УДК 519.257

ПАРАДОКС СИМПСОНА

Лящук А.О., Шавель В. А.
Научный руководитель - Шавель Н.А., к.ф.-м.н., доцент

В любом наборе статистических данных может таиться то, что способно полностью перевернуть результат с ног на голову.

Теория вероятностей и математическая статистика представляют собой области математики, необычайно богатые парадоксами, порой настолько противоречащими здравому смыслу, что поверить в них трудно даже после того, как правильность их подтверждена доказательством [1].

Условно можно выделить два типа парадоксов. Первый тип – это так называемые истинные парадоксы, которые связаны с неоднозначной интерпретацией аксиоматики теории вероятностей. Второй тип – это парадоксы, в которых происходит интуитивное понимание условий в ошибочном ключе, в результате чего решение задачи в рамках аксиоматики кажется парадоксальным. Такие парадоксы называют контринтуитивными.

Парадокс Симпсона относится к контринтуитивным явлениям в статистике, когда при наличии нескольких групп данных, в которых наблюдается одинаково направленная зависимость, при объединении групп эта зависимость исчезает или даже меняется на противоположную.

Это явление было описано Эдвардом Симпсоном в 1951 году и Удни Юлом в 1903 году. Название «парадокс Симпсона» впервые предложил Колин Блит в 1972 году. Однако, так как Симпсон фактически не был первооткрывателем этого эффекта, некоторые авторы используют безличные названия, например, «парадокс объединений».

Данный парадокс часто встречается при анализе статистических данных в социальных науках и медицине и может приводить к совершенно ошибочным выводам в исследованиях.

Рассмотрим пример возникновения парадокса Симпсона в работе менеджера. В традиционных продажах уровень конверсии — это соотношение между количеством всех клиентов, которые проявили интерес к некоторому продукту, и тех из них, кто совершил покупку. Вся работа по продвижению в компании во многом направлена именно на увеличение этого показателя.

Предположим, что фирма занимается продажами неких продуктов А и В. Менеджер по продажам оценивает результаты работы с потенциальными клиентами, которых заинтересовали данные продукты.

Всего

Продукт	Количество клиентов	Количество продаж	Конверсия продаж
A	1000	80	8%
B	1000	100	10%

Анализируя 1000 обращений по продукту A и 1000 обращений по продукту B, менеджер оценивает, что в первом случае реальными покупателями стали 80 клиентов, тогда как во втором случае реальными клиентами стали 100 человек. Соответственно, уровень конверсии для продукта A составляет 8%, а для продукта B он равен 10%. Таким образом, можно сделать вывод, что продукт B определённо предпочтительней для компании по этому показателю.

Но старательный менеджер хочет дополнительно проанализировать эти результаты с гендерной точки зрения и рассчитывает эти данные по отдельности для клиентов-мужчин и клиентов-женщин. В итоге он получает следующие таблицы.

Мужчины

Продукт	Количество клиентов	Количество продаж	Конверсия продаж
A	100	24	24%
B	300	63	21%

Женщины

Продукт	Количество клиентов	Количество продаж	Конверсия продаж
A	900	56	6,2%
B	700	37	5,3%

То, что уровень конверсии продаж при работе с клиентами-мужчинами намного выше, чем при работе с клиентами-женщинами вполне ожидаемо, так как большинство мужчин не склонно слишком долго обременять себя изучением вариантов покупок. Неожиданно другое. Теперь продукт A стал определённо предпочтительней продукта B по уровню конверсии, причём как для клиентов-мужчин (24% для продукта A против 21% для продукта B), так и для клиентов-женщин (6,2% для продукта A против 5,3% для продукта B). Налицо парадокс Симпсона.

И неизбежный вопрос – какие данные следует учитывать менеджеру в ситуации принятия решения – агрегированные или разделённые? В общем случае нет единого ответа на этот вопрос. В исследованиях было показано, что в одних случаях правильнее учитывать агрегированные данные, а в других – разделённые. Таким образом, правильное решение ситуативно, важна история и причинно-следственные связи, лежащие в основе приведенных данных, и каждая история диктует свой выбор.

В нашем примере мы можем предположить, что так называемая «скрытая» переменная – пол клиента, возможно, оказывает серьёзное влияние на его заинтересованность продуктом A или B. Мы видим, что число женщин, интересующихся продуктом A, в 9 раз превышает число интересующихся им мужчин. В то же время для продукта B этот показатель чуть больше 2. Если менеджер оценивает ситуацию таким образом, что специфика продукта A, действительно, более интересна для женщин, что наиболее вероятно в нашей ситуации, то ему следует ориентироваться на разделённые данные. Если это не так, то предпочтительней агрегированные данные.

В заключение отметим, что большинство данных представляет собой лишь некую математическую модель гораздо более сложной области. Поэтому правильная обработка этих данных лежит на стыке математики, статистики, информатики и знаний в той конкретной сфере, к которой относятся анализируемые данные. Парадокс Симпсона демонстрирует важность продуманной интерпретации данных относительно реального мира, а также демонстрирует опасность упрощения более сложной картины в попытках решить все проблемы с единой точки зрения на данные.

Литература

1. Секей Г. Парадоксы в теории вероятностей и математической статистике — М.: Мир, 1990. - 240 с.