Clase 187 — Diseño experimental

Parte: 3 — Estadística Inferencial y Causal · Fuente: Montgomery, Design and Analysis of Experiments (8ª ed.) + Kohavi, Tang & Xu (cap. 4-5). Duración estimada: 80 min.

Clase 187 — Diseño experimental

> *Parte: 3 — Estadística Inferencial y Causal · Fuente: Montgomery, Design and Analysis of Experiments (8ª ed.) + Kohavi, Tang & Xu (cap. 4-5). Duración estimada: 80 min.*

Objetivo

Pasar del A/B simple a diseños más ricos: bloques aleatorizados, factorial completo / fraccional, diseños cruzados (cross-over), switchback para experimentos con interferencia, y cluster randomization cuando la unidad de análisis no coincide con la unidad de tratamiento. Saber qué problema resuelve cada diseño y leer las consideraciones de SUTVA (Stable Unit Treatment Value Assumption).

Resultados de aprendizaje

Al finalizar, el estudiante podrá:

- Distinguir diseño completamente aleatorizado (CRD), bloques aleatorizados (RBD), factorial y fraccional 2^(k-p).
- Detectar cuándo SUTVA se viola (efectos de red, interferencia entre usuarios, fila/competencia) y aplicar el diseño correcto: cluster randomization, switchback, marketplace experiments.
- Diseñar un factorial 2^2 o 2^3 con pyDOE2 / statsmodels y descomponer efectos principales + interacciones.
- Saber cuándo usar fraccional (2^(k-p)) para reducir corridas y qué se sacrifica (confounding de interacciones de alto orden).
- Aplicar cross-over para experimentos pareados dentro de sujeto, con análisis vía test pareado o modelo mixto.

Temas

- CRD: el A/B clásico. Asume SUTVA (no interferencia entre unidades).
- RBD (bloques): bloquear por variable nuisance (ej.: día de semana, país) para reducir varianza dentro del bloque.
- Factorial 2^k: testear k factores simultáneamente. Captura interacciones; mucho más eficiente que A/B por factor.
- Fraccional 2^(k-p): corridas reducidas. Se confunden ("aliasing") efectos de alto orden con principales.
- Cross-over: cada sujeto recibe ambos tratamientos en períodos distintos. Análisis pareado, controla variabilidad inter-sujeto. Riesgo: carry-over effect.
- Cluster randomization: aleatorizar grupos (clases, ciudades) en lugar de individuos cuando hay contaminación social.
- Switchback: alternar tratamiento global en bloques de tiempo (típico de marketplaces de dos lados — Uber, DoorDash).
- SUTVA: cada unidad solo recibe una versión del tratamiento; los efectos no se propagan entre unidades.

Definiciones y características

- Aleatorización: asignación al azar a tratamiento; protege contra confounders observados y no observados.
- Bloqueo: agrupar unidades similares en bloques homogéneos; aleatorizar dentro del bloque. Reduce varianza si la variable de bloqueo importa.
- Efecto principal: efecto promedio de un factor sobre la respuesta.
- Interacción: efecto que un factor tiene sobre el efecto de otro (no aditividad).
- Confounding (en fraccional): la imposibilidad estructural de distinguir un efecto de otro debido al diseño. Se acepta confundir interacciones triples con efectos principales (asumimos triples ≈ 0).
- SUTVA: no interference + no hidden variations of treatment. Es la asunción callada de todo A/B clásico.
- Carry-over: efecto residual de un tratamiento previo en cross-over. Se mitiga con washout periods y diseños equilibrados (Latin square).
- ICC (Intra-Cluster Correlation): correlación dentro del cluster. Si ρ > 0, el n efectivo es menor; corregir con n_effective = n / (1 + (m-1)·ρ) donde m es tamaño del cluster.

Dataset / recursos

- Sintéticos para factorial 2² (e.g., A=color botón, B=texto botón → CTR).
- Iris / penguins para análisis ANOVA tipo factorial.
- Librerías: pyDOE2 (pip install pyDOE2), statsmodels.formula.api, pingouin.

Ejercicios

1. Factorial 2²: simulá CTR con ctr = 0.10 + 0.02·A + 0.015·B + 0.005·A·B + ε. Hacé el experimento con 1 000 obs por celda. Ajustá ols('ctr ~ A * B', data).fit() y reportá los 4 coeficientes (intercepto, A, B, A:B). Verificá contra el verdadero.
2. Bloqueo: simulá un experimento de uplift en tasa de retención por país con paises = ['AR','BR','MX'] con baselines distintos (p0 {0.5, 0.3, 0.4}). Compará: A/B sin estratificar vs bloqueado por país. Mostrá cómo el SE de δ cae con bloqueo.
3. Fraccional 2^(4-1): usá pyDOE2.fracfact('a b c d') y discutí qué interacciones quedan aliased. ¿Cuántas corridas vs full factorial?
4. Cluster randomization: simulá 50 escuelas con 30 alumnos c/u, ICC=0.10, efecto verdadero 0.3. Compará t-test ingenuo (n=1500) vs análisis correcto a nivel cluster (n=50). El primero infla α; el segundo es correcto.
5. Switchback: simulá precio dinámico en una ciudad con bloques de 1 h alternando A y B durante 7 días. Análisis: comparar bloques A vs B con tests pareados por hora-del-día.

Homework verificable

Sobre un dataset simulado de factorial 2³ (3 factores binarios, ej.: color × texto × posición sobre conversion):

1. Generar datos con efectos principales no nulos y una interacción A:B.
2. Ajustar OLS con todos los términos hasta triple.
3. Reportar la tabla ANOVA y identificar los términos significativos.
4. Verificar gráficamente la interacción A:B con sns.pointplot(x='A', y='conv', hue='B').
5. Discutir en 3 líneas qué pasaría si hubieras hecho 3 A/B tests separados en vez del factorial.

Criterio de aceptación: el OLS recupera los efectos verdaderos; la ANOVA marca A, B y A:B como significativos; el análisis muestra que el factorial requiere menos corridas totales que 3 A/B independientes (típicamente la mitad) y además detecta la interacción.

Errores comunes

Síntoma / mensaje	Causa y cómo arreglar
Hago A/B en una red social y "el grupo con	SUTVA violada por contaminación (el contro
Análisis a nivel individual de experimento	α inflado por correlación intra-cluster. F
Asumo interacción nula y la interpretación	Si A solo funciona cuando B está activo, l
Cross-over sin washout y carry-over inflad	El efecto del tratamiento 1 contamina la m
Hago RBD pero no bloqueo lo que importa	Si bloqueás por país pero el efecto import

Preguntas frecuentes

¿Cuándo factorial vs A/B múltiple?

Casi siempre factorial. A/B múltiple (1 factor a la vez) requiere más muestra y no detecta interacciones. Factorial 2^2 con n por celda tiene la misma precisión que dos A/B independientes con ~n total.

¿Cuántos factores antes de necesitar fraccional?

Con 2^k = 16 (k=4) ya tenés 16 celdas → si querés ~1 000 por celda son 16 000 obs. Manejable. A partir de k=5 (32 celdas) considerar fraccional, sobre todo si esperás interacciones de alto orden insignificantes.

¿Cuándo cluster randomization?

Cuando hay contaminación natural: educación (alumnos en la misma aula), salud pública (familias), marketplaces (oferta + demanda compartidas). Costo: n efectivo cae con ICC; necesitás muchos clusters.

¿Switchback es válido si hay tendencia temporal?

Sí pero requiere modelar la tendencia (ej.: incluir hora del día como covariable). Si tu métrica tiene fuerte estacionalidad y bloques son largos, mejor diseño Latin square en el tiempo.

¿Diseño antes o análisis primero?

Diseño primero, siempre. El análisis sin diseño correcto produce conclusiones sospechosas (Simpson, confounders, peeking). "You can't analyze your way out of a bad design" (Tukey).

Referencias

- Montgomery, D.C. (2017), Design and Analysis of Experiments (8ª ed.) — referencia clásica completa.
- Kohavi, R., Tang, D. & Xu, Y. (2020), Trustworthy Online Controlled Experiments, caps. 4-5 (advanced designs, interference).
- Imbens & Rubin (2015), Causal Inference for Statistics, Social, and Biomedical Sciences.
- pyDOE2 — generación de matrices de diseño.
- Athey, Eckles & Imbens (2018), Exact p-Values for Network Interference, JASA.

Siguiente clase

Clase 188 — Inferencia causal: DAGs, confounders, instrumentos

Apéndice: notebook (primer bloque)

Primera celda ejecutable del notebook de la clase.

```
# Imports y configuración inicial
```

Archivos complementarios

- notebook.ipynb