ПРИКЛАДНАЯ СТАТИСТИКА

Посиделка 2: офигительные истории про репрезентативность

Ульянкин Ппилиф *

> Простите меня за то, что я рассказываю вам и без меня всем известные очевидные вещи.
>
> *Неизвестный лектор про свою вводную лекцию*

Отбросим все предрассудки и заблуждения. Наш мир на самом деле — это сундук. Этот сундук выплёвывает данные. Никто не знает, как он устроен внутри, но каждый пытается с помощью данных в этом разобраться. Иногда люди делают это неправильно, а потом страдают. В этой посиделке мы будем смотреть на их страдания.

1 Офигительные истории

Перед тем, как перейти к офигительным историям, вспомним пару определений. **Генеральная совокупность** — это все объекты, которые нас интересуют при исследовании. **Выборка** — это та часть генеральной совокупности, по которой мы собрали данные для исследования. **Репрезентативной** называют такую выборку, по которой можно сделать корректные выводы про всю генеральную совокупность.

*https://github.com/FUlyankin/matstat_lec

1.1 Сколько ты зарабатываешь?

Юра хочет понять, сколько зарабатывают люди, живущие в Москве. В этом ему помогают его друзья: Серёга номер один и Серёга номер два. Генеральная совокупность в данном случае — все жители города. Опросить каждого человека в городе — нереально. Поэтому ребята проводят опрос среди тысячи жителей. Выводы о всём городе каждый делает по своей выборке.

Юра тусуется среди селеб, поэтому о доходах спрашивает только их. Серёга номер один ходит по разным автобусным остановкам и спрашивает людей на них. Серёга номер два раздобыл телефонную книгу со всеми жителями города и случайно отобрал из неё тысячу жителей для опроса. Чья выборка репрезентативна?

По репрезентативной выборке можно сделать вывод обо всей исследуемой генеральной совокупности. Если мы хотим понять, адекватные ли выборки собрали ребята, надо сначала понять кого они исследуют.

Если Юра исследует только селеб, у него всё будет хорошо. Селебы зарабатывают больше обычных людей, поэтому Юра не сможет обобщить своё исследование на весь город. У первого Серёги обратная проблема. На автобусных остановках он не будет встречать состоятельных людей, из-за этого его оценки будут занижены. У Серёги номер два проблем нет. Он сможет отобрать по телефонной книге людей из всех социальных страт, если они конечно согласятся с ним общаться.

Случайная выборка — это один из способов обеспечить репрезентативность. С точки зрения статистики случайность выборки означает, что

а. каждый объект генеральной совокупности в принципе может попасть в выборку;

б. все объект могут попасть в выборку равновероятно.

Серёга номер один ходит по остановкам. Не все жители Москвы могут попасть в его выборку. С точки зрения статистики, такая рандомизация не подходит. С помощью отбора по телефонной книге можно попробовать достучаться абсолютно до всех жителей города. Шанс на попадание в выборку есть у всех. С точки зрения теории, такая рандомизация подходит.

Мы в дальнейшем будем обозначать одно наблюдение как X_i. Мы будем рассматривать каждое наблюдение, как случайную величину, имеющую такое же распределение, как и генеральная совокупность. Очень часто мы будем предполагать, что все наблюдения были собраны независимо друг от друга из одного и того же распределения. Этот факт мы будем записывать как

$$X_1, X_2, \dots, X_n \sim \mathtt{iid}\ F_\theta,$$

где F_θ — какой-то закон распределения, n — размер выборки, а $\mathtt{iid}$ расшифровывается как identically independently distributed, независимы и одинаково распределены.

1.2 Внутренняя и внешняя обоснованность

Можно ли выводы о средних зарплатах, сделанные по Москве, обобщить на Нижний Новгород? Введём ещё пару определений. **Изучаемая генеральная совокупность —** генеральная совокупность объектов, из которых взята выборка. **Целевая генеральная совокупность —** это генеральная совокупность объектов, на которую распространяются результаты статистического исследования, руководствуясь причинно-следственными связями.

Когда статистики делают выводы, они обычно размышляют о том, насколько далеко можно обобщить результаты, полученные при исследовании. Если исследование было сделано корректно, его выводы обоснованны для изучаемой генеральной совокупности. В такой ситуации говорят о **внутренней обоснованности выводов.** Если полученные выводы можно обобщить и на другие генеральные совокупности, говорят про **внешнюю обоснованность.**

В случае с зарплатами, мы не можем обобщать выводы, сделанные по Москве на Нижний Новгород. Москва — не Россия. Жизнь в столичном регионе довольно сильно отличается от остальной страны. Зарплаты имеют разные распределения.

1.3 Выборы, выборы

Франклин Рузвельт сражается на выборах в президенты со своим оппонентом Альфредом Лэнданом. Происходит это аж в 1936 году. Журнал «Литерари Дайджест» рассылает всем своим подписчикам бюллетень с вопросом, за кого они собираются голосовать. В опросе принимает участие 10 миллионов человек. На основе такого огромного числа респондентов журнал предсказывает победу республиканцу Лэндану с перевесом (60 на 40). Однако на выборах побеждает демократ Рузвель — как раз с таким же перевесом, но в обратную сторону. Как думаете, почему так произошло и что журнал сделал не так?

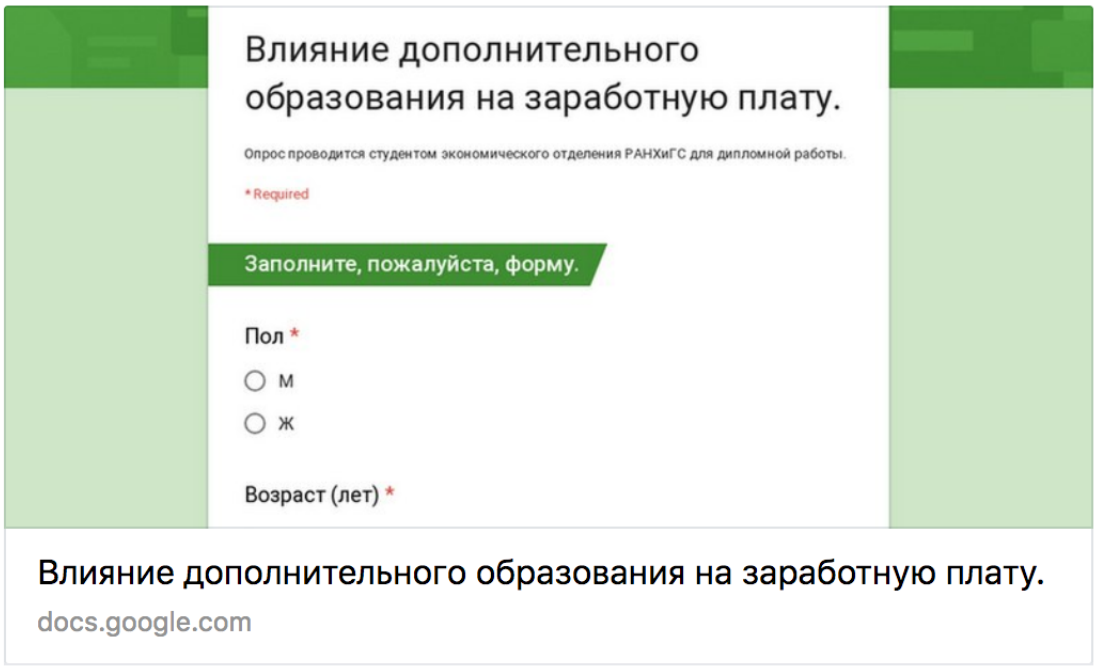
Дело в том, что выборка журнала оказалась нерепрезентативной. Большинство подписчиков журнала были республиканцами. Журнал прекрасно это понимал и пытался сгладить это смещение. Он брал телефонные книги и дополнительно рассылал бюллетени для опроса по указанным там адресам. Журнал не учел забавного факта: телефоны были доступны только среднему и высшему классу общества, а это были, в основном, тоже республиканцы. Выборка в конечном итоге оказалась **смещённой.** Простой рабочий народ не попал на радары журнала.

Никому неизвестный на тот момент социолог Джордж Гэллап тоже исследовал выборы. Он собрал ответы всего лишь 50 тысяч людей и абсолютно верно спрогнозировал результаты выборов. Когда вы собираете данные, **важнее всего не количество данных, а их качество.**

1.4 Проблема самоотбора

В прошлом примере мы поговорили про смещение. Другой важный фактор, из-за которого результаты исследования могут оказаться искаженными **— проблема самоотбора.** Не все люди хотят участвовать в исследованиях.

Помогите нашему студенту в рисерче. Пожалуйста пройдите небольшой опрос, он займет менее 5 минут.
https://docs.google.com/forms/d/e/1FAIpQLSfnp-8-jzV_Y..



Картинка 1: Мольба студента-маркетолога о помощи

Часто в социальных сетках можно увидеть посты, в которых студенты просят помочь им с курсовой и пройти небольшой опрос. Именно такой пост изображен на рисунке 1. Анкета вывешена в социальной сети на страничке студента. Как считаете, какие проблемы возникнут у студента с выборкой?

Если не предпринять никаких дополнительных усилий, а просто вывесить анкету в социальной сети, её увидит только ближайшее окружение студента. Выборка окажется смещённой. Давайте предположим, что студент знал, что такое может произойти и заплатил социальной сети за рекламу своей анкеты. Она показывается как рекламное объявление случайным людям.

В такой ситуации мы столкнёмся с проблема самоотбора. Человек должен хотеть заполнить анкету. Очень маловероятно, что люди с высокой зарплатой решат потратить своё время на заполнение анкеты. Их время слишком дорого стоит. В итог, в исследовании о влиянии образования на зарплату будут сделаны неправильные выводы.

1.5 Правда ли, что вы гей?

Другой пример — опросы, которые связаны с разными чувствительными темами. Что вы ответите интервьюеру, если он на улице спросит у вас, курили ли вы марихуану? Скорее всего, вы либо откажитесь отвечать, либо соврёте. В результате социологи получат некорректную статистику.

В 2010 году Британское управление национальной статистики (Office for National Statistic, ONS) опубликовало результаты своего исследования о том, сколько взрослых британцев являются геями[1]. Опросы показали, что это примерно 1.5%[2]. При этом исследование Кинси конца 1940-х указывало цифру около 10%[3]. Общепринятым мнением было, что в наши дни эта цифра находится в районе 5 — 7%. И кому верить?

Проблема нового опроса заключалась в том, что только 96% респондентов согласились разговаривать с социологами. Оставшиеся 4% не стали отвечать на вопрос о своей сексуальной ориентации. Возможно, что они были геями и не хотели в этом признаваться исследователям. Если бы это было так, то доля геев подскочила бы до 5.5%, величины которая находится в пределах ранее полученного диапазона. Конечно же, мы не можем знать наверняка, что это так. Возможно, что отказы отвечать связаны с чем-то другим. Также вполне может быть, что часть респондентов соврала о своей сексуальной ориентации.

А важна ли вообще эта цифра? На самом деле да. Например, правительство Тони Блэра в Британии при разработке своей социальной политики опиралось на цифру в 5 — 7%. Если бы к нему на стол попала цифра в 1.5%, политика могла бы быть совсем иной.

Чтобы избежать подобных искажений приходится придумывать разные более хитрые техники исследования. Такие, чтобы респонденты были уверенны, что они в безопасности. Например, можно поступить следующим образом.

1) Каждому респонденту мы даём монетку, кубик и листок бумаги.
2) Респондент уходит в отдельную комнату. Там он подкидывает монетку и кубик.
3) Если на монетке выпал орёл, респондент честно пишет на бумаге ответ на вопрос «Правда ли, что вы гей?»
4) Если на монетке выпала решка, респондент честно пишет на бумаге ответ на вопрос «Правда ли, что на кубике выпало чётноё число?»
5) Листок бумаги сдаётся экспериментатору. Он видит только ответ «да» либо «нет», и не знает на какой вопрос он дан. Если респондент гей и боится признаться в этом, он может сказать, что отвечал на вопрос про кубик. Никто не сможет это проверить.

Процедура вполне понятна. Не очень понятно, как можно проинтепретировать полученные результаты. На помощь в этом приходит теория вероятностей. Отталкиваясь от неё, можно очистить данные от случайности.

В среднем около половины респондентов будет отвечать на вопрос про сексуальную ориентацию. Вторая половина будет отвечать на вопрос про кубик. У половины кубик выпадет чётной стороной. Получается, что четверть всех ответов точно будет «да». Если процент ответов «да» превышает четверть, то превышение должно быть связано с тем, что респонденты ответили «да» на вопрос «Правда ли, что вы гей?».

Например, предположим, что есть 100 респондентов, и 30 из них говорят «да». Тогда, в

[1] Часть примеров позаимствованы отсюда: https://www.profmatt.com/are-you-gay
[2] https://www.theguardian.com/world/2010/sep/23/gay-bisexual-population-uk
[3] https://en.wikipedia.org/wiki/Kinsey_Reports

среднем, 25 из них говорят «да» из-за того, что у них выпало чётное число. Остальные 5 говорят «да», потому что они геи. Поскольку на вопрос «Правда ли, что вы гей?» отвечали только 50 из 100 респондентов, мы делаем вывод что 5 из 50, то есть 10% геи.

Конечно же, эти цифры приблизительны, но при большом числе респондентов, они будут довольно устойчивыми. Позже мы узнаем с вами о законе больших чисел, который обеспечивает эту устойчивость. В опросе управления национальной статистики участвовало 450000 человек. Если бы они использовали эту методологию, их результаты имели бы гораздо более высокую степень достоверности.

1.6 Самолёты, дельфины и ошибка выжевшего

Во время Второй Мировой войны американские военные собирали статистику попаданий пуль в фюзеляж самолёта. По самолётам, вернувшимся из полёта на базу, была составлена карта повреждений среднестатистического самолёта. Военные собирались укрепить дырявые части бронёй. Так бы и поступили, если бы не вмешался статистик Абрахам Вальд. Как думаете, что он сказал военным и почему?

Картинка 2: Ошибка выжевшего и самолёты Вальда

Вальд указал на то, что простреленные части самолёта — это не его слабые места. Самолёты с критическими повреждениями не возвращаются на аэродром. Поэтому укреплять надо те зоны, где пробоин нет. Если вы хотели укрепить части с пробоинами, вы совершили **ошибку выжевшего.** Это одна из разновидностей ошибки отбора. Она довольно часто мешает делать корректные выводы. Например, все знают, что иногда дельфины выносят на берег людей, потерпевших кораблекрушение. Какие же они умные и дружелюбные!

На самом деле, дельфины просто любят толкать разные предметы носом. Никому не известно, как много утопающих они затолкали ещё глубже в океан. Мы знаем только примеры чудесного спасения, потому что **жертвы дельфинов ничего никому никогда больше не смогут рассказать.**

1.7 Зефирный тест

Страшнее всего напороться на **эндогенность.** Говоря простым языком, не учесть в исследовании какие-то важные факторы и получить из-за этого неправильные выводы. Слышали когда-нибудь про зефирный тест[4]? Положите перед ребёнком зефир и дайте ему 15 минут посидеть с ним в полном одиночестве в пустой комнате. Если малыш выдержит и не съест зефир, он получит вторую зефирку. Если с искушением справиться не выйдет — тогда второй зефирки ему не полагается.

Впервые тест проводили в 1960-х в Стенфорде. В 1990-е авторы эксперимента изучили повзрослевших зефирных респондентов и заявили, что те, кто в детстве справился с искушением, оказались гораздо успешнее своих сородичей, съевших зефирку. Начиная с этого момента, зефирные тесты стали модными. В интернете можно найти довольно много видосов с этим тестом [5].

Вроде бы понятно и логично: потерпи, чтобы получить набежавшие проценты. Удержи себя от соблазна сейчас, чтобы получить что-то прекрасное в будущем. Прошло время, и зефирный тест опровергли. Оказалось, что авторы исследования немножечко налажали с репрезентативностью. Во-первых, они отобрали для исследования всего-лишь 90 детей. Да ещё и всех из детского садика Стэнфорда. Не каждый бедняк может позволить себе такой детский садик.

Новое исследование Тайлера Уаттса с коллегами блестяще показало, что дело вовсе не в силе воли, а в богатстве, которое в семье либо есть, либо нет. На этот раз ребята проверили гипотезу почти на тысяче детей, собранных из всевозможных слоёв общества. При этом они учли важные факторы, такие как, например, материальное положение семьи и образование родителей.

В итоге новое исследование обнаружило довольно слабое подтверждение базовой гипотезы (что способность откладывать удовольствие приводит к успеху в жизни). Вместо этого вышло, что возможность продержаться до второй зефирки большей частью определяется финансовым положением семьи ребёнка — и вот оно-то большей частью и объясняет будущие успехи. А

[4]Раздел нагло украден из канала Хулиномика, подписывайтесь нанего: https://t.me/hoolinomics
[5]https://www.youtube.com/watch?v=goZkgVA68iw

вовсе не умение откладывать удовольствие на потом!

Обнаружилось, что среди детей, мамы которых получили высшее образование, не было никакой разницы в дальнейшей успеваемости. А у тех детей, мамы которых не доучились, преимущество «выигравших» объяснялось социальными факторами, а не способностью дождаться вторую зефирку. Как только исследователи принимали во внимание обстановку в доме трёхлетних испытуемых (например, количество книг на полках, или реакцию мам на просьбы детей), широко воспетая способность «откладывать удовольствие на потом» внезапно переставала означать что-то особенное. Самоконтроль, взятый отдельно от всего, не позволяет детям преодолеть пропасть в благосостоянии и социальном окружении.

Если подумать, то результаты знаменитого теста можно объяснить массой причин. Например, жизнь детей из бедных семей не гарантирует им вообще какую-либо еду, не говоря уже о сладостях. В ожидании ведь есть определённый риск. Даже если родители и обещали им что-то вкусненькое на выходных, вполне вероятно, что финансовая необходимость заставит их поменять своё решение, и вместо обещанных чупа-чупсов мама купит домой пачку риса.

А если вы родились в состоятельной семье, то отложить удовольствие на потом вам не составит особого труда. Опыт подскажет, что финансовое положение семьи стабильно, и у родителей есть ресурсы, чтобы выполнить свои обещания. И даже если такой ребёнок не станет дожидаться второго зефира, у него будет шанс получить вместо него что-то вкусное чуть позже.

Исследователи поведенческой экономики из Гарварда и Принстона написали в 2013 году книгу о бедности под названием «Скудность. Почему обладание малым означает так много». Дело в том, что состояние нехватки чего-либо изменяет образ мышления людей. Бедность легко может сподвигнуть человека на сиюминутную награду. Маргиналы просто иначе оценивают то, что им доступно. Другими словами, второй зефир не имеет значения, если ты сомневаешься, что тебе вообще достанется зефир.

Эти выводы намекают на то, что бедняки стараются баловать детей при первой возможности, а состоятельные родители имеют склонность поощрять ребёнка дождаться более жирной награды. Краска для волос или конфеты могут показаться дурацким подарком, но не забывайте о том, что это порой единственное, что может себе позволить семья из Тамбовской области. А нищим детишкам какая-никакая сиюминутная радость заметно улучшит несчастную жизнь. Ведь у них нет никакой гарантии, что завтра случится хоть что-то хорошее.

2 Детсадовская аналитика

Когда людей собеседуют на аналитические позиции, для разминки им периодически накидывают ситуации, в которых с данными произошёл какой-то трэш. Например, в Яндексе на собеседованиях раньше предлагали задачу «Детсадовская аналитика». Сейчас больше не предлагают, так как её засветили на хабре в одном из разборов[6]. Попробуйте сами накидать

[6] https://habr.com/ru/post/546004/

вариантов, в чём конкретно тут возникают проблемы, а потом уже читайте решение.

Упражнение 1

Журналисты спросили у 20 детей в одном детском саду, живут ли их родители вместе. Двое детей ответили, что не живут. На следующий день была опубликована статья под заголовком «Каждая десятая семья с маленькими детьми в нашем городе разведена!» Почему выводы журналиста могут быть некорректны?

Решение:

Проблем тут вагон и маленькая тележка. Давайте перечислим основные.

- Маленькая выборка, оценка журналиста будет неточной.
- С репрезентативностью выборки большие проблемы. Детский садик может находится в неблагополучном районе. Либо наоборот, в благополучном районе. Он в выборке только один.
- Дети могут врать либо не понимать вопроса из-за маленького возраста.
- На одну семью в среднем приходится больше одного ребёнка. Оба ребёнка, сказавших что их родители не живут вместе могут относиться к одной и той же семье. Тогда оценка журналиста просаживается в два раза. Более того, в многодетных и однодетных семьях процент разводов может довольно сильно различаться.
- В поле зрения исследователя попадает не вся генеральная совокупность. Есть много благополучных семей, в которых дети не ходят в садик. Из-за этого оценка журналиста будет завышена.
- Сам по себе вопрос журналиста сформулирован некорректно. Что вообще имеется в виду под «не живут вместе»? Возможно, что отец находится в командировке. Возможно, что один из родителей погиб. Это нельзя приравнивать к семье в разводе.

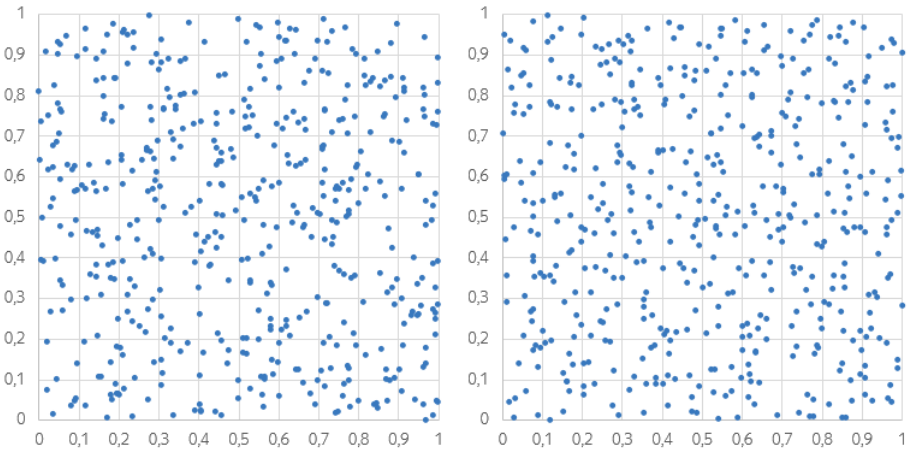
На самом деле в методике журналиста можно найти ещё кучу проблем. Подумайте о том, как эти проблемы можно решить. Как бы вы проводили такое исследование на месте журналиста?

3 Немного про природу случайности

Человек — очень плохой генератор случайных чисел. На многих курсах по статистике лектора проводят следующий эксперимент. Студентов просят написать последовательность из 100 случайных цифр, сгенерированных из их собственных голов. В процессе оказывается, что генерировать случайные последовательности из цифр — сложно. Поначалу люди пишут свои цифры довольно быстро, но вскоре замедляются. Они начинают думать о том, как сделать так, чтобы последовательность была случайной. Иногда люди сдаются и начинают писать последовательности, которые уже знают. Например, номера телефонов.

Со мной такой эксперимент проводили на курсе статистики в Яндексовой школе анализа данных, ШАД. Аудитория из сотни человек в течение 10 минут генерировала случайные последовательности из цифр. Ассистенты перенесли эти последовательности в электронный вид. Затем семинарист сгенерировал сотню случайных последовательностей и обучил классификатор, который практически идеально отделил последовательности, сгенерированные людьми от компьютерных.

Почему так происходит? Давайте взглянем на картинку 3. Какой из двух наборов точек кажется вам «более случайным»[7]?



Картинка 3: Случайные выборки

Выбрали? А вы уверены в своём выборе? Распределение на левом рисунке явно неравномерно. Есть места, в которых точки сгущаются, а есть и такие, в которых точек почти нет: из-за этого даже может показаться, что левый график более тёмный. На правом рисунке локальные сгущения и разрежения тоже присутствуют, но меньше бросаются в глаза.

Меж тем, именно левый график получен при помощи «честного» генератора случайных чисел. Правый график тоже содержит сплошь случайные точки, но эти точки сгенерированы так, чтобы все маленькие квадраты содержали равное количество точек.

Случайность — очень «комковатая» штука. Например, в последовательности

$$(1, 6, 4, 8, 8, 2, 0, 9, 1)$$

есть две восьмёрки рядом с друг другом. В списке из 100 случайных чисел, в среднем, таких двойников будет около десятка. Студенты, как правило, генерируют меньше. Им такие двойники не кажутся случайными. Точно также на картинке выше сгустки из точек показались вам довольно неестественными.

Люди склонны ассоциировать случайность с равномерностью. Если данных у нас очень много, случайность, действительно, довольно равномерна. Если мы подкидываем монетку 1000 раз, мы ожидаем увидеть орла примерно в половине случаев. Однако, в краткосрочной

[7]Идею позаимствовал у Алексея Шаграева: https://habr.com/ru/post/496750/

перспективе монетка довольно долго может выпадать орлом.

Когда на рулетке выпадает красное, все игроки ставят на чёрное. Им кажется, что красное уже выпадало и второй раз этого точно не повториться. При этом, рулетке наплевать на то, что думают люди, а также ей наплевать на свой предыдущий опыт. Каждый бросок делается независимо, и рулетка надолго может «залипнуть» на одном из цветов. Такое заблуждение людей в азартных играх называют **ошибкой игрока.**

Эти когнитивные искажения мешают нам жить[8]. Именно из-за них Apple пришлось засунуть в функцию перемешивания в iTunes костыль, который не позволяет включать две песни одного и того же исполнителя друг за другом[9]. Так последовательность из треков кажется людям более случайной.

4 Случайная выборка на практике

В реальной жизни сделать честное сэмплирование довольно сложно. Возникает много проблем. Давайте вернёмся к примеру с исследованием зарплаты. Представим себе, что в телефонной книге есть список всех людей, проживающих в Москве. Мы сделали по ней честный сэмпл. Дело осталось за малым — опросить людей.

Во-первых, люди могут отказаться с нами разговаривать. Такие люди сделают наш сэмпл непригодным. Мы никак не сможем заменить вывалившегося из него респондента на какого-то нового. Тогда нарушится предпосылка про равновероятное попадание объектов в выборку. Эти вероятность окажутся разными. Более того, такой подход может исказить результаты, как в примере с вопросом «Правда ли, что вы гей?». Тех, кто побоялся ответить «да» мы выбросим и заменим на тех, кто твёрдо говорит «нет».

[8] Немного подробнее про то, как именно: https://vas3k.ru/blog/387/
[9] https://www.telegraph.co.uk/technology/11429317/The-biggest-myths-about-technology.html

Вся социология довольно сильно страдает от такой проблемы. Любым цифрам в ней нельзя доверять. Как бы сильно не старались социологи, цифры будут искажены. Поэтому, в социологических опросах люди всегда смотрят на динамику, а не на абсолютные значения. Если был проведён опрос о популярности какого-то политика, и выяснилось, что его поддерживает 60% респондентов — этой цифре верить нельзя. При этом мы можем делать какие-то выводы по динамике этой цифры. Для этого достаточно каждый месяц проводить опрос каждый месяц по одинаковой методологии и смотреть, падает цифра или растёт.

Представим себе, что все люди из нашей выборки согласны с нами поговорить. На общение с ними может уйти вечность. Каждого человека надо найти, от каждого надо выпытать нужную нам информацию. Все живут в разных районах, а кто-то мог вообще уехать из города на время или навсегда.

Один из способов решить эту проблему — использовать **кластерную выборку.** Такой подход уместен, когда география нашей генеральной совокупности довольно разнообразна. Мы могли бы попробовать разбить город на кластеры по районам, затем отобрать несколько районов случайным образом, и только после из каждого отобранного района набирать людей. Конечно же, возникает проблема с тем, что вероятности попасть в выборку оказываются для людей разными. Если мы набираем из каждого района по 100 человек, то в более крупных районах, вероятность попадания в выборку у жителей будет меньше. Однако все эти искажения находятся под нашим контролем и мы можем взвесить на них наши итоговые результаты.

Другой широко распространённый подход для сэмплирования — сбор **стратифицированных выборок.** При таком подходе всю генеральную совокупность разбивают на несколько страт и с каждой из них работают независимо. Это помогает улучшить точность статистических результатов.

Предположим, что мы знаем, что 40% генеральной совокупности составляют мужчины, а 60% женщины. Мы берём случайную выборку из каждой из этих двух групп, страт. Размер каждой выборки делаем пропорциональным её размеру. Например, выборка размера 20 должна содержать 8 мужчин и 12 женщин. Тогда она будет репрезентативной для структуры населения.

Методы стратификации часто используются в онлайн-экспериментах. Предположим, что веб-сервис вносит какие-то изменения в свою работу. Поведение пользователей зависит от характеристик устройств, операционных систем, версий браузеров, характеристик самих пользователей и так далее. Поэтому без стратификации в А/Б-тестах легко столкнуться с тем, что, скажем, доля мобильных пользователей в разбиениях различается на 0.5% и метрики измеряют эффект от этого перекоса, а не от вносимого изменения. Подобного рода ошибки могут портить аналитику и, в конечном счёте, приводить к неверным решениям в развитии продуктов и бизнеса.

Стратифицированный подход в данном случае предписывает разбивать наблюдения на страты (по версиям устройств, ОС, браузеров и т.д.), вычислять метрики внутри страт, взвешивать их сообразно размерам этих страт и таким образом получать значения итоговых показателей. Про это мы будем говорить более подробно в будущих посиделках.

Список литературы

[1] *Matthew Handy.* Probability and statistics a guide for teachers and students at A level and beyond. // https://www.profmatt.com/statistics.

[2] *Алексей Марков.* Жлобология. // https://alexeymarkov.ru/trial/zhlobolite.pdf.