Парадокс Симпсона, Пестерев, 131-ПИо

Парадокс Симпсона — явление в статистике, когда мы видим в каждой из групп данных определенную зависимость, но при объединении этих групп зависимость исчезает или становится противоположной. Это явление было описано Эдвардом Симпсоном в 1951 году и Удни Юлом в 1903 году. Иногда это явление называют «парадоксом объединения». Удивительно, но эффект этого парадокса часто проявляется в области социологических наук и медицинской статистике, что иногда мешает проведению исследований и любой другой обработке собранных данных. Из-за этого парадокса часто могут делаться неверные выводы в исследованиях.

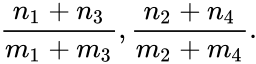
Согласно парадоксу, фактор, больше проявляющийся при любых фоновых условиях, чем противоположный ему, проигрывает менее эффективному, но относительно часто встречающемуся фактору. Эффект этого парадокса на удивление часто проявляется в области социологических наук и медицинской статистике; это происходит, когда весовая переменная не учитывается для одной группы, но должна использоваться при расчётах общих оценок.

**Пример М. Гарднера с камнями**

Пусть мы имеем четыре набора камней. Вероятность вытащить чёрный камень набора № 1 выше, чем из набора № 2. В свою очередь, вероятность вытащить чёрный камень из набора № 3 больше, чем из набора № 4. Объединим набор № 1 с набором № 3 (получим набор I), а набор № 2 — с набором № 4 (набор II). Интуитивно можно ожидать, что вероятность вытащить чёрный камень из набора I будет выше, чем из набора II. Однако, в общем случае такое утверждение неверно.

Математическое доказательство такое. Пусть {\displaystyle n_{i}}— число чёрных камней в i-ом наборе (выборке), {\displaystyle m_{i}}— общее число камней в i-ом наборе при i = 1,2,3,4. По условию:

{\displaystyle {\frac {n_{1}}{m_{1}}}>{\frac {n_{2}}{m_{3}}},{\frac {n_{3}}{m_{3}}}>{\frac {n_{4}}{m_{4}}}.}

Вероятность вытащить чёрный камень из наборов I и II, соответственно: 

Выражение для набора I не всегда больше выражения для набора II. Например: {\displaystyle n_{1}=6,~m_{1}=13,~n_{2}=4,~m_{2}=9,~n_{3}=6,~m_{3}=9,~n_{4}=9,~m_{4}=14}.

Легко проверить, что {\displaystyle 6/13>4/9,~6/9>9/14}. В то время как {\displaystyle 12/22<13/23}.

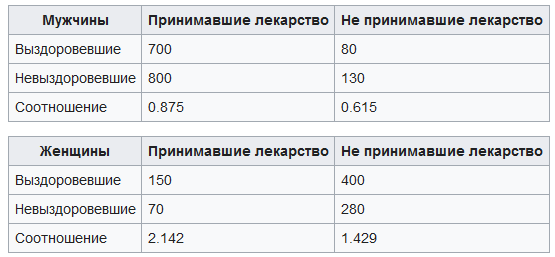
## Причины

Причина парадокса заключается в некорректном усреднении двух групп данных с различной долей контрольных наблюдений ([нерепрезентативная](https://ru.wikipedia.org/wiki/%D0%A0%D0%B5%D0%BF%D1%80%D0%B5%D0%B7%D0%B5%D0%BD%D1%82%D0%B0%D1%82%D0%B8%D0%B2%D0%BD%D0%BE%D1%81%D1%82%D1%8C) выборка). Поскольку интуитивно предполагается, что при применении найденных зависимостей доля контрольных будет одинаковой в обеих группах, а в исходных данных это не выполняется, то к ним нельзя применять арифметическое усреднение.

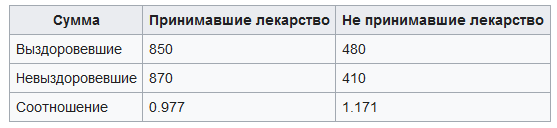
Парадокс Симпсона показывает, что выводы из результатов социологических опросов с нерепрезентативной выборкой нельзя принимать как неопровержимые, доказанные научным путём.

## Практическая значимость

Парадокс Симпсона иллюстрирует неправомерность обобщений по нерепрезентативным выборкам, иногда опасных для жизни. Так, например, в ходе эксперимента в группе мужчин и группе женщин, больных одной и той же болезнью, к стандартному лечению прибавили новый лекарственный препарат. Результат по обеим группам в отдельности подтверждал эффективность нового средства.



Интуитивно предполагается, что если в обеих группах прослеживается зависимость, она должна проявиться и при объединении этих групп. Но хотя соотношение выздоровевших и больных среди и женщин, и мужчин, принимавших лекарство, больше, чем среди тех из них, кто его не использовал, в связи с нерепрезентативностью контрольной группы в агрегированных данных эта закономерность не сохраняется.



Соотношение в агрегированных данных 850/870<480/410, то есть 0,977<1,171. Следовательно, доля выздоровевших среди принимавших лекарство меньше той же доли среди не принимавших.

Для устранения парадокса нужно обратить внимание, что отношение контрольной группы к группе воздействия в приведённых группах резко различается: у мужчин составляет (80+130)/(700+800) = 14 %, а у женщин (400+280)/(150+70) = 309 %.

Для корректного усреднения нужно обеспечить репрезентативность контрольной группы в обеих выборках, введя весовые коэффициенты так, чтобы взвешенная доля контрольных в обеих группах стала одинаковой. В данном случае достаточно количество мужчин, не принимавших лекарства, умножить на весовой коэффициент 22.07. Измененные таблицы будут выглядеть так:



Соотношение взвешенного количества выздоровевших к не выздоровевшим среди не принимавших лекарство в этом случае составит 0,685, то есть ниже, чем у принимавших лекарство. Это устраняет парадокс и показывает отношение выздоровевших к не выздоровевшим без приема лекарства для такой же пропорции мужчин и женщин, как у принимавших лекарство, что позволяет сравнивать эти цифры.