АБ Тестирование

Иван Новиков

МФТИ

7 декабря 2024 г.

Постановка

Есть две выборки, хотим проверить неравенство нулю разности математических ожиданий между ними

- $X_1, \ldots, X_n$ - выборка А (контроль)
- $Y_1, \ldots, Y_m$ - выборка В (тест), к ней применяем изменение
- $H_0 : \mathbb{E}X = (\leq, \geq)\mathbb{E}Y$ - эффекта нет
- $H_1 : \mathbb{E}X \neq (>, <)\mathbb{E}Y$ - эффект есть

t-test для двух независимых выборок 1

$$X_1, \ldots, X_n \sim \mathcal{N}(a_X, \sigma_X^2)$$
$$Y_1, \ldots, Y_m \sim \mathcal{N}(a_Y, \sigma_Y^2)$$

Случай 1 (t-test)

$\sigma = \sigma_X = \sigma_Y$ - неизвестны
S_Y^2, S_X^2 - несмещённые оценки дисперсий у Y и X

$$S^2 = \frac{(m-1)S_Y^2 + (n-1)S_X^2}{n+m-2}$$

$$T(X, Y) = \frac{\overline{Y} - \overline{X}}{S\sqrt{\frac{1}{m} + \frac{1}{n}}} \overset{H_0}{\sim} T_{n+m-2}$$

t-test для двух независимых выборок 2

$$X_1, \dots, X_n \sim \mathcal{N}(a_X, \sigma_X^2)$$

$$Y_1, \dots, Y_m \sim \mathcal{N}(a_Y, \sigma_Y^2)$$

Случай 2 (Тест Уэлча)

$\sigma_X \neq \sigma_Y$ - неизвестны

S_Y^2, S_X^2 - несмещённые оценки дисперсий у Y и X

$$\nu \approx \frac{\left(\frac{S_Y^2}{m} + \frac{S_X^2}{n}\right)^2}{\frac{S_Y^4}{m^2(m-1)} + \frac{S_X^4}{n^2(n-1)}}$$

$$T(X, Y) = \frac{\overline{Y} - \overline{X}}{\sqrt{\frac{S_Y^2}{m} + \frac{S_X^2}{n}}} \overset{H_0}{\sim} T_\nu$$

Общий случай для независимых выборок

Пусть $X_1, \dots, X_n,\ Y_1, \dots, Y_m$ - независимые выборки
Тогда

$$T(X, Y) = \frac{\overline{Y} - \overline{X}}{\sqrt{\frac{S_Y^2}{m} + \frac{S_X^2}{n}}} \xrightarrow{d_0} \mathcal{N}(0, 1)$$

Общий случай для связанных выборок

Сводим к одновыборочному тесту

$$\delta_i = Y_i - X_i$$

Требование: $\delta_1, \ldots, \delta_n$ - выборка, конечная дисперсия

$$H_0 : \mathbb{E}\delta = 0$$

$$H_1 : \mathbb{E}\delta \neq 0$$

$$T(X, Y) = \sqrt{n}\frac{\overline{\delta}}{S_\delta} \xrightarrow{d_0} \mathcal{N}(0, 1)$$

Другие тесты

Но что если данные не из $\mathcal{N}(a, \sigma^2)$?

- Если данных много, то можно вспомнить общий случай
- Также можно использовать бутстреп, построить бутстрепированную выборку статистик T, построить доверительный интервал - если 0 не попадает в интервал, то отвергаем H_0
- Если всё же данных мало - смотрим на распределение метрики и исходя из этого делаем тест (Точный тест Фишера, E-test, критерий χ^2, ...)
- Если даже распределение непонятно, то U-критерий Манна — Уитни

U-критерий Манна-Уитни

- $X_1, \ldots, X_n$, $Y_1, \ldots, Y_m$ - независимые выборки, $Z = X_1, \ldots, X_n, Y_1, \ldots, Y_m$, $R_1, \ldots, R_{n+m}$ - ранги
- $F_X(x + \mu) = F_Y(x)$
- $H_0 : \mu = 0$, $H_1 : \mu \neq 0$

$$R_X = \sum_{i=1}^{n} R_{X_i}, R_Y = \sum_{i=1}^{m} R_{Y_i}$$

$$U_X = nm + \frac{n(n+1)}{2} - R_X, U_Y = nm + \frac{m(m+1)}{2} - R_Y$$

$$U = \min\{U_X, U_Y\} \xrightarrow{d} \mathcal{N}\left(\frac{nm}{2}, \frac{nm(n+m+1)}{12}\right)$$

AA-test

Можно провалидировать тест

- **В онлайне**
  Чаще всего берут 2 группы без изменений и делают несколько раз тест, проверяют что тест не детектирует эффект
- **В офлайне**
  Берём исторические данные и делаем много раз тест, получаем выборку отвержений и неотвержений $a_1, \dots, a_n$ (например бутстрепом), проверяем $H_0 : \mathbb{E}a \leq \alpha$

Метод бакетов

Иногда у данных есть проблемы:

- Слишком много данных
- Зависимые данные (а то есть не являются выборкой) - например сессии пользователей

Метод бакетов

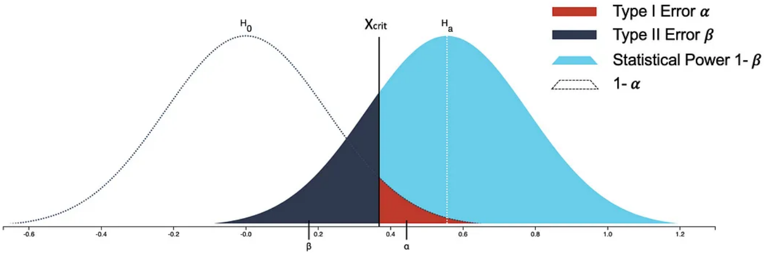
B - количество бакетов

$bucket\_id(user\_id) = hash(user\_id)\%B$

$Y_i = mean(\{X_{user\_id} | bucket\_id(user\_id) = i\})$

Получаем $Y_1, Y_2, \ldots, Y_B$ - выборка бакетов

Ошибки



	H0 is true	H0 is false
Reject H0	Type I Error	Good! (Statistical Power)
Accept H0	Good!	Type II Error

Рис. 1: Ошибки

MDE

Theorem (MDE)

1. $X_1, \ldots, X_n$, $Y_1, \ldots, Y_m$ - независимые выборки
2. α - уровень значимости критерия
3. $1 - \beta$ - мощность критерия
4. e - эффект
5. $H_0 : \mathbb{E}Y \leq \mathbb{E}X$ и $H_1 : \mathbb{E}Y > \mathbb{E}X$ при односторонней альтернативе
6. $H_0 : \mathbb{E}Y = \mathbb{E}X$ и $H_1 : \mathbb{E}Y \neq \mathbb{E}X$ при двусторонней альтернативе

Тогда

1. при односторонней альтернативе $e \geq (z_{1-\alpha} + z_{1-\beta})\sqrt{\frac{S_Y^2}{m} + \frac{S_X^2}{n}}$
2. при двусторонней альтернативе $e \geq (z_{1-\frac{\alpha}{2}} + z_{1-\beta})\sqrt{\frac{S_Y^2}{m} + \frac{S_X^2}{n}}$

MDE

Доказательство (MDE)

$$P_0\left(\frac{\overline{Y}-\overline{X}}{\sqrt{\frac{S_Y^2}{m}+\frac{S_X^2}{n}}} > z_{1-\alpha}\right) \le \alpha, \quad P_e\left(\frac{\overline{Y}-\overline{X}}{\sqrt{\frac{S_Y^2}{m}+\frac{S_X^2}{n}}} > z_{1-\alpha}\right) \ge 1-\beta$$

$$P_e\left(\frac{\overline{Y}-\overline{X}}{\sqrt{\frac{S_Y^2}{m}+\frac{S_X^2}{n}}} > z_{1-\alpha}\right) = P_e\left(\frac{\overline{Y}-\overline{X}-e}{\sqrt{\frac{S_Y^2}{m}+\frac{S_X^2}{n}}} > z_{1-\alpha} - \frac{e}{\sqrt{\frac{S_Y^2}{m}+\frac{S_X^2}{n}}}\right)$$

$$\frac{(\overline{Y}-\overline{X})-e}{\sqrt{\frac{S_Y^2}{m}+\frac{S_X^2}{n}}} \xrightarrow{d_e} \mathcal{N}(0,1) \Rightarrow P_e\left(\frac{\overline{Y}-\overline{X}}{\sqrt{\frac{S_Y^2}{m}+\frac{S_X^2}{n}}} > z_{1-\alpha}\right) \to 1-\Phi\left(z_{1-\alpha} - \frac{e}{\sqrt{\frac{S_Y^2}{m}+\frac{S_X^2}{n}}}\right)$$

MDE

Доказательство (MDE)

$$1 - \Phi\left(z_{1-\alpha} - \frac{e}{\sqrt{\frac{S_Y^2}{m} + \frac{S_X^2}{n}}}\right) \geq 1 - \beta \quad \Rightarrow \quad \Phi\left(z_{1-\alpha} - \frac{e}{\sqrt{\frac{S_Y^2}{m} + \frac{S_X^2}{n}}}\right) \leq \beta$$

$$z_{1-\alpha} - \frac{e}{\sqrt{\frac{S_Y^2}{m} + \frac{S_X^2}{n}}} \leq \Phi^{-1}(\beta) = z_\beta$$

$$e \geq (z_{1-\alpha} - z_\beta)\sqrt{\frac{S_Y^2}{m} + \frac{S_X^2}{n}} = (z_{1-\alpha} + z_{1-\beta})\sqrt{\frac{S_Y^2}{m} + \frac{S_X^2}{n}}$$

Алалогично для двусторонней альтернативы

Также для связанных выборок можно доказать аналогичную теорему

Минимальное количество данных

Узнав эффект, можно получить необходимое количество данных для теста:
Пусть $m = n$

1. при односторонней альтернативе

$$n \geq \frac{(z_{1-\alpha} + z_{1-\beta})^2 (S_X^2 + S_Y^2)}{e^2}$$

2. при двусторонней альтернативе

$$n \geq \frac{(z_{1-\frac{\alpha}{2}} + z_{1-\beta})^2 (S_X^2 + S_Y^2)}{e^2}$$

Вывод

Чтобы уменьшить необходимое количество данных, нужно снижать дисперсию метрики

Стратификация

1. $Y_{1,1}, \ldots, Y_{1,n_1}, \ldots, Y_{K,n_K}$ - метрика
2. Y_k - k-я страта
3. n_k - количество элементов в k-й страте
4. $p_k = \frac{n_k}{n}$ - доля k-й страты
5. $\overline{Y}_k = \sum_{j=1}^{n_k} Y_{k,j}$ - среднее в k-й страте
6. $\mu = \mathbb{E}Y$
7. $\mu_k = \mathbb{E}Y_k$
8. $\sigma^2 = \mathbb{V}Y$
9. $\sigma_k^2 = \mathbb{V}Y_k$

Определение средних

1. $\overline{Y} = \frac{1}{n}\sum_{k=1}^{K}\sum_{j=1}^{n_k} Y_{k,j}$ - обычное среднее
2. $\overline{Y}_{strat} = \sum_{k=1}^{K} p_k \overline{Y}_k$ - стратифицированное среднее

Стратификация

Вместо обычного случайного семплирования можно брать данные для эксперимента так, чтобы доля каждой страты была такой же как и в изначальных данных (то откуда семплируем)

Утверждение (Равенство среднего и стратифицированного среднего)

При стратифицированном семплировании

$$\overline{Y} = \overline{Y}_{strat}$$

Утверждение (Несмещённость стратифицированного среднего при стратифицированном семплировании)

$$\mathbb{E}_{strat}\overline{Y}_{strat} = \mu = \mathbb{E}\overline{Y}$$

Доказательства всех утверждений со стратификацией можно найти в [AX16]

Стратификация

Утверждение (Уменьшение дисперсии при стратификации)

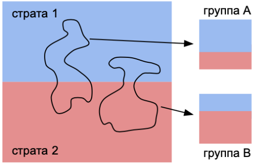
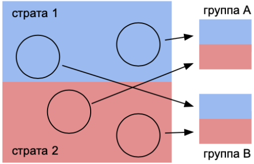
1. $\mathbb{V}\overline{Y} = \frac{1}{n}\sum_{k=1}^{K} p_k\sigma_k^2 + \frac{1}{n}\sum_{k=1}^{K} p_k(\mu_k - \mu)^2$ - обычное усреднение
2. $\mathbb{V}\overline{Y}_{strat} = \frac{1}{n}\sum_{k=1}^{K} p_k\sigma_k^2 + \frac{1}{n^2}\sum_{k=1}^{K}(1 - p_k)\sigma_k^2 + O\left(\frac{1}{n^2}\right)$ - постстратификация
3. $\mathbb{V}_{strat}\overline{Y}_{strat} = \frac{1}{n}\sum_{k=1}^{K} p_k\sigma_k^2$ - стратификация

При этом

$$\mathbb{V}_{strat}\overline{Y}_{strat} = \mathbb{V}\overline{Y}_{strat} + O\left(\frac{1}{n^2}\right) = \mathbb{V}\overline{Y} + O\left(\frac{1}{n}\right)$$

$$\mathbb{V}_{strat}\overline{Y}_{strat} \leq \mathbb{V}\overline{Y}_{strat} \leq \mathbb{V}\overline{Y}$$

Стратификация

	Обычное среднее $\overline{Y} = \frac{1}{n}\sum_{k=1}^{K}\sum_{j=1}^{n_k} Y_{kj}$	**Стратифицированное среднее** $\hat{Y}_{strat} = \sum_{k=1}^{K} w_k \overline{Y}_k$
Случайное семплирование 	Классический подход Без стратификации	**Постстратификация**
Стратифицированное семплирование 	Не контролирует вероятность ошибки I рода	**Стратификация**

CUPED

Controlled-experiment Using Pre-Experiment Data - CUPED

1. Y - метрика
2. X - ковариата (коррелируящая с метрикой величина, не затронутая изменением (которое мы применяем к тестовой выборке))

$$Y_{CUPED} = Y - \theta X + \theta \mathbb{E} X$$

Утверждение (лучший выбор θ) [Den+13]

При $\theta = \frac{cov(X,Y)}{\mathbb{V}X}$ достигается

$$\min(\mathbb{V}\overline{Y}_{CUPED}) = \mathbb{V}\overline{Y}\left(1 - \left(\frac{cov(X,Y)}{\sqrt{\mathbb{V}X\mathbb{V}Y}}\right)^2\right) = \mathbb{V}\overline{Y}(1 - corr^2(X,Y))$$

Вариации CUPED

CUPAC (Control Using Predictions as Covariate)

X - предсказание метрики ML моделью

$$Y_{CUPAC} = Y - \theta X + \theta \mathbb{E} X$$

CUMPED (Controlled-experiment Using Multiple Pre-Experiment Data)

$X^1, \dots, X^m$ - ковариаты

$$Z^1_{CUMPED} = Y - \theta_1 X^1 + \theta_1 \mathbb{E} X^1$$

$$Z^2_{CUMPED} = Z^1_{CUMPED} - \theta_2 X^2 + \theta_2 \mathbb{E} X^2$$

$$\dots$$

$$Z^m_{CUMPED} = Z^{m-1}_{CUMPED} - \theta_m X^m + \theta_m \mathbb{E} X^m$$

Несколько ковариат в CUPED

Несколько ковариат в CUPED

- $X_1, X_2, ..., X_k$ - ковариаты
- $\Sigma = cov(X_1, X_2, ..., X_k)$
- $Z = \begin{pmatrix} cov(Y, X_1) \\ \dots \\ cov(Y, X_k) \end{pmatrix}$

Тогда при преобразовании $Y_{MultiCUPED} = Y - \sum_{i=1}^{k} \theta_i (X_i - \mathbb{E}X_i)$ минимальная дисперсия достигается на

$$\theta = \Sigma^{-1} Z$$

Связь CUPED и стратификации

Связь CUPED и стратификации [Den+13]

1. $\overline{Y}_{strat} = \sum\limits_{i=1}^{k} p_i \overline{Y}_i$ - стратифицированное среднее
2. X_i - бинарная переменная о принадлежности к страте, $p_i = \mathbb{E}X_i$
3. $\overline{Y}_{CUPED} = \overline{Y} - \sum\limits_{i=1}^{k} \theta_i(\overline{X_i} - p_i)$ - CUPED

Тогда

$$\overline{Y}_{CUPED} = \overline{Y}_{strat}$$

Метрики

U - объекты

- Средние

$$M(U) = \frac{\sum\limits_{u \in U} F(u)}{\sum\limits_{u \in U} 1} = \frac{\sum\limits_{u \in U} F(u)}{|U|}$$

- Ratio-метрики

$$M(U) = \frac{\sum\limits_{u \in U} F(u)}{\sum\limits_{u \in U} G(u)}$$

Ratio-метрики

Предусреднение

Усреднить метрику вида $\dfrac{\sum_{u \in U} \frac{F(u)}{G(u)}}{|U|}$ некорректно, т.к. эффект может носить свойство разнонаправленности

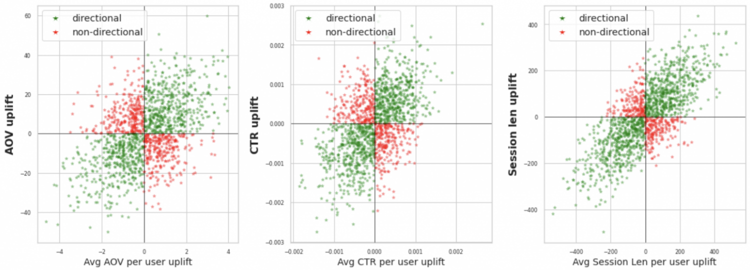


Рис. 3: Разнонаправленность, https://habr.com/ru/companies/kuper/articles/768826/

Ratio-метрики

Дисперсия ratio [Sel]

Пусть F, G - случайные величины, тогда

$$\mathbb{V}\frac{F}{G} \approx \frac{(\mathbb{E}F)^2}{(\mathbb{E}G)^2}\left(\frac{\mathbb{V}F}{(\mathbb{E}F)^2} - 2\frac{cov(F,G)}{\mathbb{E}F\mathbb{E}G} + \frac{\mathbb{V}G}{(\mathbb{E}G)^2}\right)$$

Дельта-метод

1. $X_1, \dots, X_n,\ Y_1, \dots, Y_n$ - независимые выборки,
2. M_{ratio} - ratio-метрика
3. $V_{ratio}(X)$ - оценка дисперсии для отношения

Тогда

$$\frac{M_{ratio}(Y) - M_{ratio}(X)}{\sqrt{\frac{V_{ratio}(Y)}{m} + \frac{V_{ratio}(X)}{n}}} \xrightarrow{d_0} \mathcal{N}(0,1)$$

Ratio-метрики

Линеаризация [Bud+18]

1. $X = X_A \cup X_B,\ Y = Y_A \cup Y_B,$
2. $M_{ratio}(U) = \frac{\sum\limits_{u \in U} X(u)}{\sum\limits_{u \in U} Y(u)}$ - ratio-метрика
3. $l(u) = X(u) - Y(u) M_{ratio}(X_A)$, ($M_{ratio}(X_A)$ - значение метрики в контрольной группе)
4. $L(U) = \{l(u) | u \in U\},\ \overline{L}(U) = \frac{\sum\limits_{u \in U} l(u)}{|U|}$
5. $T(L) = \frac{\overline{L}(B) - \overline{L}(A)}{\sqrt{\frac{S^2_{L(B)}}{m} + \frac{S^2_{L(A)}}{n}}},\ D(M_{ratio}) = \frac{M_{ratio}(B) - M_{ratio}(A)}{\sqrt{\frac{V_{ratio}(B)}{m} + \frac{V_{ratio}(A)}{n}}}$

Тогда

Ratio-метрики

Линеаризация [Bud+18]

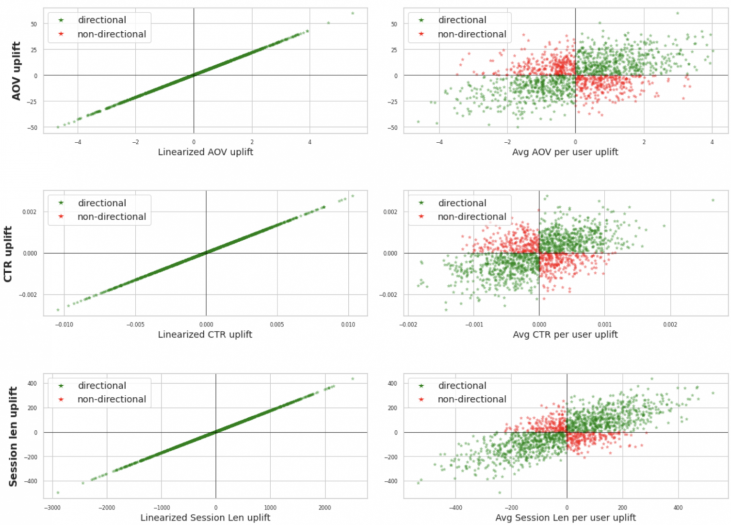
1. для некоторого γ

$$T(L) = D(M_{ratio})\sqrt{1 - \frac{\gamma}{\frac{V_{ratio}(B)}{m} + \frac{V_{ratio}(A)}{n} + \gamma}}$$

2. Если $|\widehat{corr}_B(X, Y)| < c < 1$, то $\exists C_1(c), C_2(c), \varepsilon_1(c), \varepsilon_2(c)$ т.ч. $\left|\frac{\overline{X}_B - \overline{X}_A}{\overline{X}_B}\right| < \varepsilon_1(c)$ и $\left|\frac{\overline{Y}_B - \overline{Y}_A}{\overline{Y}_B}\right| < \varepsilon_2(c)$ и $\left|\frac{T(L)}{D(M_{ratio})} - 1\right| \le C_1(c)\left|\frac{\overline{X}_B - \overline{X}_A}{\overline{X}_B}\right| + C_2(c)\left|\frac{\overline{Y}_B - \overline{Y}_A}{\overline{Y}_B}\right|$
3. Если $|corr(X, Y|B)| < c < 1$, $\mathbb{E}(X|A) \neq 0$, $\mathbb{E}(Y|A) \neq 0$, то

$$T(L) \xrightarrow{d_0} \mathcal{N}(0, 1)$$

Ratio-метрики



Литература

[Den+13] Alex Deng и др. «Improving the Sensitivity of Online Controlled Experiments by Utilizing Pre-Experiment Data». В: (2013). URL: https://exp-platform.com/Documents/2013-02-CUPED-ImprovingSensitivityOfControlledExperiments.pdf.

[AX16] Juliette Aurisset и Huizhi Xie. «Improving the Sensitivity of Online Controlled Experiments: Case Studies at Netflix». В: (2016). URL: https://www.kdd.org/kdd2016/papers/files/adp0945-xieA.pdf.

[Bud+18] Roman Budylin и др. «Consistent Transformation of Ratio Metrics for Efficient Online Controlled Experiments». В: (2018). URL: https://www.researchgate.net/publication/322969314_Consistent_Transformation_of_Ratio_Metrics_for_Efficient_Online_Controlled_Experiments#fullTextFileContent.

[Sel] Howard Seltman. «Approximations for Mean and Variance of a Ratio». В: (). URL: https://www.stat.cmu.edu/~hseltman/files/ratio.pdf.

Всё !

Иван Новиков

ivannovikov1303@gmail.com