# Прикладная статистика. Занятие 5. Последовательный анализ Вальда.

13 марта 2012 г.

# Проверка гипотез

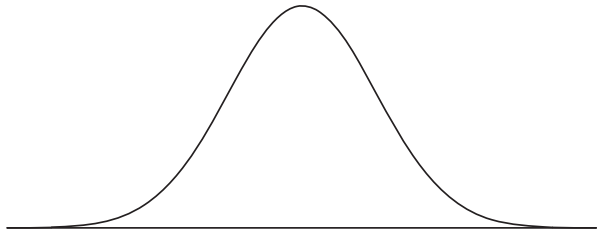
выборка: $\mathbf{X} = \{X_1, \ldots, X_n\} \sim P \in \Omega;$

нулевая гипотеза: $H_0 \colon P \in \omega, \ \ \omega \in \Omega;$

альтернатива: $H_1 \colon P \notin \omega;$

статистика: $T(\mathbf{X}), \ \ T(\mathbf{X}) \sim F(x)$ при $P \in \omega;$

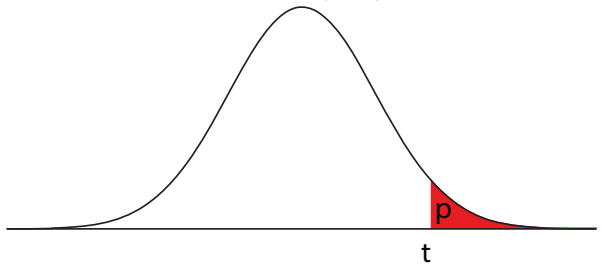
$T(\mathbf{X}) \not\sim F(x)$ при $P \notin \omega;$



реализация выборки: $\mathbf{x} = \{x_1, \ldots, x_n\};$

реализация статистики: $t = T(\mathbf{x});$

достигаемый уровень значимости: $p(\mathbf{x})$— вероятность при $H_0$ получить $T(\mathbf{X}) = t$ или ещё более экстремальное;




Гипотеза отвергается при $p(\mathbf{x}) \leq \alpha, \ \ \alpha$— уровень значимости.

# Односторонняя альтернатива

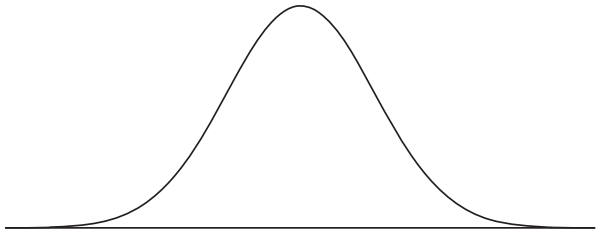
выборка: $\mathbf{X} = \{X_1, \ldots, X_n\} \sim N\left(\mu, \sigma^2\right)$;

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

альтернатива: $H_1 : \mu > \mu_0$;

статистика: $T\left(\mathbf{X}\right) = \frac{\sqrt{n}\left(\bar{X} - \mu_0\right)}{s}$;

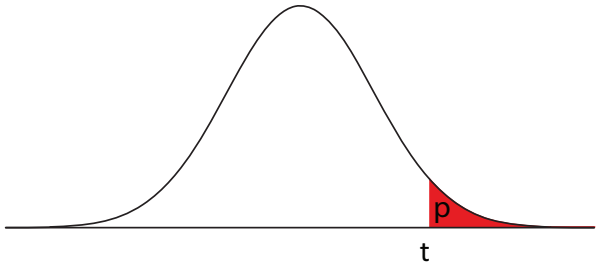
$T\left(\mathbf{X}\right) \sim t_{n-1}$ при $H_0$;



реализация выборки: $\mathbf{x} = \{x_1, \ldots, x_n\}$;

реализация статистики: $t = T\left(\mathbf{x}\right)$;

достигаемый уровень значимости: $p\left(\mathbf{x}\right) = 1 - tcdf\left(t, n-1\right)$;




Гипотеза отвергается при $p\left(\mathbf{x}\right) \leq \alpha$, $\alpha$— уровень значимости.

# Сложная гипотеза

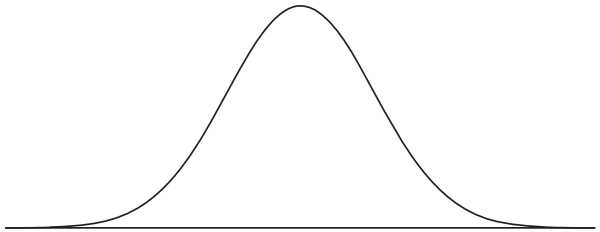
выборка: $\mathbf{X} = \{X_1, \ldots, X_n\} \sim N\left(\mu, \sigma^2\right);$

нулевая гипотеза: $H_0 : \mu \leq \mu_0;$

альтернатива: $H_1 : \mu > \mu_0;$

статистика: $T\left(\mathbf{X}\right) = \frac{\sqrt{n}\left(\bar{X} - \mu_0\right)}{s};$

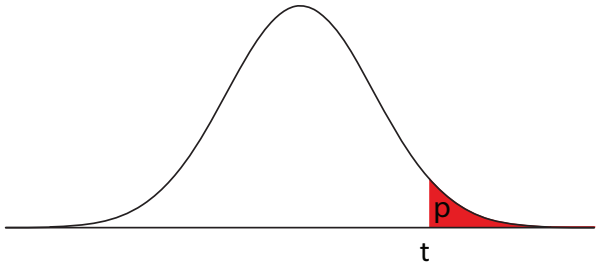
$T\left(\mathbf{X}\right) \sim t_{n-1}$ при $\mu = \mu_0;$



реализация выборки: $\mathbf{x} = \{x_1, \ldots, x_n\};$

реализация статистики: $t = T\left(\mathbf{x}\right);$

достигаемый уровень значимости: $p\left(\mathbf{x}\right) = 1 - tcdf\left(t, n-1\right);$




Гипотеза отвергается при $p\left(\mathbf{x}\right) \leq \alpha$, $\alpha$— уровень значимости.

# Постановка задачи последовательного анализа

выборка: $\mathbf{X} = \{X_1, \ldots, X_m\} \sim N\left(\mu, \sigma^2\right)$
(дисперсия известна).

Фиксируем «коридор» отклонений значений параметра $\mu$ 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: $H_1 : \mu \geq \mu_U$.

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

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

Введём следующие обозначения:

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

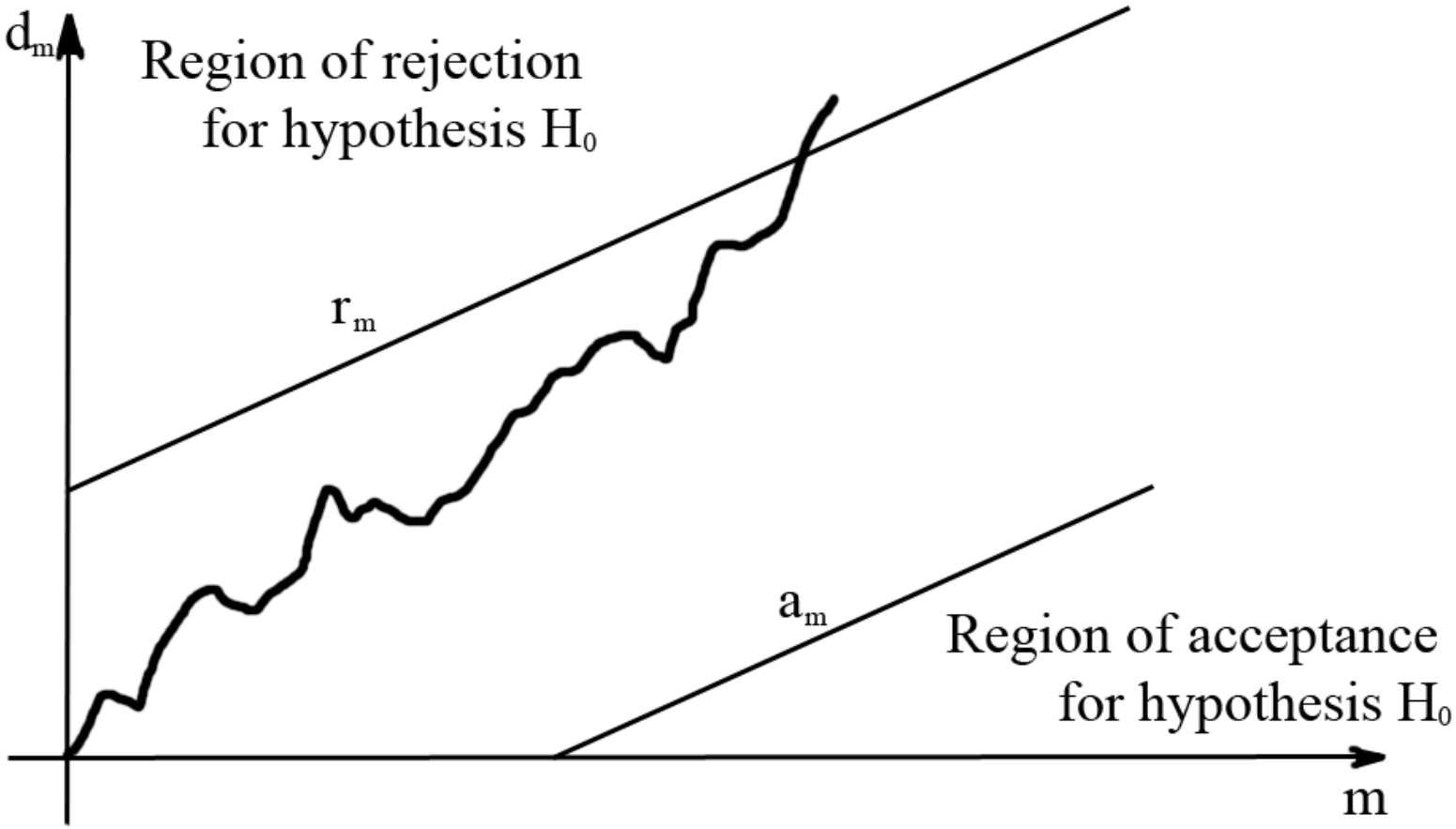
$$a_m = \frac{\sigma^2}{\mu_U - \mu_L} \ln B + m\frac{\mu_L + \mu_U}{2},$$

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

При каждом значении $n$:

- $d_m \geq r_m \Rightarrow$ отвергаем $H_0,\ \mu \geq \mu_U$;
- $d_m \leq a_m \Rightarrow$ принимаем $H_0,\ \mu \leq \mu_L$;
- $a_m < d_m < r_m \Rightarrow$ процесс продолжается, добавляем элемент выборки.

# Процедура последовательного анализа



# Момент остановки

На каком элементы выборки $n$ произойдёт остановка процедуры?

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

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

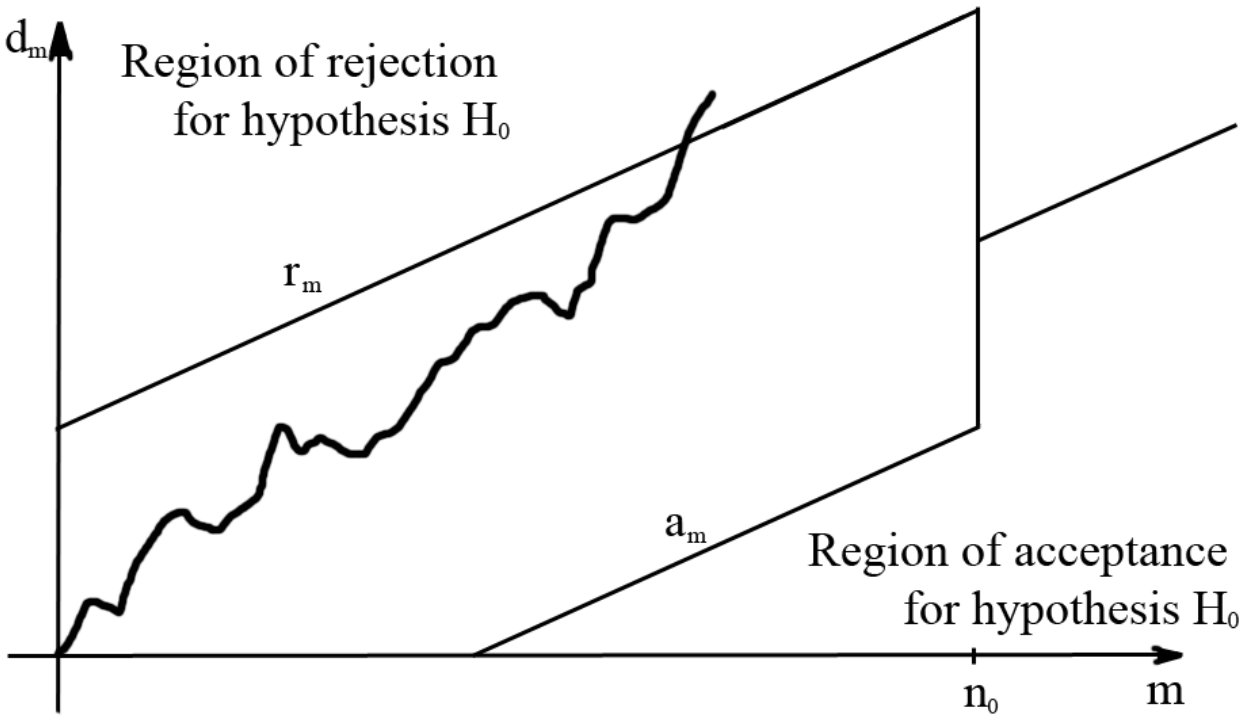
$$L(\mu) = \frac{A^h - 1}{A^h - B^h},$$

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

# Усечение

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

- $d_{n_0} \geq \frac{a_{n_0} + r_{n_0}}{2} \Rightarrow$ отвергаем $H_0,\ \mu \geq \mu_U$;
- $d_{n_0} \leq \frac{a_{n_0} + r_{n_0}}{2} \Rightarrow$ принимаем $H_0,\ \mu \leq \mu_L$.



# Критерий Стьюдента, двусторонняя альтернатива

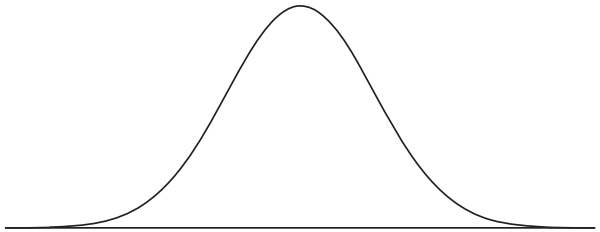
выборка: $\mathbf{X} = \{X_1, \ldots, X_n\} \sim N\left(\mu, \sigma^2\right)$;

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

альтернатива: $H_1 : \mu \neq \mu_0$;

статистика: $T\left(\mathbf{X}\right) = \frac{\sqrt{n}\left(\bar{X} - \mu_0\right)}{s}$,

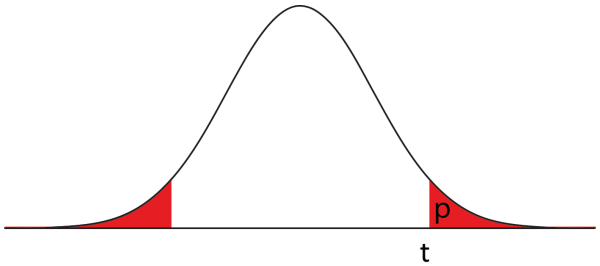
$T\left(\mathbf{X}\right) \sim t_{n-1}$ при $H_0$;



реализация выборки: $\mathbf{x} = \{x_1, \ldots, x_n\}$;

реализация статистики: $t = T\left(\mathbf{x}\right)$;

достигаемый уровень значимости: $p\left(\mathbf{x}\right) = 2\left(1 - tcdf\left(\left|t\right|, n-1\right)\right)$;




Гипотеза отвергается при $p\left(\mathbf{x}\right) \leq \alpha$, $\alpha$— уровень значимости.

# Аналог в последовательном анализе Вальда

выборка: $\mathbf{X} = \{X_1, \ldots, X_m\} \sim N\left(\mu, \sigma^2\right)$
(дисперсия известна);

Фиксируем размер отклонения значений параметра $\mu$ 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: $d_m\left(\mathbf{X}\right) = \ln \operatorname{ch}\left(\frac{\delta}{\sigma}\sum_{i=1}^{m}\left(X_i - \mu_0\right)\right).$

Константы последовательного анализа:

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

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

# Критерий хи-квадрат

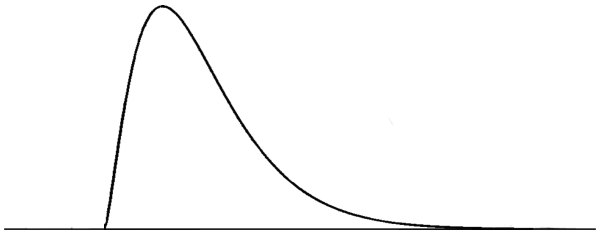
выборка: $\mathbf{X} = \{X_1, \ldots, X_n\} \sim N\left(\mu, \sigma^2\right);$

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

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

статистика: $\chi\left(\mathbf{X}\right) = \frac{(n-1)s^2}{\sigma_0^2},$

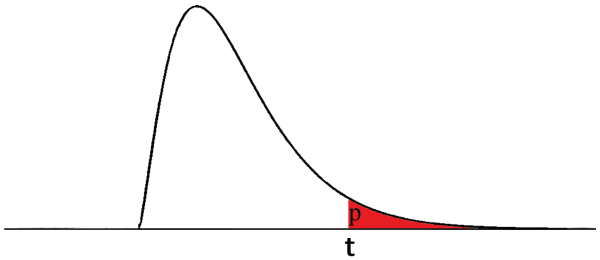
$\chi\left(\mathbf{X}\right) \sim \chi^2_{n-1}$ при $\sigma = \sigma_0;$



реализация выборки: $\mathbf{x} = \{x_1, \ldots, x_n\};$

реализация статистики: $\chi = \chi\left(\mathbf{x}\right);$

достигаемый уровень значимости: $p\left(\mathbf{x}\right) = 1 - chi2cdf\left(\chi, n-1\right);$




Гипотеза отвергается при $p\left(\mathbf{x}\right) \leq \alpha$, $\alpha$— уровень значимости.

# Аналог в последовательном анализе Вальда

выборка: $\mathbf{X} = \{X_1, \ldots, X_m\} \sim N\left(\mu, \sigma^2\right)$
(среднее известно);

Фиксируем «коридор» отклонений значений параметра $\sigma$ 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: $d_m\left(\mathbf{X}\right) = \sum_{i=1}^{m} \left(X_i - \mu\right)^2$.

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:

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

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

# z-критерий

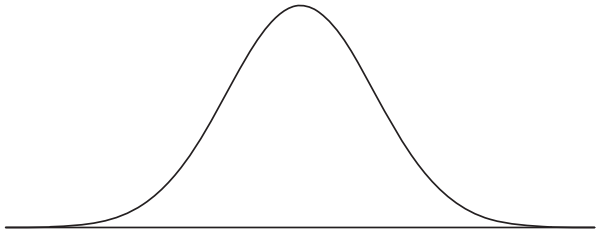
выборка: $\mathbf{X} = \{X_1, \ldots, X_n\} \sim Ber\,(p)\,;$

нулевая гипотеза: $H_0 : p \leq p_0;$

альтернатива: $H_1 : p > p_0;$

статистика: $z\,(\mathbf{X}) = \frac{\sum_{i=1}^{n} X_i - np_0}{\sqrt{np_0(1-p_0)}};$

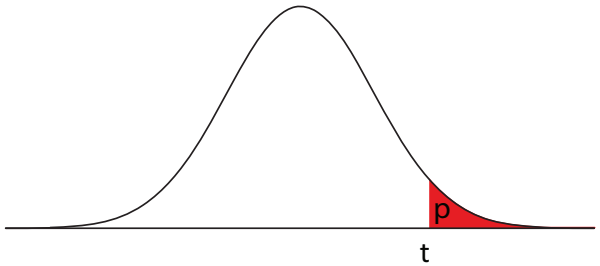
$z\,(\mathbf{X}) \approx N(0,1)$ при $p = p_0;$



реализация выборки: $\mathbf{x} = \{x_1, \ldots, x_n\};$

реализация статистики: $z = z\,(\mathbf{x})\,;$

достигаемый уровень значимости: $p\,(\mathbf{x}) = 1 - normcdf\,(z, 0, 1)\,;$




Гипотеза отвергается при $p\,(\mathbf{x}) \leq \alpha$, $\alpha$— уровень значимости.

# Аналог в последовательном анализе Вальда

выборка: $\mathbf{X} = \{X_1, \ldots, X_m\} \sim Ber\,(p)\,.$

Фиксируем «коридор» отклонений значений параметра $p$ от $p_0$, которые можно считать незначимыми:

$$p_L \leq p_0 \leq p_U$$

(хотя бы одно из неравенств — строгое).

нулевая гипотеза: $H_0 : p \leq p_L;$
альтернатива: $H_1 : p \geq p_U;$
статистика: $d_m\,(\mathbf{X}) = \sum_{i=1}^{m} X_i.$

Константы последовательного анализа:

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

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

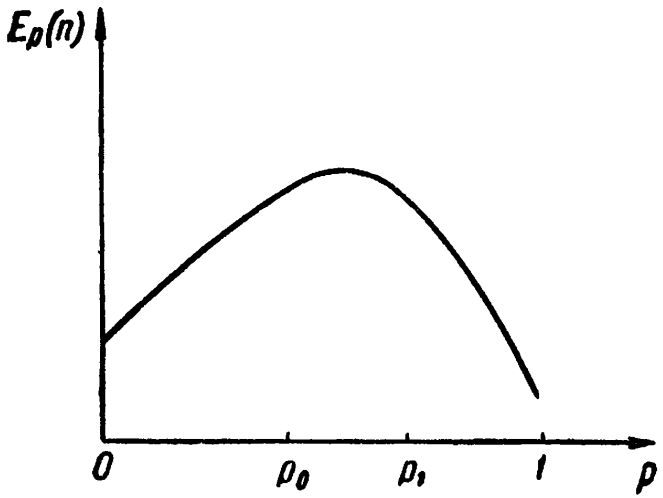
# Момент остановки

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

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

$h$ определяется как решение уравнения:

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



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

## z-критерий

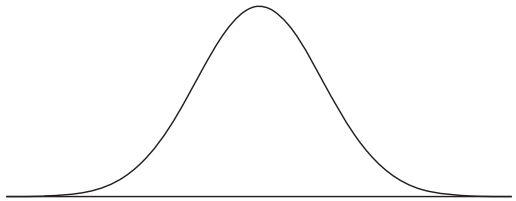
выборки: $\mathbf{X} = \{X_1, \ldots, X_n\} \sim Ber\,(p_1)\,,$
$\mathbf{Y} = \{Y_1, \ldots, Y_n\} \sim Ber\,(p_2)\,;$

нулевая гипотеза: $H_0 : p_1 \geq p_2,$

альтернатива: $H_1 : p_1 < p_2;$

статистика: $$Z\,(\mathbf{X}, \mathbf{Y}) = \frac{\sum_{i=1}^{n}(X_i - Y_i)}{\sqrt{\frac{\sum_{i=1}^{n}(X_i+Y_i)}{2n}\left(n - \frac{\sum_{i=1}^{n}(X_i+Y_i)}{2}\right)}},$$

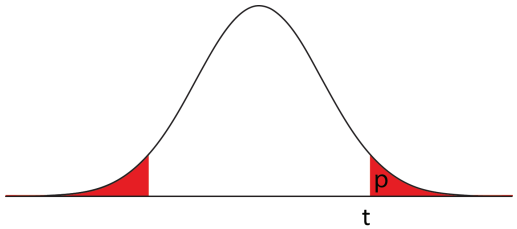
$Z\,(\mathbf{X}, \mathbf{Y}) \approx N(0,1)$ при $H_0;$



реализация выборок: $\mathbf{x} = \{x_1, \ldots, x_n\}, \mathbf{y} = \{y_1, \ldots, y_n\};$

реализация статистики: $z = Z\,(\mathbf{x}, \mathbf{y})\,;$

достигаемый уровень значимости: $p\,(\mathbf{x}, \mathbf{y}) = 1 - normcdf\,(-z, 0, 1)\,;$




Гипотеза отвергается при $p\,(\mathbf{x}, \mathbf{y}) \leq \alpha,$ $\alpha$— уровень значимости.

# Аналог в последовательном анализе Вальда

Пусть значения $x_i, y_i$ поступают парами.
Будем рассматривать только пары (0,1) и (1,0), а остальные будем отбрасывать.

$k_1 = \frac{p_1}{1-p_1}$, $k_2 = \frac{p_2}{1-p_2}$,
$u = \frac{k_2}{k_1} = \frac{p_2(1-p_1)}{p_1(1-p_2)}$ — odds ratio, относительный риск.

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

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

$$u_L \leq 1 \leq u_U$$

(хотя бы одно из неравенств — строгое).

нулевая гипотеза: $H_0 : u \geq u_U$;
альтернатива: $H_1 : u \leq u_L$;
статистика: $d_m = \sum_{i=1}^{m} (1 - X_i)\, Y_i$.

# Аналог в последовательном анализе Вальда

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:

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

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

$h$ 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 $(x_i, y_i)$, 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 $v_2(g_i) - \frac{v_1(g_i)v_2(g_i)}{v}$.

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

Последствия: аналогичные.

# Прикладная статистика
# Лекция 5. Последовательный анализ Вальда.

Рябенко Евгений
riabenko.e@gmail.com