Не хватает прав!

Не хватает прав!

**1. Виды медико-биологических данных, их особенности.**

Медико-биологические данные включают клинические, генетические, биохимические, физиологические и эпидемиологические данные. Клинические данные содержат информацию о пациентах: диагнозы, симптомы, результаты осмотров. Генетические данные представляют последовательности ДНК, РНК, SNP и другие молекулярные маркеры. Биохимические данные отражают концентрации метаболитов, белков, гормонов. Физиологические данные включают показатели работы органов (ЭКГ, ЭЭГ, давление). Эпидемиологические данные охватывают заболеваемость, распространенность патологий в популяции. Особенности таких данных: неоднородность, большой объем, наличие шумов и пропусков, необходимость стандартизации и интерпретации в контексте биологических процессов.

**2. Описательные статистики и их свойства.**

Описательные статистики используются для первичного анализа данных и включают меры центральной тенденции (среднее, медиана, мода), меры изменчивости (дисперсия, стандартное отклонение, размах) и меры формы распределения (асимметрия, эксцесс). Среднее чувствительно к выбросам, медиана устойчива к ним. Дисперсия показывает разброс данных вокруг среднего, а стандартное отклонение — его масштаб. Асимметрия указывает на скошенность распределения, эксцесс — на остроту пика. Эти статистики помогают понять структуру данных перед углубленным анализом.

**3. Основные способы организации выборки.**

Выборка может быть организована как простая случайная, стратифицированная, кластерная или систематическая. Простая случайная выборка предполагает равную вероятность отбора каждого элемента. Стратифицированная делит популяцию на однородные группы (страты), из которых делается выбор. Кластерная выборка разбивает данные на кластеры, затем случайно отбирает некоторые из них. Систематическая выборка использует фиксированный интервал (например, каждый 10-й элемент). Выбор метода зависит от структуры данных и целей исследования.

**4. Методы обработки пропущенных значений.**

Пропущенные значения можно удалить (если их мало), заменить

средним/медианой (для числовых данных) или модой (для категориальных). Более сложные методы включают множественную импутацию (предсказание пропусков на основе других переменных) и использование алгоритмов машинного обучения (KNN, регрессия). Если пропуски неслучайны (MNAR), требуется специальный анализ. Важно учитывать влияние обработки пропусков на итоговые результаты.

**5. Статистическая достоверность.**

Статистическая достоверность отражает надежность выводов, обычно выражается через p-уровень (вероятность ошибки при отклонении нулевой гипотезы). Если $p < 0.05$, результат считается статистически значимым. Доверительные интервалы показывают диапазон, в котором с заданной вероятностью находится истинный параметр. Мощность теста — вероятность обнаружить эффект, если он есть. Достоверность зависит от объема выборки, размера эффекта и вариабельности данных.

**6. Критерии для проверки на нормальность распределения, условия их применения.**

Для проверки нормальности используют графические методы (Q-Q plot, гистограммы) и статистические тесты: Шапиро-Уилка (для малых выборок), Колмогорова-Смирнова (для больших выборок), Андерсона-Дарлинга (чувствителен к хвостам распределения). Тесты применяются перед использованием параметрических методов, требующих нормальности данных. Если распределение ненормальное, применяют непараметрические тесты или преобразования данных.

**7. Способы преобразования данных.**

Для нормализации данных используют логарифмирование (для правосторонней асимметрии), квадратный корень (для умеренной асимметрии), преобразование Бокса-Кокса (оптимизирует нормальность). Для категориальных данных применяют бинаризацию или one-hot encoding. Стандартизация (z-оценка) и масштабирование (Min-Max) используют для приведения данных к единому диапазону перед анализом.

**8. Критерии определения выбросов.**

Выбросы определяют с помощью межквартильного размаха (IQR: значения за пределами $Q1 - 1.5 \cdot IQR$ или $Q3 + 1.5 \cdot IQR$), z-оценок (значения $> 3$ или $< -3$),

методов машинного обучения (Isolation Forest, DBSCAN). Визуально их можно обнаружить на boxplot или scatter plot. Выбросы могут быть артефактами или важными наблюдениями, поэтому их анализ требует осторожности.

**9. Параметрическая статистика. Тесты для зависимых и независимых выборок.**

Параметрические тесты предполагают нормальность распределения и однородность дисперсий. Для независимых выборок: t-тест Стьюдента (2 группы), ANOVA (>2 групп). Для зависимых выборок: парный t-тест, повторные измерения ANOVA. Эти тесты более мощные, чем непараметрические, но требуют строгих допущений.

**10. Непараметрическая статистика. Тесты для непрерывных и дискретных выборок.**

Непараметрические тесты не требуют нормальности. Для непрерывных данных: Манна-Уитни (независимые выборки), Вилкоксона (зависимые выборки), Краскела-Уоллиса (>2 групп). Для дискретных данных: хи-квадрат, точный тест Фишера. Эти тесты устойчивы к выбросам, но менее мощные.

**11. Преимущества и недостатки непараметрических методов.**

Преимущества: не требуют нормальности, устойчивы к выбросам, применимы к порядковым данным. Недостатки: менее мощные при выполнении условий для параметрических тестов, могут терять информацию при ранжировании данных. Используются при малых выборках или нарушении допущений параметрических тестов.

**12. Таблицы сопряженности признаков.**

Таблицы сопряженности показывают совместное распределение категориальных переменных. Анализируются с помощью теста хи-квадрат (для больших выборок) или точного теста Фишера (для малых). Используются для изучения взаимосвязей между признаками, например, генетическими маркерами и заболеваниями.

**13. Корреляционный анализ. Коэффициенты корреляции, их значимость.**

Корреляция измеряет силу и направление связи между переменными. Коэффициент Пирсона (для линейной связи), Спирмена (для монотонной), Кендалла (для ранговых данных). Значимость проверяется через p-уровень.

Корреляция не означает причинность, возможны ложные связи из-за confounding-факторов.

**14. Автокорреляция.**

Автокорреляция — связь между значениями одного временного ряда с лагом. Измеряется коэффициентом автокорреляции, используется в анализе динамических процессов (ЭКГ, экспрессия генов). Может указывать на периодичность или тренды. Учитывается в моделях ARIMA.

**15. Регрессионный анализ.**

Регрессия моделирует зависимость между предикторами и откликом. Линейная регрессия — для непрерывного отклика, логистическая — для бинарного. Множественная регрессия включает несколько предикторов. Важные аспекты: проверка мультиколлинеарности, остатков, интерпретация коэффициентов.

**16. Анализ выживаемости.**

Анализирует время до наступления события (смерть, рецидив). Методы: Каплана-Мейера (кривые выживаемости), лог-ранк тест (сравнение групп). Учитывает цензурированные данные (наблюдения, где событие не произошло). Применяется в клинических исследованиях.

**17. Модель пропорциональных интенсивностей Кокса.**

Регрессионная модель для анализа выживаемости с учетом ковариат. Предполагает пропорциональность рисков. Коэффициенты показывают влияние предикторов на hazard ratio. Используется для многомерного анализа факторов риска.

**18. Многомерный анализ данных.**

Включает методы снижения размерности (PCA, t-SNE), кластеризации (k-means, иерархическая), классификации (SVM, случайный лес). Позволяет выявлять паттерны в высокоразмерных данных (например, транскриптомных).

**19. Взаимодействие между переменными (эффект модификации).**

Взаимодействие означает, что эффект одной переменной зависит от значения другой. Анализируется через включение произведения предикторов в регрессию. Пример: влияние гена на болезнь может зависеть от пола. Важно для персонализированной медицины.