Проверка статистических гипотез и A/B тестирование

7 января 2019 г.

1 Механизм проверки статистических гипотез

1.1 Формальное определение

Для того, чтобы правильно формально записывать требуемые выражения, нам нужно ввести дополнительные обозначения. Итак, нам дана выборка: $X^n = (X_1, \ldots, X_n)$, про эту выборку известно, что $X \sim \mathbf{P}$, где $\mathbf{P}$ — это какое-то распределения из множества всех рассматриваемых распределений Ω. Мы будем проверять верность нулевой гипотезы, которая обозначается H_0 и обычно состоит в том, что $\mathbf{P} \in \omega,\ \omega \subset \Omega$, то есть что распределение обладает каким-то свойством. Для проверки гипотез также нужна альтернатива — H_1 , состоящая в том, что распределение выборки обладает каким-то другим свойством: $\mathbf{P} \in \gamma,\ \gamma \cap \omega = \emptyset$. Для того, чтобы проверить гипотезу, нам необходимо подсчитать некоторую статистику $T(X^n)$, про которую известно, что $T(X^n) \sim F(x)$ при $\mathbf{P} \in \omega,\ T(X^n) \nsim F(x)$ при $\mathbf{P} \in \gamma$. Совокупность статистики и распределения, которое статистика будет иметь при выполнении нулевой гипотезы, называется статистическим критерием.

На практике мы наблюдаем не теоретическую выборку, а конкретные значения x_i случайных величин X_i. Набор этих наблюдаемых значений называется реализацией выборки $x^n = (x_1, \ldots, x_n)$. Для полученной реализации выборки мы можем подсчитать значение статистики и, таким образом, получить реализацию статистики $t = T(x^n)$. Если бы была верна H_0, то распределение $T(X^n)$ было бы известно, поэтому мы можем оценить насколько невероятное значение реализации статистики было получено. Формально это выражение можно записать как $p(x^n) = P(T(X^n) \geqslant T(x^n) \mid H_0)$. Гипотеза отвергается, если $p(x^n) \leqslant \alpha$, где α — какой-то порог.

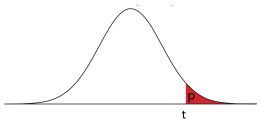


Рис. 1: Известно распределение статистики при выполнении H_0, красным отмечена зона, в которой нулевая гипотеза будет отвергнута.

1.2 Ошибки первого и второго рода

Возможны четыре варианта соотношения реальности и результата проверки гипотезы:

	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

Задача проверки гипотез не симметрична относительно пары (H_0, H_1). Во-первых, мы отвергаем или не отвергаем H_0 в пользу H_1, но не наоборот. Во-вторых, ошибки разного рода могут иметь разные последствия. Например, диагностировать смертельную болезнь у здорового человека не так плохо, как не диагностировать её у здорового. Поэтому при оценке качества статистического критерия вводится два правила:

- Вероятность ошибки первого рода ограничивается малой величиной α, которую называют уровнем значимости
- Вероятность второго рода минимизируется

На основании этих двух правил вводится два свойства критериев:

- Корректность критерия: $P(p(X^n) \leqslant \alpha \mid H_0) \leqslant \alpha$. Вероятность отвергнуть H_0, если она верна должна быть меньше уровня значимости
- Мощность критерия: $pow = P(p(X^n) \leqslant \alpha \mid H_1) \to \max$. Максимизируем вероятность отвергнуть H_0, если верна H_1

Чем ниже достигаемый уровень значимости, тем сильнее данные свидетельствуют против нулевой гипотезы в пользу альтернативы. Однако, достигаемый уровень значимости нельзя интерпретировать как вероятность справедливости нулевой гипотезы:

$$P(p(X^n) \leqslant \alpha \mid H_0) \neq P(H_0 \mid p(X^n) \leqslant \alpha)$$

Мощность критерия зависит от следующих факторов:

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

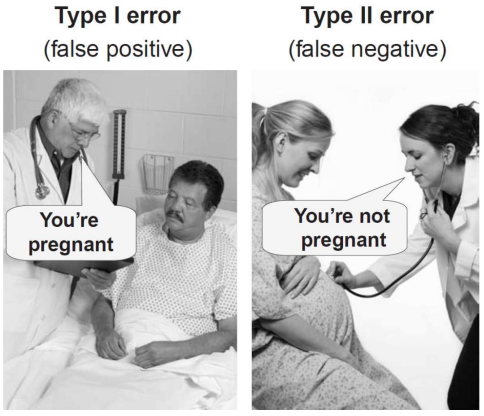


Рис. 2: Пример ошибок разного рода в реальной жизни. H_0 состоит в том, что человек не беременный.

1.3 Интерпретация результата

Значение вероятности того, что статистика примет те значения, что приняла, или ещё более невероятные (мы обозначали её $p(x^n)$) называют p-value. Заметим, что чтобы отвергнуть или не отвергнуть гипотезу нам достаточно знать только p-value, поскольку мы сравниваем это значения с заранее заданным порогом. По этой причине большинство библиотек для языков программирования для проверки статистических гипотез выдают именно эту величину. Поэтому необходимо правильно интерпретировать значения p-value:

- Если величина p-value мала, то можно отвергнуть нулевую гипотезу в пользу альтернативы.
- Если величина p-value недостаточно мала, то на основании данных нельзя отвергнуть нулевую гипотезу в пользу альтернативы.

Заметьте, что второй пункт — это не то же самое, что принять нулевую гипотезу. При помощи инструмента проверки гипотез нельзя доказать справедливость нулевой гипотезы. Тот факт, что на основании данных нельзя опровергнуть гипотезу, не означает, что она верна. Отсутствие доказательств — это не доказательство отсутствия. Для монетки, на которой орёл выпадает в 60% случаев, по результату одного броска нельзя опровергнуть нулевую гипотезу о том, что орёл выпадает в 50% случаев.

1.4 Практическая значимость

При любой проверке гипотез нужно оценивать размер эффекта — степень отличия нулевой гипотезы от истины, и оценивать его практическую значимость. Иногда статистически значимые результаты могут не нести практического смысла, а иногда отсутствие статистической значимости при наличии эффекта — это повод собрать больше данных:

- (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.

2 Примеры критериев

Критерии условно подразделены на две группы: параметрические и непараметрические критерии. Параметрические основаны на конкретном типе распределения исходной выборки или используют параметры этого распределения (среднее, дисперсию и т. д.). Непараметрические же не базируются на предположении о типе распределения выборки и не использует его параметры. К сожалению, реальные данные очень редко распределены как табличные распределения, что ограничивает применимость параметрических критериев. Однако, есть ряд популярных случаев, когда это так. В остальном же обычно применяются непараметрические критерии.

выборка:	$X^n = (X_1, \ldots, X_n), X \sim Ber(p)$
нулевая гипотеза:	$H_0 : p = p_0$
альтернатива:	$H_1 : p <\neq> p_0$
статистика:	$Z_S(X^n) = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$
нулевое распределение:	$N(0,1)$

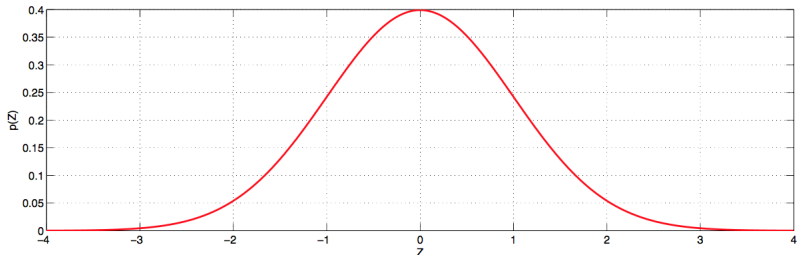


Рис. 3: Z-критерий меток для доли

выборка:	$X^n = (X_1, \ldots, X_n), X \sim Ber(p)$
нулевая гипотеза:	$H_0 : p = p_0$
альтернатива:	$H_1 : p <\neq> p_0$
статистика:	$T(X^n) = \sum_{i=1}^{n} X_i$
нулевое распределение:	$Bin(n, p_0)$

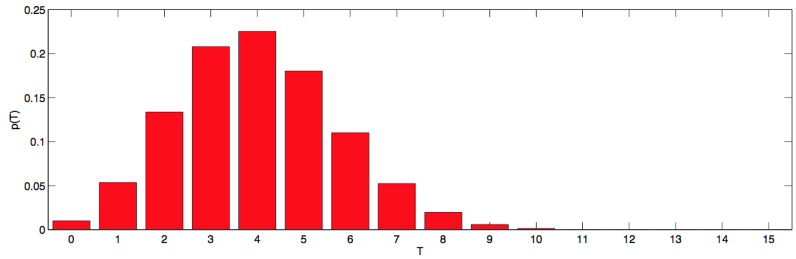


Рис. 4: Биномиальный критерий

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

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

нулевое распределение: $N(0,1)$

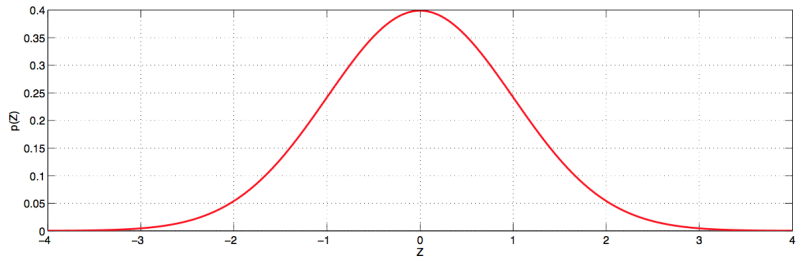


Рис. 5: Z-критерий разности долей, независимые выборки

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

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

нулевая гипотеза: $H_0 : p_1 = p_2$

альтернатива: $H_1 : p_1 <\neq> p_2$

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

нулевое распределение: $N(0,1)$

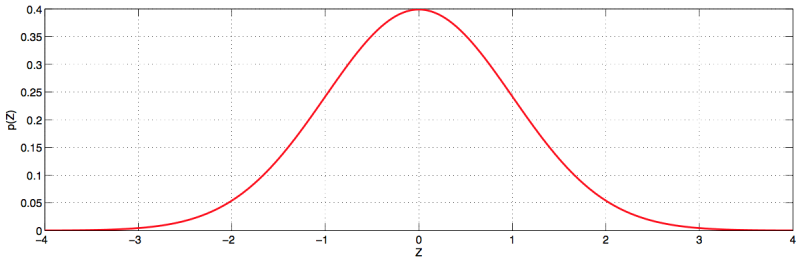


Рис. 6: Z-критерий разности долей, связанные выборки

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

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

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

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

нулевое распределение: $N(0,1)$

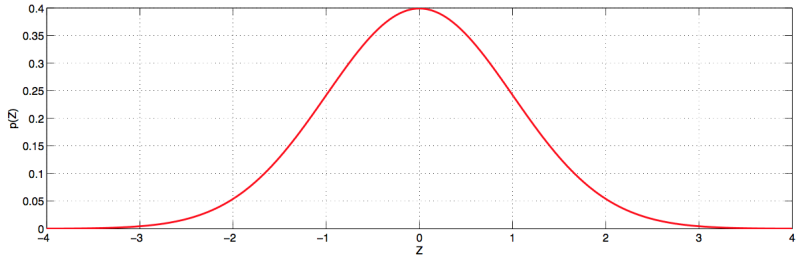


Рис. 7: Z-критерий

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

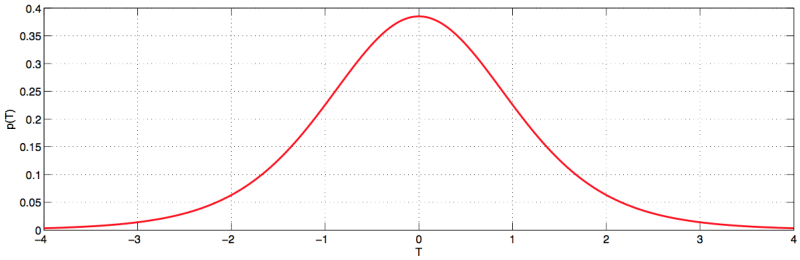
нулевая гипотеза: $H_0 \colon \mu_1 = \mu_2$

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

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

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

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



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

Рис. 8: t-критерий Стьюдента

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 $X = X_1^{n_1} \bigcup X_2^{n_2}$

$$R_1(X_1^{n_1}, X_2^{n_2}) = \sum_{i=1}^{n_1} \operatorname{rank}(X_{1i})$$

нулевое распределение: табличное

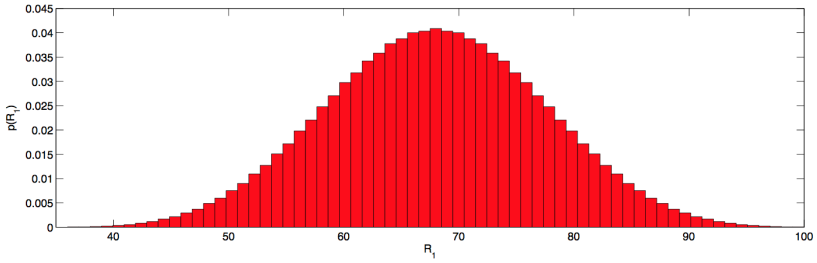


Рис. 9: Критерий Мана-Уитни

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: $H_1 \colon H_0$ неверна

Критерий Смирнова

статистика: $$D(X_1^{n_1}, X_2^{n_2}) = \sup_{-\infty<x<\infty} |F_{n_1 X_1}(x) - F_{n_2 X_2}(x)|$$

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

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

Статистики имеют табличные распределения при H_0.

Рис. 10: Критерии согласия

3 Численные критерии

Иногда прямое применение теоретических тестов невозможно, например, данных слишком много и вычисление рангов для применения критерия Мана-Уитни невозможно. Для таких ситуаций используются численные статистические критерии.

3.1 Перестановочные тесты

выборка:	$X_1^n = (X_1, \ldots, X_n)$
	$F(X)$ симметрично относительно матожидания
нулевая гипотеза:	$H_0 \colon \mathbb{E}X = m_0$
альтернатива:	$H_1 \colon \mathbb{E}X <\neq> m_0$
статистика:	$T(X^n) = \sum_{i=1}^{n} (X_i - m_0)$
нулевое распределение:	порождается перебором 2^n знаков
	перед слагаемыми $X_i - m_0$

Достигаемый уровень значимости — доля перестановок знаков, на которых получилось такое же или ещё более экстремальное значение статистики.

Рис. 11: Перестановочный тест для одной выборки

выборки:	$X_1^{n_1} = (X_{11}, \ldots, X_{1n_1})$
	$X_2^{n_2} = (X_{21}, \ldots, X_{2n_2})$
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:	$T(X_1^{n_1}, X_2^{n_2}) = \frac{1}{n_1} \sum_{i=1}^{n_1} X_{1i} - \frac{1}{n_2} \sum_{i=1}^{n_2} X_{2i}$
нулевое распределение:	порождается перебором $C_{n_1+n_2}^{n_1}$
	размещений объединённой выборки

Рис. 12: Перестановочный тест для двух независимых выборки

Множество всех возможных перестановок может быть слишком велико, поэтому для оценки нулевого распределения обычнно генерируют случайное подмножество перестановок. Если сгенерировано k перестановок, то стандартное отклонение для уровня значимости p будет $\sqrt{\frac{p(1-p)}{k}}$.

3.2 Бутстреп

Альтернативой для перестановочных тестов является бутстеп. Допустим имеется выборка $X^n = (X_1, \ldots, X_n)$, тогда мы можем сгенерировать псевдовыборку $\tilde{X}^n = (\tilde{X}_1, \ldots, \tilde{X}_n)$, где $\tilde{X}_i$ — это случайно выбранный элемент из X^n ($\tilde{X}_i$ могут повторяться). Такой процесс генерации можно повторить много раз, для каждой псевдовыборки мы можем оценить требуемую величину (среднее, дисперсия, медиана и т.д.), таким образом оценив распределение этой величины. Имея распределение статистики, можно построить доверительный интервал на разность статистик между группами.

3.3 Бутстреп и перестановочные тесты

Небольшой набор отличий бутстрепа и перестановочных тестов:

- Перестановочный критерий измеряет расстояние от 0 до статистики
- Бутстреп измеряет расстояние от статистики до 0
- Перестановочный критерий точный
- Бутстреп-критерий приближённый
- Перестановочный критерий проверяет $H_0 \colon F_{X_1}(x) = F_{X_2}(x)$ против $H_1 \colon F_{X_1}(x) = F_{X_2}(x + \Delta),\ \Delta > 0$
- Бутстреп-критерий проверяет $H_0 : \mathbf{E}X_1 = \mathbf{E}X_2$ против $H_1 : \mathbf{E}X_1 > \mathbf{E}X_2$

4 A/B тестирование

На практике часто возникают ситуации, когда есть несколько решений и нужно сравнить какое из них лучше, причём на исторических данных полное сравнение проделать невозможно. Например, есть два алгоритма рекомендательных систем, и вы хотите понять какой из них принесёт вам больше дополнительной прибыли после внедрения на сайт. Для это вам нужно разделить людей на две группы, в каждой из групп использовать только один алгоритм, а затем сравнить в какой группе было заработано больше денег. Но есть нюансы:

- Как правильно разбить пользователей на группы?
- Как отличить случайное различие между группами и неслучайное?
- Как заранее оценить срок теста?

Ответить на эти вопросы помогает аппарат проверки статистических гипотез

4.1 Разбиение на группы и применение критериев

Требуется разбить множество объектов на две тестовые группы. Хочется, чтобы это было репрезентативное разбиение. Фактически, это означает, что вы хотите, чтобы не было статистически значимых различий между группами по важным для задачи признакам. Для этого можно подсчитать эти признаки в группах и применить статистические тесты для сравнения параметров. Если они отвергают гипотезу о равенстве, то это повод переразбить пользователей. Обычно стоит смотреть следующие признаки:

- Статические фичи (пол, возраст и т.п.) распределены одинаково — Критерии согласия и др
- Исторические фичи (конверсии за какой-то период, покупки и другие важные для задачи метрики). Распределения врядли будут совсем совпадать, но стоит проверить разные статистики (среднее, медианы, дисперсии) — Непараметрические критерии
- AA-тест. Могут сказаться технические проблемы, поэтому частой практикой является запуск в продакшн двух одинаковых решений и сравнение их результатов — в группах не должно быть статистически значимых различий между целевыми метриками.

После запуска теста вы будете использовать те же самые тесты, но уже для различающих групп, таким образом честно оценивая эффект от внедрения.

4.2 Оценка сроков теста

Перед запуском теста нужно представлять на какой срок вы его запускаете и для оценки этой величины есть два подхода:

- Из соображений мощности (минимизация ошибки второго рода). Вы хотите, чтобы если эффект был, то с большой вероятностью нулевая гипотеза не отвергалась.
- Из соображений значимости результата. Вы хотите, что если вы получили какой-то эффект, то он был статистически значим.

Давайте разберём эти два подхода на примере бросания монетки (на практике это может быть конверсия посетителя в покупку на сайте): имеется монетка, по результатам бросков, мы хотим понять честная ли она (вероятность орла 50%) и какова вероятность выпадения орла. Вам нужно заранее оценить сколько бросков вам для этого потребуется.

Понятно, что если вероятность орла 50.000001%, то мы врядли сможем за разумное количество бросков его отличить от 50%, да и сама разница кажется практически не существенной. Поэтому нас интересует разница хотя бы в 1%. Также допустим, что мы используем Z-критерий меток для доли на уровне значимости 0.05.

Первый подход. Мы хотим, чтобы если настоящая вероятность монетки больше 51%, то критерий отвергал гипотезу о равенстве 50% с вероятностью хотя бы 80% (стандартное значение для вероятности ошибки второго рода). Для этого мы можно просимулировать броски с вероятностью 51% и применение критерия, на основании чего подобрать значение размера выборки. В данной задаче получится примерно 20000.

Второй подход. Мы хотим, чтобы если после оценки вероятности орла мы получили число больше 51%, то это было статистически значимое отличие. То есть, чтобы $\frac{0.51-0.5}{\sqrt{\frac{0.5(1-0.5)}{n}}} > 1.96$, откуда получаем, что примерно $n > 10000$. Таким образом, потребуется 10000 бросков, чтобы оценка вероятности орла большая 0.51 считалась статистически значимой. При таком размере выборки и ограничении на эффект мощность критерия будет 52%.

Оба способа дают похожие результаты, первый более честный и правильный, но требует симуляции и подбора параметров, что требует больший затрат времени. Второй способ менее честный, но гораздо более быстрый для применения.

4.3 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.

4.3.1 Последовательное применение критериев

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?

К сожалению, нет. 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. Чтобы корректно мониторить результаты каждый день нужно применять Статистический последовательный анализ. В нём используются специальные критерии, специально разработанные для работы с потоками данных и корректно учитывающие эту особенность.

4.3.2 Множественные проверки

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?

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:

$$\alpha \leqslant P(p_1 \leqslant \alpha \ or \ p_2 \leqslant \alpha \ or \ p_3 \leqslant \alpha | H_0) \leqslant \sum_i P(p_i \leqslant \alpha | H_0) = 3\alpha$$

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.

Ошибка первого рода вызвана не особенностью данных, а тем, что мы несколько раз её проверяем. Для корректной проверки гипотез в этом случае надо применять методы Множественной проверки гипотез. Самый простой способ — уменьшить α в число гипотез раз, но есть и более сложные подходы.