Статистика: АБ-тестирование

В этой лекции мы подробно рассмотрим, как с помощью статистических тестов можно строить серьёзную аналитику. Конкретнее, мы подробно обсудим, что такое АБ-тестирование и почему это нечто большее, чем просто проверка гипотез. Отдельное спасибо команде Филиппа Ульянкина за примеры, которые использовались также в других курсах по статистике.

11.1 АБ-тесты

Наверное, лучше всего идею АБ-тестирования представляет такой пример.

Пример 11.1. *В одной кофейне рядом с офисом Алексея не очень идут продажи пирожных. Чтобы исправить ситуацию, у них возникла мысль, что, возможно, следует добавить новую акцию: например, при покупке любого пирожного кофе в подарок в счастливые часы. В кофейне хотят понять, принесёт ли эта мера значимый успех, но проведение такого эксперимента связано с некоторыми дополнительными издержками, поэтому они хотели бы заранее всё спланировать с расчётом на стоящий результат.*

АБ-тестирование – это и есть методология проведения подобного рода экспериментов с использованием реальных данных и статистических процедур. Каждый АБ-тест имеет приблизительно следующую структуру:

1. Сбор данных (группа А получает обычный продукт без изменений, группа Б – с изменениями);
2. Подсчёт целевых метрик и проведение статистических тестов для ответа на исследовательские вопросы;
3. Принятие решения (надо или не надо продукт изменять).

На каждом этапе есть свои сложности и тонкости, которые очень важно учитывать. Когда мы планируем АБ-тест, мы начинаем с конца:

1. Какое решение нужно принять?
2. Какие целевые метрики отражают искомый эффект в данных? Каков размер эффекта, который мы хотим детектировать?
3. Какие статистические тесты мы будем использовать для проверки статистической значимости эффекта?

4. Сколько данных нужно собрать?

5. Дизайн сбора данных: как именно это делать, как должны быть устроены группы А и Б и так далее.

11.2 Метрики

В нашем примере решение, которое нужно принять, – это нужно ли оставлять счастливые часы. Дальше мы задаёмся вопросом, какая метрика может помочь понять, хорошая эта акция или нет. Например, это могут быть средние дневные продажи от выпечки и пирожных, продажи кофе, средний чек, среднее дневное количество посетителей. В онлайн-экспериментах метрики другие, но отражают похожие идеи, несколько популярных в индустрии примеров:

- CTR, clickthrough rate, считает отношение числа кликов на рекламный баннер и общего числа показов баннера;
- BR, bounce rate, количество пользователей, которые зашли на сайт и почти сразу вышли;
- CR, conversion rate, она же конверсия, количество пользователей которые пришли и сделали желаемое действие (например, приобрели рекламируемый товар);
- Retention rate, возвращаемость пользователей, сколько пользователей вернулось к вам через фиксированное количество дней;
- AOV, average order value, средний чек;
- LTV, lifetime value, время до окончательного ухода пользователя из вашего сервиса.

Метрика должна выбираться так, чтобы по результатам потом можно было чётко принять решение и чтобы её значением нельзя было манипулировать: то есть, стараться сделать так, чтобы рост (или падение) метрики – это точно для всех лучше. Странно получится, если изменение вроде полезно, но метрика то падает, то растёт в зависимости от прочих обстоятельств.

Стоит обращать внимание на долгосрочность эффектов, которые метрика пытается задетектировать. Метрика CTR измеряет краткосрочные эффекты, а метрика LTV (lifetime value, сколько денег клиент принёс компании от начала до его выхода из взаимодействия) ориентирована на долгосрочный эффект и замерить её за неделю невозможно. LTV может быть вообще очень сложно считать. Как долго вы уже пользуетесь стриминговыми сервисами?

11.3 Значимость, существенность и количество данных

В зависимости от того, как пройдёт эксперимент и как посчитаются критерии, разница в целевых метриках может быть *статистически значимой* или *незначимой*. В первом случае гипотеза отвергается на заданном уровне значимости в пользу альтернативы, во втором – наоборот; обычно гипотеза ставится со стороны отсутствия изменений. С другой стороны, разница в метриках может быть *существенной* или *несущественной*. Пирожных могли продать больше, но не настолько сильно больше, как надеялись – разница несущественна. Статистика показала, что действительно больше продажи – разница значима.

Оба свойства напрямую зависят и от того, сколько данных мы собрали. Жизнь в этом аспекте достаточно сурова.

Пример 11.2. *В 2020 году проводилось исследование эффекта витамина D в борьбе с депрессией. Эксперимент проводился 5 лет, включал в себя 18353 человека, в конце в группе А (принимали плацебо) было 609 человек с депрессией, а в группе Б (принимали витамин D) таких было 625. Но разница оказалась сильно незначимой ($p-value = 0.62$).*

Пример 11.3. *В 2018 году британские учёные (те самые) исследовали пользу раннего ужина и позднего завтрака для похудения. В исследовании принимали участие 13 человек, в группе Б 7 человек завтракали на 1.5 часа позже и ужинали на 1.5 часа раньше; исследование длилось 10 недель. Замеряли в начале и в конце эксперимента вес и объём участников, уровень глюкозы, инсулина и жиров. Уровень жира в группе Б упал на* 1.9%*, эффект оказался значимым с $p - value = 0.047$, но совершенно небольшим.*

Пример 11.4. *В 2020 году в Оксфорде исследовали полезность дексаметазона в борьбе с коронавирусом. В эксперименте участвовали 6.5 тысяч человек, часть была на аппарате ИВЛ, часть получала кислород через маску. Выяснилось, что смертность была на* 30% *ниже среди первых ($p-value = 0.0003$) и на* 20% *ниже среди вторых ($p-value = 0.0021$), если они получали терапию дексаметазона. Всего 2104 человека из 6.5 тысяч получали в течение 10 дней 6 милиграмм препарата в день. Разница оказалось статистически значимой и она была очень существенной, генеральный директор ВОЗ Тедрос Гебреисус назвал этот результат спасительным научным прорывом.*

Существенность и значимость задаются исследователем. Если продажи вырастут на 5%, но издержки на акцию в целом приведут к уменьшению прибыли, то такое увеличение будет считаться несущественным. Значимость характеризует степень уверенности: уровень значимости $\alpha = 0.3$ – достаточно грубо, а $\alpha = 0.001$ – очень надёжно.

Чтобы понять, как проверить существенность, прибегают к следующему приёму: проверяется простая гипотеза против простой альтернативы.

Пример 11.5. *Вспомним асимптотический z-тест для проверки гипотезы о равенстве долей (X_i и Y_j приходят из распределения $Ber(p_1), Ber(p_2)$). Проверяется гипотеза*

$$H_0 : p_1 = p_2, \quad H_A : p_2 - p_1 = \Delta.$$

На этапе планирования мы задаём размер эффекта Δ, уровень значимости α и вероятность ошибки второго рода β, тогда можно вычислить количество наблюдений. Предположим для простоты, что группы А и Б имеют размер $n/2$.

Ошибка второго рода – это когда мы приняли гипотезу при верной альтернативе. Принятие гипотезы – это событие

$$\frac{\overline{Y} - \overline{X}}{\sqrt{p_1(1-p_1)/n}} < z_{1-\alpha}.$$

Верна альтернатива, следовательно, мы знаем, что

$$\frac{\overline{Y} - \overline{X} - \Delta}{\sqrt{\frac{(p_1+\Delta)(1-p_1-\Delta)}{n/2} + \frac{p_1(1-p_1)}{n/2}}} \sim N(0,1),$$

обратите внимание на знаменатель! Поскольку распределение статистики известно, мы можем посчитать вероятность ошибки:

$$\mathbb{P}\left(\frac{\overline{Y} - \overline{X}}{\sqrt{p_1(1-p_1)/n}} < z_{1-\alpha}\right) = \Phi\left(z_{1-\alpha}\sqrt{\frac{p_1(1-p_1)}{2(p_1+\Delta)(1-p_1-\Delta)+2p_1(1-p_1)}} - \right.$$
$$\left. -\sqrt{n}\frac{\Delta}{\sqrt{2(p_1+\Delta)(1-p_1-\Delta)+2p_1(1-p_1)}}\right).$$

Мы требуем, чтобы эта вероятность, равнялась β, то есть, чтобы

$$z_{1-\beta} = z_{1-\alpha}\sqrt{\frac{p_1(1-p_1)}{2(p_1+\Delta)(1-p_1-\Delta)+2p_1(1-p_1)}} - \sqrt{n}\frac{\Delta}{\sqrt{2(p_1+\Delta)(1-p_1-\Delta)+2p_1(1-p_1)}}.$$

После некоторых косметических преобразований получаем с точностью до округления вверх, что

$$n = \left(\frac{z_{1-\alpha}\sqrt{p_1(1-p_1)} + z_{1-\beta}\sqrt{0.5(p_1+\Delta)(1-p_1-\Delta)+0.5p_1(1-p_1)}}{\Delta}\right)^2.$$

Есть разные функции для вычисления n, α, β при остальных двух известных. Например, функция `statsmodels.stats.power.zt_ind_solve_power` *позволяет вычислять мощность, нужное количество наблюдений и уровень значимости основываясь на двух известных и одном неизвестном параметре z-теста для равенства средних.*

Величина $\Delta = p_2 - p_1$ в примере и для других критериев определяемая похожим образом называется минимальным желаемым(обнаруживаемым) размером эффекта, или MDE

(Minimal Desirable(Discoverable) Effect). Интересно, что зная набор критериев, которые мы будем использовать, и зафиксировав вероятность ошибки первого рода α (он же уровень значимости), вероятность ошибки второго рода β (вспомним, что $1 - \beta$ – это мощность критерия) и размер эффекта Δ, который нам бы хотелось детектировать, мы можем точно сказать, сколько наблюдений нам нужно собрать. Для очень многих критериев есть специальная функция в statsmodels.

Обычно здесь помогает подобная таблица (фиксируется $\alpha = \beta$):

$\Delta \setminus \alpha$	0.01	0.02	0.03	0.04	0.05
0.15	31	24	21	18	16
0.10	78	61	51	44	39
0.05	300	234	196	170	150
0.01	7035	5483	4599	3985	3517

В зависимости от возможностей, ресурсов и желаний заказчика, можно точно ответить какой размер выборки и при каких предположениях нужно собрать, чтобы тест имел заданную мощность и контролируемую вероятность ошибки. Из формулы выше видно, что если требуется задетектировать очень маленький эффект, то n будет расти очень сильно. С другой стороны, при требовании большей точности (уменьшении вероятностей ошибок α, β) n тоже растёт. В силу того, что выборка никогда не даёт всей информации о жизни, статистическими средствами нельзя провести тест с вероятностями ошибки 0.

11.4 Важные тонкости и провалы

Вообще, всё, что угодно, может пойти не так. Согласно разным подсчётам, от 10% до трети АБ-тестов успешны, то есть, действительно помогают содержательно ответить на вопрос. Повысить успешность помогает опыт и внимательность к деталям.

Пример 11.6. *(Условия, приближенные к реальности) Фольклорная история из 2000х. Как повлияет на продажы колы увеличение содержания сахара? Фокус-группа пробует разные напитки, анализ данных показывает, что людям в целом нравится больше кола с большим содержанием сахара. Компания пробует запустить тестовую линейку с увеличенным содержанием сахара и продажи падают. АБ-тест спланирован правильно, но этап сбора данных не был продуман достаточно реалистично.*

1. *В фокус-группах наливают небольшими стаканами, тогда как в реальности колу пьют промышленно (литрами).*

2. *Фокус-группу собирали в комфортном прохладном помещении, в таких условиях сахар в небольших количествах действительно больше нравится людям; в жару всё это ощущается по-другому.*

Пример 11.7. *(Репрезентативность выборки) Пример из Фрикономики [7], который в разных формах часто можно наблюдать и сейчас. На выборах 1936г. в США журнал The Literary Digest опросил 10млн. человек, было предсказано, что с перевесом 60 на 40 победит республиканец Альф Лэндон. Как мы помним, победил демократ Рузвельт с перевесом примерно 60 на 40. Оказалось, аудитория у журнала было сильно смещена в сторону республиканцев: его читали более состоятельные люди, которые придерживались республиканской идеологии. Дополнительно журнал обзванивал читателей по телефону, но это не помогло – телефон в то время тоже был достаточно дорогой вещью.*

Пример 11.8. *(Проблема самоотбора) Наверняка вам попадались объявления в соцсетях от студентов с других специальностей с просьбой пройти опрос для дипломной или курсовой работы. Может получится так, что те люди, которым опрос адресован, просто не захотят тратить время или же просто не найдут этот опрос. То есть, изначально есть механизм, который позволяет отказаться от участия. Из-за этого выборка очень сильно смещена и результаты рискуют не быть достаточно общими.*

Пример 11.9. *(Связанные выборки) В случае с напитками или едой, восприятие может меняться в зависимости от порядка. Дать кофе для сравнения первый, потом второй, или наоборот? В примере с самой первой лекции, например, это проблема обходилась тем, что приём кофе проходил в разные дни. Но если много людей участвуют, то можно случайно каким-то людям давать напитки в одном порядке, а другим – в другом.*

Пример 11.10. *(Скрытые особенности) Исследуется удобрение и измеряется урожайность пшеницы на поле с удобрением и без него. Без удобрения результат оказался лучше. Есть другие факторы, помимо удобрения, которые надо стараться исключить:*

- *Возможно,часть участок поля без удобрения получает больше света, чем часть с удобрением;*
- *Возможно, там рядом геотеориальный источник или труба с горячей водой и почва теплее;*
- *Возможно, там сама почва лучше;*
- *...?*

Пример 11.11. *(Нарушение схемы АБ-тестирования) Нельзя по ходу изменять метрики под результаты – сами метрики становятся зависимыми от данных и распределение статистик в тестах уже не будет таким, как планировалось изначально, и ничего не гарантирует корректную работу критериев. Нельзя досрочно прерывать АБ-тест, если увидели нужный эффект (это приводит к завышению уровня значимости). Тем не менее, в последнем случае есть специальные техники, которые позволяют частично эту проблему смягчить.*

Пример 11.12. *(Проблемы метрик) Метрики подвержены трендам и сезонностям во время теста: новизна изменений привлекает пользователей, но они со временем привыкают и метрика падает. Метрика также может не отражать интересующего нас вопроса.*

11.5 Когда с АБ-тестами сложно

В офлайне тоже можно проводить АБ-тестирование, более того, оказывается, что вы наверняка много раз его видели.

Пример 11.13. *В магазине предлагается новейшая ограниченная серия круассанов с начинкой, как у известных конфет с кокосовым кремом. После окончания акции будут думать, делать ли её постоянной или оставить сезонной на основе некоторого времени продаж.*

Пример 11.14. *В магазине меняют расположение отделов в надежде лучше использовать маршрут покупателя, чтобы тот больше купил.*

Однако в офлайн-тестах есть многие естественные ограничения, которые не позволяют их проводить или создают большие сложности.

1. Если тест проводится долго, то предпочтения людей могут меняться от разных факторов: сезонности, занятости, командировок и отпусков....
2. Некоторые тесты нельзя провести по этическим соображениям: чтобы проверить, вредит ли курение здоровью, нельзя просто набрать 5000 похожих людей и половину из них заставить курить ещё 10 лет.
3. Едва ли большие решения государственного масштаба могут позволить себе АБ-тестирование. К примеру, нельзя просто попилить экономику на две части и в каждой взимать разные налоги в течение долгого времени. В более маленьких масштабах подобные эксперименты (например, ПДД, образовательные реформы, льготы для бизнеса) проводят в отдельных регионах в виде эксперимента.

Как раз из-за того, что в каких-то вопросах АБ-тестирование невозможно, люди фокусируют свои усилия на объяснимых и интерпретируемых предсказательных моделях, таких как, к примеру, линейная регрессия и решающие деревья, которые могли бы предсказать с некоторыми гарантиями, что будет, без АБ-тестирования.

Литература

[1] T. W. Anderson and D. A. Darling. Asymptotic Theory of Certain "Goodness of Fit"Criteria Based on Stochastic Processes. *The Annals of Mathematical Statistics*, 23(2):193 – 212, 1952.

[2] Peter C Austin, Muhammad M Mamdani, David N Juurlink, and Janet E Hux. Testing multiple statistical hypotheses resulted in spurious associations: a study of astrological signs and health. *J Clin Epidemiol*, 59(9):964–969, July 2006.

[3] Erika Cule, Paolo Vineis, and Maria De Iorio. Significance testing in ridge regression for genetic data. *BMC Bioinformatics*, 12(1):372, Sep 2011.

[4] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.

[5] B. Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26, 1979.

[6] R. E. Kalman. A New Approach to Linear Filtering and Prediction Problems. *Journal of Basic Engineering*, 82(1):35–45, 03 1960.

[7] S.J. Levitt, S.D. Dubner. *Freakonomics*. NY: Harper Trophy, 2006.

[8] Richard Lockhart, Jonathan Taylor, Ryan J. Tibshirani, and Robert Tibshirani. A significance test for the lasso. *The Annals of Statistics*, 42(2):413–468, 2014.

[9] J. Scott Long and Laurie H. Ervin. Using heteroscedasticity consistent standard errors in the linear regression model. *The American Statistician*, 54(3):217–224, 2000.

[10] Halbert White. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 48(4):817–838, 1980.

[11] Платонов Е.Н. Горяинова Е.Р., Панков А.Р. *Прикладные методы анализа статистических данных*. Изд. дом Высшей школы экономики, 2012.

[12] Ю.М. Кельберт, М.Я. Сухов. *Вероятность и статистика в примерах и задачах. Т.3: теория информации и кодирования*. М.: МЦНМО, 2013.

[13] А.Н. Ширяев. *Основы стохастической финансовой математики*. МЦНМО, 2016.