**Определение выбросов**

Выбросы — точки, сильно выбивающиеся из последовательности, не вписывающиеся в модель по какой-либо причине. Люди озаботились выбросами, желая избавиться от них. Выбросы оттягивают на себя средние значения и сдвигают распределение данных.[[1]](#footnote-1) Иногда выбросы – просто ошибка ввода. Поэтому первая причина обращать внимание на выбросы — это обеспечение чистоты анализа данных и моделирования. Но есть и другая. Они интересны сами по себе!



Рис. 1. Вероятность 349-дневной беременности

Представьте, что после подозрительной транзакции вам звонит компания, выпустившая вашу кредитную карту. Она определяет эту транзакцию как выброс, основываясь на вашем поведении в прошлом. И вместо игнорирования транзакции как выброса, она помечает ее как потенциально мошенническую и начинает действовать соответственно.

Как найти «выбивающиеся» точки? В кластерном анализе (см. главы [2](http://baguzin.ru/wp/?p=13802) и [5](http://baguzin.ru/wp/?p=13926)) нас интересовали группы. При определении выбросов вам важны точки, которые отличаются от групповых. Они выделяются, являясь в некотором роде исключениями.

### Захватывающее дело Хадлум против Хадлум

В 1940-х годах, британец по имени мистер Хадлум ушел на войну. Спустя 349 дней его жена, миссис Хадлум, родила. В среднем беременность длится 266 дней. Таким образом получается, что миссис Хадлум носила ребенка лишних 12 недель. Исключительно долгая беременность, не правда ли? Именно так и утверждала миссис Хадлум.

Исследования показали, что продолжительность беременности — величина с более-менее нормальным распределением со средним значением в 266 дней после оплодотворения и стандартным отклонением около 9. Так что можно вычислить значение функции нормального интегрального распределения, и получить вероятность 349-дневной беременности. В Excel для этого используется функция НОРМРАСП (рис. 1). Мы получили 0,000000 насколько хватает глаз. Другими словами, человеческого детеныша практически невозможно вынашивать так долго.

### Границы Тьюки

Концепция того, что выбросы — это маловероятные точки на колоколообразной кривой, привела к границам Тьюки. Метод Тьюки заключается в следующем:

* Рассчитать 1-й и 3-й квартиль (или, что тоже самое, 25-й и 75-й персентиль) ряда данных, в котором вы хотите найти выбросы.
* Вычесть первый квартиль из третьего — получится мера распределения данных, называемая межквартильным размахом (МР). МР устойчив к экстремальным значениям распределения (то есть робустный), в отличие от стандартной ошибки, которая более чувствительна к выбросам.
* Рассчитать нижнюю и верхнюю внутренние границы, отстоящие на 1,5МР от квартилей.
* Рассчитать нижнюю и верхнюю внешние границы, отстоящие на 3МР от квартилей.
* Значение за пределами границ – экстремально. Нормальное распределение дает одну точку на 100 вне внутренних границ, и одну на 500 000 – вне внешних.

(Метод Тьюки очень напоминает контрольные карты Шухарта; см. [У. Эдвардс Деминг. Выход из кризиса: Новая парадигма управления людьми, системами и процессами](http://baguzin.ru/wp/?p=2138), [Д. Уилер, Д. Чамберс Статистическое управление процессами](http://baguzin.ru/wp/?p=236), [Пример построения контрольной карты Шухарта в Excel](http://baguzin.ru/wp/?p=1479).)

### Применение границ Тьюки

Срок беременности миссис Хадлум выходит далеко за границы Тьюки (рис. 2).



Рис. 2. Границы Тьюки для продолжительности беременности

Примените к исходным данным условное форматирование, чтобы выделить значения, выходящие за пределы границ (рис. 3). Миссис Хадлум «покраснела», т.е., ее беременность экстремальна. Прокрутив данные до конца, вы увидите, что красных ячеек больше нет, но есть девять желтых. Это очень похоже на примерно одну точку из 100, что и предполагалось правилом Тьюки.



Рис. 3. Добавление условного форматирования выбросов

### Ограничения метода Тьюки

Границы Тьюки работают при выполнении всех трех условий:

1. Данные распределены более-менее нормально. Распределение может не быть идеальным, но кривая должна быть колоколообразной и близкой к симметричной, без разных длинных хвостов, вылезающих с одной стороны (подробнее см. [Крис Андерсон. Длинный хвост. Эффективная модель бизнеса в Интернете](http://baguzin.ru/wp/?p=4273)).
2. Выброс «отмечен» как экстремальное значение на внешней стороне распределения.
3. Данные одномерны.

Рассмотрим пример выброса, который не удовлетворяет первым двум условиям. В «Братстве Кольца», объединившись, наконец, в одну компанию (братство, в честь которого и названа книга), герои встают небольшой группой и слушают лидера эльфов, произносящего речь о том, кто они есть и какова их цель. В этой группе есть четверо высоких ребят — Гэндальф, Арагорн, Леголас и Боромир — и четверо приземистых. Это хоббиты: Фродо, Мерри, Пиппин и Сэм. Среди них есть один гном — Гимли. Гимли ниже первых на две головы, но выше вторых примерно настолько же (рис. 4). В фильме, когда мы впервые видим эту группу, Гимли явно выделяется по росту. Он не принадлежит ни к одной из групп.



Рис. 4. Гимли, сын Глойна, «гномий» выброс

Но является ли он выбросом? Его рост не ниже всех и не выше. На самом деле, его рост ближе всего к среднему в группе. Это распределение роста ничем не похоже на нормальное. Можно назвать его «мультимодальным» (распределение с несколькими пиками). И Гимли является выбросом не из-за выдающегося роста, а всего лишь оттого, что находится между двумя пиками. А если данные многомерны, то такие точки в них найти еще сложнее.

Такой вид выбросов довольно часто обнаруживается в случаях мошенничества. Кто-то слишком обычный, чтобы быть обычным. [Берни Мейдофф](https://ru.wikipedia.org/wiki/%D0%9C%D0%B5%D0%B9%D0%B4%D0%BE%D1%84%D1%84%2C_%D0%91%D0%B5%D1%80%D0%BD%D0%B0%D1%80%D0%B4) — отличный пример такого выброса. Если в большей части [схем Понци](https://ru.wikipedia.org/wiki/%D0%9C%D0%B5%D0%B6%D0%B4%D1%83%D0%BD%D0%B0%D1%80%D0%BE%D0%B4%D0%BD%D1%8B%D0%B9_%D0%BE%D1%82%D0%B2%D0%B5%D1%82%D0%BD%D1%8B%D0%B9_%D0%BA%D1%83%D0%BF%D0%BE%D0%BD#.D0.A4.D0.B8.D0.BD.D0.B0.D0.BD.D1.81.D0.BE.D0.B2.D0.B0.D1.8F_.D0.BF.D0.B8.D1.80.D0.B0.D0.BC.D0.B8.D0.B4.D0.B0_.D0.9F.D0.BE.D0.BD.D1.86.D0.B8) предлагался размер выплат более 20% сверху, больше похожий на выброс, то Мейдофф стал предлагать скромные надежные выплаты, смешивающиеся с шумом каждый год — он не перепрыгивал никаких границ Тьюки. Но постепенно эти выплаты из-за своей надежности превратились в многомерный выброс.

Как же находить выбросы в случае многомерности данных? Один из способов решения этого вопроса — отнестись к данным как к графу (см. [Кластерный анализ: сетевые графы и определение сообществ](http://baguzin.ru/wp/?p=13926)). Что определяет Гимли как выброс относительно остальных точек данных — это расстояние от него до них относительно расстояния между ними самими. Все эти расстояния, от одной точки до другой, определяют ребра графа. С его помощью можно «выманить» изолированные точки. Для этого нужно начать с создания графа k-ближайших соседей.

### Исходные данные сотрудников колл-центра

Представьте себя в роли управляющего центром поддержки клиентов. Каждый звонок, электронное письмо или сообщение в чате от покупателя создает жетон. Один работник службы поддержки должен обрабатывать не меньше 140 жетонов в день. В конце каждого диалога клиенту дается возможность оценить работника по пятибалльной шкале. Сотрудники должны поддерживать свой рейтинг не ниже 2, иначе их уволят.

Компания также следит за множеством других показателей своих работников. Сколько раз они опоздали в этом году? Сколько у них было ночных смен и смен, приходящихся на выходные? Сколько они брали больничных и какое количество из них приходилось на пятницу? Компания знает, сколько часов сотрудник тратит на внутренние обучающие курсы (компания оплачивает до 40 часов) и сколько раз он отклонял запрос на замену смен или шел навстречу просьбам коллег. Данные на каждого из 400 сотрудников службы поддержки содержатся в таблице (рис. 5). Вопрос в том, кто из них является выбросом и какие выводы о его работе можно сделать?



Рис. 5. Показатели сотрудников

### Нормализация данных

Поскольку каждый столбец имеет свои единицы измерения, нужно стандартизировать данные:

* вычесть среднее значение столбца из каждого наблюдения, а затем…
* разделить каждое наблюдение на стандартное отклонение столбца (рис. 6).

Для нормально распределенных данных среднее = 0, а стандартное отклонение = 1.

*Примечание.* Выбросы могут испортить среднее значение и стандартное отклонение, поэтому иногда специалисты предпочитают стандартизировать данные вычитанием медианы и делением на размах.



Рис. 6. Среднее значение и стандартное отклонение для каждого параметра

Создайте новый лист, и отразите на нем стандартизованные значения. Для этого можно использовать функцию НОРМАЛИЗАЦИЯ (рис. 7).



Рис. 7. Нормализованный набор данных о работе персонала

### Создание графа

Граф есть не что иное как ребра и вершины. В нашем случае каждый работник является вершиной, и для начала можно просто провести ребра между ними всеми. Длина ребра — это евклидово расстояние между двумя работниками в их нормализованных данных. Евклидово расстояние (подробнее см. [главу 2](http://baguzin.ru/wp/?p=13802)) между двумя точками — это квадратный корень из суммы квадратов разниц каждого столбца для каждого из них.

На новом листе с помощью формулы СМЕЩ создайте матрицу расстояний между сотрудниками (рис. 8).



Рис. 8. Матрица расстояний между сотрудниками

Формулы в диапазоне С3:ОL402 – это формулы массива, поэтому их нужно вводить нажатием клавиш Ctrl+Shift+Enter (подробнее о формулах массива см. [Майкл Гирвин. Ctrl+Shift+Enter. Освоение формул массива в Excel](http://baguzin.ru/wp/?p=7153)). Евклидово расстояние работника 144624 до самого себя равно 0. У вас есть граф сотрудников. Можете экспортировать его в Gephi как в [главе 5](http://baguzin.ru/wp/?p=13926), и взглянуть на то, что получится, но так как у него 16 000 ребер и только 400 вершин, это наверняка жуткая картина.

Мы сфокусируемся лишь на k ближайших соседях каждого сотрудника, чтобы найти выбросы. Первый шаг — ранжирование, то есть расположение работников согласно расстояниям относительно друг друга. Это приводит нас к первому и основному методу выделения выбросов на графе.

### Вычисляем k ближайших соседей

Создайте новый лист (рис. 9). Вставьте персональные номера от А1 вниз и от В1 вправо, чтобы получилась основу матрицы. Ранжируйте каждого работника в верхней строке в соответствии с его расстоянием до работников в столбце А. Начните с 0 таким образом, чтобы ранг 1 оказался у других работников, а 0 образовали диагональ графа (так как расстояния до самих себя всегда самые короткие).



Рис. 9. Каждый сотрудник в столбце ранжирован относительно каждой строки

### Определение выбросов на графе, метод 1: полустепень захода

Подсчитайте, как часто каждый сотрудник попадает в ТОП-5 ближайших соседей, а в ТОП-10? ТОП-20? То есть сколько человек выбирают его одним из 5 ближайших соседей, или 10, или 20? Полустепень захода — это количество ребер, входящих в любую вершину графа. Чем ниже полустепень захода, тем больше вы похожи на выброс, потому что никто не хочет быть вашим соседом (рис. 10). Конечно, вы можете выбрать любое k, но на этот раз вполне достаточно 5, 10 и 20. С помощью условного форматирования выделите ячейки, чье значение равно 0 (что значит отсутствие входящих ребер для этой вершины на графе такого размера).



Рис. 10. Полустепени захода трех разных графов (ТОП-5, ТОП-10, ТОП-20) ближайших соседей

Оказалось, что существуют всего двое сотрудников, у которых даже на графе 20 ближайших соседей нет ни одного входящего ребра. Никто не считает их соседями, даже 20 ближайших. Это довольно далеко! Персональные номера этих двоих ребят — 137155 и 143406. Переключившись обратно во лист *Сотрудники*, изучите их внимательнее (рис. 11). Сотрудник 137155 имеет большое среднее количество жетонов, высокий клиентский рейтинг и он кажется добрым самаритянином. Он брал много смен в выходные, ночью и семь раз заменял тех, кто просил его об этом. Мило! Вот кто во всех измерениях достаточно исключителен, чтобы не попасть даже в 20 ближайших расстояний до любого другого сотрудника. Удивительно! Наверное, такие работники заслуживают пиццы или чего-то подобного.



Рис. 11. Данные о работе сотрудников 137155 и 143406

Сотрудник 143406 явно контрастирует с предыдущим служащим. Никакие параметры сами по себе не являются достаточными условиями для увольнения, но количество его жетонов на два стандартных отклонения ниже среднего, клиентский рейтинг тоже примерно на пару отклонений ниже распределения. Число опозданий выше среднего и пять из шести больничных приходятся на пятницу. А еще он просил заменить себя 4 раза, но ни разу не откликнулся на просьбы коллег. Такие сотрудники остаются на местах благодаря системе. Удовлетворяя минимальным требованиям трудоустройства (заметьте, что здесь никто не переходит границ Тьюки), они плавно скатываются к краям любого распределения.

### Определение выбросов на графе, метод 2: нюансы k-расстояния

Одним из недостатков метода, описанного выше, является то, что мы лишь определили, имеет ли сотрудник ребра на графе k ближайших соседей. Было бы неплохо иметь количественную оценку «выбросовости». Рассмотрим ранжирование выбросов с помощью величины под названием k-расстояние. K-расстояние — это расстояние от сотрудника до его k-го соседа (рис. 12). Видно, что плохой работник 143604, оказывается существенно дальше, чем 137155, и оба эти значения сильно превосходят следующее по величине значение, равное 3,54.



Рис. 12. Сотрудники с k-расстоянием более 3,3

Все же у этого метода есть недостаток (рис. 13). Хотя треугольная точка — самый что ни на есть выброс, но все же его k-расстояние меньше, чем у некоторых ромбов. Дело в том, что треугольник не является глобальным выбросом, потому что он — локальный выброс. Причина, по которой ваши глаза зафиксировали в качестве странного именно его, заключается в том, что он ближе к кластеру кружков. Если бы треугольник оказался среди ромбов, все было бы иначе.



Рис. 13. k-расстояние не работает с локальными выбросами

### Определение выбросов на графе, метод 3: факторы локальных выбросов

Как и k-расстояния, факторы локальных выбросов (ФЛВ) дают по одному «баллу» за каждую точку. Чем больше баллов, тем в большей степени выбросом является точка. Но ФЛВ дает вам нечто немного большее: чем ближе балл к единице, тем более обыкновенна данная точка локально. С ростом балла точка считается все менее типичной и более относящейся к выбросам.

На высоком уровне это работает так: ты являешься выбросом, если k твоих ближайших соседей считают тебя дальше, чем их соседи считают их самих. Алгоритму интересны друзья точек и их друзья. Так он понимает слово «локальный». Анализируя рис. 13, можно понять, что именно делает треугольник выбросом. У него может не быть лучшего k-расстояния, но отношение расстояния треугольника к его ближайшим соседям по сравнению с их расстояниями друг до друга довольно велико (рис. 14).



Рис. 14. Треугольник не настолько далек от своих соседей, как далеки они друг от друга

**Расстояние достижимости.** Перед тем, как свести вместе все факторы локальных выбросов для каждого сотрудника, нужно рассчитать еще одну последовательность чисел — расстояние достижимости. Расстояние достижимости сотрудником А сотрудника В — это обычное расстояние между ними, пока А не оказывается по соседству с В, то есть удаленным на k-расстояние, которое и превращается в таком случае в расстояние достижимости. Другими словами, если А попадает в окрестность точки В, вы округляете расстояние от А до В до размера окрестности, а если же нет — то просто оставляете без изменений.

Использование расстояния достижимости вместо обычного расстояния при расчете ФЛВ помогает немного стабилизировать вычисления. Создайте новый лист *Достижимость* и замените расстояния из вкладки *k-расстояние* новыми расстояниями достижимости (рис. 15).



Рис. 15. Расстояния достижимости

Теперь все готово к вычислению факторов локальных выбросов для каждого сотрудника. Создайте новую вкладку *ФЛВ* (рис. 16). В столбце В отображается то, как выглядит каждый сотрудник с точки зрения пяти своих ближайших соседей. Фактор локального выброса для сотрудника (столбец С) — это среднее значение отношений расстояний достижимости этого сотрудника, разделенное на средние достижимости каждого из k своих соседей.



Рис. 16. Средняя достижимость каждого сотрудника относительно его соседей; фильтром выделены значения с ФЛВ > 1,4

Формула в ячейке В375: {=СУММ(ЕСЛИ(Ранг!B375:OK375<='К-расстояние'!B$1;1;0)\*ЕСЛИ (Ранг!B375:OK375>0;1;0)\*Достижимость!B377:OK377)/'К-расстояние'!B$1}

Формула в ячейке С375: {=СУММ(ЕСЛИ(Ранг!B375:OK375<='К-расстояние'!B$1;1;0)\*ЕСЛИ (Ранг!B375:OK375>0;1;0)\*B375/ТРАНСП(B$2:B$401))/'К-расстояние'!B$1}

Чем больше единицы значение в столбце С, тем больше данный сотрудник является локальным выбросом. Например, сотрудник 143406, штатный халтурщик, — самая отдаленная точка с ФЛВ, равным 1,97. Его соседи видят его вдвое дальше, чем другие соседи видят их самих. Это довольно далеко от группы.

Теперь у вас есть по одному значению, соотнесенному с каждым сотрудником, которое ранжирует его как локальный выброс и измеряется независимо от размера графа.

1. Написано по материалам книги Джона Формана [Много цифр: Анализ больших данных при помощи Excel](http://baguzin.ru/wp/?p=13802). – М.: Альпина Паблишер, 2016. – С. 383–409 [↑](#footnote-ref-1)