

Точные и приближенные вычисления при использовании перестановочных тестов

Роман Гученко

19 сентября 2020 г.

Современные электронно-вычислительные машины позволяют проводить достаточно сложный статистический анализ с использованием лишь минимальных знаний в области теории вероятностей¹. Рассмотрим сразу какую-нибудь конкретную задачу. Например, сравнение средних $\bar{x}$ и $\bar{y}$ для двух выборок^{2,3} $\{x_i\}_{i=1}^n$ и $\{y_i\}_{i=1}^m$. Допустим, что $\bar{x} < \bar{y}$. Мы хотим понять, насколько имеющаяся разница между средними значима, ведь даже если $\{x_i\}_{i=1}^n$ и $\{y_i\}_{i=1}^m$ — это на самом деле выборки из одного и того же распределения⁴, все равно существует вероятность, что мы могли получить значительно отличающиеся средние просто из-за случайности. Так как оценить эту вероятность? Составим из $\{x_i\}_{i=1}^n$ и $\{y_i\}_{i=1}^m$ одну большую выборку⁵, рассмотрим все возможные способы⁶ разделить эту составную выборку на две новые выборки $\{x'_i\}_{i=1}^n$ и $\{y'_i\}_{i=1}^m$ и посчитаем отношение разделений, в которых $\bar{y}' - \bar{x}' \geq \bar{y} - \bar{x}$, к общему числу разделений. Это будет вероятность того, что мы могли получить разницу в средних, которая $\geq \bar{y} - \bar{x}$ из-за случая⁷. Проблема в том, что число способов, которыми мы можем разделить составную выборку размера

¹В рамках первой главы книги [1]. Начать нужно будет с нее.

²Обязательно посмотрите замечательное объяснение этой задачи в картинках по ссылке: <https://www.jwilber.me/permutationtest/>.

³Чтобы не углубляться глубоко в двойственную природу выборок, давайте считать здесь, что выборка — это набор чисел, полученных в результате многократных вызовов некоторой абстрактной (псевдо)случайной функции с заданным распределением.

⁴То есть получены в результате вызова одной и той же функции.

⁵Выборку из того самого общего прото-распределения.

⁶Которые мы считаем равновероятными.

⁷Если эта вероятность маленькая (меньше наперед заданного числа в районе 0.05 или 0.1), то тест «отвергает гипотезу о том, что выборки получились из одного распределения» как маловероятную. Иначе тест «гипотезу не отвергает». Привыкаем к терминологии.

$n + m$ на две выборки размера n и m , очень быстро растет с ростом n и m . То есть оно растет как $(n + m)!$ — число перестановок из $n + m$ элементов. Отсюда, собственно, и название теста. Впрочем, интуитивно понятно, что если рассматривать не все перестановки, а только какое-то достаточно большое их случайное (равномерно выбранное) подмножество, то мы получим вполне пристойную оценку желаемой вероятности. Более того, в некоторых частных случаях можно получить и хорошую аналитическую оценку. В курсовой работе предлагается

- 1) изучить задачу сравнения средних двух выборок для последовательностей из нулей и единиц⁸ [сдача теории стоит 20/10 баллов в зависимости от курса; с этого необходимо будет начать],
- 2) написать программу, реализующую точный перестановочный тест⁹ (основанный на всех перестановках) [в двух вариантах: полным перебором и через гипергеометрическое распределение; 20/10 баллов]
- 3) и приближенный перестановочный тест (основанный на подмножестве перестановок) [еще 20/10 баллов],
- 4) а также аналитическую аппроксимацию [снова 20/10 баллов], основанную на предельной теореме 3 из документа по ссылке¹⁰.
- 5) Необходимо также будет исследовать, насколько приближенный тест и аналитическая аппроксимация отличаются от точного теста в смысле вычисляемой вероятности при различных значениях параметров [последние 20/10 баллов].

Список литературы

- [1] Ширяев А.Н. Вероятность - 1. Москва : МЦНМО, 2007, 552 стр.

⁸Количества единиц в первой и во второй выборках при разбиении в этом случае будут иметь гипергеометрические распределения. Воспользовавшись этим фактом, можно очень сильно упростить вычисления.

⁹Для этих целей хорошо подойдет Python 3 с его встроенной длинной арифметикой и возможностью считать C_{10000}^{5000} относительно безболезненно. Но нужно иметь в виду, что C_{100000}^{50000} он из коробки уже не потянет за разумное время...

¹⁰Нужно, конечно, также будет разобраться с тем, как эта аппроксимация получается. В библиографии обозначенной заметки есть отличная книга Феллера, рекомендуемая к просмотру.