ЦЕНТРАЛЬНЫЙ УНИВЕРСИТЕТ

Введение в статистику

Проект 5. Обязательный

Ковчег

Кейс, проблема, идея

Кейс

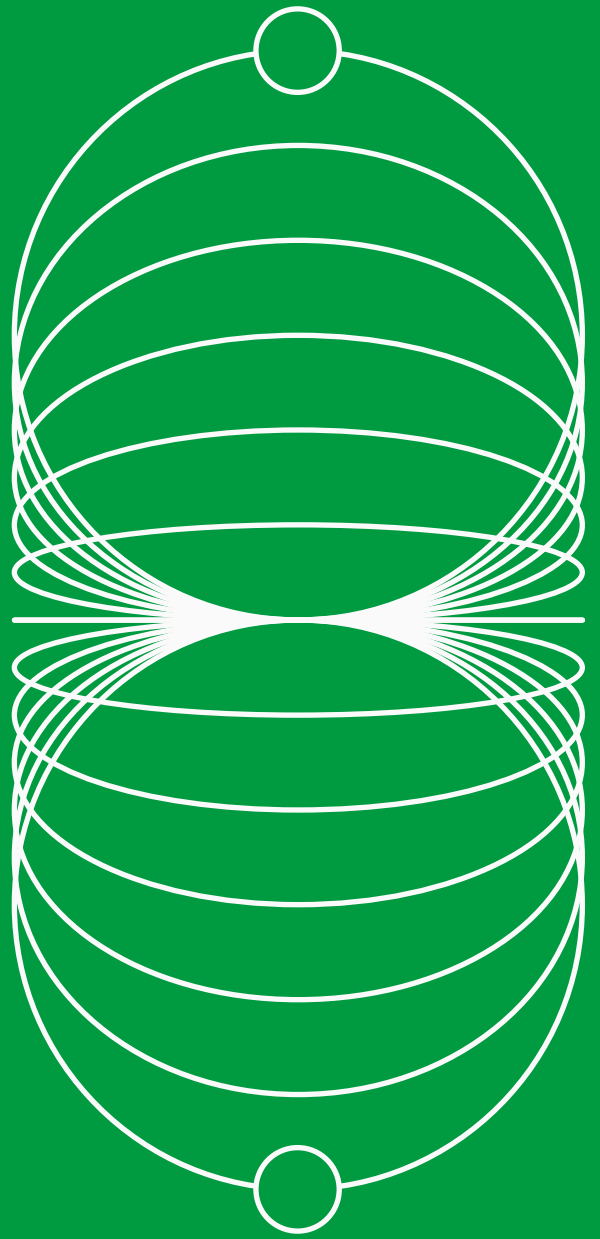
- Компания планирует проводить тестовые запуски рекламы 1 раз в день
- Нельзя, чтобы каждый клиент получал каждый день новое рекламное сообщение очередной рекламной кампании
- Тестовые запуски будут проводиться на маленьких выборках из всех клиентов
- Всего клиентов 80 тысяч, размер выборки — 400
- Выборка должна быть репрезентативной, чтобы замеры на ней позволяли делать хотя бы какие-то выводы обо всех клиентах
- Проверку репрезентативности можно осуществить на основе одной кампании, которую проводили на всех 80 тысячах клиентов

Проблема

Если один раз составить репрезентативную выборку и использовать её во всех тестах, будем наблюдать эффект очередной рекламы на человека, которому и так каждый день показывают рекламу

Идея решения

Разработать алгоритм, который будет формировать каждый раз новую выборку для тестового запуска очередной рекламной кампании



Роль, задача, результат

Роль

Аналитик отдела работы с клиентами

Задача

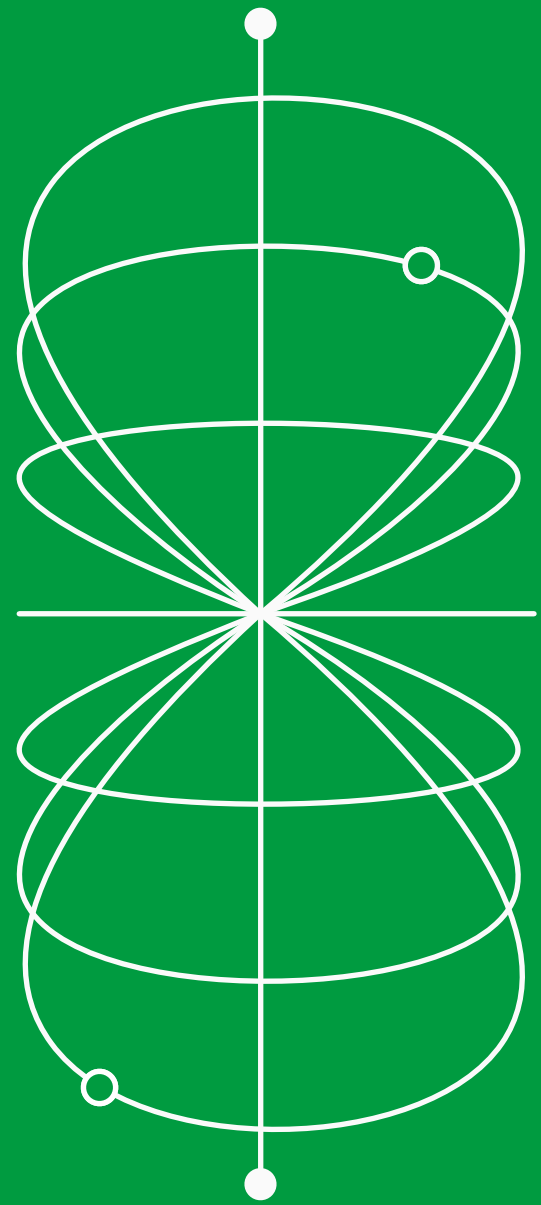
- Написать функцию на языке Python, которая формирует выборки
- Проверить, что функция формирует выборки репрезентативно
- Подготовить слайды: объяснить способ формирования выборок и показать, насколько репрезентативные выборки он формирует

Доступные данные

Датасет ark.csv содержит данные обо всех клиентах компании: возрасте, пол, совершил ли покупку по итогам рекламной кампании

Ожидаемый результат

- Презентация на слайдах в формате pdf
- Jupyter Notebook в Google Colab с расчётами



Требования к слайдам

Если слайды или Jupyter Notebook не приложен, решение кейса оценивается **в 0 баллов**

Понятность/внешний вид

- Внешний вид презентации не мешает воспринимать информацию
- Понятно на какие вопросы отвечает каждый слайд
- Содержимое таблиц, графиков понятно из слайда без необходимости открывать исходный датасет
- Выводы явно сформулированы

МАКСИМУМ 4 БАЛЛА

Обоснованность

- Выводы основаны на таблицах, графиках, показателях, полученных из данных
- Таблицы и графики получены скриншотом или картинкой из Jupyter Notebook, поэтому их можно перепроверить
- Выводы явно сформулированы

МАКСИМУМ 4 БАЛЛА

Реакция заказчика

0 баллов: не принимает, ищет другого исполнителя

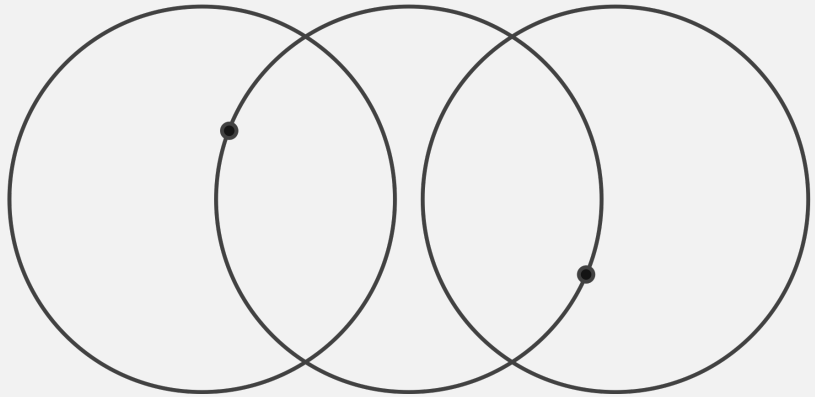
1 балл: частично принимает, считает необходимым отдать на доработку текущему исполнителю

2 балла: принимает, готов пересылать слайды от своего имени, под свою ответственность

МАКСИМУМ 2 БАЛЛА

Синий уровень

Подход к решению



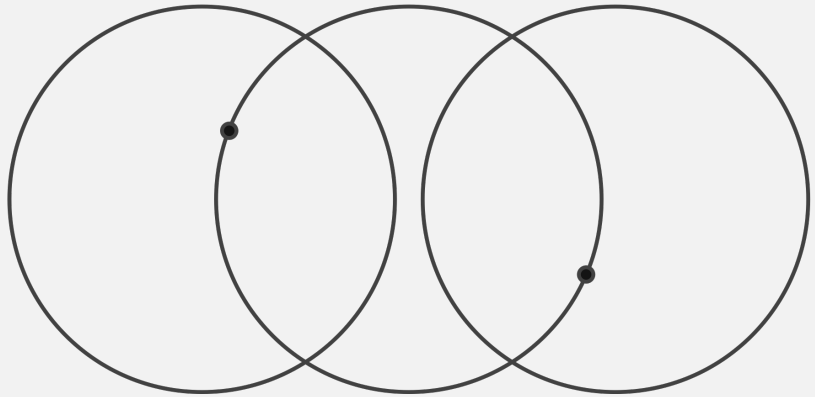
01 **Описать на слайде способ получения простой случайной выборки**

02 **На следующем слайде показать результаты симуляции этого способа формировать выборку. Для этого:**

- сделать N=10 000 повторов (сформировать выборку и посчитать по выборке долю клиентов, купивших продукт),
- построить гистограмму для полученных N значений, отметить на оси настоящую (от всех клиентов) долю клиентов, купивших продукт,
- прокомментировать, есть ли постоянное систематическое завышение или занижение ответа в методе,
- прокомментировать, может ли метод простой случайной выборки давать большую ошибку, и если может, то насколько часто.

Красный уровень

Подход к решению



01 **Сделать слайды синего уровня**

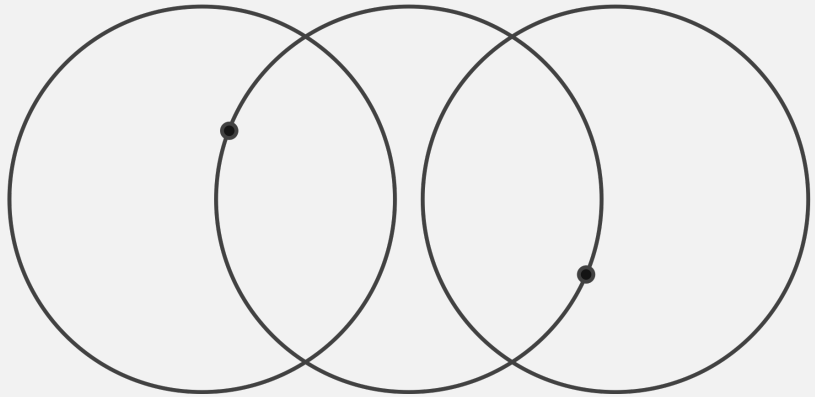
02 **Описать на слайде способ получения стратифицированной случайной выборки**

03 **На следующем слайде показать результаты симуляции этого способа формировать выборку. Для этого:**

- сделать N=10 000 повторов (сформировать выборку и посчитать по выборке долю клиентов, купивших продукт),
- построить гистограмму для полученных N значений, отметить на оси настоящую (от всех клиентов) долю клиентов, купивших продукт,
- прокомментировать, есть ли постоянное систематическое завышение или занижение ответа в методе,
- прокомментировать, может ли метод простой случайной выборки давать большую ошибку, и если может, то насколько часто.

Чёрный уровень

Подход к решению



01 **Сделать слайды синего и красного уровня**

02 **Сравнить на слайде метод простой случайной выборки и метод стратифицированной случайной выборки. Для этого:**

- показать на одном графике две гистограммы,
- сделать выводы о точности методов,
- объяснить интуитивно, за счёт чего стратифицированный метод повышает точность,
- объяснить, почему нельзя понять, какой из двух подходов лучше по одной выборке простой случайной и одной выборке стратифицированной.

Почему приходится делать симуляцию большого количества повторов обоих методов?

Что и когда нужно сдать

Что сдавать?

- Презентация в слайдах в формате pdf
- Jupyter Notebook в Google Colab с расчётами

Когда сдавать?

Сроки сдачи указаны в информационной системе

Сдача проекта – необходимое условие прохождения курса