Казахский национальный университет имени аль-Фараби
Факультет медицины и здравоохранения
Кафедра фундаментальной медицины

Лекция 15

Анализ данных: планирование и подготовка

Дисциплина: Дизайн экспериментов в разработке лекарственных средств (DERLS 5208)

Образовательная программа: 7M10117 – Фармация (магистратура)

Лектор: PhD, ассоциированный профессор Калыкова А.С.

Алматы, 2026

Цель

Сформировать навыки планирования анализа данных до начала эксперимента, подготовки данных (очистка, пропуски, выбросы), выбора статистических методов и корректной интерпретации результатов, включая ANOVA и регрессионный анализ планов эксперимента.



Вопросы для обсуждения

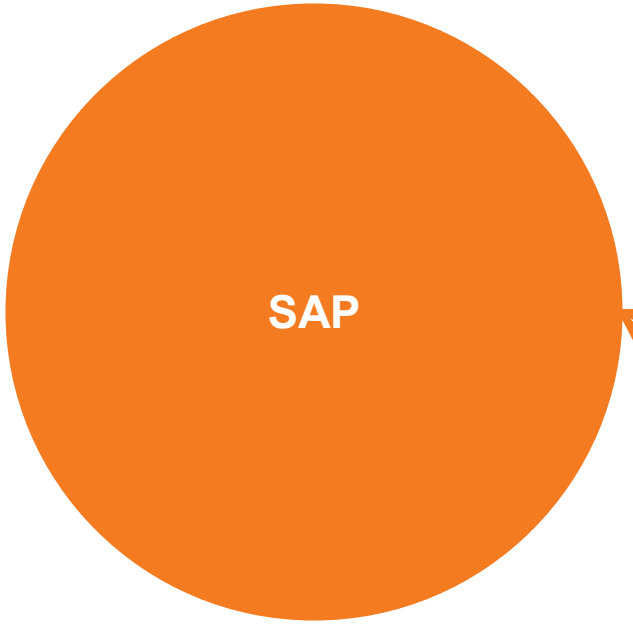
- Что должен содержать план статистического анализа (SAP)?
- Как подготовить данные и проверить допущения методов?
- Как выбрать статистический критерий и интерпретировать результат?

Анализ данных как часть дизайна эксперимента

- План анализа разрабатывается вместе с планом эксперимента и фиксируется до сбора данных.
- Это защищает от «подгонки» анализа под результат (p-hacking) и повышает воспроизводимость.
- Основные шаги: описание данных → подготовка → проверка допущений → выбор метода → интерпретация → отчёт.
- Статистика не исправляет плохо спланированный эксперимент.



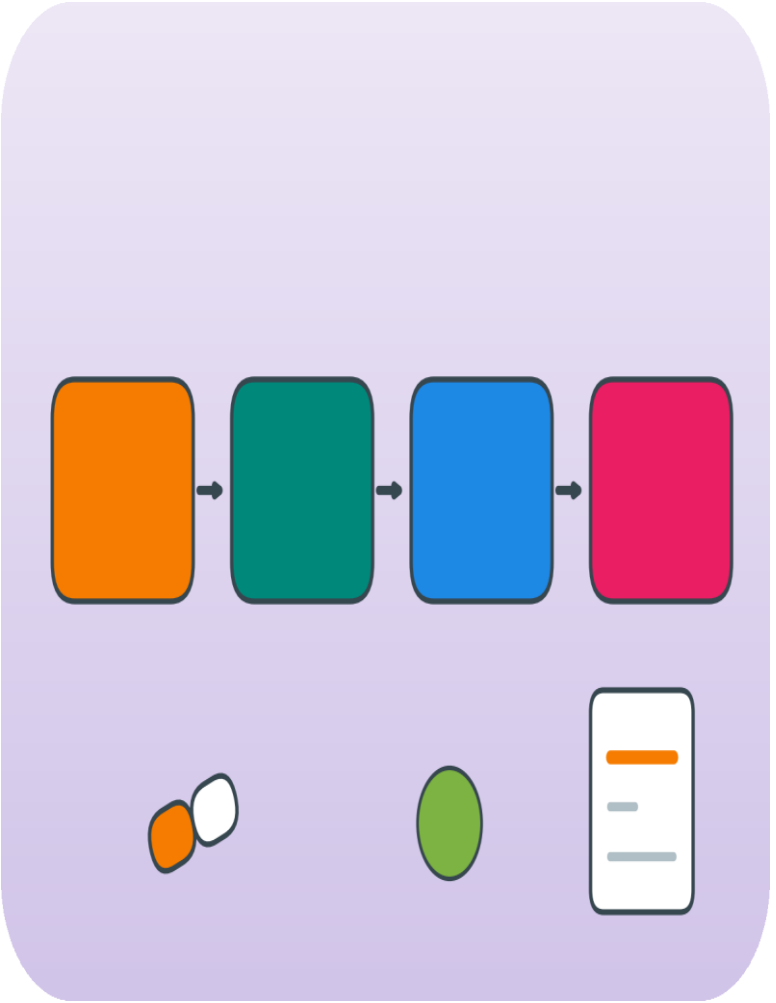
План статистического анализа (SAP)



- Цели и гипотезы; первичные и вторичные исходы
- Анализируемые наборы данных (все рандомизированные, соответствующие протоколу)
- Методы: критерии, модели, ковариаты, поправки на множественные сравнения
- Обработка пропусков и выбросов; критерии исключения
- Уровень значимости, мощность, программное обеспечение
- Таблицы, графики и форма представления результатов

Подготовка данных к анализу

1. **Проверка ввода** двойной ввод, сверка с исходными записями.
2. **Кодирование** единицы измерения, шкалы, факторы.
3. **Поиск пропусков** механизм: MCAR, MAR, MNAR (Little, Rubin, 2019).
4. **Поиск выбросов** графики, правило 1,5 IQR, причины.
5. **Преобразование** логарифмирование, стандартизация – при необходимости.
6. **Фиксация изменений** журнал очистки данных.



Типы данных и выбор критерия

Задача	Параметрический критерий	Непараметрический	Данные
Две независимые группы	t-критерий Стьюдента	Манна–Уитни	Непрерывные
Парные наблюдения	Парный t-критерий	Вилкоксона	Непрерывные
Три и более групп	ANOVA	Краскела–Уоллиса	Непрерывные
Частоты	χ^2-критерий	Точный критерий Фишера	Категориальные
Связь двух переменных	Корреляция Пирсона	Корреляция Спирмена	Непрерывные

Проверка допущений: нормальность (Шапиро–Уилка, Q–Q график), однородность дисперсий (Ливиня).

Интерпретация результатов

01

p-значение

Вероятность получить такие или более выраженные данные, если H0 верна; не вероятность гипотезы (Wasserstein, Lazar, 2016).

02

Доверительный интервал

Диапазон правдоподобных значений эффекта; показывает точность оценки.

03

Размер эффекта

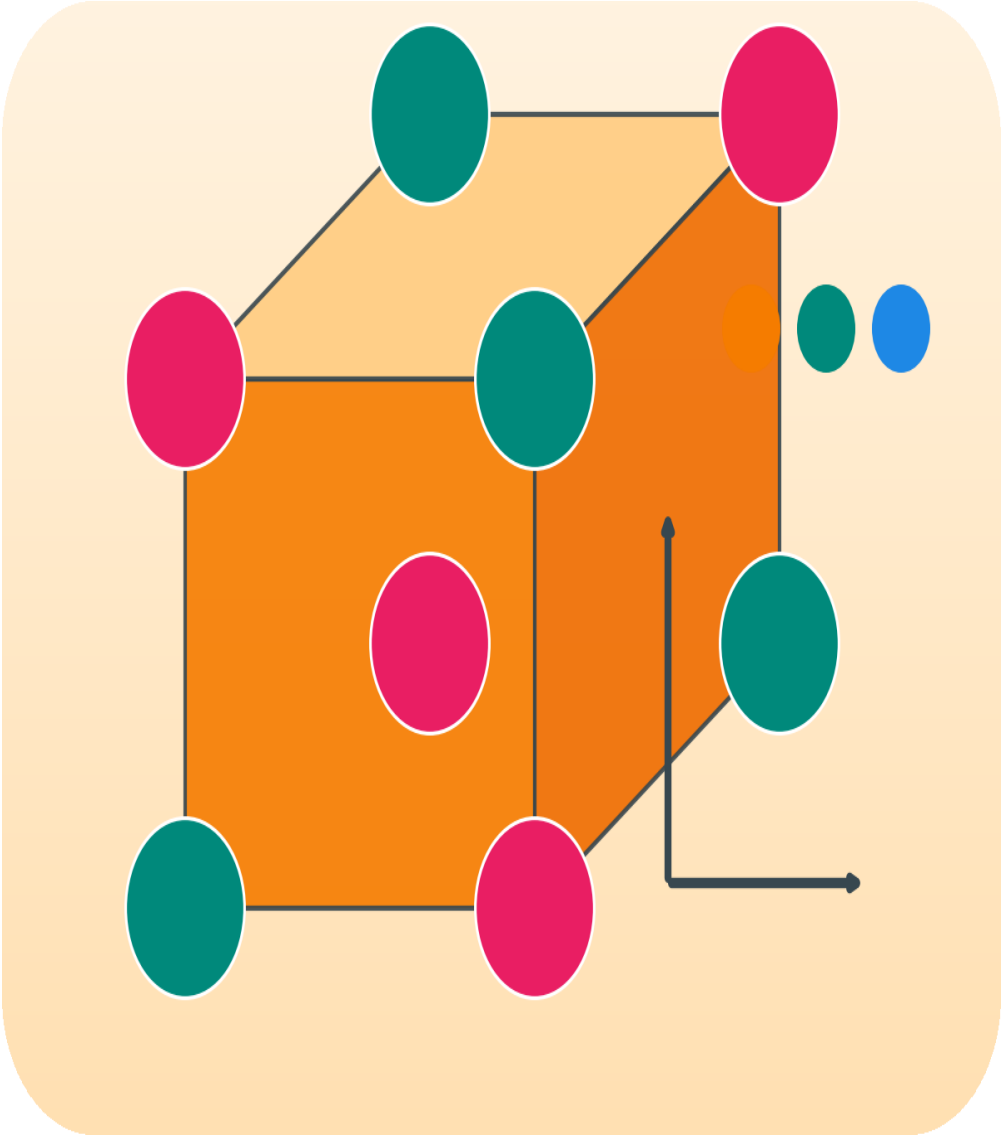
Величина различия (разность средних, d Коэна, OR) – практическая значимость.



Статистическая значимость не равна практической значимости.

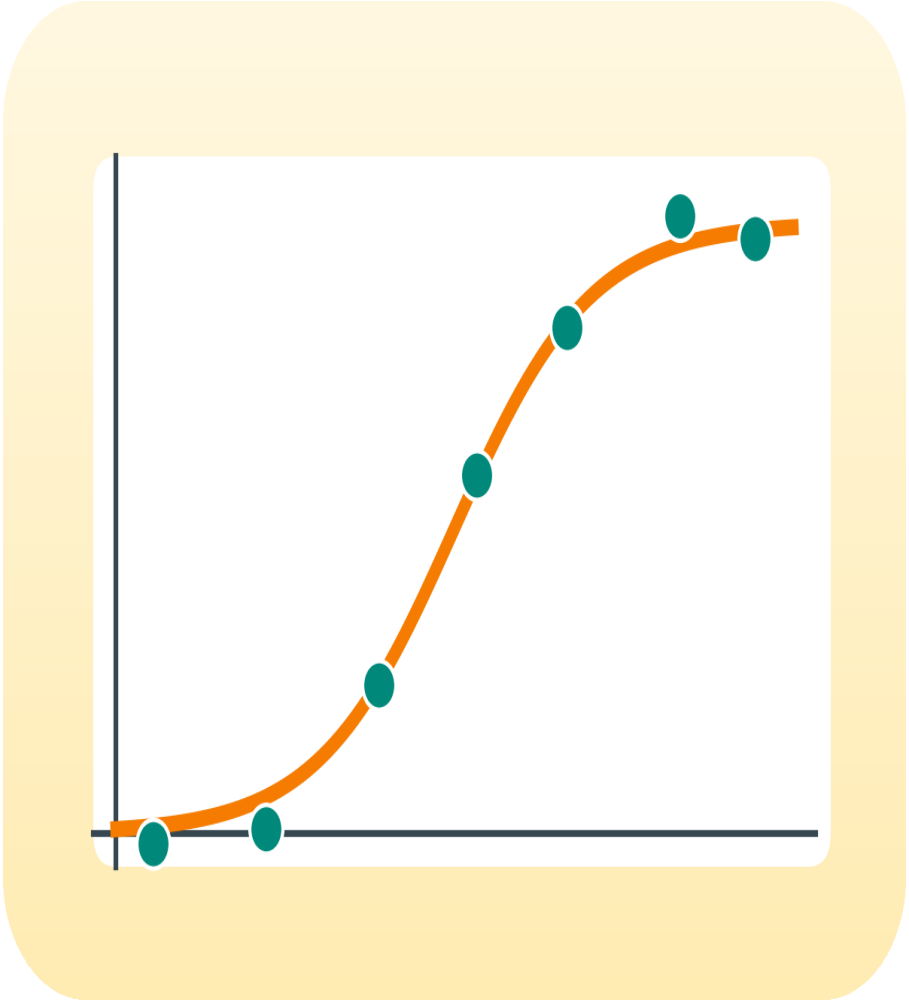
Анализ данных факторного эксперимента

- Модель: y = b0 + Σbixi + Σbijxixj; коэффициенты оцениваются методом наименьших квадратов.
- В ортогональных планах коэффициенты независимы: bj = Σxijyi / N.
- Значимость коэффициентов – t-критерий; адекватность модели – F-критерий (проверка lack-of-fit).
- ANOVA разделяет вариацию на вклад факторов и ошибку (Montgomery, 2019).



Диагностика модели и остатков

- **Остатки против предсказанных:** проверка однородности дисперсий и адекватности формы модели.
- **Нормальный график остатков:** проверка нормальности.
- **Остатки против порядка опытов:** выявление дрейфа и нарушения независимости.
- **R^2, скорректированный R^2, прогностический R^2:** качество модели.
- **Подтверждающие опыты:** проверка прогноза в новых точках.



Множественные сравнения и воспроизводимость

- При k сравнениях вероятность хотя бы одной ложной значимости растёт: 1 – (1 – α)^k.
- Поправки: Бонферрони (α/k), Тьюки для попарных сравнений, Холма.
- Воспроизводимость: скрипты анализа (R, Python), фиксация версий, публикация данных (лекция 11.1).
- Отчёт: описание методов, объёма выборки, исключений, программ и версий.



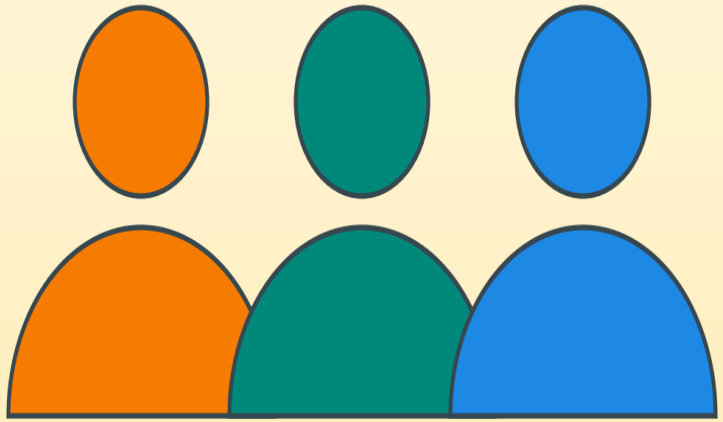
Программное обеспечение для анализа

- **R:** свободная среда; пакеты FrF2, rsm, DoE.base.
- **Python:** SciPy, statsmodels, pandas, pyDOE3.
- **Коммерческие:** Minitab, JMP, Design-Expert, STATISTICA.
- **jamovi, JASP:** графические интерфейсы, свободные.



Типичные ошибки анализа данных

- Выбор метода после просмотра результатов.
- Удаление выбросов без обоснования.
- Игнорирование пропусков и множественных сравнений.
- Интерпретация $p > 0,05$ как «эффекта нет».
- Отсутствие доверительных интервалов и размера эффекта.



Видеоматериалы и интернет-ресурсы для демонстрации

Видео (YouTube)

Data cleaning and preparation
Очистка и подготовка данных

Steps of data analysis in research (quantitative)
Шаги количественного анализа данных

Data screening: preparing raw data for analysis
Первичная проверка данных перед анализом

Сайты и базы данных

NIST/SEMATECH e-Handbook: Exploratory Data Analysis
Разведочный анализ данных

The R Project for Statistical Computing
Среда для статистического анализа

SciPy
Научные вычисления на Python

jamovi
Свободное статистическое ПО

ICH E9
Статистические принципы клинических исследований



Заключение

- План анализа фиксируется до сбора данных.
- Подготовка данных включает проверку, обработку пропусков и выбросов.
- Метод выбирают по типу данных, допущениям и задаче.
- Для факторных планов применяют регрессию и ANOVA с диагностикой остатков.
- Интерпретация включает p-значение, доверительные интервалы и размер эффекта.

Список литературы

1. Montgomery D.C. Design and Analysis of Experiments. 10th ed. Hoboken: Wiley, 2019. 688 p.
2. Little R.J.A., Rubin D.B. Statistical Analysis with Missing Data. 3rd ed. Hoboken: Wiley, 2019. 449 p.
3. Wasserstein R.L., Lazar N.A. The ASA statement on p-values: context, process, and purpose // The American Statistician. 2016. Vol. 70, No. 2. P. 129–133. https://doi.org/10.1080/00031305.2016.1154108
4. ICH E9. Statistical Principles for Clinical Trials. 1998. https://www.ich.org/page/efficacy-guidelines
5. Sidnyaev N.I. Theory of experimental design and analysis of statistical data: a textbook for universities. Moscow: Yurait, 2011. 399 p.
6. NIST/SEMATECH e-Handbook of Statistical Methods. https://www.itl.nist.gov/div898/handbook/ (дата обращения: 11.11.2026).
7. Видео: см. слайд «Видеоматериалы» (YouTube; дата обращения: 11.11.2026).