

# Статистический анализ данных. Спецкурс.

## Лекция 5. Проверка статистических гипотез

Ботанический сад-институт ДВО РАН

Кислов Д.Е.  
20 ноября 2016 г.

- Ошибки первого и второго рода. Общая структура статистических критериев;

- Ошибки первого и второго рода. Общая структура статистических критериев;
- Критерии согласия;

- Ошибки первого и второго рода. Общая структура статистических критериев;
- Критерии согласия;
- Проверка данных на соответствие нормальному распределению. Критерий Шапиро-Уилка;

- Ошибки первого и второго рода. Общая структура статистических критериев;
- Критерии согласия;
- Проверка данных на соответствие нормальному распределению. Критерий Шапиро-Уилка;
- Непараметрические критерии. Критерий хи-квадрат. Критерий Колмогорова-Смирнова;

- Ошибки первого и второго рода. Общая структура статистических критериев;
- Критерии согласия;
- Проверка данных на соответствие нормальному распределению. Критерий Шапиро-Уилка;
- Непараметрические критерии. Критерий хи-квадрат. Критерий Колмогорова-Смирнова;
- Дисперсионный анализ;

- Ошибки первого и второго рода. Общая структура статистических критериев;
- Критерии согласия;
- Проверка данных на соответствие нормальному распределению. Критерий Шапиро-Уилка;
- Непараметрические критерии. Критерий хи-квадрат. Критерий Колмогорова-Смирнова;
- Дисперсионный анализ;
- Попарное сравнение в дисперсионном анализе (post-hoc анализ);

## Понятия ошибок первого и второго рода

- Проверяемая гипотеза называется нулевой ( $H_0$ );



## Понятия ошибок первого и второго рода

- Проверяемая гипотеза называется нулевой ( $H_0$ );
- Ошибкой первого рода является вероятность отвергнуть  $H_0$ , когда она верна;

## Понятия ошибок первого и второго рода

- Проверяемая гипотеза называется нулевой ( $H_0$ );
- Ошибкой первого рода является вероятность отвергнуть  $H_0$ , когда она верна;
- Ошибкой второго рода является вероятность принять  $H_0$ , когда верна альтернативная гипотеза;

## Понятия ошибок первого и второго рода

- Проверяемая гипотеза называется нулевой ( $H_0$ );
- Ошибкой первого рода является вероятность отвергнуть  $H_0$ , когда она верна;
- Ошибкой второго рода является вероятность принять  $H_0$ , когда верна альтернативная гипотеза;

## Обозначения

- Ошибка первого рода обозначается  $\alpha$ ;

## Понятия ошибок первого и второго рода

- Проверяемая гипотеза называется нулевой ( $H_0$ );
- Ошибкой первого рода является вероятность отвергнуть  $H_0$ , когда она верна;
- Ошибкой второго рода является вероятность принять  $H_0$ , когда верна альтернативная гипотеза;

## Обозначения

- Ошибка первого рода обозначается  $\alpha$ ;
- Ошибка второго рода обозначается  $\beta$ ;

## Понятия ошибок первого и второго рода

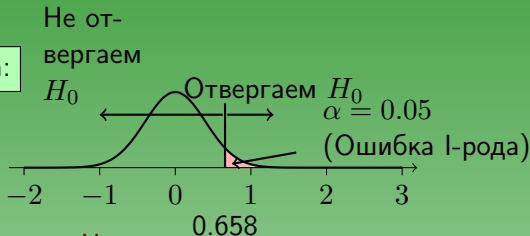
- Проверяемая гипотеза называется нулевой ( $H_0$ );
- Ошибкой первого рода является вероятность отвергнуть  $H_0$ , когда она верна;
- Ошибкой второго рода является вероятность принять  $H_0$ , когда верна альтернативная гипотеза;

## Обозначения

- Ошибка первого рода обозначается  $\alpha$ ;
- Ошибка второго рода обозначается  $\beta$ ;
- $1 - \beta$  мощность критерия (вероятность принять альтернативную гипотезу, когда она верна);

# Ошибки I-го и II-го рода. Мощность критериев.

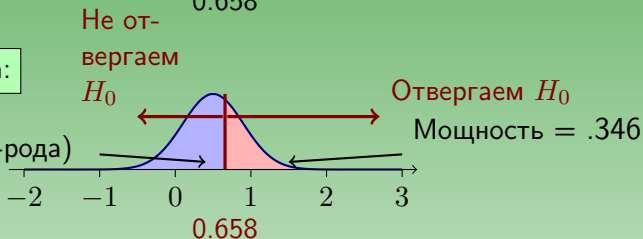
$H_0$  верна:



$H_a$  верна:

$\beta = .654$

(Ошибка II-рода)



## Назначение

Проверка соответствия наблюдаемых данных теоретическим законам распределения

## Назначение

Проверка соответствия наблюдаемых данных теоретическим законам распределения

## Наиболее распространенные критерии согласия

- Критерий  $\chi^2$ ;



## Назначение

Проверка соответствия наблюдаемых данных теоретическим законам распределения

## Наиболее распространенные критерии согласия

- Критерий  $\chi^2$ ;
- Критерий Колмогорова;

## Назначение

Проверка соответствия наблюдаемых данных теоретическим законам распределения

## Наиболее распространенные критерии согласия

- Критерий  $\chi^2$ ;
- Критерий Колмогорова;
- Критерий Шапиро-Уилка;

## Назначение

Проверка соответствия наблюдаемых данных теоретическим законам распределения

## Наиболее распространенные критерии согласия

- Критерий  $\chi^2$ ;
- Критерий Колмогорова;
- Критерий Шапиро-Уилка;
- другие специфичные критерии

# Критерий $\chi^2$ (К. Пирсон, 1900)

## Формулировка критерия

Пусть  $x_1, x_2, \dots, x_N$  – выборочные данные. Проверяется гипотеза о том, что выборка получена из закона распределения  $F(x)$ . Тогда интервал возможных значений разбивается на  $k$  частей, в каждой из которых определяется число выборочных элементов. Далее, вычисляется статистика критерия:

$$\hat{\chi}^2 = N \sum_{i=1}^k \frac{(n_i/N - P_i)^2}{P_i},$$
$$P_i = F(b_{i+1}) - F(b_i)$$

$b_i$  – границы разбиения. Величина  $\hat{\chi}^2$  имеет распределение  $\chi^2$  с  $k - 1$  степенями свободы.

Если вычисленное  $\hat{\chi}^2$  оказывается больше  $\chi^2$  квантиля для уровня  $\alpha$ , то гипотеза о соответствии данному закону распределения отвергается при уровне значимости  $\alpha$ .

## Формулировка критерия

Пусть проверяется соответствие выборочных данных  $x_1, x_2, \dots, x_N$  закону распределения  $F(x)$ . Исходя из выборочных данных формируется выборочная функция распределения  $F_n(x)$

$$F_n(x) = \frac{1}{n} \sum_{i=1}^n I_{X_i \leq x}$$

$$\forall t > 0: \lim_{n \rightarrow \infty} P(\sqrt{n}D_n \leq t) \approx K(t) = \sum_{j=-\infty}^{+\infty} (-1)^j e^{-2j^2 t^2}$$

$$D_n = \sup_x |F_n(x) - F(x)|$$

## Формулировка критерия

Пусть проверяется соответствие выборочных данных  $x_1, x_2, \dots, x_N$  закону распределения  $F(x)$ . Исходя из выборочных данных формируется выборочная функция распределения  $F_n(x)$

$$F_n(x) = \frac{1}{n} \sum_{i=1}^n I_{X_i \leq x}$$

$$\forall t > 0: \lim_{n \rightarrow \infty} P(\sqrt{n} D_n \leq t) \approx K(t) = \sum_{j=-\infty}^{+\infty} (-1)^j e^{-2j^2 t^2}$$

$$D_n = \sup_x |F_n(x) - F(x)|$$

Как быть если параметр распределения нужно оценивать по выборке?

## Формулировка критерия

Пусть проверяется соответствие выборочных данных  $x_1, x_2, \dots, x_N$  закону распределения  $F(x)$ . Исходя из выборочных данных формируется выборочная функция распределения  $F_n(x)$

$$F_n(x) = \frac{1}{n} \sum_{i=1}^n I_{X_i \leq x}$$

$$\forall t > 0: \lim_{n \rightarrow \infty} P(\sqrt{n} D_n \leq t) \approx K(t) = \sum_{j=-\infty}^{+\infty} (-1)^j e^{-2j^2 t^2}$$

$$D_n = \sup_x |F_n(x) - F(x)|$$

Как быть если параметр распределения нужно оценивать по выборке?

Для случая нормального распределения используется поправка Лильефорса...

## Формулировка критерия

Пусть проверяется соответствие выборочных данных  $x_1, x_2, \dots, x_N$  закону распределения  $F(x)$ . Исходя из выборочных данных формируется выборочная функция распределения  $F_n(x)$

$$F_n(x) = \frac{1}{n} \sum_{i=1}^n I_{X_i \leq x}$$

$$\forall t > 0: \lim_{n \rightarrow \infty} P(\sqrt{n}D_n \leq t) \approx K(t) = \sum_{j=-\infty}^{+\infty} (-1)^j e^{-2j^2 t^2}$$

$$D_n = \sup_x |F_n(x) - F(x)|$$

Для проверки однородности двух выборок применяют критерий Смирнова (построенный на базе распределения Колмогорова).



## Формулировка критерия

Статистика критерия имеет вид:

$$W = \frac{1}{s^2} \left[ \sum_{i=1}^n a_{n-i+1} (x_{n-i+1} - x_i) \right]^2$$

$$s^2 = \sum_{i=1}^n (x_i - \bar{x})^2, \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i.$$

Коэффициенты  $a_i$  находятся по таблицам; Критерий фактически основывается на отношениях оценок дисперсий распределения. Данный критерий является одним из самых мощных в плане проверки нормальности распределений.

# Критерий Стьюдента. Предпосылки. Случай известной дисперсии.

## Формулировка задачи

Пусть  $x_1, x_2, \dots, x_n$  набор независимых нормально распределенных случайных величин с известной дисперсией  $\sigma^2$ . Необходимо проверить гипотезу о равенстве среднего значения некоторому числу  $a$ . Нулевая гипотеза – среднее значение распределения есть  $a$ .

## Решение

$$z = \frac{1}{n\sigma} \sum_{i=1}^n (x_i - a)$$

$H_1 = \{z < 0\}$  (или мат. ожидание меньше  $a$ ):  $z < \Phi_\alpha$  – нулевая гипотеза отвергается в пользу  $H_1$  при уровне значимости  $\alpha$ .  
 $\Phi_\alpha$  –  $\alpha$ -квантиль стандартного нормального распределения.

# Критерий Стьюдента. Предпосылки. Случай известной дисперсии.

## Формулировка задачи

Пусть  $x_1, x_2, \dots, x_n$  набор независимых нормально распределенных случайных величин с известной дисперсией  $\sigma^2$ . Необходимо проверить гипотезу о равенстве среднего значения некоторому числу  $a$ . Нулевая гипотеза – среднее значение распределения есть  $a$ .

## Решение

$$z = \frac{1}{n\sigma} \sum_{i=1}^n (x_i - a)$$

$H_2 = \{z > 0\}$  (или мат. ожидание больше  $a$ ):  $z > \Phi_{1-\alpha}$  – нулевая гипотеза отвергается в пользу  $H_2$  при уровне значимости  $\alpha$ .

$\Phi_\alpha$  –  $\alpha$ -квантиль стандартного нормального распределения.

# Критерий Стьюдента. Предпосылки. Случай известной дисперсии.

## Формулировка задачи

Пусть  $x_1, x_2, \dots, x_n$  набор независимых нормально распределенных случайных величин с известной дисперсией  $\sigma^2$ . Необходимо проверить гипотезу о равенстве среднего значения некоторому числу  $a$ . Нулевая гипотеза – среднее значение распределения есть  $a$ .

## Решение

$$z = \frac{1}{n\sigma} \sum_{i=1}^n (x_i - a)$$

$H_3 = \{|z| > 0\}$  (или мат. ожидание не равно  $a$ ):  $z > \Phi_{1-\alpha/2}$  – нулевая гипотеза отвергается в пользу  $H_3$  при уровне значимости  $\alpha$ .

$\Phi_\alpha$  –  $\alpha$ -квантиль стандартного нормального распределения.

# Критерий Стьюдента. Случай неизвестной дисперсии

## Формулировка задачи

Пусть  $x_1, x_2, \dots, x_n$  набор независимых нормально распределенных случайных величин. Необходимо проверить гипотезу о равенстве среднего значения некоторому числу  $a$ . Нулевая гипотеза – среднее распределения равно  $a$ .

## Решение

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - m)^2,$$
$$m = \frac{1}{n} \sum_{i=1}^n x_i, t = \sqrt{n} \frac{m - a}{s}$$

$H_1 = \{t < 0\}$  (или мат. ожидание меньше  $a$ ):  $t < T_\alpha(n-1)$  – нулевая гипотеза отвергается в пользу  $H_1$  при уровне значимости  $\alpha$ .

# Критерий Стьюдента. Случай неизвестной дисперсии

## Формулировка задачи

Пусть  $x_1, x_2, \dots, x_n$  набор независимых нормально распределенных случайных величин. Необходимо проверить гипотезу о равенстве среднего значения некоторому числу  $a$ . Нулевая гипотеза – среднее распределения равно  $a$ .

## Решение

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - m)^2,$$
$$m = \frac{1}{n} \sum_{i=1}^n x_i, t = \sqrt{n} \frac{m - a}{s}$$

$H_2 = \{t > 0\}$  (или мат. ожидание больше  $a$ ):  $t > T_{1-\alpha}(n-1)$  – нулевая гипотеза отвергается в пользу  $H_2$  при уровне значимости  $\alpha$ .

# Критерий Стьюдента. Случай неизвестной дисперсии

## Формулировка задачи

Пусть  $x_1, x_2, \dots, x_n$  набор независимых нормально распределенных случайных величин. Необходимо проверить гипотезу о равенстве среднего значения некоторому числу  $a$ . Нулевая гипотеза – среднее распределения равно  $a$ .

## Решение

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - m)^2,$$
$$m = \frac{1}{n} \sum_{i=1}^n x_i, t = \sqrt{n} \frac{m - a}{s}$$

$H_3 = \{|t| > 0\}$  (или мат. ожидание не равно  $a$ ):  
 $t > T_{1-\alpha/2}(n-1)$  – нулевая гипотеза отвергается в пользу  $H_3$  при уровне значимости  $\alpha$ .

# Критерий Стьюдента. Сравнение двух средних.

## Формулировка задачи

Пусть  $x_1, x_2, \dots, x_n, y_1, y_2, \dots, y_m$  наборы независимых нормально распределенных случайных величин. Дисперсии выборок неизвестны, но равны. Необходимо проверить гипотезу о равенстве математических ожиданий, из которых получены эти выборки.

## Решение

$$t = \frac{m_x - m_y}{s} \left( \frac{m \cdot n}{m + n} \right)^{1/2},$$

$$s^2 = \frac{(n-1)s_x^2 + (m-1)s_y^2}{m + n - 2},$$

$s_x^2, s_y^2$  – несмещенные оценки дисперсий

$H_1 = \{t < 0\}$  ( $m_x < m_y$ ):  $t < T_\alpha(n + m - 2)$  – нулевая гипотеза отвергается в пользу  $H_1$  при уровне значимости  $\alpha$ .



# Критерий Стьюдента. Сравнение двух средних.

## Формулировка задачи

Пусть  $x_1, x_2, \dots, x_n, y_1, y_2, \dots, y_m$  наборы независимых нормально распределенных случайных величин. Дисперсии выборок неизвестны, но равны. Необходимо проверить гипотезу о равенстве математических ожиданий, из которых получены эти выборки.

## Решение

$$t = \frac{m_x - m_y}{s} \left( \frac{m \cdot n}{m + n} \right)^{1/2},$$

$$s^2 = \frac{(n-1)s_x^2 + (m-1)s_y^2}{m+n-2},$$

$s_x^2, s_y^2$  – несмещенные оценки дисперсий

$H_2 = \{t > 0\}$  ( $m_x > m_y$ ):  $t > T_{1-\alpha}(n+m-2)$  – нулевая гипотеза отвергается в пользу  $H_2$  при уровне значимости  $\alpha$ .

# Критерий Стьюдента. Сравнение двух средних.

## Формулировка задачи

Пусть  $x_1, x_2, \dots, x_n, y_1, y_2, \dots, y_m$  наборы независимых нормально распределенных случайных величин. Дисперсии выборок неизвестны, но равны. Необходимо проверить гипотезу о равенстве математических ожиданий, из которых получены эти выборки.

## Решение

$$t = \frac{m_x - m_y}{s} \left( \frac{m \cdot n}{m + n} \right)^{1/2},$$

$$s^2 = \frac{(n-1)s_x^2 + (m-1)s_y^2}{m+n-2},$$

$s_x^2, s_y^2$  – несмещенные оценки дисперсий

$H_3 = \{|t| > 0\}$  ( $m_x \neq m_y$ ):  $t > T_{1-\alpha/2}(n+m-2)$  – нулевая гипотеза отвергается в пользу  $H_3$  при уровне значимости  $\alpha$ .

## Задача о сравнении

Даны два набора измерений длины плодов растений двух видов  $A$  и  $B$ . Значимы ли различия в средних значениях размеров плодов для этих видов.

## Схема решения

- Проверка выборочных данных на соответствие нормальному распределению (Критерий Шапиро-Уилка, Колмогорова);
- Проверка равенства дисперсий (есть ли какие-либо основания, полагать дисперсии равными?);
- Выбор статистического теста (возможно критерий  $t$ -Стьюдента);

