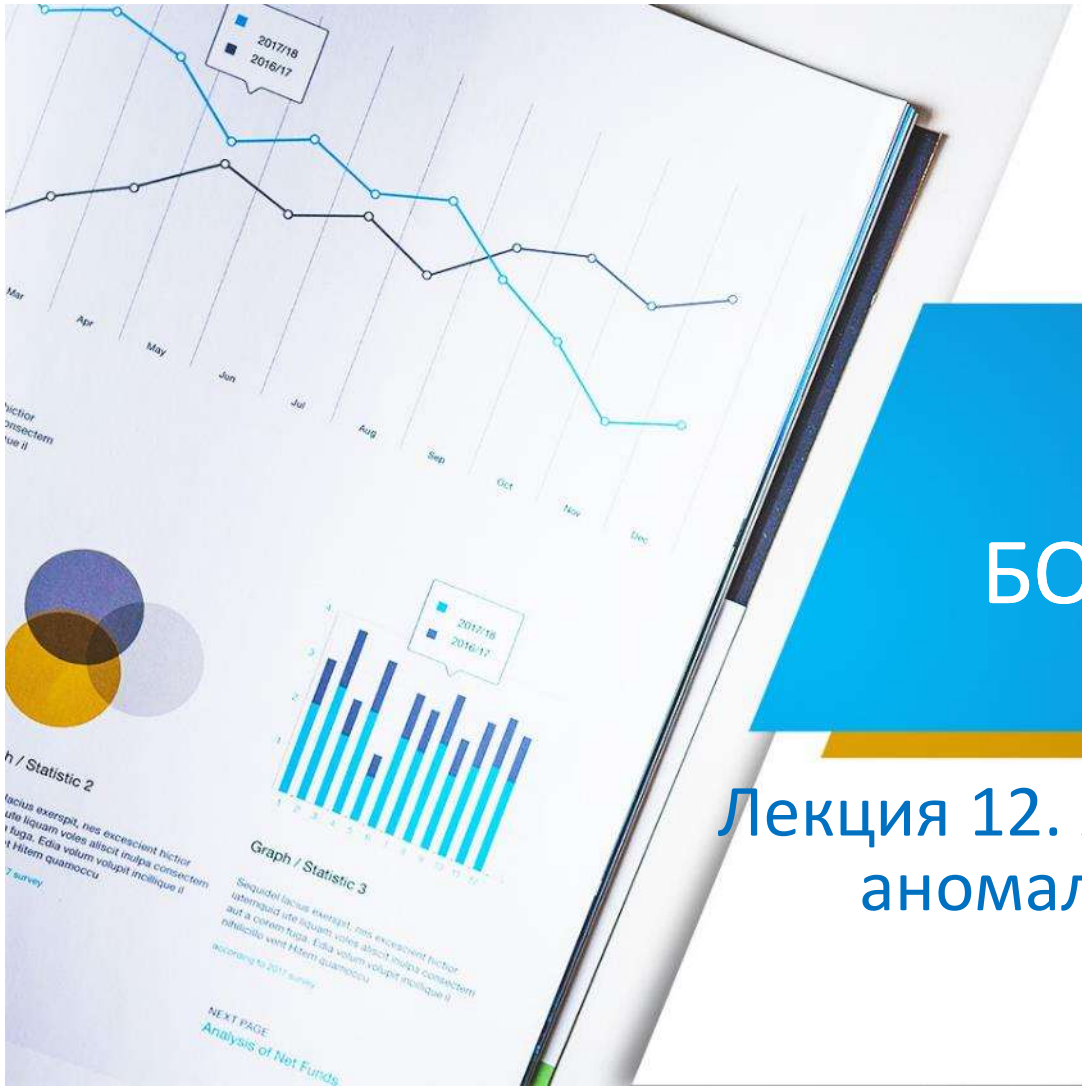


СТАТИСТИКА БОЛЬШИХ ДАННЫХ

Лекция 12. Алгоритмы обнаружения аномальных значений в данных

Введение и мотивация

Алгоритмы обнаружения аномальных значений в данных на основе использования боксплота Тьюки и его модификаций

- Введение и мотивация
- Боксплот Тьюки
- Алгоритмы обнаружения аномалий с использованием боксплота Тьюки и его модификаций.
- Заключение

Введение и мотивация

Задача отбраковки выбросов - это традиционная задача математической статистики (Chauvenet, 1863; Dixon, 1950; Grubbs 1950, 1969; Tukey, 1977; Barnett and Lewis 1978, 1994; Hawkins, 1980; Shevlyakov and Vilchevski 2002, 2011; Manoj and Senthamarai Kennan 2013). Ниже рассматриваются как классические методы отбраковки выбросов, так и сравнительно новые подходы к решению этой задачи на основе алгоритмов разведочного анализа данных Тьюки (Tukey, 1977), а именно боксплота Тьюки и его модификаций.

Аномальное значение (выброс) — это значение, которое выглядит заметно отличающимся от других элементов выборки.

An outlying observation, or outlier, is one that appears to deviate markedly from other members of the sample in which it occurs (Grubbs, 1969; Barnett, 1994; Hawkins, 1980)

Правило 3-х сигм

Рассмотрим простейший классический алгоритм отбраковки подозрительного на выброс значения x элемента выборки по правилу *3-х сигм*:

$$|x - \bar{x}| > 3s,$$

где $\bar{x}$ — выборочное среднее, а s — среднеквадратическое отклонение. При нормальном распределении данных, вероятность отбраковки подозрительного на аномально большое значение по этому правилу составляет примерно 0.003:

$$P(|x - \mu| > 3\sigma) = 0.0027.$$

Тест Граббса

Другой пример дает использование классического критерия (теста) Граббса (Grubbs, 1969) для проверки аномально больших значений элементов выборки

$$\frac{|x_{(1)}-\bar{x}|}{s} \geq \lambda_\alpha \quad \text{или} \quad \frac{x_{(n)}-\bar{x}}{s} \geq \lambda_\alpha,$$

где $x_{(1)},\ldots,x_{(n)}$ — порядковые статистики, λ_α — порог отбраковки, выбираемый из условия заданной ошибки 1-го рода $(\alpha = 0.01, 0.05, 0.1)$ в рамках подхода Неймана-Пирсона. Значения этого порога табулированы для нормального распределения данных и лежат в интервале от 2 до 3.

Подход Неймана-Пирсона

Постановка задачи Неймана-Пирсона обнаружения аномального значения как задача проверки гипотез:

- H_0: наблюдение X регулярное, $X \sim N(0,1)$,
- H_1: X является аномальным значением, $X \sim N(k,\sigma)$, $k > 0, \sigma > 1$.

Пример правила обнаружения аномального значения: H_0: $X < t$; H_1: $X \geq t$.

Мощность (вероятность правильного обнаружения аномального значения):

$$P_D = P(X \geq t | H_1),$$

а вероятность ложной тревоги (ошибки 1-го рода): $P_F = P(X \geq t | H_0)$.

Подход Неймана-Пирсона

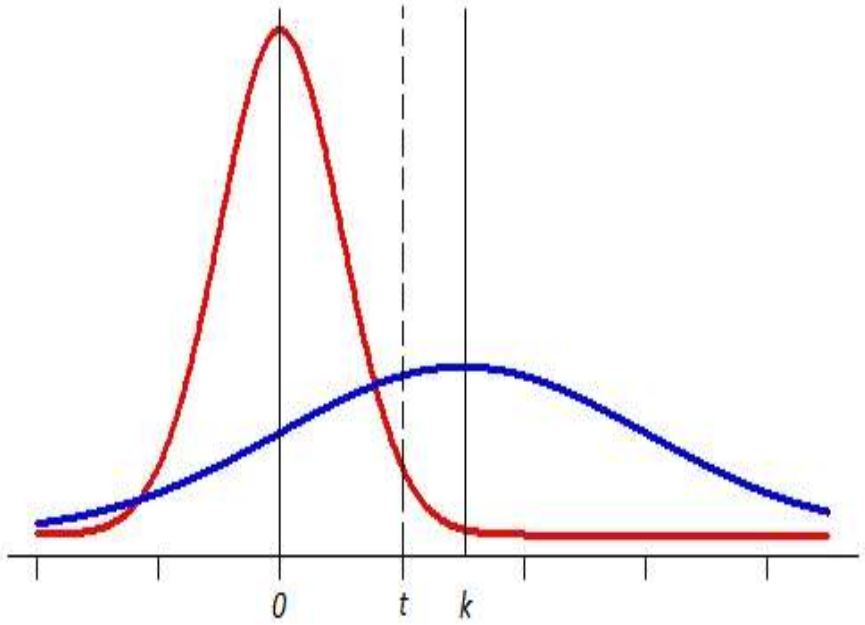
Подход Неймана-Пирсона заключается в максимизации мощности теста при ограничении вероятности ложной тревоги:

$$P_D \to \max_t,$$

$$P_F \leq \alpha,$$

где $\alpha -$ вероятность ложной тревоги (уровень значимости).

$\alpha = 0.001, 0.005, 0.01, 0.05, 0.1.$



Подход Неймана-Пирсона

В общем, в рамках подхода Неймана-Пирсона к проверке гипотез решающее правило имеет следующий вид

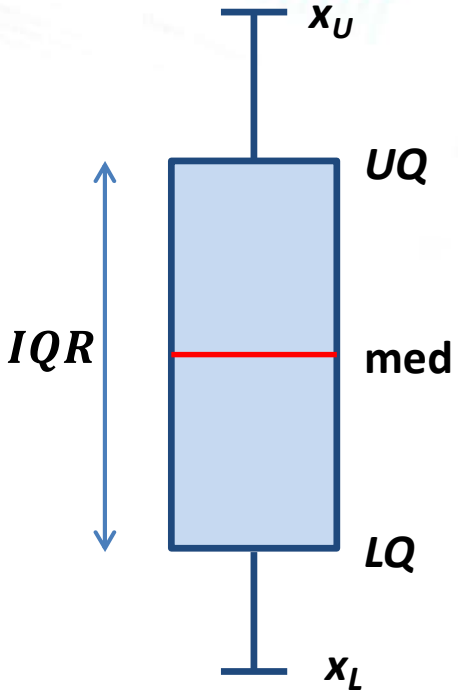
$$t(x_1, \ldots, x_n) \geq \lambda_\alpha,$$

$$P_F = \alpha,$$

где $t(x_1, \ldots, x_n)$ – статистика решающего правила (теста), λ_α – значение порога, определяемого из заданного уровня значимости.

В частности, вышеприведенные правило трех сигм и тест Граббса имеют такой вид.

Боксплот Тьюки



Нижний x_L и верхний x_U пороги отбраковки в боксплоте Тьюки задаются следующим образом:

$$\begin{cases} x_L = \max\{x_{(1)}, LQ - \frac{3}{2} IQR\} \\ x_U = \min\left\{x_{(n)}, UQ + \frac{3}{2} IQR\right\}, \end{cases}$$

где $x_{(1)}$ и $x_{(n)}$ – экстремальные порядковые статистики выборки, $IQR = UQ - LQ$ – выборочная интерквартильная широта.

Правило отбраковки: значение x_i является аномальным, если $x_i > x_U$ или $x_i < x_L$. Для нормального распределения вероятность такого события приближенно равна 0.007

Боксплот Тьюки

Робастные модификации боксплота Тьюки (Andrea and Shevlyakov, 2011): "Усы" модификации боксплота Тьюки x_L и x_U имеют следующий общий вид:

$$\begin{cases} x_L = \max\{x_{(1)}, LQ - k_S\, S\} \\ x_U = \min\{x_{(n)}, UQ + k_S\, S\}, \end{cases}$$

где S – робастная оценка масштаба и k_S коэффициент порога.

Рассмотрим следующие **робастные модификации боксплота Тьюки**: $\boldsymbol{MAD}$**-боксплот и** $\boldsymbol{FQ}$**-боксплот.** В них в качестве оценки параметра масштаба *S* (меры рассеяния) вместо интерквартильной широты *IQR* используются медиана абсолютных отклонений MAD_n и FQ-оценка. Значение коэффициента k_S будет определено ниже.

Боксплот Тьюки

Медиана абсолютных отклонений от медианы MAD_n (Hampel, 1974) определяется как:

$$MAD_n = \text{med}\,\{|x - \text{med}\,x|\}.$$

Q_n-оценка, предложенная в (Rousseeuw and Croux, 1993), представляет собой нижний квартиль распределения абсолютного значения разности случайных величин:

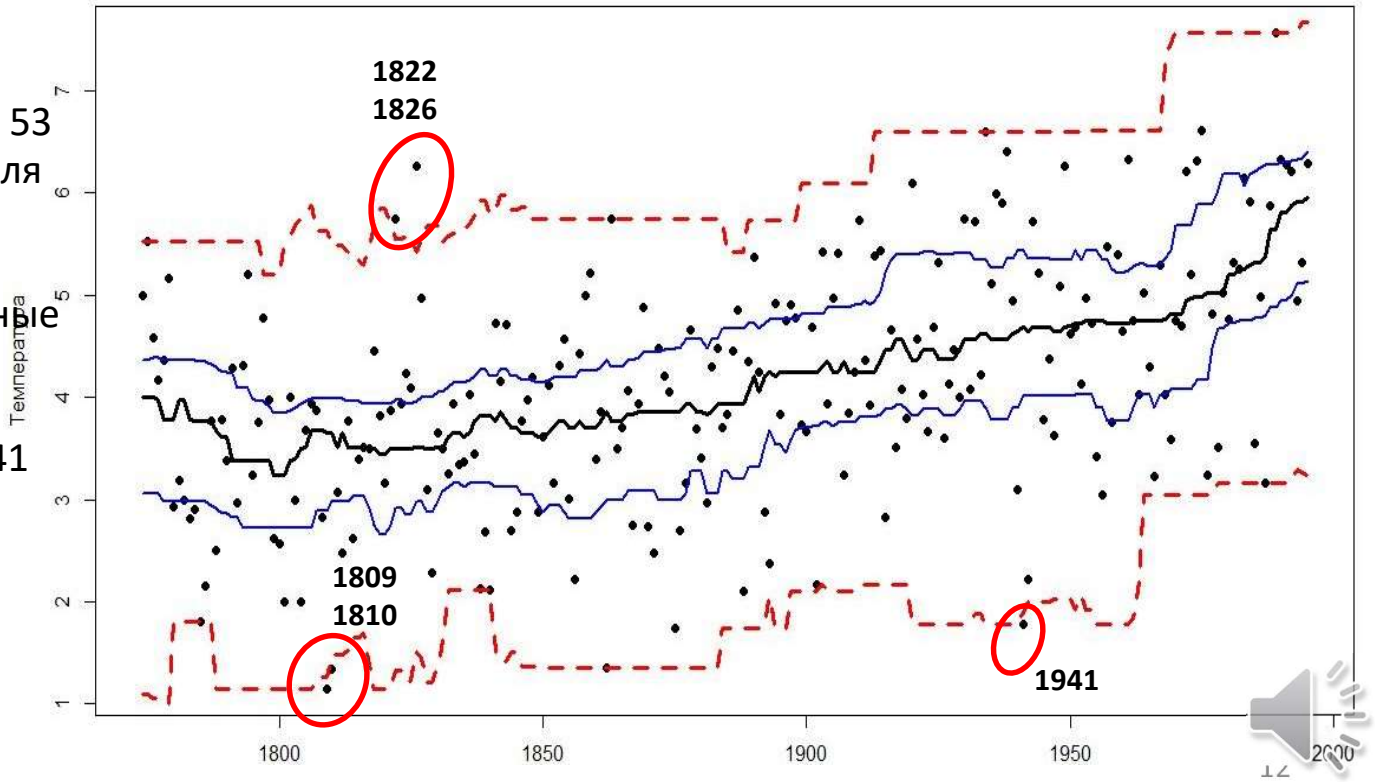
$$Q_n = c_n LQ\,\{|x_i - x_j|, \qquad i < j\},$$

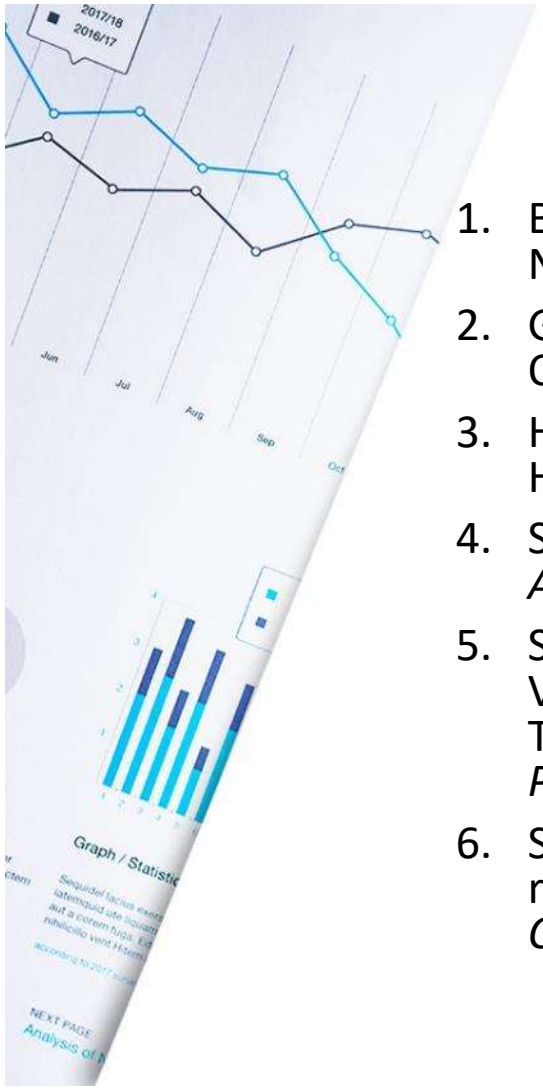
где c_n – коэффициент, обеспечивающий несмещенность оценки.

Для MAD_n оптимальное значение коэффициента порога k_S равно $k_{MAD} = 1.44$; для Q_nоценки оптимальное значение коэффициента порога k_S равно $k_Q = 0.97$.

Пример

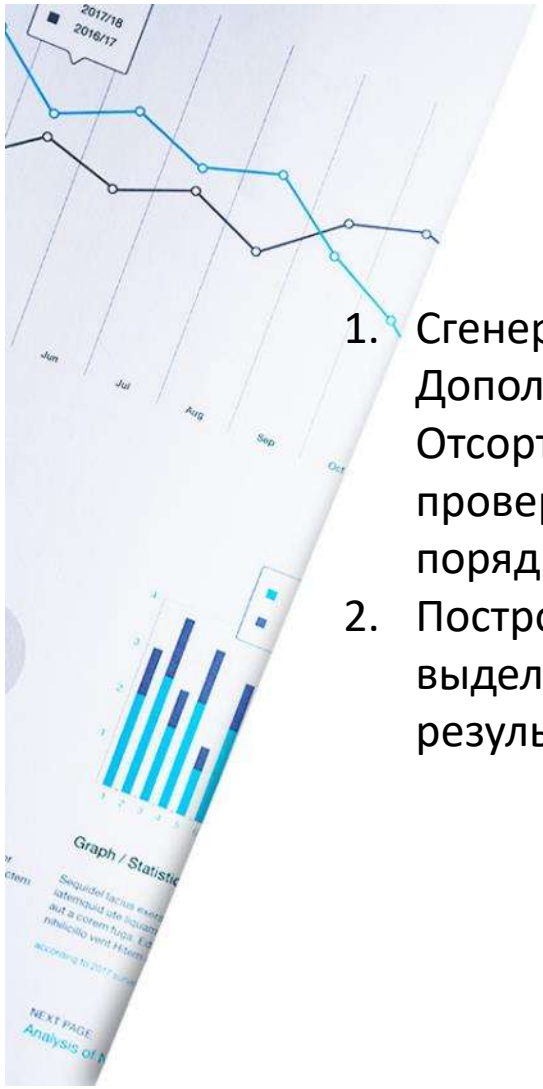
Скользящий боксплот Тьюки (Ширина окна - 53 точки) построенный для ряда среднегодовых температур Санкт-Петербурга. Аномальные значения температур наблюдались в 1809, 1810, 1822, 1826 и 1941 годах





Литература

1. Barnett V. and Lewis, T. (1994). *Outliers in Statistical Data.* Wiley, New York.
2. Grubbs F. E. (1969). Procedures for Detecting Outlying Observations in Samples. *Technometrics,* 11, 1-21.
3. Hawkins D. M. (1980). *Identification of Outliers.* Chapman and Hall, London.
4. Shevlyakov G. and Vilchevski N. (2011). *Robustness in Data Analysis: criteria and methods.* De Gruyter, Boston.
5. Shevlyakov G., Andrea K., Choudur L., Smirnov P., Ulanov A. and Vassilieva N. (2013). Robust Versions of the Tukey Boxplot with Their Application to Detection of Outliers. *ICASSP 2013 Proceedings,* Vancouver.
6. Smirnov P. and Shevlyakov G. (2014). Fast highly efficient and robust one-step M -estimators of scale based on Q_n . *Computational Statistics and Data Analysis.*



Контрольные вопросы и задания

1. Сгенерировать модельный ряд $x_k = \mathrm{norm}(0,1),\ \ k = 1, \dots, 195$; Дополнить его пятью значениями: 5, -4, 3.3, 2.99, -3. Отсортировать ряд и применить правило трех сигм для проверки на аномальность для трех первых и трех последних порядковых статистик.
2. Построить для сгенерированного ряда боксплот Тьюки, выделить обнаруженные аномальные значения. Сравнить с результатами предыдущего задания. Сделать выводы.