Resumen del capítulo: Preparación para un Test A/B

Test AA

Antes de iniciar una prueba A/B realizamos una prueba **A/A** para asegurarnos de que:

- Los resultados no se ven afectados por anomalías o valores atípicos en la población estadística
- La herramienta de división del tráfico funciona correctamente
- Los datos se envían correctamente a los sistemas analíticos

Las pruebas A/A son similares a las pruebas A/B pero, en este caso, a cada grupo se le muestra la misma versión de la página. Si el tráfico y la herramienta para realizar la prueba A/A funcionan como deberían, no habrá una diferencia (significativa) en los resultados. Las pruebas A/A también ayudan a determinar cuánto debe durar la prueba A/B y el método para analizar los resultados.

Aquí están los criterios para una prueba A/A exitosa:

- El número de usuarios en diferentes grupos no varía en más del 1%
- Para todos los grupos, los datos sobre el mismo evento se registran y se envían a sistemas analíticos
- Ninguna de las métricas clave varía en una cantidad estadísticamente significativa, por lo general, no más del 1%
- Los usuarios permanecen en sus grupos hasta el final de la prueba. Si ven diferentes versiones de la página durante el estudio, no quedará claro qué versión influyó en sus decisiones por lo que la fiabilidad de los resultados se verá comprometida.

La medida en que las métricas clave difieren entre los grupos depende de que tan sensibles deben ser los experimentos.

Errores de tipo I y tipo II en la prueba de hipótesis. Poder y significación

Un **error de tipo** I es un **resultado positivo falso**. Aquí no hay diferencia entre los grupos que se comparan pero la prueba produjo un valor p inferior al nivel de significación. En consecuencia, hay motivos para rechazar H_0. Por lo tanto, la probabilidad de cometer un error de tipo I es igual al nivel de significación, α.

Un **error de tipo** II es un **resultado negativo falso**. Esto significa que hay una diferencia entre los grupos pero la prueba produjo un valor p mayor que `α` por lo que no hay motivo para rechazar H_0. Si llamamos β a la probabilidad de cometer un error de tipo II, `1 - β` será **el poder estadístico de la prueba de hipótesis**. Si β es la probabilidad de cometer un error, `1 - β` es una probabilidad de no cometerlo o de rechazar correctamente la hipótesis nula cuando es falsa.

		Hipótesis cierta	
		H_0	H_1
Resultado de aplicar un criterio	H_0	H_0 correctamente aceptada	H_0 incorrectamente rechazada (error de tipo I)
	H_1	H_0 incorrectamente aceptada (error de tipo II)	H_0 correctamente rechazada

Comparaciones múltiples: pruebas A/B y A/B/n

A menudo, una hipótesis se prueba en diferentes variaciones. Puedes comparar varios grupos con un grupo de control pero deberás tener en cuenta la probabilidad creciente de cometer errores de tipo I y tipo II.

Hacer varias comparaciones con los mismos datos se llama **pruebas múltiples**. Lo que hay que saber al respecto es que la probabilidad de cometer un error tipo I aumenta con cada nueva prueba de hipótesis.

Si la probabilidad de cometer un error es `α` cada vez, la probabilidad de no cometer ningún error es `1-α`. Entonces, la probabilidad de no cometer ningún error en el proceso de las comparaciones `k` será:

$$(1-a)^k$$

La probabilidad de cometer al menos un error en el proceso de comparaciones `k` será:

$$1-(1-a)^k$$

Para disminuir la probabilidad de resultados falsos positivos en comparaciones múltiples, los expertos tienen varios métodos para corregir el nivel de significación, lo que ayuda a reducir la tasa de error por familia (FWER del inglés family-wise error rate).

- El procedimiento de Bonferroni (la corrección de Bonferroni):

$$\alpha_1 = \ldots = \alpha_m = \alpha/m.$$

- El método de Holm (materiales en inglés)

$$\alpha_1 = \frac{\alpha}{m},\ \alpha_2 = \frac{\alpha}{m-1},\ \ldots,\ \alpha_i = \frac{\alpha}{m-i+1},\ \ldots,\ \alpha_m = \alpha.$$

- El método de Šidák (materiales en inglés)

$$\alpha_1 = \alpha_2 = \ldots = \alpha_m = 1 - (1 - \alpha)^{\frac{1}{m}} .$$

La corrección de Bonferroni es la más común debido a su simplicidad. No es difícil dividir el nivel de significación deseado entre el número de comparaciones, que se realizan con los mismos datos, sin necesidad de realizar nuevas observaciones para cada prueba. Si recopilas nuevos datos para cada prueba de hipótesis, realiza la prueba de la manera estándar, seleccionando el valor p necesario como lo hiciste en la parte del curso sobre estadística.

Cálculo del tamaño de la muestra y la duración de la prueba

Los analistas deben tener en cuenta las condiciones en las que se generaron las muestras de la prueba A/B, incluida la duración de la prueba y si el **peeking problem** es relevante.

A la hora de determinar la duración se tienen en cuenta los cambios cíclicos en el tráfico (diario, semanal, mensual) y el tiempo que tarda un cliente en decidirse a realizar una compra.

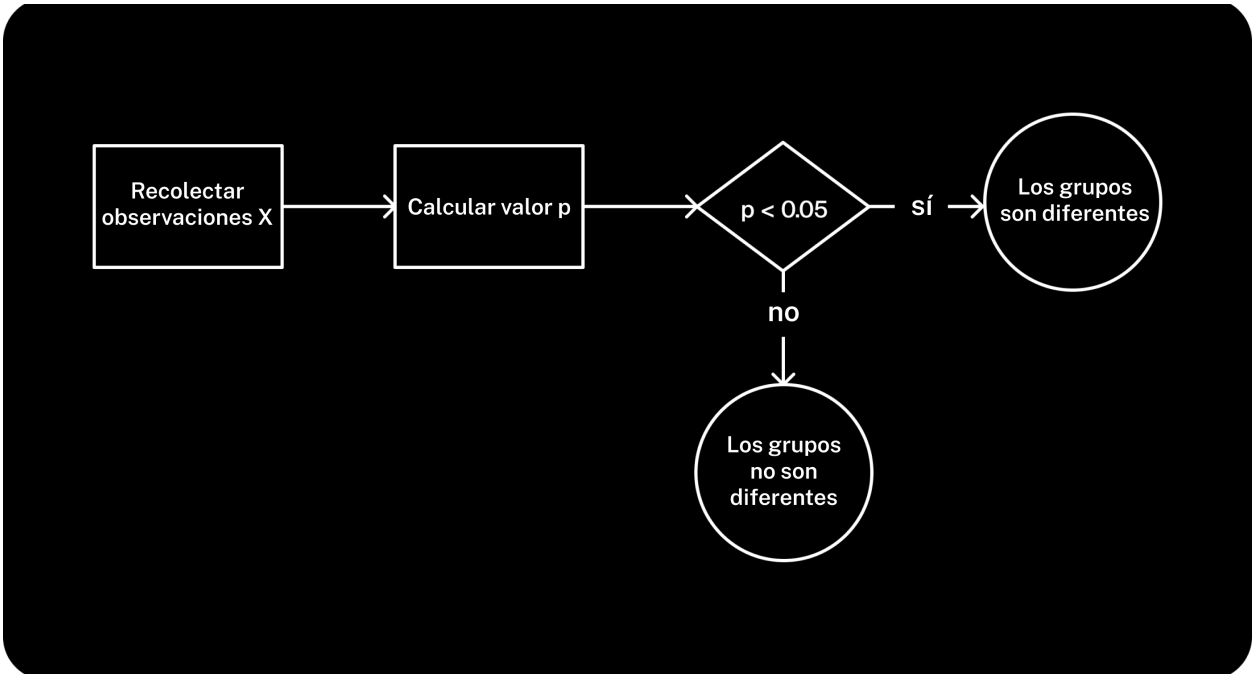
El problema de **vislumbrar los resultados** (peeking problem en inglés), aparece si la introducción de nuevos datos al comienzo de la prueba distorsiona considerablemente el resultado global. Incluso la parte más pequeña de datos nuevos es grande con respecto a los datos ya acumulados; no lleva mucho tiempo alcanzar la significación estadística.

Es una manifestación de la ley de los grandes números. Si hay pocas observaciones, su dispersión será mayor. Si hay muchos, los valores atípicos aleatorios se anulan entre sí. Esto significa que cuando la muestra es demasiado pequeña, es más probable que veas diferencias pero no serán estadísticamente significativas. (Por ejemplo, no es raro obtener cruz 8 veces de 10 al lanzar una moneda pero obtener 800 de 1000 veces sería de interés periodístico). Para una prueba estadística, esto significará disminuir el

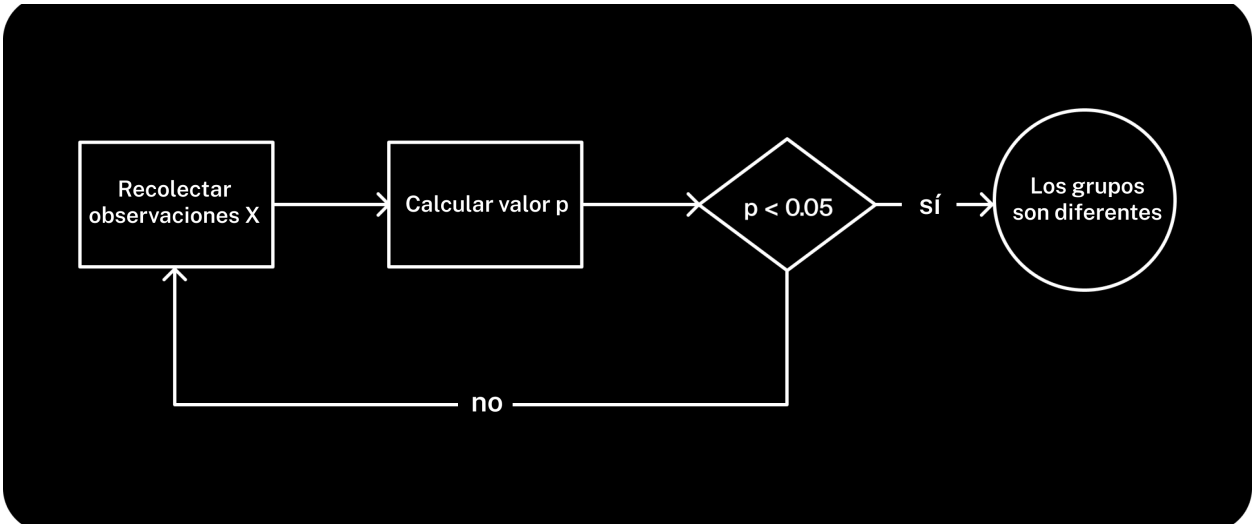
valor p **hasta que sea lo suficientemente pequeño como para rechazar la hipótesis nula.

Para compensar el peeking problem, se determina el tamaño de la muestra antes de que comience la prueba.

Aquí está el procedimiento correcto para la prueba A/B:



Y aquí está la forma incorrecta de hacerlo:



Calculadoras de duración de prueba y tamaño de muestra

Una de las formas más sencillas de determinar el tamaño de la muestra que necesitas es calcularlo en línea.

Aqui tienes varias calculadoras:

- http://www.evanmiller.org/ab-testing/sample-size.html (materiales en inglés)
- https://www.optimizely.com/sample-size-calculator/?conversion=20&effect=5&significance=95 (materiales en inglés)
- https://vwo.com/tools/ab-test-duration-calculator/ (materiales en inglés)

Estos servicios son buenos para evaluar el tamaño de muestra mínimo requerido para el cual sería notable un cambio en la métrica. Esto ayudará a calcular la duración mínima de la prueba.

Análisis gráfico de métricas y definición de alcances

Las ventajas de las calculadoras:

- Cálculo simple de conversión
- Solución del peeking problem (vislumbrar los resultados)
- Posibilidad de estimar la duración mínima de la prueba

Las desventajas:

- Tamaño de la muestra: necesario pero muchas veces no suficiente para que una prueba sea válida
- Las calculadoras no pueden tener en cuenta el hecho de que, en la vida real, la conversión y el efecto detectable mínimo relativo nunca permanecen iguales a lo largo de la prueba
- Las calculadoras solo funcionan bien para calcular el tamaño de la muestra para la conversión. También hay calculadoras para otros indicadores pero son mucho más complejas.

Determinar el momento y la duración mínima de una prueba en función de la industria

Al determinar la duración y el momento de la prueba debes saber qué tipos de aumentos en la actividad caracterizan a su audiencia.

Las posibles razones de los aumentos repentinos incluyen:

- Días laborables o fines de semana
- Vacaciones (aumento de la demanda de regalos)
- Ventas, ofertas, actividades de marketing (los descuentos aumentan la actividad del público, modificando su comportamiento de compra)
- Eventos especiales (por ejemplo, comprar artículos para el regreso a las clases en el otoño)
- Estacionalidad del producto (por ejemplo, calentadores)
- La actividad de los competidores (los competidores bajan el precio de un producto por lo que cae la actividad de tus clientes);
- Cambios en la situación política y económica (recesiones, inflación, embargos, aumento de precios por aranceles adicionales).

Además de los aumentos de actividad, también debes tener en cuenta el ciclo de **realización de la métrica que se mide**. La mayoría de las veces está relacionado con el proceso de decisión de compra o el tiempo entre la primera idea de comprar un producto y la compra en sí.