

Отделение рекламы и связей с общественностью, 2014-15 уч. год

Теория вероятностей и статистика

Лекция 8. Статистические парадоксы. Элементы анализа данных ()

И. А. Хованская, Р. Я. Будылин, И. В. Щуров, Д. А. Филимонов, К. И. Сонин (РЭШ)

Анализируя те или иные события и тенденции, мы часто обращаемся к статистическим данным. Используя данные, мы делаем по ним выводы о происходящем — порой выводы с очень серьезными последствиями. Например, если речь идет об анализе последней какой-то конкретной правительственной программы (потрачено столько-то денег, показатели сменились с таких-то на такие-то), выводы о причинно-следственных связях между затратами и предпринятыми усилиями, с одной стороны, и изменением показателей — с другой, будут играть существенную роль при принятии решений о будущих проектах.

Несмотря на то, что для обработки данных наука знает тысячи методов, математики и экономисты пишут статьи в десятки научных журналов, по всему миру работают огромные исследовательские центры, часто кажется, что все это лишнее. Многим кажется, что достаточно беглого и, уж конечно, достаточно пристального взгляда на данные, чтобы понять, какие существуют связи, какие есть тенденции, и что нужно делать. Действительно, зачастую такой взгляд на данные позволяет сделать важные, очевидные, но, к сожалению, совершенно неверные выводы.

В этой лекции мы приводим примеры нескольких совершенно элементарных «статистических парадоксов» — ситуаций, когда «здравый смысл» и «интуиция» подсказывают нам совсем не то, что нам пытаются рассказать реальные данные. Эта лекция не научит вас методам статистики — для этого необходимы как минимум полноценный курс анализа данных. Эта лекция должна научить вас осторожности.

1 Парадокс Симпсона

Рассмотрим пример, описанный известным популяризатором математики Мартином Гарднером. Пусть мы имеем четыре набора камней. Вероятность вытащить черный камень из набора 1 выше, чем из набора 2. В свою очередь, вероятность вытащить черный камень из набора 3 больше, чем из набора 4. Это значит, что в наборе 1 чаще встречаются черные камни, чем в наборе 2, а в наборе 3 чаще, чем в наборе 4. Объединим набор 1 с набором 3 (получим набор I), а набор 2 — с набором 4 (набор II). Совершенно очевидно, что вероятность вытащить черный камень из набора I больше, чем из набора II, ведь в набор I попали те наборы, где черные камни были «гуще»! Однако, в общем случае такое утверждение неверно, и мы сейчас приведем пример расположения черных и белых камней в наборах, когда это совершенно «очевидное» условие не выполнится. Пусть в наборе 1 6 черных и 7 белых камней, в наборе 4 4 черных и 5 белых камней, в наборе 3 6 черных камней и 3 белых, в наборе 4 9 черных камней и 5 белых. Тогда в наборе 1 черные камни «гуще», вероятность вытащить черный камень из первого набора выше. Действительно, в наборе 1 всего 13 камней, из них 6 черных, значит, вероятность вынуть черный камень составляет $6/13$. В наборе 2 всего 9 камней, из них 4 черных, значит, вероятность вынуть черный камень составляет $4/9$. Мы видим, что $6/13 > 4/9$, значит, вероятность вынуть черный камень из набора 1 выше. Сравним теперь наборы 3 и 4. В наборе 3 6 черных камней, а всего камней 9, значит вероятность выбрать черный камень составляет $6/9$. Для набора 4 эта вероятность равна $9/14$. Мы видим, что $6/9 > 9/14$. Итак, наборы 1, 2, 3

и 4 удовлетворяют тем требованиям, которые мы предъявили в постановке задачи. Теперь смешаем наборы 1 и 3 (т. е. те, где черные камни лежат «гуще») и наборы 2 и 4, т. е. «менее черные» наборы. Каково же соотношение черных и белых шаров в наборе Π - смеси «более черных» наборов 1 и 3? Там будет 12 черных и 10 белых камней, т. е. вероятность вынуть черный камень равна $14/22$. Во наборе Π , собранном из «менее черных» наборов 2 и 4 будет 13 черных и 10 белых шаров, т. е. вероятность вынуть черный шар составит $13/23$. Итак, набор Π , собранный из «менее черных» наборов 2 и 4 загадочным образом оказался более черным!

Если разобранный выше пример выглядит несколько загадочно, то второй пример сделает паракос Симпсона простым и понятным. Рассмотрим опять 4 набора черных и белых шаров. Только теперь соотношение черных и белых шаров наборах будет таким (см. табл. 1).

	чёрные шары	белые шары
набор 1	20	10
набор 2	10	10
набор 3	5	1
набор 4	4000	1000

Таблица 1: Распределение чёрных и белых шаров

Мы видим, что снова выполнены все требования условия задачи: набор 1 «более черный», чем набор 2, набор 3 «более черный», чем набор 4. Однако, набор 4 все же более черный и чем набор 2, и чем набор 1, и он куда больше, чем оба эти набора. Т. е. именно состав набора 4 будет решающим в смеси наборов 2 и 4. Итак, мы видим, что набор I теперь состоит из 25 черных и 11 белых шаров, вероятность вынуть черный шар оттуда составляет $25/36 = 0,69444$. Набор Π состоит из 4010 черных и 1010 белых шаров, и он гораздо «более черный» — вероятность вынуть из этого набора черный шар составляет $4010/5020 = 0,7988$.

2 Феномен Уилла Роджерса

Феномен Уилла Роджерса близок к парадоксу Симпсона, фактически, в нем описывается тот же эффект, только в других теминах. Название этого парадокса основывается на следующей цитате, приписываемой комедианту Уиллу Роджерсу: «Когда оки покинули Оклахому и переехали в Калифорнию, они повысили средний интеллект обоих штатов» (оки — презрительное или просторечное название жителей Оклахомы).

Итак, вопрос. Пусть есть два множества чисел A и B , например, наборы IQ двух классов. Может ли такое быть, что при перемещении каких-то чисел из множества A в множество B среднее значение обоих этих множеств повысится? Задумаемся: для того, чтобы после перемещения кого-то из A в B среднее значение A повысилось, множеству A нужно «избавиться» от какого-нибудь маленького числа, классу A избавиться от плохих учеников. Может ли при этом повыситься средний уровень класса B ? Очевидный ответ — нет, не может — снова неверен. Действительно, рассмотрим такие множества: $A = \{10, 20, 30, 40, 50\}$ и $B = \{1, 2, 3, 4, 5\}$

Среднее значение множества A составляет

$$\frac{10 + 20 + 30 + 40 + 50}{5} = 30$$

Ясно, что если мы выкинем из множества A самое маленькое число 10, то среднее станет больше. Действительно, без 10 оно составит

$$\frac{20 + 30 + 40 + 50}{4} = 35$$

Однако, для множества B число 10 не маленькое, а наоборот — большое. Если без него среднее в множестве B составляло $(1 + 2 + 3 + 4 + 5) / 5 = 3$, то теперь среднее будет больше:

$$\frac{1 + 2 + 3 + 4 + 5 + 10}{6} = 4\frac{1}{6}$$

Итак, повышение среднего в обоих множествах возможно, и шутка Уилла Роджерса (за которую мы никак не отвечаем) означает, что жители штата Калифорния глупее самых глупых оки.

3 Парадокс Берксона

Парадокс Берксона или ошибка Берксона состоит в неверном заключении о зависимости событий A и B , если мы наблюдаем только те исходы, для которых выполняется хотя бы одно из событий A или B .

Пусть в городе A живет 100 девушек, из них 10 красавиц и 10 умниц, причем события «случайно выбранная девушка — умница» и «случайно выбранная девушка — красавица» независимы, т. е. ровно одна девушка и умница, и красавица. Действительно, если среди всех девушек $1/10$ часть составляют умницы, то для того, чтобы эти события были независимы, среди красавиц тоже должна быть $1/10$ умниц, т. е. умниц—красавиц будет $(1/10) \times 10 = 1$.

Посмотрим теперь на эту же ситуацию с точки зрения жителя соседнего города, которому рассказывают только об интересующих его девушках: умницах или красавицах. Он знает о 19 девушках из этого города: 10 умниц, 10 красавиц, но единственная умница—красавица посчитана и в том и в другом десятке. Итак, с точки зрения этого наблюдателя, вероятность того, что девушка умница составляет $10/19$, вероятность же, что девушка умница, при условии, что она красавица, составляет все ту же $1/10$. Итак, на взгляд этого наблюдателя, события зависимы, у них есть обратная связь — среди красавиц реже встречаются умницы, думает он, чем среди всех девушек.

Тот же эффект наблюдают врачи, биологи, многие специалисты, которые вполне всерьез наимаются данными. Врач сравнивает частоту тех или иных заболеваний среди тех, кто к нему обращается, т. е. среди тех, кто чем-то болен. Подумайте сами — в каких ситуациях вы сталкивались с тем, что делаются слишком поспешные выводы о подобной зависимости.