В городе работают 200 риелторов. У каждого в каталоге ровно две квартиры, и при встрече с клиентом он случайно выбирает, какую из них показать первой. 199 обычных риелторов имеют один объект в центре и один на окраине. 1 «элитный» риелтор имеет обе квартиры в центре. Клиент случайно выбирает риелтора, тот случайным образом показывает первое объявление, и оказывается, что это квартира в центре. Клиент просит показать второе (оставшееся) объявление того же риелтора. Какова вероятность того, что и второе объявление окажется квартирой в центре?  
  
Самый простой способ решить эту задачу - просто сказать, что если 2 квартиры в центре только у элитного, то вероятность равна 2 (столько квартир в центре у него) к 201 (столько всего квартир в центре). Итого 2/201. Потому что именно такая вероятность, что посмотрев квартиру в центре, мы оказались с элитным риелтором.

Но будет правильным обосновать это с точки зрения теоремы Байеса. Можно посмотреть на итоговый вопрос (что второе объявление окажется квартирой в центре) как на вопрос о вероятности, что наш риелтор "элитный".

Тогда изначально у нас была вероятность выбрать элитного P(A) = 1/200, а обычного P(B) = 199/200. Вероятность выбрать квартиру в центре будет P(X), она должна быть равна 201/400.

При первом просмотре, если риелтор элитный, то вероятность, что квартира будет в центре P(X|A) = 1. А если обычный, то P(X|B) = 1/2.

Теперь на основании новых данных, нам нужно понять, какова вероятность, что наш риелтор элитный и вторая квартира тоже будет в центре.

Применим формулу.

P(A|X) = [P(X|A) \* P(A)] / P(X)

Проверим P(X)

P(X) = P(X|A) \* P(A) + P(X|B) \* P(B)

P(X) = 1 \* 1/200 + 1/2 \* 199/200 = 1/200 + 199/400

P(X) = 201/400

Тогда:

P(A|X) = (1 \* (1/200)) / (201/400)

P(A|X) = (1 \* 400) / (200 \* 201)

P(A|X) 2/201.

Всё сошлось. Вероятность 2/201.

Риелтор Евгений устроил показ одной и той же квартиры трём потенциальным покупателям — Анне, Владимиру и Никите. Вероятности того, что каждый из них после просмотра захочет купить кваритиру независимы и равны P(Aнна)=0,2, P(Владимир)=0,4, P(Никита)=0,8 В случае положительного решения Евгений просит написать в телеграмм. Поздно вечером Евгений видит ровна два сообщений в телеграмме. Какова вероятность того, что одним из этих двоих оказался Никита?

Итак, у нас есть 3 возможных кейса с парами покупателей (АВ, ВН, АН соответственно первым буквам имён покупателей). Для того, чтобы рассчитать эту вероятность нам нужно посчитать отношение суммы вероятностей 2 кейсов, где Никита написал (ВН, АН) к сумме вероятностей 3 кейсов (все возможные пары).

Также при расчёте вероятности каждого кейса нужно учитывать не только вероятности того, что кто-то купит квартиру, но и вероятность того, что третий не купит. Это можно проиллюстрировать так: если бы Анна захотела купить квартиру с вероятностью 1, то верятность кейса ВН была бы равна не 0,4\*0.8, а 0,4\*0,8\*0, то есть 0. Теперь посчитаем вероятности кейсов.

АВ = 0,2\*0,4\*0,2 = 0.016

ВН = 0,8\*0,4\*0,8 = 0.256

АН = 0,2\*0,6\*0,8 = 0.096

Теперь посчитаем ВН+АН и поделим сумму на АВ+АН+ВН.

(0.256+0.096)/(0.256+0.096+0.016)

0.352/0.368.

В виде десятичной дроби это примерно 0.9565.

Есть три инструмента продвижения объявления по 100, 500 и 1000 рублей. Было сделано 1000 покупок этих инструментов. Вы собрали описательные статистики:

\* Средний чек 90,8

\* Дисперсия 340,8

\* Мода 100

\* Самая большая покупка 700

Все ли с ними хорошо, поясните почему

С данными всё плохо.

Я начну разбирать не по порядку.

3) Мода = 100. Звучит абсолютно нормально. Это просто значит, что самая частая покупка - один самый дешёвый инструмент.

4) Самая большая покупка = 700. Звучит тоже норм. Если мы предполагаем, что инструменты можно покупать сразу по несколько, то мы делаем вывод, что такой max соответствует покупке 100\*7 или 100\*2+500. Это значит, что никто не покупал самый дорогой инструмент и никто не делал крупных оптовых покупок.

1) Среднее = 90,8. Это невозможно, если мы не предполагаем, что у нас могут быть скидки. Если скидки бывают и сегодня всевозможный четверг, то, такое среднее возможно. Например, если у нас всего одна покупка на 700, большая часть на 100 и ещё много покупок на число около 80 или меньше.

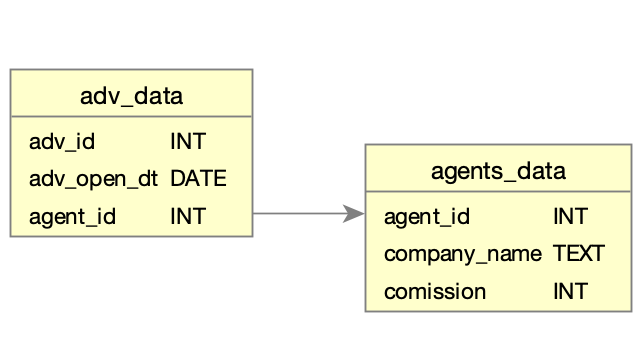
4) Если дисперсия числового ряда из 1000 элементов равна 340,8, то следовательно сумма квадратов отклонений должна быть равна 340800. (700-90,8)\*(700-90,8) = 371124.64. Это уже больше. Получается, что это невозможно.

Можно было бы проверить возможность такой дисперсии в отрыве от указанного среднего (например, приняв, что оно ошибочно и реальное среднее числового ряда будет другое). Но делать такие вычисления мне кажется излишним.

Перед тобой база реализованных объявлений о сдаче квартир в следующей структуре  
  
- adv\_id - уникальный идентификатор успешно закрытого объявления на сайте (сданных объектов)  
- adv\_open\_dt - дата закрытия объявления на сайте  
- agent\_id - уникальный идентификатор риелтора  
- company\_name - название организации, представляющей риелтора  
- comission - комиссия риелтора со сделки

Нужно вывести список риелторов c кол-вом успешных сдач выше, чем в среднем по рынку

Решение - код sql



select

ag.agent\_id

,count(case when adv.agent\_id > 0 then adv.adv\_id end) as cnt\_adv

from

agents\_data as ag

left join adv\_data as adv on ag.agent\_id = adv.agent\_id

group by ag.agent\_id

having cnt\_adv > (select

count(case when adv.agent\_id > 0 then adv.adv\_id end) / count(distinct(ag.agent\_id))

from agents\_data as ag

left join adv\_data as adv on ag.agent\_id = adv.agent\_id)

Я делал это в песочнице кликсхауса, вот ссылка.

https://play.clickhouse.com/play?user=play#

Мы решили запустить тест с 2 группами (контроль и тест). Ключевая метрика - CTR (клик/показ).

Рассчитай объем выборки для сплита 50/50. И рассчитай итоги теста по группам.

Решение - код на питоне

Поскольку это простая долевая метрика я могу использовать хи-квадрат Пирсона.

Чтобы рассчитать объём необходимой выбор нам нужно учесть альфу, мощность, mde и базовый CTR. Возьмём стандартные альфу 0,05 и мощность 0,8. Базовый ctr мы не знаем, его нужно запросить или посчитать самим из актуальных данных. Сейчас допустим будет 5%, а mde 2,5% (то есть 50% в относительном значении). Это много, но пока я кодил, я накидал данных в excel и мне будет проще так, а разницы для задачи нет.

Вот скриншот для понимания того, как выглядит таблица.

https://drive.google.com/file/d/13Qa4NdK8uslZoZib7vzW5EjfxRV-v8Tq/view?usp=sharing

Сначала сделаю последовательный расчёт тут, потом будет код для питона.

p1 = 0,05

p2 = 0,075

α = 0,05

β = 0,8

Значения для квантилей возьму из таблички.

n = (p₁(1 - p₁) + p₂(1 - p₂)) \* (Z₁-α/₂ + Z₁-β)² / (p₁ - p₂)²

n = (0,05(0,95)+0,075(0,925)) \* (1,96 + 0,84)² / (0,05 - 0,075)²

n = (0,0475 + 0,069375) \* (2,8)² / (−0,025)²

n = 0,116875 \* 7,84 / 0,000625

n = 1466,08

Таким образом выборка должна составить минимум 1467 для каждой группы

Теперь расчёт на питоне.

rate\_control = 0.05

mde = 1.5 # т.к. мы хотим увидеть увеличение от 50%, нам нужно умножить на 1.5

rate\_exp = rate\_control \* mde

n\_size = (rate\_control \* (1 - rate\_control) + rate\_exp \*(1 - rate\_exp )) \* (1.96 + 0.84)\*\*2 / (rate\_control - rate\_exp)\*\*2

print(n\_size)

Теперь расчёт итогов теста

import pandas as pd

stats = pd.read\_excel(r'C:\Users\booli\ctr\_data.xlsx')

ctrl\_shows = stats['ctrl\_shows'].sum()

ctrl\_clicks = stats['ctrl\_clicks'].sum()

test\_shows = stats['test\_shows'].sum()

test\_clicks = stats['test\_clicks'].sum()

chi\_stat = (test\_clicks - ctrl\_clicks)\*\*2 / ctrl\_clicks

print("Статистически значимо" if chi\_stat > 3.84 else "Не значимо")

# 3.84 - это значение из таблицы, соответствующее значимости 0.05 и степени свободы 1.

# это простой и рабочий код

# далее красивое дополнение, с которым мне помог дипсик (мне кажется, делать код лучше с помощью ии, это тоже полезный навык)

import numpy as np

from scipy import stats

from scipy.stats import chi2\_contingency, norm

chi2, p\_value, \_, \_ = chi2\_contingency([[ctrl\_clicks, ctrl\_shows - ctrl\_clicks],[test\_clicks, test\_shows - test\_clicks]])

print(f"CTR контроль: {ctrl\_clicks/ctrl\_shows:.3f}")

print(f"CTR тест: {test\_clicks/test\_shows:.3f}")

print(f"Хи-квадрат: {chi2:.3f}, p-значение: {p\_value:.4f}")

print("Статистически значимо" if p\_value < 0.05 else "Не значимо")

Твоя команда решила добавить персональные подборки на главной странице Циан. Вы договорились запустить тест с 2 группами (контроль и тест). Ключевая метрика - среднее количество звонков на пользователя. Рассчитай объем выборки для такого эксперимента. Опиши свой выбор. Решение - код на питоне

Для сравнения по такой метрике будем использовать t-тест. Он подходит для сравнения двух независимых средних и является типичным методом для данной задачи в силу своей универсальности и консервативности. Для t-теста нам нужно знать среднее и дисперсию (её нужно найти).

Считать мы будем по формуле:

n = 2 \* ((1,96 + 0,84)² \* σ²) / Δ²

σ² - дисперсия

Δ - mde

import numpy as np

import pandas as pd

stats = pd.read\_excel(r'C:\Users\booli\ctr\_data.xlsx')

# тут я не делал новый датасет, просто добавляю, чтобы код был полноценным

mde = 0.5 # предполагаю, что наши конверсия в звонок невысокая и даже такое увеличение будет значимым

var\_ctrl = np.var(stats['ctrl\_calls'])

n\_size = 2 \* ((1.96 + 0.84)\*\*2 \* var\_ctrl\*\*2) / mde\*\*2

print(n\_size)

Если нужны конкретные цифры, то я бы предложил считать дисперсию равной 11.25.

Тогда:

n = 2 \* ((1.96 + 0.84)\*\*2 \* 11.25\*\*2) / 0.5\*\*2

n = 7937.(9)

Итого 7938 пользователей на каждую группу.

Почему я взял дисперсию 11.25?

Если у нас будет пуассоновское распределение, то картина будет условно следующая:

85% пользователей у которых 0-3 звонка

14% пользователей у которых 4-12 звонков

1% пользователей у которых 13-50 звонков (я думаю, что есть и те у кого намного больше 100 звонков, но эти аномалии лучше просто не рассматривать).

Допустим мы считали по 10000 пользователей.

Тогда грубый подсчёт:

Вклад группы 1: 8500 \* 1.5 = 12750 звонков

Вклад группы 2: 1400 \* 7 = 9800 звонков

Вклад группы 3: 100 \* 30 = 3000 звонков

Общее количество звонков: 12750 + 9800 + 3000 = 25550

Среднее значение: 25550 / 10000 = ~2,56 звонка.

Отклонение для группы 1: 1,5 - 2,56 = -1,06

Отклонение для группы 2: 7 - 2,56 = 4,44

Отклонение для группы 3: 30 - 2,56 = 27,44

-1,06\*\*2 \* 8500 = 9550.6

4,44\*\*2 \* 1400 = 27599.04

27,44\*\*2 \* 100 = 75295.36

9550.6 + 27599.04 + 75295.36 = 112445

112445 / 10000 = ~11.25

Представь, что ты работаешь в команде умного помощника. Твоя команда решила добавился эмоджи в конце сообщения, чтобы сделать ответы живее. Вы запустили эксперимент, который нужно держать 14 дней, но на 5 день пришел продакт и сказал, что там все прокрасилось, поэтому нужно раскатить тестовую группу. Что будем делать?

Цепочка моих действий:

1. Сначала очень постараюсь объяснить продакту, в чём суть проблемы подглядывания. Может быть даже с данными из научных статей. Если он будет стоять на своём, то перейду к след. шагам.

2. Проверю, что с заградительными метриками. Мы могли допускать некоторое падение по ним, но нужно понять не является ли оно критичным.

3. Посмотрю, насколько равномерна тенденция по дням теста. Может быть у нас были очень хорошие результаты в первые пару дней, а сейчас падение. Или наоборот, вчера был аномальный всплеск и мы получили прокрас, но высок риск, что картина выровняется и перестанет быть значимой. Сравню с типичной картиной по неделям ранее. Например наличие или отсутствие выходных или других пиковых дней в тесте может иметь сильное влияние. (Говорю об этом опираясь на информацию из одного из проектов, которые делал в рамках учёбы в Яндекс Практикуме. Там исследовались данные по продажам квартир в Спб. Предполгаю, что у Циана вполне вероятно может быть похожая картина)

4. Перепроверю тест на предмет ошибок. Хочется удостовериться, что у нас хотя бы нет косяков. Например, у нас могли выгружаться неправильно даты, или выборки собрались неравномерно и в тестовой слишком мало людей. Или у пользователей из контрольной и тестовой группы и до теста было разное поведение и их некорректно сравнивать.

5. Это некотое развитие второго пункта. Если есть возможность изучить подробнее поведение пользователей из тестовой группы, то сделать это. Вдруг там какая-то ерунда, которую мы не предусмотрели. Например, пользователи ставят более высокую оценку, но задачи решаются реже.

Резюме.

Постараюсь объяснить, что его подход рискованный и ведёт к увеличению риска ошибочных выводов. Если не получается, но результаты по прошлым пунктам тоже плохие, то постараюсь донести это до продакта в структурированной форме с цифрами и картинками и выскажу, что, по моему мнению, тест нужно продолжать до конца (или хотя бы Х дней как компромисс в крайнем случае). Если вдруг в тесте ошибки, то его вообще возможно надо перезапускать.

Если продакт непреклонен, то предложу хотя бы продолжать тест до конца и не менять ничего для наших групп. Тогда мы хотя бы получим объективный результат и потом сможем это обсудить.

Ты вышел на работу в Циан в команду, которая занимается опытом пользователей, которые ищут жилье. Предложи метрики, на которые будешь смотреть в первую очередь

1) WAU/MAU - какая примерно ежедневная аудитория и сколько у нас активных пользователей в данный момент. Можно было бы DAU, но, насколько я знаю, для таких сервисов DAU очень сильно колеблется по дням недели.

(Снова говорю об этом опираясь на информацию из одного из проектов, которые делал в рамках учёбы в Яндекс Практикуме. Там исследовались данные по продажам квартир в Спб. Предполгаю, что у Циана вполне вероятно может быть похожая картина)

2) Конверсия в целевые действия: клик по карточке объявления, добавление в избранное сообщение/звонок по объявлению. Можно ещё оценить добавление в избранное.

3) Глубина просмотра (Общее число просмотров страниц / Общее количество визитов): Сколько карточек пользователь просматривает за один сеанс поиска. Помогает понять, хватает ли ему первых вариантов или он вынужден копать глубоко.

4) Отказы (Bounce Rate) и продолжительность сессии: процент сессий, где пользователь посмотрел всего 1 страницу и ушел. Высокий показатель — тревожный сигнал.