

Урок 1. Введение в статистику и основные понятия

Хакимов Р.И. + ChatGPT

Что такое статистика?

Статистика – это наука, которая занимается сбором, обработкой, анализом и интерпретацией данных. Она применяется во многих сферах жизни, включая экономику, медицину, социологию, бизнес, науку и многие другие..

Значение статистики в современном мире. Примеры использования

Статистика играет ключевую роль в современном мире, оказывая влияние на самые разные аспекты жизни. Вот несколько примеров ее использования:

Экономика и бизнес

Анализ рынка: Компании используют статистические методы для анализа рынка, определения целевой аудитории, оценки спроса и предложения, а также для прогнозирования продаж. Например, статистические данные могут помочь понять, какие продукты пользуются наибольшим спросом в определенных регионах.

Оценка рисков: В финансовом секторе статистика помогает оценивать риски инвестиций, разрабатывать стратегии управления портфелем и прогнозировать экономические тенденции.

Значение статистики в современном мире. Примеры использования

Медицина и здравоохранение

Эпидемиология: Статистика используется для анализа распространения заболеваний, оценки эффективности вакцин и лекарств, а также для прогнозирования возможных вспышек инфекций. Например, во время пандемии COVID-19 статистические модели использовались для прогнозирования распространения вируса и планирования мер по его сдерживанию.

Клинические исследования: Статистические методы применяются для оценки эффективности новых медицинских препаратов и методов лечения на основе данных клинических испытаний.

Значение статистики в современном мире. Примеры использования

Наука и исследования

Анализ данных: В науке статистические методы используются для анализа экспериментальных данных, выявления закономерностей и проверки гипотез. Например, в генетике статистика помогает в анализе генетических данных для понимания наследственных заболеваний.

Метеорология: Статистика играет важную роль в прогнозировании погоды, анализе климатических изменений и моделировании будущих климатических условий.

Значение статистики в современном мире. Примеры использования

Спорт

Анализ эффективности игроков: В спортивной аналитике статистика используется для оценки эффективности игроков и команд, разработки стратегий игры и прогнозирования результатов матчей.

Фэнтези-спорт: Статистические данные о производительности игроков используются в фэнтези-спорте для выбора команд и прогнозирования их успеха.

Основные методы сбора данных

Основные методы сбора данных в статистике включают различные подходы, которые позволяют получать информацию для последующего анализа. Эти методы могут быть качественными или количественными, в зависимости от цели исследования. Вот основные из них:

Наблюдение

Описание: Метод, при котором данные собираются через прямое или косвенное наблюдение за объектами, событиями или процессами без вмешательства в их естественное течение.

Пример: Наблюдение за поведением покупателей в магазине для определения их предпочтений.

Основные методы сбора данных

Опрос

Описание: Метод, при котором данные собираются с помощью анкет, интервью или тестов. Опросы могут быть структурированными (с заранее подготовленными вопросами) или неструктурированными (открытые вопросы).

Пример: Проведение анкетирования среди населения для выяснения уровня удовлетворенности услугами.

Основные методы сбора данных

Эксперимент

Описание: Метод, при котором исследователь активно вмешивается в процесс для изучения причинно-следственных связей. В ходе эксперимента изменяются определенные переменные, а затем наблюдаются эффекты этих изменений.

Пример: Тестирование новой формулы лекарства на группе пациентов, чтобы определить его эффективность.

Основные методы сбора данных

Анализ документов и административных данных

Описание: Сбор данных из уже существующих источников, таких как официальные документы, отчеты, административные базы данных, реестры и т. д.

Пример: Использование данных переписи населения или статистики преступности для анализа демографических тенденций.

Основные методы сбора данных

Моделирование и симуляция

Описание: Этот метод включает создание математических или компьютерных моделей, которые могут имитировать реальные процессы или явления. Данные собираются на основе выполнения этих моделей.

Пример: Симуляция климатических изменений для прогнозирования их влияния на окружающую среду.

Основные методы сбора данных

Фокус-группы

Описание: Метод, при котором группа людей собирается для обсуждения определенной темы, чтобы выявить мнения, убеждения и отношения участников.

Пример: Обсуждение новой рекламной кампании с группой потребителей для оценки их реакции на неё.

Основные методы сбора данных

Панельное исследование

Описание: Длительное исследование, при котором данные собираются от одной и той же группы людей (панели) в течение определенного времени. Это позволяет отслеживать изменения во времени.

Пример: Исследование потребительского поведения на протяжении нескольких лет для выявления долгосрочных тенденций.

Основные методы сбора данных

Биг дата и автоматизированный сбор данных

Описание: Современные методы, включающие использование больших объемов данных (big data), которые автоматически собираются с различных источников, таких как интернет, датчики, системы мониторинга и т. д.

Пример: Анализ данных с социальных сетей для предсказания трендов в обществе.

Каждый из этих методов имеет свои преимущества и ограничения и применяется в зависимости от целей и контекста исследования.

Выборка и генеральная совокупность

В статистике выборка и генеральная совокупность являются ключевыми понятиями, которые важны для проведения любого исследования и анализа данных.

Генеральная совокупность

Генеральная совокупность — это полное множество объектов, элементов или наблюдений, которые представляют интерес для исследования. Она включает в себя все возможные элементы, соответствующие определённым критериям, и может быть как конечной, так и бесконечной.

Пример: Если исследование направлено на изучение уровня доходов жителей страны, то генеральной совокупностью будут все жители этой страны.

Выборка и генеральная совокупность

Выборка

Выборка — это подмножество генеральной совокупности, которое отбирается для анализа. Выборка используется для того, чтобы сделать выводы о генеральной совокупности, поскольку в большинстве случаев исследовать всю генеральную совокупность невозможно или непрактично.

Пример: Если генеральная совокупность — это все жители страны, то выборка может включать, например, 1000 человек, случайным образом отобранных для проведения опроса.

Выборка и генеральная совокупность

Важные аспекты выборки:

Репрезентативность: Выборка должна быть репрезентативной, то есть отражать характеристики генеральной совокупности. Это важно для того, чтобы результаты анализа выборки могли быть обоснованно распространены на всю генеральную совокупность.

Размер выборки: Размер выборки должен быть достаточно большим, чтобы обеспечить точность оценок. Однако, он также должен быть сбалансирован с ресурсами, доступными для исследования.

Выборка и генеральная совокупность

Методы отбора выборки:

Случайная выборка: Каждый элемент генеральной совокупности имеет равную вероятность быть включенным в выборку. Это минимизирует предвзятость и увеличивает репрезентативность.

Стратифицированная выборка: Генеральная совокупность делится на подгруппы (страты), и из каждой подгруппы случайным образом отбирается выборка. Это помогает улучшить репрезентативность по важным характеристикам.

Выборка и генеральная совокупность

Методы отбора выборки:

Кластерная выборка: Генеральная совокупность делится на группы (кластеры), и случайным образом выбираются целые кластеры для исследования. Этот метод удобен, если невозможно собрать данные по всей генеральной совокупности.

Систематическая выборка: Элементы выборки отбираются через равные промежутки времени или пространства из упорядоченного списка. Например, каждый десятый человек из списка населения.

Выборка и генеральная совокупность

Ошибки выборки:

При работе с выборкой возможно возникновение ошибок:

Ошибка выборки: Разница между характеристиками выборки и генеральной совокупности. Она может возникнуть из-за случайных отклонений или систематических предвзятостей.

Предвзятость выборки: Систематическая ошибка, которая возникает, если выборка не является репрезентативной, например, если определённые группы населения недопредставлены.

Выборка и генеральная совокупность

Применение в статистическом анализе:

Исследователи используют данные выборки для оценки параметров генеральной совокупности, таких как среднее значение, дисперсия и т. д. Для этого применяются статистические методы, позволяющие сделать выводы о всей генеральной совокупности на основе анализа данных выборки.

Пример: Допустим, исследователь хочет узнать средний доход жителей города с населением 100 000 человек (генеральная совокупность). Проведение опроса всех жителей слишком затратно, поэтому он выбирает случайную выборку из 1000 человек. Средний доход в этой выборке может быть использован для оценки среднего дохода всего населения города, если выборка репрезентативна и достаточна по размеру.

Выборка и генеральная совокупность

Таким образом, выборка и генеральная совокупность — фундаментальные понятия в статистике, которые позволяют исследователям проводить анализ данных и делать выводы, применимые к большим группам объектов или людей.

Частоты. Распределения

В статистике частоты и распределения играют важную роль в описании и анализе данных. Эти понятия помогают структурировать данные и понять, как часто встречаются определённые значения или группы значений в выборке или генеральной совокупности.

Частоты

Частота (или абсолютная частота) — это количество раз, которое определённое значение или категория встречается в наборе данных.

Пример: В группе из 100 студентов 25 человек имеют оценку 5. В этом случае частота оценки 5 составляет 25.

Частоты. Распределения

Относительная частота — это доля или процент от общего числа наблюдений, соответствующих определённому значению.

Пример: Если из 100 студентов 25 человек получили оценку 5, то относительная частота этой оценки составляет $25/100 = 0,25$ или 25%.

Частоты. Распределения

Распределения

Распределение — это способ представления данных, показывающий, как часто встречаются различные значения или категории в выборке или генеральной совокупности. Распределение может быть представлено в виде таблиц, графиков или формул.

Типы распределений:

Дискретное распределение частот: Применяется для данных, которые принимают определённые, фиксированные значения (например, количество детей в семье).

Пример: В исследовании количества детей в семьях выборка может содержать семьи с 0, 1, 2, 3 и 4 детьми. Распределение покажет, сколько семей имеют каждый из этих вариантов.

Частоты. Распределения

Типы распределений:

Непрерывное распределение частот: Применяется для данных, которые могут принимать любые значения в определённом интервале (например, рост, вес).

Пример: Распределение роста среди студентов может показывать, что большинство студентов имеют рост в диапазоне 160-170 см, а менее распространённые значения находятся за пределами этого интервала.

Распределение частот по категориям: Применяется для категориальных данных (например, цвета, виды).

Пример: В исследовании предпочтений в цвете автомобилей распределение может показать, сколько людей предпочитают красный, синий, чёрный и другие цвета.

Частоты. Распределения

Гистограмма и полигон частот

Гистограмма — графическое представление распределения частот, где на горизонтальной оси откладываются интервалы значений, а на вертикальной — частоты (или относительные частоты). Высота столбцов показывает, сколько раз значения попадают в каждый интервал.

Пример: Гистограмма роста студентов может иметь столбцы, показывающие количество студентов в интервалах 150-160 см, 160-170 см и т. д.

Полигон частот — это линия, соединяющая точки, которые представляют частоты для каждого значения или интервала. Полигон частот обычно строится по центрам интервалов на гистограмме.