Математическая статистика в R.
Проект 1. Анализ A/B-теста для интернет-магазина

Макаров П. А.

5 февраля 2026 г.

Аннотация

Данный проект представляет собой пошаговое руководство по проведению A/B-теста в среде R. Проект предназначен для начального погружения в язык R и знакомства с основами математической статистики. Рассматривается полный цикл анализа: от генерации данных и формулировки гипотез до проверки статистической значимости и визуализации результатов. Все этапы сопровождаются подробными пояснениями и кодом на R.

Содержание

1 Постановка задачи **2**
- 1.1 Проблема . . . 2
- 1.2 Формальная задача . . . 2
- 1.3 Цели проекта . . . 2

2 Генерация данных средствами R **2**
- 2.1 Структура данных . . . 2
- 2.2 Модель генерации данных . . . 2
- 2.3 Код для генерации данных . . . 3

3 Начальное исследование и предобработка данных **3**
- 3.1 Изучение структуры данных . . . 3
- 3.2 Визуализация распределения . . . 3
- 3.3 Расчёт основных метрик . . . 4

4 Статистический анализ **5**
- 4.1 Формулировка гипотез . . . 5
- 4.2 Предварительная проверка условий для параметрических тестов . . . 5
- 4.3 Выбор и применение непараметрического теста . . . 5
- 4.4 Дополнительный анализ: бутстреп . . . 6

5 Представление итоговых результатов **6**
- 5.1 Интерпретация статистического вывода . . . 6
- 5.2 Финальная визуализация . . . 6
- 5.3 Подготовка отчета . . . 7

1 Постановка задачи

1.1 Проблема

Интернет-магазин стремится повысить средний чек покупателей. Дизайнеры предложили изменить цвет кнопки «Купить» с синего (текущий вариант, **группа A**) на красный (новый вариант, **группа B**). Маркетологи предполагают, что красная кнопка будет более заметной и побудит пользователей совершать покупки на большие суммы.

1.2 Формальная задача

Провести A/B-тест для статистической оценки гипотезы о том, что изменение цвета кнопки приводит к **статистически значимому увеличению среднего чека**.

1.3 Цели проекта

1. Сгенерировать синтетические данные, имитирующие поведение пользователей в двух группах.
2. Провести разведочный анализ данных (EDA — Exploratory Data Analysis).
3. Сформулировать статистические гипотезы.
4. Выбрать и применить подходящий статистический тест.
5. Интерпретировать результаты и сделать выводы для бизнеса.
6. Визуализировать ключевые результаты.

2 Генерация данных средствами R

Для учебных целей мы сгенерируем данные самостоятельно. Это позволит контролировать процесс и понимать структуру данных.

2.1 Структура данных

Будет создан датафрейм (таблица) со следующими столбцами:

- `user_id` (строка) — уникальный идентификатор пользователя.
- `group` (фактор) — принадлежность к группе теста: «A» (синяя кнопка) или «B» (красная кнопка).
- `purchase_amount` (числовой) — сумма покупки в условных единицах. Может быть равна 0, если покупка не совершена.

2.2 Модель генерации данных

Мы смоделируем данные так, чтобы в группе B был небольшой положительный эффект.

$$\text{Группа A (контрольная)} : X_A \sim \mathcal{N}(\mu = 50, \sigma = 15)$$
$$\text{Группа B (тестовая)} : X_B \sim \mathcal{N}(\mu = 53, \sigma = 15)$$

Кроме того, для имитации реального трафика около 30% пользователей в каждой группе не совершат покупку (`purchase_amount = 0`).

2.3 Код для генерации данных

```r
# Устанавливаем "зерно" для воспроизводимости результатов
set.seed(42)

# Количество пользователей в каждой группе
n_users <- 500

# Генерация данных для группы A
group_a <- data.frame(
  user_id = paste0("A", 1:n_users),
  group = "A",
  purchase_amount = round(rnorm(n_users, mean = 50, sd = 15), 2)
)

# Генерация данных для группы B (с небольшим положительным эффектом)
group_b <- data.frame(
  user_id = paste0("B", 1:n_users),
  group = "B",
  purchase_amount = round(rnorm(n_users, mean = 53, sd = 15), 2)
)

# Объединение данных
ab_test_data <- rbind(group_a, group_b)

# Внесение "шума": 30% пользователей не совершают покупку
zero_indices_a <- sample(1:n_users, size = 0.3 * n_users)
zero_indices_b <- sample((n_users+1):(2*n_users), size = 0.3 * n_users)
ab_test_data$purchase_amount[c(zero_indices_a, zero_indices_b)] <- 0

# Просмотр первых строк
head(ab_test_data)
```

3 Начальное исследование и предобработка данных

Прежде чем проводить анализ, необходимо изучить данные, проверить их качество и рассчитать основные метрики.

3.1 Изучение структуры данных

```r
# 1. Структура данных и основные статистики
str(ab_test_data)
summary(ab_test_data)

# 2. Проверка баланса групп
table(ab_test_data$group)
```

3.2 Визуализация распределения

```r
# Боксплот для визуального сравнения распределений
boxplot(purchase_amount ~ group, data = ab_test_data,
        main = "Распределение суммы покупки по группам",
        xlab = "Группа (A = Синяя, B = Красная)",
        ylab = "Сумма покупки",
        col = c("lightblue", "salmon"))
```

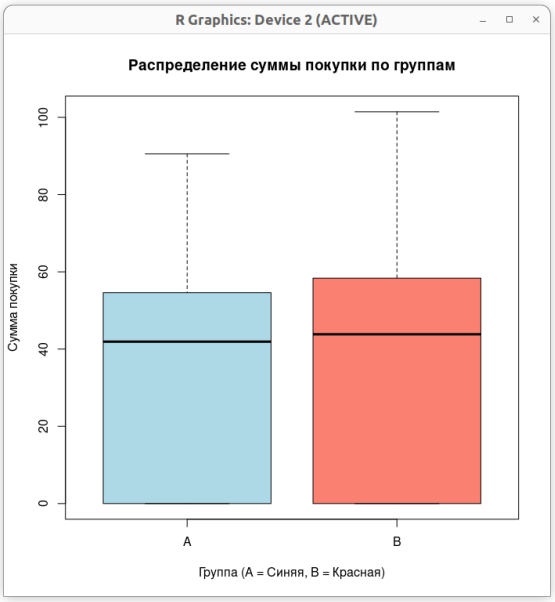

Рис. 1: Пример боксплота для сравнения распределений сумм покупок

3.3 Расчёт основных метрик

Для получения точных цифр используем пакет `dplyr`.

```r
# Установка и загрузка пакета (если не установлен)
# install.packages("dplyr")
library(dplyr)

metrics <- ab_test_data %>%
  group_by(group) %>%
  summarise(
    n = n(),
    mean_purchase = mean(purchase_amount),
    median_purchase = median(purchase_amount),
    std_dev = sd(purchase_amount),
    conversion_rate = sum(purchase_amount > 0) / n * 100
  )
print(metrics)
```

Результат можно оформить в виде таблицы.

Таблица 1: Основные метрики по группам

Группа	N	Средний чек	Медиана	Конверсия (%)
A (синяя)	500	34.7	41.9	70.0
B (красная)	500	36.9	43.8	70.0

4 Статистический анализ

4.1 Формулировка гипотез

Формализуем нулевую и альтернативную гипотезы:

$H_0 : \mu_B = \mu_A$ (Средний чек в группе B равен среднему чеку в группе A)
$H_1 : \mu_B > \mu_A$ (Средний чек в группе B **больше** среднего чека в группе A)

Гипотеза H_1 является **односторонней** (нас интересует именно увеличение).

4.2 Предварительная проверка условий для параметрических тестов

Перед применением классического t-теста необходимо проверить допущения о нормальности распределения и равенстве дисперсий.

```r
# Проверка на нормальность (тест Шапиро-Уилка на подвыборке)
shapiro.test(ab_test_data$purchase_amount[ab_test_data$group == "A"][1:100])
shapiro.test(ab_test_data$purchase_amount[ab_test_data$group == "B"][1:100])

# Проверка равенства дисперсий (тест Левена)
# Установка и загрузка пакета lawstat
install.packages("lawstat")
library(lawstat)

# Проведение теста Левена
levene_test_result <- levene.test(
  ab_test_data$purchase_amount,
  ab_test_data$group,
  location = "mean" # или "median" для более устойчивого варианта
)

print(levene_test_result)
```

Вывод: В нашем случае из-за наличия значительного числа нулей (30% данных) распределение, скорее всего, не будет нормальным (p-value теста Шапиро — Уилка < 0.05). Это делает применение параметрического t-теста некорректным.

4.3 Выбор и применение непараметрического теста

Для сравнения двух независимых групп, когда нарушено предположение о нормальности, используется **тест Манна — Уитни (Уилкоксона)**. Это непараметрический аналог t-теста, устойчивый к выбросам.

```r
# Проведение одностороннего теста Манна-Уитни
wilcox_test_result <- wilcox.test(purchase_amount ~ group,
                                  data = ab_test_data,
                                  alternative = "greater", # H1: B > A
                                  conf.int = TRUE)
print(wilcox_test_result)
```

4.4 Дополнительный анализ: бутстреп

Для получения более надежной оценки разницы средних и её доверительного интервала используем метод бутстрепа.

```
# Функция для бутстрепа разницы средних
boot_diff <- function(data, indices) {
  sample_data <- data[indices, ]
  mean_diff <- mean(sample_data$purchase_amount[sample_data$group == "B"]) -
               mean(sample_data$purchase_amount[sample_data$group == "A"])
  return(mean_diff)
}

# Установка и загрузка пакета для бутстрепа
# install.packages("boot")
library(boot)

set.seed(456)
boot_results <- boot(ab_test_data, statistic = boot_diff, R = 1000)
boot_ci <- boot.ci(boot_results, type = "perc")
print(boot_results)
print(boot_ci)
```

5 Представление итоговых результатов

5.1 Интерпретация статистического вывода

1. **Ключевой результат теста:** Смотрим на p-value из `wilcox.test`.
   - Если p-value < 0.5, то **нет оснований отвергать** H_0. Статистически значимого эффекта не обнаружено.
   - Если p-value ≥ 0.5 (уровень значимости α), то мы **отвергаем нулевую гипотезу** H_0. Существует статистически значимое свидетельство в пользу того, что красная кнопка увеличивает средний чек.
2. **Размер эффекта:** Смотрим на поле `estimate` в результате теста (смещенная медиана разностей) и на доверительный интервал `conf.int`. Также используем оценку из бутстрепа: «Средний чек в группе B выше на X единиц (95% Доверительный Интервал: [Y, Z])».

5.2 Финальная визуализация

Для отчета создадим наглядные графики с помощью пакета `ggplot2`.

```
# Установка и загрузка ggplot2
# install.packages("ggplot2")
library(ggplot2)

# График 1: Боксплот с точками (jitter)
p1 <- ggplot(ab_test_data, aes(x = group, y = purchase_amount, fill = group))
  ↪  +
  geom_boxplot(alpha = 0.7, outlier.shape = NA) +
  geom_jitter(width = 0.2, size = 0.7, alpha = 0.3) +
  labs(title = "Сравнение суммы покупки между группами",
       subtitle = "Группа A (синяя кнопка) vs. Группа B (красная кнопка)",
```

```
        x = "Группа теста",
        y = "Сумма покупки") +
   scale_fill_manual(values = c("A" = "lightblue", "B" = "salmon")) +
   theme_minimal()

# График 2: Столбчатая диаграмма средних с доверительными интервалами
p2 <- ggplot(metrics, aes(x = group, y = mean_purchase, fill = group)) +
   geom_bar(stat = "identity", alpha = 0.8) +
   geom_errorbar(aes(ymin = mean_purchase - 1.96 * std_dev/sqrt(n),
                     ymax = mean_purchase + 1.96 * std_dev/sqrt(n)),
                 width = 0.2) +
   labs(title = "Средний чек с 95% доверительными интервалами",
        x = "Группа теста",
        y = "Средняя сумма покупки") +
   scale_fill_manual(values = c("A" = "lightblue", "B" = "salmon")) +
   theme_minimal()

print(p1)
print(p2)
```

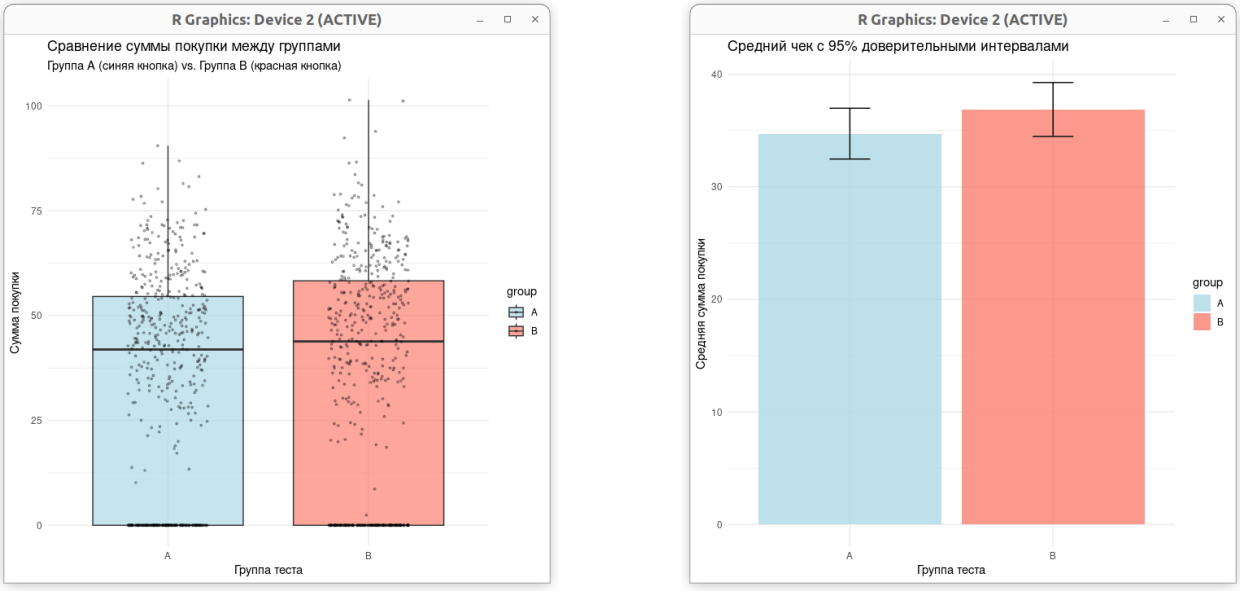

Рис. 2: Финальные графики: распределение данных с боксплотом (слева) и сравнение средних значений с доверительными интервалами (справа)

5.3 Подготовка отчета

Итоговый отчет должен содержать следующие разделы:

1. **Введение:** Описание бизнес-проблемы и цели анализа.
2. **Данные и методы:** Описание процесса генерации данных, структуры данных и примененных статистических методов (тест Манна — Уитни, бутстреп).
3. **Результаты:**
   - Сводная таблица метрик (Таблица 1).
   - p-value и значение статистики теста Манна–Уитни.

- Оценка разницы средних и её 95% доверительный интервал, полученный бутстрепом.
- Ключевые визуализации (Рис. 1, 2).

4. **Выводы и рекомендации:**
   - Статистический вывод (принята/отвергнута H_0).
   - Практическая рекомендация для бизнеса (стоит ли внедрять красную кнопку на всех страницах).
   - Ограничения анализа (например, синтетические данные, упрощенная модель поведения).

Заключение

Данный проект охватывает полный цикл работы с данными в R — от их создания до статистического вывода. Он служит отличной отправной точкой для дальнейшего изучения как языка R (пакеты `dplyr`, `ggplot2`, `boot`), так и более сложных методов статистического анализа. Для углубления знаний можно усложнить модель генерации данных, добавить больше метрик (например, время на сайте) или рассмотреть многовариантное (A/B/n) тестирование.