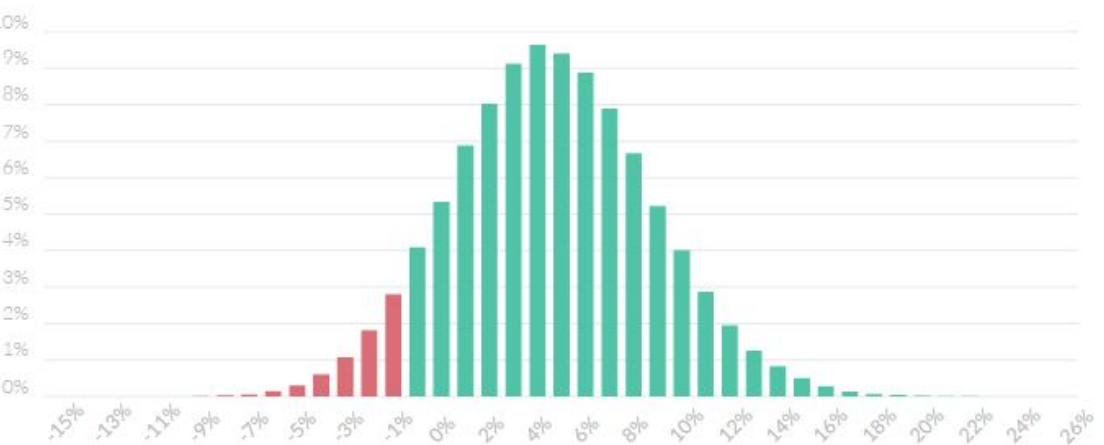


#	Users	Conversion	CR	Uplift	Chance of being best	Chance of at least €100,000 extra revenue
A	30,000	1,200	4.00%			
B	30,000	1,260	4.20%	5.00%	89.1%	84.2%

Based on 7 days of data, on average 30,000 users per variation

## Posterior simulation of difference

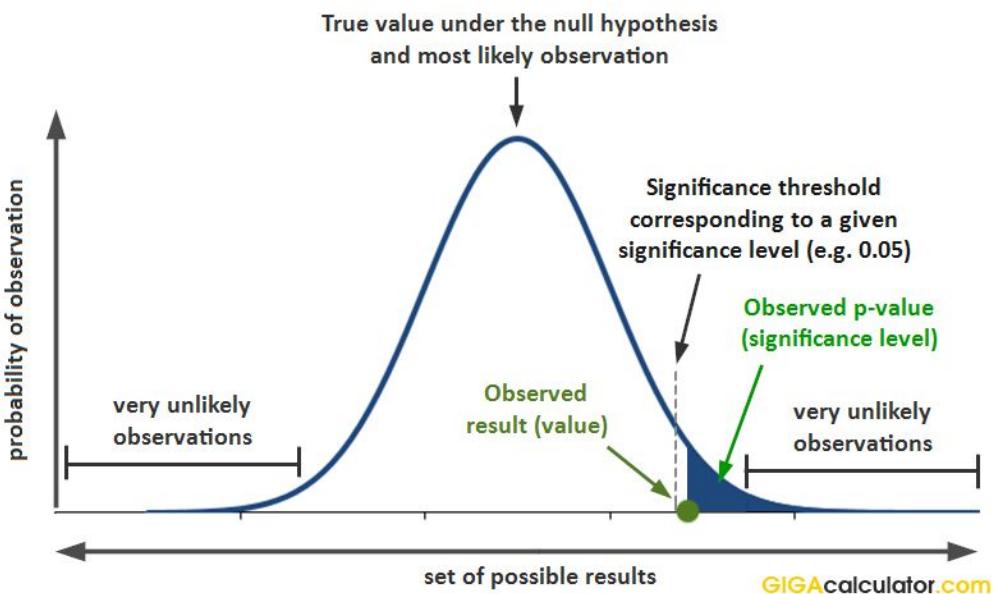
Difference in conversion rate between B and A



# INTERPRETABILIDAD

P-valor => difícil de explicar. Probabilidad => más fácil.

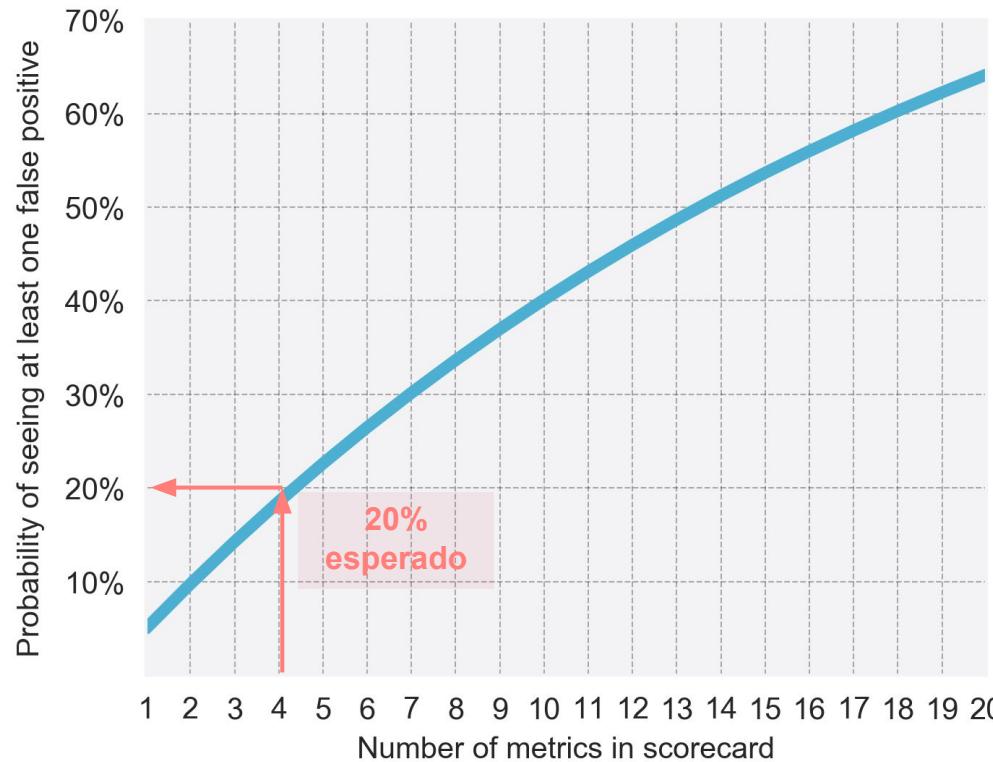
## P-values and statistical significance explained



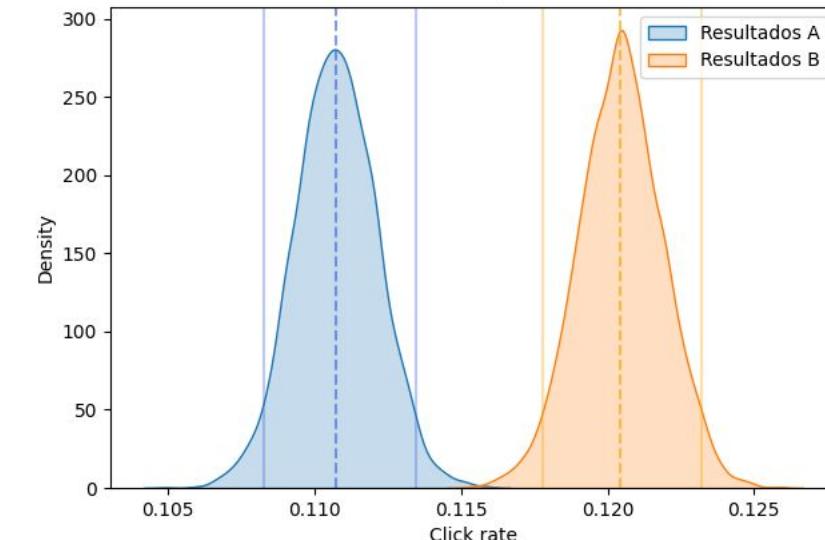
Risk assessment of implementing B		
Implement B	Probability	Effect on revenue
Expected risk	10.9%	-€202,494
Expected uplift	89.1%	€648,501
Based on an average order value of €175 and 6 months time		
		Total contribution €555,334

# CORRECCIONES

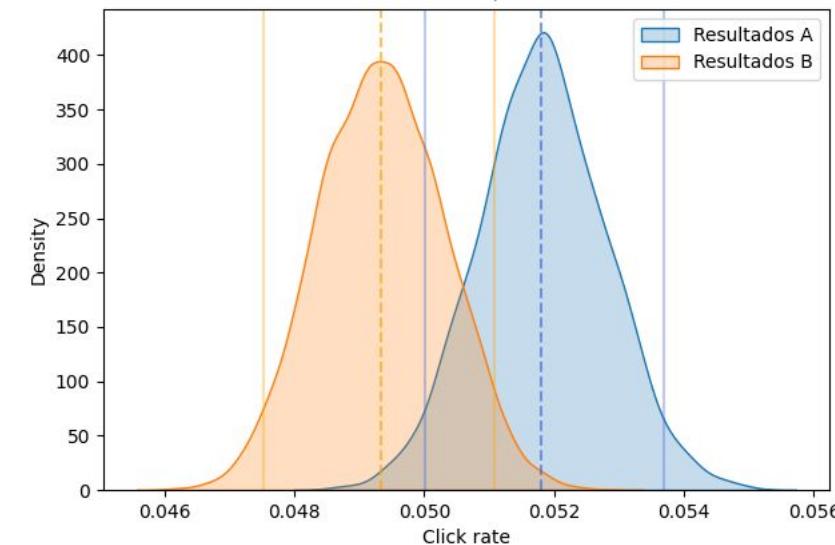
En Bayes, el volumen de datos y los resultados obtenidos, indican la incertidumbre



Sample size A: 50000, Sample size B: 50000.  
Estimated success rate A: 11.0731+/-0.1404%  
Estimated success rate B: 12.044+/-0.1437%

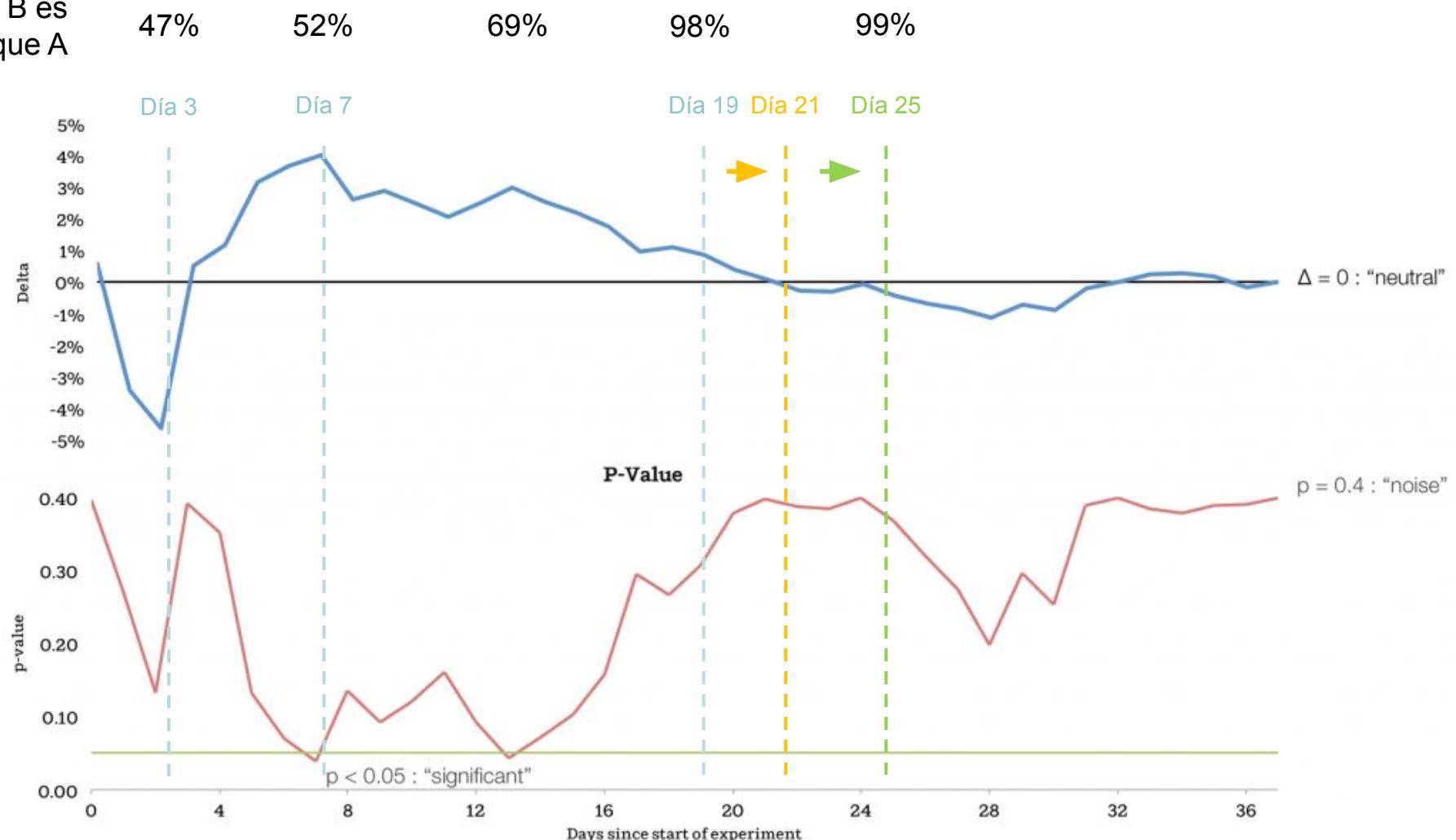


Sample size A: 50000, Sample size B: 50000.  
Estimated success rate A: 5.182+/-0.0981%  
Estimated success rate B: 4.9341+/-0.0965%



# P-EAKING

Probabilidad de que B es mejor que A



A close-up photograph of a person's hands holding two fruits against a dark background. The left hand holds a bright orange, and the right hand holds a shiny red apple. The hands are positioned to frame the central text.

# Parte 5. Bayes vs Frecuentista

# COMPARACIÓN

	Frecuentista	Bayesiano
Conocimiento previo de la tasa de éxito	Necesario	No necesario
Intuitivo	P-valor	Probabilidad de que B es mejor que A
Tamaño de muestra	Pre-calculado	No es necesario pre-calcular
P-eaking	No permitido	Permitido (pero con cautela)
Rapidez	Fija	Mucho más flexible
Incertidumbre	Intervalo de confianza (pero cuidado, estos basados en p-valor así que difícil de interpretar)	Intervalo de certidumbre (basado en probabilidades)
Correcciones	Bonferroni (u otras)	No necesario
Coste computacional	~ 0	Caro comparado con frecuentista, ya que requerimos de simulaciones MC.
Criterios de decisión	Sí	Sí, pero más difíciles de implementar

Pero entonces, ¿el  
método Frecuentista es  
peor que Bayes?

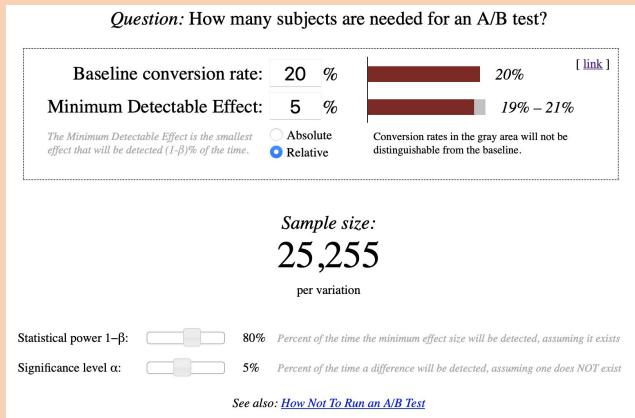
**¡NO!**

# COMPARACIÓN

	Frecuentista	Bayesiano
Conocimiento previo de la tasa de éxito	Necesario	No necesario
Intuitivo	P-valor	Probabilidad de que B es mejor que A
Tamaño de muestra	Pre-calculado	No es necesario pre-calcular
P-eaking	No permitido	Permitido (pero con cautela)
Rapidez	Fija	Mucho más flexible
Incertidumbre	Intervalo de confianza (pero cuidado, estos basados en p-valor así que difícil de interpretar)	Intervalo de certidumbre (basado en probabilidades)
Correcciones	Bonferroni (u otras)	No necesario
Coste computacional	~ 0	Caro comparado con frecuentista, ya que requerimos de simulaciones MC.
Criterios de decisión	Sí	Sí, pero más difíciles de implementar

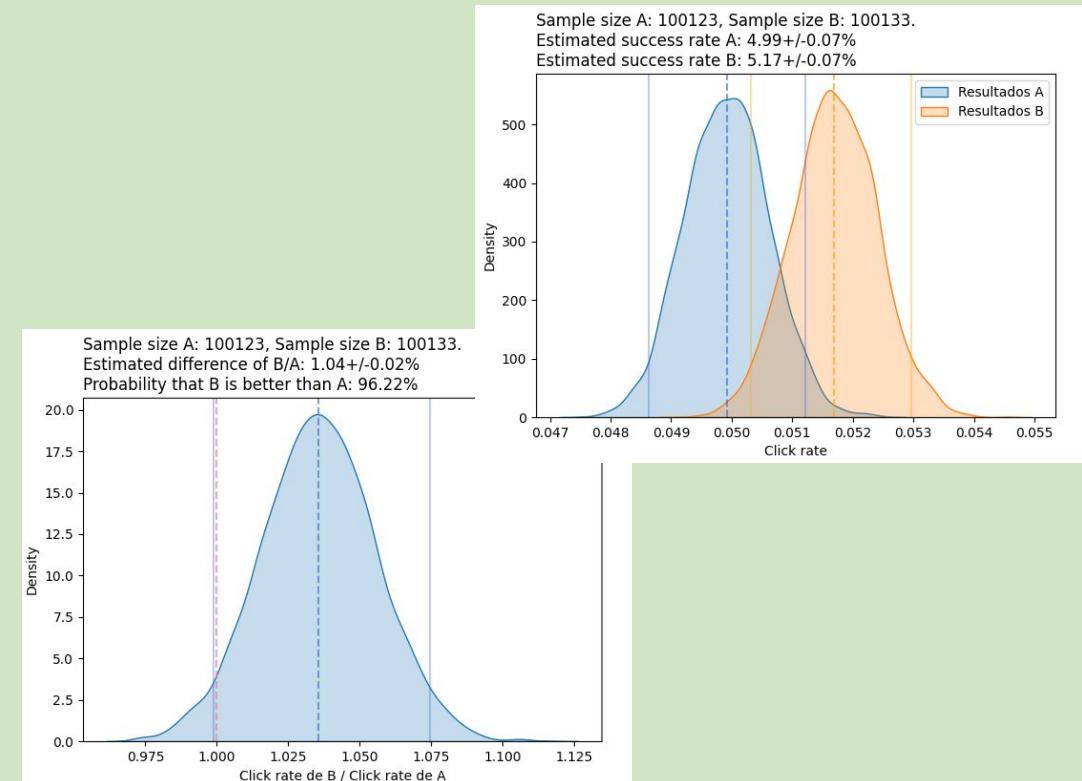
# RECOMENDACIÓN

## Diseño y Toma de decisiones



```
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
```

## Análisis en detalle y comprensivo



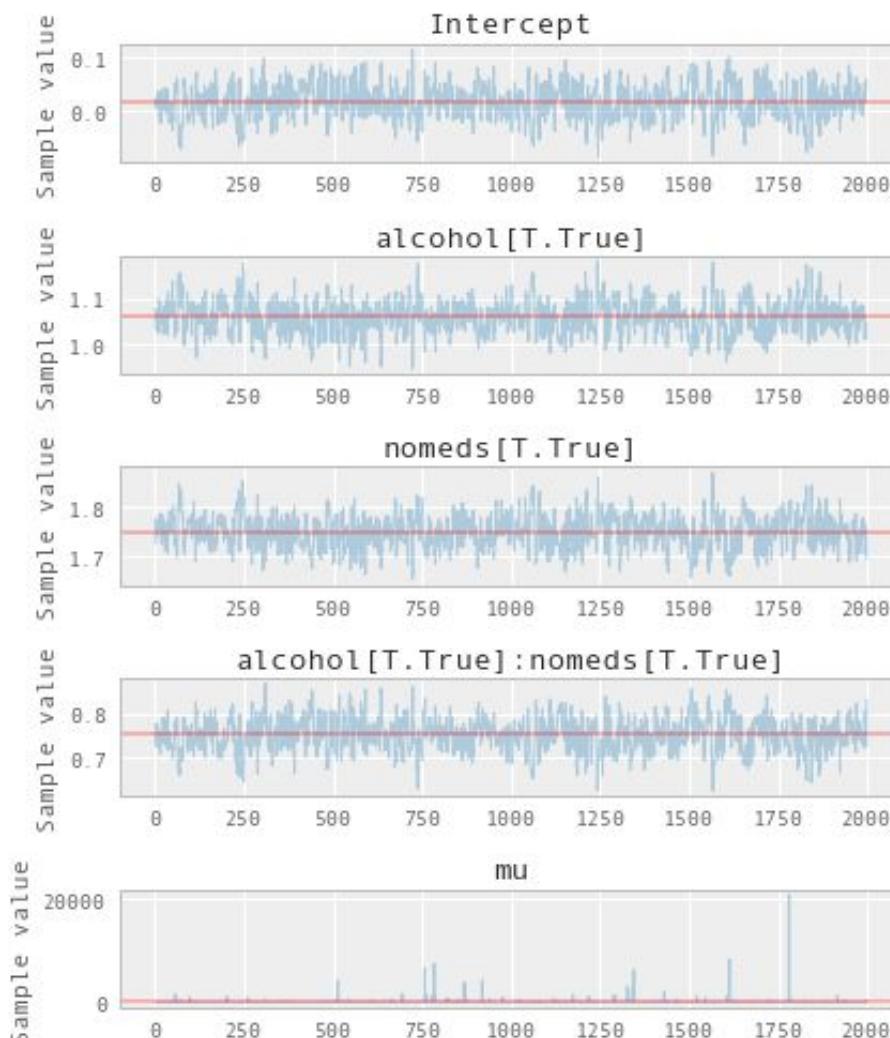
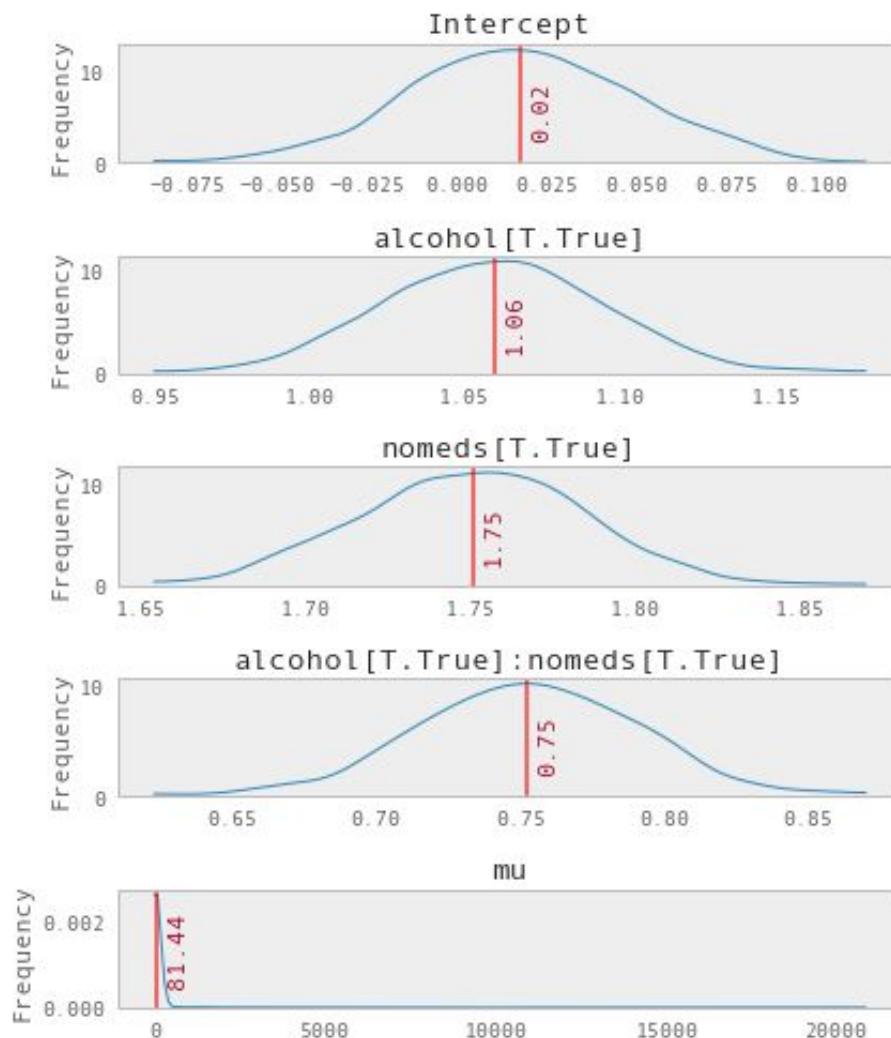
# Parte 6. Test de hipótesis y Bayes en otros usos



# LINEAR REGRESSION

```
Call:  
lm(formula = y ~ ., data = data)  
  
Residuals:  
    Min      1Q  Median      3Q     Max  
-3.1372 -0.6741 -0.0153  0.6447  3.7370  
  
Coefficients:  
              Estimate std. Error t value Pr(>|t|)  
(Intercept) 0.005056  0.026170   0.193   0.847  
x1           0.524853  0.026067  20.135 < 2e-16 ***  
x2           0.494757  0.031666  15.624 < 2e-16 ***  
x3           0.148379  0.026396   5.621  2.26e-08 ***  
x4          -0.407096  0.027101 -15.022 < 2e-16 ***  
x5          -0.205642  0.025519  -8.058 1.57e-15 ***  
x6          -0.122145  0.025631  -4.766 2.07e-06 ***  
---  
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1  
  
Residual standard error: 1.011 on 1493 degrees of freedom  
Multiple R-squared:  0.4268,    Adjusted R-squared:  0.4245  
F-statistic: 185.3 on 6 and 1493 DF,  p-value: < 2.2e-16
```

# BAYESIAN MODELLING



# AGENDA DÍA 2

- Romper el hielo.
  - Un par de recursos extra por si os apetece seguir indagando en A/B testing
  - Repaso del día 1.
- Parte 1. Más allá del test estadístico de proporciones
- Parte 2. Debilidades del método frecuentista
- Parte 3. A/B testing con análisis Bayesiano
- Parte 4. ¿Soluciona Bayes las debilidades del método frecuentista?
- Parte 5. Bayes vs Frecuentista
- Parte 6. Test de hipótesis y Bayes en otros usos

# Gracias

