КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
ІМЕНІ ТАРАСА ШЕВЧЕНКА
Кафедра прикладної статистики

Кваліфікаційна робота

на здобуття ступеня бакалавра

за спеціальністю 124 «Системний аналіз»

на тему:

СТАТИСТИЧНИЙ АНАЛІЗ ТА ПЕРЕВІРКА ГІПОТЕЗ ЗА ДОПОМОГОЮ A/B ТЕСТІВ

Студента 4 курсу
Кравцева Володимир Олексійовича
Науковий керівник:
доцент, доктор фіз.-мат. наук
Розора І.В.

Робота заслухана на засіданні кафедри прикладної статистики та рекомендована до захисту в ДЕК, протокол № 11 від 06 червня 2022 р.

Завідувач кафедри прикладної статистики
професор, доктор фіз.-мат. наук Розора І.В.

Київ – 2022

ЗМІСТ

АНОТАЦІЇ

Дипломна робота складається зі вступу, 3 розділів, висновків, списку використаних джерел (20 найменувань). Загальний обсяг роботи становить 50 сторінок, основний текст роботи викладено на 40 сторінках.

Ключові слова: A/B тест, статистична гіпотеза, статистичний експеримент, перевірка гіпотези, peeking problem, двохвибірковий тест, достатня виборка.

Реферат. В роботі розглянуто поняття A/B тестів. Досліджено та продемонстровано усі етапи проведення такого експерименту та практичні застосування, які набули великої популярності в сучасному IT-бізнесі. Продемонстрована практична програмна реалізація A/B експерименту за допомогою мови программування R. Проаналізовано переваги, недоліки та проблеми цього методу перевірки статистичніх гіпотез.

ВСТУП

Сучасна статистика знайшла багато застосувань у комерційних сферах. Сталими галузями, що базуються на використанні статистики та перевірки гіпотез є, наприклад, медицина та страхова індустрія.

Так, страхове діло застовує актуарну математику: напрям у математиці, який вивчає питання, пов'язані з оцінкою ризиків у різних сферах людської діяльності. Актуарна математика включає: принципи побудови та аналізу актуарних моделей; основні числові характеристики фінансових операцій, що використовуються у страхуванні; методи аналітичної оцінки результатів діяльності та прогнозування розвитку страхової компанії.

Статистика в медицині має багато застосувань: клінічні експерименти, моделі виживаності, тощо. Аналіз виживаності — галузь статистики, що аналізує очікуваний час до моменту відбуття події, зокрема смерті біологічних організмів або поломки механічних систем. У різних науках аналіз виживаності має назву теорія надійності або аналіз надійності (інженерія), аналіз тривалості або моделювання тривалості (економіка) та аналіз історії подій (соціологія). Аналіз виживаності намагається відповісти на такі запитання як: який відсоток популяції виживе через певний відрізок часу? З тих, хто виживе, з якою швидкістю вони гинутимуть? Чи можна брати до уваги різні причини смерті? Як різні обставини та характеристики підвищують або знижують ймовірність виживання? Клінічні випробування у свою чергу базуються на послідовній постановці та перевірці статистичних гіпотез, у тому числі за допомогою A/B тестів.

Однак, статистика набула ще більшої популярності у сучасному IT-секторі. В епоху, коли збираються терабайти даних про дії користувачів в програмі, а кожне бізнес-рішення може принести мільйонні прибутки або втрати, статистика перетворюється на вкрай ефективний інструмент. У своїй роботі я досліджуватиму проблематику та умови постановки та перевірки статистичних гіпотез у IT, а також розгляну типовий приклад використання A/B тестів для цієї задачі.

РОЗДІЛ 1

A/B ТЕСТ ЯК МЕТОД ПЕРЕВІРКИ ГІПОТЕЗИ

1.1 Що таке A/B тест

A/B тестування (також відоме як split-run testing) — це метод перевірки гіпотези, частіше за все заради дослідження поведінки користувачів в IT-продукті. A/B тести складаються з рандомізованого експерименту з двома варіантами, A та B [2, 3, 4]. A/B-тестування – це спосіб порівняння двох версій однієї змінної, як правило, шляхом порівняння реакції суб'єкта на варіант A з варіантом B і визначення того, який із двох варіантів є більш ефективним [5]. Ефективність в даному випадку - це вплив на якусь ключову бізнес-метрику, наприклад: конверсія в платіж на сайті, час використання додатку, тощо.

A/B тестування — це скорочення від "простого рандомізованого контрольованого експерименту", в якому порівнюються дві вибірки (A і B) однієї векторної змінної [1]. Ці варіанти схожі, за винятком одної детали, яка може вплинути на поведінку користувача. A/B тести широко вважаються найпростішою формою контрольованого експерименту. Однак, якщо додавати до тесту більше варіантів, його складність зростає [6]. A/B-тести корисні для розуміння поведінки користувачів і задоволення від змін в IT-продукті, таких як нова функція або зміна в інтерфейсі. Великі сайти соціальних мереж, як-от LinkedIn, Facebook та Instagram, використовують A/B тестування, щоб зробити роботу продукту більш якісною, а своїх користувачів більш задоволеними; і як спосіб оптимізації своїх послуг [7].

Сьогодні A/B тести також використовуються для проведення складних експериментів в таких предметах досліждення, як мережеві ефекти, поведінка клієнтів офлайн-бізнесів, вплив онлайн-сервісів на дії користувачів в офлайні і впливу користувачів один на одного.[7] Багато професій використовують дані A/B-тестів. Це включає аналітиків, інженерів інфраструктури даних, маркетологів, дизайнерів, інженерів програмного забезпечення та підприємців.[8] Багато бізнесів покладаються на дані A/B тестів, оскільки вони дозволяють компаніям

зрозуміти зростання компанії, збільшити дохід і покращити задоволення клієнтів. [8]

Версія А може бути версією, яка використовується на даний момент (утворюючи контрольну групу), тоді як версія В у певному відношенні модифікована в порівнянні з А (таким чином формується група впливу). Наприклад, на веб-сайті електронної комерції воронка покупок, як правило, є хорошим кандидатом для A/B тестування, оскільки навіть незначне зниження рівня відвалу на будь-якому етапі воронки може означати значний приріст продажів. Іноді значні покращення можна побачити за допомогою таких елементів тестування, як варінти тексту, макетів інтерфейсу і кольорів,[9] але не завжди. У A/B тестах користувачі бачать лише одну з двох версій, оскільки мета полягає в тому, щоб виявити, яка з двох версій є кращою.[10]

Багатоваріантне або мультиноміальне тестування подібне до тестування A/B, але може перевіряти більше двох версій одночасно або використовувати більше елементів керування. Прості A/B тести непридатні для спостережних, квазіекспериментальних чи інших неекспериментальних ситуацій – звичайне явище з даними опитування, автономними даними та іншими, більш складними явищами.

Деякі стверджують, що A/B тестування є зміною філософії та бізнес-стратегії в певних нішах, хоча підхід дуже схожий на ті, які зазвичай використовується в різноманітних дослідницьких традиціях [11]. A/B тестування як філософія веб-розробки приводить цю галузь до підходу, що базується на фактах та доказах. Вважається, що переваги A/B тестування полягають у тому, що його можна виконувати безперервно будь-де, тим більше, що більшість програмного забезпечення для автоматизації маркетингу зараз зазвичай має можливість проводити A/B тестування на постійній основі.

1.2 Історія A/B тестів

Як і майже завжди, встановити дату появи нового методу важко. Перше рандомізоване подвійне сліпе дослідження для оцінки ефективності гомеопатичного препарату відбулося в 1835 році [12]. Експерименти з рекламними

кампаніями, які порівнюють із сучасним тестуванням A/B, почалися на початку двадцятого століття [13]. Піонер реклами Клод Хопкінс використовував рекламні купони, щоб перевірити ефективність своїх кампаній. Однак цей процес, який Хопкінс описав у своїй «Науковій рекламі», не включав такі поняття, як статистична значущість і нульова гіпотеза, які використовуються при перевірці статистичних гіпотез [13]. Сучасні статистичні методи оцінки значущості вибіркових даних були розроблені окремо в цей же період. Ця робота була виконана в 1908 році Вільямом Сілі Госсетом, коли він змінив Z-тест, щоб створити t-критерій Стьюдента [14].

Із розвитком Інтернету стали доступні нові способи вибірки популяцій. Інженери Google провели свій перший тест A/B у 2000 році, намагаючись визначити оптимальну кількість результатів для відображення на сторінці результатів пошукової системи. Перший тест був невдалим через збої, які виникли через повільний час завантаження. Пізніше A/B тестування стали більш складними та глибокими, але основа та основні принципи, як правило, залишаються тими самими, і в 2011 році, через 11 років після першого тесту, Google провів понад 7000 різних тестів A/B [5].

У 2012 році співробітник Microsoft, який працював над пошуковою системою Microsoft Bing, провів експеримент, щоб перевірити різні способи відображення рекламних заголовків. За кілька годин альтернативний формат збільшив дохід на 12%, не вплинувши на показники досвіду користувачів. Сьогодні такі компанії, як Microsoft і Google, проводять понад 10 000 A/B тестів на рік [4].

1.3 Основні виклики

Проводячи A/B тестування, дослідник повинен оцінити його плюси і мінуси, щоб побачити, чи цей підхід найкраще відповідає результатам, на які він сподівається.

За допомогою A/B тестування легко отримати чітке уявлення про те, чому віддають перевагу користувачі, оскільки воно безпосередньо перевіряє одне над іншим. Підхід заснований на реальній поведінці користувачів, тому дані можуть

бути дуже корисними, особливо коли визначають, що краще працює між двома варіантами. Крім того, експеримент також може дати відповіді на дуже конкретні питання дизайну. Одним із прикладів цього є A/B тестування Google із кольорами гіперпосилань. Щоб оптимізувати дохід, вони протестували десятки різних відтінків гіперпосилань, щоб побачити, на який колір користувачі частіше натискають.

Однак у A/B тестування є кілька недоліків. Як згадувалося вище, A/B тестування добре підходить для конкретних питань дизайну, але це також може бути недоліком, оскільки експеримент працює лише для конкретних проблем проектування з дуже вимірними результатами. Це також може бути дуже дорогим і часозатратним процесом. Якщо не було знайдено значного впливу, це може закінчитися марною тратою часу та ресурсів.

1.4 Часто використовані статистики у частотному підходу до тестів

Z-тест підходить для порівняння середніх за суворих умов щодо нормальності та відомого стандартного відхилення. Т-критерії Стьюдента підходять для порівняння середніх у послаблених умовах, коли передбачення менш сильні. Найпопулярнішим непараметричним критерієм, який часто використовується для аналізу далеких від нормальності даних, є критерій Мана-Уитні. Для порівняння двох біноміальних розподілів, таких як конверсія з одного етапу в інший, можна використати точний біноміальний тест.

1.5 Класичні приклади використання

1.5.1 Email маркетинг

Компанія з базою клієнтів у 2000 осіб вирішує створити електронну кампанію зі знижкою, щоб генерувати продажі через свій веб-сайт. Дослідник створює дві версії електронного листа з різним закликом до дії (частина тексту, яка спонукає клієнтів щось зробити — у цьому випадку , зробити покупку) та ідентифікаційним промокодом.

1000 людям він надсилає електронний лист із закликом до дії: "Пропозиція діє цієї суботи! Використовуйте код A1”, а ще 1000 людям він надсилає

електронний лист із закликом до дії: "Пропозиція скоро закінчується! Використовуйте код B1". Усі інші елементи копії та макета електронних листів ідентичні. Потім дослідник відстежує, яка кампанія має вищий показник успіху, аналізуючи використання промо-кодів. Електронна пошта з кодом A1 має 5% відповіді (50 з 1000 людей, надісланих електронною поштою, використовували код, щоб купити продукт), а електронний лист із кодом B1 має 3% відповіді. Тому компанія визначає, що в цьому випадку перший заклик до дії є більш ефективним і використовуватиме його в майбутніх продажах. Правильний підхід передбачатиме застосування статистичного тестування, щоб визначити, чи були різниці в рівнях відповідей між A1 та B1 статистично значущими (тобто велика ймовірність того, що відмінності є реальними, повторюваними, а не випадковими) [16].

У наведеному вище прикладі експерименту ціль — визначити, який варіант є ефективнішим способом заохотити клієнтів зробити покупку. Проте, якби метою тесту було побачити, яка електронна пошта генеруватиме вищий рейтинг кліків, тобто кількість людей, які фактично переходять на веб-сайт після отримання листа, то результати могли б бути іншими. Наприклад, навіть незважаючи на те, що більше клієнтів, які отримали код B1, перейшли на веб-сайт, оскільки у заклику до дії не вказано дату закінчення рекламної акції, багато з них можуть не відчувати потреби негайно робити покупку. Отже, якби метою тесту було просто побачити, яка електронна пошта принесе більше трафіку на веб-сайт, то електронний лист із кодом B1 міг би бути більш успішним. A/B тест повинен мати визначений результат, який можна виміряти, наприклад кількість здійснених продажів, конверсію кліків або кількість людей, які реєструються заходять на сайт[17].

1.5.2 Визначення ціни продукту

A/B тестування можна використовувати для визначення оптимальної ціни на продукт, оскільки це, мабуть, одне з найскладніших завдань під час запуску нового продукту чи послуги. A/B тестування (особливо стосується цифрових

товарів) — це чудовий спосіб дізнатися, які ціни та пропозиції максимізують загальний дохід.

РОЗДІЛ 2

МАТЕМАТИЧНІ ВЛАСТИВОСТІ A/B ТЕСТІВ ТА ДИЗАЙН ЕКСПЕРИМЕНТУ

2.1 Необхідні умови для проведення A/B тесту

Враховуючи, що A/B тест вимагає значної кількості ресурсів і може призвести до прийняття рішень щодо продукту зі значним впливом, дуже важливо дотримуватися ключових вимог до підготовки та проведення тесту. Основні вимоги є такими:

- Наявна чітка гіпотеза, яку можливо перевірити чи спростити за домогою експерименту.
- Визначена цільова метрика, на яку ми будемо впливати, та яку будемо досліджувати.
- В одному експерименті ми тестуємо тільки одну зміну в продукті чи веб-сайті.
- Розподіл користувачів на контрольну та тестову групи є цілком випадковим, репрезантативними, та неупередженими.
- Протягом усього експерименту забезпечена цілісність та незмінність експериментального впливу на тестову групу користувачів [18].

2.2 Визначення цільової метрики

Вибір цільової метрики є однією з найважливіших частин A/B тесту, оскільки цей показник буде використовуватися для вимірювання ефективності продукту або функції для експериментальних груп, а також для визначення наявності статистично значущих відмінностей між цими двома групами.

Вибір показника успіху залежить від основної гіпотези, яка перевіряється за допомогою цього тесту. Це якщо не найбільша, то одна з найважливіших частин A/B тесту, оскільки вона визначає, як буде розроблено експеримент, а також наскільки добре для бізнеса працюють запропоновані ідеї. Вибір поганих

показників може дискваліфікувати великий обсяг роботи або призвести до неправильних висновків [2].

Дохід не завжди є кінцевою метою, тому для тесту A/B нам потрібно прив’язати основний показник до прямих цілей зміни і цілей вищого рівня продукту. Очікується, що якщо продукт приносить більше грошей, це означає, що експеримент пройшов вдало. Але для досягнення цієї мети краще рухатися маленькими кроками, та покращувати показники усіх зон продукту, наприклад усі етапи воронки зміни статуса клієнта від відвідувача сайту до покупця. Для послідовної перевірки змін конкретних цільових метрик вдало підходить A/B тест. Дуже часто цільовою метрикою є якесь значення конверсії [3]. Наприклад процент користувачів, які купили продукт серед тих, які зайшли на сторінку продукту. Далі будемо називати конверсію CR (від англ. “Conversion rate”).

$$CR = \frac{\#\text{converted}}{\#\text{converted} + \#\text{not\_converted}}$$

2.3 Визначення гіпотези експерименту

A/B тест завжди має базуватися на гіпотезі, яку необхідно перевірити. Ця гіпотеза зазвичай встановлюється в результаті мозкового штурму та співпраці відповідних людей у команді дослідників на відповідному IT-продукті. Ідея цієї гіпотези полягає в тому, щоб вирішити, як «виправити» потенційну проблему в продукті, якщо вирішення цих проблем вплине на ключові показники ефективності (KPI), які цікавлять [18]?

Визначимося зі статистичним формулюванням гіпотези. Ми вважаємо, що зміна в продукті підвищіть конверсію. Конверсія має розподіл Бернуллі, адже кожне спостереження може мати значення 0 (користувач не виконав цільову дію), або 1 (користувач виконав цільову дію). Отже, нам треба збільшити середнє вибіркове з розподілу Бернуллі.

Нульова гіпотеза (H_0) — зміна в веб-сайті не вплине на конверсію в покупку серед користувачів. Середні вибіркові контрольної та експериментальної груп не відрізняються, різниця в спостереженнях випадкова. Альтернативна гіпотеза (H_1) — зміна в веб-сайті вплине на конверсію в покупку серед користувачів. Середні вибіркові контрольної та експериментальної груп дійсно відрізняються, різниця в спостереженнях не є випадковою.

$$\begin{cases} H_0 : \mu_{con} = \mu_{exp} \\ H_1 : \mu_{con} \neq \mu_{exp} \end{cases},$$

де μ_{con} — середнє значення в контрольній виборці, μ_{exp} — середнє значення в тестовій виборці.

2.4 Підготовка параметрів A/B тесту

Щоб переконатися, що наші результати відтворювані, надійні та можуть бути узагальнені для всієї генеральної сукупності, нам потрібно уникати “проблеми підглядання”, тобто занадто частої перевиріки поточних результатів тесту та p-value, отже, щоб забезпечити реальну статистичну значущість та уникнути необ’єктивних результатів, ми хочемо переконатися, що ми збираємо достатню кількість спостереження, і ми запускаємо тест протягом мінімального заздалегідь визначеного часу. Тому перед запуском тесту нам необхідно визначити розмір вибірки контрольної та експериментальної груп і скільки часу нам потрібно провести тест. Цей процес часто називають аналізом потужності, і він включає 3 конкретні кроки: визначення потужності тесту, визначення рівня значущості тесту та визначення мінімального ефекту, який ми відстежуємо (далі — MDE, від англ. “Minimum detectable effect”). Популярним посиланням на параметри, які беруть участь у аналізі потужності для A/B тестування, є таке позначення:

α : Ймовірність помилки першого роду, рівень значущості

β : Ймовірність помилки другого роду

$(1-\beta)$: Потужність тесту

δ : MDE

2.4.1 Значущість тесту

Рівень значущості, який також є ймовірністю помилки першого роду, — це ймовірність відхилення нульового значення, отже, виявлення ефекту зміни, у випадку, коли нуль є істинним і немає статистично значущого впливу. Це значення є ймовірністю помилкового відкриття, яке часто називають хибнопозитивним результатом.

Як правило, ми використовуємо значення значущості 5%, що вказує на те, що ми маємо 5% ризик зробити висновок про наявність статистично значущої різниці між показниками експериментального та контрольного варіантів, якщо фактичної різниці немає. Отже, миримося з тим, що в 5 із 100 випадків виявляємо ефект від зміни, поки його немає. Це також означає, що у вас є значна різниця в результатах між контрольною та експериментальною групами з впевненістю 95%. Рівень значущості необов’язково має дорівнювати 5%, але це є стандартом в сучасних експерементах в IT-сфері [16].

Як і у випадку з потужністю тесту, вибір значення альфа залежить від природи тесту та бізнес-обмежень. Наприклад, якщо проведення цього A/B тесту пов’язане з високими витратами на інженерію, то компанія може вирішити вибрати високий рівень альфа, щоб точніше виявити ефект зміни. З іншого боку, якщо витрати на впровадження запропонованої версії у виробництво високі, ви можете вибрати низький рівень значущості, оскільки ця запропонована функція дійсно повинна мати великий вплив, щоб виправдати високі витрати на впровадження, тому буде важче відхилити нульовий рівень [18] .

2.4.2 Потужність тесту

Потужність статистичного тесту - це ймовірність правильного відхилення нульової гіпотези. Потужність — це ймовірність прийняти правильне рішення (відкинути нульову гіпотезу), коли нульова гіпотеза хибна.

Потужність, яка часто визначається як (1-бета), дорівнює ймовірності не зробити помилку другого роду, де помилка другого роду — це ймовірність не відхилити нульову гіпотезу, коли нуль є помилковим.

Поширеною практикою є вибирати 80% як потужність тесту A/B, тобто 20% помилки типу II, що означає, що ми миримося з тим, що не виявляємо (не відхиляємо нульовий) ефект зміни, коли насправді є ефект. Однак вибір значення цього параметра залежить від характеру тесту та бізнес-обмежень [16].

2.4.3 Мінімальний відстежуваний ефект (MDE)

З точки зору бізнесу, яке мінімальне покращення від внесеної зміни, за умови, що воно є статистично значущим, ми будемо вважати як достатнє, щоб окупити трати на розробку та запровадження зміни?

Відповідь на це питання полягає в тому, яку зміну ми прагнемо спостерігати в метриці нової версії порівняно з існуючою, щоб дати рекомендації бізнесу щодо запуску цієї функції у виробництво. MDE — це доволі суб’єктивний параметр, і зазвичай встановлюється зацікавленими сторонами: дослідником та менеджером бізнесу.

2.5 Визначення об’єму виборки

Іншою дуже важливою частиною A/B тестування є визначення мінімального розміру вибірки контрольної та експериментальної груп, який необхідно визначити за допомогою визначеної потужності тесту, рівня значущості, мінімального відстежуваного ефекту (MDE) та дисперсії двох однакових за розміром вибірок із нормально розподіленим вмістом. Розрахунок розміру вибірки залежить від основного показника, який ви вибрали для відстеження ходу контрольної та експериментальної версій. Тут розрізняють два випадки; випадок, де основний показник A/B тестування у формі двійкової змінної (наприклад,

конверсія), і випадок 2, де основний показник тесту є неперервною змінною: у вигляді пропорцій або середніх (наприклад, середня сума замовлення) [19].

2.5.1 Визначення розміру виборки у випадку двійкової змінної

Коли ми маємо справу з основним показником відстеження ефективності, який має два можливі значення, такі як значення конверсії, коли користувач може здійснити цільову дію (успіх), або не здійснити (невдача), і якщо відповіді користувачів на продукт можна визначити як незалежні події, то ми можемо розглядати це як розподіл Бернуллі, де подія успіх відбувається з імовірністю p_{con} у випадку контрольної групи та p_{exp} у випадку експериментальної групи [18]. Крім того, невдача відбувається з ймовірністю q_{con} у випадку контрольної групи та q_{exp} у випадку експериментальної групи, де:

$$q_{con} = 1 - p_{con}$$
$$q_{exp} = 1 - p_{exp}$$

Отже, випадкова величина, що описує кількість успіхів, отриманих від користувачів під час тесту, відповідає біноміальним розподілам, де розмір вибірки - це кількість разів, коли функція/продукт був показаний користувачам, а ймовірність успіху дорівнює p_{con} і p_{exp}. , для контрольної та експериментальної груп відповідно. Розмір вибірки, необхідний для порівняння цих двох біноміальних пропорцій, за допомогою двостороннього тесту з попередньо визначеним рівнем значущості, рівнем потужності, MDE, можна розрахувати таким чином:

$$N = \frac{(\sqrt{2\bar{p}\bar{q}} \ * \ z_{1-\alpha/2} \ + \ \sqrt{p_{con}q_{con} + p_{exp}q_{exp}} \ * \ z_{1-\beta})^2}{\delta^2}$$

де для визначення $\bar{p}$ і $\bar{q}$ ми можемо використати історичні дані про продукт, або (що є найкращим варіантом) провести A/A тест: це експеримент, у якому обидві групи отримають одну и те ж саму версію IT-продукту [19].

2.5.2 Визначення розміру виборки у випадку неперервної змінної

Коли ми маємо справу з основним показником відстеження ефективності, який представлений у формі середньої величини, як-от середня сума замовлення, де ми маємо намір порівняти контрольну та експериментальну групи, тоді ми можемо скористатися центральною граничною теоремою і стверджувати, що розподіл вибіркових середніх як контрольної, так і експериментальної груп відповідає нормальному розподілу. Отже, вибіркові розподіли різниці середніх цих двох груп також відповідають нормальному розподілу. Це:

$$\overline{X}_{con} \sim N(\mu_{con}, \sigma^2_{con})$$
$$\overline{X}_{exp} \sim N(\mu_{exp}, \sigma^2_{exp})$$
$$\overline{X}_{con} - \overline{X}_{exp} \sim N(\mu_{con} - \mu_{exp}, \frac{\sigma^2_{con}}{N_{con}} + \frac{\sigma^2_{exp}}{N_{exp}})$$

Отже, розмір вибірки, необхідний для порівняння середніх показників двох нормально розподілених вибірок за допомогою двостороннього тесту з попередньо визначеним рівнем значущості, рівнем потужності, MDE, можна розрахувати наступним чином:

$$N = \frac{(\sigma^2_{con} + \sigma^2_{exp})(z_{1-\alpha/2} + z_{1-\beta})^2}{\delta^2}$$

де для визначення σ^2_{con} і σ^2_{exp} (дисперсії виборок) ми можемо використати історичні дані про продукт, або провести A/A тест [19].

2.6 Визначення тривалості експерименту

Як згадувалося раніше, на це запитання потрібно відповісти перед запуском експерименту, а не під час, намагаючись зупинити тест, коли ви виявите статистичну значущість. Щоб визначити базову лінію тривалості, у випадку IT-бізнесів, коли ми тестуємо продукт, яким регулярно користуються люди, звичайним підходом є використання такої формули:

$$\text{Тривалість} = \frac{N}{\text{кількість користувачів за день}}$$

Наприклад, якщо ця формула призводить до 14, це означає, що тест потрібно проводити протягом двох тижнів. Однак дуже важливо враховувати багато специфічних для бізнесу аспектів, обираючи час і тривалість тесту, а також використовувати цю формулу з певною обережністю [18].

Наприклад, якщо хтось хотів провести експеримент на початку січня 2020 року, коли пандемія COVID-19 сколихнула світ і це вплинуло на використання інтернет-сервісів, для деяких компаній це означало високе збільшення використання сторінок, а для деяких - значне зниження самої можливості використання сервісу та його зручності. Використання A/B тесту без урахування цього призведе до неточних результатів, оскільки період активності не буде справжнім відображенням поширеного використання сторінки [8].

2.6.1 Занадто мала тривалість: ефект новизни

Користувачі, як правило, швидко та позитивно реагують на будь-які типи змін незалежно від їх природи. Цей позитивний ефект для експериментальної версії, який є повним, оскільки є зміна, незалежно від того, що це за зміна, називають ефектом новизни, і він з часом зникає і, таким чином, вважається «ілюзорним». Отже, було б неправильно приписувати цей ефект до самої експериментальної версії та очікувати, що він продовжуватиме зберігатися після того, як ефект новизни зникне. Отже, вибираючи тривалість тесту, ми повинні переконатися, що ми не запускаємо тест протягом занадто короткого періоду часу, інакше ми можемо мати ефект новизни. Ефекти новизни можуть бути серйозною

загрозою для зовнішньої валідності A/B тесту, тому важливо уникати його якомога сильніше.

Більш формально, проблема полягає в тому, як відокремити ефект нової функції від ефекту новизни, який не пов’язаний з новою функцією і завжди трапляється, коли користувач бачить щось нове. А ефект новизни — це конкретний (і найпоширеніший) приклад набагато ширшої теми: як переконатися, що я тестую лише одну конкретну річ, а не кілька речей одночасно. Скажімо, ви проводите тест, який дає деяким користувачам нижчу ціну. Як відокремити ефект нижчої ціни від збудження від отримання знижки?

Зауважте, що трапляється й навпаки. Тобто, якщо компанія надає користувачам новий досвід, спочатку вони можуть ненавидіти це, тому що це не те, до чого вони звикли, і вони відчувають, що повинні заново вчитися користуватися продуктом. Це називається неприйняттям змін [18]. Однак на практиці це набагато менша проблема з точки зору A/B тестування, оскільки вона впливає лише на основні зміни дизайну продукту, які зустрічаються рідше, ніж невеликі налаштування інтерфейсу користувача, і вони часто навіть не тестуються за допомогою A/B тесту.

Очевидним рішенням для ефекту новизни було б проводити тести довше, даючи користувачам тестів достатньо часу, щоб позбутися ефекту новизни. Однак це навряд чи є ефективним, і вартість довшого проведення тестів, ймовірно, переважить плюси, отримані від більш надійних результатів. Є декілька способів до деякої міри аналітично відокремити ефект новизни, але я не буду торкатися цієї теми у роботі

2.6.2 Занадто велика тривалість: ефект дозрівання

Плануючи A/B тест, зазвичай корисно враховувати більш тривалу тривалість тесту, щоб користувачі могли звикнути до нової функції або продукту. Таким чином, можна буде спостерігати реальний ефект зміни, надаючи більше часу користувачам, що повертаються, щоб охолонути від початкової позитивної реакції або сплеску інтересу через зміну, яка була внесена в рамках зміни. Це повинно допомогти уникнути ефекту новизни і, таким чином, покращити прогнозну

цінність для результату тесту. Однак, чим довший період тестування, тим більша ймовірність впливу зовнішніх ефектів на реакцію користувачів і, можливо, забруднення результатів тесту, ефекту дозрівання. Тому занадто довго проводити тест A/B також не рекомендується [2].

2.7 Запуск та проведення експерименту

Після того, як підготовчі роботи будуть виконані, інженерна команда продукту може починати A/B тест. По-перше, команда розробників повинна переконатися, що зберігається цілісність між контрольною та експериментальною групами. У A/B тестах наявна проблема незбалансованості класів між двома тестовими групами, але це трапляється тільки у експериментах з малими об'ємами виборки, а справді ефективного рішення цієї проблеми ще не знайшли. Отже, равномірний рандомний розподіл усіх користувачів продукту між контрольную та експерементальною групою підійде майже завжди.

По-друге, механізм зберігання даних про поведінку та зміну поведінки користувачів через зміну в продукті має бути точним і однаковим для всіх користувачів, щоб уникнути систематичної упередженості. Також неможна зупиняти тест занадто рано, як тільки дослідник знайде статистичну значущість. Це називається проблемою підглядання (peeking problem), я розгляну її трохи детальніше [14].

2.7.1 Проблема підглядання

Проведемо синтетичний A/B тест між двома однаковими монетами. Кожна має ймовірність випадіння орла, що дорівнює 0.5. Визначимо $\alpha = 0.05, \beta = 0.8$. Цільовою метрикою оберемо ймовірність випадіння орла. Сгенеруємо достатню кількість підбрасувань монет та візуалізуємо моменти, коли p-value буде дорівнювати 0.05 або менше.

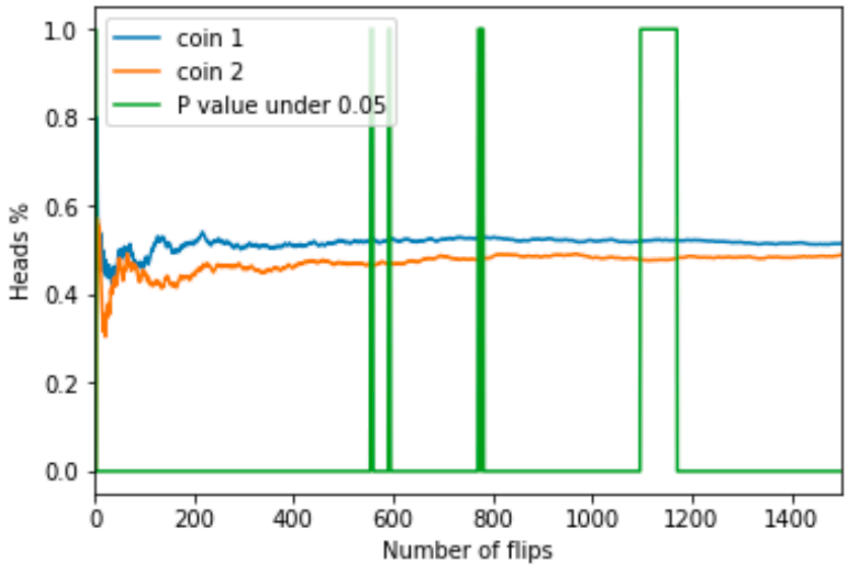

Рисунок 2.1

По вертикальній осі процент випадіння орлів для двох монет. Зелена лінія стрибає до 1, коли p-value сягає 0.05 або менше. Як видно, навіть у випадку двох однакових монет, ми деколи зможемо знайти статистично значущу різницю між ними. Це очевидно трапляється через банальну дисперсію даних, нижче ще декілька симуляцій

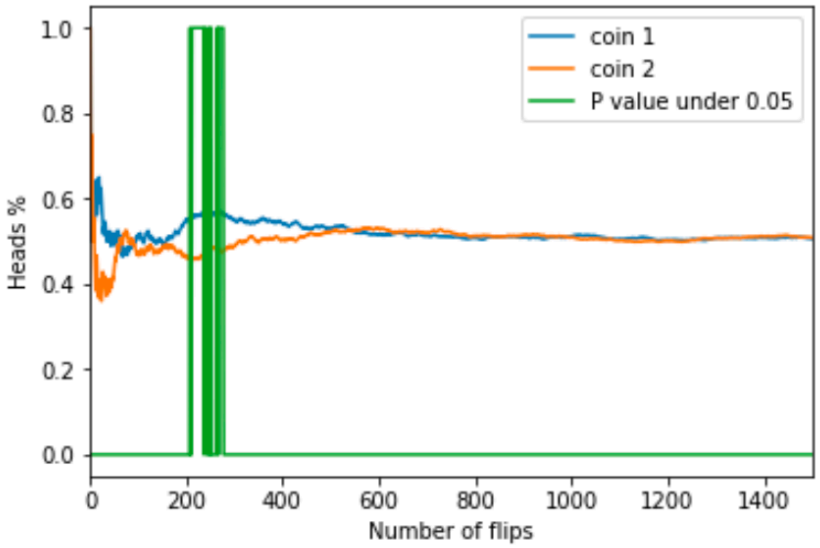

Рисунок 2.2

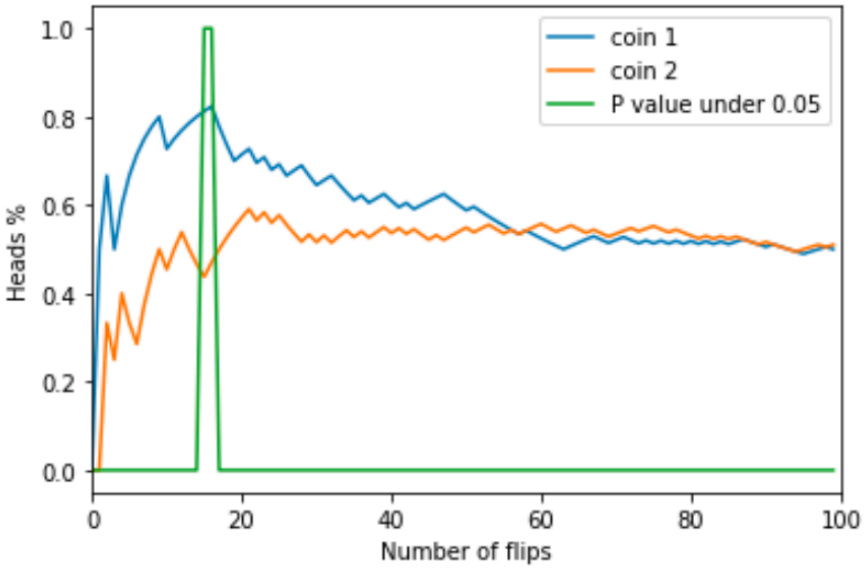

Рисунок 2.3

Тепер проведемо такий самий експеримент, але друга монета буде мати ймовірність випадіння орла 0.6, замість 0.5. Так само візуалізуємо результати експерименту.

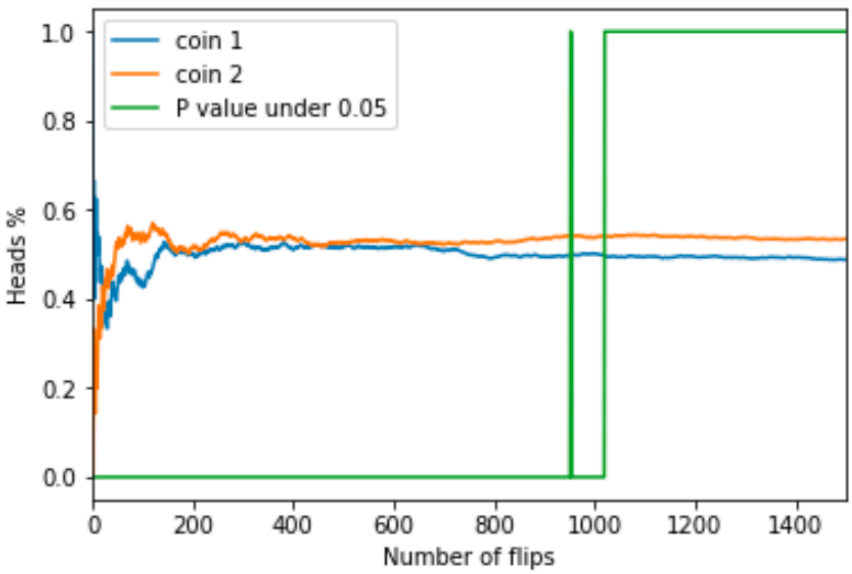

Рисунок 2.4

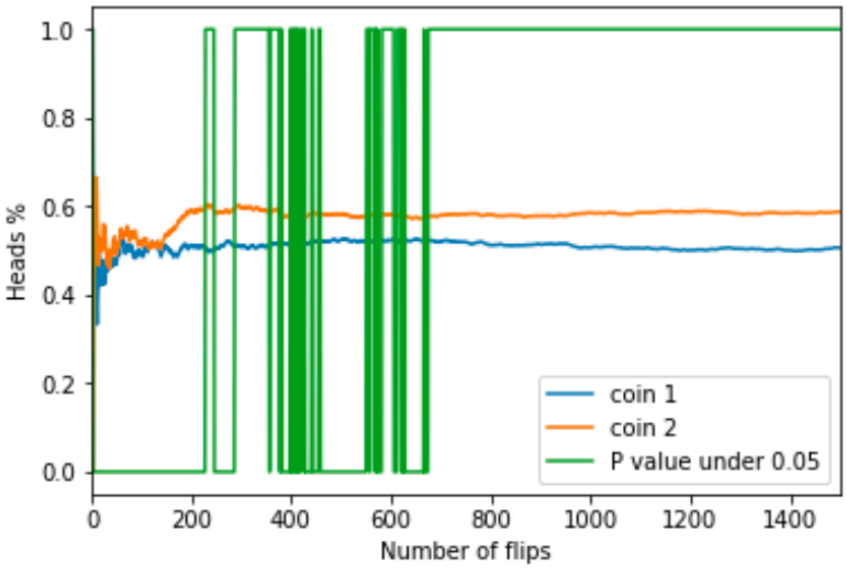

Рисунок 2.5

Іноді можна побачити справжню різницю рано, а іноді доводиться чекати довше. Через деякий момент функція кроку p-значення перевертається і залишається перевернутою. Іншими словами, через деякий час ми впевнені, що різниця справді є, і якщо ми перевіримо значущість, наші шанси дійсно високі, що ми побачимо її щоразу, коли будемо визначати результати експерименту [20].

Тепер проведемо більш наближений до реальності експеримент: спочатку визначимо необхідний об'єм виборки при $MDE = 4\,\%$, $\alpha = 0.05$, $\beta = 0.8$. За формулою, наданою вище, отримаємо 3856. Далі я продемонструю 4 випадки:

1. Коли друга монета буде мати настільки більшую ймовірність, наскільки ми очікували.
2. Коли не буде ніякої різниці.
3. Коли різниця буде більше, ніж очікувана.
4. Коли різниця буде менше, ніж очікувана.

Для кожного випадку на графіку є синя вертикальна лінія, яка повідомляє про те, що необхідний об'єм виборки був досягнутий.

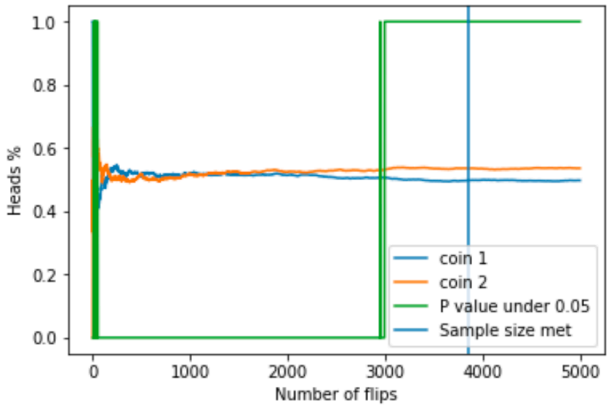

Рисунок 2.6 - Очікувана різниця

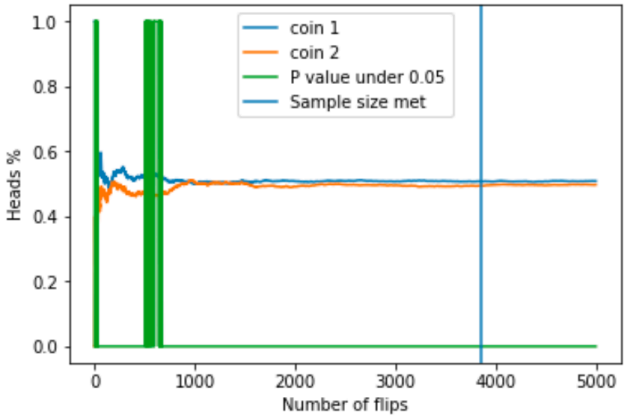

Рисунок 2.7 - Немає різниці

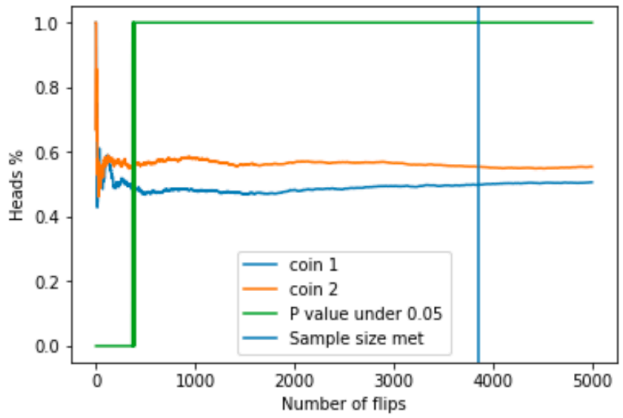

Рисунок 2.7 - Більша різниця

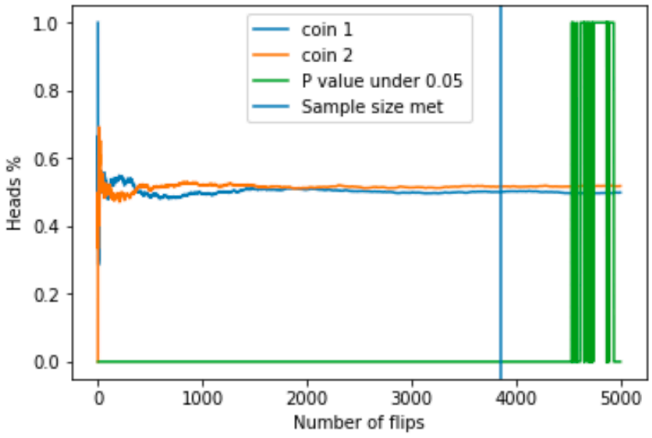

Рисунок 2.8 - Менша різниця

Завдяки цим прикладам видно, наскільки важливо знати очікуваний розмір ефекту, тобто правильно оцінювати MDE та чекати повного накопичування об’єму виборки для здійснення аналізу експерименту [20]. Також, треба відмитити, що великі компанії, які оперують терабайтами даних щодня, використовують наступну практику: тест зупиняється тільки тоді, коли p-value втрачає волатильність та залишається на одному рівні.

2.8 Аналіз результатів експерименту

Коли справа доходить до інтерпретації результатів A/B тесту, існує набір речей, які треба обчислити, щоб перевірити статистичну гіпотезу, зазначену

раніше (щоб перевірити, чи є статистично значуща різниця між контрольною та експериментальною групами). Цей набір включає:

- Вибір відповідного статистичного тесту
- Розрахунок статистики тесту
- Обчислення p-value тестової статистики
- Відхилення або прийняття статистичну гіпотезу (статистична значущість)
- Розрахувати похибку (зовнішню дійсність експерименту)
- Розрахувати довірчий інтервал (зовнішню валідність і практичну значущість експерименту)

2.8.1 Вибір критерію

Після того, як дані взаємодії контрольної та експериментальної груп будуть зібрані, ви можете перевірити статистичну гіпотезу, вибравши відповідний статистичний критерій, який зазвичай поділяється на параметричні та непараметричні критерії. Вибір критерію залежить від наступних факторів:

- Формат основної метрики
- Розмір вибірки
- Природа статистичної гіпотези (показати, що відносини між двома групами просто існують, або визначити тип відносин між групами)

Найпопулярніші параметричні тести, які використовуються в A/B тестуванні:

- Двохвибірковий T-тест (коли метрика розподіляється за розподілом Стьюдента, і дослідник хоче визначити, чи існує зв’язок і тип зв’язку між контрольною та експериментальною групами)
- Двохвибірковий Z-тесту (коли вибірка має більше 30 записів, метрика слідує асимптотичному нормальному розподілу, і дослідник хоче визначити, чи існує зв'язок і тип зв'язку між контрольною та експериментальною групами)

Найпопулярніші непараметричні тести, які використовуються в A/B тестах:

- Критерій хі-квадрат (велика вибірка, визначте, і ви хочете визначити, чи існує зв'язок між контрольною та експериментальною групами)
- Критерій Мана-Уитні (скошені розподіли вибірки, перевірка різниці в медіан між контрольною та експериментальною групами. Може також використовуватися для визначення різниці між середніми.)

2.8.1.1 Двохвибірковий Т-тест

Якщо ви хочете перевірити, чи наявна статистично значуща різниця між показниками контрольної та експериментальної груп у формі середніх (наприклад, середня сума покупки), метрика слідує за розподілом Стьюдента, Ви можете використовувати двохвибірковий Т-тест для перевірки наступної гіпотези [17]:

$$\begin{cases} H_0 : \mu_{con} = \mu_{exp} \\ H_1 : \mu_{con} \neq \mu_{exp} \end{cases}$$

$$\begin{cases} H_0 : \mu_{con} - \mu_{exp} = 0 \\ H_1 : \mu_{con} - \mu_{exp} \neq 0 \end{cases}$$

де вибірковий розподіл середніх контрольної групи слідує за розподілом Стьюдента зі ступенями свободи $N_{con} - 1$. Крім того, вибірковий розподіл середніх експериментальної групи також відповідає розподілу Стьюдента-t зі ступенями свободи $N_{exp} - 1$. Зауважте, що N_{con} і N_{exp}– це розмір виборки у контрольній та експериментальній групах відповідно.

$$\hat{\mu}_{con} \sim t(N_{con} - 1)$$
$$\hat{\mu}_{exp} \sim t(N_{exp} - 1)$$

Тоді, оцінка сукупної дисперсії двох вибірок може бути розрахована за такою формулою:

$$\hat{S}^2_{\text{сукупна}} = \frac{(N_{con}-1)*\sigma^2_{con} + (N_{exp}-1)*\sigma^2_{exp}}{N_{con}+N_{exp}-2} * \left(\frac{1}{N_{con}}+\frac{1}{N_{exp}}\right)$$

де σ^2_{con} та σ^2_{exp} – вибіркові дисперсії контрольної та експериментальної груп відповідно. Тоді стандартна помилка дорівнює квадратному кореню з оцінки об’єднаної дисперсії і може бути визначена як:

$$SE = \sqrt{\hat{S}^2_{\text{сукупна}}}$$

Отже, тестову статистику двохвибіркового Т-тесту з гіпотезою, викладеною раніше, можна розрахувати таким чином:

$$T = \frac{\hat{\mu}_{con} - \hat{\mu}_{exp}}{\sqrt{\hat{S}^2_{\text{сукупна}}}}$$

Щоб перевірити статистичну значущість спостережуваної різниці між вибірковими середніми, нам потрібно розрахувати p-value нашої тестової статистики. P-value — це ймовірність спостереження значень, принаймні таких екстремальних, як отримане значення, коли це відбувається через випадковість. Якщо говорити інакше, p-value — це ймовірність отримання ефекту, принаймні такого екстремального, як той, що є у ваших вибіркових даних, за умови, що нульова гіпотеза вірна. Тоді p-value тестової статистики можна розрахувати таким чином:

$$P_{value} = Pr[t \le -T \ \ or \ \ t \ge T] = 2 * Pr[t \ge T]$$

Інтерпретація p-value залежить від вибраного рівня значущості альфа, який був обраний перед запуском тесту під час аналізу потужності. Якщо розраховане p-value виявляється меншим, ніж дорівнює альфа (наприклад, 0,05 для 5% рівня значущості), ми можемо відхилити нульову гіпотезу і стверджувати, що існує статистично значуща різниця між первинними показниками контрольної та експериментальної груп.

Нарешті, щоб визначити, наскільки точними є отримані результати, а також прокоментувати практичну значущість отриманих результатів, ви можете обчислити довірчий інтервал тесту, використовуючи таку формулу:

$$CI = [(\hat{\mu}_{con} - \hat{\mu}_{exp)}) \; - \; t_{1-\alpha/2} * SE, \;\; (\hat{\mu}_{con} - \hat{\mu}_{exp)}) \; - \; t_{1+\alpha/2} * SE]$$

де $t_{1 \pm \alpha/2}$ є критичними значеннями тесту, що відповідає двосторонньому t-критерію з рівнем альфа-значущості, і його можна знайти за допомогою t-таблиці.

2.8.1.2 Двохвибірковий Z-тест

Якщо треба перевірити, чи є статистично значуща різниця між показниками контрольної та експериментальної груп у формі середніх (наприклад, середня сума покупки) або пропорцій (наприклад, рейтинг кліків), а також показник розподіляється за звичайним розподілом, то можна використати центральну граничну теорему, щоб стверджувати, що розподіли вибірки контрольної та експериментальної груп є асимптотично нормальними. У такому випадку можна використати двохвибірковий Z-тест. Тут ми будемо розрізняти два випадки: де основний показник у формі пропорцій (наприклад, рейтинг кліків) і де основний показник у формі середніх (наприклад, середня сума покупки).

Випадок 1: Z-тест для порівняння пропорцій (двосторонній)

Якщо треба перевірити, чи є статистично значуща різниця між показниками контрольної та експериментальної груп у формі пропорцій (наприклад, CTR), і

якщо подія кліку відбувається незалежно, мож на використовувати Z-тест із двома вибірками, щоб перевірити наступну гіпотезу:

$$\begin{cases} H_0 : p_{con} = p_{exp} \\ H_1 : p_{con} \neq p_{exp} \end{cases}$$

$$\begin{cases} H_0 : p_{con} - p_{exp} = 0 \\ H_1 : p_{con} - p_{exp} \neq 0 \end{cases}$$

де кожна подія кліку може бути описана випадковою змінною, яка може приймати два можливих значення 1 (успіх) і 0 (невдача), що слідує за розподілом Бернуллі з p_{con} і p_{exp} є ймовірністю натискання (ймовірність успіху) контрольної та експериментальної груп відповідно. Це:

$$X_{con} \sim Bern(p_{con})$$
$$X_{exp} \sim Bern(p_{exp})$$

Оскільки ми перевіряємо різницю в цих ймовірностях, нам потрібно отримати оцінку для об'єднаної ймовірності успіху та оцінку для сукупної дисперсії, яку можна зробити наступним чином:

$$\hat{p}_{\text{сукупна}} = \frac{\#\text{кліки}_{con} + \#\text{кліки}_{exp}}{\#\text{покази}_{con} + \#\text{покази}_{exp}}$$

$$\hat{S}^2_{\text{сукупна}} = \hat{p}_{\text{сукупна}}(1 - \hat{p}_{\text{сукупна}}) * \left(\frac{1}{N_{con}} + \frac{1}{N_{exp}}\right)$$

Отже, тестову статистику 2-вибіркового Z-тесту для різниці пропорцій можна розрахувати таким чином:

$$T = \frac{\hat{p}_{con} - \hat{p}_{exp}}{\sqrt{\hat{S}^2_{\text{сукупна}}}}$$

Тоді p-value цієї тестової статистики можна розрахувати таким чином:

$$P_{value} = Pr[Z \leq -T \ or \ Z \geq T] = 2 * Pr[Z \geq T]$$

Нарешті, можна обчислити довірчий інтервал тесту таким чином:

$$CI = [(\hat{p}_{con} - \hat{p}_{exp)}) - z_{1-\alpha/2} * SE, \ (\hat{p}_{con} - \hat{p}_{exp)}) - z_{1+\alpha/2} * SE]$$

де $z_{1 \pm \alpha/2}$ є критичними значеннями тесту, що відповідає двосторонньому Z-тесту з рівнем альфа-значущості, і його можна знайти за допомогою Z-таблиці. Область відхилення цього двостороннього Z-тесту з 2 вибірками можна візуалізувати за допомогою наступного графіка.

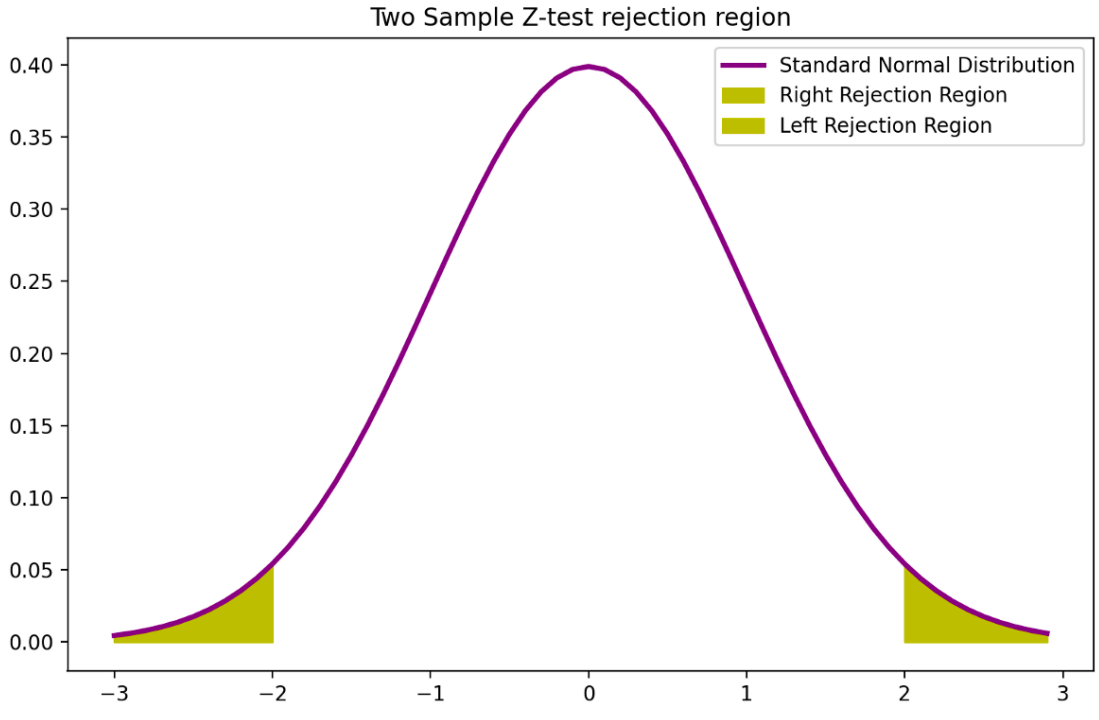

Рисунок 2.9

Випадок 2: Z-тест для порівняння середніх (двосторонній)

Якщо потрібно перевірити, чи є статистично значуща різниця між показниками контрольної та експериментальної груп у формі середніх (наприклад, середній дохід з користувача), ви можете використовувати двохвибірковий Z-тест, щоб перевірити таку гіпотезу:

$$\begin{cases} H_0 : \mu_{con} = \mu_{exp} \\ H_1 : \mu_{con} \neq \mu_{exp} \end{cases}$$

$$\begin{cases} H_0 : \mu_{con} - \mu_{exp} = 0 \\ H_1 : \mu_{con} - \mu_{exp} \neq 0 \end{cases}$$

де вибірковий розподіл середніх контрольної групи відповідає нормальному розподілу із середнім μ_{con} та $\frac{\sigma_{con}^2}{N_{con}}$. Крім того, вибірковий розподіл середніх експериментальних груп також відповідає нормальному розподілу із середнім μ_{exp} та $\frac{\sigma_{exp}^2}{N_{exp}}$. Тоді різниця в середніх значеннях контрольної та експериментальної груп також відповідає нормальним розподілам із середнім $\mu_{con} - \mu_{exp}$ та дисперсією $\frac{\sigma_{con}^2}{N_{con}} + \frac{\sigma_{exp}^2}{N_{exp}}$. Отже, тестову статистику двохвибіркового Z-тесту для різниці середніх можна розрахувати таким чином:

$$T = \frac{\hat{\mu}_{con} - \hat{\mu}_{exp}}{\sqrt{\frac{\sigma_{con}^2}{N_{con}} + \frac{\sigma_{exp}^2}{N_{exp}}}} \sim N(0,1)$$

Стандартна помилка дорівнює квадратному кореню з оцінки сукупної дисперсії і може бути визначена як:

$$SE = \sqrt{\frac{\sigma_{con}^2}{N_{con}} + \frac{\sigma_{exp}^2}{N_{exp}}}$$

Тоді p-value та довірчій інтервал цієї тестової статистики можна розрахувати таким чином:

$$P_{value} = Pr[Z \leq -T \ or \ Z \geq T] = 2 * Pr[Z \geq T]$$

$$CI = [(\hat{p}_{con} - \hat{p}_{exp)}) - z_{1-\alpha/2} * SE, \ (\hat{p}_{con} - \hat{p}_{exp)}) - z_{1+\alpha/2} * SE]$$

2.8.2 Статистична значущість та практична значущість

Під час фази статистичного аналізу A/B тестування, коли виявлено мале p-value, ми говоримо про статистичну значущість. Однак лише статистичної значущості недостатньо, щоб дати рекомендацію щодо запуску нової функції чи продукту [3]. Після виявлення статистичної значущості наступним кроком є розуміння того, чи є практична значущість. Це допоможе нам зрозуміти, чи виявлена різниця в продуктивності двох груп достатньо велика, щоб виправдати інвестиції, чи вона занадто мала, і прийняття рішення про запуск не варте вкладень.

Один із способів визначити, чи має A/B тест практичне значення, — це використовувати довірчий інтервал і порівняти його нижню межу з MDE (оцінкою економічної значущості). Точніше, якщо нижня межа CI більша за MDE (дельта), то ви можете заявити, що маєте практичне значення. Наприклад, якщо CI = [5%, 7,5%] і MDE = 3%, то можна зробити висновок, що він має практичне значення, оскільки 5% більше ніж 3% [7].

Зауважте, що треба також подивитися на ширину CI та переконатися, що вона не надто велика, оскільки занадто широка CI дає вам ознаки того, що точність ваших результатів мала і результати не будуть узагальнені для всієї сукупності.

2.9 Якість A/B тесту

A/B тестування є одним із прикладів експериментального дизайну, і, як і в будь-якому іншому типі експерименту, є також фактори, які необхідно задовольнити, щоб зробити тверді висновки та рішення щодо продукту. Такими факторами є [18]:

- Надійність та відтворюваність
- Валідність
- Потужність ефекту

2.9.1 Надійність і відтворюваність

Ідея надійності полягає в тому, що експериментальні результати мають бути більш ніж одноразовим висновком і бути за своєю природою відтворюваними та повторюваними. Останнім часом у дослідницькій індустрії з'явився феномен кризи відтворюваності, оскільки дослідники не можуть відтворити експериментальні результати. Це може статися з різних причин, таких як:

- Коли початковий експеримент був змінений або сталася проблема підглядання
- Коли в початковому експерименті була помилка вимірювання
- Коли в початковому експерименті була систематична помилка або систематичне упередження
- Відсутність документації або вихідного коду та/або даних, використаних для проведення експерименту

Що можна зробити, щоб підвищити відтворюваність експериментів:

- Зберігайте вихідний код з коментарями в захищеному хмаревому сховищі
- Зберігайте дані з коментарями в захищеному сховищі
- Скласти детальну документацію процесу та результатів
- Перевірте наявність систематичних помилок (спосіб повідомлення про дії користувачів в продукті)
- Зробіть такий же аналіз для іншого зрізу користувачів: наприклад для іншої країни, або платформи

2.9.2 Валідність

Валідність охоплює всю концепцію вашого експерименту та визначає, чи відповідають отримані результати всім вимогам рандомізованих контрольних досліджень чи ні. У випадку валідністі ми зазвичай розрізняємо два типи: внутрішня та зовнішня валідності.

Внутрішня валідність

Внутрішня валідність відноситься до спостережуваних даних і отриманих з ними результатів. Чи є ці результати дійсними та надійними, чи вони неточні й упереджені? Чи зміни залежної змінної зумовлені лише втручанням (незалежна змінна), а не іншими факторами? Нижче наведено приклади проблем, які можуть негативно вплинути на внутрішню валідність експерименту:

- Пропущене зміщення змінної
- Зворотна причинність
- Невідповідна змінна
- Використання невідповідних сурогатних змінних (треба використовувати фактичну змінну втручання)

Зовнішня валідність

Зовнішня валідність — це ступінь узагальнення ваших експериментальних результатів для всієї сукупності. Чи можна узагальнити результати на широку групу населення? Чи можна підвищити зовнішню валідність шляхом повторення експериментів у подібних, але трохи інших умовах? Нижче наведено приклади проблем, які можуть негативно вплинути на зовнішню дійсність вашого експерименту A/B. Класичними проблемами, які погано впливають на зовнішню валідність є наявніст зміщенної або нерепрезантативної виборки.

Особливо, якщо генеральна сукупність розділена на кілька субсукупностей, які якимось чином відрізняються, і дослідження вимагає, щоб кожна субсукупність була однаковою за розміром, стратифікована вибірка може бути дуже корисною. Таким чином рандомізуються одиниці кожної субсукупності, але не вся вибірка. Потім результати експерименту надійно узагальнюються від

експериментальних одиниць до більшої сукупності одиниць. Такоє можена використовувати boostrapping для обчислення стандартної помилки (похибки) ваших результатів і ширини довірчого інтервалу. А саме, якщо SE вашого A/B тесту великий або довірчий інтервал широкий, то можна зробити висновок, що точність результатів низька і результати не будуть узагальнюватися, якщо застосовувати їх до всієї сукупності.

2.9.3 Потужність ефекту

Важливо переконатися, що втручання має, або краще сказати, може мати, достатню потужність, щоб викликати вимірні зміни залежної змінної, інакше можна зробити неправильне припущення, що зміна в продукті не має ефекту (помилка другого роду). Крім того, це означає, що залежна змінна повинна бути чутливою до зміни. Чутливість можна покращити шляхом зменшення шуму (наприклад, похибки вимірювання), наприклад, шляхом повторних вимірювань та їх усереднення [10].

2.10 Поширені проблеми та підводні камені A/B тестів

Щоб не провалити онлайн-експеримент, важливо дотримуватися вказаних інструкцій і терпляче переглядати список дій, які повинні бути виконані, щоб завершитися добре підготовленим і проведеним експериментом A/B. Нижче будуть представлені поширені проблеми та підводні камені A/B тестування, яке часто проводиться з відповідними рішеннями [20].

2.10.1 Змішані ефекти

Важливо переконатися, що всі інші відомі можливі фактори, які також впливають на залежну змінну, залишаються незмінними. Тому вам потрібно контролювати якомога більше небажаних або нерівних факторів (також званих сторонніми змінними). Сторонні змінні мають значення, коли вони пов’язані як з незалежною, так і з залежною змінною [18]. Один особливий і крайній випадок цієї проблеми виникає, коли відношення між незалежною та залежною змінною повністю змінюється/інвертується, коли враховується певні помилкові змінні, на

це часто посилається парадокс Сімпсона. Незрозумілі ефекти загрожують внутрішній валідності вашого A/B експерименту. Наступні рішення можуть допомогти вам уникнути цієї проблеми:

- Контроль змішуючихся змінних
- Надійні інструменти розподіла між групами
- Відповідний вибір незалежних і залежних змінних
- Генерація випадкової вибірки

2.10.2 Упередження відбору

Одне з фундаментальних припущень A/B тестування полягає в тому, що ваша вибірка повинна бути неупередженою, і кожен тип користувачів повинен мати однакову ймовірність бути включеним до цієї вибірки. Якщо через якусь помилку ви виключили певну частину генеральної сукупності, наприклад, вибірка для середньої ваги серед жилетів США шляхом вибірки лише з одного штату, тоді це називається упередженням відбору (Selection Bias) [20].

Щоб перевірити, чи є ваша вибірка упередженою, знаючи справжній розподіл сукупності, ви створюєте вибірки бутстрепом із вашої вибірки та малюєте розподіл вибіркових середніх. Якщо цей розподіл не зосереджено навколо справжнього середнього сукупності, то вибірка є упередженою, і слід використовувати більш надійні методи вибірки, щоб випадково відбирати неупереджену вибірку.

2.10.3 Раннє припинення або підглядання

Поширеною помилкою в експерименті є дострокове припинення експерименту, як тільки ви спостерігаєте статистично значущий результат (наприклад, невелике значення p), тоді як рівень значущості та всі інші параметри моделі заздалегідь визначені на етапі аналізу потужності A/B тестування і зроблено ключове припущення про те, що експеримент триватиме до досягнення мінімального розміру вибірки. Підглядання або раннє зупинення впливає на внутрішню валідність результатів і робить їх упередженими, а також призводить до помилкових результатів [20].

2.10.4 Переливання або мережеві ефекти

Ця проблема зазвичай виникає, коли тест A/B виконується на платформах соціальних медіа, таких як Facebook, Instagram, TikTok, а також в інших продуктах, де користувачі експериментальної та контрольної груп, наприклад, знаходяться в одній групі чи спільноті і впливають на реакцію один одного на експериментальну та контрольовану версії продукту. Ця проблема призводить до упереджених результатів і неправильних висновків, оскільки порушує цілісність тесту та контрольних ефектів [2].

Щоб виявити мережеві ефекти, можна виконати стратифіковану вибірку, а потім розділити її на дві групи. Потім можна запустити A/B тест на одній вибірці, враховуючи кластеризовані вибірки, а на іншій — без. Якщо є різниця в ефектах зміни, то виникає проблема з мережевим ефектом [16].

2.10.5 Невідповідне співвідношення вибірок

Якщо є підозра, що поділ між контрольною та експериментальною групами виглядає підозрілим, що припускає, що процес призначення групи неправильно працює, оскільки більше користувачів призначено до контрольної/ експериментальної групи, ніж до експериментальної/контрольної, тоді можна виконати тест Хі-квадрат. Цей тест допоможе формально перевірити невідповідність співвідношення вибірок [18].

2.10.6 Неадекватний вибір періоду тестування

Іншою поширеною помилкою в A/B тестуванні є вибір періоду тестування. Як згадувалося раніше, одне з основних припущень A/B тестування полягає в тому, що кожен тип користувачів повинен мати однакову ймовірність, щоб бути включеним до цієї вибірки. Однак, якщо ви проводите тест у період, який не враховує свята, сезонність, вихідні та будь-які інші відповідні події, ймовірність вибору різних типів користувачів більше не буде однаковою (наприклад, покупці у вихідні дні, святкові покупці тощо). Наприклад, виконання тесту в неділю вранці відрізняється від виконання того ж тесту у вівторок об 23:00.

2.10.7 Проведення занадто великої кількості експеримент одночасно

Якщо експеримент передбачає більше, ніж одну експериментальну версію, тобто, можна назвати це не A/B тестом, а A/B/C/D тестом, то більше не можна використовувати той самий рівень значущості для перевірки статистичної значущості між одною контрольною та одною експериментальною групами. Отже, p-value або рівень значущості, з яким будуть порівнюватися результати, потрібно відкоригувати.

У цьому випадку можна використовувати поправку Бонферонні, щоб налаштувати цей рівень значущості на основі кількості вибірок n. Отже, рівень значущості, який необхідно використовувати в багатоваріантному тестуванні, має бути $\frac{\alpha}{N}$, де *N* це кількість експериментальніх варіантів. Наприклад, якщо рівень значущості становить 5%, а експериментальних груп в нашому експерименті 3, то новий скоригований рівень значущості має становити $\frac{0.05}{3}$.

РОЗДІЛ 3

ПРАКТИЧНА РЕАЛІЗАЦІЯ A/B ТЕСТУ

3.1 Контекст експерименту

Експеримент будем проводити на прикладі мобільного додатку для сканування документів xScan. Додаток заробляє за рахунок підписки, до якої закликають на самому старті користування на наступному екрані.

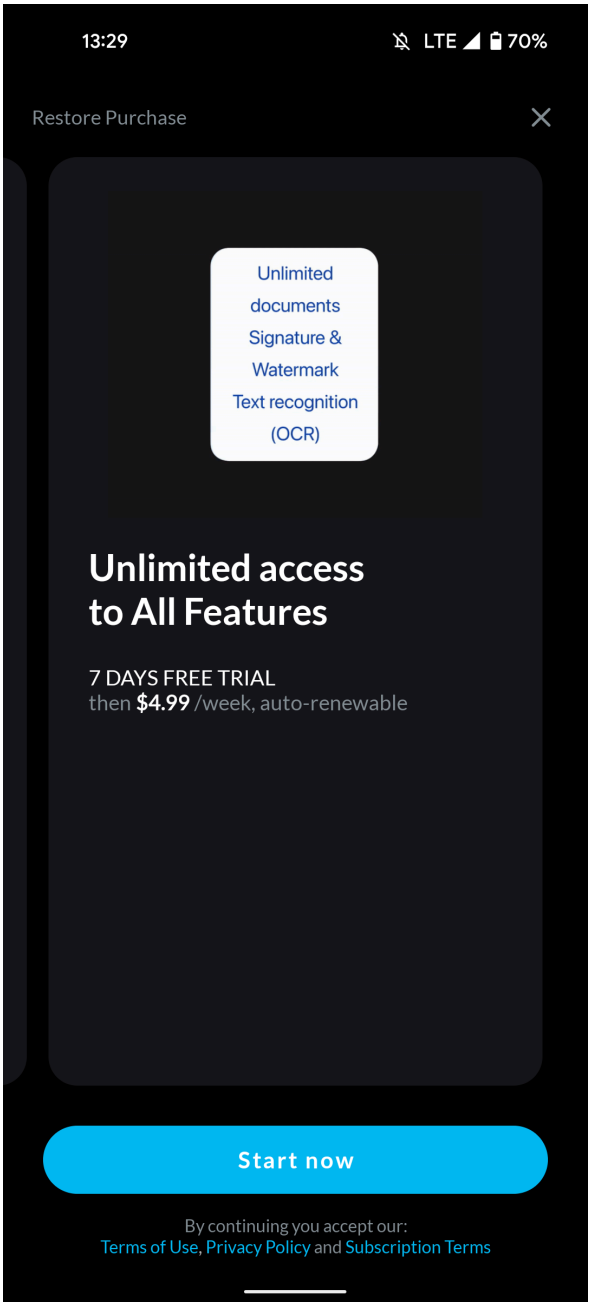

Рисунок 3.1

3.2 Цільова метрика

Нас цікавить конверсія у оформлення підписки на цьому екрані, тобто процент користувачів, які натиснули кнопку “Start now” та успішно підписались на продукт. Це одна з основних бізнес-метрик, адже вона дуже сильно корелює з доходом, а також з чистим прибутком, бо ціна залучення користувача майже стала. Отже, цільова метрика - конверсія в підписку, як і будь-яка конверсія, ця метрика підпадає під біноміальний розподіл.

3.3 Гіпотеза експерименту

Ми припускаємо, що змінивши текст кнопки з “Start now” на “Try free & subscribe”, ми отримаємо приріст в конверсії в підписку. Отже формуємо гіпотезу. Контрольна група користувачів отримує стару версію екрану, а тестова група нову версію зі зміненим текстом кнопки. Нульова гіпотеза полягає в тому, що середнє значення конверсії в підписку не зміниться, а альтернативна гіпотеза полягає в тому, що воно покращиться.

$$\begin{cases} H_0 : p_{con} = p_{exp} \\ H_1 : p_{con} \neq p_{exp} \end{cases}$$

$$\begin{cases} H_0 : p_{con} - p_{exp} = 0 \\ H_1 : p_{con} - p_{exp} \neq 0 \end{cases}$$

3.4 Аналіз потужності

Ми знаємо, що зараз конверсія в підписку становить 3.1%, а також із історії попередніх експериментів знаємо, що аналогічні зміни можуть дати приріст в 10%, тобто до 3.41%. Потужність та значущість візьмемо класичними та отримаєм такі параметри експерименту

$$\alpha : 0.05$$
$$\beta : 0.2$$
$$(1 - \beta) : 0.8$$
$$\delta : 10\,\%$$

3.5 Розмір виборки та тривалість тесту

Отже, маючи усі параметри A/B тесту, можемо підставити їх у формулу для розрахунку розміру виборки для випадку бінарних метрик, та отримаємо наступне:

$$N = \frac{(\sqrt{2 * 0.031 * 0.969} \ * \ 1.96 \ + \ \sqrt{0.031 * 0.969 + 0.0341 * 0.9659} \ * \ 0.84)^2}{0.0031^2} = 49716$$

Тобто нам потрібно залучити по 49716 користувачів у кожну групу. Ми залучаємо в середньому по 7000 користувачіва на день, отже нам потрібно 15 днів. У такому бізнесі основний фактор сезонності — це вихідні дні. Проводячи експеримент 15 діб, ми остаточно позбуваємося сезонності. Також, проблеми новизни та звикання для нас теж неактуальні, бо в експерименті беруть участь лише нові користувачі, цільову зміну вони побачать вперше та востаннє.

3.6 Аналіз результатів

Сирі дані я перетворив на таблицю такого вигляду:

	user_id	group	result	number
44171	723387	A	0	1
88692	915720	B	0	1
29670	904759	A	0	2
68259	444084	B	0	2
48940	322358	A	0	3
61488	654581	B	0	3
25139	674276	A	0	4
79225	517391	B	0	4

Рисунок 3.2

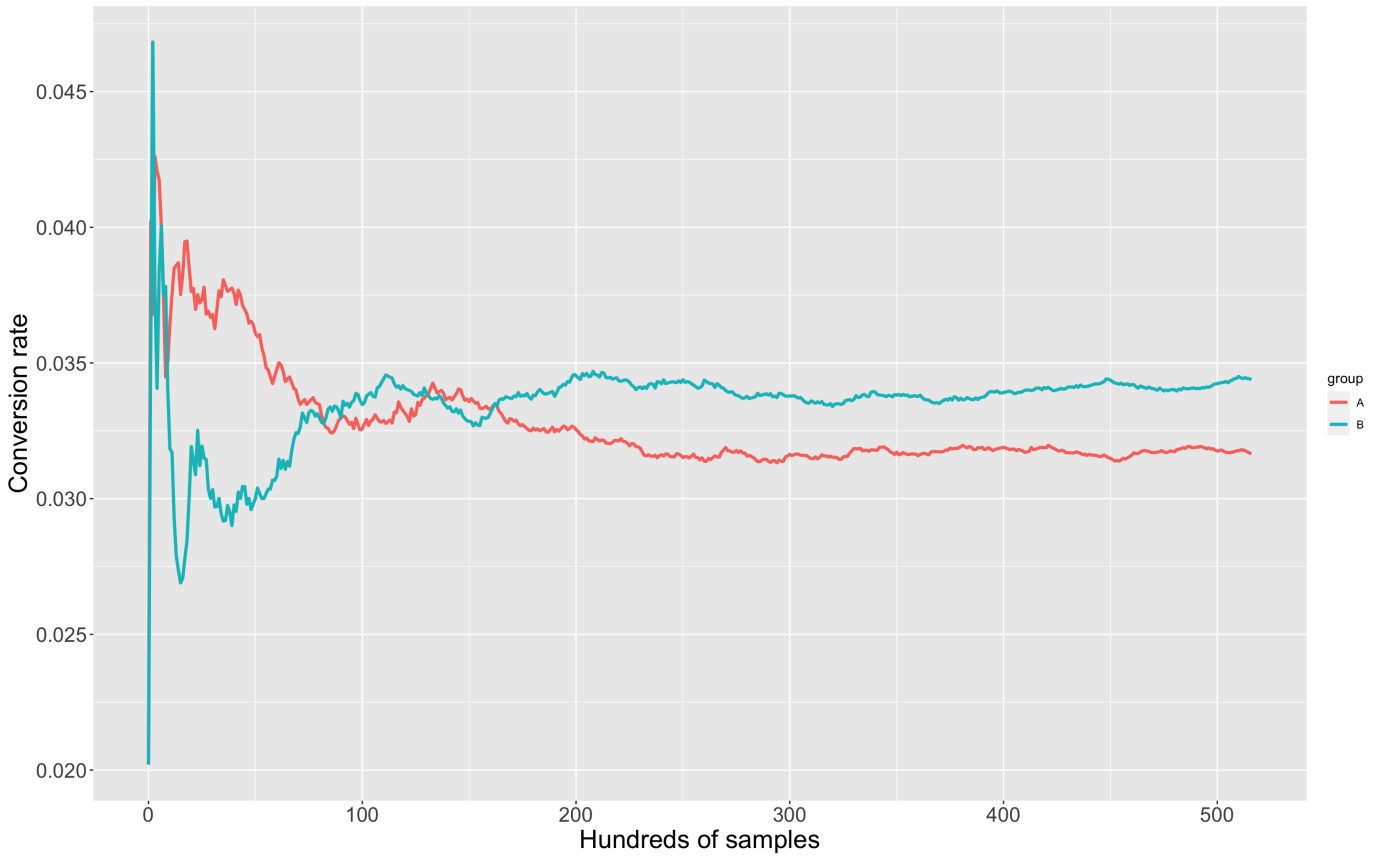

Рисунок 3.3 - Конверсії за плином часу

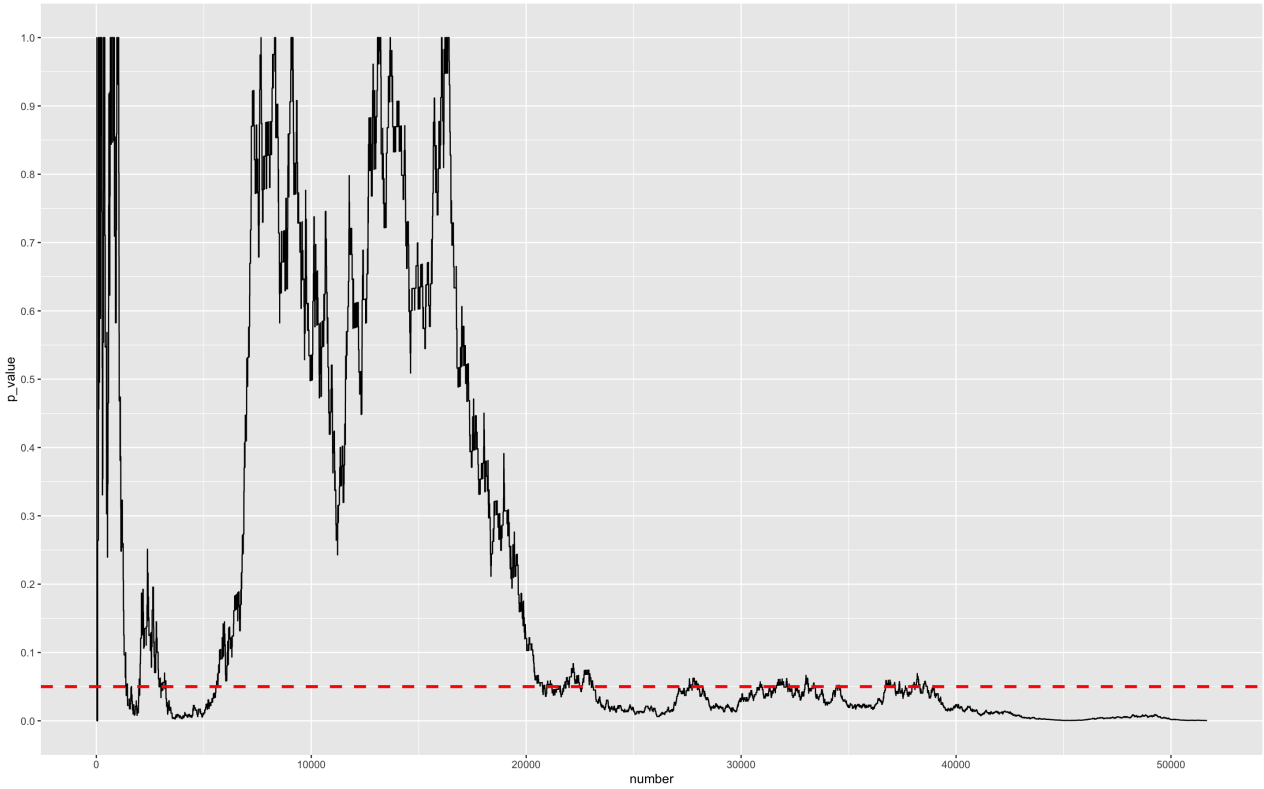

Рисунок 3.4 - P-value за плином часу

Таблиця містить наступні колонки:

- user_id — унікальний ідентифікатор користувача
- group — група в експерименті. A — контрольна, B — тестова
- result — бінарний результат дії користувача (0 — невдача, 1 — успіх)
- number — порядковий номер користувача у кожній з груп. Сформована за допомогою інформації про час виконання цільової дії

Підсумкова інформація про сформований датасет наступна:

Група	Кількість користувачів	Кількість підписавшихся	Конверсія
A	51676	1636	3.17%
B	51917	1785	3.44%

Таблиця 3.1

Конверсія має розподіл Бернуллі, отже ми можемо використати точний біноміальний тест для визначення p_value. Цей тест демонструє результат 0.00047, що є менше ніж цільовий рівень значущості 0.05. Зміна в цільовий метриці становить 8.6%, що є насправді менше, ніж MDE, отже наш експеримент є менш потужним, ніж ми очікували. Але й вибірка є трохи більшою, ніж мінімальна потрібна для заданих параметрів. Знову ж таки використуваючи формулу для визначення розміру вибірки, можемо знайти значення потужності, для якого при отриманому розміру виборки MDE буде дорівнювати нашому приросту (8.6%). Отримаємо близько 70%.

Вважати чи ні такий експеримент вдалим та рекомендуючим введення нового варіанту тексту — це дискусійне питання, але на практиці потужності в 70% буває достатньо.

Також для бінарних метрик цілком підійде критерій Стьюдента. За цим критерієм отримаємо p-value 0.0141, що також є прийнятним рівнєм значущості. Критерій Стьюдента є дуже популярним завдяки універсальності, але для конверсій все ж краще підійде точний біноміальний тест. Подивимося на те, як повели себе метрики конверсії та p-value з плином часу:

Цікаво, що тестова група вирівається уперед, починаючи з 16000 семплів у кожній групі. Також, p-value набуває сталості тільки ближче до 40000 семплів.

Загалом, експеримент можна скоріше вважати успішним, якщо звісно приріст у 8,6% є достатнім для компанії, аби окупити витрати на його запровадження.

ВИСНОВКИ

В кваліфікаційній роботі були досліджені A/B тести як метод перевірки статистичних гіпотез щодо розвитку IT-продуктів. У сучасному IT-бізнесі статистика набула багато застосувань, але A/B тести — це все ж найсучасніший спосіб прийняття рішень та валідації гіпотез, а також найбільш використоване застосування статистики. A/B тести дозволяють швидко та якісно перевіряти складні гіпотези щодо розвитку бізнесу та продукту, що в свою чергу дозволяє приймати більшу кількість чітко обґрунтованих рішеннь за меншу кількість часу. По суті, сучасні IT-продукти розвиваються за допомогою суто наукового підходу, що мабуть і дозволяє їм зростати так швидко и ефективно.

Але A/B тести мають свої очевидні обмеження: проведення такого експерименту потребує:

- Чіткої гіпотези, щодо того на яку метрику та за допомогою чого ми впливаємо
- Чіткої цільової метрики, яку можна дослідити A/в тестом
- Тестування тільки одної зміни в одному тесті
- Репрезантитвних та неупереджених контрольної та експериментальної вибірок
- Цілісного та незмінного експериментального впливу на користувачів

В практичній частині на прикладі реального експерименту було отримано не повністю однозначний результат, отже вкрай важливо пам’ятати, що A/B тести, як і будь-який інший інструмент, не є “срібною пулею” в проблемі перевірки статистичних гіпотез. Треба уважно ставитись до обмежень та вимог цього методу та розглядати кожний A/B тест як окремий незалежний експеримент, а не притримуватися шаблонів.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Young, Scott W. H. (August 2014). "Improving Library User Experience with A/B Testing: Principles and Process". Weave: Journal of Library User Experience. 1 (1).
2. Kohavi, Ron; Xu, Ya; Tang, Diane (2000). Trustworthy Online Controlled Experiments: A Practical Guide to A/B Testing. Cambridge University Press.
3. Kohavi, Ron; Longbotham, Roger (2017). "Online Controlled Experiments and A/B Tests" (PDF). In Sammut, Claude; Webb, Geoff (eds.). Encyclopedia of Machine Learning and Data Mining. Springer
4. Kohavi, Ron; Thomke, Stefan (September 2017). "The Surprising Power of Online Experiments". Harvard Business Review: pp. 74–82.
5. "The ABCs of A/B Testing - Pardot". Pardot. 12 July 2012. Retrieved 2016-02-21.
6. Kohavi, Ron; Longbotham, Roger (2017). "Online Controlled Experiments and A/B Testing". Encyclopedia of Machine Learning and Data Mining. pp. 922–929
7. Xu, Ya; Chen, Nanyu; Fernandez, Addrian; Sinno, Omar; Bhasin, Anmol (10 August 2015). "From Infrastructure to Culture: A/B Testing Challenges in Large Scale Social Networks". Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining: pp. 2227–2236.
8. Siroker, Dan; Koomen, Pete (2013-08-07). A / B Testing: The Most Powerful Way to Turn Clicks Into Customers. John Wiley & Sons. pp.
9. "Split Testing Guide for Online Stores". webics.com.au. August 27, 2012. Retrieved 2012-08-28.
10. Kaufman, Emilie (2014). "On the Complexity of A/B Testing" (PDF). 35. arXiv:1405.3224. Bibcode:2014arXiv1405.3224K – via JMLR: Workshop and Conference Proceedings.
11. Christian, Brian (2000-02-27). "The A/B Test: Inside the Technology That's Changing the Rules of Business | Wired Business". Wired.com. Retrieved 2014-03-18.
12. Stolberg, M (December 2006). "Inventing the randomized double-blind trial: the Nuremberg salt test of 1835". Journal of the Royal Society of Medicine. 99 (12): 642–643.

13. "What is A/B Testing." Convertize.
14. "Claude Hopkins Turned Advertising Into A Science." Retrieved 2019-11-01.
15. "Brief history and background for the one sample t-test". 20 June 2007.
16. Amazon.com. "The Math Behind A/B Testing". Archived from the original on 2015-09-21.
17. Kohavi, Ron; Longbotham, Roger; Sommerfield, Dan; Henne, Randal M. (February 2009). "Controlled experiments on the web: survey and practical guide". Data Mining and Knowledge Discovery. 18 (1): pp. 140–181.
18. Tatev, Karen. "Complete Guide to A/B Testing Design, Implementation and Pitfalls". [Електроний ресурс] // Towards Data Science. Medium. Режим доступу до ресурсу: https://towardsdatascience.com/simple-and-complet-guide-to-a-b-testing-c34154d0ce5a
19. Guenther, William. "A Sample Size Formula for a Non-Central t-Test". 12 Mar 2012, pp. 120-121.
20. Blaczka, David. "How not to run an A/B test". 31 Jan 2020. [Електроний ресурс] // Towards Data Science. Medium. Режим доступу до ресурсу: https://towardsdatascience.com/how-not-to-run-an-a-b-test-88637a6b921b

ДОДАТОК. КОД ПРОГРАМНОЇ РЕАЛІЗАЦІЇ ЕКСПЕРИМЕНТУ

```
df <- data.frame()

summary(df)

metrics <- df %>%
  group_by(group) %>%
  summarise(users = n_distinct(user_id),
            converted = sum(result == 1),
            cr = converted/users,
            std = sd(result)) %>% view()



users_test <- sum(df$group == 'B')
converted_test <- sum(df$result == 1 & df$group == 'B')
basic_CR <- sum(df$result == 1 & df$group == 'A') / sum(df$group == 'A')

binom.test(converted_test, users_test, basic_CR)$p.value
t.test(result ~ group, data=df)$p.value

df_filtered <- filter(df, number <= 51676)
p_values <- data.frame(number = seq(1,51676), p_value = 0)

for (i in c(1:51676)) {
  p_values$p_value[i] <- binom.test(sum(df_filtered$result == 1 &
df_filtered$group == 'B' & df_filtered$number <= i),
                                    sum(df_filtered$group == 'B' & df_filtered$number <= i),
                                    sum(df_filtered$result == 1 & df_filtered$group == 'A' &
df_filtered$number <= i) /
```

```
                                    sum(df_filtered$group == 'A' & df_filtered$number <=
i))$p.value
        }

        plot <- ggplot(p_values, aes(x = number, y = p_value)) +
          geom_line() +
          geom_hline(yintercept = 0.05, linetype="dashed", color = "red", size=1.2) +
          scale_y_continuous(breaks = seq(0,1,0.1))
        plot

        conversions <- df_filtered %>%
          group_by(group) %>%
          arrange(number) %>%
          mutate(cumulative_users = cumsum(!is.na(result)),
                 cumulative_converts = cumsum(result == 1)) %>%
          ungroup() %>%
          group_by(group, number %/% 100) %>%
          summarise(cr = max(cumulative_converts)/max(cumulative_users))

        plot_conversions <- ggplot(conversions, aes(x=`number%/%100`, y = cr, color =
group)) +
          geom_line(size = 1.2) +
          scale_y_continuous(breaks = seq(0,0.05,0.005)) +
          ylab('Conversion rate') +
          xlab('Hundreds of samples') +
          theme(axis.text.x = element_text(size=16),
             axis.text.y = element_text(size=16),
             axis.title = element_text(size=20))
        plot_conversions
```