A/B testing - Google Optimize

summary by Pablo Rodriguez-Bocca
@2019

Contenido

Estadística usada en “Google Optimize”, y su “probability to beat baseline”...

https://support.google.com/optimize/answer/7405044?hl=en
http://blog.analytics-toolkit.com/2018/google-optimize-statistical-significance-statistical-engine/
*What statistics are used in Google Optimize? “…we use **Bayesian inference** to generate our statistics”*
“With Bayesian inference, however, we’re able to use different models that adapt to each and every experiment”.
*“we’ve used”: **hierarchical, contextual, and restless models**, but at least for me it is not clear if these are still in use in the engine, if they are used alone or in combination, and what their assumptions are, aside from a one-sentence description….*

Se dice muy poco del método utilizado, por tanto no se puede conocer exactamente como se calcula “probability to beat baseline”. Sin embargo hay bibliografía suficiente como para tener una idea del método. Ver a continuación…

Usar “A/B Testing” o “Bandit Algorithms”

Hay dos investigadores fuertes: Chris Stucchio, y Evan Miller.

Conversion Rates. And how to measure them. https://www.chrisstucchio.com/pubs/slides/helpshift_2014/slides.html#1
Conclusions: Use A/B testing for long term feature decisions. Use Bandit Algorithms for Transient/high volume content

Don't use Bandit Algorithms - they probably won't work for you: https://www.chrisstucchio.com/blog/2015/dont_use_bandits.html
Why Multi-armed Bandit algorithms are superior to A/B testing:
https://www.chrisstucchio.com/blog/2012/bandit_algorithms_vs_ab.html

Bayesian Bandits - optimizing click throughs with statistics: https://www.chrisstucchio.com/blog/2013/bayesian_bandit.html

Si se tiene mucho volumen de información, y en particular si el CTR es dinámico entonces usar “Bandit algorithms”. Hay que tener más cuidado al usarlo, las hipótesis por lo general son:

1. *The samples drawn from each arm of the bandit are Independent and Identically Distributed.*
2. *The conversion rates for each arm do not change over time.*
3. *There is no delay between pulling an arm and observing the result. Or, more precisely, if there is a delay, it's shorter than the delay between opportunities to pull an arm.*

Background “Bayesian A/B Testing” vs “Frequentist statistical A/B testing”

How Not To Run an A/B Test: http://www.evanmiller.org/how-not-to-run-an-ab-test.html
Resumidamente, no parar el test antes de tiempo, definir un tamaño de muestra grande.
El tamaño de muestra puede definirse aqui: http://www.evanmiller.org/ab-testing/sample-size.html

A/B Testing Rigorously (without losing your job): http://elem.com/~btilly/ab-testing-multiple-looks/part1-rigorous.html

Una regla para evaluar calidad de AB test: Simple Sequential A/B Testing:
http://www.evanmiller.org/sequential-ab-testing.html

Is Bayesian A/B Testing Immune to Peeking? Not Exactly.
http://varianceexplained.org/r/bayesian-ab-testing/

Existen muchas críticas a los métodos basados en frecuencias, pero ambos padecen de los mismos potenciales problemas. Un resumen puede ser que la mayoría de los científicos piensan que los métodos Bayesianos son más fáciles de interrumpir sin generar perjuicio. Con muestras suficientemente grandes ninguno tiene problemas. De ahora en más describimos el método Bayesiano.

Inferencia Bayesiana en el contexto A/B testing: la sección importante del documento :)

https://web.archive.org/web/20151207201643/https://gist.github.com/stucchio/9090456
https://www.chrisstucchio.com/blog/2013/bayesian_analysis_conversion_rates.html
http://www.evanmiller.org/bayesian-ab-testing.html

De la probabilidad condicional:

$$P(A|B) = \frac{P(A \cap B)}{P(B)},$$

se obtiene la regla bayesiana:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

La regla se usa en inferencia bayesiana de la siguiente forma:

$$P(modelo|datos) = \frac{P(datos|modelo)P(modelo)}{P(datos)}$$

En nuestro A/B testing:

1. Los datos son: L = cantidad de leads, y C= cantidad de conversiones
2. Se supone un modelo, en nuestro caso una **distribución de CTR**, llamemos θ a la variable aleatoria que representa los valores posibles de CTR. Para CTR lo común es usar la distribución beta: **θ~Beta(α,β)**. Si no se sabe nada se usa Beta(1,1) que es la uniforme [0,1], si se conoce aproximadamente el CTR y su varianza entonces se pueden ajustar α y β mejor. Esto hace a la larga que se requieran menos muestras. Esta suposición del modelo determina la probabilidad apriori: **P(modelo)**. Que en nuestro caso es la densidad de la distribución beta:
$$P(\theta) = f_{\alpha,\beta}(\theta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)}$$
donde la función Beta se define como:
$$B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1}dt$$
3. Si se supone que el modelo es correcto, la realidad del experimento permite calcular haciendo cuentas la probabilidad de verosimilitud: **P(datos|modelo).** En nuestro caso se suponen todos los leads independiente, por tanto cada uno es convertido con una probabilidad θ, y la probabilidad de encontrar C conversiones en L leads es simplemente la **distribución binomial**:
$$P(L \text{ leads}, C \text{ conversiones } |\theta) = \binom{L}{C}\theta^C(1-\theta)^{L-C}$$
4. **P(datos)** no es necesario calcularlo, simplemente normalizaremos P(modelo|datos) para obtener una distribución de probabilidad.
5. Hacemos cuentas con todo lo anterior y obtenemos la probabilidad a posteriori de la evidencia: **P(modelo|datos)**. Que en nuestro caso se simplifica mucho y queda la densidad de la distribución beta pero con nuevos parámetros:
$$P(\theta|L \text{ leads}, C \text{ conversiones }) = f_{\alpha+C,\beta+L-C}(\theta)$$

Lo anterior permite determinar a partir de los datos, y algunos pocos supuestos, cuál es la distribución de probabilidad de la variable aleatoria CTR. Esto permite hacer múltiples cuentas, por ejemplo:

- Sabemos que el CTR $\theta \in [a,b]$ con un 95% de certidumbre si:

$$P(a < \theta < b) = \int_a^b f_{\alpha+C,\beta+L-C}(\theta)d\theta > 0.95$$

- Dada la cantidad de leads L', tenemos la cantidad esperada de conversiones C' para esa certidumbre:

$$C' = L' \int_a^b f_{\alpha+C,\beta+L-C}(\theta)d\theta$$

- Probabilidad de que el CTR sea mayor al 1%:

$$P(0.01 < \theta) = \int_{0.01}^1 f_{\alpha+C,\beta+L-C}(\theta)d\theta$$

- Si estamos haciendo un A/B test, entonces tenemos dos variantes de lo anterior:
  - L_A, C_A son la cantidad de leads y conversiones para la opción A de control.
  - L_B, C_B son la cantidad de leads y conversiones para la opción B de prueba.
  - Existe un CTR para cada test: θ_A y θ_B.
  - No se conoce a priori nada de los CTRs de ambos tests (por lo que se asume una distribución a priori normal: $\alpha_A = \alpha_B = \beta_A = \beta_B = 1$.
  - Entonces la probabilidad de que la opción B sea mejor que la opción A es:

$$P(\theta_B > \theta_A) = \sum_{i=0}^{C_B} \frac{B(1 + C_A + i, 1 + L_B - C_B + 1 + L_A - C_A)}{(1 + L_B - C_B + i)B(1 + i, 1 + L_B - C_B)B(1 + C_A, 1 + L_A - C_A)}$$

Esto es lo que estimo se usa para "probability to beat baseline".
Esta fórmula tiene una forma de computarse eficientemente:
https://www.chrisstucchio.com/blog/2014/bayesian_ab_decision_rule.html
Y además se conoce su expresión asintótica ($L \to \infty$):
https://www.chrisstucchio.com/blog/2014/bayesian_asymptotics.html

Retardo en obtener el CTR

Measuring Bernoulli Probabilities in the Presence of Delayed Reactions:
https://www.chrisstucchio.com/blog/2016/delayed_reactions.html

Es posible agregar al modelado la incapacidad de conocer el CTR en tiempo real. Fue estudiado en Bandit y Bayesian AB Testing.