Домашнее задание

11.1 Classical hyptest (two sample)

Сначала — две задачи на классические двухвыборочные тесты. Оформить, прислать. Лучше всего в LaTeX.

Задача 1. При испытании нового лекарства от шонибудилеза пациентов разбили на две группы по 50 человек. Одной группе давали новое лекарство, а другой — крашеный сахар (при этом все пациенты содержались вместе, остальные процедуры проводились одинаково и ни сами пациенты, ни сестры не знали, кому что дают). В результате в экспериментальной группе выздоровело 42 человека, а в контрольной 35. Проверьте с 5%-м уровнем значимости гипотезу о том, что новое лекарство эффективнее плацебо. Вычислите p-value.

Задача 2. В двух параллельных классах 25 и 28 учеников соответственно. На медосмотре всем измерили рост. получилось, что в первом классе средний рост составил 152 см со стандартным отклонением 4 см, а во втором 148 см со стандартным отклонением 5 см. Считая распределение роста в обоих классах нормальным, проверить гипотезу о совпадении роста с 5%-м уровнем значимости. Вычислить p-value.

11.2 MC tests

Напомню, что для построения статистического критерия необходимо, во-первых, знать точное распределение статистики критерия при выполнении нулевой гипотезы ("идеальное распределение"), а, во-вторых, знать *поведение* статистики в том случае, когда верна альтернатива.

Первое необходимо, чтобы построить корректный критерий (т.е. такой, ошибка первого рода для которого совпадает с заданной). Второе необходимо, чтобы правильно выбрать расположение критической области и обеспечить максимальную мощность против заданной альтернативы.

Ответ на второй вопрос, как правило, не представляет трудности. Например, для критериев типа расстояния (к которым относятся goodness-of-fit тесты типа хи-квадрат Пирсона, Колмогорова-Смиртнова, Крамера-фон Мизеса; двусторонние сравнения средних типа двустороннего t-test'а, двустронние z-test'ы сравнения пропорций и многие другие) ответ на второй вопрос выглядит достаточно просто — критическая область выбирается "на бесконечности", т.к. альтернативной гипотезе соответствует большое расстояние между выборкой и нулевой гипотезой. Для односторонних критериев критические области обычно тоже могут быть размечены "интуитивно".

Ответ же на первый вопрос нетривиален. В классических критериях мы используем некоторые частные предельные и точные теоремы о распределении статистик, общие утверждения отсутствуют.

Однако, существует универсальный способ обойти эту проблему и существенно ослабить условие. Вообще говоря, нам не нужно знать идеальное распределение статистики критерия, а достаточно только знать его квантили. Квантили же мы можем найти с любой точностью с помощью моделирования. Таким образом мы приходим снова к идее Монте-Карло моделирования (аналогично построению Монте-Карло доверительных интервалов).

Нам достаточно только потребовать, чтобы распределение статистики при верной нулевой гипотезе было фиксированным, т.е. не зависело от конкретного распределения

выборки. В случае простой нулевой гипотезы (когда нулевой гипотезе соответствует строго одно распределение выборки) это требование выполняется автоматически, в случае сложной гипотезы это требование является необходимым условием использования данной статистики в принципе, в самом деле, если оно не выполнено, то ошибка первого рода будет зависеть от конкретного распределения выборки, что недопустимо по определению корректного критерия.

В отличие от использования Монте-Карло для построения доверительных интервалов, данный способ является не нишевым, а, фактически, общепринятым. Например, в большинстве статистических пакетов в критерии Колмогорова-Смирнова для маленьких объемах выборки используются точные квантили, найденные с помощью моделирования, а распределение Колмогорова-Смирнова только для больших N.

Если необходимо просто проверить гипотезу с некоторым уровнем значимости, то достаточно просто найти соответствующий квантиль идеального распределения, вычислить статистику критерия с сравнить ее с квантилем. Если же нас интересует p-value, то необходимо *обратить* квантиль. На практике это делается очень просто — нужно промоделировать M статистик из распределения статистики при верной нулевой гипотезе, отсортировать их, а потом найти интервал обычным бинарным поиском (в `R` для этого используется функция `findInterval()`, в Python — `numpy.searchsorted` или `bisect`); затем оценить уровень квантиля как номер интервала, деленный на их количество. При достаточно большом M (порядка 1000) точность вполне достаточна для практических целей.

Применим этот подход на практике. Рассмотрим задачу проверки гипотезы согласия выборки с экспоненциальным распределением с неизвестным параметром. Такая гипотеза является сложной, поэтому необходимо придумать какую-то статистику согласия, идеальное распределение которой не будет зависеть он неизвестного параметра. К счастью, параметр масштаба для экспоненциального распределения является линейным, поэтому если его линейно оценить (например, через первый момент), нормировать выборку на полученную оценку масштаба и посчитать K-S-расстояние между нормированной выборкой и распределением $Exp(1)$, то распределение полученной величины (в случае верной нулевой гипотезы) хотя и не будет иметь распределение Колмогорова-Смирнова (даже асимтотически) но при этом не будет зависеть от неизвестного масштаба (это очевидно по линейности), и будет *асимптотически* не зависимо от объема выборки N (т.е. распределение этой величины имеет слабый предел и сходится к нему достаточно быстро, чтобы этим можно было воспользоваться на практике).

Вам предлагается собрать все вместе и построить критерий проверки согласия выборки с произвольным экспоненциальным распределением. Еще раз, коротко:

1. Фиксируем N объем выборки, M число серий моделирования

2. Моделируем M выборок длины N из экспоненциального распределения с произвольным фиксированным параметром, ОЦЕНИВАЕМ в каждой выборке параметр и нормируем на него (исходный параметр не важен, он сократиться при нормировке)

3. Для каждой выборки считаем K-S расстояние до $Exp(1)$, получаем выборку из статистик идеального распределения

4. Теперь мы можем проверять гипотезы. Проверка гипотезы: оценка параметра, нормирование, вычисление K-S статистики и вычисление p-value

5. Промоделировав M' выборок из экспоненциального распределения и проверив гипотезу M' раз, получим выборку из p-value. Нарисовав для нее выборочную функцию распределения (aka ECDF — Empirical Cumulative Distribution Function), проверим p-value на равномерность, а критерий — на корректность.

6. Затем берем вместо экспоненциального распределения Гамма и делаем все аналогично предыдущему пункту. Убеждаемся, что p-value стали распределены иначе (как?)