НЕПРАВИЛЬНЫЕ ИНТЕРПРЕТАЦИИ И ЛОЖНЫЕ ЗАКОНОМЕРНОСТИ В АНАЛИЗЕ ДАННЫХ

Дьяконов Александр (проф. ВМК МГУ)

Анализ данных в последнее время стал одной из самых популярных дисциплин, всё чаще используются английские термины «data science», «data mining» и «big data», онлайн-дисциплины по анализу данных, статистике и машинному обучению являются самыми востребованными, а профессия аналитика – высокооплачиваемой. В интернете много руководств по тому, как собирать, обрабатывать и исследовать данные. К сожалению, часто забывается, что область «data science» не случайно названа НАУКОЙ о данных. Это подразумевает строгий подход к выводам и их интерпретациям, иначе занятная инфографика может стать причиной заблуждений или даже манипуляций общественным сознанием.

Покажем, как иногда возникают неверные интерпретации. Мы обойдёмся без сложных формул, сделаем текст не слишком формальным, понятным людям далёким от прикладной математики.

Какие знаки зодиака обманывают?

Наверное, каждый из нас знает, что такое гороскоп, а некоторые даже регулярно читают свой прогноз на день. В серии интернет-публикаций [1], [2] рассказывается об одном исследовании. В нём установлено, представители каких знаков зодиака как исправно выплачивают долги по микрозаймам (см. табл. 1).

Знак зодиака	**Сколько представителей знака допускают хотя бы одну просрочку**
Овен 	35.3%
Дева 	35%
Рыбы 	34.2%

Табл. 1. Самые большие просрочки выплат по версии [2].

В [1] идёт обсуждение этой темы, в частности, высказываются точки зрения, что в скоринге (оценке кредитоспособности клиента) надо брать только классические признаки: образование, доход, кредитную историю и т.п. Но никто не говорит о главных причинах, почему зодиаку не следует доверять. Во-первых, **достаточно ли велика исследуемая выборка**[1] и **значимы отклонения** процентов в табл. 1? Во-вторых, **насколько найденные закономерности устойчивы** (например, не зависят от времени)? Давайте попробуем разобраться.

Для начала наглядно объясним, почему выборка должна быть большой. Давайте сгенерируем случайную последовательность из нулей и единиц, в которой 0 и 1 появляются с одинаковой вероятностью. Можно подбрасывать монету, в случае выпадения герба писать 0, решки – 1. Можно воспользоваться генератором псевдослучайных чисел, который есть во всех математических пакетах (мы использовали систему Matlab [3]), см. рис. 1.

Теперь посчитаем процент единиц среди первых k членов последовательности. На рис. 2 он показан в зависимости от k. Вроде бы единиц и нулей должно быть примерно поровну. Из-за случайности процент единиц постоянно отклоняется от 50%. При увеличении k отклонения становятся меньше. Заметим, что если бы наша монета была не совсем симметричная или генератор выдавал единицу не с вероятностью 0.5, а с вероятностью 0.51, то при объёме выборки 2500 мы бы этого не почувствовали (а вот при 250000 – наверное да). **Даже в простой задаче оценки вероятности или анализе симметричности монеты выборка в несколько тысяч может оказаться маленькой**.

В [1]-[2] написано, что объём выборки «более 250000» – столько человек обратилось за микрозаймами. Это кажется большим числом, но представителей конкретных знаков зодиака здесь не больше 10% [2]. Также не больше 10% людей получили микрозаймы [2]. Таким образом, проценты в каждой строке табл. 1 вычислены по не более чем 2500 заёмщикам. На рис 2 показано, какие колебания процентов здесь возможны (честно было бы генерировать 1 с вероятностью близкой к значениям из табл. 1, но там выводы примерно такие же). Поэтому говорить по табл. 1, что, например, Овны «хуже» Рыб не следует.

```
0100001110011010010110000010001101110010010100011 1...
```

Рис.1. Ряд псевдослучайных чисел.

[1] **Выборка** – это наблюдения, которые у нас есть. В данном случае – описания кредиторов.

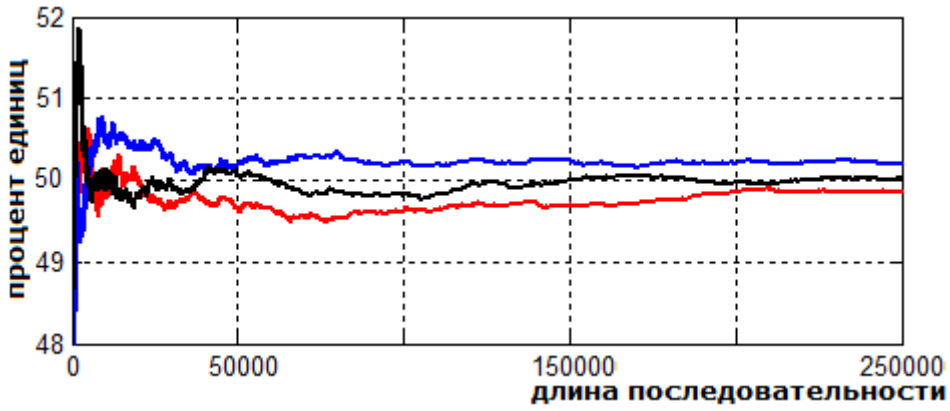


Рис.2.1. Процент единиц для трёх разных последовательностей (250000 экспериментов).

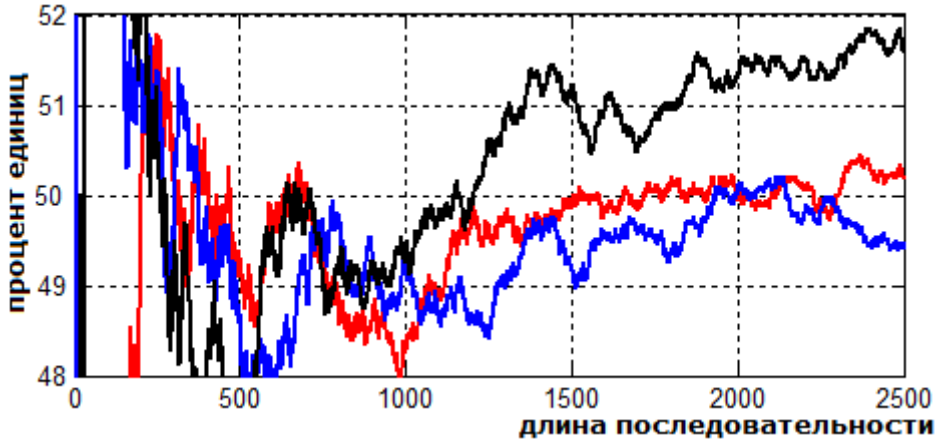


Рис.2.2. Процент единиц для трёх разных последовательностей (2500 экспериментов).

Теперь проведём более честное исследование кредитоспособности. У нас есть статистика по некоторым российским банкам[2], в ней кредитные истории около 300000 человек, которые в течение года брали кредиты. Для каждого есть полная статистика по последующим выплатам. Мы разбили всю статистику на 4 части по кварталам, в дальнейшем каждому кварталу будет соответствовать свой график.

Рассмотрим для начала, как представители разных полов выплачивают кредиты, см. рис. 3. Как видим, женщины «лучше» мужчин – примерно на 4%, лишь в одном квартале разница между процентами невыплат мужчин и женщин опускается до 2%. Теперь посмотрим, как образование влияет на действия кредитора. По логике, чем выше уровень образования, тем надёжней кредитор. Статистика подтверждает это – см. рис. 4.

Аналогично ведут себя и другие «классические» скоринговые признаки: семейный статус (женат/гражданский брак/холост/разведён/вдовец), сфера занятости, число детей и

[2] К сожалению, предоставить её читателю мы не можем (по понятным причинам).

т.п. Теперь посмотрим, как надёжность кредитора зависит от его знака зодиака – рис. 5. Здесь нет устойчивости! Знаки зодиака, которые были лучшими в одном квартале, становятся худшими в другом! Например, раки имеют самый высокий процент невыплат в первом квартале и самый низкий в последнем.

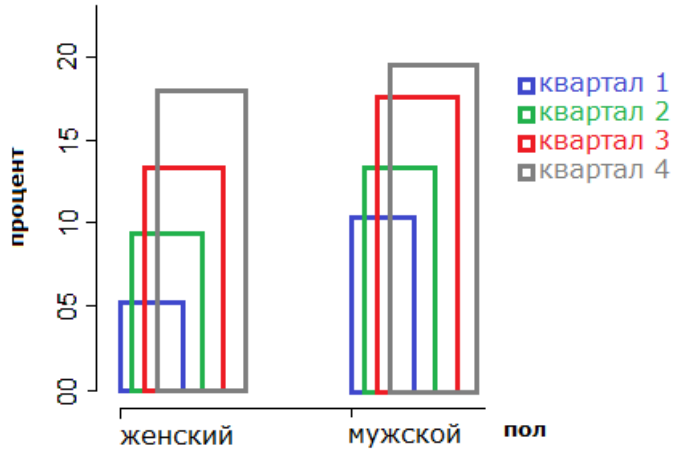


Рис. 3. Невыплата кредита в зависимости от пола.

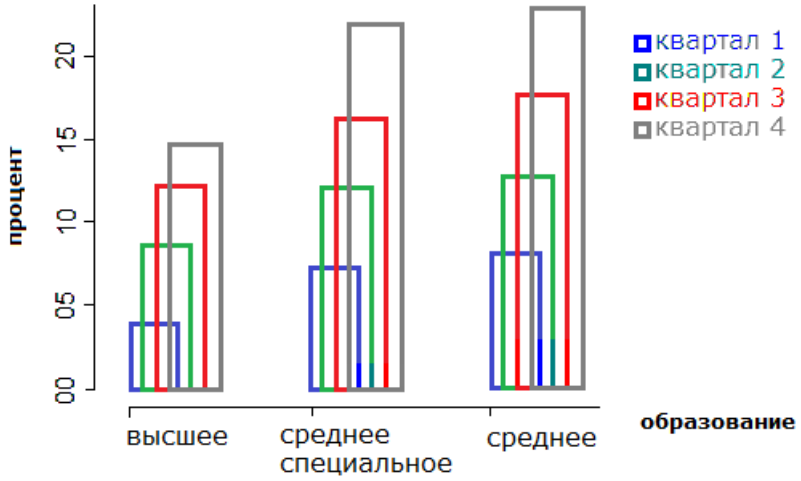


Рис. 4. Невыплата кредита в зависимости от образования.

На рис. 3–4 разница между разными группами кредиторов (мужчинами и женщинами, людьми с высшим образованием и без) слабо изменялась, по крайней мере, та группа, которая была надёжнее в одном квартале, была также надёжнее и в другом. При этом разница в процентах невыплат была существенна (часто больше 5%), а между разными знаками зодиака на рис. 5 – менее существенна (не больше 3%).

Конечно, есть проблема, что полов у нас 2, уровней образования – 3, а знаков зодиака – целых 12. И для оценки вероятности невыплат по знакам выборка должна быть больше. В каждом квартале около 75000 кредиторов делятся по 12 знакам зодиака. Поэтому процент

невыплат оценивается меньше чем по 6250 кредиторам (при равномерном распределении по знакам зодиака). Но мы честно признаём, что выборка у нас не очень большая[3], хотя более чем в два раза превосходит выборку [2]. Кстати, «плохие» знаки зодиака из табл. 1 – овен, дева и рыбы – у нас такими не являются!

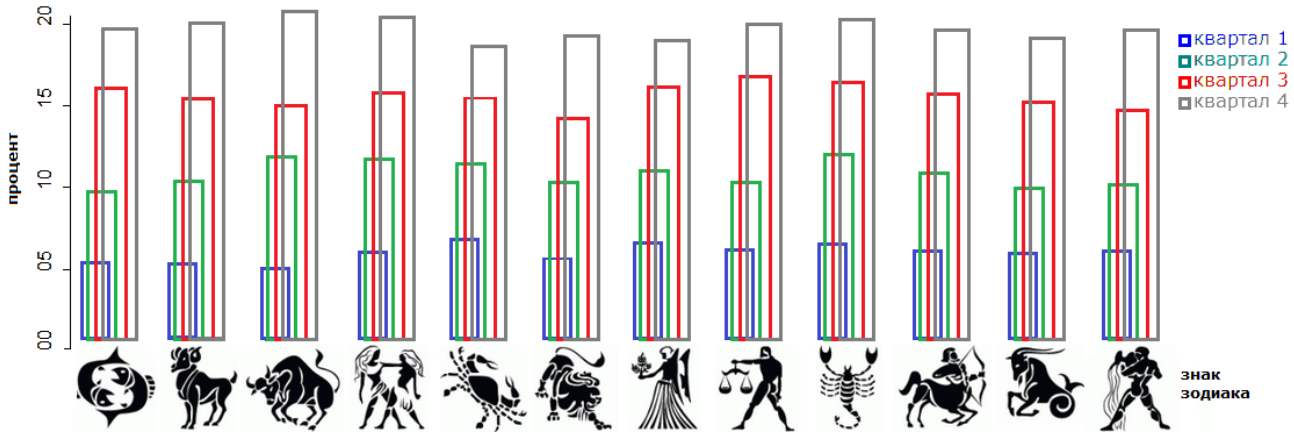


Рис. 5. Невыплата кредита в зависимости от знака зодиака.

На наш взгляд, **логическая закономерность тогда является таковой, когда с её помощью можно что-то предсказать**. Поэтому мы и разбивали статистику по кварталам. Просмотрев статистику по трём кварталам мы говорим, что в следующем кредиторы с высшим образование будут надёжнее кредиторов со средним, и мы не ошибёмся! Со знаками зодиака этого не проходит. Именно устойчивые закономерности и полезны на практике. Например, в данной задаче, банк, основываясь на прошлой статистике, предсказывает надёжность новых кредиторов в будущем.

Ещё отметим, что мы вовсе не показали, что гороскопы являются ложью, и даже, что они бесполезны при скоринге. Вдруг, например, надёжность знака зодиака зависит от месяца, в котором он выплачивает кредит. Но чтобы проверить эту гипотезу нужна совсем большая выборка. Отметим также, что в отличие от [1]-[2] наши данные были не по микрозаймам, а по кредитам в банках (поэтому выводы исследований могут различаться).

Убережём читателя ещё от одной ошибки. Парадоксально, но часто **нельзя интерпретировать данные так, как они представлены на графиках и таблицах**! Например, говорить, что 11% женщин не смогут выплатить кредит. Дело в том, что банк выдаёт кредиты не всем, а лишь тем, кто был одобрен его скоринговым алгоритмом. Поэтому у нас в

[3] Понятие «большая» зависит от задачи, которую мы решаем. См. также пост в блоге автора http://alexanderdyakonov.wordpress.com/2015/06/13/размеры-выборок/

выборке неслучайные кредиторы. Верно лишь, что 11% женщин, одобренных скоринговыми алгоритмами конкретных банков, имели проблемы с выплатами кредитов. Алгоритмы, кстати, со временем меняются и проценты тоже...

Так ли хороши оранжевые машины?

Сейчас стало популярным публиковать обзоры с «интересными открытиями в области анализа данных». Часто в них описываются просто некоторые забавные корреляции[4] в данных, но их неверные интерпретации могут стать причиной неправильных выводов. Например, один из популярных обзоров [4], [5] на первое место поставил следующее наблюдение. В одной базе данных по покупкам подержанных автомобилей явно была видна закономерность, что купленные подержанные машины оранжевого цвета меньше подвержены техническим проблемам. Были выдвинуты разные гипотезы, например, что поскольку подобная машина нестандартного цвета является средством самовыражения, то владелец лучше о ней заботится и т.п.

Отметим, что эта закономерность (как и некоторые другие) была найдена участниками соревнования по анализу данных [6] на платформе Kaggle. Сами исследователи почти никогда не обсуждают природы закономерностей. Даже в своеобразной «энциклопедии больших данных» [7] указано, что в современной аналитике **надо искать только зависимости, а их причины часто найти и обосновать невозможно**. Но вот журналисты, которые пишут обзоры, закономерности обсуждают и делают выводы...

Объяснение «экзотичностью цвета» отметается при анализе публикаций по этой теме. Скажем, в [8] речь идёт о тех же данных, и оказывается, что жёлтые машины являются «самыми плохими», хотя цвет тоже достаточно экзотический. Давайте попробуем разобраться в чём же дело, тем более, что аналогичную статистику по цветам автомобилей часто приводят в разных отчётах ГИБДД (по авариям, угонам и т.п.) и тоже подвергают интерпретациям.

Допустим, мы живём в мире, в котором всего два производителя машин: первый производит очень надёжные машины исключительно белого цвета, они ещё ни разу не ломались в течение гарантийного срока. Второй – машины белого и чёрного цвета, примерно в одинаковом количестве. В среднем 20% его машин (как белых так и чёрных)

[4] Если не знаете значение слова «корреляция» – замените его на «зависимость».

подвержены поломкам в течение гарантийного срока. Оба производителя выпускают машины в одном количестве (в год)[5].

Рис. 6. Схематичное распределение цветов машин и поломок в нашем модельном мире.

Что же в этом случае нам скажет статистика? Ответ показан в табл. 2 и легко следует из рис. 6. Следуя обычной обывательской логике, мы скажем, что чёрные машины ломаются примерно в три раза чаще белых, а дальше начнём искать причины этого явления. Возможно даже, придумаем что-нибудь вроде того, что чёрный цвет лучше скрывает дефекты кузова, поэтому им красят уже дефективные машины и т.п. Но на самом деле, главное – кто выпустил машину. Если первый производитель, то она не ломается, а если второй – то вероятность поломки 0.2 независимо от цвета. Из данных табл. 2 это, естественно, не следует! **Не всегда предоставленные данные говорят о причинах тех или иных значений.**

машины:	**белые**	**чёрные**
поломок	6.7%	20%

Табл. 2. Процент поломок машин.

Хорошо, а если у нас есть не табл. 2, а полная статистика: описания всех машин с пометкой были ли поломки? Мы можем посмотреть, как зависят друг от друга различные признаки: факт поломки, цвет машины, вес машины, стоимость, производитель, модель и т.д. Но как понять (используя только данные, а не наши знания об их природе), какие признаки первичны, т.е. от них может что-то зависеть, а какие вторичны, т.е. это они зависят от других? Вот здесь у Kaggle есть чему поучиться. В 2013 году на этой платформе проводилось соревнование по поиску зависимостей [9]. Пусть есть таблица с двумя колонками (см. табл. 3 или 4). Необходимо разработать алгоритм, который определяет,

[5] Математики называют такие примеры вырожденными, поскольку некоторые параметры здесь доведены до крайности (один производитель выпускает 0% чёрных машин, его надёжность – 100%). Можно рассмотреть и невырожденный абсолютно реальный пример, но он запутает читателя.

зависит ли A от B ($B \to A$), B от A ($A \to B$), есть ли взаимная зависимость ($B \leftrightarrow A$) или они независимы ($A \mid B$). Данные в соревновании реальные (хотя были и модельные – специально сгенерированные). Возникает естественный вопрос, а что такое «зависимость»? Допустим, имеется в виду функциональная зависимость, скажем

$$A = F(B), \tag{1}$$

где F – некоторая функция. Подобные закономерности есть в реальных данных, поскольку вид (1) напоминает нам элементарные физические законы. Скажем, напряжение зависит от силы тока (закон Ома), удлинение стержня зависит от силы его растяжения (закон Гука) и т.п. С точки зрения математики, если $A = F(B)$, то $B = F^{-1}(A)$, если функция F – обратима, а F^{-1} – её обратная. Так в табл. 3

$$B = 2*A, \text{ но и } A = B/2.$$

Даже если функция не обратима (табл. 4), то всё равно с точки зрения логики зависимость часто бывает взаимной, скажем в табл. 4 $B=A^2$, но и $A=c\sqrt{B}$, где c – коэффициент, который случайно принимает значения ±1. Скажем, Вы говорите положительное число, а я извлекаю из него корень, потом подбрасываю монету и если она выпадает орлом, то ставлю перед корнем знак минус. Тогда ответ зависит от того, какое число Вы мне дадите (пусть и не определяется однозначно), а не наоборот.

A	***B***
1	2
2	4
3	6
4	8

Табл. 3. Пример одной таблицы.

A	***B***
2	4
-1	1
3	9
–2	4

Табл. 4. Пример другой таблицы.

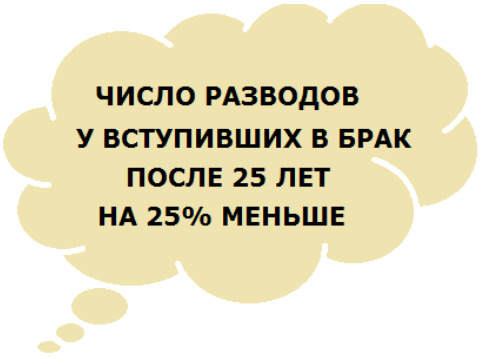
Организаторы соревнования [9] сделали очень просто. Они решили, что в реальных данных, снабжённых контекстом (описанием, как они получены), очевидно что от чего зависит. Действительно, пусть в левом столбце таблицы – высота горы, в правом – температура на этой высоте, измеренная с помощью термометра. Вопрос, что от чего зависит? Очевидно, что температура от высоты, а не наоборот. Почему? Мы можем поменять высоту (просто спустившись с термометром с горы или поднявшись на гору), при этом меняется температура (и показания термометра, который её фиксирует). Но нельзя

сделать наоборот: поменять температуру, чтобы изменилась высота. Таким образом, мы можем нагенерировать много двухстолбцовых таблиц с реальными данными (длина тормозного пути и скорость автомобиля, материал детали и срок службы, возраст пациента и давление и т.п.), в которых мы чётко знаем, что от чего зависит. И теперь задаться целью построить алгоритм, который сам по этим таблицам понимает, есть ли зависимости и какие они[6]. Этот алгоритм можно будет потом использовать для данных, природа которых недостаточно изучена. Он подскажет, зависит ли скорость реабилитации больного от дозы приёма лекарства, что из социальных и экономических показателей влияет на настроения в обществе и т.д. Весь вопрос: какова будет точность алгоритма? Для этого и проводилось соревнование. Этот вопрос также изучает целое направление в теории построения обучающихся алгоритмов «Causal Relations». На наш взгляд, **это очень оригинальные постановка вопроса и направление исследований**. Кстати, алгоритм удалось построить достаточно неплохой (точность больше 70%).

Возвращаясь к нашей задаче об оранжевых машинах, если мы нашли корреляции в данных и объявляем их зависимостями (не факт, что мы имеем право так делать), то надо хотя бы правильно указывать направления зависимостей. Скажем, если есть зависимости между качеством машины, производителем и цветом, то, наверное, первичный признак здесь производитель.

Когда лучше вступать в брак?

Во многих группах в социальных сетях участникам рассылают удивительные статистические факты, например такой:



Опять же, они подвергаются обывательской интерпретации. Например, что человек, который

[6] Такие алгоритмы строятся методами **машинного обучения**. К сожалению, подробнее об их построении мы здесь писать не будем.

женится после 25, уже взрослый и подходит ко всему более ответственно. Не будем подвергать такие выводы критике, проведём лишь один эксперимент.

Для начала отметим, что наш «статистический факт» должен быть пояснён: как вычисляются эти проценты? Наверное, «вступившие в брак после 25» это не только те, кто сделал это впервые, но и разведённые, после 25 снова вступившие в брак. Иначе было бы написано «...впервые вступивших...». Поэтому с интерпретацией надо быть осторожней, поскольку такой малый процент разводов в поздних браках может быть связан с тем, что «после первого брака люди с большей ответственностью подходят к семейной жизни / выбору партнёра и т.п.» Говорить, какая из интерпретаций верна (и насколько), можно лишь имея на руках всю статистику по бракам и разводам. Как и раньше: **объективно есть только данные, а выводы могут быть и необъективны**.

Но сейчас мы обойдёмся без данных! Просто проведём опыт. Допустим, где-то живут инопланетяне. Жизнь каждого из них длится ровно 60 лет. В любой год своей жизни среднестатистический инопланетянин меняет свой семейный статус с вероятностью 0.05 (т.е. вступает в брак, если в нём не был, и разводится, если был). Это, конечно, не означает, что инопланетяне подбрасывают многогранник с 20 гранями, одна из которых помечена, и меняют семейный статус, когда выпадает помеченная грань. Просто статистика такая, что среди всех женатых каждый год примерно 5% разводится, а среди неженатых – 5% женится.

Не будем забивать голову читателя формулами, просто сделаем компьютерный эксперимент: сгенерируем матрицу из 0 и 1 размера 100000×60. Идём по строкам слева-направо. Первый элемент можно положить равным нулю (или единице с вероятностью 0.05), следующие элементы заполняем так: если слева стоит элемент x, то его же пишем справа с вероятностью 0.95, а с вероятностью 0.05 – элемент $(1–x)$. На рис. 7 показано, что получилось. Нетрудно видеть, что матрица моделирует изменение социальных статусов для 100000 инопланетян, в её i-й строке и j-м столбце элемент равен 0, если i-й инопланетянин на j-м году жизни не женат, и 1 – в противном случае.

000000000000000000001111111110111111000000000000000000000
000000000000000000000000000000000000011111111111111111111
001111111111111111111111111111100000000000000000000001000000

Рис. 7. Первые строки, полученной матрицы.

Что получилось: 4.6% инопланетян никогда не вступали в брак (это соответствует строкам, состоящим из одних нулей). Удалим их из матрицы, поскольку процент неудачных

браков надо вычислять из общего числа браков. Около 84.8% инопланетян в своей жизни хотя бы раз разводятся (это соответствует тому, что в 84.8% строк есть фрагмент ...10... – на рис. 7 выделен красным). Тех, кто до 25 лет не женился – около 24% (в строке 25 первых нулей – на рис. 7 показаны зелёным цветом), среди них тех, кто развёлся – 63.6% (в строке 25 первых нулей, а потом есть фрагмент ...10...).

Сравните: 84.8% и 63.6% (см. рис. 8). Эта разница не объясняется сознательностью инопланетян после 25 или после первого брака. В мире, где решения о браке могут приниматься выбрасыванием многогранника, нет сознательности! Просто чем раньше вступишь в брак, тем больше шансов развестись (при постоянной вероятности развода и ограниченной длине жизни) – **разница объясняется устройством мира!**

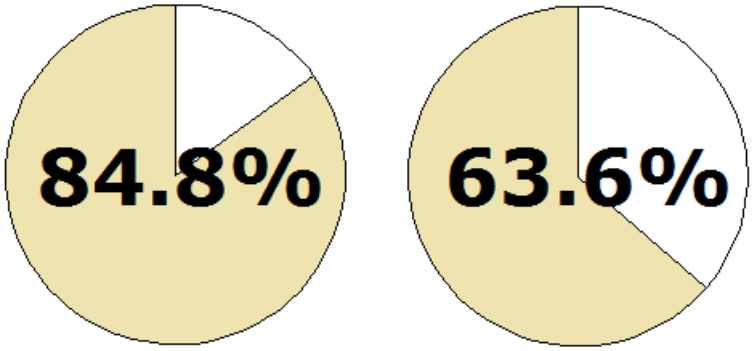


Рис. 8. Средний процент разводов (слева) и процент разводов у вступивших в брак после 25 (справа).

Конечно, наш пример искусственный. Люди, в отличие от вымышленных инопланетян, не могут вступать в брак уже на первом году жизни, не живут фиксированное число лет и т.п. Но мы просто продемонстрировали существование определённого эффекта.

Те, кто знает теорию вероятностей или комбинаторику, понимают, что можно обойтись без генерации матриц на компьютере. Вероятность остаться холостым у инопланетянина

$$(1-0.05)^{60} = 0.046\ldots,$$

что соответствует нашим 4.6%. Все остальные проценты можно также вывести аналитически. При желании и умении читатель может это сделать самостоятельно.

Заключение

Итак, не верьте всем интерпретациям данных или найденным закономерностям, которыми пестрят СМИ и Интернет. Понять по набору чисел причины явлений и истинные закономерности порой очень сложно или даже невозможно. Всегда надо думать своей головой и лучше оперировать с исходными данными, а не посчитанными «специально для Вас» характеристиками этих данных. Тем более не стоит, начитавшись статей, покупать оранжевые машины, тянуть с предложением любимой или бежать в банк за кредитом.

Удачи! Всем правильных выводов!

Ссылки и литература

[1] Зодиакальный скоринг http://www.banki.ru/news/daytheme/?id=7408493

[2] Исследование MoneyMan: нужны ли займы Львам, Рыбам и Скорпионам http://moneyman.ru/articles/goroskop-moneyman

[3] Система математических вычислений Matlab http://www.mathworks.com/

[4] 20 неожиданных открытий, сделанных благодаря анализу данных http://slon.ru/specials/data-economics/articles/20-unexpected-discoveries/

[5] Большие данные: новый облик человечества http://ichip.ru/bolshie-dannye-novyjj-oblik-chelovechestva.html

[6] Соревнование по анализу данных http://www.kaggle.com/c/DontGetKicked/

[7] В. Майер-Шенбергер, К. Кукьер Большие данные: Революция, которая изменит то, как мы живем, работаем и мыслим // Изд-во Манн, Иванов и Фербер, 2013 г.

[8] Жёлтые подержанные машины чаще подвергаются поломкам http://www.telegraph.co.uk/motoring/news/9060066/Yellow-second-hand-cars-most-likely-to-be-defective.html

[9] Соревнование по анализу данных http://www.kaggle.com/c/cause-effect-pairs

Работа является финалистом конкурса STRF научно-популярных статей и опубликована (23.07.2015) в блоге автора по адресу

https://alexanderdyakonov.wordpress.com/2015/07/23/неправильные-интерпретации/

– здесь в комментариях можно писать о замеченных ошибках. Спасибо!