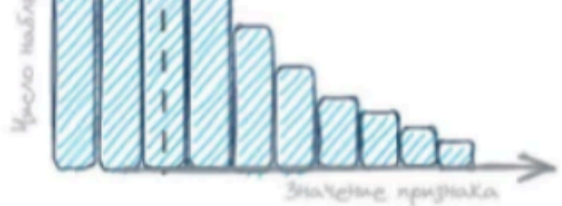
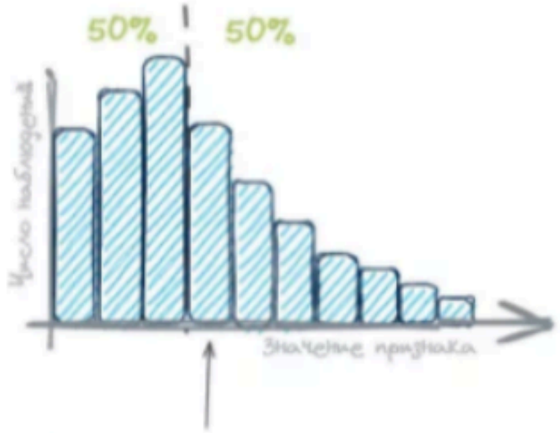
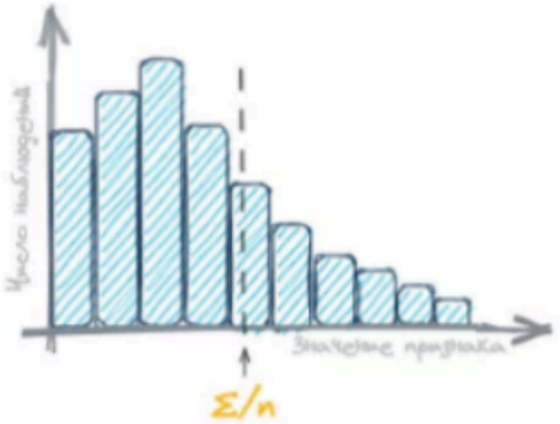






Статистика, введение

Выборка и генеральная совокупность.

Методы формирования выборки.

Дискретные и номинативные переменные.

Меры центральной тенденции.

Цикл **PPDAC**.

Цифры сами по себе не умеют говорить. Именно мы говорим за них. Мы наполняем их смыслом.

Нейт Сильвер,“Сигнал и шум”

Статистика — наука, которая изучает массовые явления, определяет их развитие и ищет связи между ними. Она занимается сбором, анализом, интерпретацией и организацией данных. Это обширная область знаний, которая охватывает разные предметы

~~Есть лож, есть наглая лож, есть статистика~~

Сбор данных – процесс получения информации из различных источников, включая опросы, датчики, базы данных или веб-сайты.

Анализ данных – процесс изучения, обработки и выявления закономерностей в собранной информации.

Интерпретация данных – объяснение и представление результатов анализа в понятной форме.

Организация данных – упорядочивание и структурирование данных для их удобного хранения и использования.

Генеральная совокупность

Генеральная совокупность — это множество всех объектов, относительно которых предполагается делать выводы в рамках конкретного исследования.

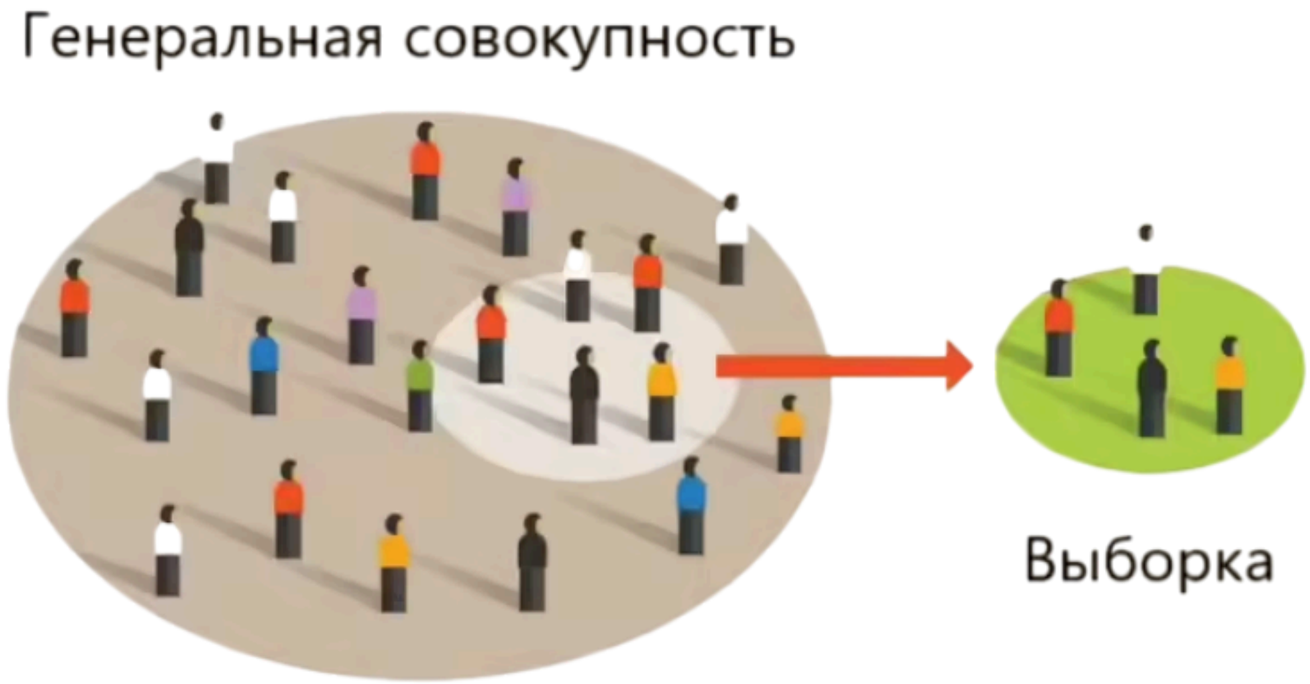
Пример 1 — мы хотим узнать средний рост жителей города Душанбе. Все жители города Душанбе без исключения будут представлять для нас генеральную совокупность.

Пример 2 — если мы врачи и тестируем лекарство от какого-либо заболевания, то все пациенты, страдающие данной болезнью, будут представлять для нас

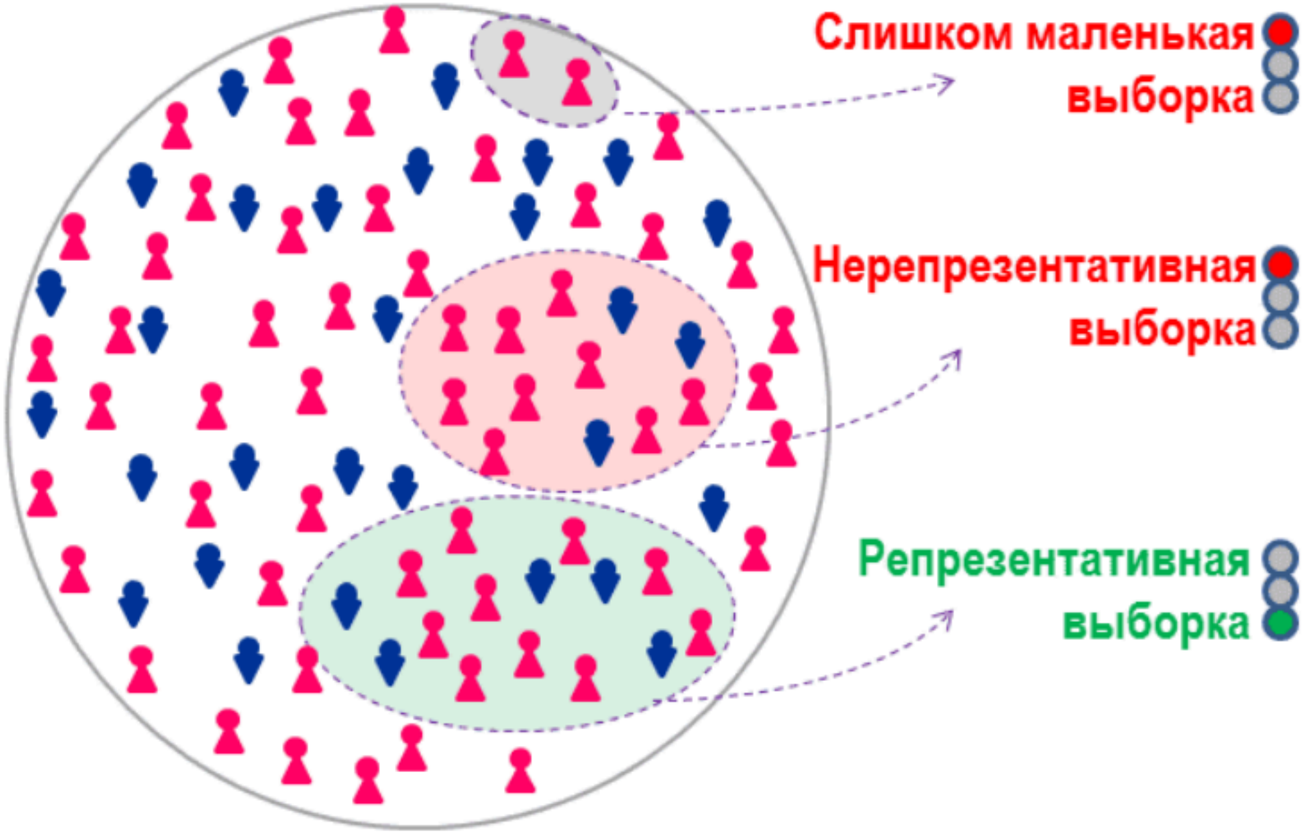
генеральную совокупность.

Выборка

Выборка — это часть генеральной совокупности, отражающая ее свойства.



Репр



Репрезентативная выборка – это такая

выборка, в которой все основные

признаки генеральной совокупности

представлены приблизительно в той же

пропорции или с той же частотой, что и

в самой генеральной совокупности.

Но мы не знаем, совпадает ли распределение отдельных характеристик выборки с генеральной совокупностью.

Иными словами, можно ли доверять выборке?

Это действительно важный вопрос, но для ответа на него с учётом всех нюансов потребуется углубиться в теорию вероятностей.

Поэтому вот упрощённый ответ: можно только предполагать, что они соответствуют друг другу. Но чем больше данных мы соберём, тем более похожим будет распределение выборки и генеральной совокупности (это одно из следствий центральной предельной теоремы).

Если это объяснение вас не удовлетворило и вы всё же хотите забраться в дебри, то переходите по ссылке.

https://onlinestatbook.com/2/normal_distribution/intro.html

Методы формирования выборки

Есть несколько способов собрать репрезентативную выборку:

Простая случайная выборка (simple random sample)

Стратифицированная выборка (stratified sample)

Групповая выборка (cluster sample)

Простая случайная выборка (simple random sample)

Случайным образом выбираем объекты из нашей генеральной совокупности. При этом, чем больше случайных объектов мы выбираем,

тем лучше выборка отражает свойства генеральной совокупности.

Например: идём в людное место города Душанбе и опрашиваем всех людей, которые нам попадаются. Среди опрошенных будут люди разного роста, возраста, пола и в разных пропорциях.

Стратифицированная выборка (stratified sample)

Разделяем нашу генеральную совокупность на группы (страты) на основе определённого признака/признаков.

Чтобы эти группы были равновероятно представлены в выборке, берём случайным образом элементы из каждой группы с равной вероятностью.

Например: делим жителей города на группы по возрасту и полу. Идём в группу

“Мужчины” возраста “30-40 лет”, случайно опрашиваем представителей данной группы, затем идём в группу “Женщины” возраста “20-30 лет”, случайно опрашиваем представителей данной группы и т.д.

В таблице суммируются принципиальные различия между случайной и стратифицированной выборками:

Простая случайная выборка	Стратифицированная выборка (страты)
Выбираем элементы из генеральной совокупности случайным образом	Мы уже на основе определённых признаков разделили нашу генеральную совокупность и добавляем в каждую подгруппу примерно равное количество элементов. Так наша выборка будет хорошо отражать особенности генеральной совокупности.
Чем больше элементов мы берём из генеральной совокупности, тем лучше выборка отражает её особенности	Выбираем элементы из каждой группы

Групповая выборка (cluster sample)

Делим нашу генеральную совокупность на группы, но эти группы должны быть относительно похожи между собой (в качестве примера можно взять районы Душанбе и считать, что в них примерно одинаковое число жителей)

Выбираем только те группы, которые нас интересуют.

Из выбранных групп случайным образом выбираем элементы.

Чтобы лучше понять, чем стратифицированная выборка отличается от групповой, рассмотрим таблицу:

Стратифицированная выборка Групповая выборка

Выбираем элементы из каждой группы (страты)

Внутри группы элементы однородны, а между группами — различаются

Выбираем элементы только из выбранных групп (страт)

В пределах группы элементы разнородны, но при этом все группы имеют схожесть

Повышает точность

Повышает эффективность выборки, уменьшая её стоимость

Насколько важна репрезентативность?

Сбор репрезентативной выборки — это нетривиальная задача, которая включает в себя выбор метода и параметров сбора (например, подбор страт).

Аккуратно собранная выборка — обязательное условие для проведения дальнейшего исследования. Использование нерепрезентативных данных приводит к ложным или неполным выводам, поэтому крайне важно обращать внимание на данные, на которых проводилось то или иное исследование.

Типы переменных

Разумеется, мы формируем выборку не просто так — нас интересуют определённые характеристики генеральной совокупности, которые мы решили исследовать с помощью нашей выборки.

Переменные — это то, что вы измеряете, манипулируете и контролируете в статистике и исследованиях. Все исследования анализируют переменные, которые могут описывать человека, место, вещь или идею.

В целом все типы переменных, с которыми так или иначе мы столкнёмся, можно

разделить на две большие группы:

Количественные

непрерывные

дискретные

Номинативные

(категориальные

)

Номинативные переменные

Номинативные (также называются категориальные) — это цвет, пол, категория продукции, профессия и всё то, что не выразить цифрами. С ними можно проводить статистические операции, например, **подсчет частоты** и **доли**.

Информационные технологии 170 24%

Маркетинг 203 29%

Медицина 162 22%

Финансы 174 25%

Итого 709 100%

Количественные переменные

Количественные переменные представляют собой непосредственное измеренное значение некоторого признака.
Если наша переменная может принимать абсолютно любое значение на некотором промежутке, то такая переменная будет называться непрерывной.

Непрерывная (рост наших респондентов) - [160.7, 174.5, 168.9, 184.3, 178.6]

В случае дискретных количественных переменных мы ожидаем, что переменные будут принимать только определённые значения.

Например, если мы считаем количество детей в семье, то, скорее всего, полученная информация будет выглядеть так: [0, 1, 2, 3] ребёнка в семье.

И 3.5 ребёнка в семье как минимум вызовет у нас некоторые подозрения.

Меры центральной тенденции

С количественными переменными можно производить гораздо больше статистических изысканий, чем с номинативными.

Например, для описания количественных переменных одним числом, мы можем использовать одну из мер центральной тенденции.

Мода

Медиана

Среднее арифметическое

Мода

Мода – значение, которое встречается чаще всего среди остальных значений совокупности.

В нашем числовом ряде [2, 7, 2, 4, 2, 6, 7, 7, 3, 7] модой является цифра 7. Она встречается четыре раза.

Цифра Количество 7 (мода) 4 6 1

4 1

3 1

2 3

Медиана

Медиана – значение, которое находится посередине совокупности, упорядоченной по возрастанию или убыванию. Она разделяет исследуемую совокупность на две равные части.

Если мы возьмем все тот же числовой ряд, по которому мы рассчитали моду, и отсортируем его по возрастанию, то теперь он будет выглядеть следующим образом:

[2, 2, 2, 3, 4, 6, 7, 7, 7, 7]

Поскольку у нас четное количество данных (10 чисел), то посередине у нас находится не одно число, а два. В таком случае вычислять медиану мы будем через среднее значение 5-го и 6-го элементов: (4 + 6) / 2 = 5.

Среднее арифметическое

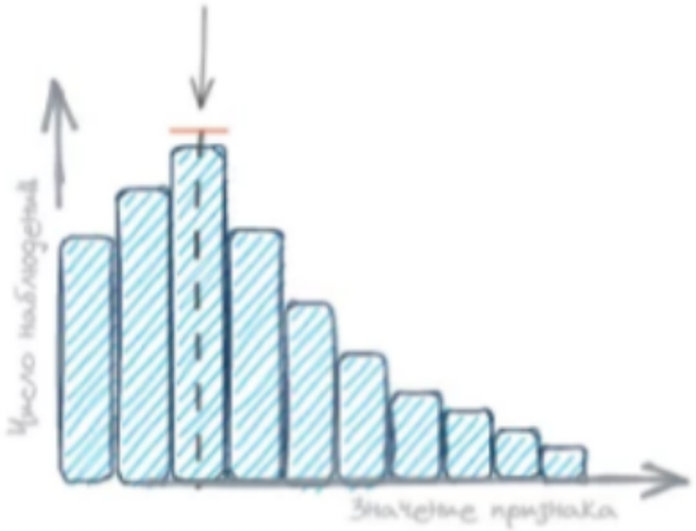
Среднее арифметическое значение – одна из самых популярных величин, используемая в качестве меры центральной тенденции. Вычисляется следующим образом: сумма всех переменных разделенная на их количество.

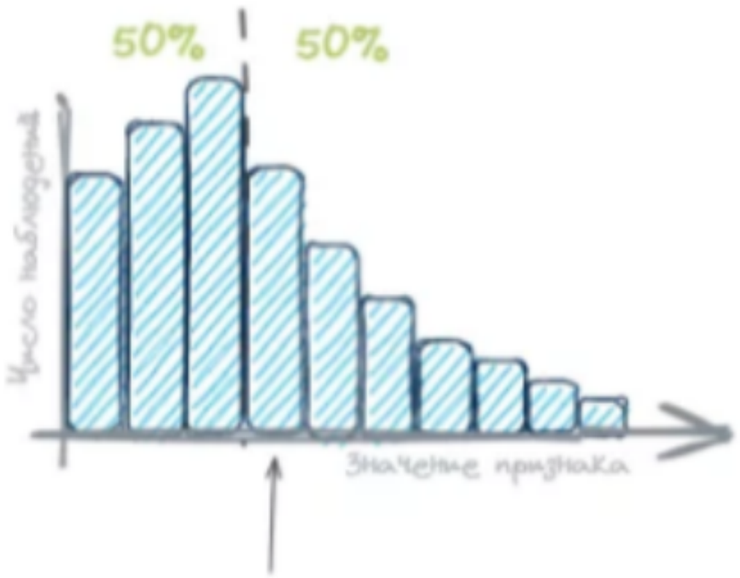
Например в числовом ряде [2, 2, 2, 3, 4, 6, 7, 7, 7, 7] среднее мы посчитаем так: (2 + 2 + 2 + 3 + 4 + 6 + 7 + 7 + 7 + 7) / 10 = 4,7.

Важный аспект: среднее крайне неустойчиво к выбросам, т.е. чрезмерно высоким/низким значениям.

Классическим примером является показатель средней заработной платы, когда экстремально высокие зарплаты маскируют реальное положение дел. В этом случае, более объективной величиной будет медиана, а не среднее.

Графическое представление





МЕДИАНА СРЕДНЕЕ МОДА

Аналитические циклы

Техника, будь то Excel, SQL, языки программирования или BI-инструменты – это всё очень увлекательные вещи, но всякое серьезное дело начинается с

методологии.

В аналитике тоже существуют свои циклы, которые превращают аналитический процесс из кустарного производства в конвейерное.

Одним из аналитических циклов которые мы рассмотрим, будет цикл - PPDAC

P - Problem

P - Plan

D - Data

A - Analysis

C - Conclusion

ПРОБЛЕМА

Интерпретация Выводы
Новые идеи Коммуникация
Понимание и определение проблемы
Как мы можем ответить на этот вопрос?

Разработка исследования

Что и как измерять?

ЗАКЛЮЧЕНИЕ ПЛАН ЦИКЛ PPDAC

закономерностей Выдвижение Управление

гипотез

АНАЛИЗ ДАННЫЕ

Сбор

Построение таблиц, графиков Поиск Очистка

Посмотрим, как можно применить этот цикл в HR аналитике на простом примере.

1. **Проблема** – компания теряет прибыль из-за нехватки персонала. Предполагаем, что мы может отыскать причины этого исследовав текучесть.

2. **План** – разрабатываем план исследования. К примеру, мы решили, что будем измерять метрики текучести и удовлетворенности персонала.

3. **Данные** – организуем сбор данных, которые нам понадобятся для ответа на вопрос, на постоянной основе в HR системах, обеспечиваем их чистоту и полноту.

4. **Анализ** – строим дашборд, визуализируем динамику метрик в различных вариантах и разрезах, выдвигаем и проверяем гипотезы о связях текучести и удовлетворенности.

5. **Заключение** – интерпретируем результаты, даём заказчику рекомендации по снижению текучести. Получив ответ на этот вопрос, у нас могут возникнуть новые вопросы и цикл вновь начинается с определения проблемы.