1 Про корреляции

(10 баллов). Ответьте на следующие вопросы:

1) Что можно сказать про случайные величины X и Y, если $Corr(X, Y) = 1$?

2) Чем отличаются коэффициенты корреляции Пирсона и Спирмена? В каких случаях лучше пользоваться коэффициентом корреляции Спирмена?

3) Дана случайная величина X, равномерно распределенная на отрезке $[-1, 1]$ и величина $Y = |X|$. Чему равен их коэффициент корреляции (вычислите $Cov(X, Y)/\sqrt{Var(X)Var(Y)}$) ?

4) Дана случайная величина X, равномерно распределенная на отрезке $[-1, 1]$ и величина $Y = X^2$. Сгенерируйте выборку длины 1000 величины X и на основании нее получите соответствующую выборку для величины Y. По полученным выборка посчитайте коэффициенты корреляции Пирсона и Спирмена.

2 Про тесты с метриками-отношениями

(10 баллов). Ответьте на следующие вопросы:

1) В чем причина того, что биномиальный тест для доли часто плохо работает на практике и в чем это выражается (см. гифку про биномиальный тест)?

2) Верно ли, что Пуассоновский бутстреп для общегруппового CTR для всех рассмотренных нами случаев показывал лучшие результаты, чем тест Манна-Уитни на кликах (без бакетного преобразования)?

3) Для чего в задаче об оценке разницы в общегрупповом CTR был необходим дельта-метод (см.ноутбук и статью)?

4) Объясните, по как проводятся тесты над метриками с использованием бакетного преобразования? Какого рода метрики можно рассчитывать внутри бакета?

3 Более практические задачи

Упражнение 1 (20 баллов). В файле lifeline.xls содержатся 50 пар наблюдений из исследования докторов Л. Матера и М. Уилсона. В нем рассматривались следующие переменные: X — длина «линии жизни» на левой руке в сантиметрах (с точностью до 0.15 см) и Y — продолжительность жизни человека (округленная до ближайшего целого года). Изучите корреляцию X и Y . Верно ли, что X и Y связаны линейной регрессионной зависимостью?

Упражнение 2 (20 баллов). Сгенерируйте 100 выборок длины 1000 из распределения $N(0,1)+exponential(1)$. Для каждой из выборок $X_1, X_2....X_{100}$ постройте по 5 выборок Y вида:

$Y^{(1)} = X_i^2 + 0.1N(0,1)]$

$Y^{(2)} = \sqrt{|X_i|} + 0.1N(0,1)$

$Y^{(3)} = X_i * sin(X_i) + 0.1N(0,1)]$

$Y^{(4)} = X_i^3 + 0.1N(0,1)$

$Y^{(5)} = X_i^3 cos(X_i) + 0.1N(0,1)$

для каждой пары выборок X_i, соответствующая ей Y_i, посчитать коэффициенты корреляции Спирмена и Кендалла. (Всего должно получиться 500 значений для Спирмена и 500 значений для Кендалла). Нарисуйте диаграмму рассеяния (scatterplot), где по оси X будет корреляция Спирмена, а по оси Y - Кендалла. Можно ли сказать на основании собранных нами данных, что эти коэффициенты корреляции как-то связаны друг с другом?

4 Проект в закрытой формулировке

(80 баллов) (Обязательно попробовать) Представьте, что ваша компания занимается продажей товаров. Покупка устроена так: человек заходит на сайт (аналогично нашему показателю views), затем совершает покупку с некоторой вероятностью (аналогично нашему success-rate с бета-распределением) и далее совершает покупку (аналогично нашему clicks). Каждая покупка характеризуется каким-то "чеком"(стоимостью товара), его можно моделировать с помощью экспоненциального распределения вида $Ce^{-\lambda x}$ на интервале $[100, 2500]$ (подберите для этого распределения правильную нормировочную константу C и некоторую разумную λ, помните - это распределение чеков на покупку каких-то недорогих, частых товаров)

Допустим, что в обеих группах базовый success-rate = 0.03 , а в группе B его uplift равен 0.1. Кроме того, в группе B распределение чеков имеет меньшее значение λ: $\lambda_A = 1.2\lambda_B$.

Мы хотим измерить, значимо ли меняется значение 80 квантиля для чека пользователей в группе. Проверьте, какой из тестов для этого лучше подойдет: тест Манна-Уитни поверх бакетного преобразования или Пуассоновский бутстреп? Как их работоспособность зависит от скошенности распределений истинного CTR (бета-распределения) пользователей и показов (логнормального распределения). Для оценки можно использовать те же параметры, что и в лекции.

5 Проект в открытой формулировке

(Опционально)(100 баллов) Выберите некоторую метрику, которую вы могли бы анализировать в своей профессиональной деятельности в АБ-тестах. Подумайте, какие в ней можно ожидать изменения (сдвиг среднего, сдвиги

в каких-то квантилях) и попробуйте применить для измерения статистической значимости между группами А и Б ту же парадигму, что мы применяли для CTR'ов. Попробуйте Т-тест, Манна-Уитни, тесты с бакетными преобразованиями и бутстреп.