Статистический анализ данных. Спецкурс.

Лекция 5. Проверка статистических гипотез

Ботанический сад-институт ДВО РАН

Кислов Д.Е.

13 ноября 2016 г.

Структура лекции

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

Структура лекции

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

Структура лекции

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

Структура лекции

- 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;

Структура лекции

- 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;

Структура лекции

- 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);

Проверка гипотез

Понятия ошибок первого и второго рода

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

Проверка гипотез

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 α;

Проверка гипотез

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 α;
- Ошбика второго рода обозначается β;

Проверка гипотез

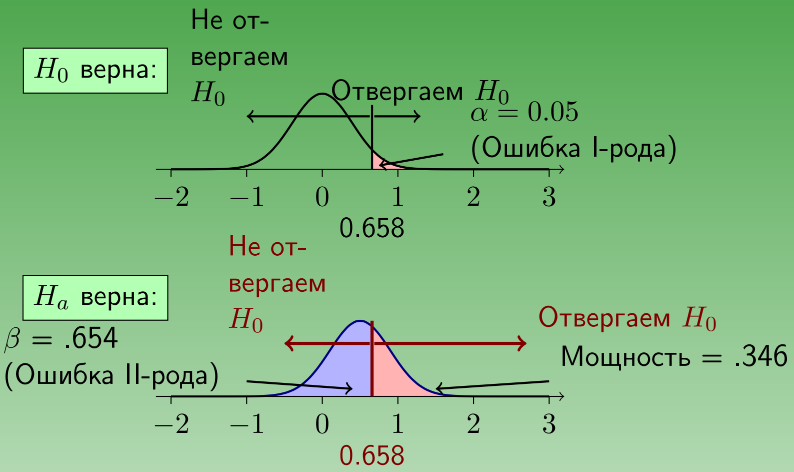
Понятия ошибок первого и второго рода

- 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 α;
- Ошбика второго рода обозначается β;
- $1 - \beta$ мощность критерия (вероятность принять альтернативную гипотезу, когда она верна);

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



Критерии согласия

Назначение

Проверка соответствия наблюдаемых данных теоретическим законам распределения

Критерии согласия

Назначение

Проверка соответствия наблюдаемых данных теоретическим законам распределения

Наиболее распространенные критерии согласия

- Критерий χ^2;

Критерии согласия

Назначение

Проверка соответствия наблюдаемых данных теоретическим законам распределения

Наиболее распространенные критерии согласия

- Критерий χ^2;
- Критерий Колмогорова;

Критерии согласия

Назначение

Проверка соответствия наблюдаемых данных теоретическим законам распределения

Наиболее распространенные критерии согласия

- Критерий χ^2;
- Критерий Колмогорова;
- Критерий Шапиро-Уилка;

Критерии согласия

Назначение

Проверка соответствия наблюдаемых данных теоретическим законам распределения

Наиболее распространенные критерии согласия

- Критерий χ^2;
- Критерий Колмогорова;
- Критерий Шапиро-Уилка;
- другие специфичные критерии

Критерий χ^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 χ^2 с $k-1$ степенями свободы.

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

Критерий Колмогорова

Формулировка критерия

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

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

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

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

Критерий Колмогорова

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 $F_n(x)$

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

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

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

Как быть если параметр распределения нужно оценивать по выборке?

Критерий Колмогорова

Формулировка критерия

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

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

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

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

Как быть если параметр распределения нужно оценивать по выборке?

Для случая нормального распределения используется поправка Лильефорса...

Критерий Колмогорова

Формулировка критерия

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

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

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

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

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

Критерий Шапиро-Уилка

Формулировка критерия

Статистика критерия имеет вид:

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

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

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.