**Анализ эффективности удержания**

**1. Достигаемые уровни значимости**

* Подробный разбор техники построения таблицы сопряженности рассмотрен в видео “**Пример: поправки на множественную проверку при корреляционном анализе**” на 3 уроке 3 недели (вам нужна функция **pivot\_table**)
* Для того чтобы из таблицы сопряженности выделить какие-либо два штата, можете просто указать их имена в функции **pt.loc[["CT", "DC"], :]**
* Обратите внимание, что вам нужно взять всевозможные пары штатов. Формально говоря, это сочетания без повторений из n по 2. Количество этих сочетаний равно n!2!(n−2)! Для того чтобы в питоне получить сочетания из n по k можно воспользоваться функциейcombinations из модуля itertools, передав в функцию параметр k.

**3. Уровни значимости с включенной коррекцией**

* Обратите внимание, что на этот раз функцию **chi2\_contingency** необходимо вызвать с параметром correction=True

**4. Критерий Фишера**

* Как в первом и третьем заданиях, используйте всевозможные пары штатов, а также функцию **fisher\_exact** из модуля **scipy.stats**.

**5, 6, 7. Разные виды корреляции и выводы**

* Если не получается сдать задание, попробуйте пересмотреть видео “**Корреляция Пирсона**” и “**Корреляция Спирмена**” на 2 уроке 3 недели.

**8. Коэффициент V Крамера**

* Обратите внимание, что для подсчета коэффициента V Крамера, необходимо использовать формулу из видео “**Корреляция Мэтьюса и коэффициент Крамера**”. Для получения значения χ можете воспользоваться функцией **chi2\_contingency**

*Посчитайте значение коэффицента корреляции Крамера между двумя признаками: штатом (state) и оттоком пользователей (churn) для всех пользователей, которые находились в контрольной группе (treatment=1). Что можно сказать о достигаемом уровне значимости при проверке гипотезы о равенство нулю этого коэффициента?*

Правильным почему-то является пункт:

* Для вычисления коэффициента Крамера используется значение статистики xи-квадрат, на которую мы не можем положиться применительно к нашим данным.

У меня ответ по первому пункту прошёл, но вообще, меня немного смутило условие:

*Постройте таблицы сопряженности между каждой из всех 1275 возможных* ***неупорядоченных*** *пар штатов и значением признака churn.*

То есть, как эти пары нужно неупорядочивать, случайно смешивать?

1) Например, получаем список пар с помощью itertools.combinations.

2) Получаем случайные индексы.

3) Получаем таблицы сопраженности.

4) Смешиваем таблицы сопраженности случайным образом, говорится же, пары должны быть неупорядоченными.

5) Получаем каждый раз разные ответы на 1 вопрос.

Если взять стандартный код, генерируемый combinations и не смешивать ничего, то всё проходит нормально.

Нужна таблица сопряженности для признаков state и churn. Из нее надо последовательно брать подтаблицы сопряженности для неупорядоченных пар штатов. В каком порядке эти 1275 пар генерируются, не важно, результат детерминирован с точностью до перестановки пар.

Что такое таблица сопряжённости? Это когда по строкам у нас один признак (штаты в нашем случае), по столбцам другой признак (отток в нашем случае), и на пересечении строк-столбцов стоят числа встретить первый признак при наличии второго. Т.е. в нашем случае число churn=false и число сhurn=true для каждого штата в каждой паре.

Вопрос по 9 заданию. Для сравнения эффективности методов удержания 0 и 2 я использовал Z-критерий для двух долей независимых выборок. Ответы сошлись. Я применил поправку Холма на множественною проверку гипотез для 3-х гипотез: 0-1, 2-1 и 0-2. Если смысл применять поправку в таком случае, когда гипотез мало? Ответы значения pvalue до поправки и после отличаются, но статистическая значимость отличий до поправки и после не меняется.

Если требуется неслучайный результат проверки нескольких гипотез, то нужна поправка на множественную проверку - всегда. Иначе значимость результатов будет преувеличена.