Лукьянов Павел Борисович
профессор Департамента математики

Программирование в среде R

Лекция

Статистический анализ данных. Специальные графические функции

Финансовый университет, 2021

Типы данных, встречающиеся в исследованиях, экспериментах, деятельности

Первичные данные (записаны, сохранены, не обработаны)

Можно ли их отбрасывать ?

Выбросы

Числовые

- Данные, меняющиеся непрерывно
- Данные, меняющиеся дискретно

Пропущенные

- NA

Нечисловые

- Категориальные данные
  - **Факторы**
    - Порядковые (шкальные)
    - Номинальные
- Текстовая информация

Вторичные данные

Результаты обработки и преобразований

доли, проценты, расчетные значения, коэффициенты, числовые параметры и т.д.

Первичные данные: числовые данные

Числовые данные

Данные, меняющиеся непрерывно (интервальные данные)

- **Примеры**: масса, расстояние, температура, процентная ставка, зарплата и т.д.
- Непрерывные данные выражаются **действительными числами**, могут принимать любое значение внутри некоторого диапазона
- Значения двух переменных можно **упорядочить**, узнать, насколько одно значение больше другого
- Переменные можно **отобразить на мерной шкале**
- Больше всего методов обработки разработано для **интервальных данных**

R: тип double

Данные, меняющиеся дискретно (целочисленные данные)

- **Примеры**: годы, количество сотрудников в организации, число детей в семье, количество мест в самолете, число планет и т.д.
- Выражаются, как правило, целыми числами
- Промежуточных значений может не быть
- Значения двух переменных можно **упорядочить**
- Переменные можно **отобразить на мерной шкале**

R: тип integer

Первичные данные: нечисловые данные, порядковые

- **Примеры шкальных данных, данные разбиты на уровни**
  Состояние больного (тяжелое, средней тяжести, легкое недомогание, здоров)
  Погода (ясно, облачно, пасмурно, дождь, буря, ураган)
  Посещаемость занятий (всегда, часто, средне, редко, никогда)
- Придумывается **шкала, которая что-то отражает** (успеваемость, удобство, качество обслуживания и т.д.).
- Различным **значениям данных** на шкале **присваиваются баллы**
- Баллы **условны**: оценки школьнику, студенту в России – от 2 до 5, в Италии – от 1 до 30
- Каждому баллу соответствует некоторое описание, дается ответ о совпадении
- Вводятся **отношения порядка** (холодно -> прохладно -> тепло -> жарко), но сами значения нельзя оценить количественно
- Существует проблема совместимости данных из разных шкал
- **Общая задача:** уходить от шкальных данных к интервальным, т.к. для интервальных данных гораздо больше различных методов обработки данных, качество анализа выше
- **Пример.** Глубина моря около пляжа, факторы: от «Мелко», «Глубоко», «Очень глубоко», «Обрыв», «Бездна» переход к точным измерениям глубины

Функции factor(), as.factor() – из любого вектора делают набор категориальных данных

Первичные данные: нечисловые данные, номинальные

Нечисловые данные

Категориальные данные

Факторы

Номинальные данные

- **Примеры**
  Цвета: красный, белый, синий, черный
  Пол: мужской, женский
  Времена года: лето, осень, зима, весна
  Дни недели: понедельник, вторник, …
- Номинальные данные **не упорядочиваются.**
  Если присвоить данным цифровую кодировку 1, 2, …., то сравнение «больше – меньше» не имеет смысла
- В номинальных данных **цифры смысла не имеют** и не используются
- Для регистрации номинальных данных **используют буквы**, **сокращения слов, значки**, аббревиатуры и т.д.
- **Обработка** номинальных данных — **подсчет количеств совпадений** номинальных данных в общем массиве данных

Какой тип данных для факторов в R?

Бинарные данные

- Да / Нет
- 0 / 1
- Наличие / Отсутствие
- TRUE / FALSE

R: тип logical

Задача на расчет показателей выборки

Постановка задачи

Получена выборка: 2, 14, 5, 7, -3, 7, 11, 6, 0.
Рассчитать следующие показатели: медиану, первый и третий квартили, IQR.
Определить выбросы. Отметить рассчитанные значения на графике boxplot.

Наблюдения, исследования, эксперименты

Объекты исследования. Как называются

- **Генеральная совокупность**
  Совокупность всех объектов или их характеристик, по которым проводится анализ или исследование
- **Выборка**
  Некоторая часть генеральной совокупности

Результаты исследования. Где хранятся

Данные записываются и хранятся в

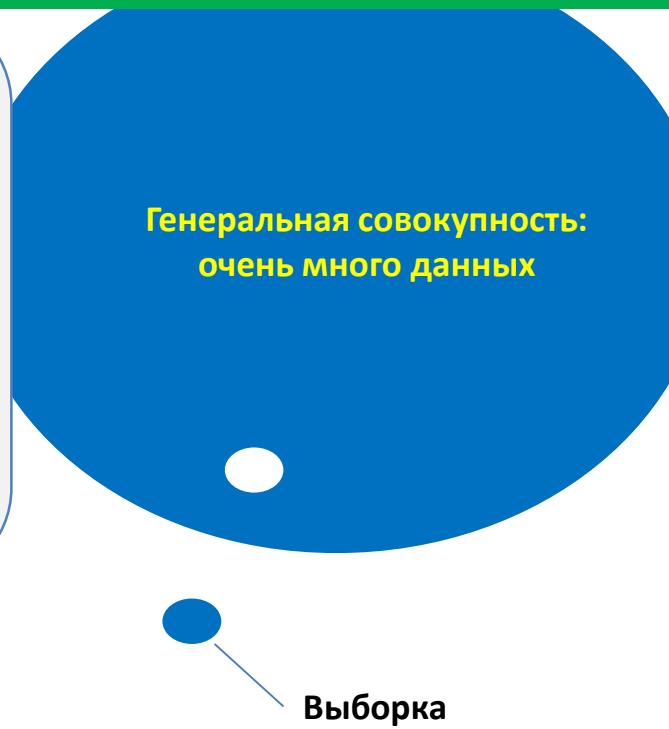
- **векторах (одномерные)**
- **таблицах, матрицах (двумерные)**
- **массивах (многомерные)**
- **списках (сложные структуры)**

Генеральная совокупность и выборка

Генеральная совокупность (ГС): совокупность всех объектов, по которым будут сделаны выводы при проведении исследования. Другой термин – популяция.

Как правило, **ГС** задается указанием нескольких условий, например

- Женщины + пенсионеры + пенсия > 10 тыс. руб + ходят в Пятерочку чаще 1 раза в неделю
- Взрослое население региона (область, город) + имеющее право голоса
- Семьи + наличие маленького ребенка + проживающие с родителями + доход выше 60 тыс на человека



В идеальном случае **нужно исследовать ВСЕ объекты ГС**
но ЭТО НЕРЕАЛЬНО: дорого, долго, физически невозможно и т.д.

Пример: Прогноз результатов выборов в регионе

Поэтому

- исследуют лишь **часть объектов (создают выборку)**
- по результатам анализа выборки делаются выводы
- полученные выводы переносят на всю **ГС**

Но корректно ли так делать?

- **При выполнении определенных условий - ДА**
- В разделе математики, занимающемся анализом выборок и условиями применимости выводов на **ГС**, разработана теория выборочных исследований. Занимается этим **математическая статистика**
- **Выборка** или **выборочная совокупность** — часть генеральной совокупности элементов, которая охватывается экспериментом или наблюдением

Что такое правильная выборка?

Характеристики выборки

- **Способ формирования**: по каким правилам из ГС извлекаются элементы для выборки
  - Репрезентативность
  - Повторность
  - Рандомизация

- **Объем**: сколько элементов из ГС включается в выборку

Репрезентативность

- Репрезентативная выборка – выборка конечного объёма, обладающая всеми свойствами ГС, значимыми с точки зрения задач исследования

- Репрезентативность определяет, возможно ли обобщать результаты исследования, выполненные на выборке, на всю генеральную совокупность, из которой она была собрана

- Необходимым условием построения репрезентативной выборки является равная вероятность включения в нее каждого элемента генеральной совокупности

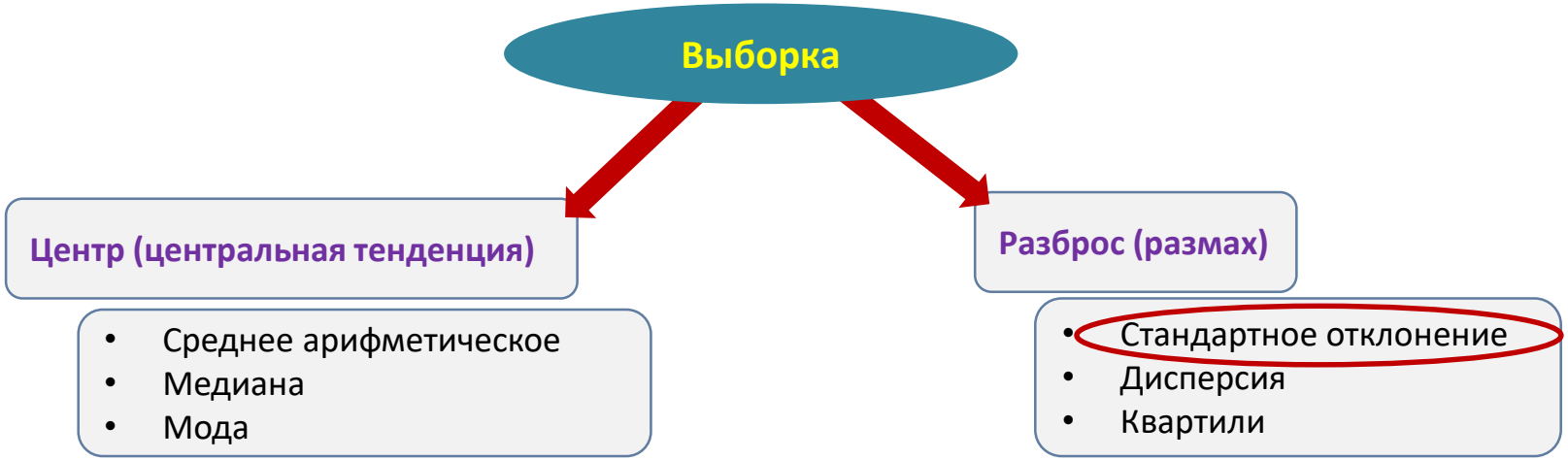
Повторность

- В выборке должно быть относительно много объектов из ГС
- Повторности должны быть независимы друг от друга. (Телефоны с конвейера должны быть из разных партий)

Рандомизация

- Каждый объект генеральной совокупности должен иметь равные шансы попасть в выборку

Характеристики выборки



- Стандартное отклонение – характеристика величины разброса значений выборки
- Стандартное отклонение показывает, как в выборке распределены значения относительно среднего значения

Расчет стандартного отклонения

1. Находим среднее значение выборки
2. Определяем разность между каждым значением и средним
3. Возводим разность в квадрат
4. Суммируем квадраты разностей
5. Делим сумму квадратов на (количество элементов в выборке – 1)
6. Извлекаем корень

Показатели выборки – функция summary()

```
>
> # регистрация роста сотрудников отдела
>
> h <- c(174, 162, 188, 192, 165, 168, 172.5)
> h
[1] 174.0 162.0 188.0 192.0 165.0 168.0 172.5
>
> # отсортируем выборку
> sort(h)
[1] 162.0 165.0 168.0 172.5 174.0 188.0 192.0
>
```

Структура данных – функция str()

```
>
> # узнаем структуру наших данных
> str(h)
 num [1:7] 174 162 188 192 165 ...
>
```

Общие показатели выборки функция summary()

```
>
> # узнаем общие показатели выборки h
> summary(h)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  162.0   166.5   172.5   174.5   181.0   192.0
>
```

минимальное значение

первый квартиль

медиана

среднее арифметическое

третий квартиль

максимальное значение

Примеры выборок с разными характеристиками

Задание имен значениям выборки

```
> # зададим имена каждому значению выборки h
> names(h) <- c("Николай", "Евгений", "Петр", "Александр", "Екатерина", "Василий", "Георгий")
> h
  Николай   Евгений      Петр Александр Екатерина   Василий   Георгий
    174.0     162.0     188.0     192.0     165.0     168.0     172.5
```

Добавим зарплаты сотрудников

```
> salary <- c(21, 19, 27, 11, 102, 25, 21)
> names(salary)<-c("Грузчик", "Курьер", "Менеджер", "Уборщица", "Директор","Бухгалтер","Экспедитор")
> salary
   Грузчик     Курьер   Менеджер   Уборщица   Директор  Бухгалтер Экспедитор
        21         19         27         11        102         25         21
```

Создание таблицы

```
> # создадим таблицу, объединив рост сотрудников и их зарплату
> company <-data.frame(h, names(salary), salary)
> names(company) <- c("Рост,см", "Должность", "Зарплата,тыс.руб")
> company
           Рост,см  Должность Зарплата,тыс.руб
Николай      174.0    Грузчик                21
Евгений      162.0     Курьер                19
Петр         188.0   Менеджер                27
Александр    192.0   Уборщица                11
Екатерина    165.0   Директор               102
Василий      168.0  Бухгалтер                25
Георгий      172.5 Экспедитор                21
```

Медиана и среднее арифметическое выборки

Общие показатели по таблице

```
> summary(company)
    Рост,см              Должность Зарплата,тыс.руб
 Min.   :162.0   Бухгалтер :1    Min.   : 11.00
 1st Qu.:166.5   Грузчик   :1    1st Qu.: 20.00
 Median :172.5   Директор  :1    Median : 21.00
 Mean   :174.5   Курьер    :1    Mean   : 32.29
 3rd Qu.:181.0   Менеджер  :1    3rd Qu.: 26.00
 Max.   :192.0   Уборщица  :1    Max.   :102.00
                 Экспедитор:1
```

медиана и среднее отличаются значительно

медиана и среднее близки

Сравнение выборок роста и зарплат

```
> sort(h)
  Евгений Екатерина    Василий    Георгий    Николай      Петр Александр
    162.0      165.0      168.0      172.5      174.0      188.0      192.0
>
> sort(salary)
  Уборщица     Курьер    Грузчик Экспедитор  Бухгалтер   Менеджер   Директор
        11         19         21         21         25         27        102
>
```

Расчет медианы. Медиана – центральная характеристика данных

Определение медианы

- Медиана – величина, находящаяся в центре ранжированной (отсортированной) выборки
- Медиана более устойчива (робастна) к выбросам и ошибочным данным по сравнению со средним арифметическим

Алгоритм вычисления медианы

- Выборка сортируется от меньшего значения к большему
- Пусть **N** – количество элементов выборки

  Если **N** нечетное:

  медиана = элементу с порядковым номером = (N + 1) / 2

  Если **N** четное:

  медиана = среднему арифметическому элементов с порядковыми номерами (N / 2) и (N / 2) +1

Свойства медианы

- медиана делит выборку пополам
- слева и справа от медианы находится одинаковое количество элементов

Пример. Рассчитать медиану для выборок

7, 3, 12, 21, 9, 5, 5, 17, 14
2, 7, 3, 9, 6, 7
16, 4, 17, 14, 11, 9, 30, 6

Расчет медианы в R: median(числовой вектор)

Центральные характеристики распределения данных. Мода

- Мода – значение в выборке, которое встречается чаще всего
- Мода, как правило, применяется для номинальных данных

Примеры номинальных данных
- Пол
- Цвет
- Ответы «Да» / «Нет»
- Наличие / Отсутствие
- Варианты выбора из заданного множества

```
>
> gender <- c("male", "female", "male", "male", "female", "male", "male")
> t.gender <- table(gender)
> t.gender
gender
female   male
     2      5
>
> mode <-t.gender[which.max(t.gender)]
> mode
male
   5
>
```

Расчет стандартного отклонения

Расчет стандартного отклонения

1. Находим среднее значение выборки
2. Определяем разность между каждым значением и средним
3. Возводим разность в квадрат
4. Суммируем квадраты разностей
5. Делим сумму квадратов на (количество элементов в выборке – 1)
6. Извлекаем корень

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - x_{ср})^2}{n-1}}$$

```
>
> nh <- length(h)
> mean.h <- sum(h)/nh
> s.h <- sum((h - mean.h)^2)
> sd.h <- sqrt(s.h / (nh-1))
> mean.h
[1] 174.5
> s.h
[1] 781.5
> sd.h
[1] 11.41271
>
```

```
x<-sample(15)
x
sd(x)
```

```
> sd(h)
[1] 11.41271
>
```

Расчет стандартного отклонения в R: sd(числовой вектор)

Стандартное отклонение. Пример использования

							Среднее
Остатки по неделям, склад № 1	33	31	32	36	31	31	32,3
Остатки по неделям, склад № 2	22	34	58	52	10	21	32,8

							Среднее	Станд откл
Склад № 1	33	31	32	36	31	31	32,3	2,0
Склад № 2	22	34	58	52	10	21	32,8	18,9

работа менеджмента на складе № 2 значительно хуже, чем на складе № 1;
склад № 2 «лихорадит»

Расчет в R:

```
> store1 = c(33, 31, 32, 36, 31, 31)
> store2 = c(22, 34, 58, 52, 10, 21)
>
> mean(store1)
[1] 32.33333
> mean(store2)
[1] 32.83333
>
> median(store1)
[1] 31.5
> median(store2)
[1] 28
>
> sd(store1)
[1] 1.966384
> sd(store2)
[1] 18.87238
```

Дисперсия?

d – дисперсия

sd – стандартное отклонение

d = sd * sd

Расчет в R: var(вектор)

Характеристики разброса данных: квартили, процентили, IQR

Определение квартиля

Квартили (кварта, четверть) – значения, которые делят отсортированную выборку на четыре группы, содержащие приблизительно равное количество наблюдений.

Общий объем выборки делится на четыре равные части: 25%, 50%, 75% 100%.

- Первый (нижний) квартиль **Q1** отсекает слева 25 % выборки
- Второй (средний) квартиль **Q2** отсекает слева 50 % выборки
- Третий (верхний) квартиль **Q3** отсекает слева 75 % выборки

Определение процентиля

n-й процентиль - это значение, ниже которого расположено **n** процентов элементов выборки

Первый процентиль – значение, ниже которого располагается **1** процент элементов выборки

25-й процентиль совпадает с первым квартилем **Q1**

90-й процентиль – значение, ниже которого расположено **90** процентов всей выборки

Определение IQR

Интерквартильный размах (интервал, отрезок) **InterQuartile Range**
IQR: IQR = Q3–Q1

IQR —интервал, содержащий центральные 50% наблюдений выборки, т.е. интервал между 25-м и 75-м процентилями

Расчет квартилей: два способа

1 способ. Простой, грубый

1. Сортируем выборку от меньших значений к большим
2. Находим медиану
3. Рассматриваем левую половину выборки. Находим медиану, это **Q1**
4. Рассматриваем правую половину выборки. Находим медиану, это **Q3**

Пример 1. y <- c (1, 7, 4, 3)

Отсортированная выборка = (1, 3, 4, 7)

Медиана = (3 + 4) / 2 = 3.5

Первый квартиль Q1 = медиана выборки (1, 3) = (1 + 3) / 2 = 2

Третий квартиль Q3 = медиана выборки (4, 7) = (4 + 7) / 2 = 5.5

Значения квартилей не совпадают

Расчет в R

```
>
> y<-c(1,7,4,3)
> summary(y)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1.00    2.50    3.50    3.75    4.75    7.00
>
```

Точный алгоритм расчета квартилей

1. Сортируем выборку от меньших значений к большим
2. Каждому значению выборки сопоставляем действительное число **f** в диапазоне от 0 до 1 по формуле **f(i) = (i – 1) / (N – 1)**, где
   i – номер элемента в выборке
   N – общее число элементов выборки
3. Первый квартиль **Q1,** соответствующий **f** = 0.25, вычисляется по двум соседним **f**-значениям, находящимся ниже и выше 0.25
4. Второй квартиль **Q2,** соответствующий **f** = 0.5 вычисляется по двум соседним **f**-значениям, находящимся ниже и выше 0.5
5. Третий квартиль **Q3** ,соответствующий **f** = 0.75 вычисляется по двум соседним **f**-значениям, находящимся ниже и выше 0.75

Любой процентиль рассчитывается аналогично

Пример 2. Дана выборка

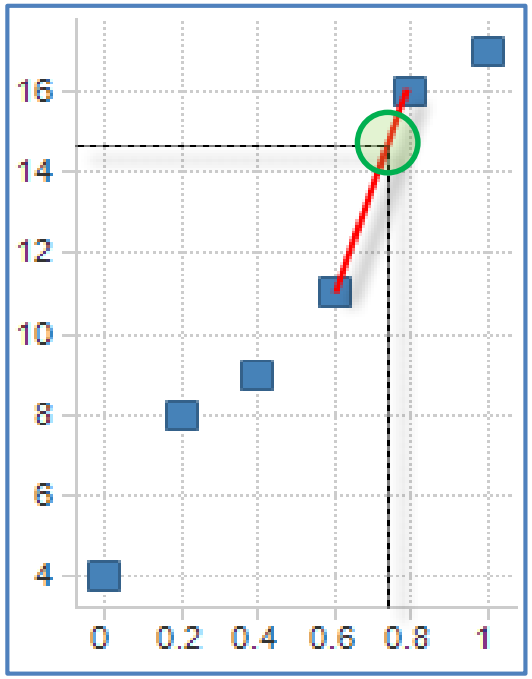
11, 17, 8, 16, 4, 9

- **Рассчитать таблицу f-значений**
- **Найти Q1 и Q3**
- **Определить IQR**

Решение примера 2

Выборка	f-значение
4	0
8	0,2
9	0,4
11	0,6
16	0,8
17	1

Поиск **Q3**



Для нахождения точки пересечения отрезка с координатами [**(x = 0,6; y = 11), (x=0,8; y = 16)**] с прямой **x=0,75** используем формулу линейной интерполяции

$$y = Y_1 + \frac{Y_2 - Y_1}{X_2 - X_1}(x - X_1)$$

X1	0,6
X2	0,8

Y1	11
Y2	16

Ответ

Q3 = 14.75

Q1 = 8.25

IQR = 6.5

```
> z<- c(11, 17, 8, 16, 4, 9)
> summary(z)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   4.00    8.25   10.00   10.83   14.75   17.00
>
```

Пример 3

Пример. Дана выборка

30, 44, 4, 1, 52, 17, 19, 35, 41, 8, 11

- **Рассчитать таблицу f-значений**
- **Найти Q1, Q2, Q3**
- **Определить IQR**

Выборка	1	4	8	11	17	19	30	35	41	44	52
f-значение	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1

$$y = Y_1 + \frac{Y_2 - Y_1}{X_2 - X_1}(x - X_1)$$

Ответы

Q1 = 9.5

Q2 = 19

Q3 = 38

IQR = 28.5

Правильное решение примера 1

Пример 1. Дана выборка 1, 3, 4, 7

- Рассчитать таблицу f-значений
- Найти Q1, Q2, Q3
- Определить IQR

Выборка (Y)	f-значение (X)
1	0
3	1/3
4	2/3
7	1

Решение

Q1 = 1 + (3 – 1) / (1/3 - 0) * (0.25 – 0) = 2.5

Q2 = 3 + 1/ (1/3) * (0.5 – 1/3) = 3 + 3/2 – 1= 3.5

Q3 = 4 + (7 – 4)/(1 – 2/3) * (0.75 – 2/3) = 4 + 27 /4 – 6 = (27 - 8) / 4 = 4.75

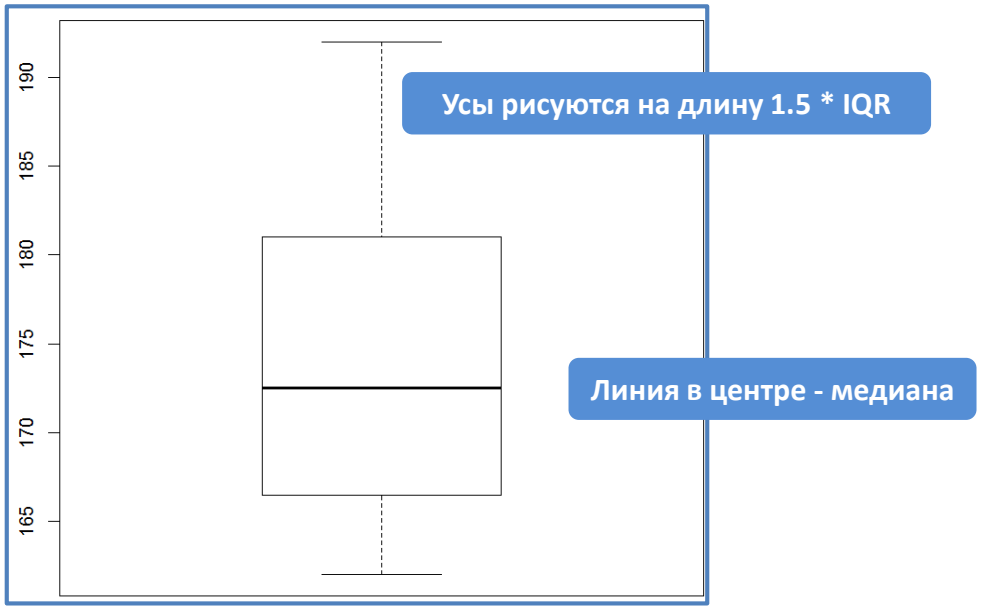
```
>
> y<-c(1,7,4,3)
> summary(y)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1.00    2.50    3.50    3.75    4.75    7.00
>
```

Графическое представление центральной характеристики и разброса

Характеристики выборок роста и зарплат

```
> summary(h)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  162.0   166.5   172.5   174.5   181.0   192.0
> summary(salary)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  11.00   20.00   21.00   32.29   26.00  102.00
>
```

Ящик с усами (боксплот) для выборки h

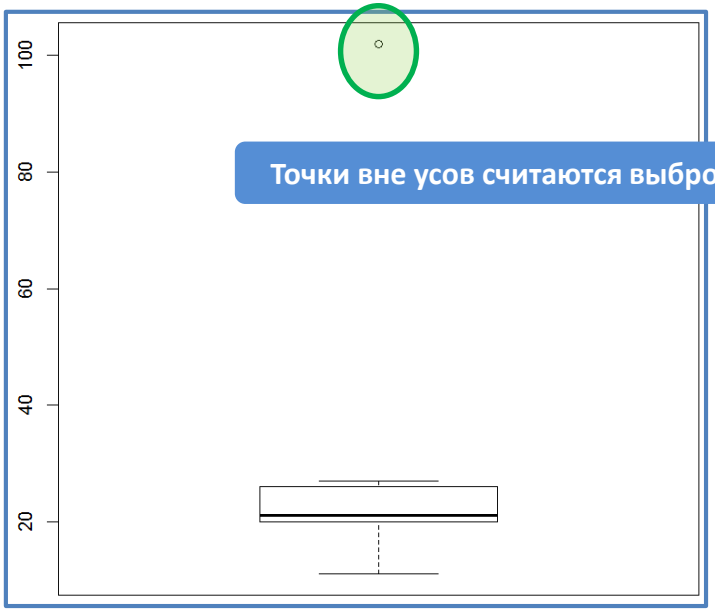


Усы рисуются на длину 1.5 * IQR

Линия в центре - медиана

Высота ящика = IQR

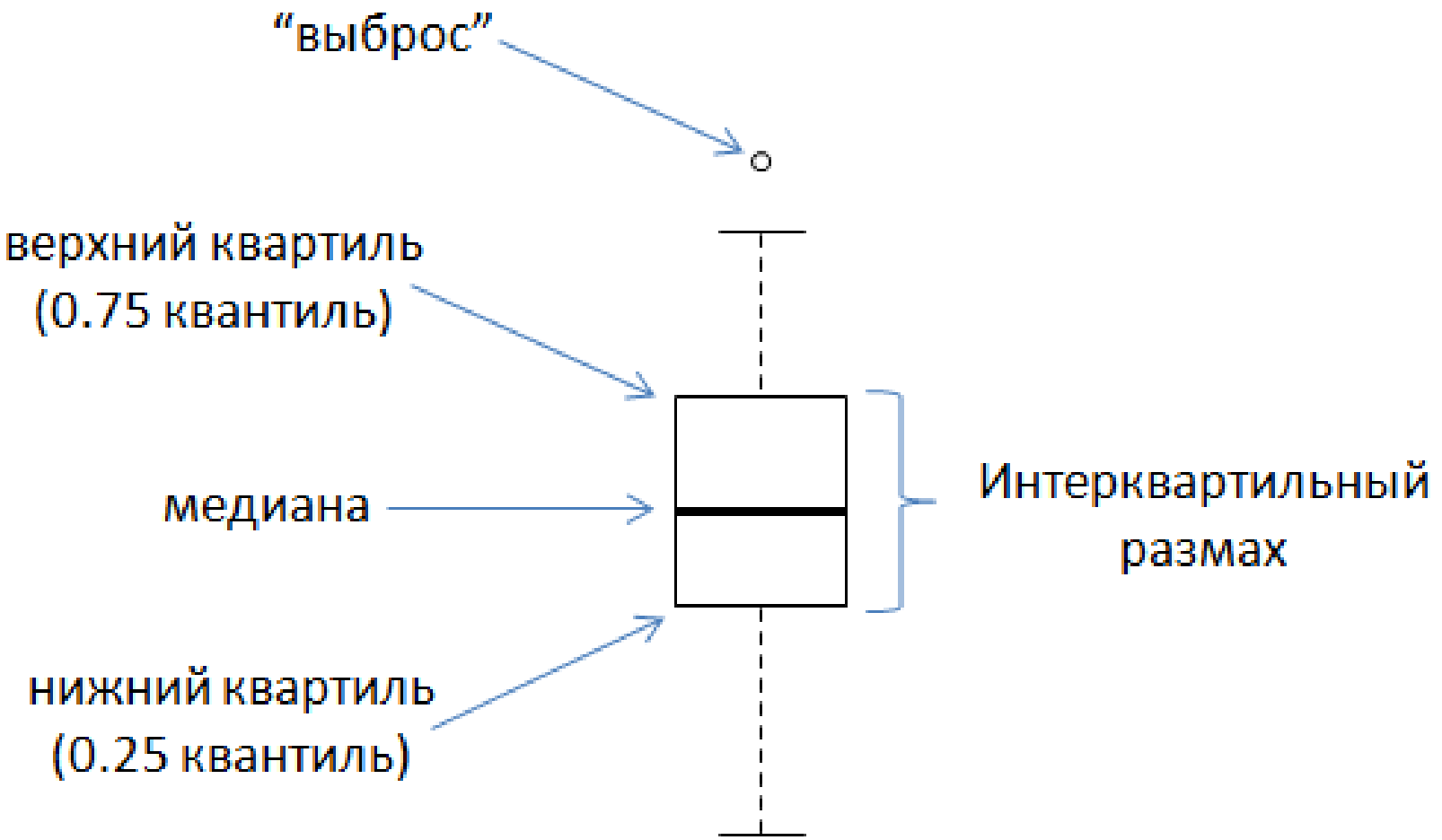
Ящик с усами (боксплот) для выборки salary



Точки вне усов считаются выбросами

Функция boxplot() – диаграмма размаха

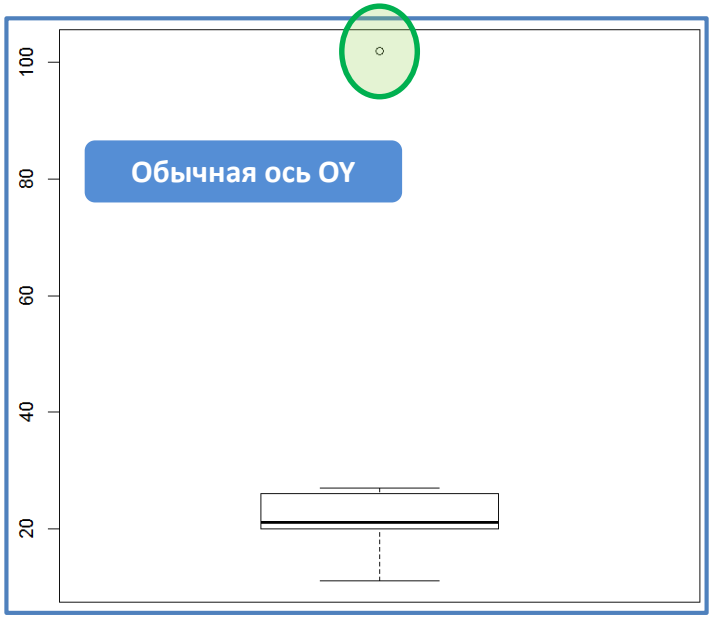
Диаграмма размаха (ящик с усами, изобретен 50 лет назад) – график, использующийся в описательной статистике, отображающий одномерное распределение вероятностей некоторой выборки данных

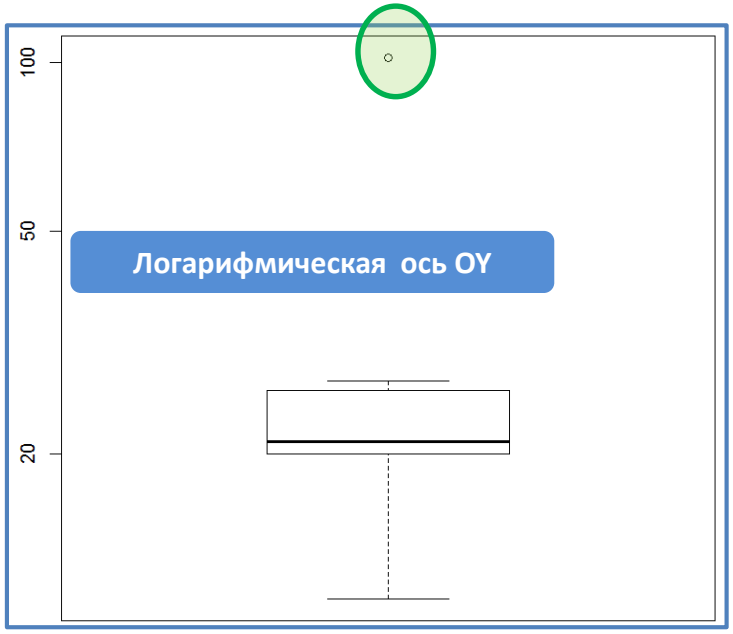


Команды для рисования боксплота

```
>
> boxplot(h)
>
> boxplot(salary)
> boxplot(salary, log="y")
>
```

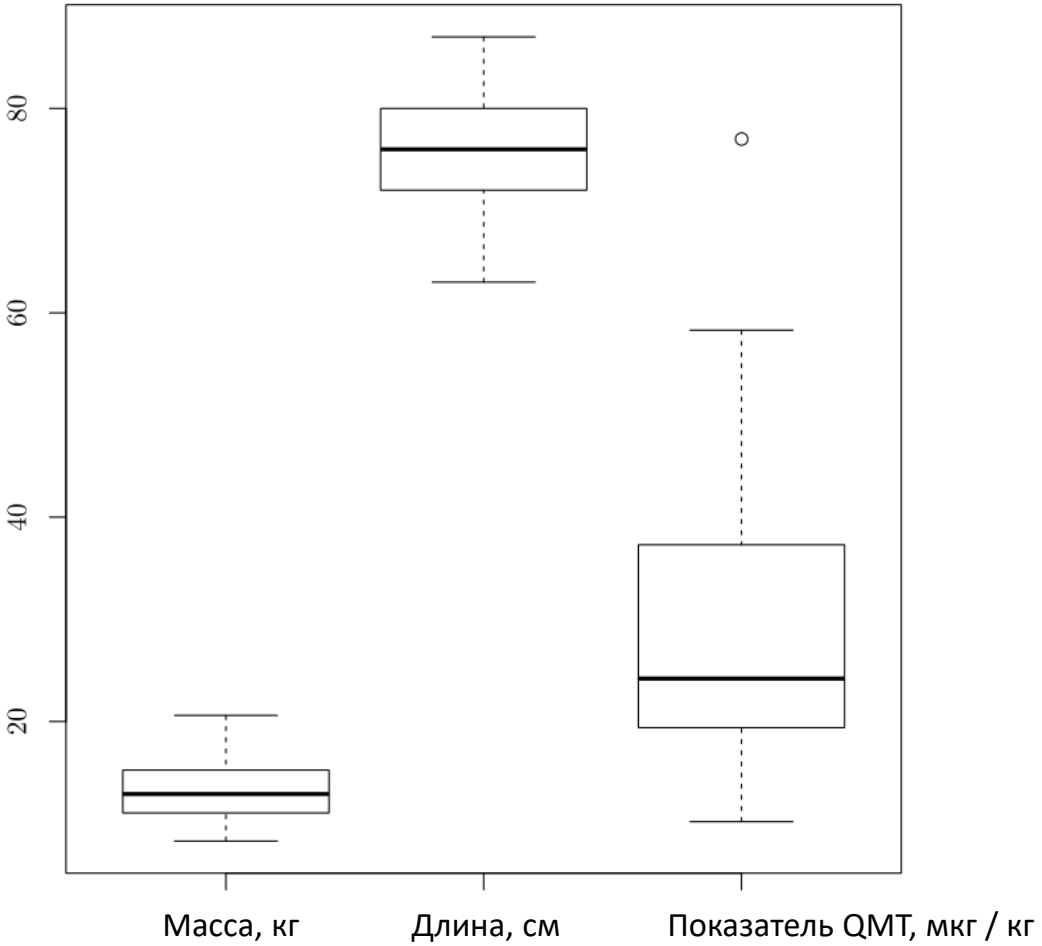
Ящики с усами (боксплоты) для выборки salary





Варианты размещения боксплотов

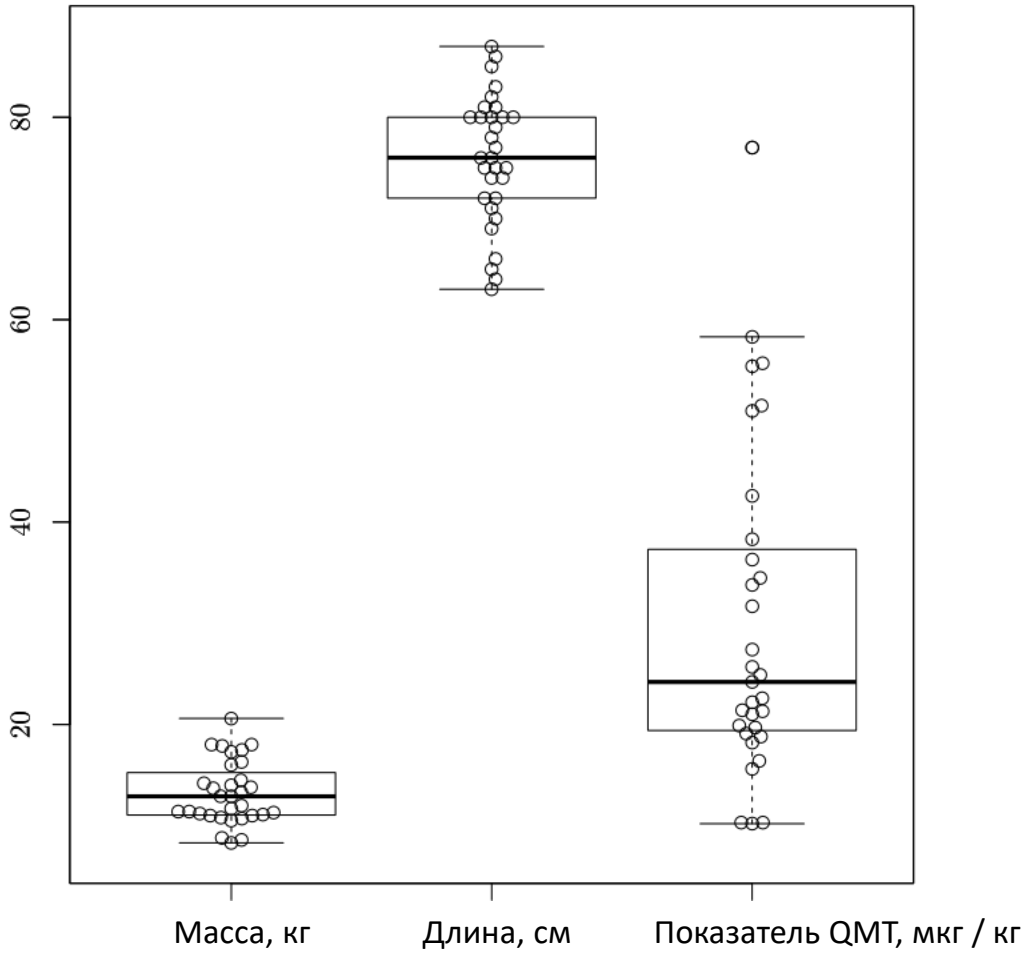
Представление данных, имеющих разный масштаб и ед. измерения



Результаты исследования диких обезьян, пораженных вирусом Quarella Mourus
атолл Баа, 2018

Варианты размещения боксплотов

Бокспоты -ульи



Результаты исследования диких обезьян, пораженных вирусом Quarella Mourus
атолл Баа, 2018

Рисуем усы у boxplot

```
y<-c(1,7,4,3)
boxplot(y)
```

```
y<-c(11, 17, 8, 16, 4, 9)
boxplot(y)
```

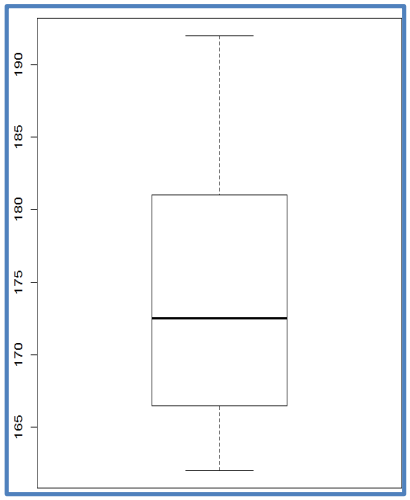
```
y<-c(30, 44, 4, 1, 52, 17, 19, 35, 41, 8, 11)
boxplot(y)
```

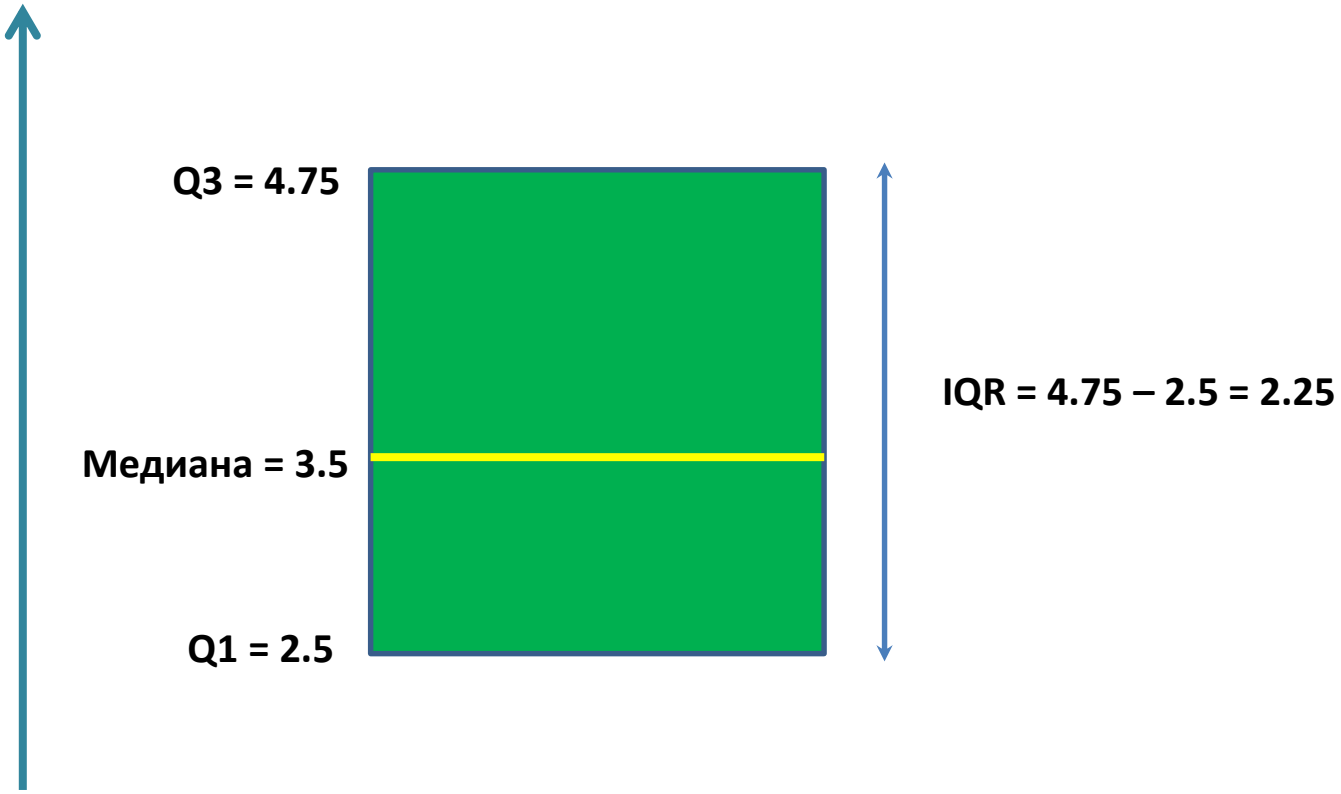
Все значения выборки находятся внутри диапазона [L1, L2]

[L1, L2] = Q3 + IQR * 1.5 – (Q1 – IQR * 1.5)

[L1, L2] = Q3 – Q1 + IQR * 3 = 4 * IQR

Начинаем рисовать boxplot



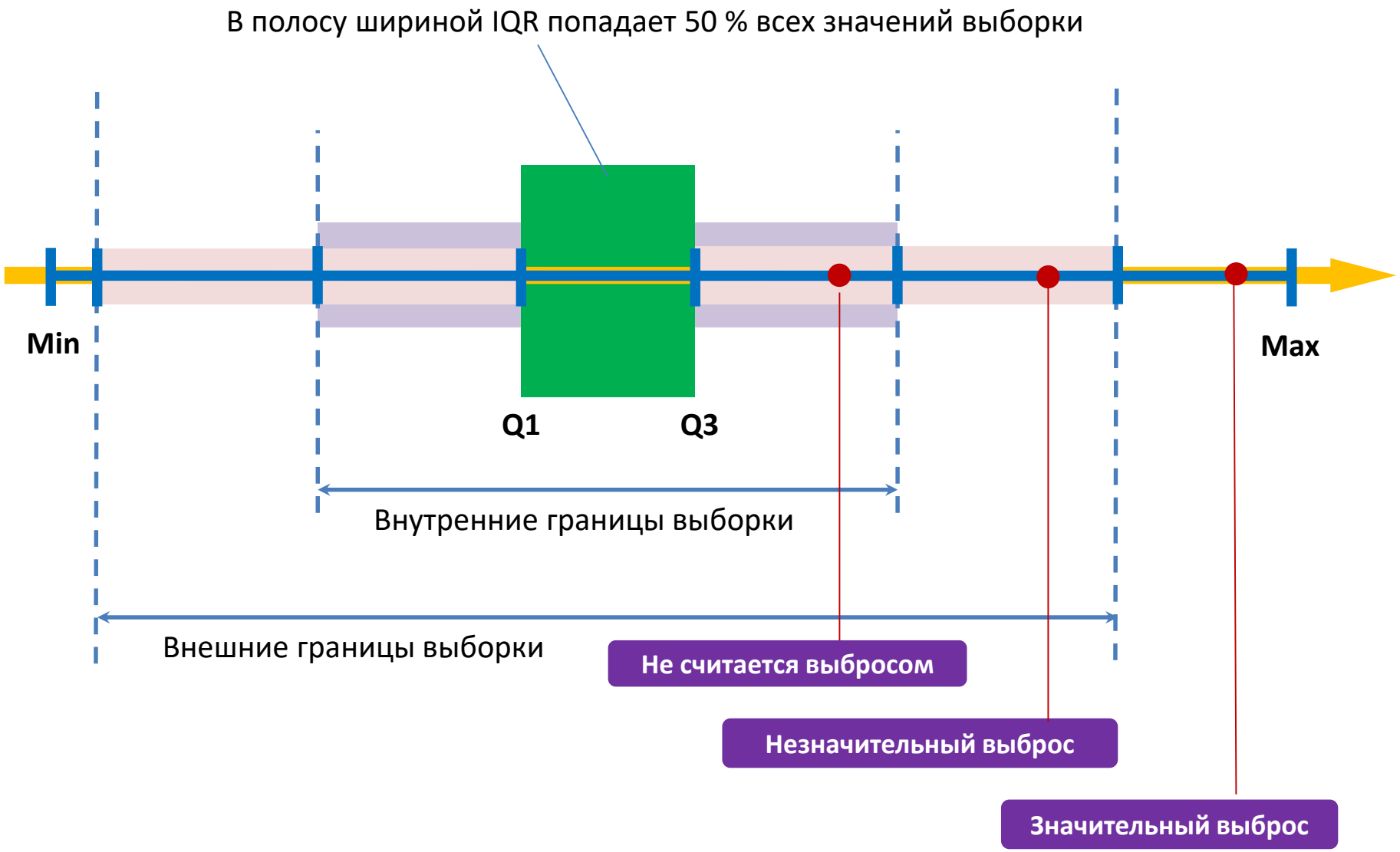


Вопросы

Что такое усы у boxplot ?
В каких случаях усы есть, в каких случаях их нет?
Как их рисовать?
Что такое выбросы?
Как их рисовать?

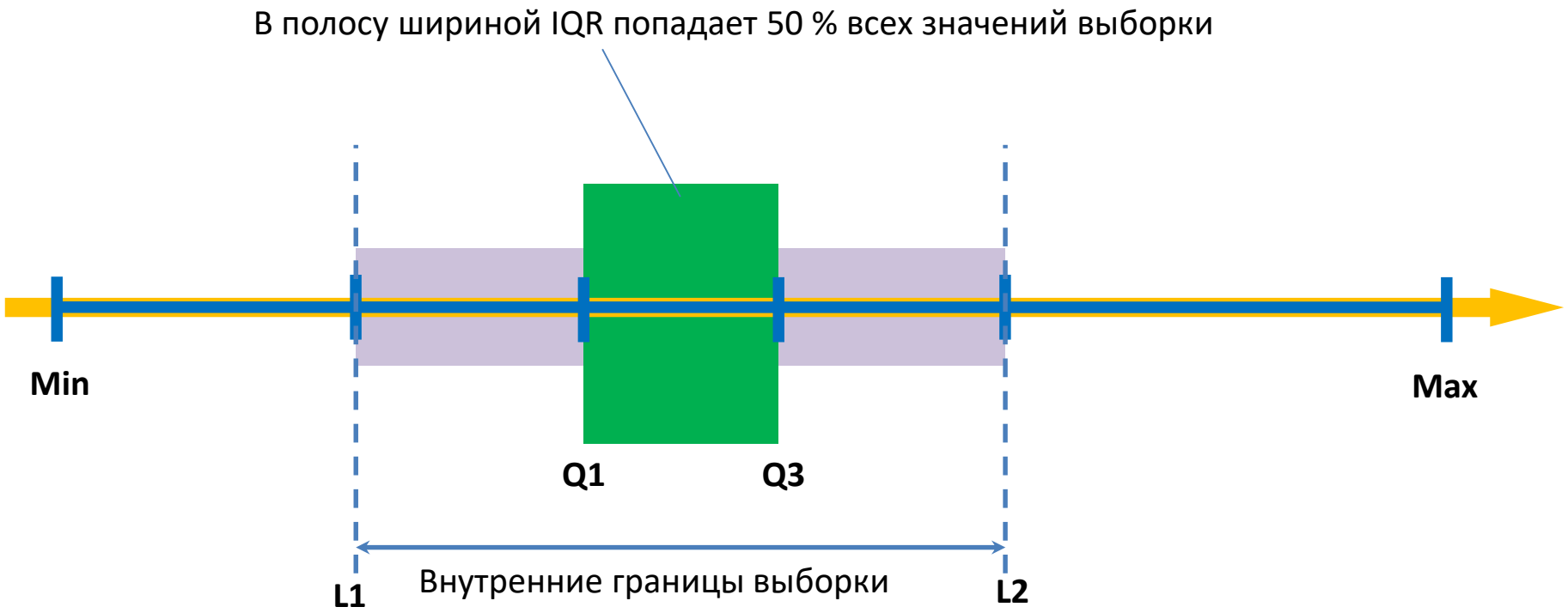
Внутренние и внешние границы выборки

Диапазон значений выборки



Внутренние границы выборки – усы у boxplot

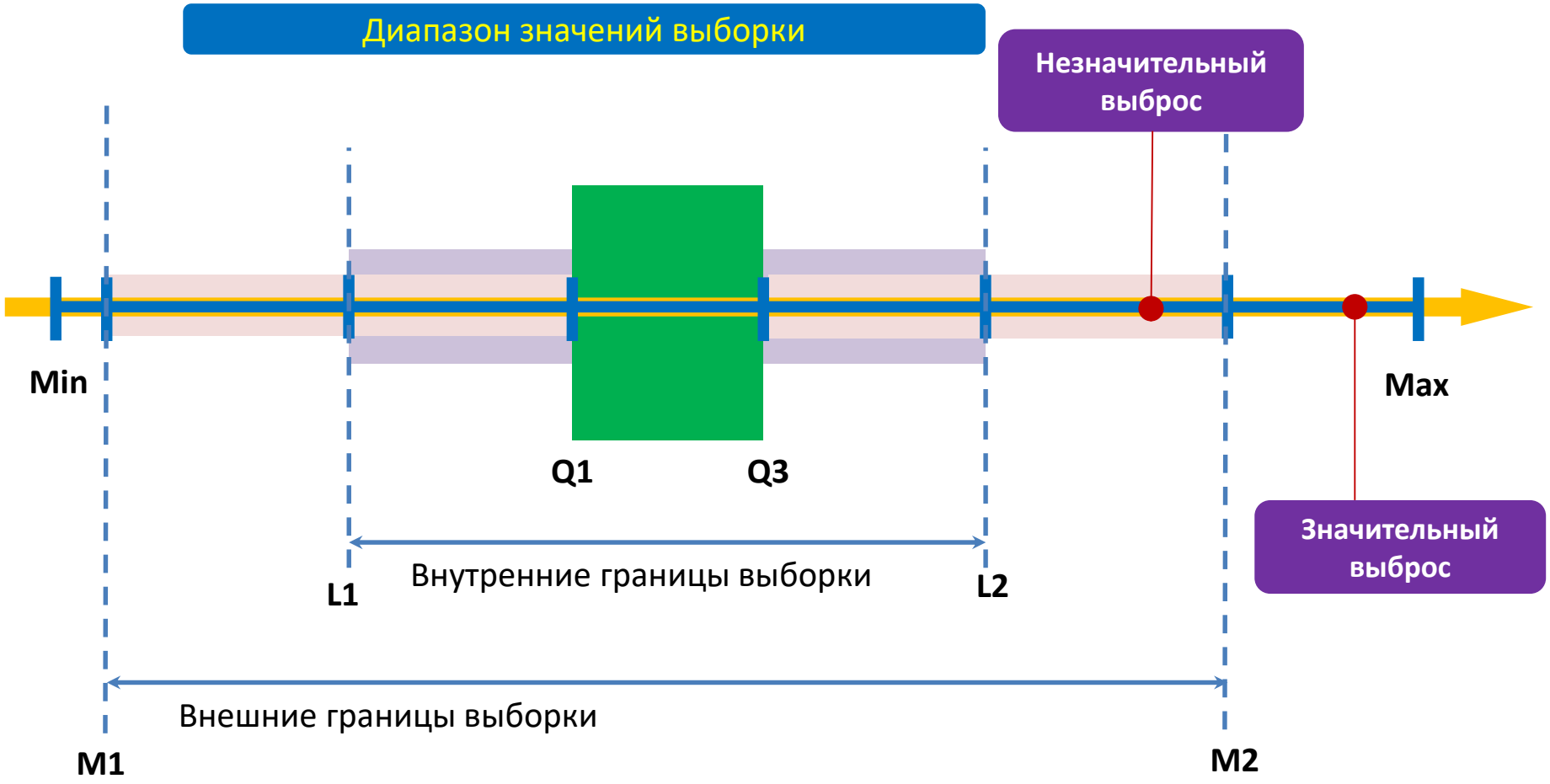
Диапазон значений выборки



Ширина прямоугольника = IQR * 1.5
L1 = Q1 – IQR * 1.5
L2 = Q3 + IQR * 1.5

Все значения из диапазонов [L1, Q1], [Q3, L2] рисуются усами

Внешние границы выборки. Выбросы



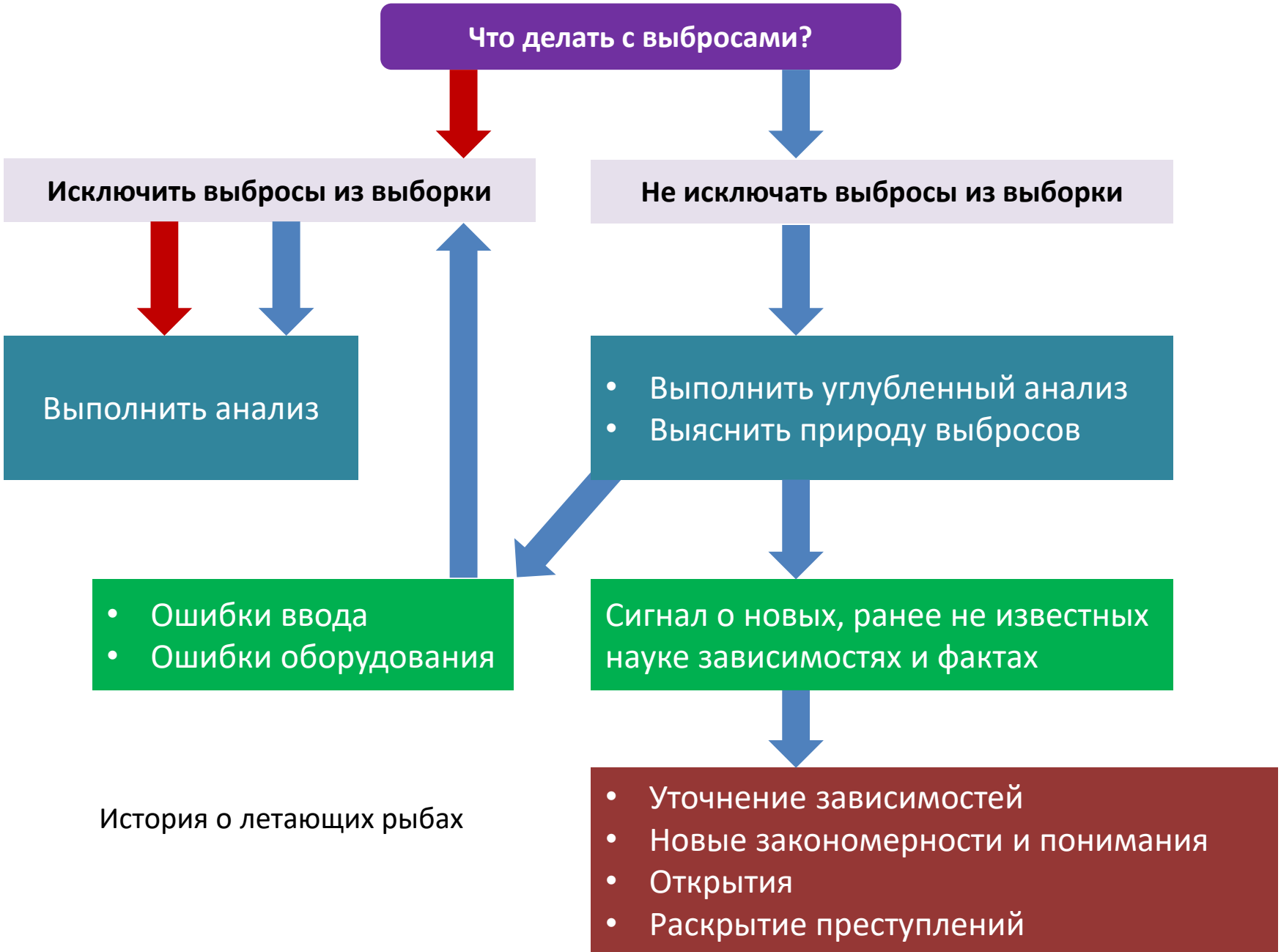
Ширина прямоугольника = IQR * 3

M1 = Q1 – IQR * 3

M2 = Q3 + IQR * 3

Все значения из диапазонов [Min, L1], [L2, Max] рисуются точками

Природа выбросов



История о летающих рыбах