

О.О. Ящук, Я.С. Бондаренко

Дніпропетровський національний університет імені Олеся Гончара

ПОСЛІДОВНИЙ АНАЛІЗ В ЗАДАЧІ А/В-ТЕСТУВАННЯ

Розглянуто послідовний аналіз у задачі А/В-тестування сторінок сайту, який потребує меншого числа спостережень порівняно з класичною перевіркою статистичних гіпотез. Це дозволяє прискорити тестування та приймати рішення щодо ефективності варіацій веб-сторінок якнайшвидше. Розроблено програмні реалізації А/В-тестування у випадку класичної та послідовної перевірки статистичних гіпотез на мові програмування R.

Рассмотрен последовательный анализ в задаче А/В-тестирования страниц сайта, который требует меньшего числа наблюдений по сравнению с классической проверкой статистических гипотез. Это позволяет ускорить тестирование и принимать решение об эффективности вариаций веб-страниц быстрее. Разработаны программные реализации А/В-тестирования при классической и последовательной проверке статистических гипотез на языке программирования R.

Sequential testing with likelihood ratios which requires fewer observations than classical hypothesis testing is proposed. It allows to speed up A/B testing. Decisions about effectiveness of variations of web pages are made as soon as possible. Software implementation of A/B testing has been developed in R.

Ключові слова: точкове оцінювання параметрів, побудова довірчих інтервалів для невідомих параметрів, перевірка статистичних гіпотез, А/В-тестування.

1. Вступ

А/В тестування є потужний інструмент для підвищення якості сайту. А/В-тестування дозволяє оцінювати кількісні показники роботи двох варіацій веб-сторінки, а також порівнювати їх між собою. Спліт-тестування допомагає досліджувати ефективність змін сторінки, наприклад додавання нових елементів дизайну або закликів до дії. Практичний сенс застосування цього методу полягає в пошуці та впровадженні компонентів сторінки, що збільшує її результативність. За допомогою А/В-тестування можна впливати на конверсію, стимулювати збут і підвищувати прибутковість веб-проекту. Отже, актуальність проблематики, пов'язаної з вивченням математичної моделі А/В тестування, є безперечна.

Кожен бізнес веб-сайт прагне, щоб відвідувачі перетворювалися з простих відвідувачів на відвідувачів, які б виконували бажані для власників сайту дії. А/В тестування – метод, що визначає, яка варіація веб-сторінки є більш ефективна. Конверсія – це відношення кількості відвідувачів сторінки, які виконали цільову дію, до загальної кількості відвідувачів сторінки. Під час А/В тестування потік відвідувач поділяється навпіл. Дві версії веб-сторінки А та В в один і той самий час пропонують двом групам відвідувачів. Версію веб-сторінки, яка має більший процент конверсії, вважають кращою. Веб-сторінку, на яку спрямовують потік відвідувачів для визначення процента конверсії називають цільовою сторінкою. Збільшення продажів та збільшення кількості зареєстрованих користувачів (збільшення процента конверсії) складає мету А/В тестування.

Дослідженням з даної теми присвячена робота [1] вчених Стенфордського університету, які дійшли висновку щодо доцільності застосування послідовної перевірки статистичних гіпотез [2 – 5] за неперервного моніторингу результатів А/В-тестування, зокрема, вони застосували підхід J.Ville и A.Wald. Послідовний аналіз – це метод перевірки статистичних гіпотез, характерною рисою якого є те, що кількість спостережень, необхідних для процесу перевірки, не задана наперед, а залежить від результату самих спостережень, тобто є випадкова величина, на відміну від традиційних критеріїв. Перевага даного методу полягає в тому, що він потребує в середньому приблизно в два рази менше спостережень, ніж рівносильна йому перевірка з фіксованою кількістю спостережень.

2. Постановка задачі

Метою дослідження є реалізація класичної та послідовної перевірки статистичних гіпотез, запропонованої М. Girshick [6], та ілюстрація переваг послідовного аналізу в задачі А/В тестування сторінок сайту.

3. Класична перевірка статистичних гіпотез під час А/В-тестування

У двох групах (контрольній та експериментальній) проведемо незалежні випробування Бернуллі з імовірністю успіху p_{cont} в одному випробуванні в контрольній групі та ймовірністю успіху p_{exp} в одному випробуванні в експериментальній групі. Імовірності успіху p_{cont} та p_{exp} невідомі. Відносно невідомих параметрів висуваємо нульову гіпотезу $H_0 : p_{cont} = p_{exp}$ (експериментальна та контрольна групи мають однакову ймовірність успіху). Відтак побудуємо критерій для перевірки нульової

гіпотези. Нульова гіпотеза може мати різні альтернативи: двосторонню $H_1 : p_{cont} \neq p_{exp}$ (експериментальна та контрольна групи мають різні ймовірності успіху) або односторонні $H_1 : p_{cont} > p_{exp}$ та

$H_1 : p_{cont} < p_{exp}$. Оцінка $\frac{\mu_{cont}}{n_{cont}}$ (частота успіху) є незміщена, спроможна

оцінка для параметра p_{cont} в контрольній групі; оцінка $\frac{\mu_{exp}}{n_{exp}}$ (частота

успіху) є незміщена, спроможна оцінка для параметра p_{exp} в експеримен-

тальній групі. Оцінка $\hat{p} = \frac{\mu_{cont} + \mu_{exp}}{n_{cont} + n_{exp}}$ виступає об'єднаною частотою

успіху в двох групах.

Критерій χ^2 порівняння двох ймовірностей [7]. Якщо гіпотезу $H_0 : p_{cont} = p_{exp}$ відхиляти за

$$\chi^2 = \frac{\left| \frac{\mu_{cont}}{n_{cont}} - \frac{\mu_{exp}}{n_{exp}} \right|}{\sqrt{\hat{p}(1-\hat{p}) \left(\frac{1}{n_{cont}} + \frac{1}{n_{exp}} \right)}} \geq z_{1-\alpha/2}$$

і не відхиляти в супротивному разі, то з ймовірністю α гіпотеза буде відхилятися, коли вона справедлива (альтернатива $H_1 : p_{cont} \neq p_{exp}$).

Або, що те ж саме, гіпотеза $H_0 : p_{cont} - p_{exp} = 0$ відхилятиметься,

якщо $\hat{\theta} = \frac{\mu_{cont}}{n_{cont}} - \frac{\mu_{exp}}{n_{exp}}$ лежить за межами довірчого інтервалу

$$\left(-z_{1-\alpha/2} \sqrt{\hat{p}(1-\hat{p}) \left(\frac{1}{n_{cont}} + \frac{1}{n_{exp}} \right)}; z_{1-\alpha/2} \sqrt{\hat{p}(1-\hat{p}) \left(\frac{1}{n_{cont}} + \frac{1}{n_{exp}} \right)} \right),$$

і не відхилятиметься в супротивному разі.

Обчислення обсягу вибірок контрольної та експериментальної груп. Мінімальний обсяг вибірок контрольної та експериментальної груп, який забезпечує задані ймовірність помилки I роду α , потужність критерію $1 - \beta$, різницю рівнів конверсії θ та базовий рівень конверсії p у разі перевірки простої гіпотези $H_0 : p_{cont} - p_{exp} = 0$ проти простої альтернативи $H_0 : p_{cont} - p_{exp} = \theta$, дорівнює

$$n = \frac{2(z_{1-\alpha/2} + z_{1-\beta})^2 p(1-p)}{\theta^2}.$$

Наприклад, за $\alpha = 0,05$, $1 - \beta = 0,8$, $p = 0,1$, $\theta = 0,02$ мінімальний обсяг вибірки кожної групи дорівнюватиме 3533.

Реалізація А/В-тестування. Під час А/В тестування контрольній групі відвідувачів пропонують для перегляду сторінку А, експериментальній групі відвідувачів – сторінку В (варіацію А). Для збереження чистоти експерименту будемо ідентифікувати відвідувачів під час А/В тестування, а за повторних відвідувань сайту пропонуватимемо їм для перегляду ту саму сторінку А або В з попереднього відвідування. Висуваємо нульову гіпотезу про однаковий рівень конверсії в обох групах $H_0 : p_{cont} = p_{exp}$. Моделюємо потік відвідувачів сайту. Кожен відвідувач з імовірністю $\frac{1}{2}$ може потрапити до контрольної групи і з імовірністю $\frac{1}{2}$ може потрапити до експериментальної групи. Після того як відвідувач опинився в одній із двох груп, змодельуємо його поведінку. Поведінку відвідувача однозначно визначаємо двома подіями: успіх – відвідувач виконав цільову дію, неуспіх – відвідувач не виконав цільову дію. Якщо відвідувач опинився в контрольній групі, то подія «успіх» відбувається з імовірністю p_{cont} , якщо ж відвідувач опинився в експериментальній групі, то подія «успіх» відбувається з імовірністю p_{exp} . За кожного $n = n_{cont} + n_{exp}$ слід перевіряти нульову гіпотезу $H_0 : p_{cont} - p_{exp} = 0$.

На рис. 1, 2 зображено результати реалізацій А/В тестування за таких параметрів: $p_{cont} = 0,1$, $p_{exp} = 0,12$, $n = 20000$, $\alpha = 0,05$.

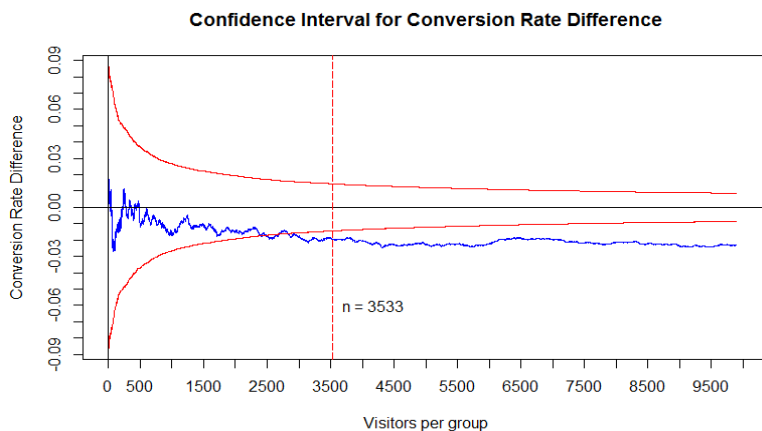


Рис. 1. Тест 1. Довірчий інтервал для Conversion Rate Difference

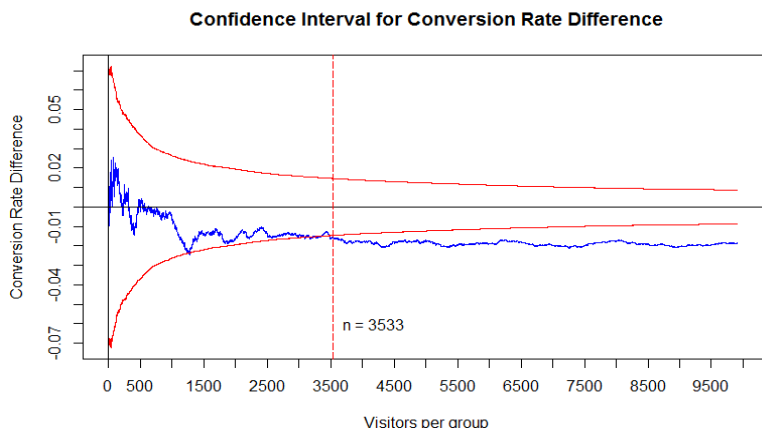


Рис. 2. Тест 2. Довірчий інтервал для Conversion Rate Difference

Тепер необхідно дочекатися, доки мінімальний обсяг однієї з груп до-рівнюватиме 3533 і тільки тоді приймати рішення – відхилити чи не відхи-ляти нульову гіпотезу.

4. Послідовна перевірка статистичних гіпотез під час А/В-тестування

Розглянемо метод послідовної перевірки гіпотез, для якого фіксовані помилки першого та другого роду та який потребує меншої кількості спостережень, що дозволить прискорювати А/В тестування та якнайшвидше приймати рішення щодо ефективності варіації веб-сторінки. Далі позначимо : $P_{cont} = P_1, P_{exp} = P_2$.

Нехай ξ_1, \dots, ξ_n - вибірка з розподілу $P(x_1, p_1) = p_1^{x_1}(1 - p_1)^{1-x_1}$, $x_1 = 0, 1$; η_1, \dots, η_n - вибірка з розподілу $P(x_2, p_2) = p_2^{x_2}(1 - p_2)^{1-x_2}$, $x_2 = 0, 1$. Параметри p_1, p_2 невідомі. Відносно параметрів p_1, p_2 висуваємо гіпотезу $H_0 : p_1 \leq p_2$, альтернативна їй гіпотеза $H_1 : p_1 > p_2$. Необхідно побудувати послідовний критерій для перевірки гіпотези H_0 . Виберемо деяку величину p_1^0 параметра p_1 і деяку величину p_2^0 параметра p_2 , при цьому величини p_1^0 та p_2^0 обиремо такими, що $p_1^0 < p_2^0$.

Нехай нульова гіпотеза $H_0 : p_1 = p_1^0, p_2 = p_2^0$ - це гіпотеза про те, що спільний розподіл величин ξ та η задається добутком

$$P(\xi, p_1^0)P(\eta, p_2^0) = \left(p_1^0\right)^{x_1} \left(1 - p_1^0\right)^{1-x_1} \left(p_2^0\right)^{x_2} \left(1 - p_2^0\right)^{1-x_2}, x_1 = 0, 1; x_2 = 0, 1.$$

Альтернативна гіпотеза $H_1 : p_1 = p_2^0, p_2 = p_1^0$ - це гіпотеза про те, що спільний розподіл величин ξ та η задається добутком

$$P(\xi, p_2^0)P(\eta, p_1^0) = \left(p_2^0\right)^{x_1} \left(1 - p_2^0\right)^{1-x_1} \left(p_1^0\right)^{x_2} \left(1 - p_1^0\right)^{1-x_2}, x_1 = 0, 1; x_2 = 0, 1.$$

Припустимо, що спостереження проводимо парами, кожна пара складається з одного спостереження величини ξ та одного спостереження величини η . Для проведення перевірки необхідно вибрати дві постійні B та A і на кожній стадії експерименту рахувати величину відношення

$$\frac{p_{1n}}{p_{0n}} = \frac{P(x_{11}, p_2^0)P(x_{21}, p_1^0)P(x_{12}, p_2^0)P(x_{22}, p_1^0) \dots P(x_{1n}, p_2^0)P(x_{2n}, p_1^0)}{P(x_{11}, p_1^0)P(x_{21}, p_2^0)P(x_{12}, p_1^0)P(x_{22}, p_2^0) \dots P(x_{1n}, p_1^0)P(x_{2n}, p_2^0)},$$

де x_{11} - перше спостереження величини ξ ; x_{21} - перше спостереження величини η ; x_{12} - друге спостереження величини ξ ; x_{22} - друге спостереження величини η і т. д.

Експеримент продовжуємо до тих пір, поки відношення $\frac{P_{1n}}{P_{0n}}$ залишиться в проміжку між B та A . Гіпотеза H_0 приймається, якщо

$$\frac{P_{1n}}{P_{0n}} \leq B, \text{ і відхиляється, якщо } \frac{P_{1n}}{P_{0n}} \geq A.$$

Спільний розподіл за справедливої гіпотези H_0 дорівнює

$$\begin{aligned} P_{1n} &= P(x_{11}, p_2^0)P(x_{21}, p_1^0) \dots P(x_{1n}, p_2^0)P(x_{2n}, p_1^0) = \\ &= (p_2^0)^{x_{11}} (1 - p_2^0)^{1-x_{11}} (p_1^0)^{x_{21}} (1 - p_1^0)^{1-x_{21}} \dots (p_2^0)^{x_{1n}} (1 - p_2^0)^{1-x_{1n}} (p_1^0)^{x_{2n}} (1 - p_1^0)^{1-x_{2n}}. \end{aligned}$$

Спільний розподіл за справедливої гіпотези H_1 дорівнює

$$\begin{aligned} P_{0n} &= P(x_{11}, p_1^0)P(x_{21}, p_2^0) \dots P(x_{1n}, p_1^0)P(x_{2n}, p_2^0) = \\ &= (p_1^0)^{x_{11}} (1 - p_1^0)^{1-x_{11}} (p_2^0)^{x_{21}} (1 - p_2^0)^{1-x_{21}} \dots (p_1^0)^{x_{1n}} (1 - p_1^0)^{1-x_{1n}} (p_2^0)^{x_{2n}} (1 - p_2^0)^{1-x_{2n}}. \end{aligned}$$

Відношення ймовірностей набуває вигляду

$$\frac{P_{1n}}{P_{0n}} = \left(\frac{p_2^0(1 - p_1^0)}{p_1^0(1 - p_2^0)} \right)^{\sum x_{1i} - \sum x_{2i}}.$$

Логарифм відношення ймовірностей дорівнює

$$\ln \frac{P_{1n}}{P_{0n}} = (\sum x_{1i} - \sum x_{2i}) \ln \left(\frac{p_2^0(1 - p_1^0)}{p_1^0(1 - p_2^0)} \right).$$

М. Girshick [6] установив, що ймовірність того, що послідовний критерій відношення ймовірностей закінчиться прийняттям гіпотези H_0 , залежить лише від величини

$$v(p_1, p_2) = \ln \left(\frac{p_1(1 - p_2)}{p_2(1 - p_1)} \right).$$

Цю величину можна вважати допустимою мірою відхилення p_1 від p_2 .

Функція $v(p_1, p_2)$ задовольняє умови: $v(p_1, p_2) = 0$, коли $p_1 = p_2$,

$v(p_1, p_2) < 0$, коли $p_2 > p_1$, $v(p_1, p_2) = -v(p_2, p_1)$.

Вибір чотирьох величин p_1^0, p_2^0, A, B можна зробити, вдаючись до таких міркувань. Нехай d така додатна величина, що прийняття гіпотези H_0 можна вважати помилкою для всіх $v(p_1, p_2) \geq d$, а відхилення гіпотези H_0 – помилкою для всіх $v(p_1, p_2) \leq -d$. Ми прагнемо здобути таку методику перевірки, для якої б імовірність відхилення гіпотези H_0 для всіх $v(p_1, p_2) \leq -d$ не перевищувала заздалегідь заданої величини α , а імовірність прийняття гіпотези H_0 для всіх $v(p_1, p_2) \geq d$ не перевищувала заздалегідь заданої величини β . Методика перевірки матиме такі властивості, якщо величини p_1^0, p_2^0, A, B вибрані так, що

$$v(p_1^0, p_2^0) = \ln \left(\frac{p_1^0(1-p_2^0)}{p_2^0(1-p_1^0)} \right) = -d.$$

І послідовний критерій відношення ймовірностей, застосований до перевірки гіпотези H_0 відносно гіпотези H_1 має силу (α, β) .

Логарифм відношення ймовірностей набуває вигляду

$$\ln \frac{p_{1n}}{p_{0n}} = (\sum x_{1i} - \sum x_{2i}) \ln \left(\frac{p_2^0(1-p_1^0)}{p_1^0(1-p_2^0)} \right) = d(\sum x_{1i} - \sum x_{2i}).$$

Продовжуємо моделювати пари спостережень, поки

$$\frac{\ln B}{d} < \sum_{i=1}^n (x_{1i} - x_{2i}) < \frac{\ln A}{d},$$

$$\frac{1}{d} \ln \left(\frac{\beta}{1-\alpha} \right) < \sum_{i=1}^n (x_{1i} - x_{2i}) < \frac{1}{d} \ln \left(\frac{1-\beta}{\alpha} \right).$$

Приймаємо гіпотезу $H_0 : p_1 = p_1^0, p_2 = p_2^0$, якщо

$$\sum_{i=1}^n (x_{1i} - x_{2i}) \leq \frac{1}{d} \ln \left(\frac{\beta}{1-\alpha} \right),$$

і відхиляємо гіпотезу $H_0 : p_1 = p_1^0, p_2 = p_2^0$, якщо

$$\sum_{i=1}^n (x_{1i} - x_{2i}) \geq \frac{1}{d} \ln \left(\frac{1-\beta}{\alpha} \right).$$

Реалізація А/В-тестування. У разі А/В тестування контрольній групі відвідувачів пропонуємо до перегляду сторінку А, експериментальній групі відвідувачів пропонують для перегляду сторінку В (варіацію А). Ви-суваємо нульову гіпотезу про однаковий рівень конверсії в обох групах $H_0 : p_1 \leq p_2$. Моделюємо потік відвідувачів сайту. Кожен відвідувач з імовірністю $\frac{1}{2}$ може потрапити до контрольної групи і з такою ж імовірністю може потрапити до експериментальної групи. Після того як відвідувач опинився в одній із двох груп моделюємо його поведінку. Поведінку відвідувача однозначно визначаємо двома подіями: успіх – відвідувач виконав цільову дію, неуспіх – відвідувач не виконав цільову дію. Якщо відвідувач опинився в контрольній групі, то подія «успіх» відбувається з імовірністю p_{cont} . Якщо відвідувач опинився в експериментальній групі, то подія «успіх» відбувається з імовірністю p_{exp} . За кожного $n = n_{cont} + n_{exp}$ перевіряємо нульову гіпотезу $H_0 : p_1 = p_1^0, p_2 = p_2^0$.

На рис. 3 зображено результати реалізацій А/В тестування за таких параметрів: $p_1^0 = 0,1, p_2^0 = 0,11, \alpha = 0,05, \beta = 0,2$.

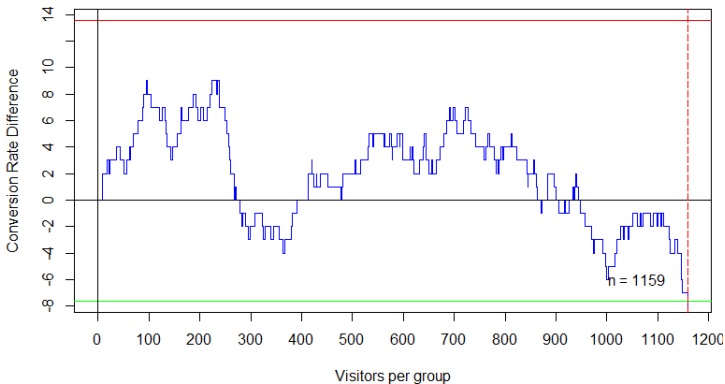


Рис. 3. Тест 1. Довірчий інтервал для Conversion Rate Difference. Тестування завершено прийняттям нульової гіпотези на 1159 кроці

На рис. 4 зображено результати реалізацій A/B тестування за таких параметрів: $p_1^0 = 0,1$, $p_2^0 = 0,12$, $\alpha = 0,05$, $\beta = 0,2$.

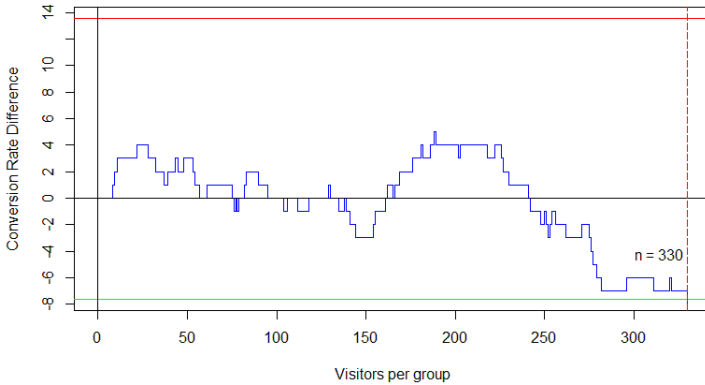


Рис. 4. Тест 2. Довірчий інтервал для Conversion Rate Difference. Тестування завершено прийняттям нульової гіпотези на 330 кроці

На рис. 5 зображено результати реалізацій A/B тестування за таких параметрів: $p_1^0 = 0,1$, $p_2^0 = 0,14$, $\alpha = 0,05$, $\beta = 0,2$.

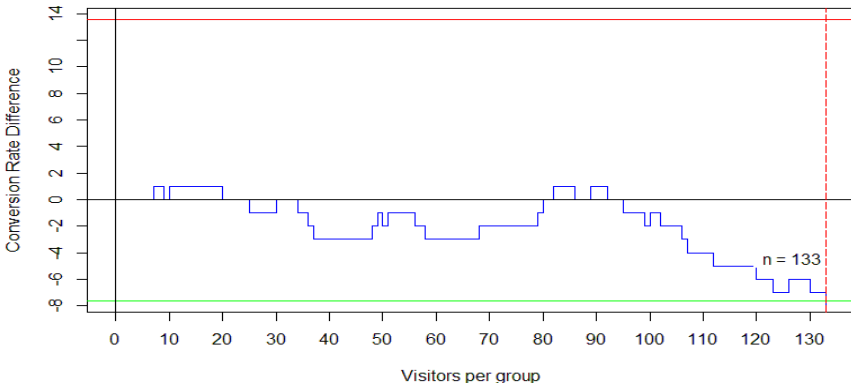


Рис. 5. Тест 3. Довірчий інтервал для Conversion Rate Difference. Тестування завершено прийняттям нульової гіпотези на 133 кроці

5. Порівняння класичної та послідовної перевірок статистичних гіпотез у разі А/В-тестування

За класичної перевірки статистичних гіпотез за допомогою критерію χ^2 мінімальний обсяг вибірок контрольної та експериментальної груп, який забезпечує задані ймовірність помилки I роду α , потужність критерію $1 - \beta$, різницю рівнів конверсії θ та базовий рівень конверсії p у випадку перевірки простої гіпотези $H_0 : p_{cont} - p_{exp} = 0$ проти простої альтернативи $H_0 : p_{cont} - p_{exp} = \theta$, дорівнює

$$n = \frac{2(z_{1-\alpha/2} + z_{1-\beta})^2 p(1-p)}{\theta^2}.$$

Цю фіксовану кількість спостережень n , необхідних у випадку класичної перевірки гіпотез, порівнюємо з середньою кількістю спостережень \hat{n} , необхідних за послідовної перевірки гіпотез (див. таблицю).

Порівняння критерію χ^2 та послідовного критерію для 1000 тестувань

Гіпотеза, що перевіряється $H_0 : p_{cont} < p_{exp}$		$\alpha = 0,05 \quad \beta = 0,2$		Показник зменшення кількості відвідувачів, %
		Критерій χ^2 n	Критерій Girshick \hat{n}	
$p_{cont}^0 = 0,1$	$p_{exp}^0 = 0,11$	14132	1017	92
$p_{cont}^0 = 0,1$	$p_{exp}^0 = 0,12$	3533	337	90
$p_{cont}^0 = 0,1$	$p_{exp}^0 = 0,13$	1571	175	88
$p_{cont}^0 = 0,1$	$p_{exp}^0 = 0,14$	884	105	88
$p_{cont}^0 = 0,1$	$p_{exp}^0 = 0,15$	566	70	87
$p_{cont}^0 = 0,1$	$p_{exp}^0 = 0,16$	393	47	88
$p_{cont}^0 = 0,1$	$p_{exp}^0 = 0,17$	289	36	87

Закінчення таблиці

$p_{cont}^0 = 0,1$	$p_{exp}^0 = 0,18$	221	35	84
$p_{cont}^0 = 0,1$	$p_{exp}^0 = 0,19$	175	31	82
$p_{cont}^0 = 0,1$	$p_{exp}^0 = 0,20$	142	19	86
$p_{cont}^0 = 0,1$	$p_{exp}^0 = 0,21$	117	18	84
$p_{cont}^0 = 0,1$	$p_{exp}^0 = 0,22$	99	15	84
$p_{cont}^0 = 0,1$	$p_{exp}^0 = 0,23$	84	14	83
$p_{cont}^0 = 0,1$	$p_{exp}^0 = 0,24$	73	14	80
$p_{cont}^0 = 0,1$	$p_{exp}^0 = 0,11$	63	13	79

6. Висновки

У роботі розглянуто класичну та послідовну перевірки статистичних гіпотез та проілюстровано переваги послідовного аналізу в задачі А/В-тестування сторінок сайту, а саме: у разі застосування послідовного критерію в середньому необхідна менша кількість спостережень, ніж за класичного критерію χ^2 з фіксованою кількістю спостережень тієї ж сили. Тому метод послідовного аналізу має широкую сферу застосування і є ефективним в різних галузях, наприклад під час вибіркового контролю продукції або за статистичної обробки результатів фізичних експериментів.

Бібліографічні посилання

1. **Pekelis, L.** The New Stats Engine [Електронний ресурс] / L. Pekelis, D. Walsh, R. Johari. – Режим доступу: http://pages.optimizely.com/rs/optimizely/images/stats_engine_technical_paper.pdf
2. **Lai, T.** Sequential analysis: Some classical problems and new challenges [Text] / T. Lai. – Statistica Sinica, 2001. – P. 303–408.
3. **Robbins, H.** Statistical methods related to the law of the iterated logarithm [Text] / H. Robbins // Ann. Math. Statist. – 1970. – P. 1397–1409.
4. **Siegmund, D.** Sequential Analysis: Tests and Confidence Intervals [Text] / D. Siegmund. – Springer, 1985.
5. **Вальд, А.** Последовательный анализ [Текст] / А. Вальд. – М., 1960.
6. **Girshick, M.A.** Contributions of the theory of sequential analysis [Електронний ресурс] / M.A. Girshick. – Режим доступу: <http://projecteuclid.org/euclid.aoms/1177730976>. – Загл. з екрана.
7. **Ван дер Варден, Б.Л.** Математическая статистика [Текст] / Б.Л. Ван дер Варден. – М., 1960.

Надійшла до редколегії 21.05.2016