

# Прикладной статистический анализ данных

## Последовательный анализ

Бахтеев Олег  
psad@phystech.edu

2021

# История задачи

## Задача

ВМФ США хочет проверить эффективность двух типов снарядов. Производится  $N$  ( $N \approx 1000$ ) раундов обстрелов целей двумя типами снарядов, после чего производится статистический тест.

## Проблема

Статистический тест может быть слишком расточительным. Часто знающий офицер может “на глаз” понять, какой снаряд лучше после нескольких сотен раундов.

## 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

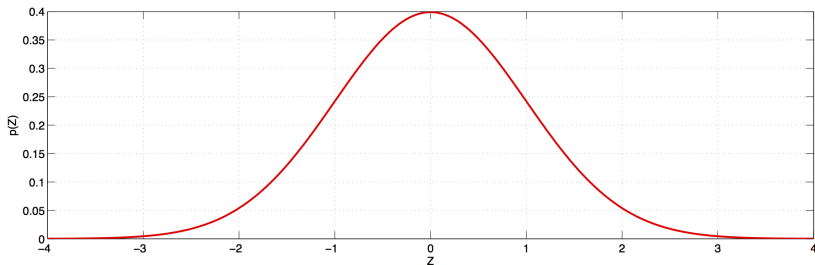
выборка:  $X^n = (X_1, \dots, X_n), X \sim \text{Ber}(p)$

нулевая гипотеза:  $H_0: p = p_0$

альтернатива:  $H_1: p > p_0$

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

нулевое распределение:  $N(0, 1)$



## Z-критерий меток для доли

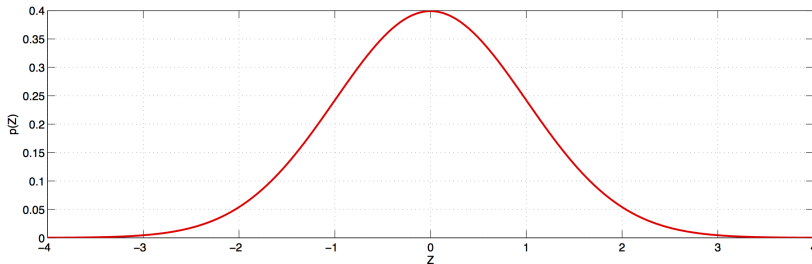
выборка:  $X^n = (X_1, \dots, X_n), X \sim \text{Ber}(p)$

нулевая гипотеза:  $H_0: p \leq p_0$

альтернатива:  $H_1: p > p_0$

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

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



## Z-критерий меток для доли

Как выбрать наименьший достаточный объём выборки?

# Как работают последовательные тесты

Пусть  $\alpha, \beta$  — вероятности ошибки первого и второго рода.

При фиксированном размере выборки  $m$   $\beta$  параметризуется через  $\alpha$  (или наоборот):

$$\beta = \beta(\alpha, m)$$

Что будет, если  $m$  не фиксировано?

Наша задача сводится к отысканию наименьшего  $m$  для нашей выборки, соответствующая наступлению события [ошибка первого рода  $\leq \alpha$ , ошибка второго рода  $\leq \beta$ ].

## Оперативная характеристика

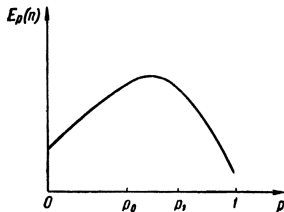
Оперативная характеристика  $L(w)$  — вероятность принять нулевую гипотезу при условии, что  $w$  — истинное значение параметра.

Пусть наш последовательный критерий устроен так: будем считать, что если **первые**  $m_0$  респондентов узнали продукт, то продукт узнаваемый.

$$L(p) = (1 - p)^{m_0}$$

Теперь требуемая мощность выборки  $n$  — это случайная величина, мы можем оценить ожидаемое завершение последовательного критерия:

$$E_p(n) = \sum_{i=1}^{m_0-1} ip(1-p)^{i-1} + m_0(1-p)^{m_0-1}.$$





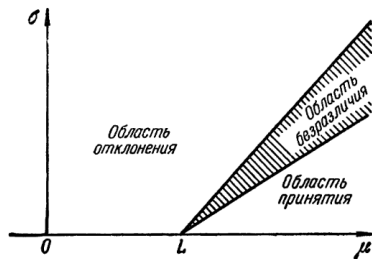
## Последовательный анализ Вальда: ключевая идея

Вместо порога  $p_0$  введем два порога:  $p_L, p_U$ . Будем полагать, что отклонение значения параметра  $p$  от  $p_0$  несущественно, если:

$$p_L \leq p_0 \leq p_U.$$

### Алгоритм

- ❶ Если  $\frac{p_U}{p_L}(x_1, \dots, x_m) \geq A$  для очередной выборки  $x_1, \dots, x_m$ : отклонить гипотезу  $H_0$ .
- ❷ Если  $\frac{p_U}{p_L}(x_1, \dots, x_m) \leq B$  для очередной выборки  $x_1, \dots, x_m$ : принять гипотезу  $H_0$ .
- ❸ Если  $A < \frac{p_U}{p_L}(x_1, \dots, x_m) < B$ , пополнить выборку:  $m := m + 1$ .



## Последовательный анализ Вальда: ключевая идея

Пусть:

$\alpha$  — уровень значимости — допускаемая вероятность ошибки первого рода,

$\beta$  — допускаемая вероятность ошибки второго рода.

Вероятность получения такой выборки  $x_1, \dots, x_n$ , что

$$B < \frac{p_U}{p_L} < A \text{ (для меньших подвыборок), } \frac{p_U}{p_L} \geq A$$

при выполнении гипотезы  $H_1(p = p_U)$  больше в  $A$  раз, чем при выполнении альтернативы  $H_0(p = p_L)$ .

Но эта вероятность равна  $\alpha$  при верности  $H_0$  и  $1 - \beta$  при верности  $H_1$ :

$$1 - \beta \geq \alpha A.$$

## Постановка задачи последовательного анализа

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

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

$$p_L \leq p_0 \leq p_U$$

(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.

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:

$\alpha$  — уровень значимости — допускаемая вероятность ошибки первого рода,

$\beta$  — допускаемая вероятность ошибки второго рода.

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

Введём следующие обозначения:

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

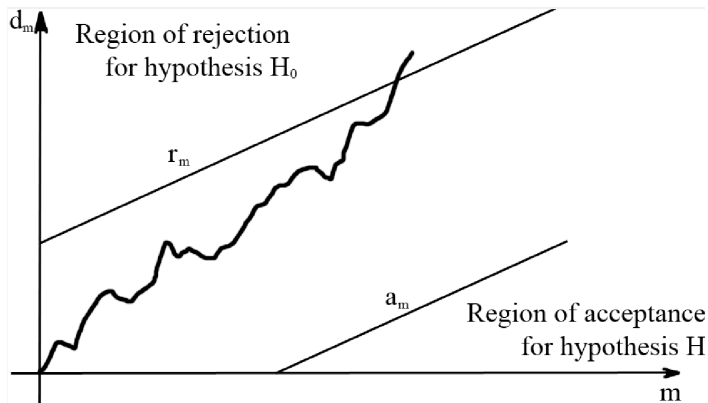
$$a_m = \frac{\ln B + m \ln \frac{1-p_L}{1-p_U}}{\ln \frac{p_U}{p_L} - \ln \frac{1-p_U}{1-p_L}},$$

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

# 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 элементе выборки  $n$  произойдёт остановка процедуры?

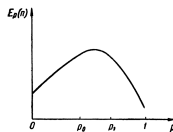
$n$  — случайная величина, можно говорить о её матожидании:

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

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

$h$  определяется как решение уравнения:

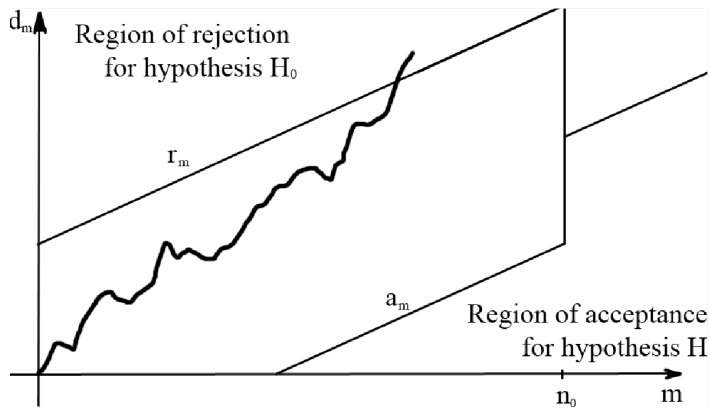
$$p = \frac{1 - \left(\frac{1-p_U}{1-p_L}\right)^h}{\left(\frac{p_U}{p_L}\right)^h - \left(\frac{1-p_U}{1-p_L}\right)^h}.$$



## Усечение

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

- $d_m \geq \frac{a_{n_0} + r_{n_0}}{2} \Rightarrow$  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  $g_1, g_2, \dots$  по  $v$  элементов. Тогда значения статистики  $d_m$  сравниваются с  $a_m, r_m$  только при  $m = v, 2v, \dots$ .

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  $\alpha$  и  $\beta$  обычно малы, отклонением можно пренебречь.



## Z-критерий для разности двух долей, связанные выборки

выборки:  $X_1^n = (X_{11}, \dots, X_{1n}), X_1 \sim \text{Ber}(p_1)$

$X_2^n = (X_{21}, \dots, X_{2n}), X_2 \sim \text{Ber}(p_2)$

выборки связанные

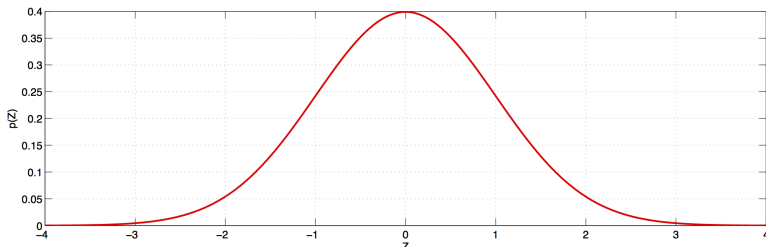
$X_1^n X_2^n$	1	0
1	$e$	$f$
0	$g$	$h$

нулевая гипотеза:  $H_0: p_1 \geq p_2$

альтернатива:  $H_1: p_1 < p_2$

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:

$$\bullet u = 1 \Leftrightarrow p_1 = p_2,$$

$$\bullet u > 1 \Leftrightarrow p_1 > p_2,$$

$$\bullet u < 1 \Leftrightarrow p_1 < p_2.$$

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:  $H_0: u \geqslant u_U$

альтернатива:  $H_1: u \leqslant u_L$

статистика:  $d_m(X_1^m, X_2^m) = \sum_{i=1}^m (1 - X_{1i}) X_{2i}$

## 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:

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

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

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

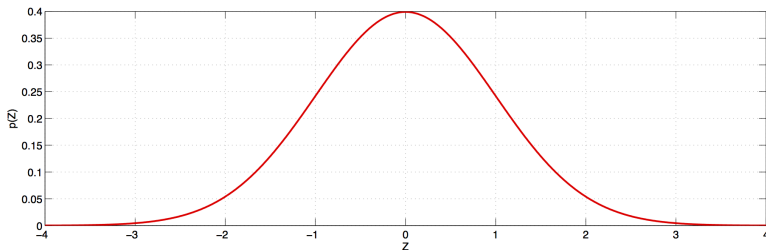
Пусть  $v_1(g_i)$  — число успехов в выборке из  $v$  наблюдений над первой биномиальной совокупностью в группе  $g_i$ ,  $v_2(g_i)$  — над второй. Тогда для этой пары групп в качестве оценки числа пар  $(0, 1)$  примем величину  $v_2(g_i) - \frac{v_1(g_i)v_2(g_i)}{v}$ .

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

Последствия: аналогичные.

## Z-критерий для среднего нормального распределения, односторонняя альтернатива

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



## Z-критерий для среднего нормального распределения, односторонняя альтернатива

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

## Аналог в последовательном анализе

Фиксируем «коридор» отклонений  $\mu$  от  $\mu_0$ , которые можно считать незначимыми:

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

(хотя бы одно из неравенств — строгое).

нулевая гипотеза:  $H_0: \mu \leq \mu_L$

альтернатива:  $H_1: \mu \geq \mu_U$

статистика:  $d_m(X^m) = \sum_{i=1}^m X_i$



## 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:

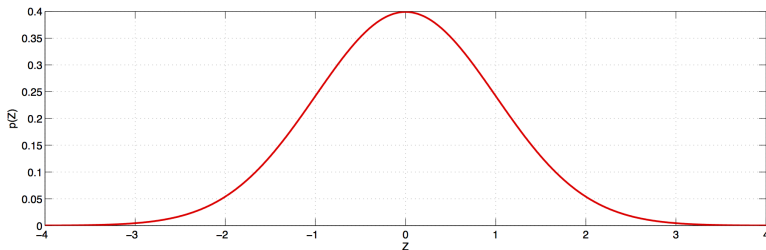
$$\mathbb{E}_\mu(n) = \frac{L(\mu) \ln B (1 - L(\mu)) \ln A}{\mu_L^2 - \mu_U^2 + 2(\mu_U - \mu_L)\mu},$$

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

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

## Z-критерий для среднего нормального распределения, двусторонняя альтернатива

выборка:  $X^n = (X_1, \dots, X_n), X \sim N(\mu, \sigma^2), \sigma$  известна  
нулевая гипотеза:  $H_0: \mu = \mu_0$   
альтернатива:  $H_1: \mu \neq \mu_0$   
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  $\mu$  от  $\mu_0$ , которые можно считать незначимыми:

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

нулевая гипотеза:  $H_0: \left| \frac{\mu - \mu_0}{\sigma} \right| \leq \delta$

альтернатива:  $H_1: \left| \frac{\mu - \mu_0}{\sigma} \right| > \delta$

статистика:  $d_m(X^m) = \ln \operatorname{ch} \left( \frac{\delta}{\sigma} \sum_{i=1}^m (X_i - \mu_0) \right)$

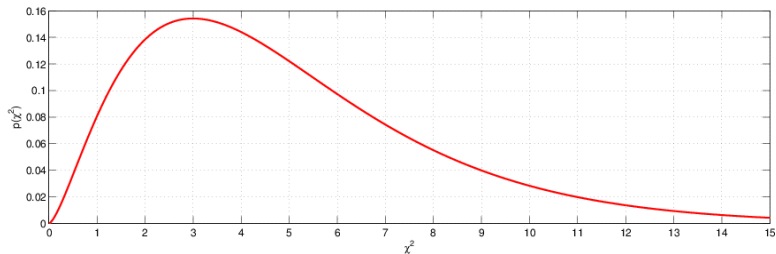
Константы последовательного анализа:

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

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

# Критерий хи-квадрат

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



## Критерий хи-квадрат

**Пример:** не превышает ли погрешность прибора заявленного уровня?

## Аналог в последовательном анализе

Фиксируем «коридор» отклонений  $\sigma$  от  $\sigma_0$ , которые можно считать незначимыми:

$$\sigma_L \leq \sigma_0 \leq \sigma_U$$

(хотя бы одно из неравенств — строгое).

нулевая гипотеза:  $H_0: \sigma \leq \sigma_L$

альтернатива:  $H_1: \sigma \geq \sigma_U$

статистика:  $d_m(X^m) = \sum_{i=1}^m (X_i - \mu)^2$

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:

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

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



## Доверительный интервал для среднего

Дано:  $X_1, \dots, X_n$ ,  $X \sim N(\mu, \sigma^2)$ ,  $\mu, \sigma$  неизвестны.

Требуется: построить доверительный интервал  $J$  для среднего  $\mu$  фиксированной длины  $2d$ :

$$P(\mu \in J) \geq 1 - \alpha \quad \forall \mu, \sigma.$$

При фиксированном  $n$  и неизвестном  $\sigma$  решения не существует (Данциг, 1940).

## Последовательные доверительные интервалы для среднего

При известном  $\sigma$ :

$$J_n = [\bar{X}_n - d, \bar{X}_n + d],$$

$$P(\mu \in J_n) = P(|\bar{X}_n - \mu| \leq d) = P\left(\frac{\sqrt{n}|\bar{X}_n - \mu|}{\sigma} \leq \frac{\sqrt{n}d}{\sigma}\right) = 2\Phi(\sqrt{n}d/\sigma) - 1;$$

$$2\Phi(\sqrt{n}d/\sigma) - 1 \geq 1 - \alpha = 2\Phi(z_{1-\alpha/2}) - 1;$$

так как  $\Phi$  монотонна,

$$\frac{\sqrt{n}d}{\sigma} \geq z_{1-\alpha/2} \Rightarrow n \geq \frac{z_{1-\alpha/2}^2 \sigma^2}{d^2} \equiv C.$$

$C$  — минимальный размер выборки, при котором  $J_n$  имеет уровень доверия  $1 - \alpha$ .

## Последовательные доверительные интервалы для среднего

### Двухэтапная процедура Стейна.

- ①  $X_1, \dots, X_m$  — пилотная выборка,  $m \geq 2$ ,  $S_m^2$  — её выборочная дисперсия.
- ② Определим, сколько нужно добавить наблюдений:

$$\hat{C} = \frac{t_{1-\alpha/2, m-1}^2 S_m^2}{d^2},$$
$$N = \max \left( \left[ \hat{C} \right] + 1, m \right),$$

$J_N = [\bar{X}_N - d, \bar{X}_N + d]$  — искомый  $100(1 - \alpha)\%$  доверительный интервал для  $\mu$ .

$$\frac{t_{1-\alpha/2, m-1}^2}{z_{1-\alpha/2}^2} C \leq \mathbb{E}_{\mu, \sigma} (N) \leq \frac{t_{1-\alpha/2, m-1}^2}{z_{1-\alpha/2}^2} C + m.$$

# Последовательные доверительные интервалы для среднего

Двухэтапная процедура состоятельна:

$$P_{\mu,\sigma}(\mu \in J_N) \geq 1 - \alpha,$$

и асимптотически состоятельна:

$$\lim_{d \rightarrow 0} P_{\mu,\sigma}(\mu \in J_N) = 1 - \alpha,$$

но асимптотически неэффективна:

$$\lim_{d \rightarrow 0} \mathbb{E}_{\mu,\sigma} \left( \frac{N}{C} \right) = \frac{t_{1-\alpha/2, m-1}^2}{z_{1-\alpha/2}^2} > 1.$$

# Последовательные доверительные интервалы для среднего

## Полностью последовательная процедура.

- $X_1, \dots, X_m$  — пилотная выборка,  $m \geq 2$ .
- Для каждого  $n = m, m+1, \dots$  вычисляем

$$\hat{C}_n = \frac{z_{1-\alpha/2}^2 S_n^2}{d^2}.$$

- Продолжаем набирать выборку, если  $n < \hat{C}_n$ .

$N$  — наименьшее целое  $n \geq \hat{C}_n$ .

$$\mathbb{E}_{\mu, \sigma}(N) \leq C + m.$$

## Последовательные доверительные интервалы для среднего

Полностью последовательная процедура только асимптотически состоятельна:

$$\lim_{d \rightarrow 0} P_{\mu, \sigma} (\mu \in J_N) = 1 - \alpha,$$

но зато асимптотически эффективна:

$$\lim_{d \rightarrow 0} \mathbb{E}_{\mu, \sigma} \left( \frac{N}{C} \right) = 1.$$

## Доверительный интервал для разности двух средних

Дано:  $X_{i1}, \dots, X_{in_i}, X_i \sim N(\mu_i, \sigma_i^2), i = 1, 2$ .

Требуется: построить доверительный интервал  $J$  для разности средних  $\mu_1 - \mu_2$  фиксированной длины  $2d$ :

$$P(\mu_1 - \mu_2 \in J) \geq 1 - \alpha \quad \forall \mu_1, \mu_2, \sigma_1, \sigma_2.$$

Введём следующие обозначения:

$$\begin{aligned} n &= (n_1, n_2), \\ T_n &= \bar{X}_{1n_1} - \bar{X}_{2n_2}, \\ U_n^2 &= \frac{(n_1 - 1) S_{1n_1}^2 + (n_2 - 1) S_{2n_2}^2}{n_1 + n_2 - 2}. \end{aligned}$$

Будем рассматривать доверительные интервалы  $J_n = [T_n - d, T_n + d]$ .

**Случай**  $\sigma_1 = \sigma_2 = \sigma$

Поскольку дисперсии равны, будем брать выборки одинакового размера:

$$n_1 = n_2 = n,$$
$$U_n^2 = \frac{S_{1n}^2 + S_{2n}^2}{2}.$$

**Двухэтапная процедура.**

- ①  $X_{i1}, \dots, X_{im}$  — пилотные выборки,  $m \geq 2$ ,  $U_m^2$  — оценка дисперсии по ним.
- ② Определим, сколько нужно добавить пар наблюдений:

$$\hat{C} = \frac{2t_{1-\alpha/2, 2m-2}^2 U_m^2}{d^2},$$
$$N = \max \left( \left[ \hat{C} \right] + 1, m \right),$$

$J_N = [\bar{T}_N - d, \bar{T}_N + d]$  — искомый  $100(1 - \alpha)\%$  доверительный интервал для  $\mu_1 - \mu_2$ .



Случай  $\sigma_1 = \sigma_2 = \sigma$

$$\frac{t_{1-\alpha/2, 2m-2}^2}{z_{1-\alpha/2}^2} C \leq \mathbb{E}_{\mu_1, \mu_2, \sigma} (N) \leq \frac{t_{1-\alpha/2, 2m-2}^2}{z_{1-\alpha/2}^2} C + m.$$

Двухэтапная процедура состоятельна и асимптотически состоятельна, но асимптотически неэффективна (по сравнению с  $C = \frac{2z_{1-\alpha/2}^2 \sigma^2}{d^2}$ ).

## Случай $\sigma_1 \neq \sigma_2$

Пусть  $W_1, W_2 \sim St(m-1)$  независимы,  $h_{m,1-\alpha/2}$  —  $(1-\alpha/2)$ -квантиль распределения  $W_1 - W_2$ :

$$P(W_1 - W_2 \leq h_{m,1-\alpha/2}) = 1 - \frac{\alpha}{2}.$$

$$h_{m,1-\alpha/2} \approx \sqrt{2}z_{1-\alpha/2}.$$

### Двухэтапная процедура.

- ①  $X_{i1}, \dots, X_{im}$  — пилотные выборки,  $m \geq 2$ ,  $S_{1m}^2, S_{2m}^2$  — оценки дисперсий по ним;
- ② Определим, сколько нужно добавить наблюдений в каждую выборку:

$$\hat{C}_1 = \frac{h_{m,1-\alpha/2} S_{1m}^2}{d^2}, \quad \hat{C}_2 = \frac{h_{m,1-\alpha/2} S_{2m}^2}{d^2},$$
$$N_1 = \max\left(\left[\hat{C}_1\right] + 1, m\right), \quad N_2 = \max\left(\left[\hat{C}_2\right] + 1, m\right),$$

$J_N = [\bar{T}_N - d, \bar{T}_N + d]$ ,  $N = (N_1, N_2)$  — искомый  $100(1-\alpha)\%$  доверительный интервал для  $\mu_1 - \mu_2$ .

Случай  $\sigma_1 \neq \sigma_2$

$$\mathbb{E}N_i \approx \frac{h_{m,1-\alpha/2}^2 \sigma_i^2}{d^2}, \quad i = 1, 2.$$

Двухэтапная процедура состоятельна.

В рамках последовательного анализа удалось найти точное решение проблемы Беренца-Фишера!

## Доверительный интервал для матожидания

Дано:  $X_1, X_2, \dots, X \sim F(x)$ ,  $\mathbb{E}X, \mathbb{D}X$  конечны.

Требуется: построить доверительный интервал  $J$  для среднего  $\mathbb{E}X$  фиксированной длины  $2d$ :

$$P(\mathbb{E}X \in J) \geq 1 - \alpha \quad \forall F(x).$$

Если известно  $\mathbb{D}X$ , по центральной предельной теореме приближённым решением является интервал  $J_n = [\bar{X}_n - d, \bar{X}_n + d]$ , где  $n$  — наименьшее целое, удовлетворяющее условию

$$n \geq \frac{z_{1-\alpha/2} (\mathbb{D}X)^2}{d^2} \equiv C.$$

# Последовательные доверительные интервалы для матожидания

## Полностью последовательная процедура.

- $X_1, \dots, X_m$  — пилотная выборка,  $m \geq 2$ ,  $S_m^2$  — оценка дисперсии по ней.
- Для каждого  $n = m, m+1, \dots$  вычисляем

$$\hat{C}_n = \frac{z_{1-\alpha/2}^2 (S_n^2 + \frac{1}{n})}{d^2}.$$

- Продолжаем набирать выборку, если  $n < \hat{C}_n$ .

$\frac{1}{n}$  — поправка на случай, если распределение  $F(x)$  дискретно.

$N$  — наименьшее целое  $n \geq \hat{C}_n$ .

$$\mathbb{E}_F(N) \leq C + m + 2.$$

Процедура асимптотически состоятельна и асимптотически эффективна.

## Доверительный интервал для медианы

Дано:  $X_1, X_2, \dots, X \sim F(x - \theta)$ ,  $\theta = \text{med } X$ ,

- $F(x)$  симметрична относительно нуля,
- $F(x)$  дважды дифференцируема в окрестности нуля  $\mathcal{N}$ ;
- $F''(x)$  ограничена вне  $\mathcal{N}$ .

Требуется: построить доверительный интервал  $J$  для медианы  $\theta$  фиксированной длины  $2d$ :

$$P(\theta \in J) \geq 1 - \alpha \quad \forall F(x).$$

## Доверительный интервал для медианы

При фиксированном  $n$  асимптотический доверительный интервал задаётся порядковыми статистиками:

$$b(n) = \max \left( 1, \left\lceil \frac{1}{2} (n - z_{1-\alpha/2} \sqrt{n} - 1) \right\rceil \right),$$

$$a(n) = n - b(n) + 1,$$

$$J_n = [X_{n:b(n)}, X_{n:a(n)}],$$

$$\lim_{n \rightarrow \infty} P(\theta \in J_n) = 1 - \alpha.$$

Чтобы построить доверительный интервал фиксированной длины  $2d$  с помощью полностью последовательной процедуры, нужно продолжать набирать выборку, пока  $X_{n:a(n)} - X_{n:b(n)} > 2d$ .

# 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.