**федеральное государственное автономное образовательное учреждение высшего образования**

**Первый Московский государственный медицинский университет им. И.М. СеченоваМинистерства здравоохранения Российской Федерации**

**(Сеченовский Университет)**

Кафедра Высшей математики, механики и
математического моделирования ИКНиММ НТПБ

# Методические материалы по дисциплине:

## Основы прикладной статистики и анализа данных

основная профессиональная Высшее образование - магистратура - программа магистратуры
45.04.04 Интеллектуальные системы в гуманитарной среде

# РАЗДЕЛ 1. Байесовская статистика (5 вопросов)

**1. В чем заключается основная идея байесовского подхода в статистике?**

**Ответ:** Байесовский подход рассматривает неизвестные параметры как случайные величины, имеющие собственное распределение вероятностей. В отличие от классической статистики, он объединяет априорную информацию (доопытные знания) с данными наблюдений и получает апостериорное распределение параметра.

**2. Что такое априорное распределение в байесовской статистике?**

**Ответ:** Априорное распределение — это распределение вероятностей, которое выражает наши знания о параметре до того, как мы увидели данные. Оно отражает субъективное мнение или объективную информацию, известную заранее.

**3. Что такое апостериорное распределение и как оно получается?**

**Ответ:** Апостериорное распределение — это обновленное распределение параметра после учета наблюдаемых данных. Оно получается путем объединения априорного распределения с функцией правдоподобия данных через теорему Байеса.

**4. Чем байесовский подход отличается от частотного (классического) подхода к оценке параметров?**

**Ответ:** При частотном подходе параметры считаются фиксированными неизвестными константами, а вероятность трактуется как доля событий при бесконечном повторении опыта. При байесовском подходе параметры случайны, а вероятность выражает степень уверенности исследователя. Кроме того, байесовский подход явно использует априорную информацию.

**5. Что такое сопряженное априорное распределение?**

**Ответ:** Это такое априорное распределение, которое в сочетании с определенным типом данных дает апостериорное распределение того же семейства, что и априорное. Это сильно упрощает вычисления.

---

# РАЗДЕЛ 2. Вероятность. Случайные величины (6 вопросов)

**6. Что такое пространство элементарных исходов в теории вероятностей?**

**Ответ:** Пространство элементарных исходов — это множество всех возможных результатов случайного эксперимента. Например, при бросании игрального кубика оно состоит из шести исходов $\{1, 2, 3, 4, 5, 6\}$.

**7. Чем отличаются совместное, маргинальное и условное распределения нескольких случайных величин?**

**Ответ:** Совместное распределение описывает вероятности всех комбинаций значений нескольких величин одновременно. Маргинальное распределение — это распределение

одной величины, полученное суммированием (или интегрированием) по всем остальным. Условное распределение — это распределение одной величины при фиксированном значении другой.

**8. Что такое независимость двух случайных величин?**

**Ответ:** Две случайные величины независимы, если их совместное распределение распадается в произведение их маргинальных распределений. Это означает, что знание значения одной не дает никакой информации о значении другой.

**9. Что такое ковариация между двумя случайными величинами и что она показывает?**

**Ответ:** Ковариация — это мера совместной изменчивости двух случайных величин. Положительная ковариация означает, что при увеличении одной величины другая в среднем тоже увеличивается; отрицательная — что одна растет, когда другая убывает. Нулевая ковариация указывает на отсутствие линейной связи.

**10. Чем ковариация отличается от коэффициента корреляции?**

**Ответ:** Ковариация зависит от единиц измерения величин и может быть любой по величине. Коэффициент корреляции — это нормированная версия ковариации, которая принимает значения от -1 до 1 и не зависит от масштаба, что делает его удобным для сравнения силы связи между разными парами величин.

**11. Что такое условное математическое ожидание?**

**Ответ:** Условное математическое ожидание — это ожидаемое значение одной случайной величины при условии, что другая случайная величина приняла определенное значение. Оно показывает среднее значение Y для фиксированного X.

---

# РАЗДЕЛ 3. Дескриптивная статистика (5 вопросов)

**12. Какие основные характеристики центральной тенденции используются в описательной статистике?**

**Ответ:** Основные характеристики центральной тенденции — это среднее арифметическое (сумма значений, деленная на их количество), медиана (значение, стоящее в середине упорядоченного ряда) и мода (наиболее часто встречающееся значение).

**13. В чем преимущество медианы перед средним арифметическим?**

**Ответ:** Медиана устойчива к выбросам и асимметрии распределения. Если в данных есть сильно отклоняющиеся значения, среднее может сильно искажаться, а медиана останется репрезентативной для типичного наблюдения.

**14. Какие основные меры разброса используются в описательной статистике?**

**Ответ:** Основные меры разброса — это дисперсия (средний квадрат отклонений от среднего), стандартное отклонение (корень из дисперсии), размах (разность между

максимумом и минимумом), межквартильный размах (разность между третьим и первым квартилями).

**15. Что такое квартили и для чего они используются?**

**Ответ:** Квартили — это значения, которые делят упорядоченную выборку на четыре равные части. Первый квартиль (Q1) отделяет 25% наименьших значений, третий квартиль (Q3) отделяет 75% наименьших значений. Они используются для описания распределения и выявления выбросов.

**16. Что такое ящик с усами (boxplot) и какую информацию он дает?**

**Ответ:** Боксплот — это графический способ представления распределения данных. Он показывает медиану, квартили, размах и явно выделяет выбросы. Наглядно демонстрирует симметричность или асимметричность распределения.

---

# РАЗДЕЛ 4. Задачи кластеризации (5 вопросов)

**17. Что такое кластеризация и какова ее цель?**

**Ответ:** Кластеризация — это метод машинного обучения, который группирует объекты в кластеры так, чтобы объекты внутри одного кластера были максимально похожи друг на друга, а объекты из разных кластеров — максимально различны. Цель — обнаружение скрытых групп в данных.

**18. В чем отличие кластеризации от классификации?**

**Ответ:** Классификация относится к обучению с учителем — нам известны метки классов, и мы строим правило для предсказания меток новых объектов. Кластеризация — это обучение без учителя: метки заранее неизвестны, алгоритм сам находит структуру в данных.

**19. Опишите основной принцип работы алгоритма k-средних (k-means).**

**Ответ:** Алгоритм k-средних разбивает данные на заранее заданное число кластеров. Он случайно выбирает центры кластеров, затем относит каждый объект к ближайшему центру, после чего пересчитывает центры как средние по каждому кластеру. Эти шаги повторяются до стабилизации центров.

**20. В чем слабость алгоритма k-средних?**

**Ответ:** Основные слабости: необходимо заранее знать число кластеров k; он чувствителен к начальному выбору центров и может сходиться к локальному оптимуму; он плохо работает с кластерами неправильной формы, разной плотности и при наличии выбросов.

**21. Что такое иерархическая кластеризация и чем она отличается от k-средних?**

**Ответ:** Иерархическая кластеризация строит дерево (дендрограмму), последовательно объединяя или разделяя кластеры. Она не требует заранее заданного числа кластеров,

позволяет выбрать уровень детализации и лучше подходит для данных сложной структуры, но более вычислительно затратна.

---

# РАЗДЕЛ 5. Корреляция. Модели линейной регрессии (6 вопросов)

**22. Что такое корреляция и какую связь между переменными она измеряет?**

**Ответ:** Корреляция — это статистическая мера, которая показывает наличие и силу линейной связи между двумя количественными переменными. Она не указывает на причинно-следственную связь, а лишь на взаимосвязь в данных.

**23. Что означает значение коэффициента корреляции, близкое к нулю?**

**Ответ:** Значение, близкое к нулю, означает отсутствие линейной связи между переменными. Однако это не исключает наличия нелинейной связи (например, параболической) или сложной зависимости.

**24. В чем разница между корреляцией и регрессией?**

**Ответ:** Корреляция просто измеряет силу и направление линейной связи между двумя переменными без разделения на зависимую и независимую. Регрессия строит уравнение, которое позволяет предсказывать значение одной переменной (зависимой) на основе другой или нескольких других (независимых).

**25. Что такое модель линейной регрессии?**

**Ответ:** Это статистическая модель, которая описывает связь между зависимой переменной и одной или несколькими независимыми переменными через линейное уравнение. Она предполагает, что изменение независимой переменной на единицу приводит к постоянному изменению зависимой переменной.

**26. Что такое коэффициент детерминации (R-квадрат) в линейной регрессии и как его интерпретировать?**

**Ответ:** R-квадрат — это доля дисперсии зависимой переменной, объясненная регрессионной моделью. Если R-квадрат равен 0,85, это означает, что модель объясняет 85% вариации данных, а 15% остается необъясненными (ошибками).

**27. Какие основные предположения должны выполняться для корректного применения линейной регрессии?**

**Ответ:** Основные предположения: линейность связи, независимость и нормальность распределения остатков, гомоскедастичность (постоянная дисперсия ошибок при разных значениях предикторов) и отсутствие сильной мультиколлинеарности между предикторами.

---

# РАЗДЕЛ 6. Метод максимального правдоподобия. Дисперсионный анализ (5 вопросов)

**28. Какова основная идея метода максимального правдоподобия?**

**Ответ:** Метод максимального правдоподобия состоит в выборе таких значений параметров модели, при которых вероятность получить именно те данные, которые мы наблюдали, является максимальной. То есть мы ищем параметры, при которых наша выборка становится «наиболее правдоподобной».

**29. Что такое функция правдоподобия?**

**Ответ:** Функция правдоподобия — это вероятность наблюдаемых данных, выраженная как функция неизвестных параметров модели. Она показывает, насколько вероятны данные при различных значениях параметров.

**30. В чем отличие метода максимального правдоподобия от метода наименьших квадратов?**

**Ответ:** Метод наименьших квадратов минимизирует сумму квадратов отклонений и не требует предположений о распределении ошибок. Метод максимального правдоподобия требует задания вероятностной модели данных и при нормальном распределении ошибок дает те же оценки, что и метод наименьших квадратов, но более гибок для разных распределений.

**31. Какова цель дисперсионного анализа?**

**Ответ:** Дисперсионный анализ (ANOVA) используется для сравнения средних значений трех и более групп и выявления того, влияет ли некоторый фактор (или несколько факторов) на зависимую переменную. Он оценивает, значимо ли различие между группами или оно обусловлено случайностью.

**32. В чем принципиальное отличие однофакторного и двухфакторного дисперсионного анализа?**

**Ответ:** Однофакторный дисперсионный анализ изучает влияние одного фактора на зависимую переменную. Двухфакторный анализ позволяет одновременно оценивать влияние двух факторов и их взаимодействие (то есть зависит ли эффект одного фактора от уровня другого).

---

# РАЗДЕЛ 7. Методы регуляризации. Выбор оптимальной модели (5 вопросов)

**33. Что такое регуляризация в машинном обучении и зачем она нужна?**

**Ответ:** Регуляризация — это метод, который добавляет штраф к сложности модели в процессе обучения. Она нужна для борьбы с переобучением, когда модель слишком

хорошо подстраивается под обучающие данные и теряет способность обобщения на новых данных.

**34. В чем суть метода регуляризации L1 (лассо-регрессия)?**

**Ответ:** Лассо-регрессия добавляет к функции потерь штраф, равный сумме абсолютных значений коэффициентов. Это приводит к тому, что некоторые коэффициенты становятся равными нулю, что делает модель более интерпретируемой и позволяет отбирать наиболее важные признаки.

**35. В чем суть метода регуляризации L2 (ридж-регрессия)?**

**Ответ:** Ридж-регрессия добавляет к функции потерь штраф, равный сумме квадратов коэффициентов. Это сглаживает значения коэффициентов, уменьшая их разброс и делая модель устойчивее к мультиколлинеарности, но все коэффициенты остаются ненулевыми.

**36. В чем основное различие между L1 и L2 регуляризацией с точки зрения отбора признаков?**

**Ответ:** L1-регуляризация (лассо) может обнулять коэффициенты, фактически удаляя неважные признаки из модели. L2-регуляризация (ридж) уменьшает коэффициенты, но не обнуляет их, поэтому все признаки сохраняются, но их влияние сглаживается.

**37. Что такое перекрестная проверка (кросс-валидация) и для чего она используется?**

**Ответ:** Кросс-валидация — это метод оценки качества модели, при котором данные разбиваются на несколько частей (фолдов). Модель обучается на части данных и проверяется на оставшейся, процедура повторяется несколько раз. Она используется для выбора оптимальной модели и настройки гиперпараметров без использования тестовых данных.

---

# РАЗДЕЛ 8. Программирование на R (5 вопросов)

**38. Что представляет собой R как среда для статистических вычислений?**

**Ответ:** R — это язык программирования и среда с открытым исходным кодом, предназначенная для статистической обработки данных, анализа данных и визуализации. Он широко используется в науке, финансах, биоинформатике и машинном обучении благодаря огромному количеству пакетов.

**39. Как в R называются основные структуры данных?**

**Ответ:** Основные структуры данных в R: вектор (одномерный массив одного типа), матрица (двумерный массив одного типа), список (может содержать элементы разных типов), дата-фрейм (таблица с колонками разных типов, аналогичная электронной таблице).

**40. Что такое пакеты в R и как их установить?**

**Ответ:** Пакеты — это наборы функций, данных и документации, расширяющие возможности базового R. Они устанавливаются командой install.packages("имя_пакета") и подключаются командой library(имя_пакета).

**41. Для чего используется функция ggplot2 в R?**

**Ответ:** ggplot2 — это мощный пакет для создания статической графики в R, основанный на грамматике графики. Он позволяет строить сложные многослойные графики с высокой степенью настройки.

**42. Что такое tidyverse и какие пакеты в него входят?**

**Ответ:** Tidyverse — это коллекция пакетов для обработки и анализа данных, объединенных общей философией «чистых данных». В него входят: dplyr (манипуляция данными), tidyr (приведение данных), ggplot2 (визуализация), readr (импорт данных) и другие.

---

# РАЗДЕЛ 9. Статистическое тестирование гипотез (8 вопросов)

**43. Что такое нулевая гипотеза в статистическом тестировании?**

**Ответ:** Нулевая гипотеза (H0) — это исходное предположение об отсутствии эффекта, различий или связи. Например, что средние значения двух групп равны, или что коэффициент регрессии равен нулю. Это утверждение, которое мы пытаемся опровергнуть.

**44. Что такое альтернативная гипотеза?**

**Ответ:** Альтернативная гипотеза (H1) — это утверждение, противоположное нулевой гипотезе, которое принимается, если нулевая гипотеза отвергается. Например, что средние значения групп не равны, или что коэффициент регрессии отличается от нуля.

**45. Что такое p-значение (p-value) и как его интерпретировать?**

**Ответ:** P-значение — это вероятность получить наблюдаемые или более экстремальные результаты при условии, что нулевая гипотеза верна. Если p-значение меньше заранее заданного уровня значимости, мы отвергаем нулевую гипотезу.

**46. Что такое уровень значимости (α) и как его выбирают?**

**Ответ:** Уровень значимости — это вероятность ошибки первого рода, то есть вероятность отвергнуть нулевую гипотезу, когда она на самом деле верна. Обычно выбирают $\alpha = 0,05$ или $\alpha = 0,01$ в зависимости от требуемой строгости исследования.

**47. В чем разница между ошибкой первого и второго рода?**

**Ответ:** Ошибка первого рода — это ложное отклонение нулевой гипотезы (ложное обнаружение эффекта). Ошибка второго рода — это неверное принятие нулевой гипотезы, когда на самом деле верна альтернативная (пропуск реального эффекта).

**48. Что такое мощность статистического теста?**

**Ответ:** Мощность теста — это вероятность правильно отвергнуть нулевую гипотезу, когда на самом деле верна альтернативная. Она равна единице минус вероятность ошибки второго рода. Чем выше мощность, тем лучше тест обнаруживает реальные эффекты.

**49. В чем разница между параметрическими и непараметрическими тестами?**

**Ответ:** Параметрические тесты предполагают, что данные имеют определенное распределение (чаще всего нормальное) и опираются на параметры этого распределения. Непараметрические тесты не делают предположений о распределении данных и работают с рангами или порядками, что делает их более устойчивыми, но часто менее мощными.

**50. Для чего используется t-тест (критерий Стьюдента)?**

**Ответ:** Т-тест используется для проверки гипотезы о равенстве средних значений в двух группах. Применяется в трех вариантах: для независимых выборок, для зависимых (парных) выборок и для сравнения среднего выборки с известным теоретическим значением.