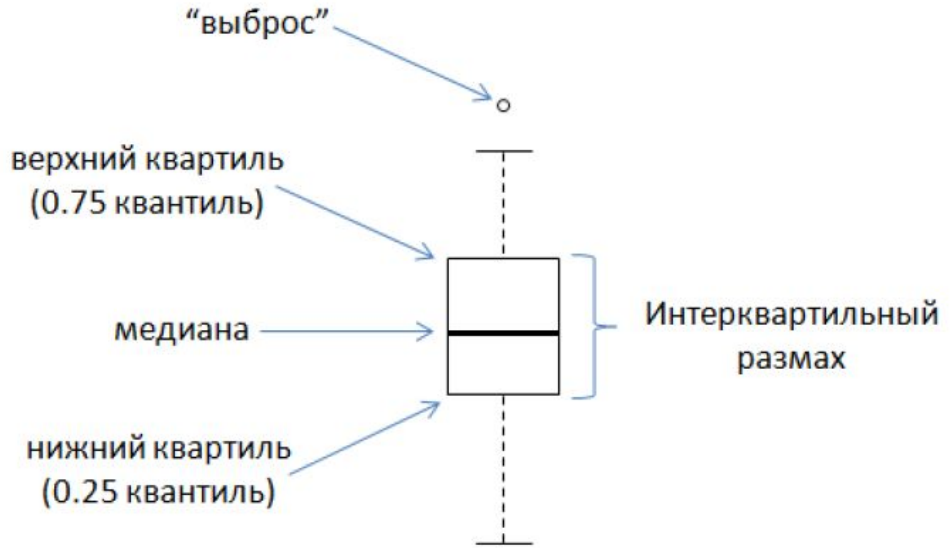
Анализ выбросов

Диаграммы размахов (ящичковые диаграммы, диаграммы с усами)

- Функция boxplot()



- Наблюдения, находящиеся за пределами "усов", потенциально могут быть выбросами. Однако всегда следует внимательно относиться к такого рода нестандартным наблюдениям – они вполне могут оказаться "нормальными" для исследуемой совокупности, и поэтому не должны удаляться из анализа без дополнительного расследования причин их появления.

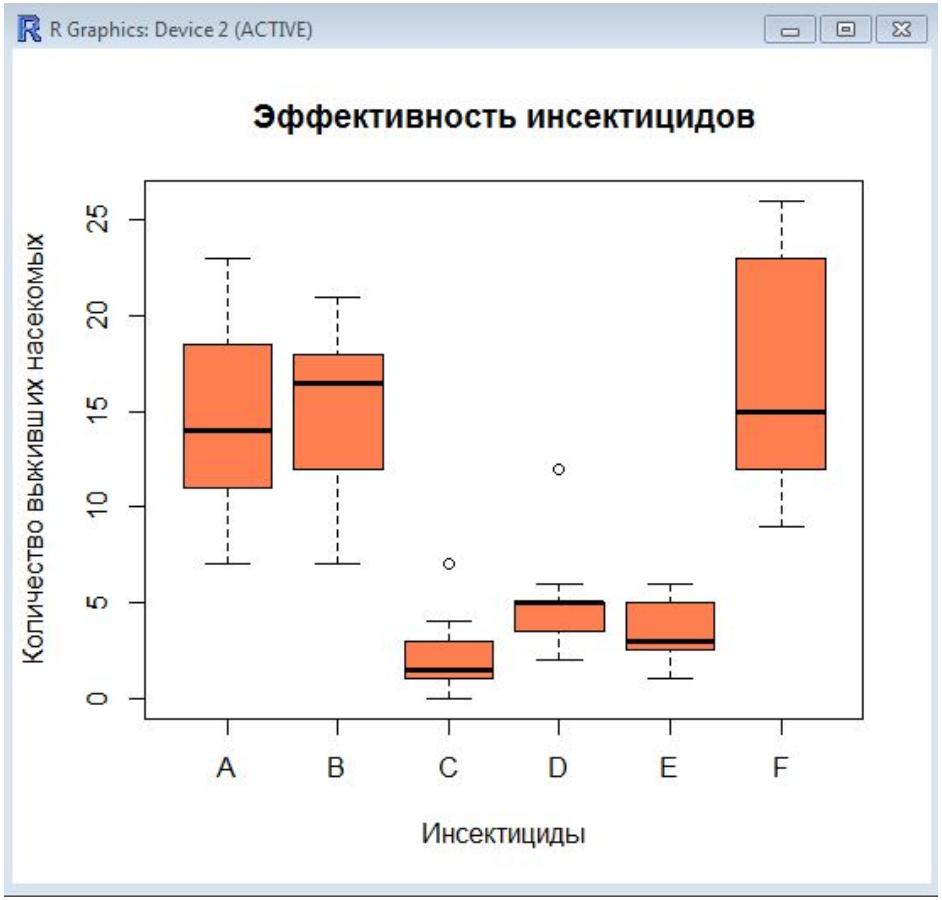
Применение boxplot()

- Рассмотрим 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. Данные этого эксперимента входят в состав стандартного набора данных R и доступны по команде data(InsectSprays). В таблице InsectSprays имеется два столбца: count, содержащий результаты подсчета насекомых, и spray, содержащий коды инсектицидных средств (от A до F).
- Для построения графика, на котором будут представлены "ящики с усами" для каждого инсектицида, достаточно выполнить команду

  boxplot(count ~ spray, data = InsectSprays)
- Улучшим график

  boxplot(count ~ spray, xlab = "Инсектициды", ylab = "Количество выживших насекомых", main = "Эффективность инсектицидов", col = "coral", data = InsectSprays)

Результат

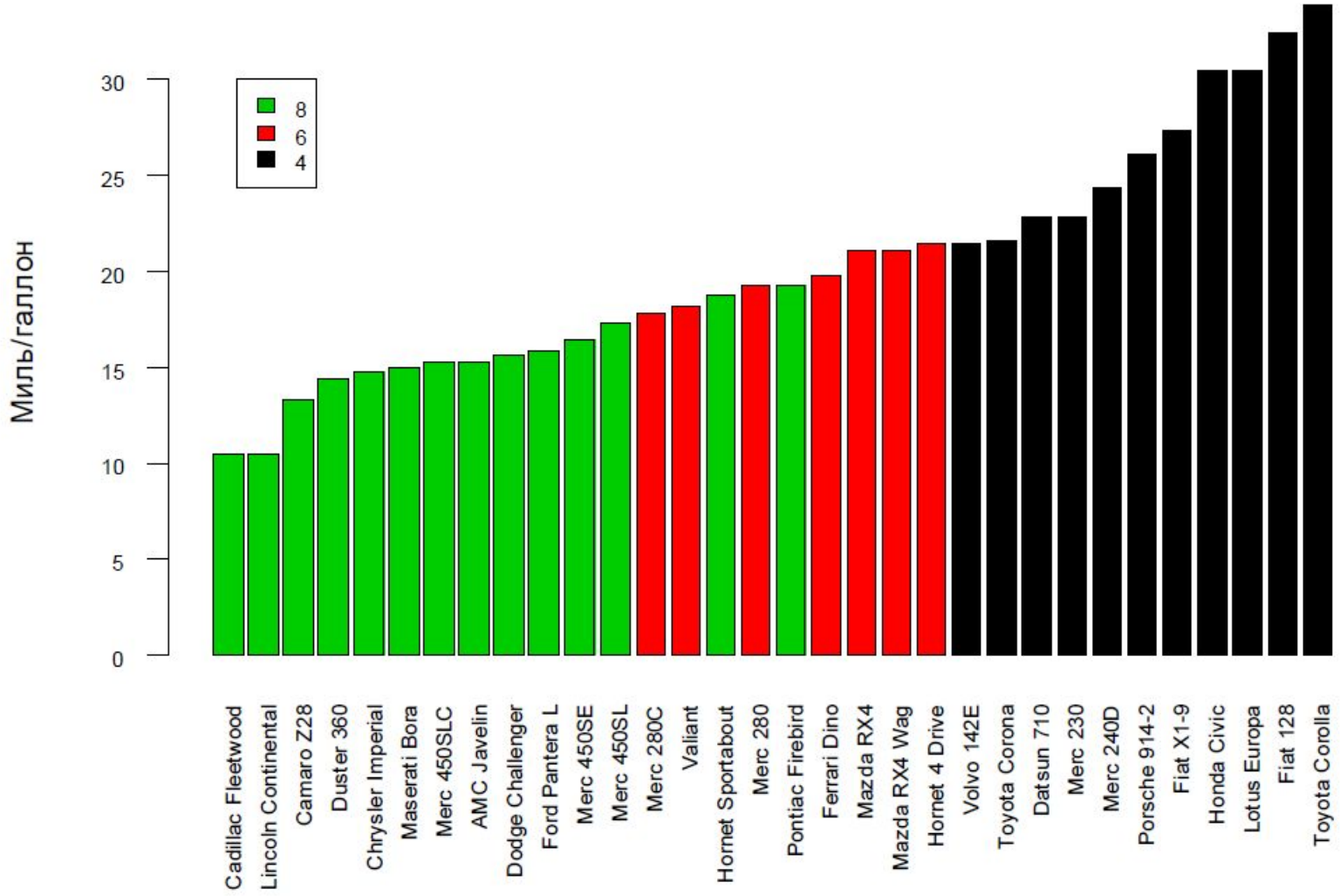


- Как видим, количество насекомых на растениях, обработанных инсектицидами C, D и E было наиболее низким, что говорит о высокой эффективности этих препаратов по сравнению с тремя другими средствами. На растениях, обработанных средствами C и D, были отмечены необычно высокие количества насекомых (см. точки над "усами").

Диаграммы Кливленда

- ***Точечные диаграммы Кливленда*** представляют собой графики, на которых точки используются для отображения значений некоторой количественной переменной (или переменных), разбитых на группы в соответствии с уровнями некоторой номинальной переменной (или переменных). Считается, что такие диаграммы визуально лучше воспринимаются.
- Построим столбиковую диаграмму, изображающую распределение 32 моделей автомобилей 1973-1974 годов выпуска по экономичности двигателя (выражается как количество миль, которое автомобиль проезжает на одном галлоне топлива). Данные, использованные для построения диаграммы, были опубликованы в американском журнале *Motor Trend в 1974 г. и входят в стандартный* набор данных R (доступны по команде data(mtcars)).

Пример столбчатой диаграммы



Точечная диаграмма

dotchart(mtcars$mpg, labels = row.names(mtcars), main=" Экономия топлива у 32 моделей автомобилей", xlab="Миль/галлон", cex = 0.8)

- Параметры функции dotchart():

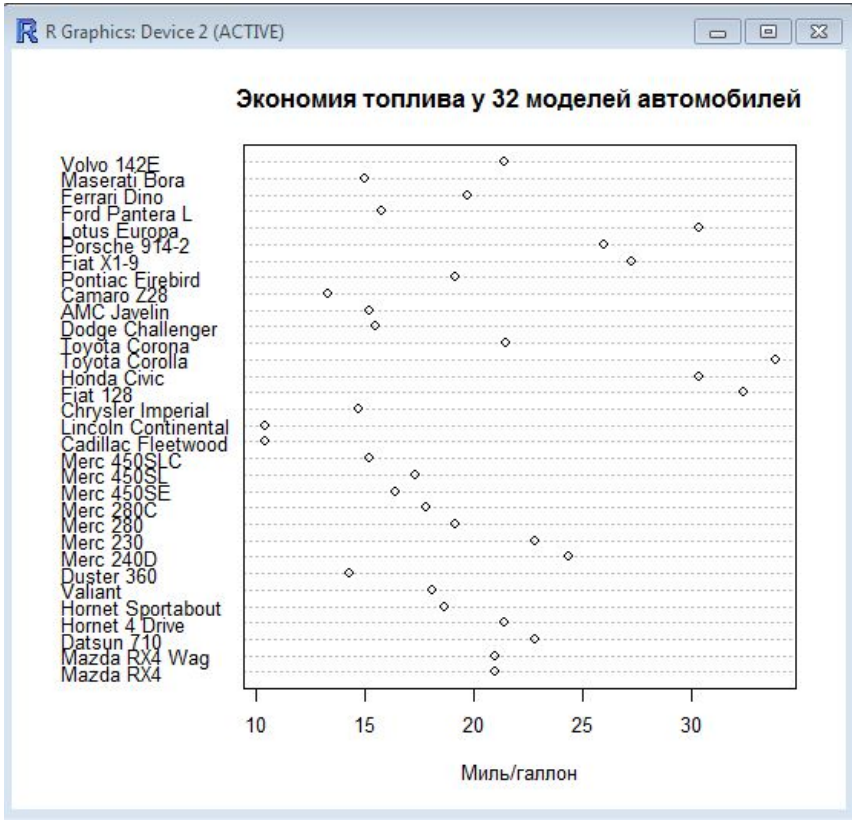
а) переменная, для которой строится график (mtcars$mpg);

б) текстовый вектор, содержащий названия моделей автомобилей (в данном случае они являются названиями строк таблицы – row.names(mtcars));

в) заголовок графика (аргумент main) и название оси X (аргумент xlab);

г) размер точек на графике и одновременно размер шрифта для названий моделей (cex = 0.8).

Результат



- Картина станет гораздо более ясной, если мы отсортируем данные по возрастанию пробега, сгруппируем данные по количеству цилиндров в двигателе и раскрасим соответствующие группы разными цветами.

Подготовка данных

- Отсортируем исходную таблицу по возрастанию mpg с использованием функции order()) и сохраним результат в виде новой таблицы данных с именем x:

```
x <- mtcars[order(mtcars$mpg), ]
```

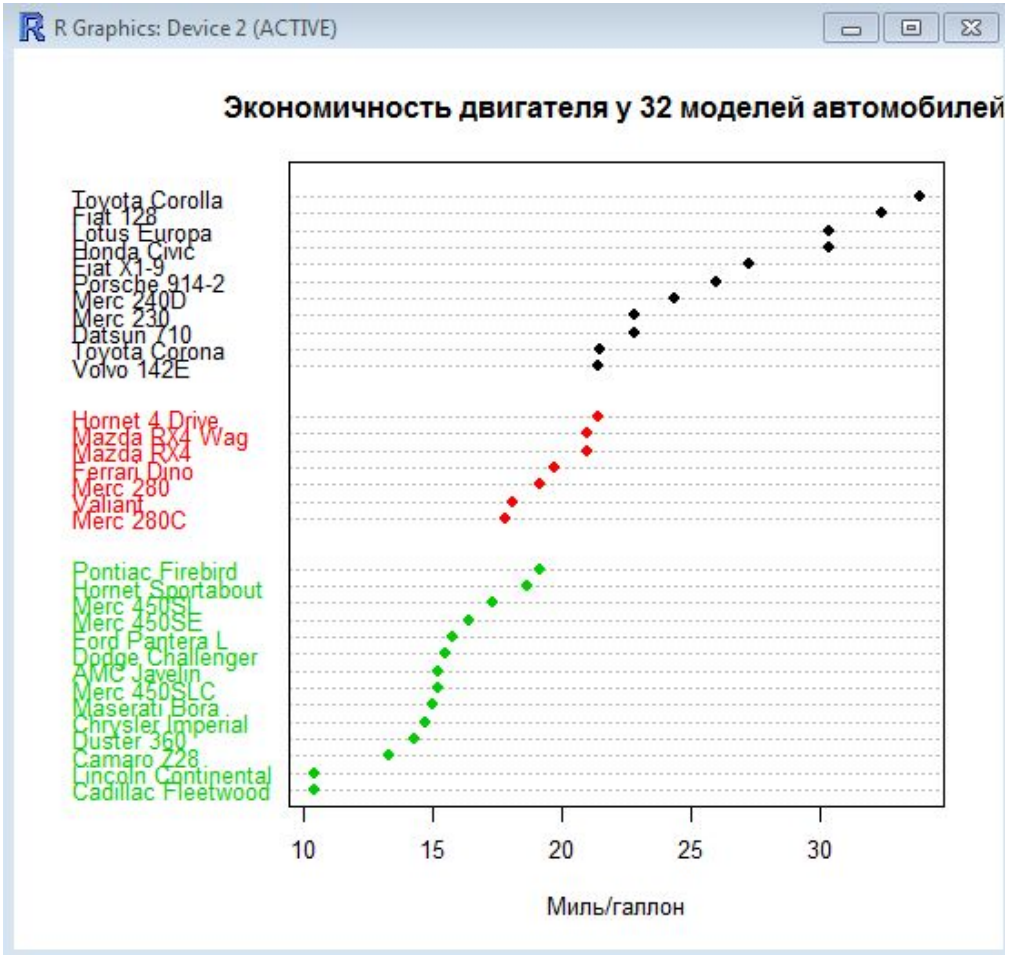
- Создадим новый столбец color в таблице x, который будет содержать числовые коды цветов для каждой из трех групп автомобилей:

```
x$color[x$cyl==4] <- 1
x$color[x$cyl==6] <- 2
x$color[x$cyl==8] <- 3
```

- Теперь у нас есть все необходимое для построения желаемого графика:

```
dotchart(x$mpg, labels = row.names(x), groups = x$cyl, gcolor = "blue", pch = 16, main="Экономичность двигателя у 32 моделей   автомобилей", xlab="Миль/галлон", cex = 0.8, color = x$color)
```

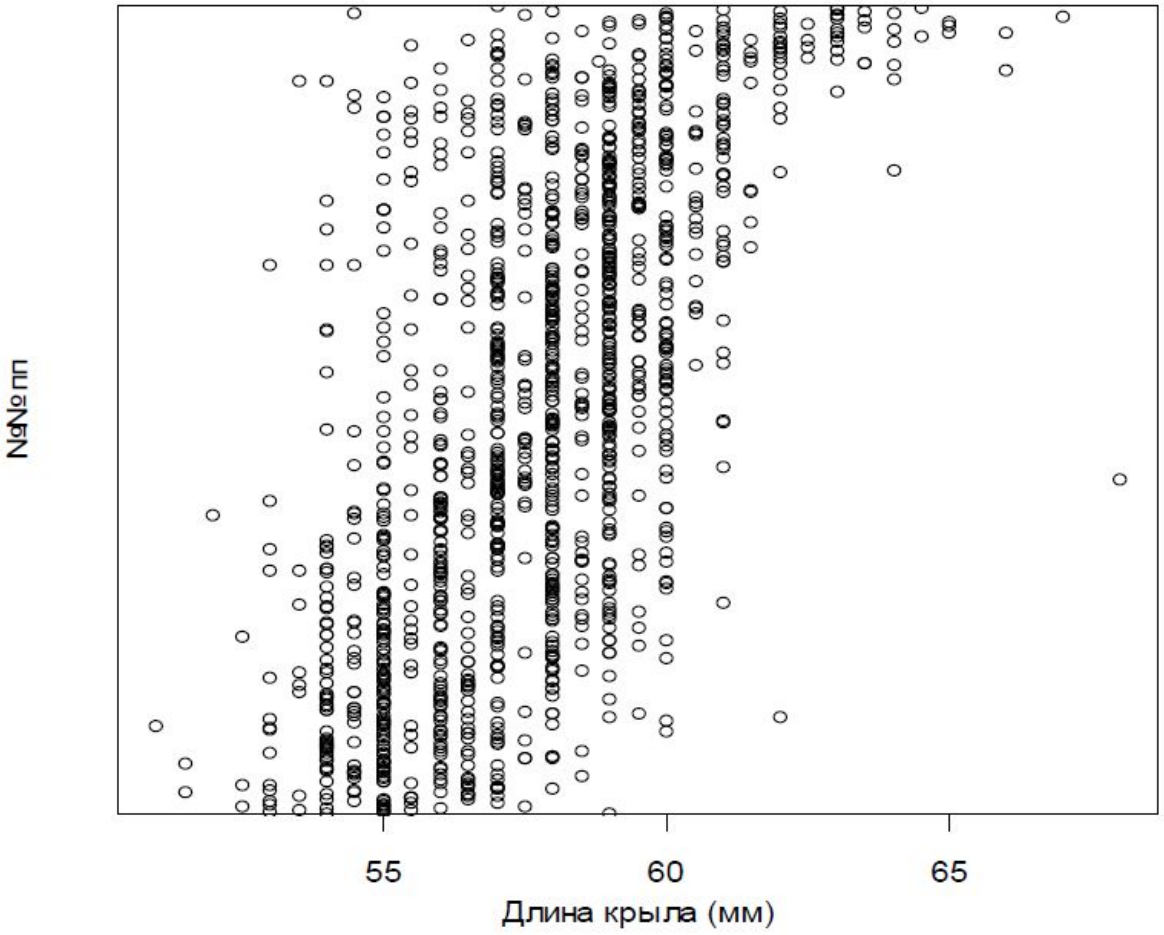
Результат


R Graphics: Device 2 (ACTIVE)
Экономичность двигателя у 32 моделей автомобилей
Toyota Corolla
Fiat 128
Lotus Europa
Honda Civic
Fiat X1-9
Porsche 914-2
Merc 240D
Merc 230
Datsun 710
Toyota Corona
Volvo 142E
Hornet 4 Drive
Mazda RX4 Wag
Mazda RX4
Ferrari Dino
Merc 280
Valiant
Merc 280C
Pontiac Firebird
Hornet Sportabout
Merc 450SL
Merc 450SE
Ford Pantera L
Dodge Challenger
AMC Javelin
Merc 450SLC
Maserati Bora
Chrysler Imperial
Duster 360
Camaro Z28
Lincoln Continental
Cadillac Fleetwood
10
15
20
25
30
Миль/галлон

Выбросы

- Под "выбросом" мы будем понимать наблюдение, которое "слишком" велико или "слишком" мало по сравнению с большинством других имеющихся наблюдений.
- Обычно для выявления выбросов используют диаграмму размахов или точечную диаграмму Кливленда.
- Например, на следующей диаграмме Кливленда, представлены данные о длине крыла у 1295 воробьев. Здесь таблица была предварительно упорядочена в соответствии с весом птиц, и поэтому облако точек имеет примерно S-образную форму.

Диаграмма Кливленда распределения длин крыла воробьев

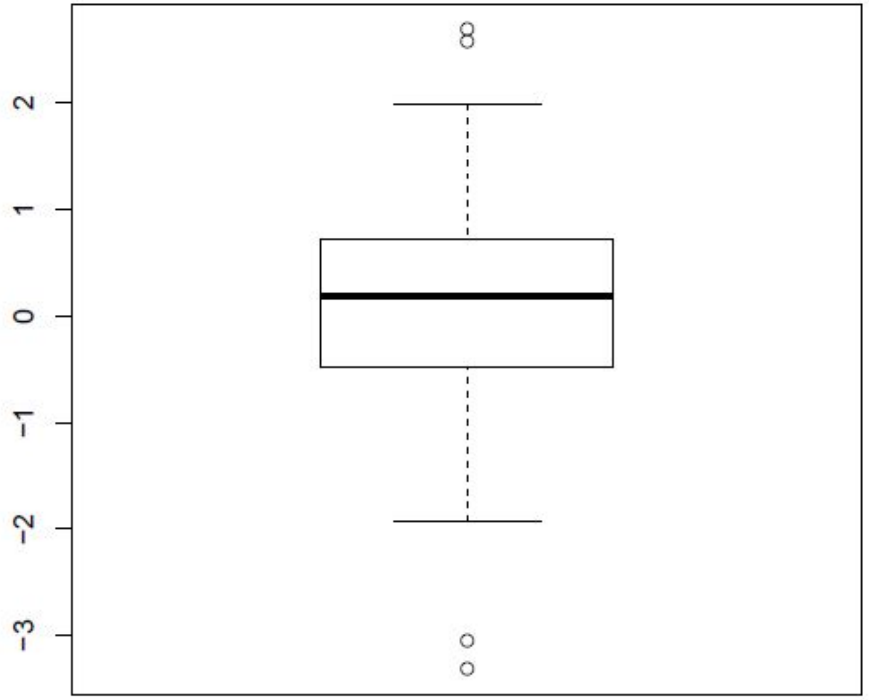


Интерпретация результатов

- На рисунке хорошо выделяется точка, соответствующая длине крыла 68 мм. Однако это значение длины крыла не следует рассматривать в качестве выброса, поскольку оно лишь незначительно отличается от других значений длины. Эта точка выделяется на общем фоне лишь потому, что исходные значения длины крыла были упорядочены по весу птиц. Соответственно, выброс скорее стоит искать среди значений веса (т.е. очень высокое значение длины крыла (68 мм) было отмечено у воробья, необычно мало весящего для этого).

Поиск выбросов с помощью диаграммы размаха

```
> set.seed(3147)
> x <- rnorm(100)
> summary(x)
Min. 1st Qu. Median Mean 3rd Qu. Max.
-3.3150 -0.4837 0.1867 0.1098 0.7120 2.6860
> # outliers
> boxplot.stats(x)$out
[1] -3.315391 2.685922 -3.055717 2.571203
> boxplot(x)
```



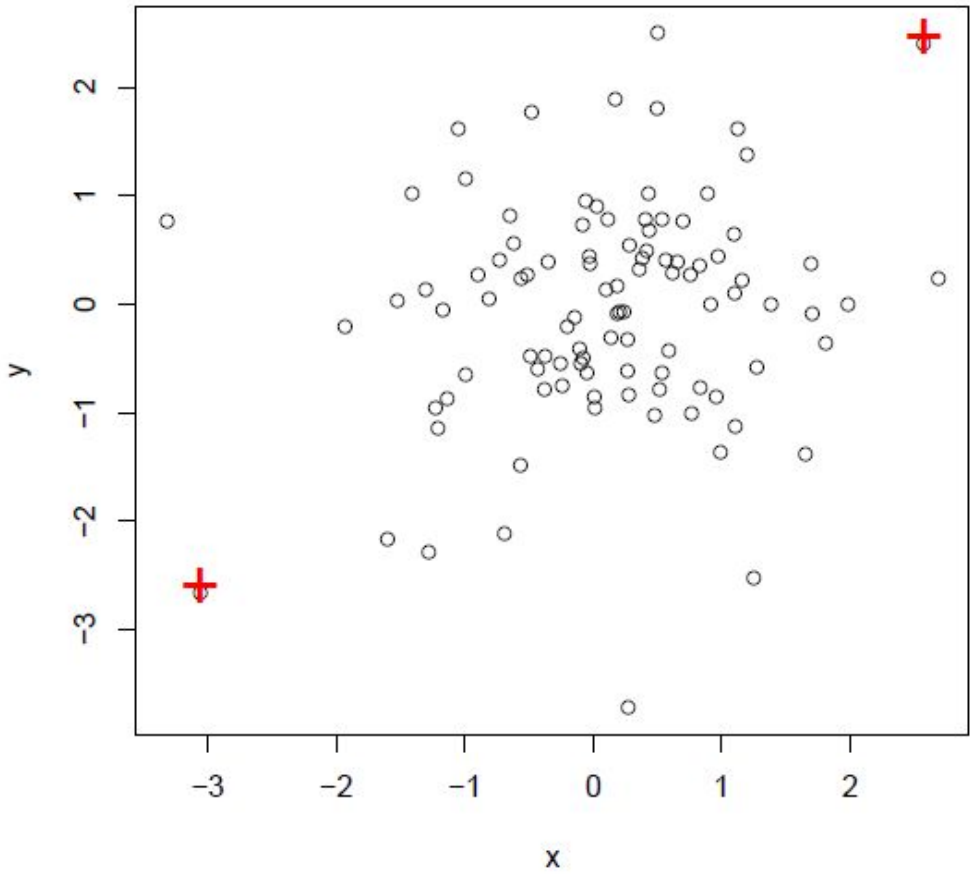
Поиск выбросов с помощью диаграммы размаха

```
> y <- rnorm(100)
> df <- data.frame(x, y)
> rm(x, y)
> head(df)
x y
1 -3.31539150 0.7619774
2 -0.04765067 -0.6404403
3 0.69720806 0.7645655
4 0.35979073 0.3131930
5 0.18644193 0.1709528
6 0.27493834 -0.8441813
```

Поиск выбросов с помощью диаграммы размаха

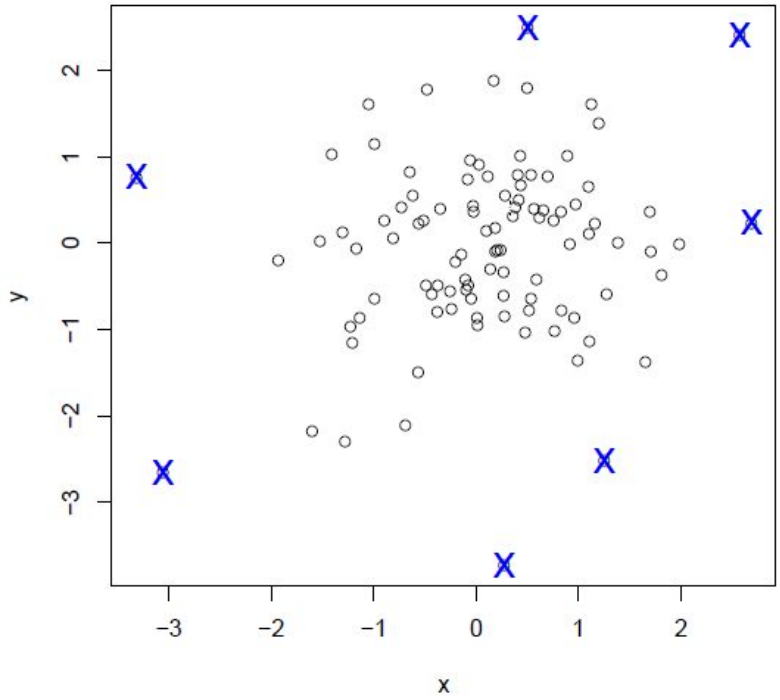
```
> attach(df)
> # find the index of outliers from x
> (a <- which(x %in% boxplot.stats(x)$out))
[1] 1 33 64 74
> # find the index of outliers from y
> (b <- which(y %in% boxplot.stats(y)$out))
[1] 24 25 49 64 74
> detach(df)
> # outliers in both x and y
> (outlier.list1 <- intersect(a,b))
[1] 64 74
> plot(df)
> points(df[outlier.list1,], col="red", pch="+", cex=2.5)
```

Результат



Поиск выбросов

```
> # outliers in either x or y
> (outlier.list2 <- union(a,b))
[1] 1 33 64 74 24 25 49
> plot(df)
> points(df[outlier.list2,], col="blue", pch="x“, cex=2)
```



Удаление выбросов

- Более строгий подход к определению выбросов состоит в оценке того, какое влияние эти необычные наблюдения оказывают на результаты анализа. При этом следует делать различие между необычными наблюдениями для зависимых и независимых переменных (предикторов). Например, при изучении зависимости численности какого-либо биологического вида от температуры большинство значений температуры может лежать в пределах от 15 до 20 °C, и лишь одно значение может оказаться равным 25 °C. Такой план эксперимента, мягко говоря, неидеален, поскольку диапазон температур от 20 до 25 °C будет исследован неравномерно. Однако при проведении реальных полевых исследований возможность выполнить измерения для высокой температуры может представиться только однажды. Что же тогда делать с этим необычным измерением, выполненным при 25 °C? При большом объеме наблюдений подобные редкие наблюдения можно исключить из анализа. Однако при относительно небольшом объеме данных еще большее его уменьшение может быть нежелательным с точки зрения статистической значимости получаемых результатов.

Нормализующее преобразование

- Альтернативой удалению необычных значений предиктора является нормализующее преобразование (чаще всего, логарифмирование). В общем случае, найти оптимальное решение позволяет так называемое преобразование Бокса-Кокса.
- Универсальное семейство преобразований Бокса-Кокса (БК) случайной величины *x* является степенным преобразованием

$$x' = \frac{x^{\lambda} - 1}{\lambda}$$

с произвольным положительным или отрицательным показателем степени λ. Поскольку деление на нуль приводит к неопределенности, то при λ = 0 используется логарифмическое преобразование *x'(λ) = ln(x). Найти значение* λ *можно, например, найдя максимум логарифма функции* максимального правдоподобия.

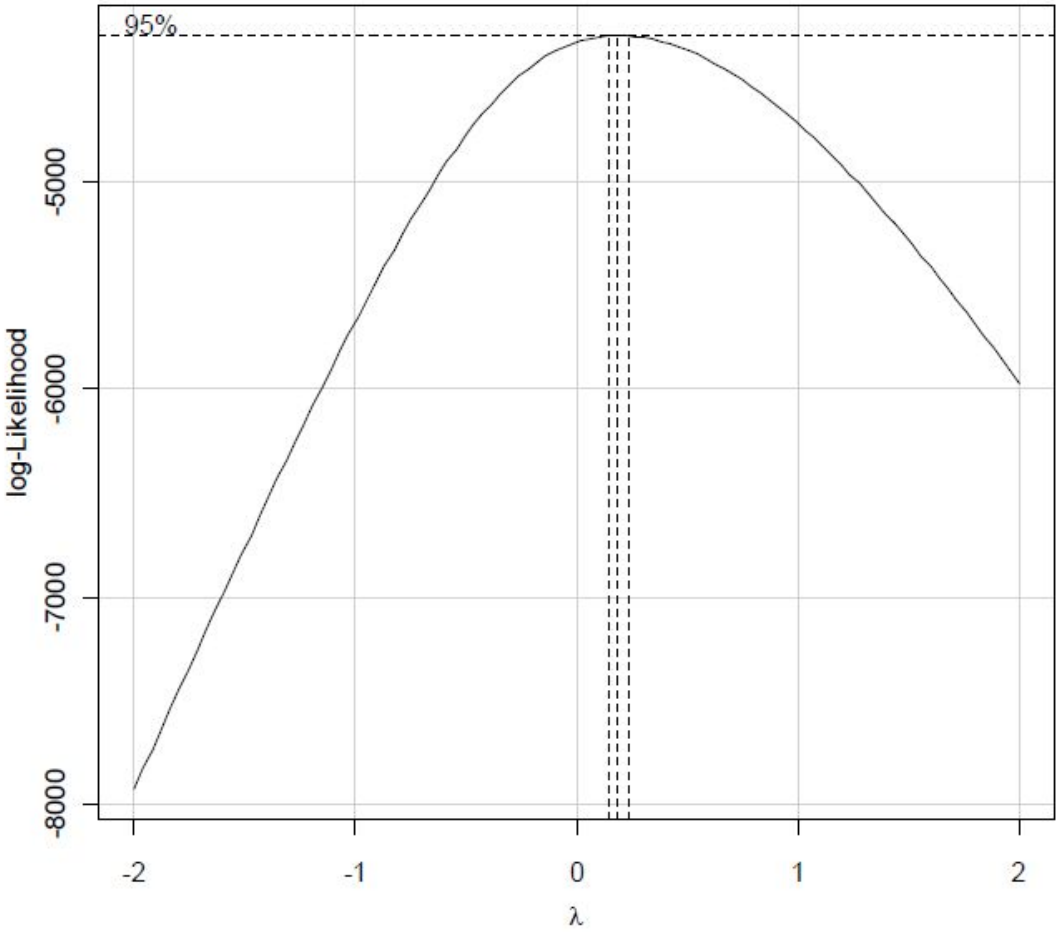
Пример

- Проанализируем уровень зараженности двустворчатого моллюска *Dreissena polymorpha инфузорией Conchophthirus acuminatus в трех озерах Беларуси. Данные возьмем с сайта* figshare. В таблице данных нас будут интересовать две переменные: длина раковины моллюска (ZMlength, мм) и число обнаруженных в моллюске инфузорий (CAnumber).

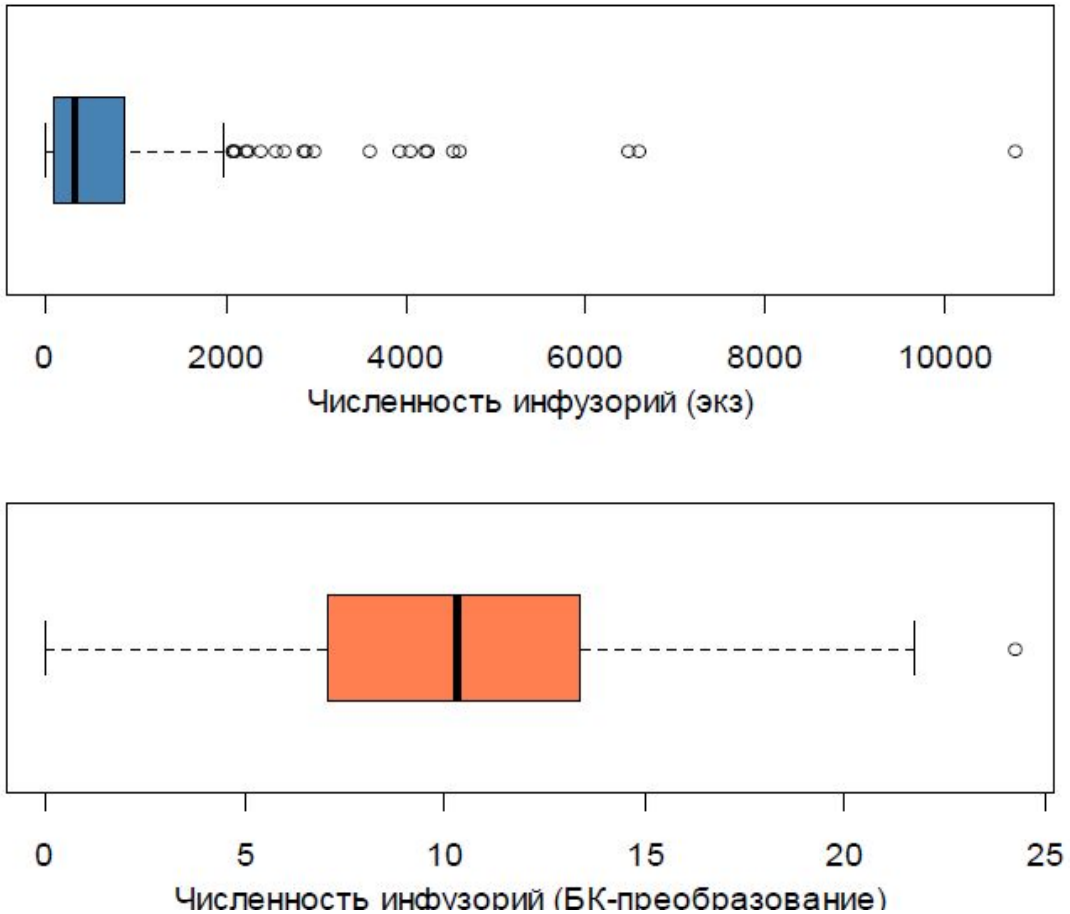
Пример

```
Моллюски <-
read.table("http://figshare.com/media/download/98923/97987",
header=TRUE, sep="\t", strip.white=TRUE)
library(car)
# Поиск максимума функции правдоподобия и построение графика
# изменения параметра БК-трансформации для заданной модели
m.null <- lm(Моллюски$CAnumber+1~1)
bc.null <- boxCox(m.null)
bc.null.opt <- bc.null$x[which.max(bc.null$y)]
paste("Оптимальная лямбда БК-преобразования:",bc.null.opt)
[1] "Оптимальная лямбда БК-преобразования: 0.181818181818182"
CAnumber_bc <- bcPower(Моллюски$CAnumber+1, bc.null.opt)
par(mfrow=c(2,1))
boxplot(Моллюски$CAnumber,horizontal = TRUE,col = "steelblue",
xlab = "Численность инфузорий (экз)")
boxplot(CAnumber_bc,horizontal = TRUE,col = "coral",
xlab = "Численность инфузорий (БК-преобразование)")
```

Оптимальное значение λ = 0.182



Эффект преобразования CAnumber по методу Бокса-Кокса



После БК-преобразования распределение значений CAnumber приблизилось к нормальному, в связи с чем за пределами указанного интервала оказалось только одно наблюдение.

Тест Граббса

- С использованием теста Граббса можно проверить нулевую гипотезу о том, что максимальное значение не является выбросом:

library(outliers)

grubbs.test(Моллюски$CAnumber, type = 10)

Grubbs test for one outlier

data: Моллюски$CAnumber

G = 10.8336, U = 0.7524, p-value < 2.2e-16

alternative hypothesis: highest value 10782 is an outlier

grubbs.test(CAnumber_bc, type = 10)

Grubbs test for one outlier

data: CAnumber_bc

G = 3.3267, U = 0.9767, p-value = 0.1961

alternative hypothesis: highest value 24.2569 is an outlier