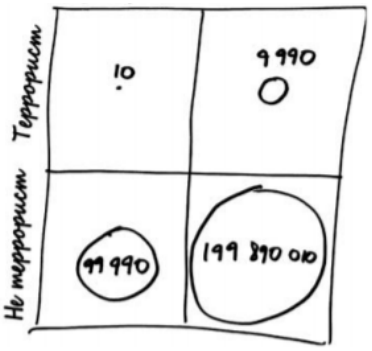
Статистика 2 (16.03.21)

На семинаре была рассмотрена таблица, составленная на основе результатов действия алгоритма Вконтакте.



Содержимое этой таблицы — около 200 миллионов пользователей сети Вконтакте. Линия между верхней и нижней частями матрицы отделяет будущих террористов (верхняя часть) от невиновных (нижняя часть). Безусловно, любая террористическая ячейка довольно немногочисленна. Скажем, если быть максимально подозрительными, в стране есть около 10 тысяч людей, за которыми федералам действительно стоит присматривать. Это один из каждых 20 тысяч пользователей общей пользовательской базы. Разделение матрицы на левую и правую часть, собственно, и есть то, что делает Вконтакте: с левой стороны находится сотня тысяч людей, которых специалисты считают с высокой степенью вероятности связанными с терроризмом.

Ответили на 2 простых вопроса:

Вопрос 1: какова вероятность, что человек попадет в список Вконтакте, при условии что он не террорист? (0,05%)

Вопрос 2: какова вероятность, что человек не террорист, при условии что он входит в список Вконтакте? (99,99%)

Прочувствовали парадокс:

невиновные редко попадают в список Вконтакте, казалось бы, их там должно быть мало, а их там 99,99%

Поняли, что так происходит из-за того, что заявленная нам «в 2 раза бОльшая вероятность» на самом деле ничтожно мала из-за того что вероятность того, что обычный пользователь окажется террористом совсем незначительна.

Дали философское определение p-значения:

p-значение — это ответ на 1 вопрос:

«Вероятность, что наблюдаемый результат эксперимента будет иметь место при условии, что нулевая гипотеза правильна».

Определили что такое — ответ на второй вопрос:

Описали его тем же философским языком для бОльшего понимания

«Вероятность, что нулевая гипотеза верна при условии наблюдения определенного результата эксперимента»

Увидели зависимость вероятности <<второго вопроса>> от p-value:

P-value	0.20	0.15	0.10	0.05	0.02	0.01	0.005	0.001
α(p)	0.465	0.436	0.385	0.289	0.175	0.111	0.067	0.018

Рассмотрели эксперимент с двумя играми и поняли, что для того, чтобы делать выводы о гипотезах нужно делать много экспериментов на больших выборках и обсудили проблему того, что сейчас в мире экспериментов делают недостаточно в силу тех или иных причин (этичность, например).

Посмотрели на распределение p-значения при верной основной и альтернативной гипотезе.