Основы статистики и статистические критерии

Александр Сахнов
linkedin.com/in/amsakhnov

Staff MLE at Alibaba Group

2 сентября 2021 г.

Оглавление

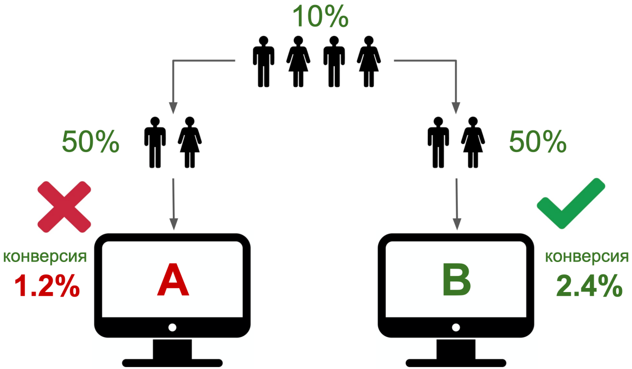
Как принимали решения раньше

- Опрос пользователей
- Сравнение метрики до и во время эксперимента
- Мнение эксперта

Что такое AB тестирование?

Definition

AB тестирование — метод, который позволяет на основе сравнения пилотной и контрольной групп оценивать изолированный эффект внедряемых изменений.



Статистическая гипотеза

Definition

Статистическая гипотеза — любое предположение о распределении и свойствах случайной величины.

Example (Примеры)

- Несколько простых гипотез: $H_0 = \{F = F_0\},\ H_1 = \{F = F_1\}$;
- Простая основная гипотеза и сложная альтернатива: $H_0 = \{\mathbb{E}X = \mathbb{E}Y\},\ H_1 = \{\mathbb{E}X \neq \mathbb{E}Y\}$;
- Гипотеза независимости: $H_0 = \{\mathbb{P}(X) = \mathbb{P}(X|Y)\},\ H_1 = \{H_0 \text{ неверна}\}$;

Статистический критерий

Definition

Статистический критерий — математическое правило, позволяющее по реализациям выборок отвергнуть или не отвергнуть нулевую гипотезу с заданным уровнем значимости.

Дана выборка $X_1, \dots, X_n \sim F$.
Хотим проверить простую гипотезу H_0 против сложной альтернативы H_1.

Пусть можно задать функцию $t(X^n)$, обладающую свойствами:

1. если H_0 верна, то $t(X^n) \Rightarrow G$, где G - непрерывное распределение;
2. если H_0 неверна, то $|t(X^n)| \xrightarrow{P} \infty$ при $n \to \infty$.

Для СВ $Y \sim G$ определим постоянную C из равенства $\alpha = \mathbb{P}(|Y| \geq C)$.
Тогда критерий:

$$\delta(X^n) = \begin{cases} H_0, & \text{если } t(X^n) < C, \\ H_1, & \text{если } t(X^n) \geq C \end{cases}$$

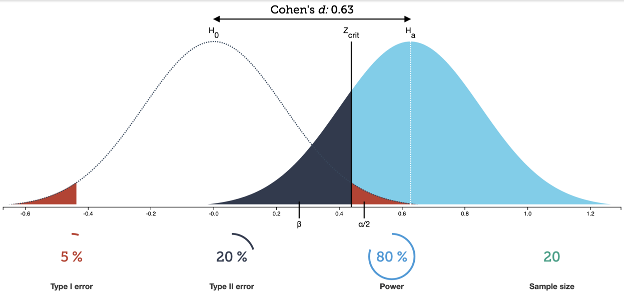
Уровень значимости и мощность теста

Definition

Уровень значимости — вероятность отклонить нулевую гипотезу при условии её истинности, вероятность совершения *ошибки первого рода*.

Definition

Статистическая мощность — вероятность отклонения основной гипотезы в случае, когда альтернативная гипотеза верна. Чем выше мощность теста, тем меньше вероятность совершить *ошибку второго рода*.



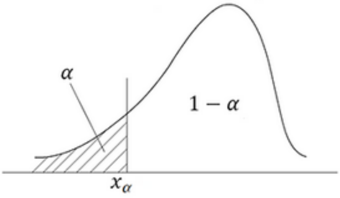
Тест на равенство средних. Критическая область

Есть две выборки $X_1, \dots, X_n \sim F_1$ и $Y_1, \dots, Y_n \sim F_2$.
Определим гипотезы $H_0 : \mathbb{E}X = \mathbb{E}Y$ и $H_1 : \mathbb{E}X \neq \mathbb{E}Y$
Рассмотрим распределение случайной величины $t = \langle X^n \rangle - \langle Y^n \rangle$.

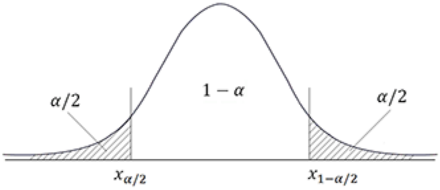
Definition

Критическая область — область выборочного пространства, при попадании в которую нулевая гипотеза отклоняется.

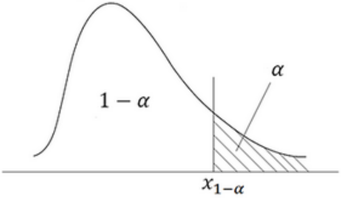
левосторонняя к.о.



двусторонняя к.о.



правосторонняя к.о.



Этапы проверки статистических гипотез

1. Выдвижение основной гипотезы H_0 и альтернативной гипотезы H_1.
2. Выбор уровня значимости α, на котором будет сделан вывод о справедливости гипотезы. Он равен вероятности допустить ошибку первого рода.
3. Расчет статистики критерия такой, что она зависит от выборки и по её значению можно сделать вывод об истинности нулевой гипотезы.
4. Построение критической области.
5. По попаданию или непопаданию значения статистики в критическую область делается вывод о истинности выдвинутой гипотезы на выбранном уровне значимости.

Тест Стьюдента

Есть две выборки: $X_1, \ldots, X_{n_1} \sim N(\mu_1, \sigma_1)$ и $Y_1, \ldots, Y_{n_2} \sim N(\mu_2, \sigma_2)$.
Гипотезы: $H_0 : \mathbb{E}X = \mathbb{E}Y$ и $H_1 : \mathbb{E}X \neq \mathbb{E}Y$.

Средние выборок

$$\overline{X} = \frac{1}{n_1}\sum_{i=1}^{n_1} X_i, \qquad \overline{Y} = \frac{1}{n_2}\sum_{i=1}^{n_2} Y_i$$

Оценки дисперсий

$$S_X^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(X_i - \overline{X})^2, \qquad S_Y^2 = \frac{1}{n_2 - 1}\sum_{i=1}^{n_2}(Y_i - \overline{Y})^2$$

Статистика теста

$$t(X^n, Y^n) = \frac{\overline{Y} - \overline{X}}{\sqrt{\frac{S_X^2}{n_1} + \frac{S_Y^2}{n_2}}} \approx St(\nu), \qquad \nu = \frac{\left(\frac{S_X^2}{n_1} + \frac{S_Y^2}{n_2}\right)^2}{\frac{S_X^4}{n_1^2(n_1 - 1)} + \frac{S_Y^4}{n_2^2(n_2 - 1)}}$$

Тест Стьюдента

Предположения

- Средние значения выборок распределены нормально
- Дисперсии выборок равны
- Выборки независимы друг от друга

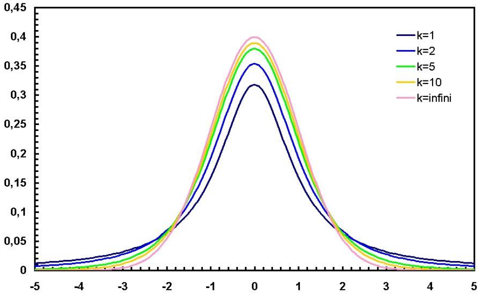


Рис.: Плотность распределения Стьюдента

При неизвестных дисперсиях распределение статистики t имеет приближенное распределение (проблема Беренса - Фишера).
Приближение достаточно точное при выполнении одного из следующих условий

- $n_1 = n_2$
- $\mathbb{I}(n_1 > n_2) = \mathbb{I}(\sigma_1 > \sigma_2)$

Тест Манна-Уитни

Есть две выборки: $X_1, \ldots, X_{n_1} \sim F_X$ и $Y_1, \ldots, Y_{n_2} \sim F_Y$.
Гипотезы: $H_0 : F_X(t) = F_Y(t)$ и $H_1 : F_X(t) = F_Y(t + \Delta), \Delta \neq 0$.

Составить единый ранжированный ряд из обеих сопоставляемых выборок.
Подсчитать отдельно сумму рангов выборок R_1 и R_2.
Вычислить

$$U_1 = R_1 - \frac{n_1(n_1+1)}{2}, \qquad U_2 = R_2 - \frac{n_2(n_2+1)}{2}$$

Значение U-статистики Манна-Уитни

$$U = \min\{U_1, U_2\}$$

Пример

Две выборки $X = \{1, 4\}$, $Y = \{2, 5, 7\}$. Объединим выборки $\{1, 2, 4, 5, 7\}$. Суммы рангов $R_1 = 4$, $R_2 = 11$.
$U_1 = 4 - 3 = 1$, $U_2 = 11 - 6 = 5$
$U = \min\{1, 5\} = 1$

Тест Манна-Уитни

Свойства

- $0 \leq U_1 \leq n_1 n_2$, $0 \leq U_2 \leq n_1 n_2$
- $U_1 + U_2 = n_1 n_2$
- При $n_1, n_2 \geq 20$ распределение $U(X^n, Y^n) \sim N\left(\frac{n_1 n_2}{2}, \frac{n_1 n_2 (n_1 + n_2 + 1)}{12}\right)$

Предположения

- X и Y из одного распределения с точностью до сдвига
- Элементы внутри выборок независимы
- Выборки независимы друг от друга

Замечания

- Если есть дублирующиеся значения, то для них нужно проставить их средний ранг и внести корректировку в аппроксимирующее нормальное распределение.
- Устойчив к выбросам, результат может отличаться от теста Стьюдента.

Критерий Колмогорова

Дана выборка $X_1, \dots, X_n \sim F$.
Гипотезы: $H_0 : F = F_0$ и $H_1 : F \neq F_0$.

Если F_0 непрерывная, то можно пользоваться критерием Колмогорова.
Определим статистику $t(X^n) = \sqrt{n} \sup\limits_x |\hat{F}_n(x) - F_0(x)|$.

Если H_0 верна, то статистика $t(X^n)$ имеет распределение Колмогорова

$$K(x) = \begin{cases} \sum\limits_{i=-\infty}^{\infty} (-1)^i e^{-2i^2x^2} & , \; x > 0 \\ 0 & , \; x \leq 0 \end{cases}$$

Распределение Колмогорова табулировано.
Критерий Колмогорова

$$\delta(X^n) = \begin{cases} H_0, \text{ если } t(X^n) < C, \\ H_1, \text{ если } t(X^n) \geq C \end{cases}$$

Этапы проверки статистических гипотез

Критерий отношения правдоподобия

Функция правдоподобия $L(\theta) = \prod_{i=1}^{n} p(X_i; \theta)$.

Гипотезы: $H_0 : \theta \in \Theta_0$ и $H_1 : \theta \notin \Theta_0$.

Статистика отношения правдоподобия

$$\lambda(X^n) = 2 \ln \left(\frac{\sup_{\theta \in \Theta} L(\theta)}{\sup_{\theta \in \Theta_0} L(\theta)} \right) = 2 \ln \left(\frac{L(\hat{\theta})}{L(\hat{\theta}_0)} \right)$$

где $\hat{\theta}$ - ОМП, $\hat{\theta}_0$ - ОМП при условии $\theta \in \Theta_0$.

Допустим $\theta = (\theta_1, \dots, \theta_q, \theta_{q+1}, \dots, \theta_r)$. Пусть $\Theta_0 = \{\theta : (\theta_{q+1}, \dots, \theta_r) = (\theta_{0,q+1}, \dots, \theta_{0,r})\}$.
Если H_0 верна, то

$$\lambda(X^n) \sim \chi^2_{r-q,\alpha}$$

где $r - q$ — размерность Θ минус размерность Θ_0, α - уровень значимости.
Критерий отношения правдоподобия

$$\delta(X^n) = \begin{cases} H_0, & \text{если } \lambda(X^n) < C, \\ H_1, & \text{если } \lambda(X^n) \geq C \end{cases}$$

Доверительный интервал

Доверительным интервалом с доверительной вероятностью $1-\alpha$ для параметра θ называется интервал $C_n = (a, b)$, где $a = a(X_1, \ldots, X_n)$ и $b = b(X_1, \ldots, X_n)$ - такие функции выборки, что $\mathbb{P}(\theta \in C_n) \geq 1-\alpha$.

Возьмём в качестве параметра θ разность средних распределений. Нулевая гипотеза $H_0 : \mathbb{E}X = \mathbb{E}Y$. Тогда критерий проверки гипотезы о равенстве средних будет иметь вид

$$0 \notin (a, b) \quad \Leftrightarrow \quad \text{отвергнуть гипотезу } H_0$$

Примеры расположения доверительного интервала относительно нуля:

В первом случае значимых отличий нет, ноль внутри доверительного интервала.
Во втором случае значимые отличия есть, ноль вне доверительного интервала.

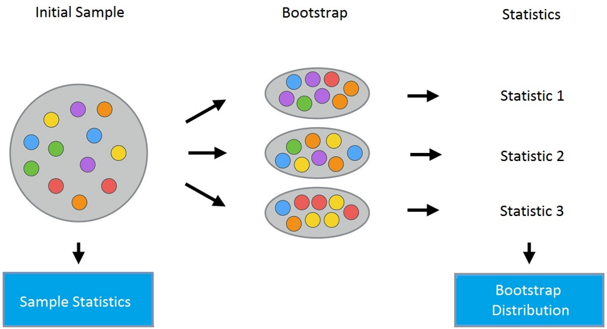
Бутстреп

Definition

Бутстреп — это метод для подсчета стандартных ошибок и нахождения доверительных интервалов статистических функционалов.

Хотим оценить распределение статистики T по выборке $X_1, \dots, X_n$.

- Генерируем B подвыборок из выборки X^n;
- Вычисляем статистику Т для каждой подвыборки;
- Оцениваем распределение по получившемуся множеству статистик.

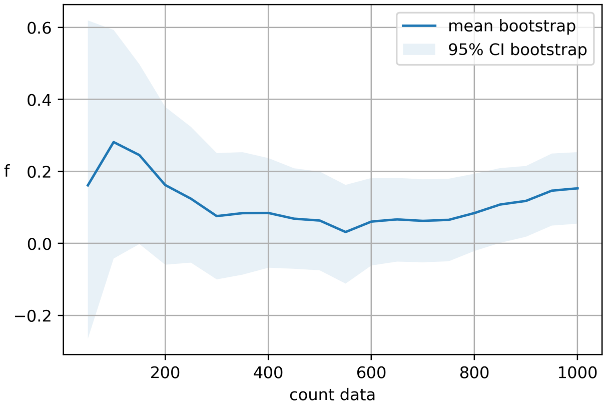


Пример

Дана функция

$$f(x) = x\cos(71x) + \frac{\sin(13x^2)}{x}$$

Хотим оценить $m = \mathbb{E}(f(X))$ по выборке $X_1, \ldots, X_n \sim N(1,1)$.
Точечная оценка $\hat{m} = n^{-1}\sum_i f(X_i)$.
Для оценки разброса оценки воспользуемся бутстрепом.



Проверка гипотезы о равенстве средних

Есть две выборки: $X_1, \dots, X_n \sim F_X$ и $Y_1, \dots, Y_n \sim F_Y$.
Гипотезы: $H_0 : \mathbb{E}X = \mathbb{E}Y$ и $H_1 : \mathbb{E}X \neq \mathbb{E}Y$.

Генерируем B пар подвыборок из выборок X^n, Y^n размером n.
Для каждой пары считаем разность выборочных средних $\{T_{n,1}, \dots, T_{n,B}\}$.
По получившимуся множеству разностей строим доверительный интервал и проверяем гипотезу.

Нормальный интервал

Предположим, что полученное множество статистик, посчитанных на бутстрепных данных, имеет нормальное распределение. Тогда

$$C_n = \left(T - z_{\alpha/2}\hat{se}_{boot},\, T + z_{\alpha/2}\hat{se}_{boot}\right)$$

где разность выборочных средних $T = \overline{Y^n} - \overline{X^n}$, оценка стандартной ошибки на основе бутстрепа $\hat{se}_{boot} = \sqrt{1/B\sum_{i=1}^{B}(T_{n,i} - \overline{T_n^B})^2}$, модуль квантиля стандартного нормального распределения $z_{\alpha/2}$.

Интервал на основе процентилей

$$C_n = (\theta^*_{\alpha/2}, \theta^*_{1-\alpha/2})$$

где θ^*_α - квантили посчитанные по множеству $\{T_{n,1}, \dots, T_{n,B}\}$.
Можно подобрать монотонное преобразование, которое преобразует распределение статистики T к распределению похожее на нормальное. Так как монотонное преобразование сохраняет квантили, то квантили $\theta_{\alpha/2}, \theta_{1-\alpha/2}$ перейдут в соответствующие кванитили нормального распределения. Тогда легко показать, что вероятность попасть в определённый выше интервал равна $1-\alpha$.

Центральный интервал

Пусть $\theta = T(F)$ и $\hat{\theta}_n = T(\hat{F}_n)$.
Введём $R_n = \hat{\theta}_n - \theta$ с распределением $H(r) = \mathbb{P}_F(R_n \leq r)$.
Определим доверительный интервал

$$C_n = (a, b) = \left(\hat{\theta}_n - H^{-1}\left(1 - \frac{\alpha}{2}\right), \hat{\theta}_n - H^{-1}\left(\frac{\alpha}{2}\right)\right)$$

Легко показать, что $\mathbb{P}(\theta \in C_n) = 1 - \alpha$, но мы не знаем $H(r)$.
Оценим его с помощью бутстрепа

$$\hat{H}(r) = \frac{1}{B}\sum_{i=1}^{B}\mathbb{I}(R^*_{n,i} \leq r), \qquad R^*_{n,i} = \hat{\theta}^*_{n,i} - \hat{\theta}_n$$

Пусть r^*_β - β выборочная квантиль, посчитанная по выборке $(R^*_{n,1}, \ldots, R^*_{n,B})$, а θ^*_β - β выборочная квантиль, посчитанная по выборке $(\hat{\theta}^*_{n,1}, \ldots, \hat{\theta}^*_{n,B})$, тогда

Центральный интервал

$$\hat{a} = \hat{\theta}_n - \hat{H}^{-1}\left(1 - \frac{\alpha}{2}\right) = \hat{\theta}_n - r^*_{1-\alpha/2} = 2\hat{\theta}_n - \theta^*_{1-\alpha/2}$$
$$\hat{b} = \hat{\theta}_n - \hat{H}^{-1}\left(\frac{\alpha}{2}\right) = \hat{\theta}_n - r^*_{\alpha/2} = 2\hat{\theta}_n - \theta^*_{\alpha/2}$$

Получаем центральный доверительный интервал

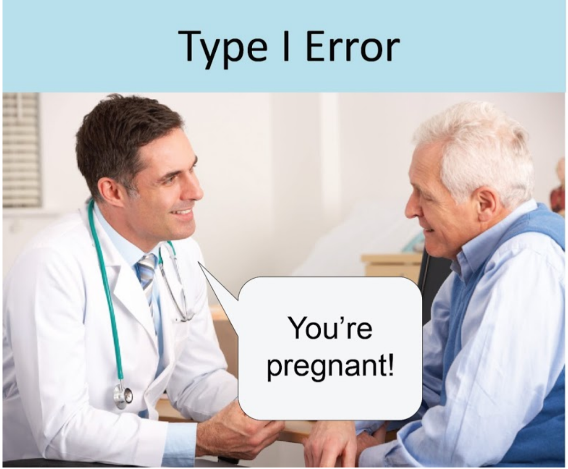
$$C_n = (2\hat{\theta}_n - \theta^*_{1-\alpha/2}, 2\hat{\theta}_n - \theta^*_{\alpha/2})$$

Заметим, что $\mathbb{P}(T(F) \in C_n) \to 1 - \alpha$ при $n \to \infty$.

Итого, Бутстреп

- Позволяет оценить распределение некоторой функции от случайной выборки
- Не делает предположений о виде распределения
- Много вычислений

Ошибки I и II рода



Ошибки I и II рода

Нашли эффект, когда его нет

(ошибка I рода)

“Новые стеллажи увеличат средний чек на 1%.”

Установить 14000 стеллажей, которые ничего не меняют.

Не нашли эффект, когда он был

(ошибка II рода)

Смена ассортимента увеличивает выручку на 2%

Провели пилот на 5 магазинах, тк дорого оборудовать.
Не увидели стат значимого эффекта.

Оценка ошибок I и II рода

Оценка ошибки I рода

1. генерируем пилотную и контрольную группы
2. на исторических данных, где не был запущен эксперимент, считаем метрики для групп
3. оцениваем значимость отличия средних и запоминаем результат
4. повторяем первые три пункт, чтобы набрать статистику
5. считаем долю случаев, когда средние значения отличались значимо

Оценка ошибки II рода

1. генерируем пилотную и контрольную группы
2. на исторических данных, где не был запущен эксперимент, считаем метрики для групп, к метрикам пилотной группы добавляем эффект
3. оцениваем значимость отличия средних и запоминаем результат
4. повторяем первые три пункт, чтобы набрать статистику
5. считаем долю случаев, когда средние значения не отличались значимо

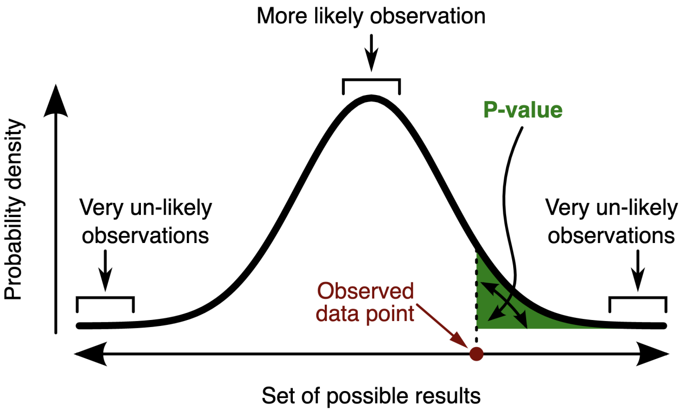
p-value

p_{value} - вероятность при нулевой гипотезе наблюдать полученное или более экстремальное значение статистики.

$$p_{value} = \mathbb{P}(T > t | H_0)$$

Статистический критерий можно записать как

$$p_{value} < \alpha \quad \Leftrightarrow \quad \text{отвергнуть гипотезу}$$



Распределение p-value

Theorem

Пусть тест размера α имеет вид: отвергнуть $H_0 \Leftrightarrow T(x^n) > c(\alpha)$,
где x^n - наблюдаемая выборка.
Если H_0 верна, то $p_{value}(x^n) = \mathbb{P}(T(X^n) > T(x^n)|H_0)$.
Из последнего свойства также следует, что $p_{value}(X^n) \sim Uniform(0,1)$.

Утверждение

Пусть случайная величина X имеет распределение $F(x)$, и $F(x)$ обратима. Тогда случайная величина $Y = F(X)$ имеет распределение $Uniform(0,1)$.

Док-во: $\mathbb{P}(F(X) < x) = \mathbb{P}(X < F^{-1}(x)) = F(F^{-1}(x)) = x, x \in (0,1)$. □

Случайная величина $Y = 1 - F(X)$ также является равномерной, причём:

$$Y = 1 - F(X) = \mathbb{P}_{Z \sim F(.)}(Z > X),$$

Теперь p_{value} подходит под роль Y в утверждении выше: $Y = p_{value}(X^n),\ X = T(X^n)$.

Что дальше?

Пайплайн АВ теста

- Гипотеза
- Метрики и алгоритм принятия решений
- Ожидаемый эффект и размер групп
- Подбор групп
- Проведение пилота
- Обработка результатов

Материалы

Материалы для самостоятельного изучения

1. Larry Wasserman. All of Statistics.