Возможности замещения метрик отношений при проведении A/B тестирования

Ю. В. Хицкова

Аннотация— **Необходимость тестирований заключается в проверке корректности работы продукта на малом количестве данных в период его реализации, чтобы избежать ошибок при последующем использовании. Основными видами тестирования информационных ресурсов являются юзабилити и А/Б тестирование.**

Сходство А/Б-тестов и Юзабилити тестирования состоит в том, что:

– Оба метода направлены на улучшение пользовательского опыта и эффективности продукта.

– Используются для оптимизации интерфейсов и контента продуктов на основе реальных данных.

– Позволяют определить проблемные области и выявить точки роста.

Различия А/В-тестов и Юзабилити тестирования состоит в том, что:

– Юзабилити тестирование фокусируется на оценке удобства использования продукта, в то время как А/В тесты сравнивают разные версии продукта для определения эффективности.

– В юзабилити тестировании пользователи выполняют задачи и исследователи наблюдают за их действиями, в то время как в А/Б тестах сравниваются результаты использования разных версий продукта.

– Юзабилити тестирование относится к категории «качественных», когда А/Б-тесты в свою очередь - к «количественным».

Ключевые слова— **А/Б тестирование, анализ данных, python, библиотеки python для анализа данных, алгоритм А/Б тестирования.**

I. Введение

АБ-тестирование (A/B testing) – это статистический метод исследования, который позволяет сравнивать два или более версии IT продукта (веб-страницы, приложения или иного продукта), чтобы определить, какая из них эффективнее влияет на поведение пользователей [1-10].

Актуальность А/Б – тестирования в сфере IT обусловлена необходимостью оптимизации пользовательского опыта и увеличения конверсии.

Проведение А/Б-тестов позволяет компаниям, аналитикам и разработчикам оптимизировать интерфейсы, контент и функциональность продуктов на основе реальных данных о поведении пользователей.

Кроме того, А/Б-тестирование позволяет снизить риски при внедрении изменений в продукт, так как позволяет оценить их эффективность на небольшой группе пользователей перед их массовым запуском. Таким образом, А/Б-тестирование является важным инструментом для оптимизации продуктов в сфере IT, позволяя компаниям принимать обоснованные решения на основе данных и повышать эффективность своих продуктов.

Тестирование должно удовлетворять требованиям.

1. Проверка точности данных: данные должны отображать реальные значения и соответствовать ожидаемым результатам (метрикам), группы должны быть распределены случайно и в каждой представлены одни и те же сегменты пользователей.
2. Проверка соответствия требованиям: данные должны соответствовать требованиям заказчика или проекта, выводы, сделанные в результате их анализа, должны быть полезны для бизнеса.

Необходимость тестирований заключается в проверке корректности работы продукта на малом количестве данных в период его реализации, чтобы избежать ошибок при последующем использовании новых характеристик приложения на всех пользователях.

В работе реализован стандартный алгоритм проведения А/Б-теста с целью выяснения влияния изменения приложения на пользователей, на основе полученных данных [11-12].

II. Виды тестирования

Существует несколько видов исследования и тестирования продуктов, все они заключатся в проверке и оценке работоспособности приложения с точки зрения взаимодействия пользователя с ним. В данном контексте необходимо рассмотреть два из них и сделать выбор тестирований с помощью которого будет реализована работа. Это юзабилити тестирование и А/Б-тесты [12]. В данной работе применено А/Б тестирование.

III. Этапы AB тестирования

Как у любого масштабного метода исследования у АВ-тестов есть несколько этапов, а именно:

1. Определение цели тестирования.
2. Разработка вариантов приложения.
3. Разделение трафика (пользователей) на группы.
4. Проведение тестирования.
5. Анализ результатов.
6. Принятие решения.

Рассмотрим подробнее некоторые этапы.

Перед началом исследования необходимо определить период проведения (время) и количество ожидаемых пользователей (или других данных). АБ- тест проводится на ограниченном количестве пользователей, а не на всей аудитории продукта сразу.

Определение цели тестирования является одним из важнейших этапов. На нем определяются цели и формулируются гипотезы тестирования. Заранее четко определяются метрики, которые будут измеряться в ходе теста. Определяется время, выделенное на тест.

Разработка вариантов включает необходимость дизайна и определения образа каждого из вариантов. Помимо АБ-тестов существуют АВС-тесты и так далее, но они малоприменимы в промышленной разработке из-за больших трат ресурсов, денег и времени. Создаются различные варианты (А и Б) продукта, которые будут тестироваться. Каждая из версий содержит определенные изменения, предположительное влияющие на пользователя. Изменения могут быть любыми от изменения цвета веб-страницы, до внедрения нового алгоритма категоризации.

Определяются метрики продукта – это количественные показатели, которые используются для измерения и оценки различных аспектов продукта. Они позволяют количественно оценить эффективность, результативность, качество и т.д.

Гипотезы (основная и альтернативная) – противоположные друг другу.

Например, H0 (основная гипотеза) и H1 (альтернативная гипотеза):

H0 – изменения не повлияют на продукт

H1 – изменения повлияют на продукт

В конце исследования одна из гипотез опровергается, а другая принимается. На основе чего делается вывод о необходимости внедрения изменений в продукт.

Разделение трафика: определяется и переопределяется по ходу тестирования. Трафик на сайт или продукт разделяется между тестовой и контрольной группами, чтобы пользователи случайным образом попадали на одну из версий. Это позволяет провести сравнение результатов между группами непредвзято, так как попадание пользователя в группу определяется случайно.

Проведение тестирования*:* основная фаза АБ-тестирования. На ней пользователи взаимодействуют с разными версиями продукта, в то время как аналитики и разработчики собирают данные о поведении пользователей, их действиях и метриках на основе взаимодействия.

Анализ результатов: наиболее важный этап при подведении итогов тестирования. После окончания выделенного времени и последовательного завершения тестирования оценивается эффективность каждой из версий по ранее заданным метрикам. Определяется какая версия показала наилучшие результаты и была наиболее эффективной.

На основе анализа результатов тестирования принимается решение, которое влияет на дальнейшее развитие продукта. Решение принимается с помощью принятия/опровержения гипотез о том, какая версия будет использоваться в дальнейшем (при необходимости проводятся дополнительные исследования). Именно она будет внедряться в продукт и оптимизироваться.

После выбора оптимальной версии продукта происходит масштабирование изменений на всю аудиторию или продукт.

В данной работе рассмотрен этап – анализ результатов.

Для принятия решения по итогам тестирования необходимо количественно определить результаты на основе полученных данных. Для наиболее удобного и стандартизированного подсчета используются метрики.

Существует различное множество метрик, некоторые универсальные, другие наоборот характерны для определенных сфер. Наиболее популярны IT сфере: конверсия, retention (возвращаемость), ARPPU, ARPU, DAU, WAU, MAU – это метрики активности пользователей или покупателей, их возвращаемости и т.д., используемые в аналитике веб-сайтов и мобильных приложений. Также распространены метрики среднего чека, количества заказов. Некоторые из них мы будем использовать в нашем примере.

Определение метрики для подсчета результатов одна из основных задач А/Б тестирования.

Статистические методы для оценки результатов

После подсчета метрик необходимо сравнить их значения в обеих группах:

1 тестовая группа – группа, взаимодействующая с версией продукта без изменений

0 контрольная группа – группа, взаимодействующая с измененной версией продукта

Следующим шагом необходимо выяснить значимы ли различия в значениях метрик. Для этого используются статистические методы такие как ANOVA, Т-тест, Хи-квадрат и подобные. Необходимо определить какой именно метод наиболее применим к полученным данным, для этого существует несколько алгоритмов.

IV. Особенности А/Б тестирования приложения доставки

Данные и условие взяты нами с ресурса karpov.courses [13]. Необходимо проанализировать результаты А/Б теста для проверки системы рекомендаций в приложении доставки (с заранее полученными данными).

Данные были выгружены с karpov.courses в формате csv. Обработка данных будет производиться в среде Anaconda на языке программирования Python.

Условие задачи: в приложение доставки был внедрен новый алгоритм рекомендаций продуктов, необходимо выяснить как изменение повлияло на взаимодействие пользователей с приложением на основе проведения АВ-теста.

Средства реализации.

В качестве среды разработки был взят Jupyter notebook и библиотеки для анализа данных:

- Pandas – используется для чтения и преобразования

данных из различных форматов, а так же для таких манипуляций как группировка, сортировка, фильтрация и объединение;

- Numpy – предназначена для работы с многомерными массивами данных и выполнения математических операций над ними, хорошо интегрируется с другими библиотеками, создавая мощные инструменты для работы с большим количеством данных;

- Scipy.stats – включает различные статистические тесты для проверки гипотез, такие как t-тесты, используется для вычисления статистических характеристик (среднее, медиана, дисперсия и прочие);

- Plotly.express – предназначена для визуализации данных;

- Matplotlib.pyplot – расширяет возможности предыдущей библиотеки и позволяет добавлять подписи, наименование осей и метки на графики;

- Seaborn – предоставляет возможности для визуализации статистических связей между переменными, включая корреляции, регрессионные анализы и другие статистические методы.

С помощью использования перечисленных библиотек и модулей был проведен анализ результатов А/Б теста.

Для анализа результатов АВ-теста был использован следующий алгоритм [18-24]:

1. Подключение библиотек и выгрузка файлов
2. Предобработка данных.
3. Подбор метрик.
4. Выдвижение гипотез.
5. Расчет метрики для каждой из групп.
6. Сравнение полученных значений метрики.
7. Проведение теста, выявление статистической значимости результатов (по p-value)
8. Вывод о влиянии изменений.

Рассмотрим некоторые шаги подробнее:

1. Пользователи разбиты на две группы: контролную тестовую. Так как данные были загружены из нескольких файлов, необходимо объединить таблицы по общим столбцам.

```
user_order = users.merge(orders, how = 'inner', on
='order_id')
```

2. Далее проведен предварительный анализ данных. Определение и переопределение типов – для очищения данных и корректной работы библиотечных функций.

```
user_order = user_order.astype({'product_id': 'int64'})
user_order.dtypes
```

```
user_id          int64
order_id         int64
action          object
time            object
date            object
group            int64
creation_time   object
product_ids     object
product_id       int64
dtype: object
```

Было обнаружено, что в дата фрейме существуют отменённые заказы, то есть какие-то заказы - create_order, стали в итоге cancel_order. С помощью метода nunique() было обнаружено число уникальных пользователей (user_order_prod.user_id.nunique()), которые делали заказы и число уникальных заказов (user_order_prod.order_id.nunique()). На каждого пользователя приходится в среднем более 4 заказов.

3. Мы можем посмотреть на гистограмму распределения суммы заказов по группами и гистограмму плотности распределения (см. рисунок 1 и рисунок 2).

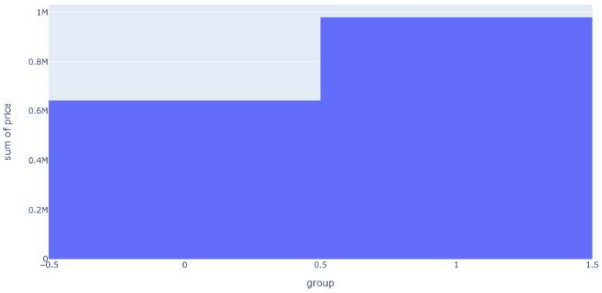

Рис. 1: Гистограмма распределение суммы заказов по группам

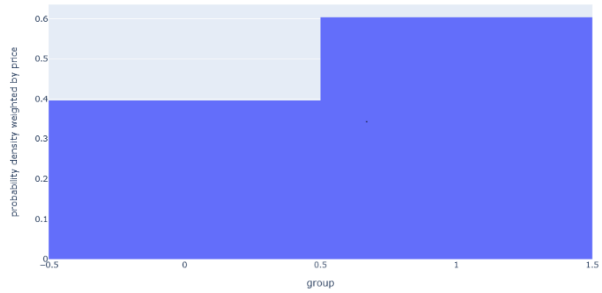

Рис. 2: Гистограмма плотности распределение суммы заказов по группам

Из графиков видно, что общая выручка в тестовой группе выше, чем в контрольной. Посмотрим на медиану, квартили и выбросы нашего распределения с помощью графика боксплот (см. рисунок 3).

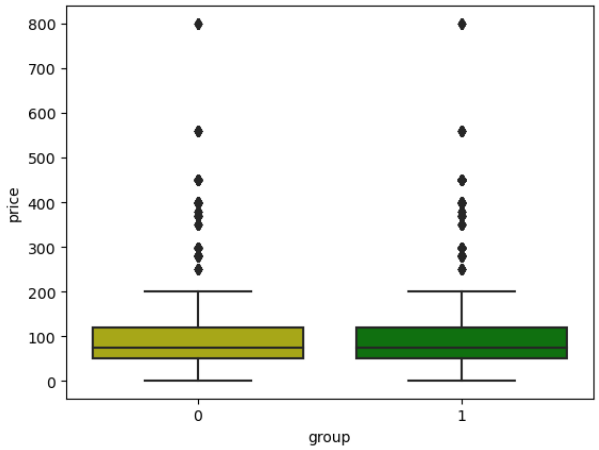

Рис. 3: Медиана, квартили и выбросы суммы заказов, в том числе отмененных

Из рисунка видно, что значимых отличий нет. Далее разделим данные по группам на тестовую и

контрольную: user_order_prod_cont = user_order_prod.query('group == 0'), также поступим для контрольной группы.

4. Подбор метрик – важный этап AB-тестирования. В начале необходимо из ряда существующих метрик выбрать необходимые таким образом, чтобы в конце периода тестирования по ним можно было сделать корректные выводы об изменении приложения. Для рассматриваемого случая важными является средний чек - AVO, сумма заказа, средняя выручка от одного покупателя – это по сути ARPPU. Также можно посмотреть среднее количество заказов на 1 покупателя. Соответственно у нас будет три гипотезы.

Гипотеза 1. H0 – метрика средний чек не изменилась статистически значимо. Остается старый вариант приложения.

H1 – метрика средний чек выросла статистически значимо, система рекомендаций меняется на новую.

Посчитаем, как и насколько изменился средний чек.

Развернем таблицу, чтобы убрать из нее отмененные заказы.: .pivot(index = 'order_id', values = 'price', columns = 'action').fillna(0)) для контрольной и тестовой групп.

Рассчитаем средний чек (AVO). Для контрольной группы он будет равен 115,0, для тестовой – 110. Медиана суммы чека равна – 115,03 и 90,33 руб. соответственно. Мода контрольной группы - 60, тестовой – 50. Мы видим, что все показатели, касающиеся средней суммы заказа изменились в худшую сторону. Но так как AVO является ratio-метрикой, мы должны либо привести ее к независимым параметрам с помощью одного из известных методов: линеаризация, cuped, бакетный метод, либо использовать другие метрики – средние для одного пользователя или же конверсии пользователей, не забывая, что они могу изменяться не сонаправлено нашей метрики [26]. Ratio-метрика — это метрика отношения не пользовательского уровня (non-user level metric) с зависимыми наблюдениям, но которая явным образом выражается через отношение сумм соответствующих пользовательских сигналов [26]. Для бакетного метод у нас недостаточно наблюдений, их должно быть не менее 1000 в каждой группе. Посмотрим на распределение данных нашей метрики: рисунок 4 – распределение в контрольной группе, рисунок 5 – распределение в тестовой группе.

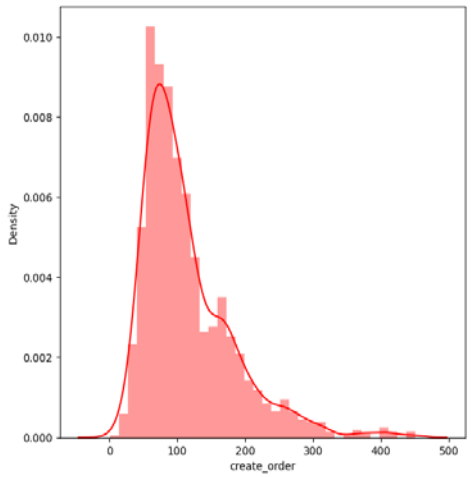

Рис. 4: Распределение суммы чека в контрольной группе

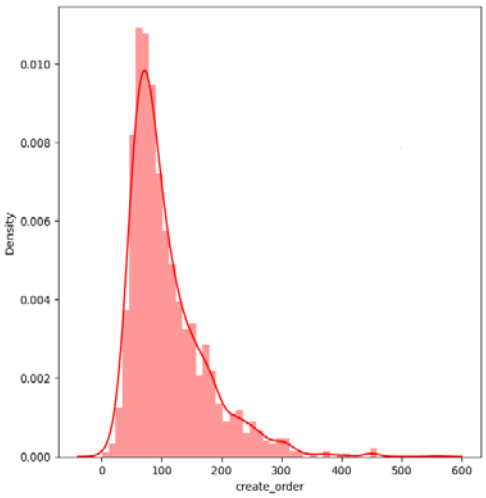

Рис. 5: Распределение суммы чека в тестовой группе

Для сравнения построим боксплот (см. рисунок 6). Видно, что и в данном случае значительных различий нет.

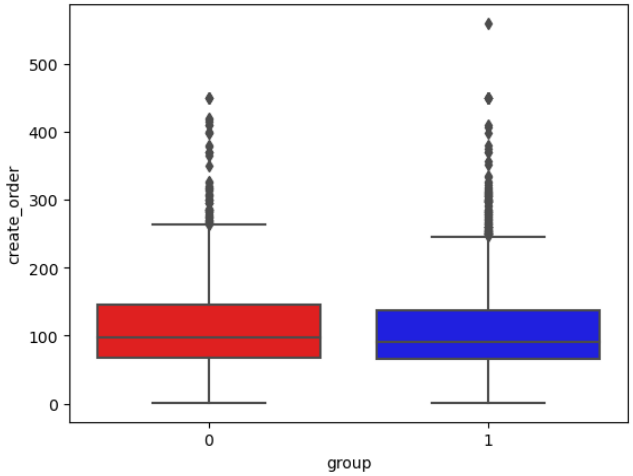

Рис. 6: Медиана, квартили и выбросы среднего чека только выполненных заказов

Но необходимо провести статистический тест. Для анализа статистической значимости ratio метрики нельзя использовать t-test, так как он предполагает независимость данных, а у нас зависимые наблюдения.

Для зависимых наблюдений с высокой корреляцией между собой мы можем использовать бутстрап. К тому же он позволяет сравнить нам не только средние значения, но и медиану, в нашем случае нам необходимо среднее значение. Результаты bootstrap в контрольной группе представлены на рисунке 7, в тестовой группе на рисунке 8.

```
]: boot_strap_contr_m = stats.bootstrap((user_order_prod_cont_p_n_canc.create_order, ), np.mean, n_resamples=15000)
   boot_strap_contr_m

]: BootstrapResult(confidence_interval=ConfidenceInterval(low=111.72835216624541, high=118.46194393009661), bootstrap_distrib
   ution=array([118.1150488 , 116.59908878, 113.95555672, ..., 112.94614261,
          115.4001054 , 119.25651526]), standard_error=1.7129815201807175)
```

Рис. 7: Результаты статистического теста в контрольной группе

```
i]: boot_strap_test_m = stats.bootstrap((user_order_prod_cont_test_p_n_canc.create_order, ), np.mean, n_resamples=15000)
   boot_strap_test_m

i]: BootstrapResult(confidence_interval=ConfidenceInterval(low=107.48820649332413, high=112.61636865137025), bootstrap_distrib
   ution=array([110.15881587, 109.54093414, 108.66820799, ..., 110.89089829,
          110.91448503, 110.7997662 ]), standard_error=1.307994216294764)
```

Рис. 8: Результаты статистического теста в тестовой группе

Совпадение интервала по средней примерно – 1/8. На этом основании нельзя сказать, что это статистически значимые изменения, так как вероятность непопадания в доверительный интервал составляет примерно более 80 процентов. То есть сокращение среднего чека произошло не статистически значимо.

Далее смотрим метрики пользователей. Пользовательских конверсий в данных нет, поэтому будем использовать средние метрики пользователя. В данном случае рассмотрим количество заказов на пользователя и выручку на пользователя.

Гипотеза 2. H0 – (ARPPU) средняя выручка с одного покупателя изменилась не значимо статистически. Остается старая система рекомендаций.

H1 – ARPPU выросла статистически значимо, будет внедрена новая система рекомендаций.

Средняя выручка на пользователя в тестовой группе выросла на 617,4. Здесь для сравнения средних можно использовать t -тест.

ss.stats.ttest_ind(user_order_cont_revenue['revenue'], user_order_test_revenue['revenue']). P-value = .226663465852294e-28

Для увеличения точности результат прологорифмируем данные и сравним уже сглаженный ряд.

ss.stats.ttest_ind(user_order_cont_revenue_log, user_order_test_revenue_log). P-value = 6.108540480706531e-39

Графики простого и прологарифмированного распределений говорят о том, что выбросы были незначительные и заметного сглаживания данных не произошло, пример для контрольной группы (рисунки 9-10).

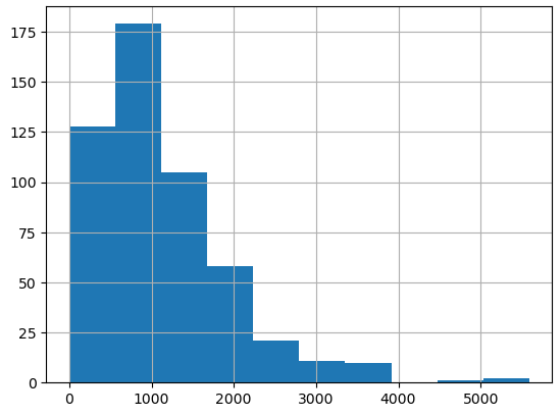

Рис. 9: Доход от пользователей

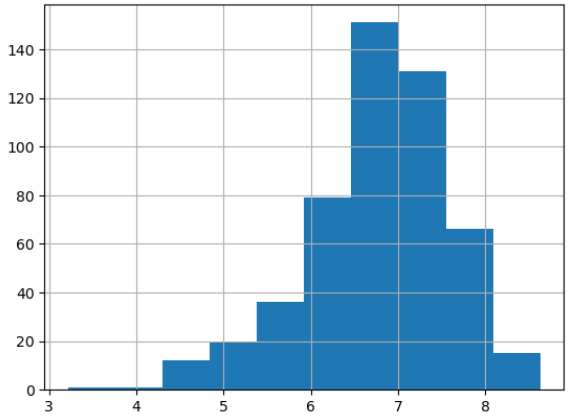

Рис. 10: Прологарифмированный доход от пользователей

В результате получаем, что p-value бесконечно малая величина, как в первом, так и во втором случае. Следовательно, изменения привели к статистически значимому значительному увеличению ARPPU. Значит необходимо внедрять новую систему рекомендаций.

Гипотеза 3. H0 – количество заказов на одного пользователя не изменилось статистически значимо, остаётся старый вариант рекомендаций.

H1 – количество заказов увеличилось статистически значимо, будет внедрен новый вариант рекомендаций.

Количество заказов на пользователя в тестовой групп выросло на 6,7. С помощью t – теста выявили, что изменение статистически значимо (p-value – бесконечно малая величина), соответственно будет внедрена новая система рекомендаций.

5. Вывод о влиянии изменений – на основании прошлых шагов формулируется вывод о том, что нулевую гипотезу можно отвергать, следовательно приложение улучшилось и его необходимо распространять на всех пользователей.

V. Заключение

Проведение A/B теста с p-value < 0.05 означает, что статистически значимые различия между группами обнаружены на выбранном уровне значимости. В данном случае, нулевая гипотеза о равенстве средних значений двух групп была отвергнута, что может

указывать на присутствие статистически значимых различий между вариантами A и B.

Однако, важно помнить, что есть еще метрика отношений, которую можно изучить с помощью других методов, указанных выше. В сфере IT A/B тесты играют ключевую роль при оптимизации пользовательского опыта, улучшении продуктов и услуг, а также принятии обоснованных решений на основе данных. A/B тестирование позволяет компаниям проверять гипотезы, оценивать влияние изменений на поведение пользователей, улучшать функционал и дизайн продукта.

Библиография

[1] Козырева, Н. Е., & Рахманова, А. Ю. (2021). A/B тестирование как инструмент оценки взаимодействия бренда с потребителями в диджитал среде // Экономика и бизнес: тенденции и инновации. С. 295-303.

[2] Высоцкая, А. И., Комарчева, А. Р., & Чжен, А. А. A/B Тестирование в Digital // ЕО IPSO. 2024, № 5, С. 75.

[3] Бобко, Д. В., & Шинкевич, К. А. (2020). Маркетинговые исследования на основе A/B тестирования в цифровых компаниях.

[4] Жуковский, В. А. (2019). Повышение эффективности организации посредством разработки фреймворка автоматизированного A/B тестирования // Международный академический вестник. № 10, С. 88-91.

[5] Тюринова, В. А., & Мауритс, В. Г. Методы A/B-тестирования информационных ресурсов. Финансовая система в национальной экономике: предпосыл, 175.

[6] Бычков, И. В., & Дедкова, С. Н. (2013). Централизованное тестирование по математике: нестандартный способ решения уравнений, содержащих переменную под знаком модуля.

[7] Союнов, Х. Т. (2019). Интернет-маркетинг: стратегии, инструменты и тренды // Качество управленческих кадров и экономическая безопасность организации. С. 110-113.

[8] Бажан, З. И. (2016). Тестирование как один из эффективных способов проверки теоретической и методической подготовки обучающихся в вузе. Проблемы современного педагогического образования, (51-2), 34-40.

[9] Claeys, E., Gancarski, P., Maumy-Bertrand, M., Wassner, H.: Dynamic allocation optimization in A/B-tests using classification-based preprocessing // IEEE TKDE. 2021, 35(1), P. 335-349.

[10] Fabijan, A., Dmitriev, P., Arai, B., Drake, A., Kohlmeier, S., Kwong, A.: A/B integrations: 7 lessons learned from enabling A/B testing (2023)

[11] Kaufmann, E., Cappé, O., Garivier, A.: On the complexity of A/B testing. ArXiv e-prints, May 2014

[12] Астахова И. Ф., Маковий К. А., Хицкова Ю. В. Система интеллектуализации юзабилити-тестирования информационных ресурсов // Актуальные проблемы прикладной математики, информатики и механики. 2022. С. 1719-1726.

[13] Карпов курс: сайт. URL: https://karpov.courses/ (дата обращения 10.09.2024)

[14] Kaluza, B., Mirchevska, V., Dovgan, E., Lustrek, M., Gams, M.: UCI machine learning repository, an agent-based approach to care in independent living (2010)

[15] Grushka-Cockayne, Yael, et al. "A/B Testing at Vungle." Darden Business Publishing Cases (2015): 1-7.

[16] Gui, H., Xu, Y., Bhasin, A., & Han, J. (2015, May). Network a/b testing: From sampling to estimation. In: Proceedings of the 24th International Conference on World Wide Web. P. 399-409.

[17] Siroker, D., & Koomen, P. (2015). A/B testing: The most powerful way to turn clicks into customers. John Wiley & Sons.

[18] King, R., Churchill, E. F., & Tan, C. (2017). Designing with data: Improving the user experience with A/B testing. O'Reilly Media, Inc.

[19] Nguyen, H. Q. (2001). Testing applications on the Web: Test planning for Internet-based systems. John Wiley & Sons.

[20] Johari, R., Koomen, P., Pekelis, L., & Walsh, D. (2017, August). Peeking at a/b tests: Why it matters, and what to do about it. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. P. 1517-1525.

[21] Quin, F., Weyns, D., Galster, M., & Silva, C. C. (2024). A/B testing: A systematic literature review // Journal of Systems and Software, 112011.

[22] Deng, A., & Shi, X. (2016, August). Data-driven metric development for online controlled experiments: Seven lessons learned. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. P. 77-86.

[23] Gui, H., Xu, Y., Bhasin, A., & Han, J. (2015, May). Network a/b testing: From sampling to estimation. In Proceedings of the 24th International Conference on World Wide Web. P. 399-409.

[24] Fabijan, A., Dmitriev, P., McFarland, C., Vermeer, L., Holmström Olsson, H., & Bosch, J. (2018). Experimentation growth: Evolving trustworthy A/B testing capabilities in online software companies // Journal of Software: Evolution and Process. 30(12), e2113.

[25] Kohavi, R., & Longbotham, R. (2015). Online controlled experiments and A/B tests // Encyclopedia of machine learning and data mining, 1-11.

[26] Мосин П. «Линеаризация: зачем и как укрощать ratio-метрики в A/B-тестах». URL: https://habr.com/ru/companies/kuper/articles/768826/, (дата обращения 15.09.2024)

Статья получена 17 ноября 2024.

Ю.В. Хицкова – доцент кафедры систем управления и информационных технологий в строительстве факультета информационных технологий и компьютерной безопасности, кандидат экономических наук, доцент, Воронежский государственный технический университет (email: prosvetovau@list.ru)

Possibilities of Replacing Ration Metrics When Conducting A/B Testing

Julia V. Khitskova

***Abstract*– The need for testing is to check the correctness of the product's operation on a small amount of data during its implementation in order to avoid errors during subsequent use. The main types of testing of information resources are usability and A/B testing.**

The similarity between A/B tests and usability testing is that:

– Both methods are aimed at improving the user experience and efficiency of the product.

– They are used to optimize the interfaces and content of products based on real data.

– They allow you to identify problem areas and identify growth points.

The differences between A/B tests and usability testing are that:

– Usability testing focuses on assessing the ease of use of a product, while A/B tests compare different versions of a product to determine effectiveness.

– In usability testing, users perform tasks and researchers observe their actions, while A/B tests compare the results of using different versions of a product.

– Usability testing falls into the category of "qualitative", while A/B testing, in turn, falls into the category of "quantitative".

***Keywords*– A/B testing, data analysis, Python, Python libraries for data analysis, A/B testing algorithm.**

REFERENCES

[1] Kozyreva N. E., Rahmanova A. Yu. (2021) A/B testirovanie kak instrument ocenki vzaimodejstviya Brenda s potrebitelyami v didzhital srede. *Ekonomika i biznes tendencii i innovacii.* P. 295-303.

[2] Vysockaya A.I. Komarcheva A.R., Chzhen A.A. A/B testirovanie v digital // EO IPSO. 2024, No.5, P. 75.

[3] Bobko D.V., Shinkevich K.A. 2020 marketingovye issledovaniya na osnove a v testirovaniya v cifrovyh kompaniyah.

[4] Zhukovskij V.A. 2019 Povyshenie effektivnosti organizacii posredstvom razrabotki frejmvorka avtomatizirovannogo A/B testirovaniya. *Mezhdunarodnyj akademicheskij vestnik.* 10-88-91

[5] Tyurinova V.A., Maurits V.G. Metody A/B testirovaniya informacionnyh resursov i58 finansovaya sistema v nacionalnoje konomike predposyl 175.

[6] Bychkov I.V., Dedkova S.N. 2013 centralizovannoe testirovanie po-matematike nestandartnyj sposob resheniyauravnenij soderzhaschih-peremennuyu pod znakom modulya.

[7] Soyunov H.T. Internet-marketing strategii instrument I trendy in kachestvo-upravlencheskih kadrov I ekonomicheskaya bezopasnost organizacii. 2019. p-110-113.

[8] Bazhan Z.I. 2016 testirovanie kak odin iz effektivnyh sposobov proverki teoreticheskoj I metodicheskoj podgotovki obuchayuschihsya v vuze problem sovremennogo pedagogicheskogo obrazovaniya. 51-2-34-40.

[9] Claeys, E., Gancarski, P., Maumy-Bertrand, M., Wassner, H.: Dynamic allocation optimization in A/B-tests using classification-based preprocessing. IEEE TKDE **35**(1), 335–349 (2021)

[10] Fabijan, A., Dmitriev, P., Arai, B., Drake, A., Kohlmeier, S., Kwong, A.: A/B integrations: 7 lessons learned from enabling A/B testing (2023)

[11] Kaufmann, E., Cappé, O., Garivier, A.: On the complexity of A/B testing. ArXiv e-prints, May 2014?

[12] Astakhova I. F., Makoviy K. A., Khitskova Yu. V. Intellectualization system for usability testing of information resources // Actual problems of applied mathematics, computer science and mechanics. 2022. P. 1719-1726.

[13] Karpov Courses: website. URL: https://karpov.courses/ (date of access 10.09.2024)

[14] Kaluza, B., Mirchevska, V., Dovgan, E., Lustrek, M., Gams, M.: UCI machine learning repository, an agent-based approach to care in independent living (2010) /

[15] Grushka-Cockayne, Yael, et al. "A/B Testing at Vungle." Darden Business Publishing Cases (2015): 1-7.

[16] Gui, H., Xu, Y., Bhasin, A., & Han, J. (2015, May). Network a/b testing: From sampling to estimation. In *Proceedings of the 24th International Conference on World Wide Web* (pp. 399-409).

[17] Siroker, D., & Koomen, P. (2015). *A/B testing: The most powerful way to turn clicks into customers*. John Wiley & Sons.

[18] King, R., Churchill, E. F., & Tan, C. (2017). *Designing with data: Improving the user experience with A/B testing*. " O'Reilly Media, Inc.".

[19] Nguyen, H. Q. (2001). *Testing applications on the Web: Test planning for Internet-based systems*. John Wiley & Sons.

[20] Johari, R., Koomen, P., Pekelis, L., & Walsh, D. (2017, August). Peeking at a/b tests: Why it matters, and what to do about it. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1517-1525).

[21] Quin, F., Weyns, D., Galster, M., & Silva, C. C. (2024). A/B testing: A systematic literature review. *Journal of Systems and Software*, 112011.

[22] Deng, A., & Shi, X. (2016, August). Data-driven metric development for online controlled experiments: Seven lessons learned. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 77-86).

[23] Gui, H., Xu, Y., Bhasin, A., & Han, J. (2015, May). Network a/b testing: From sampling to estimation. In Proceedings of the 24th International Conference on World Wide Web. P. 399-409

[24] Fabijan, A., Dmitriev, P., McFarland, C., Vermeer, L., Holmström Olsson, H., & Bosch, J. (2018). Experimentation growth: Evolving trustworthy A/B testing capabilities in online software companies. *Journal of Software: Evolution and Process*, *30*(12), e2113.

[25] Kohavi, R., & Longbotham, R. (2015). Online controlled experiments and A/B tests. *Encyclopedia of machine learning and data mining*, 1-11.

[26] Mosin P. "Linearization: why and how to tame ratio metrics in A/B tests." URL: https://habr.com/ru/companies/kuper/articles/768826/, (date accessed 09/15/2024)