Машинное обучение

Лекция 6. Проверка статистических гипотез и A/B тестирование.

ADSC
Applied Data Science Center

1 февраля 2018

Краткое содержание

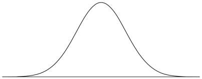
Проверка гипотез

выборка: $X^n = (X_1, \dots, X_n)$, $X \sim P \in \Omega$

нулевая гипотеза H_0: $P \in \omega$, $\omega \subset \Omega$

альтернатива H_1: $P \notin \omega$

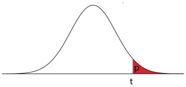
статистика: $T(X^n)$, $T(X^n) \sim F(x)$ при $P \in \omega$, $T(X^n) \nsim F(x)$ при $P \notin \omega$



реализация выборки: $x^n = (x_1, \dots, x_n)$

реализация статистики: $t = T(x^n)$

достигаемый уровень значимости: $p(x^n)$ — вероятность при H_0 получить $T(X^n) = t$ или ещё более экстремальное

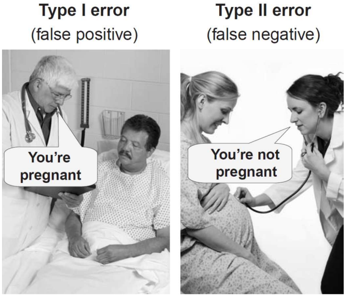


$$p(x^n) = P(T \geqslant t | H_0)$$

Гипотеза отвергается при $p(x^n) \leqslant \alpha$, α — уровень значимости

Ошибки первого и второго рода

	H_0 верна	H_0 неверна
H_0 принимается	H_0 верно принята	Ошибка второго рода (False negative)
H_0 отвергается	Ошибка первого рода (False positive)	H_0 верно отвергнута



Ошибки первого и второго рода

Задача проверки гипотез несмметрична относительно пары $(\mathrm{H}_0, \mathrm{H}_1)$: вероятность ошибки первого рода ограничивается малой величиной α, второго рода — минимизируется путём выбора критерия.

Корректный критерий: $\mathrm{P}(\mathrm{p}(\mathrm{T}) \leqslant \alpha | \mathrm{H}_0) \leqslant \alpha$
Мощность: $\mathrm{pow} = \mathrm{P}(\mathrm{p}(\mathrm{T}) \leqslant \alpha | \mathrm{H}_1) \to \mathsf{max}$

Интерпретация результата

Если величина p достаточно мала, то данные свидетельствуют против нулевой гиптезы в пользу альтернативы.
Если величина p недостаточно мала, то данные не свидетельствуют против нулевой гипотезы в пользу альтернативы.

При помощи инструмента проверки гипотез нельзя доказать справедливость нулевой гипотезы. В философии подобные рассуждения встречаются в критерии научности Поппера.

Отсутствие доказательств $\neq$ доказательство отсутствия.

Статистическая и практическая значимости

Вероятность отвергнуть нулевую гипотезу зависит не только от того, насколько она отличается от истины, но и от размера выборки.
По мере увеличения n нулевая гипотеза может сначала приниматься, но потом выявятся более тонкие несоответствия выборки гипотезе H_0, и она будет отвергнута.

При любой проверке гипотез нужно оценивать размер эффекта — степень отличия нулевой гипотезы отличается от истины, и оценивать его практическую значимость.

Статистическая и практическая значимости

- (Lee et al, 2010): за три года женщины, упражнявшиеся не меньше часа в день, набрали значимо меньше веса, чем женщины, упражнявшиеся меньше 20 минут в день ($p < 0.001$). Разница в набранном весе составила 150 г. Практическая значимость такого эффекта сомнительна.
- (Ellis, 2010, гл. 2): в 2002 году клинические испытания гормонального препарата Премарин, облегчающего симптомы менопаузы, были досрочно прерваны. Было обнаружено, что его приём ведёт к значимому увеличению риска развития рака груди на 0.08%, риска инсульта на 0.08% и инфаркта на 0.07%. Формально эффект крайне мал, но с учётом численности населения он превращается в тысячи дополнительных смертей.
- (Kirk, 1996): если при испытании гипотетического лекарства, позволяющего замедлить прогресс ослабления интеллекта больных Альцгеймером, оказывается, что разница в IQ контрольной и тестовой групп составляет 13 пунктов, возможно, изучение лекарства стоит продолжить, даже если эта разница статистически незначима.

Уровень значимости

Чем ниже достигаемый уровень значимости, тем сильнее данные свидетельствуют против нулевой гипотезы в пользу альтернативы.

Достигаемый уровень значимости нельзя интерпретировать как вероятность справедливости нулевой гипотезы!

$$\mathrm{p} = \mathrm{P}(\mathrm{T} \geqslant \mathrm{t} | \mathrm{H}_0) \neq \mathrm{P}(\mathrm{H}_0 | \mathrm{T} \geqslant \mathrm{t})$$

Отсутствие свидетельств = свидетельство отсутствия.

Мощность

$\mathrm{pow} = \mathrm{P}(\mathrm{p}(\mathrm{T}) \leqslant \alpha | \mathrm{H}_1)$ — вероятность отвергнуть H_0, если верна альтернатива

Мощность критерия зависит от следующих факторов:

- размер выборки;
- размер отклонения от нулевой гипотезы;
- чувствительность статистики критерия;
- тип альтернативы.

Размер выборки

На выборке из 10 бросков монетки вы не отличите честную от смещённой с вероятностью орла 0.51.

Обеспечение требуемой мощности: размеры выборки подбирается так, чтобы при размере отклонения от нулевой гипотезы не меньше заданного (например, вероятность орла не меньше 0.51) мощность была не меньше заданной.

Руководствуясь этим правилом, оценивается время АБ тестирования. Например, вы хотите показать увеличение конверсии с 0.51 до 0.53, значит, нужно собрать столько событий, чтобы при конверсии не менее 0.53 гипотеза о её равенстве 0.51 отвергалась с вероятностью более 85%.

Параметрические критерии

Параметрические критерии проверки гипотез допускают дополнительное знание о распределении выборок, что позволяет составлять более мощные критерии.

К сожалению, реальные данные очень редко распределены как табличные распределения. Но есть ряд популярных случаев, когда это так, их и разберём.

Z-критерий меток для доли

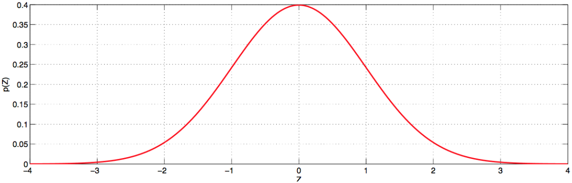
выборка: $X^n = (X_1, \ldots, X_n), X \sim Ber(p)$

нулевая гипотеза: $H_0\colon p = p_0$

альтернатива: $H_1\colon p <\neq> p_0$

статистика: $Z_S(X^n) = \frac{\hat{p}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$

нулевое распределение: $N(0,1)$



Биномиальный критерий

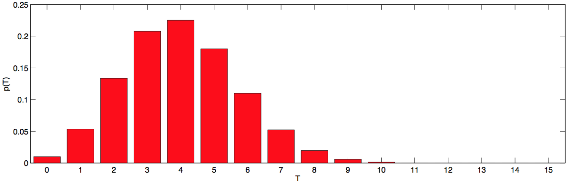
выборка: $X^n = (X_1, \dots, X_n), X \sim Ber(p)$

нулевая гипотеза: $H_0\colon p = p_0$

альтернатива: $H_1\colon p <\neq> p_0$

статистика: $T(X^n) = \sum\limits_{i=1}^{n} X_i$

нулевое распределение: $Bin(n, p_0)$



Z-критерий разности долей, независимые выборки

выборки: $X_1^{n_1} = (X_{11}, \ldots, X_{1n_1}), X_1 \sim Ber(p_1)$
$X_2^{n_2} = (X_{21}, \ldots, X_{2n_2}), X_2 \sim Ber(p_2)$
выборки независимы

Исход \ Выборка	$X_1^{n_1}$	$X_2^{n_2}$
1	a	b
0	c	d
$\sum$	n_1	n_2

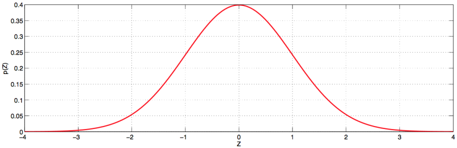
нулевая гипотеза: $H_0: p_1 = p_2$

альтернатива: $H_1: p_1 <\neq> p_2$

статистика: $Z(X_1^{n_1}, X_2^{n_2}) = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{P(1-P)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$

$P = \frac{\hat{p}_1 n_1 + \hat{p}_2 n_2}{n_1 + n_2}, \ \hat{p}_1 = \frac{a}{n_1}, \ \hat{p}_2 = \frac{b}{n_2}$

нулевое распределение: $N(0,1)$



Z-критерий разности долей, связанные выборки

выборки: $X_1^n = (X_{11}, \dots, X_{1n}), X_1 \sim Ber(p_1)$
$X_2^n = (X_{21}, \dots, X_{2n}), X_2 \sim Ber(p_2)$
выборки связанные

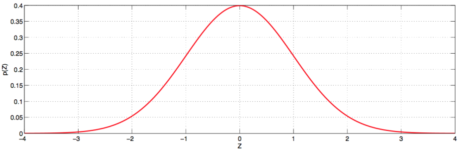
X_1^n \ X_2^n	1	0
1	e	f
0	g	h

нулевая гипотеза: $H_0\colon p_1 = p_2$

альтернатива: $H_1\colon p_1 <\neq> p_2$

статистика: $Z(X_1^n, X_2^n) = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{f+g}{n^2} - \frac{(f-g)^2}{n^3}}} = \frac{f-g}{\sqrt{f+g-\frac{(f-g)^2}{n}}}$

нулевое распределение: $N(0,1)$



Z-критерий

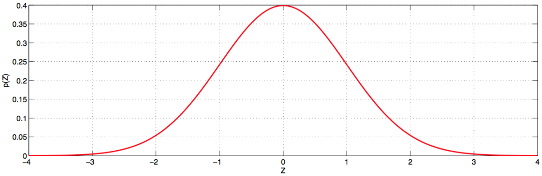
выборки: $X_1^{n_1} = (X_{11}, \dots, X_{1n_1}), X_1 \sim N(\mu_1, \sigma_1^2)$
$X_2^{n_2} = (X_{21}, \dots, X_{2n_2}), X_2 \sim N(\mu_2, \sigma_2^2)$
σ_1, σ_2 известны

нулевая гипотеза: $H_0\colon \mu_1 = \mu_2$

альтернатива: $H_1\colon \mu_1 <\neq> \mu_2$

статистика: $Z(X_1^{n_1}, X_2^{n_2}) = \dfrac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$

нулевое распределение: $N(0,1)$



t-критерий Стьюдента

выборки: $X_1^{n_1} = (X_{11}, \ldots, X_{1n_1}), X_1 \sim N(\mu_1, \sigma_1^2)$
$X_2^{n_2} = (X_{21}, \ldots, X_{2n_2}), X_2 \sim N(\mu_2, \sigma_2^2)$
σ_1, σ_2 неизвестны

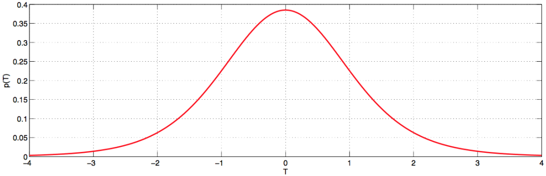
нулевая гипотеза: $H_0\colon \mu_1 = \mu_2$

альтернатива: $H_1\colon \mu_1 <\neq> \mu_2$

статистика: $T(X_1^{n_1}, X_2^{n_2}) = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$

$$\nu = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{S_1^4}{n_1^2(n_1-1)} + \frac{S_2^4}{n_2^2(n_2-1)}}$$

нулевое распределение: $\approx St(\nu)$



Приближение достаточно точно при $n_1 = n_2$ или $[n_1 > n_2] = [\sigma_1 > \sigma_2]$.

Достойны упоминания

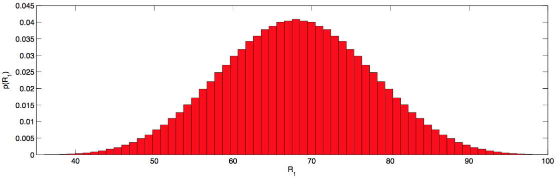
1. Доверительные интервалы Вальда, Уилсона — доверительные интервалы для Z-тестов
2. Критерий Харке-Бера, критерий согласия Пирсона — проверка данных на нормальность

Непараметрические критерии

Существует специальный набор критериев, которые можно применять, не зная точного распределения выборки.

Критерий Мана-Уитни

выборки:	$X_1^{n_1} = (X_{11}, \ldots, X_{1n_1})$
	$X_2^{n_2} = (X_{21}, \ldots, X_{2n_2})$
	выборки независимые
нулевая гипотеза:	$H_0\colon F_{X_1}(x) = F_{X_2}(x)$
альтернатива:	$H_1\colon F_{X_1}(x) = F_{X_2}(x+\Delta), \Delta <\neq> 0$
статистика:	$X_{(1)} \leq \ldots \leq X_{(n_1+n_2)}$ — вариационный ряд объединённой выборки $X = X_1^{n_1} \bigcup X_2^{n_2}$
	$R_1(X_1^{n_1}, X_2^{n_2}) = \sum\limits_{i=1}^{n_1} \operatorname{rank}(X_{1i})$
нулевое распределение:	табличное



Критерии согласия

выборки: $X_1^{n_1} = (X_{11}, \dots, X_{1n_1})$

$X_2^{n_2} = (X_{21}, \dots, X_{2n_2})$

выборки независимые

нулевая гипотеза: $H_0\colon F_{X_1}(x) = F_{X_2}(x)$

альтернатива: $H_1\colon H_0$ неверна

Критерий Смирнова

статистика: $D\left(X_1^{n_1}, X_2^{n_2}\right) = \sup\limits_{-\infty<x<\infty} \left|F_{n_1X_1}(x) - F_{n_2X_2}(x)\right|$

Критерий Андерсона (модификация критерия Смирнова-Крамера-фон Мизеса)

статистика:
$$T\left(X_1^{n_1}, X_2^{n_2}\right) = \frac{1}{n_1n_2(n_1+n_2)}\left(n_1\sum_{i=1}^{n_1}\left(\operatorname{rank}\left(X_{1i}\right) - i\right)^2 + \right.$$
$$\left. + n_2\sum_{j=1}^{n_1}\left(\operatorname{rank}\left(X_{2j}\right) - j\right)^2\right) - \frac{4n_1n_2-1}{6(n_1+n_2)}$$

Статистики имеют табличные распределения при H_0.

Разбиение на тестовые группы

Требуется разбить множество объектов на две тестовые группы. Хочется, чтобы это было репрезентативное разбиение. Что нужно проверить:

- Статические фичи (пол, возраст и т.п.) распределены одинаково — Критерии согласия и др
- Исторические фичи (конверсии за какой-то период, покупки и и.д.). Распределения врядли будут прям совпадать, но стоит проверить разные статистики (среднее, медианы, дисперсии) — Непараметрические критерии
- АА-тест. Разбиение всё равно может оказаться плохим, поэтому стоит провести АА тест, в рамках которого убедиться, что в группах нет значимых различий между целевыми метриками.

Измерение эффекта

Вы делаете рекламу аксессуаров к заказу в интернет магазине. Вам требуется проверить наличие и оценить экономический эффект от использования вашей модели.

Прежде всего нужно ответить на следующие вопросы:

1. Что является целевой метрикой?
2. На какое увеличение мы рассчитываем?
3. Как проверять статистическую значимость результата?
4. Как оценить эффект?
5. Как долго должен идти AB тест?

Подводные камни

В статистике очень легко самообмануться. Поэтому надо всегда понимать формально какую гипотезу мы проверяем и какими предположениями пользуемся. Сейчас мы приведём несколько неочевидных примеров некорректного применение аппарата мат. статистики.

Последовательный анализ

Вы распланировали АБ тест. По вашим оценкам(вы использовали биномиальный тест) за 81 день отклонение изменяемой величины на 1 процент является значимым. Вы ежедневно мониторили результаты теста и через 9 дней обнаружили отклонение в 5 процентов, что является значимым для теста длиной 9 дней. Можно в этом случае досрочно завершить АБ тест?

Последовательный анализ

Из закона повторного логарифма следует, что отклонения по ходу теста могут быть сколь угодно большими (особенно, если тест долго длится). Так как состояние мониторилось каждый день, то мы просто специально подобрали момент, когда отклонения было большим, поэтому тест нельзя останавливать. Гипотеза стала зависима от данных.

Как с этим жить: Нужно применять Статистический последовательный анализ (гуглите Sequential analysis). Он даёт во-первых корректные, а во-вторых более мощные критерии, пользуясь дополнительным знанием о потоковости данных.

Множественная проверка гипотез

Допустим, что вы проверяете средний чек, среднее число товаров в чеке, среднее число аксессуаров в чеке. Для каждой из этих величин вы составили свой критерий для проверки гипотезы о наличие эффекта. Каков уровень значимости для такой одновременной проверки гипотез?

Множественная проверка гипотез

Поскольку величина чека, число товаров в нём и число аксессуаров в нём — зависимые величины, то нельзя в точности найти уровень значимости, но можно его оценить:

$$\alpha \leqslant \mathrm{P}(\mathrm{p}_1 \leqslant \alpha \text{ or } \mathrm{p}_2 \leqslant \alpha \text{ or } \mathrm{p}_3 \leqslant \alpha | \mathrm{H}_0) \leqslant \sum_{\mathrm{i}} \mathrm{P}(\mathrm{p}_{\mathrm{i}} \leqslant \alpha | \mathrm{H}_0) = 3\alpha$$

Причём скорее всего самое первое неравенство строгое. Получается, что из-за того, что мы проверяем несколько гипотез, вероятность ошибки первого рода повышается. Она будет вызвана не особенностью данных, а тем, что мы несколько раз её проверяем.

Множественная проверка гипотез

Ошибка первого рода вызвана не особенностью данных, а тем, что мы несколько раз её проверяем.

Как с этим жить: Нужно применять методы Множественной проверки гипотез (гуглите Multiple comparisons problem). Самый простой способ — уменьшить α в число гипотез раз, но есть и более сложные подходы. Есть хорошая реализация в Python — statsmodels.sandbox.stats.multicomp.multipletests.

Итоги

1. Существует концепция проверски статистических гипотез.
2. Для проверки гипотез приеняются различные статистические критерии.
3. Бывают параметрические и непараметрические критерии.
4. На практике машинного обучения это всё применяют для AB и AA тестирования.
5. Существуют разные подводные камни.

Полезные ссылки (они кликабельны)

1. Лекции ВШЭ по прикладной статистике. Презентация основана на этих материалах. Здесь есть технические подробности всего, о чём рассказывалось.
2. Лагутин наглядная математическая статистика. Если заинтересуетесь статистикой.