Урок 2. Описательные статистики

Хакимов Р.И. + ChatGPT

Статистические методы анализа данных

Статистические методы анализа данных представляют собой набор техник и процедур, которые позволяют исследователям изучать, интерпретировать и делать выводы из данных. Эти методы варьируются от простых описательных статистик до сложных моделей и алгоритмов. Вот основные категории и примеры статистических методов анализа данных:

Статистические методы анализа данных

Описательная статистика

Цель: Обобщение и описание характеристик данных.

Основные методы:

Среднее значение (среднее арифметическое): Среднее всех значений в наборе данных.

Медиана: Центральное значение в упорядоченном наборе данных.

Мода: Наиболее часто встречающееся значение.

Размах: Разница между максимальным и минимальным значениями.

Описательная статистика

Основные методы:

Дисперсия и стандартное отклонение: Измерение разброса данных относительно среднего значения.

Квартиль: Деление данных на четыре равные части, позволяющее понять распределение данных.

Статистические методы анализа данных

Инференциальная статистика / Статистический метод

Цель: Сделать выводы о генеральной совокупности на основе данных выборки.

Основные методы:

Доверительный интервал: Интервал, в котором с определенной вероятностью находится истинное значение параметра генеральной совокупности.

Тестирование гипотез: Проверка предположений о параметрах генеральной совокупности на основе выборочных данных.

t-тест: Используется для сравнения средних значений двух групп.

ANOVA (дисперсионный анализ): Для сравнения средних значений более чем двух групп.

χ^2*-тест (хи-квадрат тест):* Для проверки взаимосвязи между двумя категориальными переменными.

Инференциальная статистика / Статистический метод

Основные методы:

Регрессионный анализ: Метод, позволяющий оценить влияние одной или нескольких независимых переменных на зависимую переменную.

Линейная регрессия: Модель, предполагающая линейную связь между переменными.

Логистическая регрессия: Используется для моделирования вероятности бинарного исхода.

Статистические методы анализа данных

Корреляционный анализ

Цель: Оценка степени и направления связи между двумя переменными.

Основные методы:

Коэффициент корреляции Пирсона: Измеряет линейную зависимость между двумя количественными переменными.

Коэффициент корреляции Спирмена: Не параметрическая мера связи между ранжированными переменными.

Ковариация: Показатель, описывающий, как две переменные изменяются вместе.

Статистические методы анализа данных

Многомерные методы анализа

Цель: Изучение зависимости и связей между несколькими переменными одновременно.

Основные методы:

Множественная регрессия: Расширение линейной регрессии для случая, когда есть несколько независимых переменных.

Кластерный анализ: Метод группировки объектов или наблюдений в группы (кластеры) на основе сходства между ними.

Статистические методы анализа данных

Временные ряды

Цель: Анализ данных, упорядоченных во времени.

Основные методы:

Автокорреляция: Измерение зависимости данных от своих прошлых значений.

Модели ARIMA (авторегрессия, интегрированная модель скользящего среднего): Для прогнозирования временных рядов.

Экспоненциальное сглаживание: Метод для сглаживания временных рядов и прогноза будущих значений.

Статистические методы анализа данных

Байесовские методы

Цель: Обновление вероятности гипотезы на основе новых данных.

Основные методы:

Байесовская регрессия: Обновление распределения параметров модели по мере поступления новых данных.

Байесовская сеть: Графическая модель, которая представляет зависимость между переменными.

Статистические методы анализа данных

Методы машинного обучения и статистическое обучение

Цель: Создание моделей, которые могут предсказывать или классифицировать данные на основе обучения на существующих данных.

Основные методы:

Методы классификации: Логистическая регрессия, деревья решений, случайные леса, нейронные сети.

Методы регрессии: Линейная и нелинейная регрессия, опорные векторы.

Методы кластеризации: k-средних, иерархическая кластеризация.

Статистические методы анализа данных

Эти методы и техники позволяют исследователям и аналитикам собирать, анализировать и интерпретировать данные, делая обоснованные выводы и принимая решения. Выбор конкретного метода зависит от целей исследования, природы данных и гипотез, которые нужно проверить.

Меры центральной тенденции: среднее значение, медиана, мода

Меры центральной тенденции — это статистические показатели, которые описывают центр распределения данных. Они помогают понять, где сосредоточены значения в наборе данных. Вот основные меры центральной тенденции:

Меры центральной тенденции: среднее значение, медиана, мода

Среднее значение (среднее арифметическое)

Описание: Среднее значение рассчитывается как сумма всех значений в наборе данных, деленная на количество этих значений. Это наиболее распространенная мера центральной тенденции.

Формула:

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

где $\bar{X}$ — среднее значение, X_i — отдельные значения, n — количество значений.

Пример: Если у вас есть набор данных: 4, 7, 8, 10, то среднее значение будет:

$$\bar{X} = \frac{4 + 7 + 8 + 10}{4} = 7.25$$

Особенности: Среднее значение чувствительно к экстремальным значениям (выбросам), что может искажать представление о центральной тенденции в случае наличия аномально больших или маленьких значений.

Меры центральной тенденции: среднее значение, медиана, мода

Медиана

Описание: Медиана — это значение, которое делит упорядоченный набор данных пополам. Половина значений в наборе меньше медианы, а другая половина больше.

Формула:

- Для нечетного числа значений медиана — это срединное значение.
- Для четного числа значений медиана — это среднее арифметическое двух срединных значений.

Пример:

- Если есть набор данных: 3, 7, 8, 12, 14, то медиана будет 8 (третье значение в упорядоченном наборе).
- Если есть набор данных: 3, 7, 8, 12, 14, 20, то медиана будет $\frac{8+12}{2} = 10$ (среднее арифметическое двух срединных значений).

Особенности: Медиана не чувствительна к выбросам и экстраординарным значениям, поэтому она лучше отражает центральную тенденцию в случаях, когда данные имеют сильные отклонения.

Меры центральной тенденции: среднее значение, медиана, мода

Мода

Описание: Мода — это значение, которое встречается в наборе данных наиболее часто. В отличие от среднего значения и медианы, мода может быть не уникальной: набор данных может иметь несколько мод (многомодальный) или не иметь мод вообще.

Пример:

- Если есть набор данных: 1, 2, 2, 3, 4, то мода будет 2 (значение, которое встречается чаще всего).
- Если есть набор данных: 1, 2, 3, 4, 5, то в этом случае мода отсутствует, так как все значения встречаются одинаково часто.

Особенности: Мода полезна для категориальных данных, где вычисление среднего значения или медианы может быть неуместным. Также мода может давать представление о наиболее частом событии или явлении в данных.

Меры центральной тенденции: среднее значение, медиана, мода

Сравнение методов

Среднее значение часто используется, когда данные распределены нормально и нет значительных выбросов. Оно предоставляет хорошее представление о "центре"данных.

Медиана предпочтительна, когда данные имеют выбросы или распределение несимметрично, так как она не искажается экстремальными значениями.

Мода полезна для категориальных данных и для понимания наиболее частых значений в данных, но может быть менее информативной для количественных данных, особенно если значения распределены равномерно.

В разных ситуациях различные меры центральной тенденции могут предоставлять более полезную информацию, поэтому важно учитывать особенности данных при выборе подходящего метода.

Меры изменчивости: диапазон, дисперсия, стандартное отклонение

Меры изменчивости помогают оценить, насколько данные в наборе варьируются или отклоняются от центра распределения. Они дают представление о том, насколько значения данных разбросаны вокруг центральной тенденции (например, среднего значения). Вот основные меры изменчивости:

Меры изменчивости: диапазон, дисперсия, стандартное отклонение

Диапазон, размах

Описание: Диапазон — это разница между максимальным и минимальным значениями в наборе данных. Это самая простая мера изменчивости.

Формула:

$$\text{Диапазон} = X_{\text{max}} - X_{\text{min}}$$

где X_{max} — максимальное значение, а X_{min} — минимальное значение.

Пример: Для набора данных 5, 8, 12, 15 диапазон будет:

$$15 - 5 = 10$$

Особенности: Диапазон прост в расчетах и интерпретации, но он чувствителен к выбросам, так как зависит только от крайних значений.

Меры изменчивости: диапазон, дисперсия, стандартное отклонение

Дисперсия

Описание: Дисперсия — это среднее значение квадратов отклонений наблюдений от их среднего значения. Она показывает, насколько сильно значения данных отклоняются от среднего.

Формула:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})^2$$

где σ^2 — дисперсия, X_i — отдельные значения, $\bar{X}$ — среднее значение, N — количество значений.

Для выборки формула дисперсии немного изменяется:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

где s^2 — выборочная дисперсия, n — размер выборки.

Меры изменчивости: диапазон, дисперсия, стандартное отклонение

Дисперсия

Пример: Для набора данных 4, 8, 6, 5, 9:

$$\bar{X} = \frac{4+8+6+5+9}{5} = 6.4$$

дисперсия равна:

$$\sigma^2 = \frac{(4-6.4)^2 + (8-6.4)^2 + (6-6.4)^2 + (5-6.4)^2 + (9-6.4)^2}{5}$$

$$= \frac{5.76 + 2.56 + 0.16 + 1.96 + 6.76}{5} = 3.44$$

Особенности: Дисперсия выражается в квадрате единиц измерения исходных данных, что может затруднять интерпретацию.

Меры изменчивости: диапазон, дисперсия, стандартное отклонение

Стандартное отклонение

Описание: Стандартное отклонение — это квадратный корень из дисперсии. Оно измеряет среднее отклонение значений от среднего значения и выражается в тех же единицах измерения, что и исходные данные.

Формула:

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})^2}$$

где σ — стандартное отклонение, X_i — отдельные значения, $\bar{X}$ — среднее значение, N — количество значений.

Для выборки:

$$\sigma = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

где σ — стандартное отклонение, n — размер выборки.

Меры изменчивости: диапазон, дисперсия, стандартное отклонение

Стандартное отклонение

Пример: Для вышеупомянутого набора данных, если дисперсия равна 3.44, то стандартное отклонение будет:

$$\sigma = \sqrt{3.44} \approx 1.85$$

Особенности: Стандартное отклонение более интуитивно понятно, чем дисперсия, так как оно выражается в тех же единицах измерения, что и исходные данные.

Меры изменчивости: диапазон, дисперсия, стандартное отклонение

Сравнение мер изменчивости

Диапазон предоставляет общую информацию о разбросе данных, но не учитывает распределение внутри интервала и чувствителен к выбросам.

Дисперсия дает более полное представление о том, насколько данные варьируются вокруг среднего значения, но ее интерпретация может быть затруднена из-за квадратов единиц измерения.

Стандартное отклонение является более удобной мерой для интерпретации, так как оно находится в тех же единицах, что и данные, и дает представление о средней степени отклонения значений от среднего.

Эти меры помогают исследователям и аналитикам понять, насколько данные разбросаны и как это может влиять на интерпретацию результатов и выводы из анализа.