

Основы анализа и визуализации данных для медиа

Лекция 3
12 сентября 2019

Рогович Татьяна
rogovich@gmail.com

Статистика выводов и описательные статистика

Структура лекции

- Поговорим, чем отличается описательная статистика от индуктивной статистики.
- Разберемся с генеральной совокупностью и выборкой.
- Поговорим о базовых представлениях данных в одном измерении - частотные таблицы, гистограммы, столбчатые диаграммы. Посмотрим на разные типы распределений.
- А что делать если переменных две? Посмотрим на базовые представления бивариативного графика.

Описательная (дескриптивная) статистика

Использование методов статистики и графических подходов для представления информации об изучаемых данных.

Позволяет обобщать первичные результаты, полученные при наблюдении или в эксперименте. Применение описательной статистики включает следующие этапы:

- Сбор данных
- Категоризация данных
- Обобщение данных
- Представление данных

Индуктивная статистика

Статистическая техника, основанная на теории математической вероятности и используемая для получения обобщений относительно генеральной совокупности на основе выборки, взятой из этой же совокупности.

Описательная статистика

Генеральная совокупность

Population

Состоит из всех людей или других объектов (например, атлантических лососей или частей самолетов), которые исследователи хотели бы изучить, если бы обладали бесконечными ресурсами.

Описательная статистика

Перепись населения Census

Результат периодического сбора данных обо всей генеральной совокупности называется переписью. Во многих странах государственные организации проводят перепись населения.

Описательная статистика

Выборка Sample

Часть генеральной совокупности элементов, которая охватывается экспериментом (наблюдением, опросом).

Описательная и индуктивная статистика



Описательная и индуктивная статистика

Параметры

описывают генеральную совокупность.

Например, среднее μ

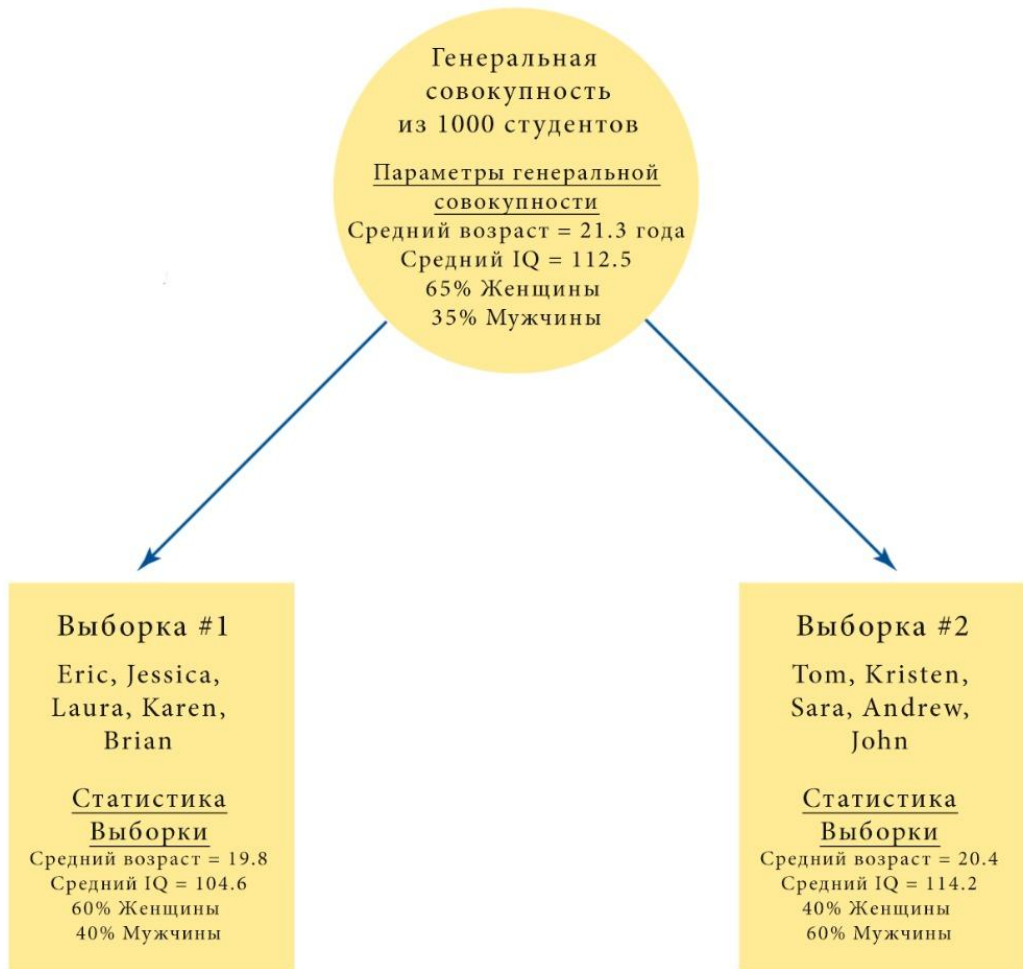
Статистики

описывают выборку

Например, среднее M

Статистики выборок из одной и той же совокупности могут отличаться*

*Но теоретически
при правильно
сделанных выборках
не должны значимо
отличаться



Как делается выборка?

Выборка



```
graph TD; A[Выборка] --> B[Случайная]; A --> C[Детерминированная]
```

Случайная

1. Простое случайное извлечение
2. Стратифицированная выборка
3. Кластерная выборка

Детерминированная

1. Волонтеры
2. Квотная (разбиваем на страты, но выбираем не случайно)
3. Нерепрезентативная (работаем с тем, кто есть)

Как делается выборка?

Стратифицированная выборка

то процесс, состоящий из двух этапов, в котором совокупность делится на подгруппы (слои, страты, strata). Слои должны взаимно исключать и взаимно дополнять один другого, чтобы каждый элемент совокупности относился к одному и только одному слою, и ни один элемент не был упущен. Далее, из каждого слоя случайным образом выбираются элементы, при этом обычно используется метод простой случайной выборки.

Цель стратифицированной выборки — увеличение точности.

Когда используем?

Стратифицированная выборка

Например, распределение дохода семей неравномерно, так доход большинства семей составляет меньше 100 тысяч рублей в месяц, и лишь немногие семьи имеют доход выше 300 тысяч рублей в месяц. Если применить простую случайную выборку, семьи с доходом 300 тысяч рублей в месяц и выше могут не быть адекватно представлены.

Как делается выборка?

Кластерная выборка

В кластерной выборке (cluster sampling) изучаемая совокупность сначала делится на взаимоисключающие и взаимодополняющие подгруппы, или кластеры. Затем с помощью вероятностного метода выборки формируется случайная выборка кластеров. В выборку включаются либо все элементы отобранного кластера, либо проводится их отбор вероятностным методом.

Цель кластерной выборки — увеличить эффективность выборки, уменьшив затраты на ее проведение.

Когда используем?

Кластерная выборка

Распространенная форма кластерной выборки — территориальная выборка (area sampling), в которой кластеры состоят из географических территорий, таких как округа, жилые районы или кварталы.

Пример. Выборка ВЦИОМ

ВЫБОРКА ИССЛЕДОВАНИЯ

Как мы выбираем номера телефонов

Для построения выборки мы берём **все диапазоны стационарных и мобильных телефонных номеров с сайта Россвязи**. Совокупная ёмкость этих диапазонов составляет более 650 млн. телефонных номеров – в неё попадают АБСОЛЮТНО ВСЕ телефонные номера, которые уже используются операторами или которые могут быть введены по действующим правилам.

Для опроса мы случайно (датчиком случайных чисел) отбираем из этой базы 40-45 тысяч номеров в зависимости от сезона (летом и в новогодние каникулы больше, а осенью и весной меньше) – это наша основа выборки. **Телефонные номера для прозвона выбираются из нее также датчиком случайных чисел.**

<https://wciom.ru/research/research/sputnik/>

Пример. Выборка эксперимента в сети Facebook

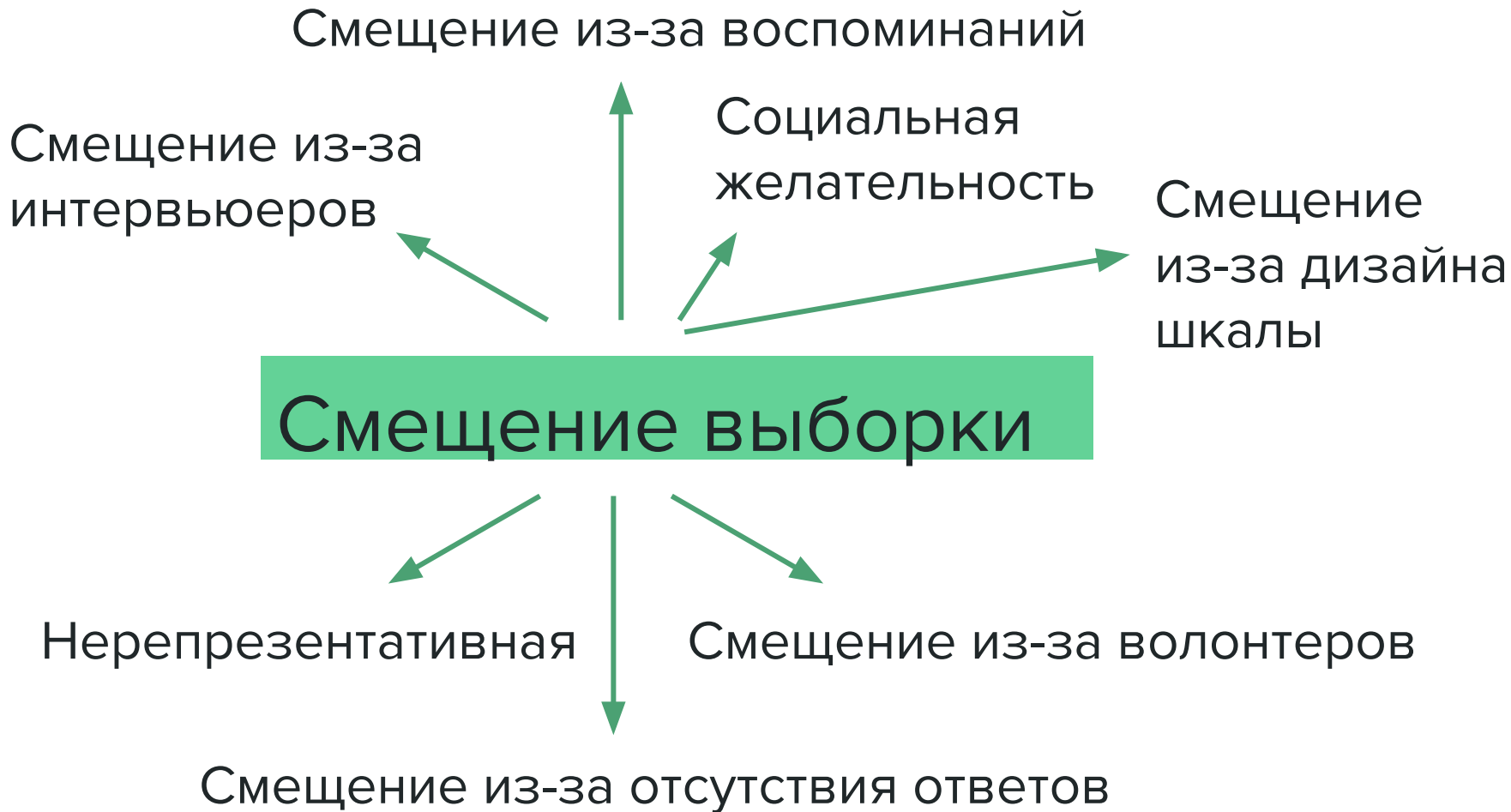
September 13, 2012

A 61-million-person experiment in social influence and political mobilization

Nature

By: Robert Bond, Christopher J. Fariss, Jason J. Jones, Adam D. I. Kramer, Cameron Marlow, Jaime Settle

<https://research.fb.com/publications/a-61-million-person-experiment-in-social-influence-and-political-mobilization/>



Одни и те же данные могут быть как выборкой, так и генеральной совокупностью

Например, оценки за полугодие всех учеников школы.

Анализ генеральной совокупности - абсолютные утверждения.

Анализ выборки - вероятностное суждение (best guess).

От чего зависит размер выборки?

- Размер изменений (хотим мы узнать, как что-то меняется на 0.5 % или на 5 %)
- Сила эффекта (статистическая сила)
- Надежность результата (статистическая значимость)
- Бюджет (менее важен для исследований в корпорациях на внутренних данных)

<https://www.evanmiller.org/ab-testing/sample-size.html>