НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

Data Analysis National Olympiad - DANO

ГЕНЕРАЛЬНАЯ СОВОКУПНОСТЬ И МЕТОДЫ ФОРМИРОВАНИЯ ВЫБОРОК

Илья Слаболицкий
Матвей Зехов

Москва, 2021

ГЕНЕРАЛЬНАЯ СОВОКУПНОСТЬ vs ВЫБОРКА

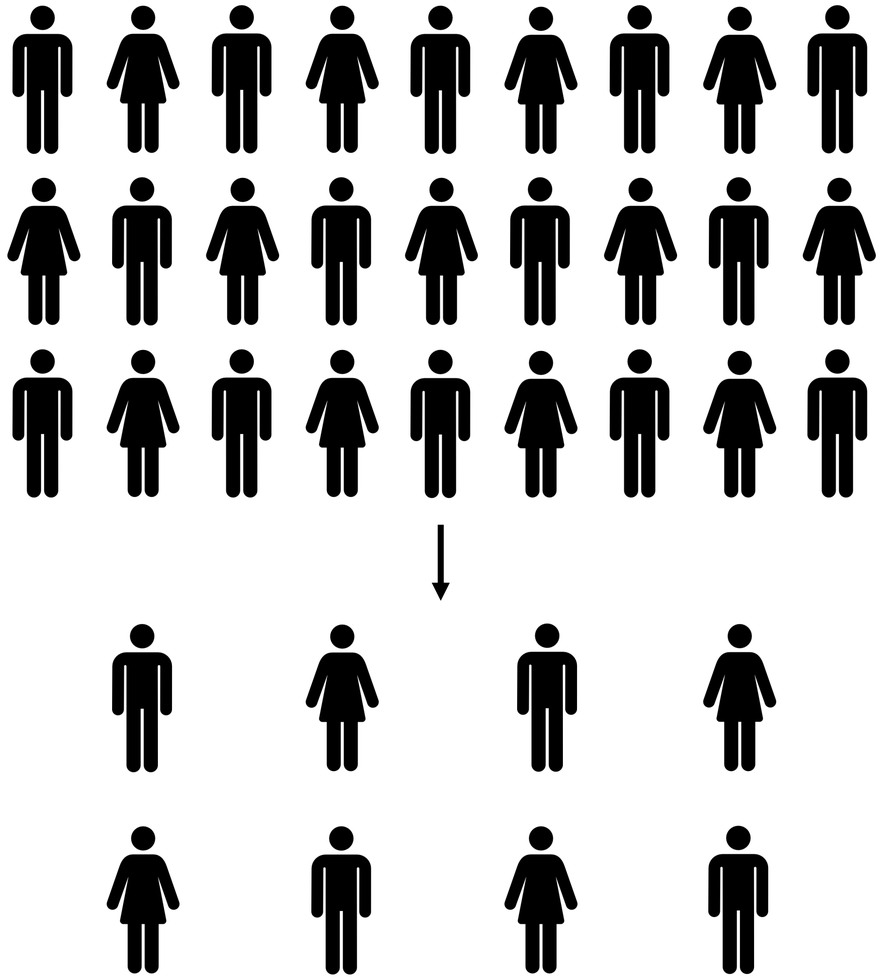
В чем отличие?

- **Генеральная совокупность** (population) – это совокупность всех объектов, обладающих общими признаками и относительно которых нам хочется делать какие-либо выводы при анализе некоторой конкретной задачи.

  Признаки, по которым мы разделяем объекты, могут быть абсолютно любыми: территориальное положение, возраст, уровень доходов и многие-многие другие.

- **Выборочная совокупность** или **выборка** (sample) – та часть генеральной совокупности, которую мы отбираем в рамках эксперимента и на основе которой мы будем описывать или охарактеризовывать генеральную совокупность.

- Некоторые объекты могут быть одновременно генеральной совокупностью и выборкой: например, все студенты Вышки.



ЗАЧЕМ НУЖНЫ ВЫБОРКИ?

Для чего вообще нужны выборки, если есть генеральная совокупность?

- Зачастую (практически всегда!) получить информацию и собрать данные о каждом или даже почти каждом элементе генеральной совокупности не представляется возможным. В связи с этим возникает необходимость в формировании выборки, то есть некоторой репрезентативной подгруппы генеральной совокупности.

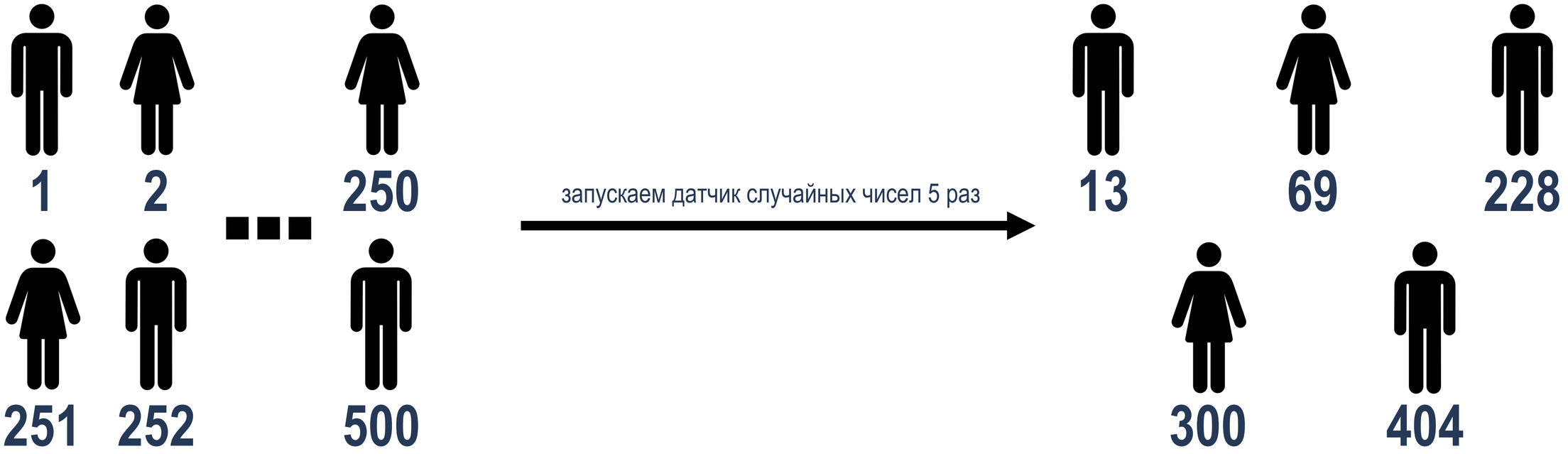
  Такой процесс формирования выборки называется отбором или **семплированием** (sampling).

Какие существуют методы семплирования?

<table>
<tr><td colspan="2">Вероятностные методы семплирования
(probability sampling)
• каждый элемент генеральной совокупности имеет шанс быть выбранным
• возможность добиться репрезентативной выборки
• большое количество тонкостей реализации процедуры отбора</td><td colspan="2">Невероятностные методы семплирования
(non-probability sampling)
• неслучайный критерий отбора элементов генеральной совокупности
• простота процедуры отбора
• высокий риск возникновения выборочного смещения (sample bias)</td></tr>
<tr><td>Случайный отбор
(simple random sample)</td><td>Систематический отбор
(systematic sample)</td><td>Convenience sample</td><td>Voluntary response sample</td></tr>
<tr><td>Стратифицированный отбор
(stratified sample)</td><td>Кластерный отбор
(cluster sample)</td><td>Purposive sample</td><td>Snowball sample</td></tr>
</table>

СЛУЧАЙНЫЙ ОТБОР

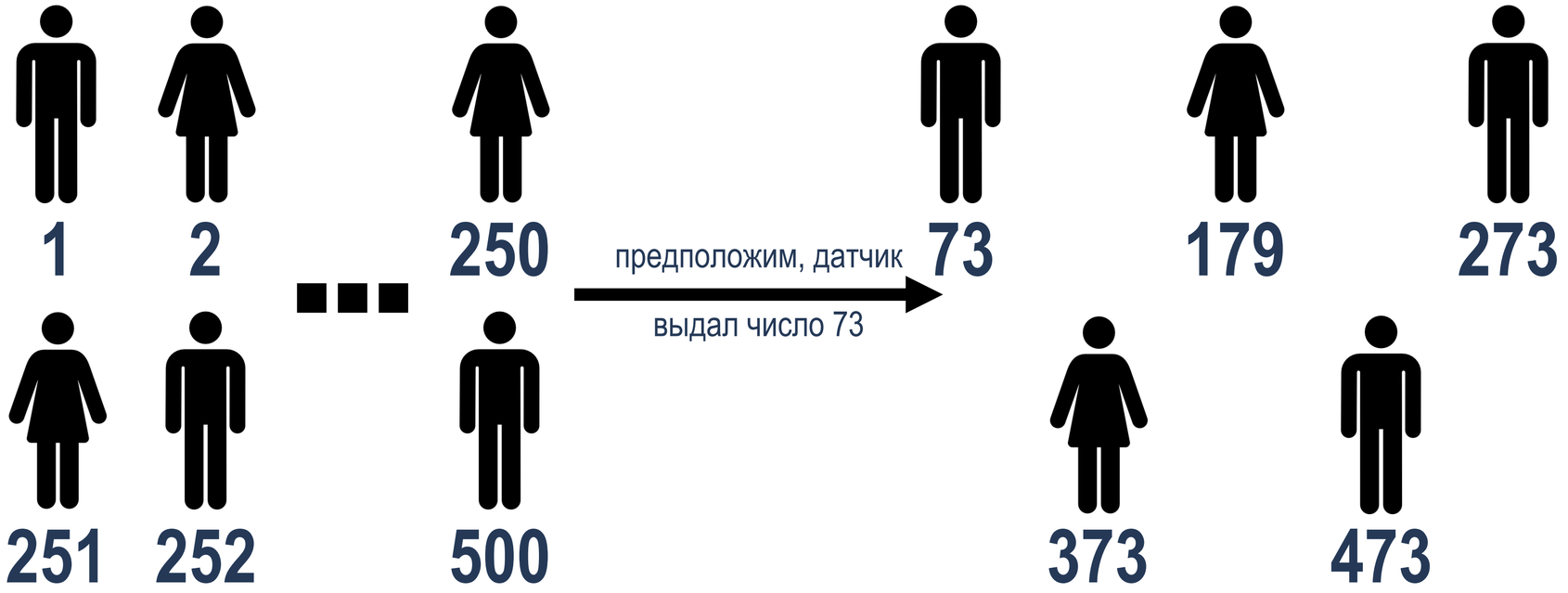
- Тип метода: вероятностный, каждый элемент имеет равные шансы быть отобранным
- Суть метода: каждый из N элементов генеральной совокупности нумеруется, и случайным образом отбираются n элементов из них
- **Пример.** В некоторой научной лаборатории одного крупного университета работает 500 сотрудников. Исследовательница Яна хочет получить выборку из 5 случайно отобранных сотрудников (не привязывайтесь к цифрам, это всего лишь пример!) при помощи случайного отбора. Как ей осуществить данную процедуру?



СИСТЕМАТИЧЕСКИЙ ОТБОР

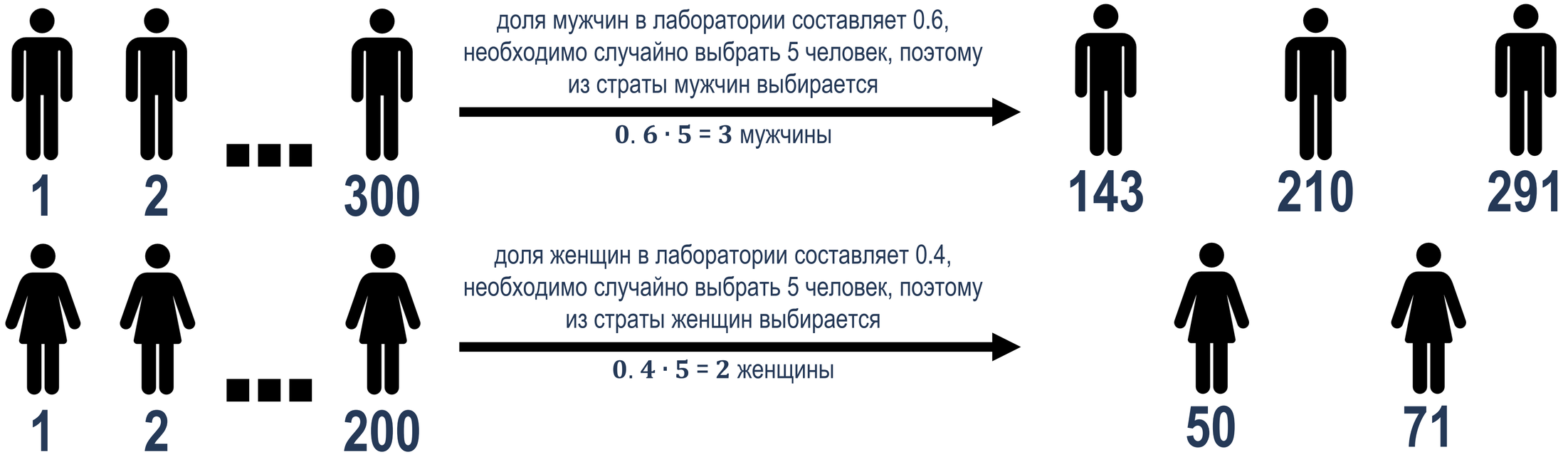
- Тип метода: вероятностный, каждый элемент имеет шанс быть отобранным
- Суть метода: каждый из N элементов генеральной совокупности нумеруется, и вся генеральная совокупность делится на n равных (по возможности) интервалов; затем из первого (по возрастанию) интервала случайно отбирается один элемент и с шагом, равным длине интервала, отбираются другие элементы
- **N.B!** Используя данный метод есть риск «пропустить» некоторые подгруппы элементов
- **Пример.** Рассмотрим ту же научную лабораторию. Помогите исследовательнице Яне провести систематический отбор.

При помощи датчика случайных чисел выберем число из интервала [1, 100] и начнем систематический отбор всех сотрудников с него. Далее с шагом шагом 100 отберем 5 сотрудников.



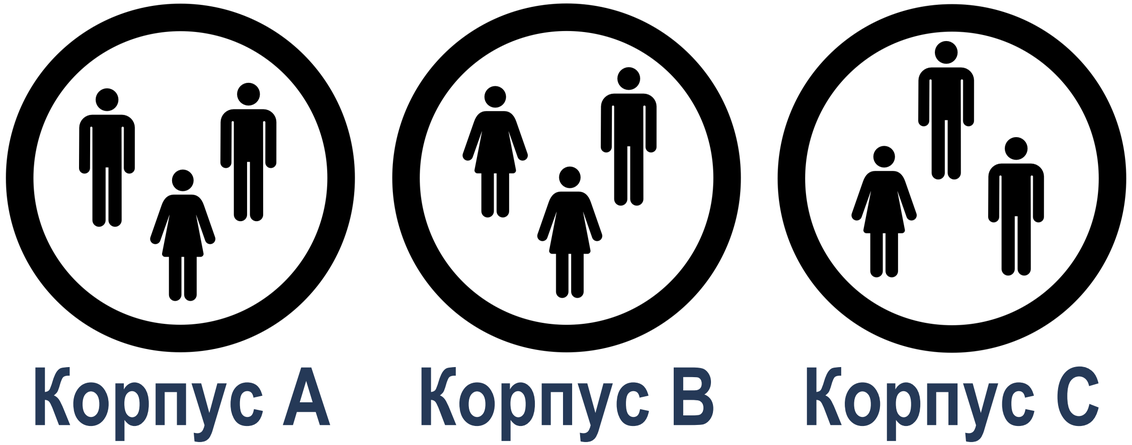
СТРАТИФИЦИРОВАННЫЙ ОТБОР

- Тип метода: вероятностный, каждый элемент имеет шанс быть отобранным
- Суть метода: генеральная совокупность разделяется на несколько подгрупп или **страт** (strata) в соответствии с каким-либо признаком (гендер, возраст или любой другой признак), а затем производится случайный отбор в каждой страте пропорционально доле данной страты от всей генеральной совокупности
- **Пример.** Снова обратимся к научной лаборатории, про сотрудников которой дополнительно известно, что среди них 300 мужчин и 200 женщин. Как при данных условиях исследовательнице Яне провести стратифицированный отбор?



КЛАСТЕРНЫЙ ОТБОР

- Тип метода: вероятностный, каждый элемент имеет шанс быть отобранным
- Суть метода: генеральная совокупность, состоящая из N элементов, разделяется на l кластеров таким образом, чтобы в каждом кластере было поровну (по возможности) элементов; затем случайны образом выбирается один кластер и из него при помощи случайного, систематического или стратифицированного отбора отбирается n элементов
- **N.B!** Кластеры должны быть организованы таким образом, чтобы элементы различных кластеров практически не различались своими характеристиками
- **Пример.** Уже знакомую нам научную лабораторию расформировали, создав три отдельных подразделения: корпус А, корпус В и корпус С. Численность сотрудников в каждом из них составила 200, 150 и 150 человек, соответственно. Известно, что каждый из корпусов занимается схожими задачами, а сотрудники в них набирались случайным образом. Помогите исследовательнице Яне провести кластерный отбор.



предположим, датчик случайных чисел выбрал Корпус С,

далее уже по одной из трех известных схем выбираем 5 сотрудников из Корпуса С

СПАСИБО ЗА ВНИМАНИЕ!

НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

Электронная почта для связи: islabolitskiy@hse.ru

Давайте поработаем! Юные исследователи, вашему вниманию предлагается небольшое задание для самостоятельного решения по теме «Генеральная совокупность и методы формирования выборок»[1]

Вас позвали в один из филиалов крупного аналитического агентства «Неслучайные случайности»[2]. Из-за сложной эпидемиологической ситуации все 27 сотрудников этого филиала работают удаленно. Вам поручили организовать небольшой опрос, результаты которого будут использованы для проведения исследования. Однако работы у сотрудников филиала очень много, почту они просматривают крайне редко, так что отправить письмо каждому лично с просьбой пройти опрос у вас не получится. Неожиданно вы вспоминаете, что изучали различные методы формирования выборок. Вы также решаете, что трети всех сотрудников будет достаточно для получения репрезентативной выборки. Примените ваши теоретические знания на практике!

1. Скачайте данные о сотрудниках филиала. Файл называется *random_data*.
2. Проведите случайный отбор. Для удобства вы можете использовать любой рандомайзер (например, сайт random.org).
3. Проведите систематический отбор. Решайте сами, стоит ли как-нибудь упорядочивать сотрудников филиала или нет.
4. Проведите стратифицированный отбор. Решайте сами, по какому/каким признаку/признакам разделить сотрудников.
5. Проведите кластерный отбор. Организуйте кластеры самостоятельно. Решайте сами, каким способом отбирать сотрудников внутри кластера.
6. Объясните, какой из способов позволил организовать наиболее репрезентативную выборку и почему.

Желаю успехов!

---

[1] К данному заданию не прикладывается решение. Задание рассчитано на самостоятельную работу (желательно в небольших группах). Однако, если же вас заинтересовала данная задача и вы очень хотите себя проверить, напишите письмо автору видео и составителю задачи, Слаболицкому Илье Сергеевичу: islabolitskiy@hse.ru.

[2] Внимание! Все персонажи вымышленные! Любые совпадения с реальными людьми случайны!