Задание 8

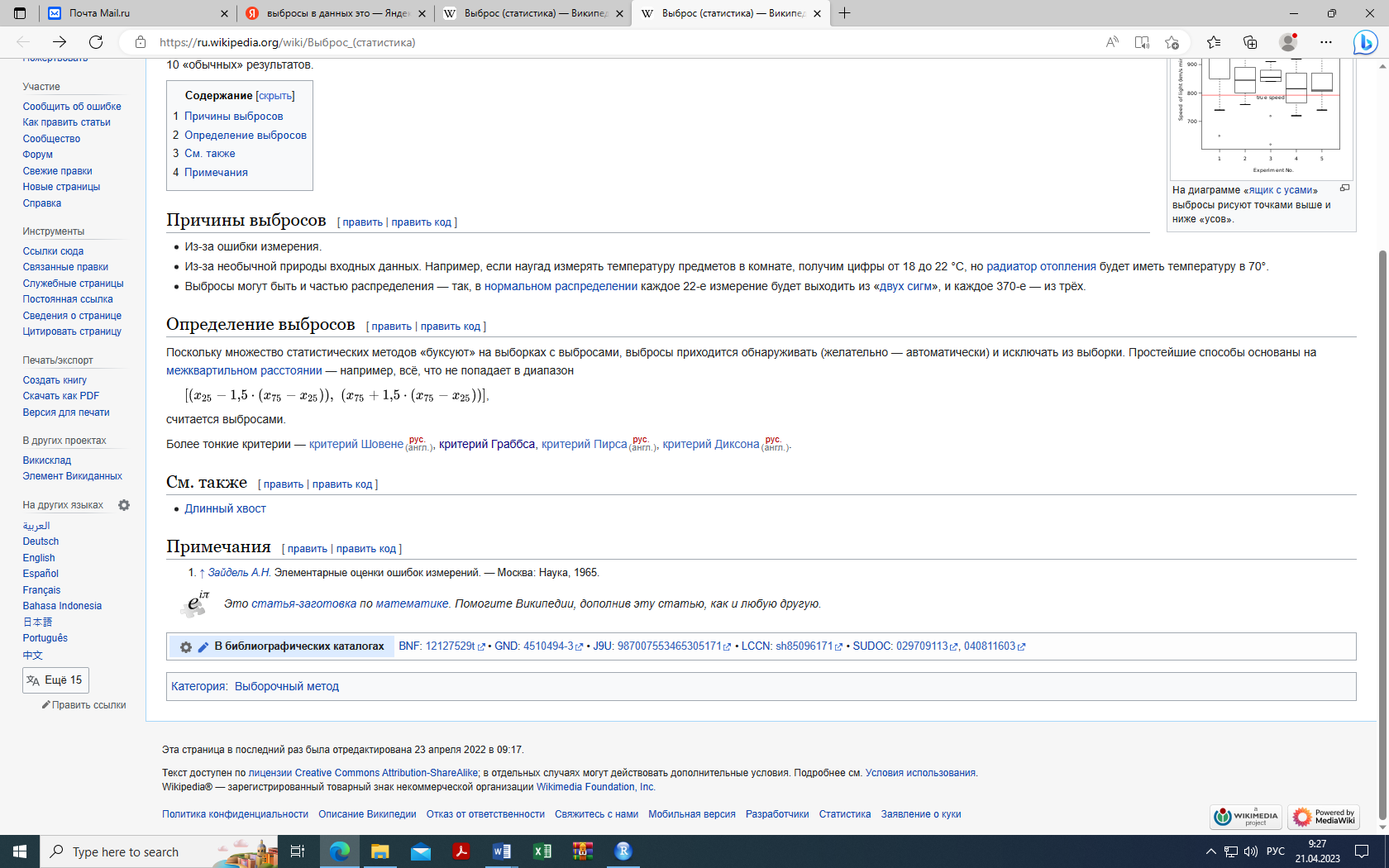
Выбросы

**Выброс** (*outlier*), **промах** – в [статистике](https://ru.wikipedia.org/wiki/%D0%A1%D1%82%D0%B0%D1%82%D0%B8%D1%81%D1%82%D0%B8%D0%BA%D0%B0) результат [измерения](https://ru.wikipedia.org/wiki/%D0%98%D0%B7%D0%BC%D0%B5%D1%80%D0%B5%D0%BD%D0%B8%D0%B5), выделяющийся из общей [выборки](https://ru.wikipedia.org/wiki/%D0%92%D1%8B%D0%B1%D0%BE%D1%80%D0%BA%D0%B0).

Обнаружение выбросов и аномалий всегда актуально для исследователя. Если решаются задачи прогнозирования, то удаление нетипичных значений, как правило, повышает точность предсказаний, поскольку данные без аномалий представляют собой типичный объект. Кроме того, статистические характеристики чувствительны к наличию выбросов.

Всегда нужно сравнивать наблюдение с другими значениями, полученными тем же способом, прежде чем называть их выбросами. Действительно, человек с ростом 200 см, скорее всего, будет считаться отклонением по сравнению с остальным населением, но этот же человек не будет считаться статистическим выбросом, если мы измерим рост баскетболистов.

Поскольку множество статистических методов «буксуют» на выборках с выбросами, выбросы приходится обнаруживать (желательно автоматически) и исключать из выборки. Простейшие способы основаны на [межквартильном расстоянии](https://ru.wikipedia.org/wiki/%D0%9C%D0%B5%D0%B6%D0%BA%D0%B2%D0%B0%D1%80%D1%82%D0%B8%D0%BB%D1%8C%D0%BD%D0%BE%D0%B5_%D1%80%D0%B0%D1%81%D1%81%D1%82%D0%BE%D1%8F%D0%BD%D0%B8%D0%B5) — например, всё, что не попадает в диапазон считается выбросами:



Другой простой способ обнаружить выбросы – использовать функцию boxplot. Рассмотрим на примере.

Допустим у нас есть следующие данные:

x <- c(

4.27296076175273, 4.3253044335640, 4.55202839528403,

4.33285651824668, 4.0128340310945, 4.07155452370293,

4.04113475664987, 2.7753693269563, 2.49186430883914,

2.39823431758359, 2.3789955162936, 2.14825292752989,

2.19583315284264, 2.1947104626036, 2.14216783486566,

2.28399128121205, 2.3048696257819, 2.23703873535593,

2.28669486582313, 2.4163169738317, 2.03782758779637,

2.61237703056071, 3.47620332697605)

y <- c(

9.097731,9.512591,9.740439,9.910364,9.865059,9.935519,

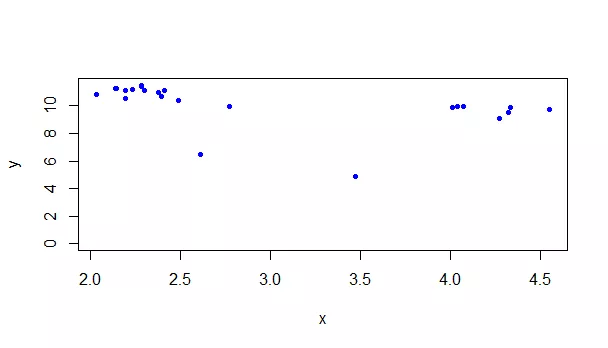
9.972640,9.920197,10.367693,10.680861,10.999012,11.246248,

10.532816,11.144033,11.261961,11.400160,11.080695,11.173922,

11.506877,11.105483,10.807685,6.489205,4.882802)

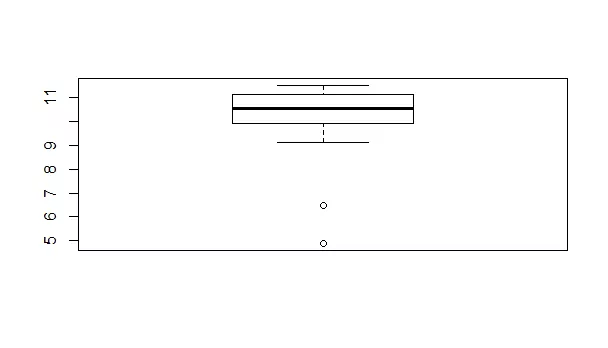
Выведем графическое отображение данных:

plot(x,y)



Скорее всего здесь есть выбросы. Если эти данные моделировать линейной регрессией, то эти точки очень сильно исказят результаты предсказания для множества остальных точек. Давайте проверим на выбросы:

boxplot(y)



Присутствуют две точки выброса.

Самый простой способ – удалить эти точки. boxplot не только рисует график, но и сохраняет все его параметры в объекте. Из которого мы можем достать их. Например, выбросы хранятся тут: boxplot.stats(y)$out

Получим индексы точек выбросов в наших векторах:

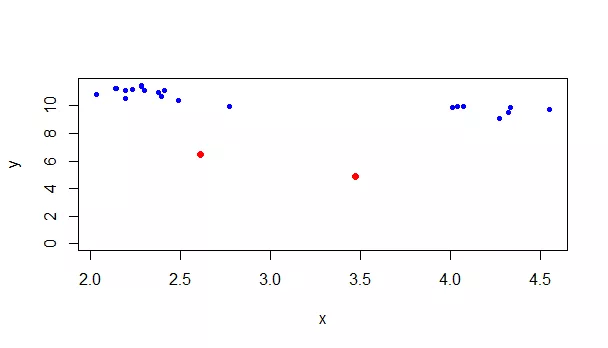
ind <- which(y %in% boxplot.stats(y)$out)

Сохраним координаты точек выбросов в отдельном dataframe:

outler <- data.frame(x=x[ind], y=y[ind])

Теперь проверим - те ли вообще точки мы нашли?

plot(x,y,col='blue', pch=20, ylim=c(0,max(y)))  
points(outler$x, outler$y, col='red',pch=19)



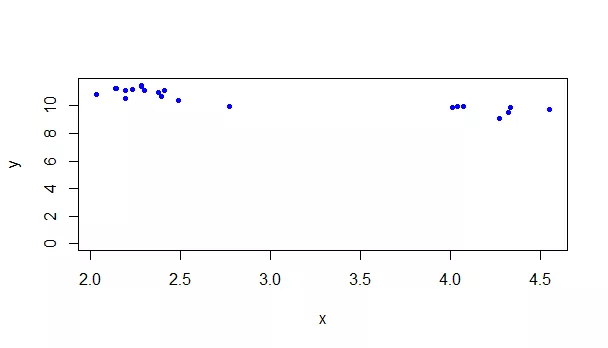
Дальше очистим данные от выбросов, проверим, что статистических выбросов не наблюдается и посмотрим на новые "чистые" данные.

x <- x[-ind]

y <- y[-ind]

boxplot(y)

plot(x,y,col='blue', pch=20, ylim=c(0,max(y)))



**Реализация критериев определения выбросов в статистическом пакете R**

**Тест Граббса** — это статистический тест, который можно использовать для выявления наличия выбросов в наборе данных. Чтобы использовать этот тест, набор данных должен иметь примерно нормальное распределение и иметь не менее 7 наблюдений.

Для выполнения теста Граббса в R используется функция grubbs.test{ outliers}. Нулевая гипотеза: в данных нет выбросов.

Пример

library(outliers)

data <- c(5, 14, 15, 15, 14, 13, 19, 17, 16, 20, 22, 8, 21, 28, 11, 9, 29, 40)

grubbs.test(data)

Grubbs test for one outlier

data: data

G = 2.65990, U = 0.55935, p-value = 0.02398

alternative hypothesis: highest value 40 is an outlier

Так как, полученное значение p-value = 0.02398<0.05, нулевую гипотезу отвергаем, выбросы присутствуют в данных. Делаем вывод, что максимальное значение 40 является выбросом.

Если бы вместо этого мы хотели проверить, является ли наименьшее значение «5» выбросом, то:

grubbs.test(data, opposite= TRUE )

Grubbs test for one outlier

data: data

G = 1.4879, U = 0.8621, p-value = 1

alternative hypothesis: lowest value 5 is an outlier

Нулевую гипотезу принимаем. Минимальное значение «5» не является выбросом.

Предположим, что у нас есть два больших значения на одном конце набора данных: 40 и 42. Чтобы проверить, являются ли оба этих значения выбросами, мы могли бы выполнить тест Граббса и указать, что type=20:

data <- c(5, 14, 15, 15, 14, 13, 19, 17, 16, 20, 22, 8, 21, 28, 11, 9, 29, 40, 42)

grubbs.test(data, type=20)

Grubbs test for two outliers

data: data

U = 0.38111, p-value = 0.01195

alternative hypothesis: highest values 40, 42 are outliers

Значение p теста составляет 0.01195. Поскольку это меньше 0,05, отклоняем нулевую гипотезу и делаем вывод, что значения 40 и 42 являются выбросами.

Иногда значения, которые отображаются как выбросы в наборах данных, являются просто опечатками, сделанными человеком при вводе данных. Вернитесь и убедитесь, что значение было введено правильно, прежде чем принимать какие-либо дальнейшие решения.

Если выброс окажется результатом опечатки или ошибки ввода данных, вы можете решить присвоить ему новое значение, например, среднее значение или медиану набора данных.

Если значение является истинным выбросом, вы можете удалить его, если оно окажет значительное влияние на общий анализ.

**Фильтр Хэмпеля**

Другой метод, известный как фильтр Хэмпеля, заключается в том, чтобы рассматривать как выбросы значения вне интервала, которые формируются медианным значением плюс-минус 3 медианы абсолютных отклонений (MAD):

I = [median - 3 \* MAD; median + 3 \* MAD]

В которых MAD — это медианное абсолютное отклонение и определяется как медиана абсолютных отклонений от медианы данных:

https://leftjoin.ru/pictures/--2021-03-24--14.09.11.png

Для этого метода сначала устанавливаем пределы интервала с помощью функций median() и mad() на примере:

x=rnorm(seq(1,500),23,1)

boxplot(x)

summary(x)

Min. 1st Qu. Median Mean 3rd Qu. Max.

19.81 22.36 23.04 23.04 23.72 25.96

lower\_bound <- median(x) - 3 \* mad(x, constant=1)

lower\_bound

[1] 21.02341

upper\_bound <- median(x) + 3 \* mad(x, constant=1)

upper\_bound [1] 25.06453

Все наблюдения меньше 21.02343 и больше 25.06453 будут рассматриваться как потенциальные выбросы. Номера строк наблюдений за пределами интервала затем могут быть извлечены с помощью функции which():

outlier\_ind <- which(x < lower\_bound | x > upper\_bound)

outlier\_ind

[1] 1 7 19 41 59 136 194 220 250 258 271 278 282 317 345 356 396 432 443 446 447

> x[14]

[1] 22.74829

> x[498] [1] 21.02343

Согласно фильтру Хэмпеля, для переменной x есть 21 потенциальных выброса.

**Тест Диксона (Dixon’s test)**

Подобно тесту Граббса, тест Диксона используется для того, чтобы проверить, является ли самое высокое или самое низкое значение выбросом. Таким образом, если под сомнением находятся более одного выброса, тест необходимо проводить индивидуально для этих предполагаемых значений.

Тест Диксона используется для выборки небольшого объема (обычно когда n <= 25). Нулевая гипотеза: наименьшее значение выборки не является выбросом.

Чтобы выполнить тест Диксона в R, используем функцию dixon.test{outliers}. Если необходимо применить гипотезу к наибольшему значению, необходимо добавить в функцию аргумент opposite = TRUE).

Test1 <- dixon.test(x[1:25])

Test1

Dixon test for outliers

data: x[1:25]

Q = 0.17751, p-value = 0.868

alternative hypothesis: lowest value 20.8007768943081 is an outlier

Результаты показывают, что самое наименьшее значение 20.8007768943081 не является выбросом (p-value = 0.868).

Чтобы проверить максимальное значение, просто добавьте аргумент opposite = TRUE к функции dixon.test():

test2 <- dixon.test(x[1:25],opposite = TRUE)

test2

Dixon test for outliers

data: x[1:25]

Q = 0.14488, p-value = 0.6453

alternative hypothesis: highest value 25.3421161469636 is an outlier

Результаты показывают, что максимальное значение 25.3421161469636 не является выбросом (p-value = 0.6453).

Рекомендуется всегда сверять результаты статистического теста на выбросы с диаграммой, чтобы убедиться, что проверили все потенциальные выбросы:

boxplot(x[1:25])

или

boxplot(x[1:25])

out <- boxplot.stats(x[1:25])$out

boxplot(x[1:25],ylab = "x")

mtext(paste("Outliers: ", paste(out, collapse = ", ")))

**Задание**

Определите наличие выбросов в количестве выживших на растениях насекомых в зависимости от эффективности инсектицидных средств (**data(InsectSprays)).**

Данные, полученные в ходе эксперимента по изучению эффективности шести видов инсектицидных средств. Каждым из этих средств обработали по 12 растений, после чего подсчитали количество выживших на растениях насекомых. В таблице InsectSprays имеется два столбца - count, содержащий результаты подсчета насекомых, и spray, содержащий [коды](http://r-analytics.blogspot.ru/2011/11/r_08.html) инсектицидных средств (от А до F).

1. Постройте диаграмму размаха показателя InsectSprays$count, а также диаграмму размаха показателя InsectSprays$count по видам инсектицидных средств.
2. Определите выбросы в данных по видам инсектицидных средств, используя разные методы.