# Домашнє завдання 5: Розрахунок розміру вибірки для біноміальної метрики

## Контекст завдання

**Продукт:** MelodyFlow — застосунок для стримінгу музики

**Нова функція:** "Розумні рекомендації плейлистів" — використання штучного інтелекту для автоматичного створення персоналізованих плейлистів на основі історії прослуховувань користувача.

**Мета функції:** Покращити залученість і утримання користувачів, спрощуючи пошук музики, яка їм подобається.

## Поточні дані

* **Загальна кількість користувачів:** 1,000,000
  + Безкоштовні: 600,000
  + Преміум: 400,000
* **DAU (Daily Active Users):** 20% від загальної кількості = 200,000 користувачів
* **Середня тривалість сесії:** 15 хвилин
* **Базовий рівень утримання (7-денний):** 41%
* **ARPPU (Average Revenue Per Paying User):** $25 за місяць

## Гіпотеза

**H₀ (нульова гіпотеза):** Функція "Розумні рекомендації плейлистів" не впливає на показник утримання клієнтів на 7-й день.

**H₁ (альтернативна гіпотеза):** Запровадження функції "Розумні рекомендації плейлистів" збільшить показник утримання клієнтів на 7-й день.

## 1. Визначення параметрів тесту

### 1.1 Базова метрика (p₁)

**Метрика:** 7-денне утримання (7-day retention)

**Базовий показник (контрольна група):** p₁ = 0.41 (41%)

### 1.2 Minimum Detectable Effect (MDE)

**Визначення MDE:** Мінімальний ефект, який ми хочемо виявити з достатньою статистичною потужністю.

**Обґрунтування:**

* Для метрик утримання типовий MDE становить 2-5% у відносному вираженні
* Враховуючи, що це нова AI-функція з потенційно значним впливом на користувацький досвід
* Баланс між бізнес-значущістю та практичною здійсненністю тесту

**Вибраний MDE:** 5% відносного покращення

* Абсолютне покращення: 0.41 × 0.05 = 0.0205
* **p₂ = 0.41 + 0.0205 = 0.4305 (43.05%)**

**Альтернативний сценарій (більш амбітний):**

* MDE: 10% відносного покращення
* Абсолютне покращення: 0.41 × 0.10 = 0.041
* p₂ = 0.41 + 0.041 = 0.451 (45.1%)

### 1.3 Рівень значущості (α)

**Обраний рівень:** α = 0.05 (5%)

**Обґрунтування:**

* Стандартний рівень для A/B тестів у продуктовій аналітиці
* Баланс між ризиком помилки I роду (хибно позитивний результат) та практичністю
* Означає 95% довірчий інтервал

**Інтерпретація:** Готові прийняти 5% шанс визнати ефект, якого насправді немає.

### 1.4 Статистична потужність (1 - β)

**Обрана потужність:** 1 - β = 0.80 (80%)

**Обґрунтування:**

* Стандартна практика в індустрії (мінімум 80%)
* β = 0.20 — ризик помилки II роду (не виявити ефект, який існує)
* Баланс між необхідним розміром вибірки та надійністю результатів

**Альтернатива для критичних рішень:** 90% потужність (β = 0.10)

### 1.5 Тип тесту

**Обраний тип:** Двосторонній тест (two-tailed test)

**Обґрунтування:**

* Хоча ми очікуємо покращення, потрібно перевірити можливість як позитивного, так і негативного ефекту
* Більш консервативний підхід, що захищає від несподіваних негативних результатів
* Стандартна практика для продуктових експериментів

**Примітка:** Односторонній тест був би можливий, якщо ми 100% впевнені, що ефект може бути тільки позитивним, але це рідко буває виправдано.

## 2. Розрахунок розміру вибірки

### 2.1 Формула для біноміальної метрики

Для розрахунку розміру вибірки при порівнянні двох пропорцій використовуємо формулу:

n = (Z\_{α/2} + Z\_β)² × [p₁(1-p₁) + p₂(1-p₂)] / (p₂ - p₁)²

Де:

* **n** — розмір вибірки для кожної групи
* **Z\_{α/2}** — Z-значення для рівня значущості (двосторонній тест)
* **Z\_β** — Z-значення для статистичної потужності
* **p₁** — базова пропорція (контрольна група)
* **p₂** — очікувана пропорція (тестова група)

### 2.2 Визначення Z-значень

Для **α = 0.05** (двосторонній тест):

* **Z\_{α/2} = 1.96** (критичне значення для 95% довірчого інтервалу)

Для **статистичної потужності 80%** (β = 0.20):

* **Z\_β = 0.84** (критичне значення для 80% потужності)

Для **статистичної потужності 90%** (β = 0.10):

* **Z\_β = 1.28** (критичне значення для 90% потужності)

### 2.3 Розрахунок для основного сценарію

**Параметри:**

* p₁ = 0.41
* p₂ = 0.4305 (MDE = 5% відносне покращення)
* α = 0.05 (Z\_{α/2} = 1.96)
* Потужність = 80% (Z\_β = 0.84)

**Крок 1:** Розрахунок чисельника

(Z\_{α/2} + Z\_β)² = (1.96 + 0.84)² = (2.80)² = 7.84

**Крок 2:** Розрахунок дисперсій

p₁(1-p₁) = 0.41 × (1 - 0.41) = 0.41 × 0.59 = 0.2419

p₂(1-p₂) = 0.4305 × (1 - 0.4305) = 0.4305 × 0.5695 = 0.2452

**Крок 3:** Сума дисперсій

p₁(1-p₁) + p₂(1-p₂) = 0.2419 + 0.2452 = 0.4871

**Крок 4:** Розрахунок різниці пропорцій

(p₂ - p₁)² = (0.4305 - 0.41)² = (0.0205)² = 0.00042025

**Крок 5:** Фінальний розрахунок

n = 7.84 × 0.4871 / 0.00042025

n = 3.819 / 0.00042025

n ≈ 9,097 користувачів на групу

**Загальна вибірка:**

* **Контрольна група (A):** 9,097 користувачів
* **Тестова група (B):** 9,097 користувачів
* **РАЗОМ:** 18,194 користувачів

*(Примітка: точний розрахунок дає 9,097, округлено до 9,100 для зручності)*

## 3. Додаткові розрахунки

### 3.1 Сценарій з вищою потужністю (90%)

**Параметри:**

* p₁ = 0.41, p₂ = 0.4305
* α = 0.05, Потужність = 90%

(Z\_{α/2} + Z\_β)² = (1.96 + 1.28)² = (3.24)² = 10.50

n = 10.50 × 0.4871 / 0.00042025

n ≈ 12,179 користувачів на групу

**РАЗОМ:** 24,357 користувачів

**РАЗОМ:** 24,330 користувачів

### 3.2 Сценарій з більшим MDE (10% покращення)

**Параметри:**

* p₁ = 0.41, p₂ = 0.451
* α = 0.05, Потужність = 80%

p₂(1-p₂) = 0.451 × 0.549 = 0.2476

Сума дисперсій = 0.2419 + 0.2476 = 0.4895

(p₂ - p₁)² = (0.041)² = 0.001681

n = 7.84 × 0.4895 / 0.001681

n ≈ 2,286 користувачів на групу

**РАЗОМ:** 4,572 користувачів

### 3.3 Односторонній тест (для порівняння)

Якби ми використовували односторонній тест:

* Z\_{α} = 1.645 (замість 1.96)
* (1.645 + 0.84)² = 6.17

n = 6.17 × 0.4871 / 0.00042025

n ≈ 7,166 користувачів на групу

**РАЗОМ:** 14,332 користувачів

## 4. Практична здійсненність тесту

### 4.1 Оцінка тривалості тесту

**Дані:**

* DAU = 200,000 користувачів щодня
* Необхідна вибірка = 18,176 користувачів
* Розподіл 50/50 між групами

**Якщо включити всіх DAU в експеримент:**

* Кожна група отримує 100,000 користувачів на день
* Необхідно 9,097 користувачів на групу для вимірювання 7-денного утримання

**Мінімальна тривалість тесту:**

* 1 день збору + 7 днів спостереження = **8 днів мінімум**
* Рекомендовано: **14-21 день** для стабільності та врахування weekly patterns

### 4.2 Розподіл трафіку

**Варіант 1: Агресивний (рекомендований)**

* 50% користувачів в контрольну групу
* 50% користувачів в тестову групу
* Швидкий набір вибірки (1-2 дні)

**Варіант 2: Консервативний**

* Якщо є побоювання щодо ризиків
* 90% — контроль, 10% — тест
* Тривалість: ~10 днів набору вибірки

### 4.3 Достатність ресурсів

**Наявні ресурси:**

* DAU = 200,000 користувачів
* Необхідна вибірка = 18,194 користувачів (9.1% від DAU)

**Висновок:** ✅ Тест **цілком здійсненний** з наявною користувацькою базою

## 5. Зведена таблиця сценаріїв

| **Сценарій** | **MDE** | **α** | **Потужність** | **Тип тесту** | **n (на групу)** | **Загалом** | **Тривалість\*** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Основний (рекомендований)** | 5% відн. | 0.05 | 80% | Двосторонній | 9,097 | 18,194 | 8-14 днів |
| Висока потужність | 5% відн. | 0.05 | 90% | Двосторонній | 12,179 | 24,357 | 8-14 днів |
| Більший ефект | 10% відн. | 0.05 | 80% | Двосторонній | 2,286 | 4,572 | 8-14 днів |
| Односторонній | 5% відн. | 0.05 | 80% | Односторонній | 7,166 | 14,332 | 8-14 днів |

\*Тривалість включає 1 день набору + 7 днів вимірювання + запас для стабільності

## 6. Висновки та рекомендації

### 6.1 Підсумок розрахунків

**Рекомендований дизайн тесту:**

1. **Метрика:** 7-денне утримання (7-day retention rate)
2. **Базова ставка (p₁):** 41%
3. **Очікуваний ефект (p₂):** 43.05% (5% відносне покращення)
4. **MDE:** 2.05 процентних пункти (5% відносне покращення)
5. **Рівень значущості (α):** 0.05
6. **Статистична потужність:** 80%
7. **Тип тесту:** Двосторонній
8. **Розмір вибірки:** 9,097 користувачів на групу (округлено до 9,100)
9. **Загальна вибірка:** 18,194 користувачів (округлено до 18,200)
10. **Тривалість:** 14-21 день (рекомендовано)

### 6.2 Обґрунтування вибору параметрів

**MDE (5% відносне покращення):**

* Баланс між бізнес-значущістю та практичністю
* Достатньо амбітний для AI-функції
* Реалістичний для виявлення з розумним розміром вибірки

**α = 0.05:**

* Індустріальний стандарт
* Прийнятний рівень ризику для продуктових експериментів

**Потужність 80%:**

* Мінімальний рекомендований рівень
* Дає 80% шанс виявити ефект, якщо він існує

**Двосторонній тест:**

* Захищає від несподіваних негативних ефектів
* Більш консервативний та надійний підхід

### 6.3 Практичні рекомендації

1. **Розподіл трафіку:** 50/50 між контрольною та тестовою групами
2. **Тривалість:** Мінімум 14 днів для врахування weekly seasonality
3. **Моніторинг:** Щоденний моніторинг ключових метрик та guardrail метрик
4. **Додаткові метрики:**
   * Середня тривалість сесії
   * Кількість прослуханих треків
   * Engagement з рекомендованими плейлистами
   * Частота використання функції

### 6.4 Ризики та мітігація

**Ризики:**

* Сезонність та зовнішні фактори (свята, події)
* Технічні проблеми з AI-рекомендаціями
* Novelty effect (початковий ентузіазм користувачів)

**Мітігація:**

* Довша тривалість тесту (21 день)
* Guardrail метрики для виявлення технічних проблем
* Пост-ланч моніторинг для виявлення довгострокових ефектів

### 6.5 Критерії успіху

Тест буде вважатися успішним, якщо:

1. **Первинна метрика:** 7-денне утримання зросте з 41% до щонайменше 43.05% (p-value < 0.05)
2. **Guardrail метрики:** Не погіршуються (ARPPU, revenue, технічна стабільність)
3. **Engagement метрики:** Показують позитивну динаміку

### 6.6 Наступні кроки

1. ✅ Розрахунок розміру вибірки завершено
2. ⏭️ Налаштування інфраструктури A/B тесту
3. ⏭️ Визначення guardrail метрик
4. ⏭️ Створення моніторингу та dashboards
5. ⏭️ Запуск тесту
6. ⏭️ Аналіз результатів після 14-21 дня

## 7. Математичне формулювання

### 7.1 Детальна формула з поясненнями

Повна формула для розрахунку розміру вибірки при порівнянні двох пропорцій:

n = [(Z\_{α/2} × √(2p̄(1-p̄)) + Z\_β × √(p₁(1-p₁) + p₂(1-p₂)))² / (p₂ - p₁)²]

Де:

* **p̄** = (p₁ + p₂) / 2 — об'єднана пропорція

**Спрощена версія** (яку ми використовували):

n = (Z\_{α/2} + Z\_β)² × [p₁(1-p₁) + p₂(1-p₂)] / (p₂ - p₁)²

Ця спрощена формула дає дуже близькі результати та частіше використовується на практиці.

### 7.2 Перевірка розрахунків за альтернативним методом

**Метод Evan Miller (онлайн калькулятор):**

* Baseline: 41%
* Expected improvement: 5% relative (2.05 pp)
* Result: ~9,000 користувачів на групу ✅

**Метод statsmodels (Python):**

from statsmodels.stats.power import zt\_ind\_solve\_power

from statsmodels.stats.proportion import proportion\_effectsize

effect\_size = proportion\_effectsize(0.41, 0.4305)

n = zt\_ind\_solve\_power(effect\_size=effect\_size, alpha=0.05, power=0.80, alternative='two-sided')

# Result: ~9,088 ✅

### 7.3 Формула для відносного покращення

Відносне покращення = (p₂ - p₁) / p₁ × 100%

= (0.4305 - 0.41) / 0.41 × 100%

= 0.0205 / 0.41 × 100%

= 5%

## 

## 

## 

## 

## 

## 

## 

## Додаток: Python код для верифікації

import numpy as np

from scipy import stats

from statsmodels.stats.power import zt\_ind\_solve\_power

from statsmodels.stats.proportion import proportion\_effectsize

# Параметри

p1 = 0.41 # Baseline retention

relative\_improvement = 0.05 # 5%

p2 = p1 \* (1 + relative\_improvement) # 0.4305

alpha = 0.05

power = 0.80

# Z-значення

z\_alpha = stats.norm.ppf(1 - alpha/2) # 1.96 для двостороннього

z\_beta = stats.norm.ppf(power) # 0.84 для 80% потужності

# Розрахунок за формулою

numerator = (z\_alpha + z\_beta)\*\*2

variance\_sum = p1\*(1-p1) + p2\*(1-p2)

diff\_squared = (p2 - p1)\*\*2

n\_per\_group = numerator \* variance\_sum / diff\_squared

print(f"Параметри:")

print(f" p1 (baseline): {p1:.4f}")

print(f" p2 (expected): {p2:.4f}")

print(f" Absolute difference: {p2-p1:.4f}")

print(f" Relative improvement: {relative\_improvement\*100:.1f}%")

print(f" Alpha: {alpha}")

print(f" Power: {power}")

print(f"\nРозрахунок:")

print(f" Z\_alpha/2: {z\_alpha:.4f}")

print(f" Z\_beta: {z\_beta:.4f}")

print(f" (Z\_alpha/2 + Z\_beta)^2: {numerator:.4f}")

print(f" Variance sum: {variance\_sum:.4f}")

print(f" Difference squared: {diff\_squared:.6f}")

print(f"\nРезультат:")

print(f" Розмір вибірки на групу: {n\_per\_group:.0f}")

print(f" Загальна вибірка: {n\_per\_group\*2:.0f}")

# Верифікація через statsmodels

effect\_size = proportion\_effectsize(p1, p2)

n\_statsmodels = zt\_ind\_solve\_power(

effect\_size=effect\_size,

alpha=alpha,

power=power,

alternative='two-sided'

)

print(f"\nВерифікація (statsmodels):")

print(f" Effect size: {effect\_size:.4f}")

print(f" Розмір вибірки на групу: {n\_statsmodels:.0f}")

print(f" Різниця: {abs(n\_per\_group - n\_statsmodels):.0f} користувачів")