RON KOHAVI DIANE TANG YA XU

KONTROLOWANE EKSPERYMENTY ONLINE

PRAKTYCZNY PRZEWODNIK PO TESTACH A/B

Helion

Tytuł oryginału: Trustworthy Online Controlled Experiments: A Practical Guide to A/B Testing

Tłumaczenie: Katarzyna Ellerik

Projekt okładki: Studio Gravite / Olsztyn; Obarek, Pokoński, Pazdrijowski, Zaprucki
Materiały graficzne na okładce zostały wykorzystane za zgodą Adobe Stock.

ISBN: 978-83-289-0763-8

Drogi Czytelniku!
Jeżeli chcesz ocenić tę książkę, zajrzyj pod adres
https://helion.pl/user/opinie/koekon
Możesz tam wpisać swoje uwagi, spostrzeżenia, recenzję.

Helion S.A.
ul. Kościuszki 1c, 44-100 Gliwice
tel. 32 230 98 63
e-mail: *helion@helion.pl*
WWW: *https://helion.pl* (księgarnia internetowa, katalog książek)

Printed in Poland.

Spis treści

CZĘŚĆ III. Techniki uzupełniające i alternatywne w stosunku do eksperymentów kontrolowanych

CZĘŚĆ IV. Zaawansowane zagadnienia związane z budowaniem platformy do eksperymentowania

Rozdział 2.

Przeprowadzanie i analizowanie eksperymentów — kompleksowy przykład

> Im mniej faktów, tym silniejsze opinie.
>
> — *Arnold Glasow*

W rozdziale 1. wyjaśniliśmy, czym są eksperymenty kontrolowane, i przedstawiliśmy wagę pozyskiwania rzeczywistych danych w podejmowaniu decyzji zamiast polegania na intuicji. Przykład w tym rozdziale obrazuje podstawowe zasady projektowania, przeprowadzania i analizowania eksperymentów. Mają one zastosowanie niezależnie od tego, jak wdrożone jest oprogramowanie: na serwerach webowych czy w przeglądarkach, w postaci aplikacji desktopowych czy mobilnych, konsol do gier czy cyfrowych asystentów. Aby zachować prostotę i przejrzystość, skupiamy się na przykładzie optymalizacji strony internetowej. W rozdziale 12. omawiamy różnice w przeprowadzaniu eksperymentów na tzw. *grubych klientach*, czyli na przykład natywnych aplikacjach desktopowych czy mobilnych.

Kontekst przykładu

Naszym konkretnym przykładem jest fikcyjna strona sklepu internetowego sprzedającego widgety. Zmian, które możemy przetestować, jest wiele: wprowadzenie nowej funkcjonalności, zmiany w interfejsie użytkownika, aktualizacje serwerowe itd.

W naszym przykładzie zespół ds. marketingu chce podnieść statystyki sprzedaży poprzez rozsyłanie promocyjnych wiadomości e-mail zawierających kody zniżkowe do zrealizowania przy zakupie widgetów. Ta zmiana to potencjalna zmiana modelu biznesowego, ponieważ firma nie oferowała wcześniej kuponów rabatowych. Jeden z pracowników przeczytał jednak ostatnio o firmie Dr. Footcare, która straciła znaczną część przychodów po wprowadzeniu takich zniżek (Kohavi, Longbottom i inni, 2009), a także

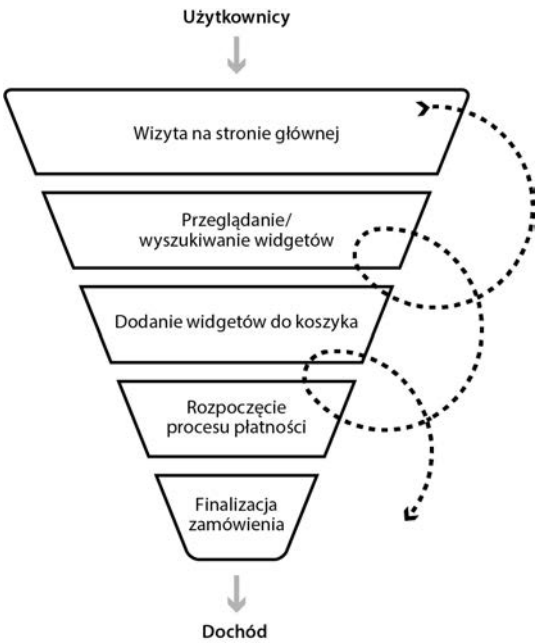
że *usuwanie* kuponów rabatowych to dobra praktyka według GoodUI.org (Linowski, 2018). Biorąc pod uwagę te zewnętrzne dane, powstały wątpliwości, czy dodanie pola zniżkowego przy zakupach nie obniży przychodu nawet w sytuacji, gdy kuponów nie będzie — czyli gdy użytkownicy tylko zobaczą pole, zwolnią interakcje w poszukiwaniu zniżek, a może nawet zupełnie je porzucą.

Chcemy ocenić wpływ samego dodania pola kodu zniżkowego. Możemy skorzystać z podejścia atrapy drzwi czy malowanych drzwi (Lee, 2013) — analogia opiera się na zbudowaniu lub namalowaniu niedziałających drzwi i sprawdzeniu, ile osób będzie próbowało je otworzyć. W tym przypadku implementujemy trywialną zmianę w postaci dodania pola z kodem rabatowym w procesie finalizowania zamówienia. Nie implementujemy całego, prawdziwego systemu rabatowego, bo nie ma żadnych aktywnych kuponów. Jakąkolwiek podałby użytkownik, system zwróci komunikat typu „Niepoprawny format kodu". Naszym celem jest tylko oszacowanie wpływu posiadania takiego pola na dochód i odniesienie się do zagrożenia, że odwróci ono uwagę użytkowników od dokończenia procesu zakupowego. Ponieważ jest to prosta zmiana, przetestujemy dwie implementacje interfejsu użytkownika. Powszechną praktyką jest jednoczesne testowanie wielu grup eksperymentalnych, by móc ocenić pomysł, a nie implementację. W naszym przykładzie pomysłem jest dodanie kodu rabatowego, a implementacją konkretna zmiana w UI.

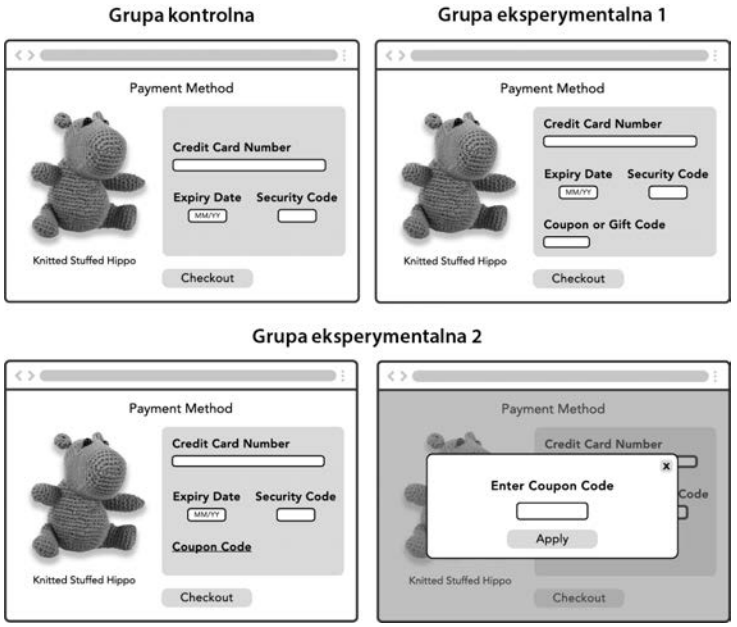
Ten prosty test A/B jest podstawowym krokiem oceny wykonalności nowego modelu biznesowego.

Podczas przenoszenia proponowanej zmiany interfejsu na hipotezę warto widzieć proces zakupowy jako lejek (ang. *funnel*), który został przedstawiony na rysunku 2.1. Klient zaczyna proces na stronie głównej, przegląda kilka widgetów, dodaje któryś z nich do koszyka, zaczyna proces płatności i wreszcie finalizuje zamówienie. Oczywiście koncepcja lejka jest uproszczeniem — klienci rzadko przechodzą przez te kroki całkowicie liniowo. W rzeczywistości pojawia się wiele pętli między stanami, niektórzy powracający użytkownicy pomijają pewne kroki itd. Niemniej ten prosty model jest pożyteczny przy przemyśleniu projektu eksperymentu i analizie, ponieważ eksperymenty zwykle mają na celu optymalizację któregoś z poziomów lejka (McClure, 2007).

W naszym eksperymencie dodajemy pole z kodem rabatowym na stronie podsumowania zamówienia i testujemy dwa różne interfejsy użytkownika (rysunek 2.2). Chcemy oszacować wpływ zmiany dochód. Nasza hipoteza brzmi: dodanie pola z kodem rabatowym na stronie podsumowania zamówienia obniży dochód.



Rysunek 2.1. Lejek zakupowy w przypadku użytkownika online. Użytkownicy nie muszą przechodzić przez niego liniowo — mogą pomijać lub powtarzać kroki, a także krążyć między nimi



Rysunek 2.2. (1) Grupa kontrolna — stara wersja strony. (2) Grupa eksperymentalna 1 — pole kodu rabatowego lub karty podarunkowej poniżej danych karty płatniczej. (3) Grupa eksperymentalna 2 — pole kodu rabatowego lub karty podarunkowej jako wyskakujące okno

Aby zmierzyć wpływ zmiany, musimy zdefiniować wskaźnik celu (wskaźnik powodzenia). Jeśli mamy tylko jeden taki wskaźnik, możemy bezpośrednio użyć go jako OKE (zobacz rozdział 7.). Możliwością, która szybko przychodzi do głowy, jest oczywiście dochód. Zwróć uwagę, że chociaż chcemy zwiększyć ogólny dochód, to nie zalecamy posługiwania się samą sumą dochodu, ponieważ zależy ona od liczby użytkowników w każdym wariancie. Nawet jeśli wariantom przypisze się równy rozkład ruchu, to rzeczywista liczba użytkowników może się różnić choćby przez przypadek. Sugerujemy normalizowanie wskaźników przez wielkość próbki, co oznacza, że *dochód na użytkownika* będzie dobrym OKE.

Następnym ważnym pytaniem jest określenie, których użytkowników brać pod uwagę w mianowniku wskaźnika dochodu na użytkownika:

- *Wszyscy użytkownicy, którzy odwiedzili stronę.* Wydaje się to uzasadnionym zachowaniem, ale generuje też spory szum, ponieważ obejmuje także użytkowników, którzy nie rozpoczęli nawet procesu zakupowego, w którym wprowadzono zmianę. Wiemy, że na tych użytkowników, którzy nie rozpoczęli procesu, zmiana nie mogła mieć wpływu. Wyłączenie tych użytkowników da nam bardziej czuły test A/B (zobacz rozdział 20.).
- *Tylko użytkownicy, którzy zakończyli proces zakupowy.* Ten wybór byłby nieprawidłowy, ponieważ zakłada, że zmiana wpłynie na kupowane ilości, a nie procent użytkowników kończących transakcję. Jeśli więcej użytkowników będzie kupować, dochód na użytkownika może spaść, mimo że ogólny dochód wzrośnie.
- *Tylko użytkownicy, którzy rozpoczęli proces zakupowy.* To najlepszy wybór, biorąc pod uwagę umiejscowienie zmiany w lejku sprzedażowym. Obejmujemy analizą wszystkich użytkowników, na których potencjalnie wpłynęła zmiana, ale wykluczamy tych, którzy nie mieli z nią styczności (którzy nie rozpoczęli procesu składania zamówienia), co daje nam najbardziej skoncentrowany wynik.

Nasza doprecyzowana hipoteza brzmi teraz tak: dodanie pola z kodem rabatowym na stronie podsumowania zamówienia obniży dochód w przeliczeniu na użytkownika wśród użytkowników, którzy rozpoczęli proces zakupowy.

Testowanie hipotezy — zdobycie wagi statystycznej

Zanim zaczniemy projektować, przeprowadzać czy analizować eksperyment, powinniśmy zapoznać się z kilkoma podstawowymi pojęciami związanymi ze statystycznym testowaniem hipotez.

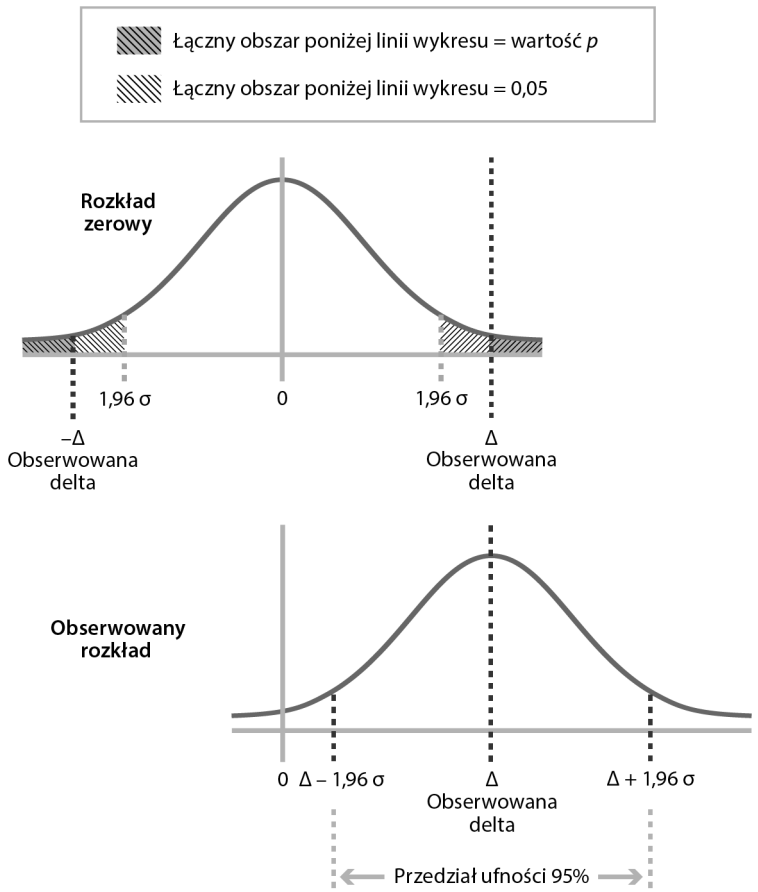
Przede wszystkim charakteryzujemy wskaźnik przez zrozumienie linii bazowej *średniej* wartości oraz *standardowego błędu* średniej, czyli tego, jak będzie się zmieniać szacowana

wartość naszego wskaźnika. Musimy znać tę zmienność, by prawidłowo dobrać rozmiar eksperymentu i wyliczyć istotność statystyczną podczas analizy. W przypadku większości eksperymentów liczy się średnią, ale można posługiwać się także inną statystyką zbiorczą, np. centylami. Czułość, czy też zdolność wykrywania statystycznie istotnych różnic, poprawia się, im mniejszy jest standardowy błąd. Zwykle można to osiągnąć przez przypisanie wariantom większego odsetka ruchu lub wydłużenie czasu trwania eksperymentu, ponieważ przeważnie liczba użytkowników rośnie z czasem. To drugie rozwiązanie może jednak nie być tak skuteczne jak pierwszych kilka tygodni, gdyż przyrost nowych użytkowników przebiega podliniowo ze względu na powracających użytkowników, podczas gdy niektóre wskaźniki same mają *rosnącą* zmienność w czasie (Kohavi i inni, 2012).

Podczas przeprowadzania eksperymentu nie opisujemy wskaźnika dla jednej próbki, lecz dla wielu. W szczególności dla eksperymentu kontrolowanego mamy jedną próbkę dla *grupy kontrolnej* i po jednej próbce dla każdej *grupy eksperymentalnej*. Ilościowo testujemy, czy różnica między parą grup eksperymentalnych i grupą kontrolną zachodzi przy założeniu **hipotezy zerowej**, że średnie są takie same. Jeśli zachodzi, to obalamy hipotezę zerową i możemy twierdzić, że różnica jest statystycznie istotna. W tym przypadku, biorąc pod uwagę szacunki dochodu na użytkownika w próbkach grup eksperymentalnych i grupy kontrolnej, obliczmy **wartość różnicy *p***, czyli prawdopodobieństwo zaobserwowania takiej różnicy. Obalamy hipotezę zerową i postulujemy, że eksperyment ma znaczenie (rezultat jest statystycznie istotny), gdy wartość *p* jest odpowiednio niska. „Odpowiednio" — czyli jak?

Standardem naukowym jest posługiwanie się wartością *p* poniżej 0,05, co oznacza, że jeśli rzeczywiście nie ma efektu, to możemy poprawnie wnioskować, iż nie będzie go w 95 przypadkach na 100. Innym sposobem sprawdzenia, czy różnica jest statystycznie istotna, jest sprawdzenie, czy **przedział ufności** jest jak najkrótszy. Przedział ufności rzędu 95% oznacza zakres obejmujący rzeczywistą różnicę w 95% przypadków i dla dość dużych próbek zwykle koncentruje się wokół obserwowanej delty między grupą kontrolną i grupą eksperymentalną poszerzoną o wartość 1,96 błędu standardowego po każdej ze stron. Rysunek 2.3 przedstawia równorzędność tych wskazań.

Moc statystyczna to prawdopodobieństwo wykrycia różnicy pomiędzy wariantami, gdy taka różnica rzeczywiście zachodzi (statystycznie obalamy hipotezę zerową, gdy jest różnica). W praktyce moc Twojego eksperymentu powinna pozwolić z dużym prawdopodobieństwem dojść do wniosku, że eksperyment poskutkował zmianą większą niż minimum, na jakie zwracałeś uwagę. Zwykle moc testu rośnie wraz z wielkością próbki. Powszechną praktyką jest projektowanie eksperymentów z mocą 80–90%. Więcej szczegółów statystycznych znajdziesz w rozdziale 17.



Rysunek 2.3. U góry: użycie wartości p do oceny, czy obserwowana delta jest statystycznie istotna. Jeśli wartość p jest mniejsza niż 0,05, różnicę uważa się za statystycznie istotną. Na dole: równorzędny pogląd posługujący się 95-procentowym przedziałem ufności [Δ – 1,96 σ, Δ + 1,96 σ] do oceny istotności statystycznej. Jeśli zero znajduje się poza przedziałem, przyjmuje się istotność statystyczną

Choć „istotność statystyczna" mierzy, jak prawdopodobne jest, że rezultat, jaki zaobserwowałeś (lub bardziej radykalny), jest dziełem przypadku, to nie każdy statystycznie istotny wynik będzie praktycznie znaczący. W przypadku dochodu na użytkownika jak duża różnica będzie miała dla nas znaczenie z biznesowego punktu widzenia? Innymi słowy, jaka różnica jest *praktycznie istotna*?

Ustalenie konkretnej granicy jest ważne dla zrozumienia, czy różnica warta jest poniesienia kosztów wprowadzenia zmiany. Jeśli strona generuje miliardy dolarów, jak Google czy Bing, to różnica 0,02% jest praktycznie istotna. Dla porównania startup może uznać nawet 2-procentową zmianę za zbyt małą, ponieważ celuje w różnice 10% lub nawet więcej. W naszym przykładzie przyjmijmy, że z perspektywy biznesowej wzrost dochodu w przeliczeniu na użytkownika 1% lub więcej będzie praktycznie istotny.

Projektowanie eksperymentu

Jesteśmy teraz gotowi, żeby zająć się projektowaniem eksperymentu. Mamy hipotezę i granicę praktycznej istotności i określiliśmy wskaźnik. Posłużymy się tym zestawem decyzji, by dokończyć projekt:

1. Jaką jednostkę randomizacji przyjmiemy?
2. Jaką populację jednostek randomizacji będziemy brali za cel?
3. Jak duży musi być eksperyment?
4. Jak długo powinien on trwać?

Na razie przyjmijmy, że jednostką randomizacji są *użytkownicy*. W rozdziale 14. omawiamy alternatywy, ale to właśnie *użytkownicy* są zdecydowanie najczęstszym wyborem.

Branie za cel określonej populacji oznacza, że chcesz poddać eksperymentowi tylko użytkowników o konkretnej charakterystyce. Przykładowo interesuje Cię przetestowanie nowego tekstu na stronie, ale w tym momencie dysponujesz tylko kilkoma wersjami językowymi nowej treści, więc możesz wybrać kierowanie eksperymentu tylko do użytkowników, których język interfejsu ustawiony jest na jeden z tych, które obecnie obsługujesz. Do innych popularnych kryteriów targetowania należą: region geograficzny, używana platforma i rodzaj urządzenia. W naszym przykładzie jednak zdecydujemy się targetować wszystkich użytkowników.

Rozmiar eksperymentu (w naszym przypadku liczba użytkowników) ma bezpośredni wpływ na precyzję wyników. Jeśli zależy Ci na wykryciu nawet niewielkiej zmiany lub pogłębionej pewności co do konkluzji, przeprowadź większy eksperyment z większą liczbą użytkowników. Oto kilka zmian, nad którymi warto się zastanowić:

- Jeśli posłużymy się *wskaźnikiem zakupu* (tzn. odpowiedzią tak/ nie na pytanie, czy użytkownik kupił produkt, bez związku z kwotą zakupu) zamiast *dochodem na użytkownika* jako OKE, standardowy błąd będzie mniejszy, co oznacza, że zachowując tę samą precyzję, będziemy mogli przeprowadzić eksperyment na mniejszej liczbie użytkowników.
- Jeśli podniesiemy poziom praktycznej istotności, co będzie oznaczało, że różnica 1% jest teraz za mała i interesują nas tylko większe zmiany, również będziemy mogli przeprowadzić eksperyment na mniejszej liczbie użytkowników, bo większe zmiany łatwiej wykryć.
- Jeśli chcemy obniżyć wartość p na przykład do 0,01, aby mieć większą pewność, że zmiana rzeczywiście zachodzi przed odrzucaniem hipotezy zerowej, musimy zwiększyć próbkę.

Oto kilka dodatkowych uwag na temat wielkości eksperymentu:

- Na ile bezpieczny jest eksperyment? W przypadku dużych zmian, co do których nie jesteś pewien reakcji użytkowników, możesz rozpocząć od mniejszej proporcji użytkowników w grupie eksperymentalnej. To rozumowanie nie powinno mieć wpływu na wybór ostatecznego rozmiaru eksperymentu, lecz jedynie na sposób jego rozwijania się w czasie (w rozdziale 15. znajdziesz więcej szczegółów na ten temat).
- Czy eksperyment musi współdzielić ruch z innymi testami? Jeśli tak, jak rozłożyć wymagania względem tego ruchu? Patrząc ogólnie, jeśli masz do przetestowania kilka zmian, możesz wybrać zrobienie tego jednocześnie lub sekwencyjnie. Jeśli musisz podzielić ruch na kilka oddzielnych, równoległych testów, do każdego z nich zostanie skierowany proporcjonalnie mniejszy ruch. W rozdziale 4. omawiamy przeprowadzanie eksperymentów w jednej warstwie lub w sposób nachodzący na siebie oraz, co ważniejsze, jak budować infrastrukturę do zeskalowania wszystkich eksperymentów.

Kolejnym ważnym pytaniem jest kwestia czasu trwania eksperymentu. W tym przypadku należy zastanowić się nad następującymi kwestiami:

- **Większa liczba użytkowników.** Ze względu na fakt, że użytkownicy są włączani do eksperymentów online z biegiem czasu, im dłużej trwa eksperyment, tym więcej użytkowników go zobaczy. Zwykle oznacza to wyższą moc statystyczną eksperymentu (wyjątek stanowi sytuacja, gdy mierzony wskaźnik ma charakter zbiorczy, np. liczba sesji, a zmienność również rośnie; w rozdziale 18. znajdziesz więcej szczegółów). Poziom napływu użytkowników w czasie również może być podliniowy, ponieważ ci sami użytkownicy mogą powracać: jeśli pewnego dnia masz N użytkowników, kolejnego dnia będziesz mieć mniej niż 2N użytkowników, ponieważ niektórzy z nich odwiedzą stronę w oba dni.
- **Efekt dni tygodnia.** Populacja użytkowników może być inna w ciągu tygodnia niż w weekendy. Nawet ci sami użytkownicy mogą zachowywać się inaczej. Eksperyment powinien brać pod uwagę cykl tygodniowy. Zalecamy przeprowadzanie eksperymentów przez co najmniej tydzień.
- **Sezonowość**. Mogą zdarzyć się również inne okresy, gdy zachowanie użytkowników odbiega od normy, np. wakacje i święta. Jeśli baza użytkowników obejmuje cały świat, musisz liczyć się z wpływem kalendarza świąt w USA i innych miejscach globu. Przykładowo sprzedaż kart upominkowych może wypaść dobrze przed świętami Bożego Narodzenia, ale nie w pozostałej części roku. Mamy tu do czynienia ze zjawiskiem *trafności zewnętrznej* — stopnia, do którego możemy generalizować wyniki, w naszym przypadku na inny zakres dat.

- **Efekt pierwszeństwa i nowości**. Zdarzają się eksperymenty, które miewają mniejszy lub większy wpływ początkowo i wymagają czasu na stabilizację. Przykładowo użytkownicy mogą wypróbować nowy, rzucający się w oczy przycisk i dojść do wniosku, że nie jest przydatny, więc jego klikalność z czasem spadnie. Z drugiej strony funkcjonalności, które wymagają przyjęcia się, potrzebują czasu na zbudowanie bazy użytkowników.

Biorąc to wszystko pod uwagę, w tej chwili nasz eksperyment można opisać następująco:

1. Jednostką randomizacji jest użytkownik.
2. Będziemy brali na cel wszystkich użytkowników, ale analizie będą podlegali tylko ci, którzy weszli na stronę podsumowania zamówienia.
3. Aby osiągnąć 80-procentową moc wykrycia 1% różnicy w dochodzie na użytkownika, przeprowadzimy analizę mocy, by ustalić rozmiar eksperymentu.
4. Oznacza to przeprowadzenie eksperymentu przez co najmniej cztery dni z podziałem 34% – 33% – 33% między grupę kontrolną, grupę eksperymentalną 1 i grupę eksperymentalną 2. Eksperyment potrwa jednak cały tydzień, by była pewność, że uwzględnia efekt dni tygodnia, a potencjalnie nawet dłużej, jeśli będziemy obserwować efekt pierwszeństwa lub nowości.

Uogólniając, uzyskanie wyższej mocy eksperymentu niż to niezbędne nie jest szkodliwe, a może być wręcz zalecane, ponieważ czasami może pojawić się potrzeba zbadania pewnego segmentu (np. regionu geograficznego lub platformy) lub zagwarantowania, że eksperyment wystarczy do wykrycia zmian w kilku kluczowych wskaźnikach. Przykładowo być może będziemy dysponowali wystarczającą mocą, by wykryć wpływ na dochód wśród wszystkich użytkowników, ale zbyt małą, gdybyśmy chcieli ocenić ten wpływ tylko wśród populacji w Kanadzie. Zwróć też uwagę, że wybraliśmy mniej więcej równe rozmiary grupy kontrolnej i grup eksperymentalnych. Jednak jeśli grup eksperymentalnych będzie więcej, warto pomyśleć nad poszerzeniem grupy kontrolnej tak, by była większa niż eksperymentalne (rozdział 18. zawiera szersze omówienie tego tematu).

Przeprowadzanie eksperymentu i pozyskanie danych

Pora przystąpić do eksperymentu i zebrać niezbędne dane. Tutaj zamieszczamy krótki opis elementów wchodzących w jego skład, a więcej szczegółów znajdziesz w rozdziale 4., w punkcie „Skalowanie eksperymentów — szczegóły przypisywania wariantów".

Aby przeprowadzić eksperyment, potrzebujemy:

- **Instrumentacji**, by uzyskać dane z logów na temat sposobu interakcji użytkowników ze stroną oraz na temat tego, do którego eksperymentu te interakcje się kwalifikują.
- **Infrastruktury** umożliwiającej wykonanie testów, od konfiguracji eksperymentu po przypisanie wariantów. Więcej szczegółów znajdziesz w rozdziale 4., „Platforma i kultura eksperymentowania".

Po przeprowadzeniu eksperymentu i zebraniu danych z logów za pomocą niezbędnej instrumentacji przychodzi czas na przetworzenie danych, wyliczenie podsumowującej statystyki i zwizualizowanie rezultatów (zobacz rozdział 4. i rozdział 16.).

Interpretowanie rezultatów

Mamy dane z eksperymentu! Zanim przyjrzymy się wskaźnikowi dochodu w przeliczeniu na użytkownika, warto przeprowadzić kilka weryfikacji, by upewnić się, że eksperyment przebiegł prawidłowo.

Jest wiele miejsc, w które mogą wkraść się usterki skutkujące unieważnieniem wyników eksperymentu. Aby je dostrzec, musimy śledzić *wskaźniki ochronne* czy *niezmienniki*. Te wskaźniki nie powinny różnić się między grupą kontrolną i grupą eksperymentalną. Jeśli się jednak zmieniają, jakakolwiek zmierzona różnica jest prawdopodobnie wynikiem innych wprowadzonych modyfikacji, a nie testowanej funkcjonalności.

Wyróżniamy dwa rodzaje wskaźników niezmiennych:

1. Wskaźniki ochronne oparte na zaufaniu, takie jak oczekiwanie, że próbki w grupie kontrolnej i grupie eksperymentalnej będą miały wielkość zgodną z konfiguracją czy ten sam stosunek odczytu z pamięci podręcznej.
2. Organizacyjne wskaźniki ochronne — takie jak opóźnienie — czyli te, które są ważne dla organizacji i oczekuje się, że pozostaną niezmienne dla wielu eksperymentów. W eksperymencie dotyczącym pola kodu rabatowego zmiana opóźnienia byłaby wielkim zaskoczeniem.

Jeśli te weryfikacje nie wypadają pozytywnie, prawdopodobnie problem tkwi w projekcie eksperymentu, infrastrukturze lub przetwarzaniu danych. Więcej informacji na ten temat znajdziesz w rozdziale 21.

Po przeprowadzeniu weryfikacji opartych na wskaźnikach ochronnych możemy przyjrzeć się rezultatom (zobacz tabelę 2.1).

Ponieważ wartość p jest mniejsza niż 0,05 dla obu grup eksperymentalnych, obalamy hipotezę zerową, że grupy eksperymentalne i grupa kontrolna mają taką samą średnią.

Tabela 2.1. Wyniki wskaźnika dochodu na użytkownika z eksperymentu dotyczącego strony podsumowania zamówienia

	Dochód na użytkownika w grupie eksperymentalnej (w dol.)	**Dochód na użytkownika w grupie kontrolnej (w dol.)**	**Różnica**	**Wartość *p***	**Przedział ufności**
Grupa eksperymentalna 1 a grupa kontrolna	3,12	3,21	–0,09 (–2,8%)	0,0003	[–4,3%, –1,3%]
Grupa eksperymentalna 2 a grupa kontrolna	2,96	3,21	–0,25 (–7,8%)	1,5e-23	[–9,3%, –6,3%]

Co to znaczy? Otóż potwierdziliśmy wzorzec zakładający, że dodanie pola kodu rabatowego w interfejsie użytkownika obniży dochód. Jeśli pochylimy się bardziej nad liczbami, okaże się, że wyniki wskazują, iż spadek spowodowany jest mniejszą liczbą użytkowników kończących proces zakupowy. Co za tym idzie, każdy marketingowy e-mail z kodem rabatowym musi wynagradzać nie tylko koszty implementacji mechanizmu obsługi tych kodów oraz jego utrzymania, ale także negatywny wpływ samego dodania zniżek. W związku z tym, że model marketingowy przewidywał niewielki wzrost dochodu, ale test A/B pokazuje znaczny spadek dochodu w przypadku wszystkich użytkowników, zapada decyzja, by rozstać się z pomysłem kodów promocyjnych. Test A/B z malowanymi drzwiami oszczędził nam sporo pracy!

Od rezultatów do decyzji

Celem przeprowadzania testów A/B jest zebranie danych, by w oparciu o nie podjąć decyzję. Wiele wysiłku trzeba włożyć w to, aby wyniki były powtarzalne i wiarygodne tak, by ostateczny wybór był właściwy. Zastanówmy się nad procesem decyzyjnym dla kilku przypadków, z jakimi możemy się spotkać.

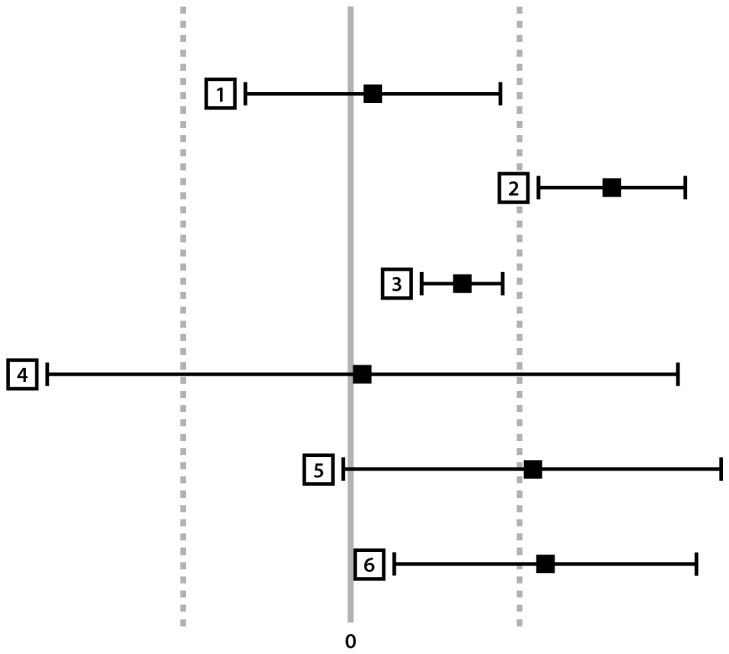
W każdym z nich dysponujemy rezultatami eksperymentu, a przyświeca nam cel przełożenia tych wyników na decyzję o wydaniu lub niewydaniu produktu. Powodem podkreślania kwestii procesu decyzyjnego jest fakt, że ostateczny wybór musi brać pod uwagę zarówno konkluzje płynące z pomiarów, jak i szerszy kontekst, który obejmuje:

- **Konieczność kompromisu między wskaźnikami**. Przykładowo jeśli zaangażowanie użytkowników rośnie, ale dochód spada, to czy powinniśmy wydawać produkt? Innym przykładem jest wzrost zużycia mocy procesora CPU, który może spowodować, że koszt funkcjonowania usługi przewyższy ewentualną korzyść ze zmiany.

- **Koszt wdrożenia zmiany**. Składa się na niego:
  - **Koszt pełnej implementacji produktu przed wydaniem**. Pewne funkcjonalności mogły zostać wytworzone w całości przed eksperymentem. W takiej sytuacji koszt wdrożenia funkcjonalności od 1% do 100% jest równy zeru. Jednak nie zawsze tak jest. W naszym przykładzie nakład związany z malowanymi drzwiami był niewielki, ale pełna implementacja będzie kosztowna.
  - **Koszt utrzymania po wdrożeniu**. Ten wydatek może wzrosnąć ze względu na nowy kod. Nowy kod zawiera zwykle więcej defektów i jest słabiej pokryty testowo dla przypadków brzegowych. Jeśli ponadto wprowadza dodatkową złożoność, może też zwiększyć koszt i uciążliwość budowania nowych funkcjonalności w przyszłości. Jeśli te wydatki są wysokie, musisz upewnić się, że oczekiwane zyski je pokryją. W takich przypadkach pilnuj, aby granica praktycznej istotności była wystarczająco restrykcyjna, by to zapewnić. Z drugiej strony, jeśli koszt jest bliski zeru, możesz wydać każdą dodatnią zmianę, więc granica praktycznej istotności będzie ustawiona nisko.
- **Konsekwencje błędnej decyzji**. Nie wszystkie decyzje są takie same i nie wszystkie pomyłki są sobie równe. Być może nie będzie negatywnych konsekwencji wydania zmiany, która nie przynosi żadnego efektu. Jednak konsekwencje utraty szansy w przypadku niewydania zmiany, która ma wpływ, mogą być spore. Przykładowo możesz testować dwie możliwe wiodące oferty na swojej stronie, ale każda z nich pozostanie aktywna tylko kilka dni. W tym przypadku dokonanie niewłaściwego wyboru może nie mieć konsekwencji, bo zmiana jest krótkotrwała. W takich okolicznościach możesz pozwolić sobie na obniżenie granicy zarówno praktycznej, jak i statystycznej istotności.

Musisz wziąć ten kontekst pod uwagę przy ustalaniu pułapu statystycznej i praktycznej istotności. Będzie on odgrywał kluczową rolę przy przechodzeniu od wyników eksperymentu do decyzji czy działań. Przy założeniu, że zaktualizowaliśmy ten poziom przed rozpoczęciem eksperymentu, by oddawał szerszy kontekst, przyjrzymy się przykładom z rysunku 2.4, aby wskazać, jak posługiwać się granicami w podejmowaniu decyzji.

1. Wynik nie jest statystycznie istotny. Jasne jest także, że nie ma żadnej praktycznej wagi. Prowadzi to do prostej konkluzji, że zmiana na niewiele się zdaje. Możesz zdecydować o dalszej pracy nad nią w kolejnej iteracji lub o porzuceniu pomysłu.
2. Wynik jest statystycznie i praktycznie istotny. To prosta decyzja: wdrażamy zmianę!



Rysunek 2.4. Przykłady zrozumienia praktycznej i statystycznej istotności w procesie podejmowania decyzji. Granice praktycznej istotności wyznaczają dwie przerywane linie. Szacowana różnica dla rezultatów każdego przypadku to czarny punkt na wykresie wraz ze swoim przedziałem ufności

3. Wynik jest statystycznie istotny, ale nie praktycznie. W tym przypadku masz pewność co do znaczenia zmiany, ale ta wielkość może nie być wystarczająca, by przeważyć inne czynniki, takie jak na przykład koszty. Prawdopodobnie nie warto wdrażać tej zmiany.
4. Ten przykład może być neutralny, tak jak pierwszy, ale przedziały ufności leżą poza granicą tego, co praktycznie istotne. Jeśli przeprowadzisz eksperyment i dowiesz się, że możesz zwiększyć lub zmniejszyć dochód o 10%, czy naprawdę przyjąłbyś wyniki eksperymentu i uznał zmianę za neutralną? Lepiej przyznać, że nie dysponujesz odpowiednią mocą do wyciągnięcia mocnych wniosków i nie masz wystarczających danych, żeby podjąć decyzję o wdrożeniu. Dla takiego przypadku zalecamy przeprowadzanie dalszych testów na większej liczbie jednostek, co da większą moc statystyczną.
5. Ten rezultat oznacza prawdopodobną istotność praktyczną, ale nie statystyczną. Co za tym idzie, choć w szczęśliwym scenariuszu możesz mieć nadzieję, że zmiana ma wpływ, to może się też okazać, iż nie ma żadnego. Z perspektywy pomiarowej najlepszym zaleceniem byłoby powtórzyć ten test, ale z większą mocą, by uzyskać większą precyzję rezultatu.

6. Rezultat jest statystycznie istotny i prawdopodobnie praktycznie istotny. Podobnie jak w powyższym przypadku, jest możliwe, że zmiana nie jest praktycznie istotna. Tym samym, tak jak wyżej, również tu zalecamy powtórzenie testu z większą mocą. Jednak z perspektywy decyzji o wdrożeniu zdecydowanie się na wdrożenie będzie rozsądne.

Najważniejsze jest, aby pamiętać, że dojdzie do sytuacji, w których będziesz musiał podjąć decyzję, mimo że wyniki eksperymentu mogą nie dawać jasnej odpowiedzi. W takich przypadkach musisz otwarcie mówić o czynnikach, jakie bierzesz pod uwagę, a zwłaszcza o tym, jak przekładają się one na granice statystycznej i praktycznej istotności. Będzie to stanowiło podstawę do przyszłych decyzji, a nie jedynie chwilowy kaprys.

Skorowidz

F

G

H

I

J

K

L

M

N

O

P

R

S

T

W

Z

TRUDNO OD RAZU OCENIĆ WARTOŚĆ POMYSŁU. Tymczasem nawet małe zmiany mogą mieć olbrzymi wpływ na dochody organizacji. Aby się przekonać, jak zmiana sposobu prezentowania treści użytkownikowi wpłynie na jego wrażenia, najlepiej przeprowadzić eksperyment kontrolowany online. Tego rodzaju badania, zwane również testami A/B, są regularnie prowadzone przez największe firmy. Przed wykorzystaniem otrzymanych wyników w działalności biznesowej trzeba jednak poddać je walidacji.

TA KSIĄŻKA ZAWIERA szereg praktycznych wskazówek dotyczących przygotowania, przeprowadzania i oceniania wyników kontrolowanych eksperymentów online. Dzięki niej nauczysz się stosować naukowe podejście do formułowania założeń i oceny hipotez w testach A/B, dowiesz się także, jak sprawdzać wiarygodność wyników i interpretować je do celów dalszej pracy. Omówiono tu takie pułapki jak efekt przeniesienia, prawo Twymana, paradoks Simpsona i interakcji sieciowych, zaprezentowano również informacje ułatwiające zrozumienie praktycznych konsekwencji statystyki. Interesującą częścią książki jest opis skalowalnej platformy, która radykalnie zmniejsza całościowy koszt eksperymentu. Publikację docenią zarówno początkujący, jak i zaawansowani eksperymentatorzy, którzy wymagają wysokiej pewności uzyskanych wyników.

DR RON KOHAVI jest wiceprezesem i partnerem technologicznym w Airbnb. Wcześniej pracował w Microsofcie i Amazonie. Jego artykuły cytowano ponad 40 tysięcy razy.

DIANE TANG jest ekspertką Google w dziedzinie zeskalowanej analizy danych i infrastruktury, kontrolowanych eksperymentów online, a także systemów reklamowych. Ma na koncie liczne publikacje i patenty.

DR YA XU kieruje działem data science i eksperymentami w LinkedIn. Opublikowała szereg artykułów na temat eksperymentów. Często występuje na prestiżowych konferencjach, wykłada też na uczelniach wyższych.

POZYSKUJ DANE, KTÓRYM ZAUFASZ!

Helion

helion.pl

HELION SA
ul. Kościuszki 1c
44-100 Gliwice
tel.: 32 230 98 63
helion@helion.pl

KOD KORZYŚCI
Sięgnij po więcej! ▶

ISBN 978-83-289-0763-8

9 788328 907638

Cena: 79,00 zł