Clase 185 — A/B testing: tamaño de muestra, poder estadístico

Parte: 3 — Estadística Inferencial y Causal · Fuente: Bruce & Bruce, cap. 3 A/B Testing + Kohavi, Tang & Xu, Trustworthy Online Controlled Experiments (2020). Duración estimada: 90 min.

Clase 185 — A/B testing: tamaño de muestra, poder estadístico

> *Parte: 3 — Estadística Inferencial y Causal · Fuente: Bruce & Bruce, cap. 3 A/B Testing + Kohavi, Tang & Xu, Trustworthy Online Controlled Experiments (2020). Duración estimada: 90 min.*

Objetivo

Diseñar y analizar un A/B test end-to-end: definir hipótesis y métrica primaria, calcular tamaño de muestra con el poder estadístico deseado, randomizar correctamente, analizar resultados sin peeking y reportar con effect size + IC. Conocer tres herramientas modernas que reducen muestra requerida o eliminan el problema de peeking: CUPED, sequential testing (always-valid p-values) y A/B bayesiano.

Resultados de aprendizaje

Al finalizar, el estudiante podrá:

- Calcular n requerido con statsmodels.stats.power.TTestIndPower o NormalIndPower para una MDE (minimum detectable effect) dada, α y poder.
- Implementar el análisis: t-test (continua) o z-test de proporciones (binaria), con effect size + IC95 %.
- Identificar y evitar 5 errores clásicos: peeking, p-hacking, no estratificación, SRM (sample ratio mismatch), Simpson's paradox.
- Aplicar CUPED para reducir varianza usando una covariable pre-experimento.
- Diseñar un test secuencial con always-valid p-values (Howard et al. 2021) o GST (group sequential testing) que permita parar antes sin inflar α.
- Comparar A/B clásico (frecuentista) con A/B bayesiano (PyMC o bayesab) y entender ventajas (interpretación directa, parar cuando alcance precisión).

Temas

- Hipótesis nula vs alternativa en A/B; métrica primaria, guardrails (no degradar latencia, error rate).
- Poder estadístico: P(rechazar H | H verdadera). Convención: 80 %.
- Sample size: depende de α (0.05), poder (0.8), σ y MDE.
- Aleatorización a nivel correcto (usuario vs sesión vs request).
- Peeking problem: mirar el resultado intermedio e inflar α.
- SRM (Sample Ratio Mismatch): si el ratio observado A/B se aleja del esperado 50/50, hay bug de asignación.
- Simpson's paradox: la tendencia global se invierte al estratificar.
- Complemento moderno: CUPED, sequential testing, A/B bayesiano.

Versión profundizada — 2026

El tema moderno que antes vivía como complemento dentro de esta clase ahora tiene su(s) clase(s) propia(s) con patrón completo, ejercicios y homework:

- Clase 154a — CUPED, sequential testing, always-valid p-values

Definiciones y características

- MDE (Minimum Detectable Effect): el efecto más chico que considerás relevante de detectar. Lo elegís antes del experimento.
- Poder estadístico (1-β): probabilidad de detectar el MDE si es real. Convención: 0.80.
- α: error tipo I. Para A/B testing, 0.05 es estándar; 0.01 para decisiones críticas.
- SRM (Sample Ratio Mismatch): cuando el ratio de asignación observado se desvía significativamente del esperado. Indica bug en la randomización. Test: chi² sobre conteos A vs B.
- Stratificación: balancear covariables (país, plataforma) entre A y B para evitar Simpson's paradox.
- CUPED: reducción de varianza usando covariable pre.
- Always-valid p-value: válido bajo cualquier tiempo de parada; permite peeking sin inflación de α.
- Novelty effect: los usuarios reaccionan al cambio en sí, no al diseño. Confirma con análisis por cohortes/tiempo.

Dataset / recursos

- Simular A/B: rng.binomial(1, 0.10, n) vs rng.binomial(1, 0.12, n) → MDE de 2 pp absoluto.
- Para CUPED: simular X (pre) y Y = 0.5*X + ε + δ·tratamiento.
- Librerías: statsmodels.stats.power, scipy.stats, confseq, pingouin.

Ejercicios

1. Sample size: querés detectar un uplift de tasa de conversión de 10 % → 11 % con poder 0.8 y α=0.05. Usá statsmodels.stats.proportion.samplesize_proportions_2indep_onetail o power.NormalIndPower().solve_power. ¿Cuánto necesitás por grupo?
2. Análisis clásico: simulá el experimento (n=8 000 por grupo, p_A=0.10, p_B=0.108), aplicá z-test de proporciones, reportá p, IC95 % de la diferencia y poder post-hoc.
3. CUPED: simulá X = rng.normal(50, 10, 2000) y Y = X + ε + 2·tratamiento con ε ~ N(0,5). Calculá n requerido con y sin CUPED para detectar el efecto de 2.
4. Peeking simulado: bajo H verdadera, simulá 1 000 experimentos donde "parás temprano" si p < 0.05 mirando cada 100 obs hasta 5 000. Mostrá cómo el α real se infla a ≈ 0.25.
5. A/B bayesiano: con A=(1000, 80), B=(1000, 100), calculá P(p_B > p_A) con priors Beta(1,1). Interpretá.

Homework verificable

Diseñar y analizar un A/B test simulado:

1. Calcular n por grupo para MDE=1 pp absoluto, baseline 10 %, α=0.05, poder=0.8.
2. Simular el experimento con n calculado y true uplift de 1.2 pp.
3. Reportar: z-test (p, IC), Cohen's h (effect size para proporciones), análisis bayesiano (P(B > A), expected uplift).
4. Repetir simulando peeking cada 1 000 obs sin corrección → mostrar inflación de α.
5. Concluir en 4 líneas: recomendación para producción (frecuentista clásico vs bayesiano vs always-valid).

Criterio de aceptación: n ≈ 14 800 por grupo. El z-test debe rechazar H con p < 0.05, P(B > A) debe ser > 0.95 en bayesiano, y el ejercicio de peeking debe mostrar α real entre 0.20 y 0.30.

Errores comunes

Síntoma / mensaje	Causa y cómo arreglar
Pareo temprano cuando vi p < 0.05	Peeking → α inflado. Fix: pre-registrar fe
El test A/B muestra p < 0.05 pero el negoc	Novelty effect, sesgo de selección, no est
Conteos A vs B son 48/52 con n=10⁶ y "es c	SRM serio. Fix: chi² sobre conteos; si p <
Resultado positivo global, negativo por ca	Simpson's paradox por estratificación disp
Reporto p < 0.05 sin effect size ni IC	Inutilizable para decisión de negocio. Fix

Preguntas frecuentes

¿Cuánto dura un A/B test?

Hasta alcanzar el n calculado y cubrir al menos un ciclo completo del negocio (típicamente 1-2 semanas para capturar weekday/weekend effects, holidays, etc.). Parar antes solo con sequential testing válido.

¿Si mi MDE es muy chico, qué hago?

Necesitás más muestra ($1/\text{MDE}^2$). Si no podés conseguirla: (a) usá CUPED para reducir varianza, (b) re-evaluá si ese MDE realmente importa para el negocio (un 0.5 % de uplift puede no compensar el costo de desplegar).

¿A/B bayesiano necesita pre-registro?

Conceptualmente menos: la decisión bayesiana (P(B > A) > threshold) es coherente con cualquier tiempo de parada si el prior es honesto. En práctica industrial igual conviene pre-registrar el threshold para evitar racionalizaciones.

¿Aleatorizo a nivel de usuario o de sesión?

A nivel de unidad de tratamiento: si el cambio afecta al usuario (ej.: redesign de homepage), por usuario. Si es un cambio que el usuario puede experimentar múltiples veces sin "aprenderlo" (ej.: ranking de búsqueda), por sesión o request — pero con cuidado por correlación intra-usuario.

¿Qué hago si n calculado es prohibitivo?

Opciones: (a) aumentar MDE (¿es realista el efecto que esperás?), (b) reducir varianza con CUPED, (c) reducir α a 0.10 si el costo de un falso positivo es bajo, (d) hacer un quasi-experiment con synthetic controls (Clase 157).

Referencias

- Bruce & Bruce, cap. 3 — A/B Testing.
- Kohavi, R., Tang, D. & Xu, Y. (2020), Trustworthy Online Controlled Experiments, Cambridge University Press — la biblia industrial.
- Deng, A. et al. (2013), Improving the Sensitivity of Online Controlled Experiments by Utilizing Pre-Experiment Data, WSDM (CUPED original).
- Howard, Ramdas, McAuliffe & Sekhon (2021), Time-Uniform, Nonparametric, Nonasymptotic Confidence Sequences, Annals of Statistics.
- statsmodels.stats.power.
- confseq — always-valid CIs.

Siguiente clase

Clase 186 — CUPED, sequential testing, always-valid p-values

Apéndice: notebook (primer bloque)

Primera celda ejecutable del notebook de la clase.

```
# Imports y configuración inicial
```

Archivos complementarios

- notebook.ipynb