Прикладная статистика 15. Последовательный анализ.

20 декабря 2013 г.

Z-критерий для доли

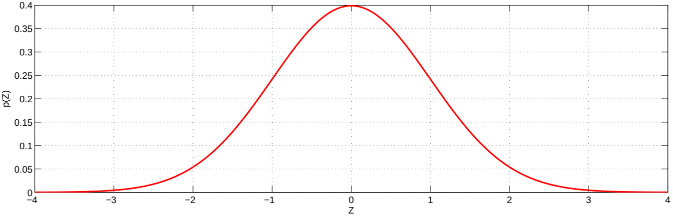
выборка: $X^n = (X_1, \dots, X_n)\,, X_i \sim Ber\,(p)\,;$

нулевая гипотеза: $H_0\colon p = p_0;$

альтернатива: $H_1\colon p > p_0;$

статистика: $Z\,(X^n) = \frac{\hat{p} - p_0 - \frac{1}{2n}}{\sqrt{\frac{p_0(1-p_0)}{n}}}, \quad \hat{p} = \frac{1}{n}\sum\limits_{i=1}^n X_i;$

$Z\,(X^n) \sim N(0,1)$ при H_0;



достигаемый уровень значимости:

$$p\,(z) = 1 - ncdf\,(z, 0, 1).$$

Z-критерий для доли

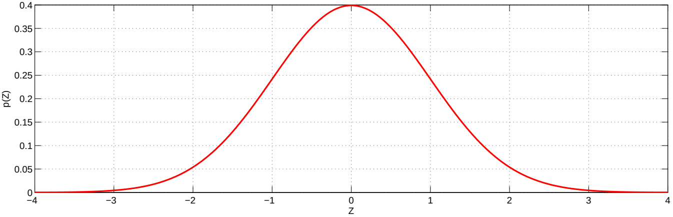
выборка: $X^n = (X_1, \dots, X_n), X_i \sim Ber(p);$

нулевая гипотеза: $H_0\colon p \leq p_0;$

альтернатива: $H_1\colon p > p_0;$

статистика: $Z(X^n) = \frac{\hat{p} - p_0 - \frac{1}{2n}}{\sqrt{\frac{p_0(1-p_0)}{n}}}, \quad \hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i;$

$Z(X^n) \sim N(0,1)$ при $p = p_0$ (наименее благоприятная конфигурация);



достигаемый уровень значимости:

$$p(z) = 1 - ncdf(z, 0, 1).$$

Z-критерий для доли

Пример: 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%.

Как выбрать наименьший достаточный объём выборки?
Например, если истинная узнаваемость составляет 35%, то при уровне значимости 0.05 для обеспечения мощности 0.8 нужно опросить 534 человек.

Постановка задачи последовательного анализа

выборка: $X^m = (X_1, \ldots, X_m)\,, X_i \sim Ber\,(p)\,.$

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

$$p_L \leq p \leq p_U$$

(хотя бы одно из неравенств — строгое).

нулевая гипотеза: $H_0 : p \leq p_L;$

альтернатива: $H_1 : p \geq p_U\,.$

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

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

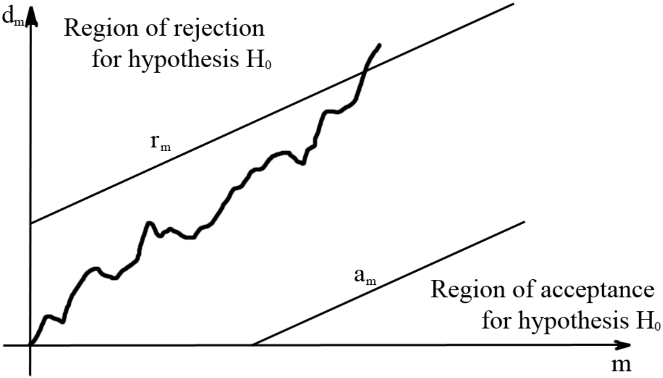
$$a_m = \frac{\ln B + m \ln \frac{1-p_L}{1-p_U}}{\ln \frac{p_U}{p_L} - \ln \frac{1-p_U}{1-p_L}},$$

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

Процедура последовательного анализа

При каждом значении m:

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



Момент остановки

На каком элементы выборки n произойдёт остановка процедуры?

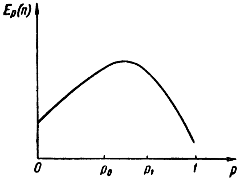
n — случайная величина, можно говорить о её матожидании:

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

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

h определяется как решение уравнения:

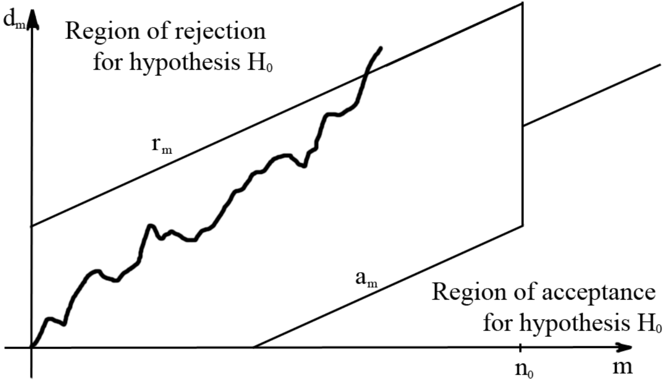
$$p = \frac{1 - \left(\frac{1-p_U}{1-p_L}\right)^h}{\left(\frac{p_U}{p_L}\right)^h - \left(\frac{1-p_U}{1-p_L}\right)^h}.$$



Усечение

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

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



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
$$\alpha' \leq \frac{\alpha}{1-\beta}, \quad \beta' \leq \frac{\beta}{1-\alpha}.$$
Так как величины α и β обычно малы, отклонением можно пренебречь.

Z-критерий для двух долей

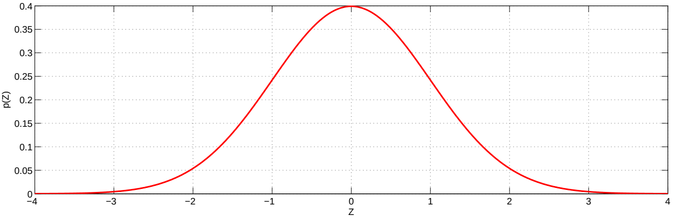
выборки: $X_1^{n_1} = (X_{11}, \dots, X_{1n_1}), X_{1i} \sim Ber(p_1);$
$X_2^{n_2} = (X_{21}, \dots, X_{2n_2}), X_{2i} \sim Ber(p_2);$

нулевая гипотеза: $H_0\colon p_1 \geq p_2;$

альтернатива: $H_1\colon p_1 < p_2;$

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

$Z(X_1^{n_1}, X_2^{n_2}) \sim N(0,1)$ при $p_1 = p_2$ (наименее благоприятная конфигурация);



достигаемый уровень значимости:

$$p(z) = 1 - ncdf(z, 0, 1).$$

Z-критерий для двух долей

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

$$u_L \leq 1 \leq u_U$$

(хотя бы одно из неравенств — строгое).

$$\begin{aligned} \text{нулевая гипотеза:} \quad & H_0\colon u \geq u_U; \\ \text{альтернатива:} \quad & H_1\colon u \leq u_L; \\ \text{статистика:} \quad & d_m\left(X_1^m, X_2^m\right) = \sum_{i=1}^{m} \left(1 - X_{1i}\right) X_{2i}. \end{aligned}$$

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

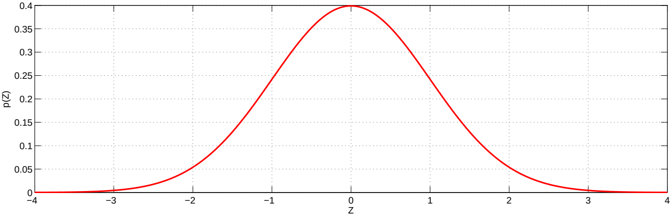
выборка: $X^n = (X_1, \dots, X_n), X_i \sim N(\mu, \sigma^2), \sigma$ известна;

нулевая гипотеза: $H_0\colon \mu \le \mu_0$;

альтернатива: $H_1\colon \mu > \mu_0$;

статистика: $Z(X^n) = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$;

$Z(X^n) \sim N(0,1)$ при $\mu = \mu_0$ (наименее благоприятная конфигурация);



достигаемый уровень значимости:

$$p(z) = 1 - ncdf(z, 0, 1).$$

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:

$$\mu_L \leq \mu_0 \leq \mu_U$$

(хотя бы одно из неравенств — строгое).

нулевая гипотеза:	$H_0\colon \mu \leq \mu_L;$
альтернатива:	$H_1\colon \mu \geq \mu_U;$
статистика:	$d_m(X^m) = \sum\limits_{i=1}^{m} X_i.$

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:

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

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

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

Z-критерий для среднего нормального распределения, двусторонняя альтернатива

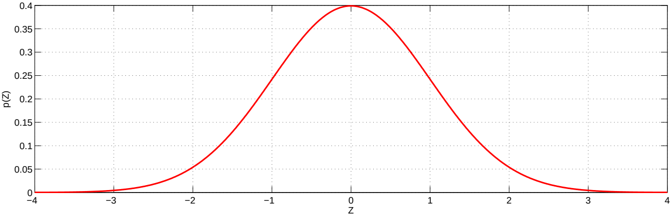
выборка: $X^n = (X_1, \dots, X_n), X_i \sim N(\mu, \sigma^2), \sigma$ известна;

нулевая гипотеза: $H_0\colon \mu = \mu_0$;

альтернатива: $H_1\colon \mu \neq \mu_0$;

статистика: $Z(X^n) = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}}$;

$Z(X^n) \sim N(0,1)$ при $\mu = \mu_0$ (наименее благоприятная конфигурация);



достигаемый уровень значимости:

$$p(z) = 2\left(1 - ncdf(|z|, 0, 1)\right).$$

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:

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

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

Константы последовательного анализа:

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

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

Критерий хи-квадрат

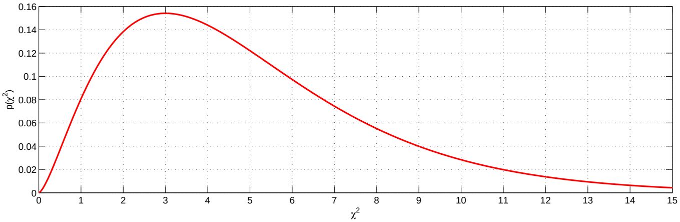
выборка: $X^n = (X_1, \ldots, X_n), X_i \sim N(\mu, \sigma^2), \mu$ известно;

нулевая гипотеза: $H_0 \colon \sigma \leq \sigma_0$;

альтернатива: $H_1 \colon \sigma > \sigma_0$;

статистика: $\chi^2(X^n) = \frac{(n-1)S^2}{\sigma_0^2}$;

$\chi^2(X^n) \sim \chi^2_{n-1}$ при $\sigma = \sigma_0$ (наименее благоприятная конфигурация);



достигаемый уровень значимости:

$$p(\chi^2) = 1 - chi2cdf(\chi^2, n-1).$$

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:

$$\sigma_L \le \sigma_0 \le \sigma_U$$

(хотя бы одно из неравенств — строгое).

$$\begin{aligned} \text{нулевая гипотеза:} \quad & H_0\colon \sigma \le \sigma_L; \\ \text{альтернатива:} \quad & H_1\colon \sigma \ge \sigma_U; \\ \text{статистика:} \quad & d_m\left(X^m\right) = \sum_{i=1}^{m} \left(X_i - \mu\right)^2. \end{aligned}$$

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}.

Прикладная статистика
15. Последовательный анализ.

Рябенко Евгений
riabenko.e@gmail.com