

МЕЖГОСУДАРСТВЕННОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ
ВЫСШЕГО ОБРАЗОВАНИЯ
«БЕЛОРУССКО-РОССИЙСКИЙ УНИВЕРСИТЕТ»

Кафедра «Маркетинг и менеджмент»

СТАТИСТИЧЕСКИЕ МЕТОДЫ АНАЛИЗА ДАННЫХ

*Методические рекомендации к лабораторным работам
для студентов специальности
6-05-0611-04 «Электронная экономика»
очной и заочной форм обучения*



Могилев 2026

УДК 311
ББК 606
С78

Рекомендовано к изданию
учебно-методическим отделом
Белорусско-Российского университета

Одобрено кафедрой «Маркетинг и менеджмент» «8» мая 2026 г.,
протокол № 11

Составитель канд. физ.-мат. наук, доц. В. А. Ливинская

Рецензент канд. экон. наук Т. В. Романькова

Методические рекомендации предназначены для формирования практических навыков анализа данных в цифровой экономике с помощью современных инструментов анализа данных.

Учебное издание

СТАТИСТИЧЕСКИЕ МЕТОДЫ АНАЛИЗА ДАННЫХ

Ответственный за выпуск	А. В. Александров
Корректор	А. А. Подошевка
Компьютерная верстка	Н. П. Полевничая

Подписано в печать . Формат 60×84/16. Бумага офсетная. Гарнитура Таймс.
Печать трафаретная. Усл. печ. л. . Уч.- изд. л. . Тираж 36 экз. Заказ №

Издатель и полиграфическое исполнение:
Межгосударственное образовательное учреждение высшего образования
«Белорусско-Российский университет».
Свидетельство о государственной регистрации издателя,
изготовителя, распространителя печатных изданий
№ 1/156 от 07.03.2019.
Пр-т Мира, 43, 212022, г. Могилев.

© Белорусско-Российский
университет, 2026

Содержание

Введение.....	4
Лабораторная работа № 1. Общая характеристика методов статистического анализа данных. Предварительный статистический анализ данных.....	5
Лабораторная работа № 2. Методы статистического оценивания и проверки гипотез	6
Лабораторная работа № 3. Корреляционный анализ.....	10
Лабораторная работа № 4. Регрессионный анализ.....	14
Лабораторная работа № 5. Временные ряды.....	17
Лабораторная работа № 6. Методы статистической классификации многомерных данных.....	19
Список литературы.....	24

Введение

Методические рекомендации к проведению лабораторных работ по дисциплине «Статистические методы анализа данных» предназначены для студентов специальности 6-05-0611-04 «Электронная экономика». Методические рекомендации разработаны в соответствии с учебной программой.

Целью лабораторных работ является формирование у студентов практических навыков применения статистических методов для решения задач электронного маркетинга. Студенты научатся обрабатывать и анализировать маркетинговые данные, строить статистические модели, проверять гипотезы и интерпретировать полученные результаты для принятия обоснованных маркетинговых решений.

В результате выполнения лабораторных работ студент будет знать основные методы статистического анализа данных, уметь применять программные средства для статистической обработки данных, владеть навыками интерпретации результатов статистического анализа в контексте маркетинговых задач.

Лабораторные работы выполняются с использованием следующих программных средств: Microsoft Excel, Python (библиотеки pandas, scipy, sklearn) или R. Выбор программного средства согласовывается с преподавателем.

Содержание отчета : титульный лист, цель работы, краткие теоретические сведения, результаты выполнения каждого задания с расчётами и графиками, выводы по каждому заданию и общие выводы по работе. Все расчёты выполняются с использованием программных средств (Excel, Python, R). Графики оформляются в соответствии с требованиями к научной визуализации данных.

Лабораторная работа № 1. Общая характеристика методов статистического анализа данных. Предварительный статистический анализ данных

Цель работы

1 Сформировать представление о типах данных и шкалах измерения в маркетинговых исследованиях.

2 Освоить методы предварительного анализа данных: оценка пропущенных значений, выбросов, распределения.

3 Научиться вычислять и интерпретировать основные описательные статистики.

4 Развить навыки визуализации данных для выявления паттернов и аномалий.

Теоретический минимум

Типы данных в маркетинге:

– категориальные (номинальные: пол, регион; порядковые: оценка удовлетворенности);

– количественные (дискретные: число покупок; непрерывные: время на сайте, сумма чека).

Основные описательные статистики:

– меры центральной тенденции: среднее, медиана, мода;

– меры вариации: дисперсия, стандартное отклонение, размах, межквартильный размах;

– меры формы распределения: асимметрия, эксцесс.

Предварительный анализ включает:

1) проверку полноты данных (пропущенные значения);

2) выявление выбросов (методы: межквартильный размах, z-оценки);

3) оценку нормальности распределения (гистограммы, Q–Q plots, тесты Шапиро – Уилка);

4) анализ корреляций между переменными;

5) визуализацию: boxplot, гистограммы, scatterplot, barplot.

Инструменты реализации: пакеты dplyr, ggplot2, summary(), psych в R; pandas, seaborn, matplotlib в Python.

Задание к выполнению

Исходные данные. Датасет с информацией о клиентах интернет-магазина (переменные: возраст, пол, регион, средний чек, частота покупок, время на сайте, конверсия).

Этапы работы.

1 Загрузите данные в среду анализа (R/Python). Проверьте структуру: типы переменных, количество наблюдений.

2 Выполните первичную оценку качества данных. Подсчитайте количество и долю пропущенных значений по каждой переменной. Предложите стратегию обработки пропусков (удаление, импутация).

3 Рассчитайте описательные статистики для количественных переменных (среднее, медиана, SD, min, max, квартили).

4 Для категориальных переменных постройте таблицы частот и относительных частот.

5 Визуализируйте распределение ключевых метрик с помощью следующих инструментов:

- гистограмма + плотность распределения для «среднего чека»;
- boxplot для «времени на сайте» по группам «конверсия/не конверсия»;
- столбчатая диаграмма распределения клиентов по регионам.

6 Выявите потенциальные выбросы в переменной «сумма покупки» с помощью правила $1,5 \times IQR$. Проанализируйте, являются ли они ошибками ввода или реальными кейсами.

Контрольные вопросы

1 В чем разница между номинальной, порядковой и интервальной шкалами измерения? Приведите примеры из маркетинга.

2 Когда целесообразно использовать медиану вместо среднего арифметического?

3 Как интерпретировать значение коэффициента асимметрии, равное 1,5?

4 Опишите алгоритм выявления выбросов с помощью межквартильного размаха.

5 Какие визуализации наиболее информативны для анализа распределения одной количественной переменной?

6 Почему важно анализировать пропущенные значения перед построением моделей?

7 Какие особенности предварительного анализа данных характерны именно для маркетинговых исследований?

Лабораторная работа № 2. Методы статистического оценивания и проверки гипотез

Цель работы

1 Освоить построение доверительных интервалов для средних и долей.

2 Научиться формулировать нулевую и альтернативную гипотезы в контексте маркетинговых задач.

3 Применять параметрические и непараметрические критерии для сравнения групп.

4 Интерпретировать p -значения и принимать обоснованные решения на основе статистических тестов.

Теоретический минимум

В статистическом анализе данных оценивание параметров генеральной совокупности на основе выборочных наблюдений осуществляется двумя взаимодополняющими способами.

Точечная оценка предоставляет единственное численное значение в качестве приближения к неизвестному параметру: например, выборочное среднее $\bar{x}$ служит оценкой математического ожидания, а выборочная доля $\hat{p}$ – оценкой вероятности успеха в генеральной совокупности.

Однако любая точечная оценка неизбежно сопряжена со случайной ошибкой, обусловленной вариабельностью выборки, поэтому для более полной характеристики параметра используют **интервальную оценку**.

Доверительный интервал строится вокруг точечной оценки по формуле

$$\bar{x} \pm t_{\alpha/2} \cdot \frac{s}{\sqrt{n}},$$

где $t_{\alpha/2}$ – критическое значение t -распределения Стьюдента;

s – выборочное стандартное отклонение;

n – объём выборки.

Ширина получаемого интервала напрямую связана с выбранным **уровнем доверия** (традиционно 95 % или 99 %): повышение уровня доверия ведёт к расширению интервала, поскольку исследователь стремится с большей уверенностью «накрыть» истинное значение параметра, жертвуя при этом точностью локализации. Таким образом, 95-процентный доверительный интервал интерпретируется не как вероятность попадания параметра в конкретный рассчитанный интервал, а как свойство процедуры построения: при многократном повторении выборок 95 % построенных таким образом интервалов будут содержать истинное значение оцениваемого параметра.

Процедура **проверки статистических гипотез** представляет собой формализованный механизм принятия решений в условиях неопределённости на основе выборочных данных. Исследование начинается с формулировки двух взаимоисключающих утверждений: **нулевой гипотезы** H_0 , отражающей статус-кво или отсутствие эффекта (например, «средние значения в двух группах равны»), и **альтернативной гипотезы** H_1 , выражающей наличие исследуемого эффекта или различия.

В процессе проверки возможно совершение двух типов ошибок:

1) ошибка I рода заключается в необоснованном отклонении верной нулевой гипотезы (ложное открытие), вероятность чего контролируется исследователем через **уровень значимости** α (обычно 0,05 или 0,01);

2) ошибка II рода состоит в сохранении ложной нулевой гипотезы (пропуск

реального эффекта), вероятность которой обозначается β , а величина $1 - \beta$ называется мощностью теста.

Современный подход к интерпретации результатов опирается на p -значение – вероятность получить наблюдаемый в выборке результат или более экстремальный при условии, что нулевая гипотеза истинна.

Если рассчитанное p -значение оказывается меньше заранее установленного уровня значимости α , нулевая гипотеза отвергается в пользу альтернативной; в противном случае у исследователя недостаточно оснований для отклонения H_0 .

При этом важно помнить, что статистическая значимость не тождественна практической важности: даже малый, несущественный для бизнеса эффект может стать статистически значимым при достаточно большом объеме выборки, поэтому интерпретацию результатов следует дополнять оценкой размера эффекта и доверительными интервалами.

Параметрические критерии – это статистические тесты, которые предполагают, что данные принадлежат к распределению с известной функциональной формой, как правило, к нормальному распределению, и делают допущения о параметрах этого распределения (среднее, дисперсия). Применяют когда выборка достаточно велика ($n \geq 30$) или распределение близко к нормальному, дисперсии в сравниваемых группах однородны (проверка тестом Левена, Бартлетта), данные измерены в интервальной или относительной шкале.

Непараметрические критерии – это статистические тесты, которые не требуют предположений о виде распределения генеральной совокупности и работают с рангами, частотами или знаками отклонений, а не с исходными числовыми значениями. Применяют, когда:

- распределение существенно отличается от нормального (асимметрия, эксцесс);
- при обнаружении выбросов, которые невозможно обоснованно исключить;
- при малом объеме выборки ($n < 30$) проверку нормальности провести ненадежно;
- данные измерены в порядковой (ранговой) или номинальной шкале;
- дисперсии в группах неоднородны и не выравниваются преобразованиями.

В таблице 1 представлена сравнительная характеристика применения параметрических и непараметрических критериев

Задание к выполнению

Исходные данные. Результаты А/В-теста двух версий лендинга (группы А и В), включающие: время на странице, конверсия (0/1), демографические характеристики.

Этапы работы.

1 Сформулируйте исследовательский вопрос: «Влияет ли новая версия лендинга на конверсию?»

2 Рассчитайте точечные оценки и 95-процентные доверительные интервалы для:

- доли конверсии в каждой группе;

– среднего времени на странице в каждой группе.

Таблица 1 – Параметрические и непараметрические аналоги проверки статистических гипотез

Задача	Параметрический критерий	Непараметрический аналог
Проверка нормальности	–	Тесты Шапиро – Уилка, Колмогорова – Смирнова, Андерсона – Дарлинга
Сравнение среднего с константой	Одновыборочный t -критерий	Знаковый критерий, критерий Уилкоксона для одной выборки
Сравнение двух независимых групп	Двухвыборочный t -критерий	Критерий Манна – Уитни (Уилкоксона ранговый)
Сравнение дисперсий двух выборок	F -критерий Фишера	–
Сравнение двух зависимых групп	Парный t -критерий	Критерий Уилкоксона для парных выборок
Сравнение более трех независимых групп	Однофакторный ANOVA	Критерий Крускала – Уоллиса
Сравнение более трех зависимых групп	Повторные измерения ANOVA	Критерий Фридмана
Оценка линейной связи	Коэффициент Пирсона	Коэффициент Спирмена или Кендалла
Связь двух категориальных переменных	–	χ^2 -критерий Пирсона, точный тест Фишера

3 Проверьте предпосылки для применения t -критерия:

- нормальность распределения времени на странице (гистограмма, тест Шапиро – Уилка);
- равенство дисперсий (тест Левена).

4 При выполнении предпосылок примените двухвыборочный t -критерий для сравнения среднего времени. При нарушении – критерий Манна – Уитни.

5 Для сравнения долей конверсии используйте:

- двухвыборочный z -критерий для пропорций;
- χ^2 -критерий независимости в таблице сопряженности 2×2 .

6 Проведите парный анализ: если данные содержат повторные измерения (одни и те же пользователи до/после изменений), примените парный t -критерий или критерий Уилкоксона.

Результат. Отчет с пошаговым обоснованием выбора критериев, таблицами результатов тестов, интерпретацией в бизнес-терминах («новая версия увеличивает конверсию на X п. п., эффект статистически значим»).

Контрольные вопросы

- 1 В чём различие между точечной и интервальной оценкой параметра?
- 2 Что показывает доверительная вероятность и как она связана с уровнем значимости?
- 3 Как построить доверительный интервал для средней при известной и

неизвестной дисперсии?

4 Дайте определение нулевой и альтернативной гипотезы. Приведите примеры из маркетинговых исследований.

5 Объясните различие между ошибками первого и второго рода. Как они связаны с уровнем значимости и мощностью критерия?

6 Что такое p -значение и как оно используется при принятии решения?

7 Опишите алгоритм проверки гипотезы о равенстве средних двух независимых выборок.

8 Чем отличается парный t -критерий от двухвыборочного для независимых выборок?

Лабораторная работа № 3. Корреляционный анализ

Цель работы

1 Сформировать представление о типах взаимосвязей между переменными и методах их количественной оценки.

2 Освоить расчёт и интерпретацию коэффициентов корреляции Пирсона и Спирмена для количественных признаков.

3 Изучить методы оценки связи между категориальными переменными: критерий χ^2 , коэффициенты ϕ , Крамера V , контингенции.

4 Научиться проверять статистическую значимость коэффициентов корреляции и строить доверительные интервалы.

5 Применить корреляционный анализ для решения практических задач: выявление драйверов конверсии, анализ взаимосвязи каналов привлечения и поведения клиентов.

Теоретический минимум

Корреляционный анализ представляет собой совокупность методов, направленных на выявление и количественную оценку статистической связи между двумя или более переменными. Важно подчеркнуть, что корреляция отражает лишь согласованность изменений признаков, но не устанавливает причинно-следственную зависимость: наличие сильной связи между переменными не означает, что одна из них является причиной изменения другой.

При работе с **количественными переменными**, измеренными в интервальной или относительной шкале, наиболее распространённым инструментом является **коэффициент корреляции Пирсона r** . Он измеряет силу и направление *линейной* связи между двумя переменными и принимает значения от -1 до +1. Положительные значения указывают на прямую связь (с ростом одной переменной растёт другая), отрицательные – на обратную. Значения, близкие к нулю, свидетельствуют об отсутствии линейной зависимости, однако не исключают наличия нелинейной связи. Коэффициент Пирсона вычисляется как отношение ковариации двух переменных к произведению их стандартных

отклонений. Статистическая значимость коэффициента проверяется с помощью t -критерия, а для оценки устойчивости оценки целесообразно строить доверительные интервалы, например, методом Фишера (z -преобразование). Применение коэффициента Пирсона предполагает выполнение ряда условий: линейность связи, нормальность распределения хотя бы одной из переменных, отсутствие выраженных выбросов, гомоскедастичность.

Когда предпосылки для применения корреляции Пирсона нарушены (распределение отличается от нормального, наличие выбросов, связь носит монотонный, но нелинейный характер) или данные измерены в порядковой шкале (например, оценки удовлетворённости по шкале Лайкерта), целесообразно использовать **ранговый коэффициент корреляции Спирмена ρ** . Данный коэффициент оценивает степень согласованности рангов наблюдений по двум переменным и также изменяется в пределах от -1 до $+1$. Поскольку Спирмен работает с рангами, а не с исходными значениями, он устойчив к выбросам и не требует предположения о нормальности распределения. Интерпретация значения коэффициента аналогична Пирсону: чем ближе модуль к единице, тем теснее монотонная связь. Проверка значимости осуществляется на основе таблиц критических значений или асимптотического приближения.

Для анализа взаимосвязи между **категориальными (номинальными) переменными** применяются иные подходы, основанные на анализе таблиц сопряжённости. Исходным этапом служит построение таблицы частот, отражающей совместное распределение двух признаков. Статистическая проверка гипотезы о независимости переменных осуществляется с помощью **критерия χ^2 Пирсона**, который сравнивает наблюдаемые частоты с ожидаемыми при условии независимости признаков. Значимое значение статистики χ^2 (при $p < \alpha$) позволяет отвергнуть гипотезу о независимости и сделать вывод о наличии связи. Однако сам по себе χ^2 не измеряет силу связи, поскольку зависит от объёма выборки. Для количественной оценки тесноты ассоциации используются специальные коэффициенты: **коэффициент ϕ** для таблиц 2×2 , **коэффициент Крамера V** для таблиц произвольной размерности, а также **коэффициент контингенции** и **коэффициент Чупрова**. Значения этих коэффициентов варьируются от 0 (отсутствие связи) до 1 (полная зависимость), при этом интерпретация пороговых значений (слабая, умеренная, сильная связь) зависит от предметной области и размерности таблицы.

Важным аспектом корреляционного анализа является визуализация результатов. Для количественных переменных информативны диаграммы рассеяния (scatterplot) с добавлением линии регрессии, для ранговых данных – графики согласованности рангов, для категориальных – мозаичные диаграммы (mosaic plot) или тепловые карты таблиц сопряжённости. При множественном анализе корреляций целесообразно строить корреляционные матрицы с визуальным кодированием значений (например, с помощью цветовой шкалы), что позволяет быстро выявлять паттерны взаимосвязей в многомерных данных.

Задание к выполнению

Исходные данные. Датасет с информацией о клиентах интернет-магазина:

- customer_id (уникальный идентификатор);
- region (регион проживания: Центр, Север, Юг, Восток, Запад);
- age, income (демографические признаки);
- visits_month (число визитов на сайт за месяц);
- time_on_site (время на сайте);
- avg_order_value (средний чек (p.));
- total_spent (общая сумма покупок за период);
- conversion_rate (доля визитов с покупкой);
- last_purchase_days (дней с последней покупки);
- channel (основной канал привлечения (organic, paid, social, email));
- pages_viewed (количество просмотренных страниц);
- device_type (тип устройства).

Часть А. Корреляция количественных признаков: Пирсона и Спирмена.

1 Загрузите данные в среду анализа (R/Python). Выполните первичный осмотр: проверьте типы переменных, пропущенные значения, базовые описательные статистики.

2 Сформулируйте исследовательские гипотезы:

- а) «Время на сайте положительно связано с объемом покупок»;
- б) «Доход клиента коррелирует с частотой покупок».

3 Постройте матрицу диаграмм рассеяния (scatterplot matrix) для ключевых количественных переменных. Визуально оцените характер связей: линейный, нелинейный, наличие выбросов.

4 Рассчитайте коэффициенты корреляции Пирсона для всех пар количественных переменных. Постройте тепловую карту корреляционной матрицы с подписями значений.

5 Для каждой пары проверьте предпосылки применения Пирсона.

5.1 Оцените нормальность распределения (гистограммы, тест Шапиро – Уилка).

5.2 Проанализируйте наличие выбросов (boxplot, z-оценки).

5.3 Визуально проверьте линейность связи по scatterplot.

6 Для пар, где предпосылки нарушены, рассчитайте коэффициент корреляции Спирмена. Сравните значения Пирсона и Спирмена: расхождения могут указывать на нелинейность или влияние выбросов.

7 Проверьте статистическую значимость коэффициентов (расчёт p -значений). Отберите значимые связи на уровне $\alpha = 0,05$.

8 Для двух-трех ключевых коэффициентов постройте 95-процентные доверительные интервалы методом Фишера. Интерпретируйте результаты в контексте маркетинга: какие поведенческие метрики наиболее тесно связаны с целевыми показателями.

Часть Б. Анализ связи категориальных признаков.

- 1 Выберите две категориальные переменные для анализа, например, `channel` (канал привлечения) и `conversion` (конверсия).
 - 2 Постройте таблицу сопряжённости абсолютных и относительных частот. Визуализируйте распределение с помощью столбчатой диаграммы с группировкой или мозаичного графика.
 - 3 Сформулируйте гипотезы:
 - а) H_0 : канал привлечения и конверсия независимы;
 - б) H_1 : между признаками существует статистическая связь;
 - 4 Проверьте условия применимости критерия χ^2 : ожидаемые частоты в ячейках таблицы должны быть не менее пяти (в противном случае используйте точный тест Фишера).
 - 4.1 Выполните расчёт критерия χ^2 : получите значение статистики, степени свободы, p -значение. Сделайте вывод о значимости связи.
 - 4.2 Рассчитайте меру силы связи:
 - а) для таблицы 2×2 – коэффициент ϕ ;
 - б) для таблиц большей размерности – коэффициент Крамера V .
- Интерпретируйте величину коэффициента (слабая/умеренная/сильная связь).
- Повторите анализ для другой пары категориальных переменных, например, `region` и `device_type`. Сравните результаты.

Часть В. Интеграционный анализ и интерпретация.

- 1 Объедините результаты: постройте сводную таблицу значимых корреляций (количественные + категориальные признаки) с указанием коэффициентов, p -значений и размера эффекта.
- 2 Выявите потенциальные проблемы мультиколлинеарности: пары количественных предикторов с $|r| > 0,7$ могут дублировать информацию в регрессионных моделях.
- 3 Сформулируйте практические выводы для электронного маркетинга:
 - 3.1 Какие каналы привлечения ассоциированы с более высокой конверсией?
 - 3.2 Какие поведенческие метрики наиболее информативны для прогнозирования размера чека?
 - 3.3 Есть ли региональные различия в предпочтениях устройств, и как это влияет на дизайн коммуникаций?
- 4 Подготовьте визуальный отчёт: корреляционная тепловая карта, ключевые scatterplot, мозаичные диаграммы для категориальных связей.

Контрольные вопросы

- 1 В чём принципиальное различие между коэффициентами корреляции Пирсона и Спирмена? Когда целесообразно использовать каждый из них?

2 Почему коэффициент корреляции, близкий к нулю, не гарантирует отсутствие связи между переменными? Приведите пример нелинейной зависимости.

3 Как интерпретировать значение коэффициента Крамера $V = 0,35$ для таблицы 3×4 ? Какие пороговые значения принято использовать для оценки силы связи?

4 Почему при анализе таблиц сопряжённости необходимо проверять ожидаемые частоты? Что делать, если в более чем 20 % ячеек ожидаемая частота менее пяти?

5 Как построить доверительный интервал для коэффициента корреляции Пирсона? Почему для этого применяется z -преобразование Фишера?

6 В чём ограничения корреляционного анализа при выявлении причинно-следственных связей? Какие дополнительные методы могут помочь установить причинность?

7 Как влияет объём выборки на статистическую значимость коэффициента корреляции? Может ли слабая корреляция быть значимой при большом n ?

8 Почему при множественном анализе корреляций требуется коррекция уровня значимости? Какие методы коррекции вы знаете и когда их применять?

9 Как визуализировать корреляционную матрицу для более 10 переменных, чтобы обеспечить наглядность и информативность?

10 Какие особенности интерпретации корреляций характерны для маркетинговых данных: асимметрия распределений, наличие выбросов, смешивающие факторы?

Лабораторная работа № 4. Регрессионный анализ

Цель работы

1 Освоить построение и интерпретацию простой и множественной линейной регрессии.

2 Научиться оценивать качество модели: R^2 , скорректированный R^2 , F -статистика.

3 Проверять предпосылки метода наименьших квадратов: линейность, гомоскедастичность, нормальность остатков, отсутствие мультиколлинеарности.

Теоретический минимум

Линейная регрессия: функция, позволяющая по величине изменения одного коррелируемого признака определить среднюю величину другого признака.

$$Y = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p + \varepsilon.$$

Для оценки параметров регрессионного уравнения традиционно применяется метод наименьших квадратов, который находит такие значения коэффициентов, при которых сумма квадратов отклонений фактических наблюдений от предсказанных моделью значений является минимальной. Корректность и статистическая надёжность получаемых оценок напрямую зависят от выполнения ряда фундаментальных предпосылок.

Во-первых, предполагается линейность связи между зависимой и независимыми переменными, что проверяется визуально с помощью диаграмм рассеяния и графиков зависимости остатков от предсказанных значений.

Во-вторых, остатки модели должны быть независимыми, что особенно критично при работе с последовательными данными и формально оценивается, например, тестом Дарбина – Уотсона.

В-третьих, требуется гомоскедастичность, то есть постоянство дисперсии ошибок на всём диапазоне значений предикторов; её нарушение выявляется анализом графиков остатков или тестом Бреуша – Пагана.

В-четвёртых, распределение остатков должно приближаться к нормальному, что диагностируется по гистограммам, Q–Q plot'ам и тесту Шапиро – Уилка.

Наконец, модель не должна страдать от сильной мультиколлинеарности, когда предикторы тесно коррелируют между собой; для контроля этого явления используется фактор инфляции дисперсии (VIF), значения которого в качественной модели не должны превышать порогов в 5–10 единиц.

При соблюдении указанных условий коэффициенты регрессии β_j приобретают чёткую содержательную интерпретацию: каждый из них показывает, на сколько единиц изменится зависимая переменная при изменении X_j на одну единицу при фиксированных остальных переменных.

Это позволяет количественно оценить изолированное влияние каждого маркетингового или экономического фактора на целевой показатель.

Качество построенной модели оценивается на нескольких уровнях. Коэффициент детерминации R^2 показывает долю объясненной дисперсии результативного фактора за счет совокупной вариации предикторов. Однако его значение искусственно завышается при добавлении новых переменных. Для устранения этого недостатка используют скорректированный R^2 (штраф за количество предикторов).

Статистическую значимость модели в целом проверяют с помощью F -теста: значимость модели в целом, тогда как значимость отдельных коэффициентов оценивается через t -тесты.

Помимо показателей внутренней согласованности, точность прогнозных свойств модели оценивают с помощью метрик ошибок, рассчитываемых на тестовой или валидационной выборке.

Средняя абсолютная ошибка (MAE) вычисляется как среднее арифметическое абсолютных разностей между фактическими и предсказанными значениями, предоставляя интуитивно понятную меру отклонения.

Среднеквадратичная ошибка (MSE) усредняет квадраты этих разностей, тем самым усиливая вес крупных расхождений и делая метрику особенно чувствительной к выбросам.

Для более наглядной интерпретации в исходных единицах измерения применяют корень из среднеквадратичной ошибки (RMSE), который возвращает метрику к масштабу исходных данных и отражает среднее типичное отклонение прогнозов модели от реальности.

Задание к выполнению

Исходные данные. Датасет с метриками цифрового маркетинга.

Этапы работы.

1 Постройте простую линейную регрессию: зависимость «общего дохода» от «расходов на контекстную рекламу».

1.1 Визуализируйте данные и линию регрессии.

1.2 Интерпретируйте коэффициент наклона и R^2 .

2 Постройте множественную регрессию: доход $\sim$ расходы на каналы + трафик + средний чек.

2.1 Проверьте статистическую значимость каждого предиктора.

2.2 Рассчитайте и сравните R^2 и скорректированный R^2 .

3 Проведите диагностику остатков модели.

3.1 Постройте график остатков от предсказанных значений

3.2 Оцените нормальность (Q–Q plot).

3.3 Рассчитайте VIF для проверки мультиколлинеарности.

4 При выявлении нарушений предпосылок предложите корректировки: преобразование переменных (логарифмирование), удаление выбросов, использование робастных стандартных ошибок.

Результат. Отчет с уравнениями моделей, таблицами коэффициентов, диагностическими графиками, интерпретацией в терминах маркетинга («увеличение бюджета на соцсети на 10 % при прочих равных ведет к росту дохода на X %»).

Контрольные вопросы

1 Почему коэффициент регрессии нельзя интерпретировать как причинно-следственный эффект без дополнительных условий?

2 В чем разница между R^2 и скорректированным R^2 ? Когда последний предпочтительнее?

3 Как по графику остатков выявить гетероскедастичность?

4 Что показывает значение $VIF = 8$ для предиктора?

Лабораторная работа № 5. Временные ряды

Цель работы

- 1 Освоить визуализацию и декомпозицию временных рядов (тренд, сезонность, шум).
- 2 Научиться проверять ряд на стационарность и применять преобразования.
- 3 Построить прогнозные модели: скользящее среднее, экспоненциальное сглаживание, ARIMA.
- 4 Оценить точность прогнозов с использованием метрик (MAE, RMSE, MAPE).

Теоретический минимум

Любой временной ряд представляет собой последовательность наблюдений, упорядоченных во времени, и его внутренняя структура обычно складывается из нескольких взаимосвязанных компонент.

Основу динамики часто составляет тренд – долгосрочная направленная тенденция изменения показателя, отражающая устойчивый рост, снижение или стагнацию. Поверх тренда могут накладываться сезонные колебания – регулярные и повторяющиеся паттерны с фиксированным периодом (суточные, недельные или годовые циклы), которые характерны для потребительского спроса и маркетинговых активностей. Помимо сезонности, ряды могут демонстрировать цикличность – более длительные и непериодические колебания, обусловленные макроэкономическими факторами или фазами делового цикла. Оставшаяся, необъяснённая часть вариации относится к случайной компоненте (шуму), которая отражает влияние непредсказуемых факторов, внешние шоки и измерительные погрешности.

Ключевым условием применения большинства классических моделей прогнозирования является стационарность ряда. В слабой форме стационарность подразумевает, что математическое ожидание и дисперсия процесса остаются постоянными во времени, а автоковариация между наблюдениями зависит исключительно от временного лага, а не от конкретного момента наблюдения.

На практике стационарность проверяют с помощью визуального анализа графиков ряда и автокорреляционных функций, а также формальных статистических тестов, среди которых наиболее распространён расширенный тест Дики – Фуллера (ADF). Если ряд оказывается нестационарным, его приводят к стационарному виду с помощью специальных преобразований: логарифмирования для стабилизации дисперсии или дифференцирования (вычисления разностей между последовательными наблюдениями) для устранения тренда.

В зависимости от сложности структуры ряда и целей исследования выбирают различные модели прогнозирования. На начальном этапе часто применяются наивные методы, такие как использование последнего известного значения, простого среднего или скользящего среднего, которые служат базисом для оценки более сложных алгоритмов.

Для учёта трендовых и сезонных паттернов широко используется экспоненциальное сглаживание, в частности модель Хольта - Винтерса, адаптивно взвешивающая прошлые наблюдения с учётом уровня, тренда и сезонности. Более гибким и мощным инструментом являются модели класса $ARIMA(p, d, q)$, сочетающие авторегрессионную компоненту $AR(p)$, учитывающую линейную зависимость текущего значения от прошлых; интегрирование $I(d)$, отвечающее за порядок дифференцирования; и компоненту скользящего среднего $MA(q)$, моделирующую влияние прошлых ошибок прогноза. Когда ряд обладает выраженной сезонностью, применяется расширенная версия модели – $SARIMA$, которая включает дополнительные сезонные параметры авторегрессии и скользящего среднего.

Качество построенных моделей оценивается на основе их способности точно предсказывать будущие значения. Для этого данные разделяют на обучающую и тестовую выборки с обязательным сохранением временного порядка, часто используя специализированные процедуры кросс-валидации для временных рядов (time-series cross-validation), которая последовательно расширяет обучающую выборку и проверяет прогноз на следующих периодах.

Точность прогнозов измеряют с помощью стандартных метрик ошибок:

- средней абсолютной ошибки (MAE), отражающей среднее абсолютное отклонение прогноза от факта;
- корня из средней квадратичной ошибки (RMSE), которая более чувствительна к крупным выбросам;
- средней абсолютной процентной ошибки (MAPE), удобной для интерпретации в относительных величинах.

Сравнение этих метрик на тестовой выборке позволяет выбрать наиболее адекватную модель, оценить её устойчивость и определить горизонт практического применения в маркетинговом планировании.

Задание к выполнению

Исходные данные. Ежедневные/еженедельные данные о трафике сайта, количестве заказов или выручке интернет-магазина за 1–2 года.

Этапы работы.

1 Загрузите временной ряд, преобразуйте индекс в формат даты, постройте график ряда.

2 Выполните декомпозицию ряда (аддитивная или мультипликативная модель) для выделения тренда и сезонности.

3 Проверьте ряд на стационарность с помощью теста Дики – Фуллера. При необходимости выполните дифференцирование.

4 Постройте автокорреляционную (ACF) и частичную автокорреляционную (PACF) функции для подбора параметров $ARIMA$.

5 Обучите несколько моделей:

- простое экспоненциальное сглаживание;
- модель Хольта – Винтерса (с трендом и сезонностью);

– модель ARIMA или SARIMA (подберите p , d , q через AIC или автоматический подбор).

6 Разделите данные на обучающую (80 %) и тестовую (20 %) выборки. Постройте прогноз на горизонт 7–30 дней.

7 Оцените точность прогнозов на тестовой выборке с помощью MAE, RMSE, MAPE. Сравните модели.

8 Визуализируйте фактические значения, прогноз и доверительный интервал.

9 Сформулируйте рекомендации для планирования маркетинговых активностей на основе выявленной сезонности и прогноза.

Результат. Отчет с графиками декомпозиции, подбором модели, таблицей метрик ошибок и прогнозом на ближайший период. Экспорт прогноза в Excel для интеграции в отчетность.

Контрольные вопросы

1 Дайте определение временного ряда и приведите примеры из практики электронного маркетинга.

2 Назовите основные компоненты временного ряда и дайте им характеристику.

3 В чём различие между аддитивной и мультипликативной моделями декомпозиции?

4 Опишите метод скользящей средней и его применение для сглаживания.

5 Что показывает параметр α в экспоненциальном сглаживании?

6 Как выбрать форму тренда для моделирования?

7 Как рассчитываются сезонные индексы?

8 Опишите алгоритм прогнозирования с учётом тренда и сезонности.

9 Какие показатели используются для оценки точности прогноза?

10 Как интерпретируется значение MAPE?

Лабораторная работа № 6. Методы статистической классификации многомерных данных

Цель работы

1 Освоить методику проведения однофакторного дисперсионного анализа (ANOVA) для сравнения средних значений целевых метрик по нескольким группам.

2 Научиться применять пост-хок тесты (Тьюки, Бонферрони) для попарных сравнений и интерпретировать их результаты.

3 Изучить основы кластерного анализа: алгоритмы k -средних и иерархической кластеризации.

4 Сформировать навыки интерпретации кластеров и профилирования сегментов пользователей для маркетинговых решений.

5 Применить полученные методы для решения практической задачи: сегментация клиентов интернет-магазина и сравнение эффективности регионов продаж.

Теоретический минимум

Однофакторный дисперсионный анализ (One-Way ANOVA).

Назначение. Сравнение средних значений количественной зависимой переменной по трём и более независимым группам, определяемым одним категориальным фактором.

Модель.

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij},$$

где Y_{ij} – значение зависимой переменной у j -го наблюдения в i -й группе;

μ – общее среднее;

α_i – эффект i -й группы ($\sum \alpha_i = 0$);

ε_{ij} – случайная ошибка, $\varepsilon \sim N(0, \sigma^2)$.

Гипотезы.

H_0 : $\mu_1 = \mu_2 = \dots = \mu_k$ (все групповые средние равны);

H_1 : хотя бы одно среднее отличается.

F-статистика.

$$F = \frac{MS_{between}}{MS_{within}} = \frac{SS_{between} / (k - 1)}{SS_{within} / (N - k)},$$

где $SS_{between}$ – межгрупповая дисперсия;

SS_{within} – внутригрупповая дисперсия;

k – количество групп;

N – общее количество наблюдений.

Для корректного применения однофакторного дисперсионного анализа необходима проверка выполнимости следующих предпосылок ANOVA.

1 Независимость наблюдений.

2 Нормальность распределения остатков в каждой группе (проверка тестами Шапиро – Уилка, Q–Q plot).

3 Гомогенность дисперсий (проверка тестами Левена, Бартлетта).

Если предпосылки нарушены, можно осуществить следующие действия:

– преобразование зависимой переменной (логарифм, квадратный корень);

– использование непараметрического аналога: критерий Крускала –

Уоллиса;

– применение робастных версий ANOVA (Welch's ANOVA при неравенстве дисперсий).

В таблице 2 представлены тесты, которые проводятся только при отклонении H_0 в ANOVA для выявления конкретных различий между парами групп.

Таблица 2 – Пост-хок тесты (множественные сравнения)

Тест	Когда применять
Тьюки (Tukey HSD)	Сбалансированные или умеренно несбалансированные выборки; равные дисперсии
Бонферрони	Малое число сравнений; требуется строгий контроль ошибки I рода
Геймса – Хауэлла	Нарушена предпосылка гомогенности дисперсий

Кластерный анализ. Цель – группировка объектов (клиентов) в однородные сегменты на основе схожести по множеству признаков без заранее известных меток (обучение без учителя). В качестве меры близости между i и j объектами в работе использовать Евклидово расстояние:

$$d_{ij} = \sqrt{\sum_{k=1}^p (x_{ik} - x_{jk})^2}.$$

Алгоритм метода k -средних (k -means), минимизирующий внутрикластерную сумму квадратов расстояний до центра кластера (WSS), состоит из следующих шагов.

- 1 Задать число кластеров k .
- 2 Инициализировать центроиды.
- 3 Повторять до сходимости.
 - 3.1 Назначить каждый объект ближайшему центроиду.
 - 3.2 Пересчитать центроиды как среднее по объектам кластера.

Оптимальное число кластеров k можно определить с помощью методов локтя и силуэта.

Метод локтя (Elbow method) предполагает построения графика зависимости WSS от k . Оптимальным считается количество кластеров, соответствующее «излому» графика.

Силуэтный коэффициент вычисляется по формуле

$$s(i) = \frac{b(i) - a(i)}{\max \{a(i), b(i)\}},$$

где $a(i)$ – среднее расстояние внутри кластера;

$b(i)$ – минимальное среднее расстояние до другого кластера.

Если значения $s(i) > 0,5$, то это указывает на разумную кластеризацию.

Иерархическая кластеризация реализуется с помощью агломеративного подхода, суть которого предполагает последовательное снижение размерности матрицы расстояний между объектами. В самом начале каждый объект считается

отдельным кластером, далее происходит последовательное объединение ближайших с использованием различных мер связи.

Существует ряд различных методов иерархической кластеризации.

- 1 Метод Уорда (англ. Ward's method).
- 2 Метод одиночной связи (англ. single linkage).
- 3 Метод полной связи (англ. complete linkage).
- 4 Метод средней связи (англ. pair-group method using arithmetic averages).

Результат визуализируется дендрограммой; число кластеров выбирается по высоте «разрезания»

Интерпретация кластеров.

- 1 Расчёт средних/медианных значений признаков по кластерам.
- 2 Визуализация: параллельные координаты.
- 3 Профилирование: присвоение содержательных названий сегментам («лояльные», «ценовые», «новые» и т. д.).

Задание к выполнению

Исходные данные. Датасет с информацией о клиентах интернет-магазина:

- customer_id: уникальный идентификатор;
- region: регион проживания (Центр, Север, Юг, Восток, Запад);
- age, income: демографические признаки;
- visits_month: число визитов на сайт за месяц;
- avg_order_value: средний чек (руб.);
- total_spent: общая сумма покупок за период;
- conversion_rate: доля визитов с покупкой;
- last_purchase_days: дней с последней покупки;
- channel: основной канал привлечения (organic, paid, social, email).

Часть А. Однофакторный дисперсионный анализ: сравнение эффективности регионов

- 1 Загрузите данные в R/Python. Проверьте структуру, типы переменных, пропущенные значения.
- 2 Сформулируйте исследовательский вопрос: «Различается ли средний чек (avg_order_value) клиентов из разных регионов?»
- 3 Визуализируйте распределение avg_order_value по регионам с помощью boxplot с русскими подписями осей.
- 4 Проверьте предпосылки ANOVA:
 - а) нормальность остатков в каждой группе (тест Шапиро – Уилка, гистограммы);
 - б) гомогенность дисперсий (тест Левена).
- 5 Если предпосылки выполнены, проведите однофакторный ANOVA. Если нарушены, используйте критерий Крускала – Уоллиса или преобразуйте переменную.
- 6 При значимом результате ANOVA ($p < 0,05$) выполните пост-хок сравнения:
 - а) тест Тьюки для всех попарных сравнений;

б) коррекция Бонферрони для выбранных пар регионов (по гипотезе исследования).

7 Представьте результаты в виде таблицы с групповыми средними, доверительными интервалами разностей и скорректированными p -значениями (используйте *kableExtra* в R или *pandas styler* в Python).

8 Сформулируйте выводы. Какие регионы статистически значимо различаются по среднему чеку? Какие маркетинговые рекомендации можно дать для регионов с низкими показателями?

Часть Б. Кластерный анализ: сегментация клиентов по поведенческим признакам.

1 Подготовьте данные для кластеризации.

1.1 Отберите количественные поведенческие переменные: `visits_month`, `avg_order_value`, `conversion_rate`, `last_purchase_days`, `total_spent`.

1.2 Обработайте пропуски (медианная импутация или удаление).

1.3 Стандартизируйте признаки (z -score), так как k -means чувствителен к масштабу.

2 Определите оптимальное число кластеров.

2.1 Постройте график «метод локтя» (WSS от k для $k = 1 \dots 10$).

2.2 Рассчитайте силуэтный коэффициент для $k = 2 \dots 6$.

2.3 Обоснуйте выбор финального k .

3 Примените метод k -средних с выбранным k .

4 Проведите иерархическую кластеризацию для сравнения.

4.1 Постройте дендрограмму с использованием метода Уорда.

4.2 «Разрежьте» дендрограмму на то же число кластеров.

4.3 Оцените согласованность с k -means (таблица сопряженности).

5 Проинтерпретируйте кластеры.

5.1 Рассчитайте средние значения исходных (нестандартизированных) признаков по каждому кластеру.

5.2 Визуализируйте профили кластеров с помощью параллельных координат или столбчатых диаграмм.

5.3 Присвойте каждому кластеру содержательное название (например, «Высокоценные лояльные», «Редкие покупатели с низким чеком», «Новые клиенты с потенциалом»).

6 Проанализируйте распределение кластеров по регионам и каналам привлечения (таблицы сопряженности, χ^2 -тест).

7 Сформулируйте маркетинговые рекомендации для каждого сегмента.

7.1 Какие каналы коммуникации предпочтительны?

7.2 Какие типы предложений могут повысить конверсию или средний чек?

7.3 Как выстраивать стратегию удержания для разных групп?

Контрольные вопросы

- 1 В чем принципиальное различие между дисперсионным анализом и серией попарных t -критериев? Почему ANOVA предпочтительнее при сравнении трёх и более групп?
- 2 Как интерпретировать результат теста Левена с $p = 0,02$ при проведении ANOVA? Какие альтернативы доступны исследователю?
- 3 Почему пост-хок тесты требуют коррекции уровня значимости?
- 4 Когда целесообразно использовать тест Тьюки, а когда – поправку Бонферрони?
- 5 Почему стандартизация признаков обязательна при применении метода k -средних? Что произойдет, если её не выполнить?
- 6 Как метод «локтя» и силуэтный коэффициент дополняют друг друга при выборе числа кластеров?
- 7 В чем преимущества и недостатки иерархической кластеризации по сравнению с k -means?
- 8 Как проверить устойчивость результатов кластеризации к изменению начальных условий или подвыборки данных?
- 9 Почему интерпретация кластеров должна проводиться на исходных (нестандартизированных) значениях признаков?
- 10 Какие этические и практические ограничения следует учитывать при использовании результатов сегментации для персонализации маркетинговых предложений?

Список литературы

- 1 **Кабаков, Р. Р.** *Р в действии. Анализ и визуализация данных на языке R : практическое руководство* / Р. Кабаков. – 2-е изд. – М. : ДМК Пресс, 2023. – 590 с.
- 2 **Лонг, Д. Р.** *Книга рецептов. Проверенные рецепты для статистики, анализа и визуализации данных : практическое руководство* / Д. Лонг, П. Титор. – М. : ДМК Пресс, 2020. – 510 с.
- 3 **Масицкий, С. Э.** *Статистический анализ и визуализация данных с помощью R : практическое руководство* / С. Э. Масицкий, В. К. Шитиков. – 2-е изд. – М. : ДМК Пресс, 2023. – 497 с.
- 4 **Гайдель, А. В.** *Лабораторный практикум по курсу «Интеллектуальный анализ данных» : практикум* / А. В. Гайдель, А. Г. Храмов. – Самара : Самар. ун-т, 2019. – 104 с.