Прикладной статистический анализ данных.
11. Последовательный анализ

Бахтеев Олег
mipt.psad19@gmail.com

2019

Z-критерий меток для доли

Задача: рекламная кампания планировалась так, чтобы обеспечить узнаваемость продукта среди целевой аудитории более 30%. После окончания кампании проводится опрос с целью оценки узнаваемости.

H_0: узнаваемость продукта не превышает 30%.
H_1: узнаваемость продукта превышает 30%.

Z-критерий меток для доли

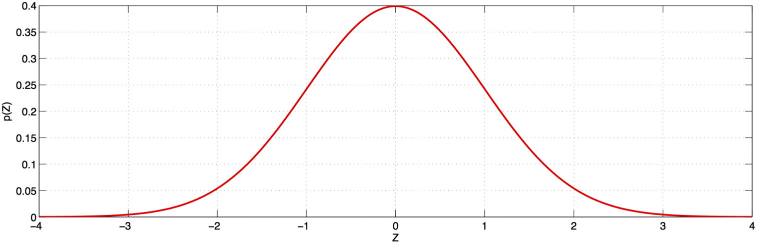
выборка: $X^n = (X_1, \ldots, X_n), X \sim Ber(p)$

нулевая гипотеза: $H_0 \colon p = p_0$

альтернатива: $H_1 \colon p > p_0$

статистика: $Z(X^n) = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}, \quad \hat{p} = \frac{1}{n} \sum\limits_{i=1}^{n} X_i$

нулевое распределение: $N(0, 1)$



Z-критерий меток для доли

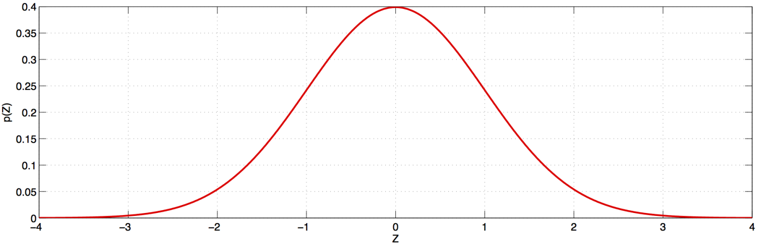
выборка: $X^n = (X_1, \ldots, X_n), X \sim Ber(p)$

нулевая гипотеза: $H_0\colon p \leqslant p_0$

альтернатива: $H_1\colon p > p_0$

статистика: $Z(X^n) = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}, \quad \hat{p} = \frac{1}{n}\sum\limits_{i=1}^{n} X_i$

нулевое распределение: $N(0,1)$ при $p = p_0$



Z-критерий меток для доли

Как выбрать наименьший достаточный объём выборки?

Постановка задачи последовательного анализа

$$\text{выборка:} \quad X^m = (X_1, \ldots, X_m), X \sim Ber(p).$$

Фиксируем «коридор» отклонений значения параметра p от p_0, которые можно считать несущественными:

$$p_L \leqslant p_0 \leqslant p_U$$

(хотя бы одно из неравенств — строгое).

$$\begin{aligned} \text{нулевая гипотеза:} \quad & H_0\colon p \leqslant p_L; \\ \text{альтернатива:} \quad & H_1\colon p \geqslant p_U. \end{aligned}$$

Пусть данные поступают постепенно.
Задача: построить проверку гипотез так, чтобы обойтись как можно меньшим объёмом выборки.

Анонс: процедура последовательного анализа при тех же значениях мощности и уровня значимости позволяет обойтись меньшим (иногда вдвое) объёмом выборки.

Процедура последовательного анализа

Поскольку размер выборки не фиксирован, мы можем фиксировать вероятности ошибок обоих родов:
α — уровень значимости — допускаемая вероятность ошибки первого рода,
β — допускаемая вероятность ошибки второго рода.

$$\text{статистика:} \quad d_m\left(X^m\right) = \sum_{i=1}^{m} X_i.$$

Введём следующие обозначения:

$$A = \frac{1-\beta}{\alpha}, \quad B = \frac{\beta}{1-\alpha},$$

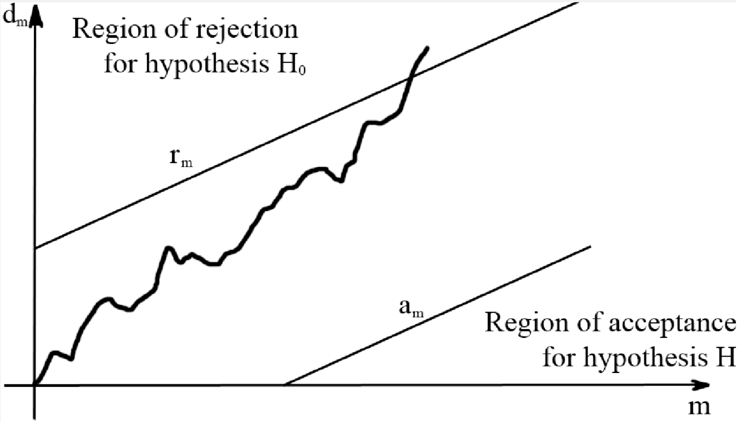
$$a_m = \frac{\ln B + m \ln \frac{1-p_L}{1-p_U}}{\ln \frac{p_U}{p_L} - \ln \frac{1-p_U}{1-p_L}},$$

$$r_m = \frac{\ln A + m \ln \frac{1-p_L}{1-p_U}}{\ln \frac{p_U}{p_L} - \ln \frac{1-p_U}{1-p_L}}.$$

Процедура последовательного анализа

При каждом значении m:

- $d_m \geqslant r_m \Rightarrow$ отвергаем H_0, $p \geqslant p_U$;
- $d_m \leqslant a_m \Rightarrow$ принимаем H_0, $p \leqslant p_L$;
- $a_m < d_m < r_m \Rightarrow$ процесс продолжается, добавляем элемент выборки.



Момент остановки

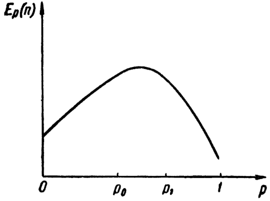
На каком элементы выборки n произойдёт остановка процедуры?

n — случайная величина, можно говорить о её матожидании:

$$\mathbb{E}_p(n) = \frac{L(p)\ln B + (1 - L(p))\ln A}{p \ln \frac{p_U}{p_L} + (1-p)\ln \frac{1-p_U}{1-p_L}},$$

$L(p) = \frac{A^h - 1}{A^h - B^h}$ — оперативная характеристика — вероятность принять нулевую гипотезу при условии, что p — истинное значение параметра;
h определяется как решение уравнения:

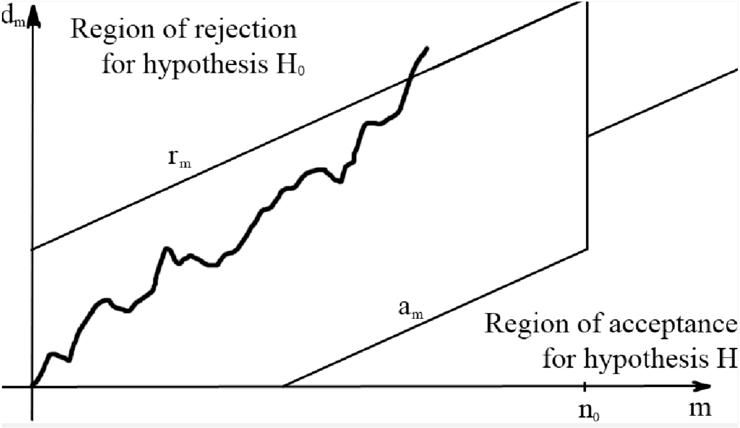
$$p = \frac{1 - \left(\frac{1-p_U}{1-p_L}\right)^h}{\left(\frac{p_U}{p_L}\right)^h - \left(\frac{1-p_U}{1-p_L}\right)^h}.$$



Усечение

Если при $m = n_0$ решение ещё не принято, но возможности добавлять элементы выборки больше нет, используем следующий критерий:

- $d_m \geqslant \frac{a_{n_0}+r_{n_0}}{2} \Rightarrow$ отвергаем H_0, $p \geqslant p_U$;
- $d_m \leqslant \frac{a_{n_0}+r_{n_0}}{2} \Rightarrow$ принимаем H_0, $p \leqslant p_L$.



Группировка наблюдений

Наблюдения могут поступать группами $g_1, g_2, \ldots$ по v элементов. Тогда значения статистики d_m сравниваются с a_m, r_m только при $m = v, 2v, \ldots$.

Последствия:

- увеличивается размер выборки, при котором происходит остановка;
- истинные вероятности ошибок могут оказаться больше номинальных, но при этом
$$\alpha' \leqslant \frac{\alpha}{1-\beta}, \;\; \beta' \leqslant \frac{\beta}{1-\alpha}.$$
Так как величины α и β обычно малы, отклонением можно пренебречь.

Z-критерий для разности двух 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

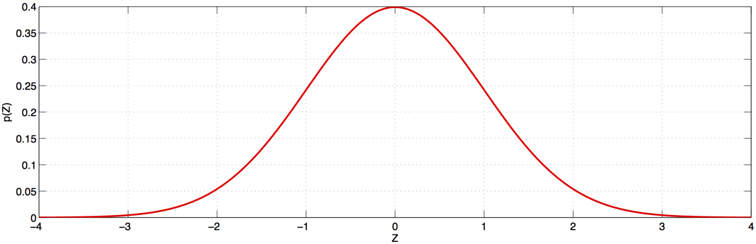
$X_1^n \backslash X_2^n$	1	0
1	e	f
0	g	h

нулевая гипотеза: $H_0 \colon p_1 \geqslant p_2$

альтернатива: $H_1 \colon p_1 < p_2$

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

нулевое распределение: $N(0,1)$ при $p_1 = p_2$



Z-критерий для разности двух долей, связанные выборки

Пример: имеются два технологических процесса, классический и модернизированный, p_1, p_2 — доли брака в них.

H_0: доля брака в классическом процессе не меньше доли брака в модернизированном.
H_1: доля брака в классическом процессе меньше доли брака в модернизированном.

Аналог в последовательном анализе

Пусть значения x_{1i}, x_{2i} поступают парами.
Будем рассматривать только различающиеся пары — $(0, 1)$ и $(1, 0)$, а остальные будем отбрасывать.

$k_1 = \frac{p_1}{1-p_1}$, $k_2 = \frac{p_2}{1-p_2}$ — риски,
$u = \frac{k_1}{k_2} = \frac{p_1(1-p_2)}{p_2(1-p_1)}$ — относительный риск:

- $u = 1 \Leftrightarrow p_1 = p_2$,
- $u > 1 \Leftrightarrow p_1 > p_2$,
- $u < 1 \Leftrightarrow p_1 < p_2$.

Фиксируем «коридор» отклонений u от 1, которые можно считать незначимыми:

$$u_L \leqslant 1 \leqslant u_U$$

(хотя бы одно из неравенств — строгое).

нулевая гипотеза:	$H_0\colon u \geqslant u_U$
альтернатива:	$H_1\colon u \leqslant u_L$
статистика:	$d_m\left(X_1^m, X_2^m\right) = \sum\limits_{i=1}^{m} \left(1 - X_{1i}\right) X_{2i}$

Аналог в последовательном анализе

Константы последовательного анализа:

$$a_m = \frac{\ln B + m \ln \frac{1-u_U}{1-u_L}}{\ln u_U - \ln u_L},$$

$$r_m = \frac{\ln A + m \ln \frac{1-u_U}{1-u_L}}{\ln u_U - \ln u_L}.$$

Момент остановки:

$$\mathbb{E}_u(n) = \frac{L(u) \ln B + (1 - L(u)) \ln A}{\frac{u}{u+1} \ln \frac{u_U(1+u_L)}{u_L(1+u_U)} + \frac{1}{u+1} \ln \frac{1+u_L}{1+u_U}} \Bigg/ \left(p_1(1-p_2) + p_2(1-p_1)\right),$$

$$L(u) = \frac{A^h - 1}{A^h - B^h},$$

h определяется как решение уравнения

$$\frac{u}{u+1} = \frac{1 - \left(\frac{1+u_L}{1+u_U}\right)^h}{\left(\frac{u_U(1+u_L)}{u_l(1+u_U)}\right)^h - \left(\frac{1+u_L}{1+u_U}\right)^h}.$$

Группировка наблюдений

Наблюдения могут поступать группами $g_1, g_2, \ldots$ пар выборок по v элементов. Если при этом внутри пар выборок не указаны соответствия элементов (x_{1i}, x_{2i}), статистику d_m вычислить невозможно.

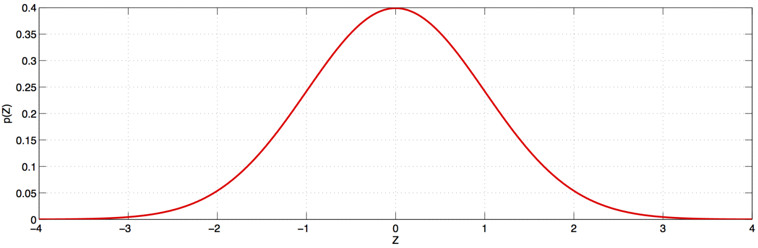
Пусть $v_1\left(g_i\right)$ — число успехов в выборке из v наблюдений над первой биномиальной совокупностью в группе g_i, $v_2(g_i)$ — над второй. Тогда для этой пары групп в качестве оценки числа пар $(0, 1)$ примем величину $v_2(g_i) - \frac{v_1(g_i)v_2(g_i)}{v}$.

$$d_{g_m} = \sum_{i=1}^{g_m} \left(v_2(g_i) - \frac{v_1\left(g_i\right) v_2\left(g_i\right)}{v} \right).$$

Последствия: аналогичные.

Z-критерий для среднего нормального распределения, односторонняя альтернатива

выборка:	$X^n = (X_1, \ldots, X_n)\,, X \sim N\left(\mu, \sigma^2\right), \sigma$ известна
нулевая гипотеза:	$H_0\colon \mu \leqslant \mu_0$
альтернатива:	$H_1\colon \mu > \mu_0$
статистика:	$Z\left(X^n\right) = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$
нулевое распределение:	$N(0,1)$ при $\mu = \mu_0$



Z-критерий для среднего нормального распределения, односторонняя альтернатива

Пример: при помощи прибора с известной погрешностью σ измеряется концентрация вредного вещества в образце. Необходимо проверить, что она не превышает предельно допустимой.

Аналог в последовательном анализе

Фиксируем «коридор» отклонений μ от μ_0, которые можно считать незначимыми:

$$\mu_L \leqslant \mu_0 \leqslant \mu_U$$

(хотя бы одно из неравенств — строгое).

$$\begin{aligned} \text{нулевая гипотеза:} \quad & H_0\colon \mu \leqslant \mu_L \\ \text{альтернатива:} \quad & H_1\colon \mu \geqslant \mu_U \\ \text{статистика:} \quad & d_m\left(X^m\right) = \sum_{i=1}^{m} X_i \end{aligned}$$

Аналог в последовательном анализе

Константы последовательного анализа:

$$a_m = \frac{\sigma^2}{\mu_U - \mu_L} \ln B + m \frac{\mu_U + \mu_L}{2},$$

$$r_m = \frac{\sigma^2}{\mu_U - \mu_L} \ln A + m \frac{\mu_U + \mu_L}{2},$$

Момент остановки:

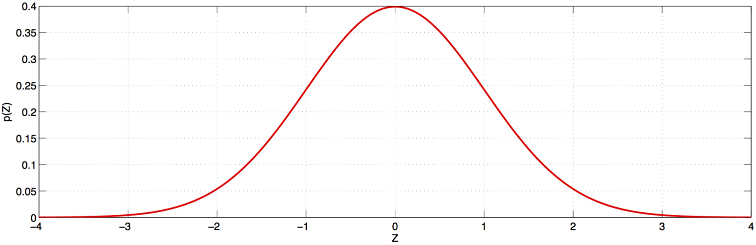
$$\mathbb{E}_\mu\left(n\right) = \frac{L(\mu) \ln B \left(1 - L(\mu)\right) \ln A}{\mu_L^2 - \mu_U^2 + 2\left(\mu_U - \mu_L\right)\mu},$$

$$L\left(\mu\right) = \frac{A^h - 1}{A^h - B^h},$$

$$h = \frac{\mu_U + \mu_L - 2\mu}{\mu_U - \mu_L}.$$

Z-критерий для среднего нормального распределения, двусторонняя альтернатива

выборка: $X^n = (X_1, \ldots, X_n), X \sim N(\mu, \sigma^2), \sigma$ известна
нулевая гипотеза: $H_0\colon \mu = \mu_0$
альтернатива: $H_1\colon \mu \neq \mu_0$
статистика: $Z(X^n) = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$
нулевое распределение: $N(0, 1)$



Z-критерий для среднего нормального распределения, двусторонняя альтернатива

Пример: многократные измерения прибором с известной погрешностью для проверки наличия у прибора смещения.

Аналог в последовательном анализе

Фиксируем симметричный «коридор» отклонений μ от μ_0, которые можно считать незначимыми:

$$\left|\frac{\mu-\mu_0}{\sigma}\right| \leqslant \delta.$$

$$\begin{array}{rl}
\text{нулевая гипотеза:} & H_0: \left|\frac{\mu-\mu_0}{\sigma}\right| \leqslant \delta \\
\text{альтернатива:} & H_1: \left|\frac{\mu-\mu_0}{\sigma}\right| > \delta \\
\text{статистика:} & d_m\left(X^m\right) = \ln\operatorname{ch}\left(\frac{\delta}{\sigma}\sum_{i=1}^{m}\left(X_i-\mu_0\right)\right)
\end{array}$$

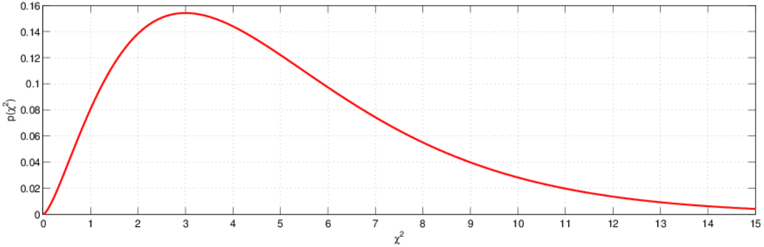
Константы последовательного анализа:

$$a_m = \ln B + m\frac{\delta^2}{2},$$

$$r_m = \ln A + m\frac{\delta^2}{2}.$$

Критерий хи-квадрат

выборка:	$X^n = (X_1, \dots, X_n)\,, X \sim N\left(\mu, \sigma^2\right), \mu$ известно
нулевая гипотеза:	$H_0 \colon \sigma \leqslant \sigma_0$
альтернатива:	$H_1 \colon \sigma > \sigma_0$
статистика:	$\chi^2\left(X^n\right) = \frac{\sum_{i=1}^{n}(X_i - \mu)}{\sigma_0^2}$
нулевое распределение:	χ^2_n при $\sigma = \sigma_0$



Критерий хи-квадрат

Пример: не превышает ли погрешность прибора заявленного уровня?

Аналог в последовательном анализе

Фиксируем «коридор» отклонений σ от σ_0, которые можно считать незначимыми:

$$\sigma_L \leqslant \sigma_0 \leqslant \sigma_U$$

(хотя бы одно из неравенств — строгое).

нулевая гипотеза: $H_0\colon \sigma \leqslant \sigma_L$

альтернатива: $H_1\colon \sigma \geqslant \sigma_U$

статистика: $d_m\left(X^m\right) = \sum\limits_{i=1}^{m}\left(X_i - \mu\right)^2$

Константы последовательного анализа:

$$a_m = \frac{2\ln B + m\ln\frac{\sigma_U^2}{\sigma_L^2}}{\frac{1}{\sigma_L^2} - \frac{1}{\sigma_U^2}},$$

$$r_m = \frac{2\ln A + m\ln\frac{\sigma_U^2}{\sigma_L^2}}{\frac{1}{\sigma_L^2} - \frac{1}{\sigma_U^2}}.$$

Случай неизвестного среднего

Если среднее неизвестно, предлагается использовать его выборочную оценку:

$$\text{статистика:} \quad d_m\left(X^m\right) = \sum_{i=1}^{m}\left(X_i - \bar{X}\right)^2$$

При этом в последовательном анализе на m-м шаге вместо констант a_m, r_m необходимо использовать a_{m-1}, r_{m-1}.

Литература

- последовательная проверка гипотез — Вальд;
- последовательные доверительные интервалы — Mukhopadhyay.

Вальд, А. *Последовательный анализ*, 1960.

Mukhopadhyay, N., de Silva, B. M. *Sequential methods and their applications*, 2009.