RUDN university

5100

Программирование в среде R

Шевцов Василий Викторович,
директор ДИТ РУДН, shevtsov_vv@rudn.university

Набор данных

```
dreissena <- read.delim(
          "http://files.figshare.com/1360878/Dreissena.txt")
```

- Month - качественная переменная с тремя уровнями, соответствующими времени отбора проб дрейссены: May (май), July (июль) и September (сентябрь);
- Day — день отбора проб (с даты начала проведения исследований);
- Lake — качественная переменная с тремя уровнями, обозначающими изученные озера: Batorino, Myastro и Naroch;
- Site — качественная переменная с девятью уровнями, обозначающими места отбора проб (SI - S9). В каждом озере моллюсков собирали на трех постоянных станциях;
- Length — длина раковины моллюсков (мм);
- Infection — количество инфузорий, обнаруженных в каждом моллюске («интенсивность инвазии», «уровень инвазии»).

Набор данных

```
> dreissena <- read.delim("http://files.figshare.com/1360878/Dreissena.txt")
> str(dreissena)
'data.frame':   476 obs. of  6 variables:
 $ Month    : Factor w/ 3 levels "July","May","September": 2 2 2 2 2 2 2 2 2 2 ...
 $ Day      : int  1 1 1 1 1 1 1 1 1 1 ...
 $ Lake     : Factor w/ 3 levels "Batorino","Myastro",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ Site     : Factor w/ 9 levels "S1","S2","S3",..: 3 3 3 3 3 3 3 3 3 3 ...
 $ Length   : num  14.9 14 13 14 12 14 12 19 16.5 18 ...
 $ Infection: int  36 30 331 110 4 171 31 887 525 497 ...
```

	Month	Day	Lake	Site	Length	Infection
1	May	1	Batorino	S3	14.9	36
2	May	1	Batorino	S3	14.0	30
3	May	1	Batorino	S3	13.0	331
4	May	1	Batorino	S3	14.0	110
5	May	1	Batorino	S3	12.0	4
6	May	1	Batorino	S3	14.0	171
7	May	1	Batorino	S3	12.0	31
8	May	1	Batorino	S3	19.0	887
9	May	1	Batorino	S3	16.5	525
10	May	1	Batorino	S3	18.0	497
11	May	1	Batorino	S3	19.0	56

функция qplot()

Аргументы функции qplot()

- x и y — переменные X и Y соответственно;
- data — таблица данных («data frame» в терминах R), содержащая переменные X и Y. Если этот аргумент не указан, то функция qplot () попытается автоматически извлечь векторы x и y из текущей рабочей среды и объединить их в таблицу;
- facets - формула, определяющая способ разбиения рисунка на отдельные подобласти при создании категоризованных графиков;
- margins — аргумент, используемый при создании категоризованных графиков. Позволяет включать (TRUE) или отключать (FALSE) отображаемые по краям графика названия уровней качественной переменной, в соответствии с которыми рисунок разбивается на подобласти;
- geom — текстовый вектор с названиям геометрических объектов, используемых для изображения данных. Если на функцию qplot () поданы две переменные - X и Y, то аргумент geom по умолчанию примет значение "point" («точка»). Если же подана только количественная переменная Y, то значением по умолчанию будет "histogram" («гистограмма»). Возможно совмещение нескольких типов геометрических объектов на одном рисунке;

Аргументы функции qplot()

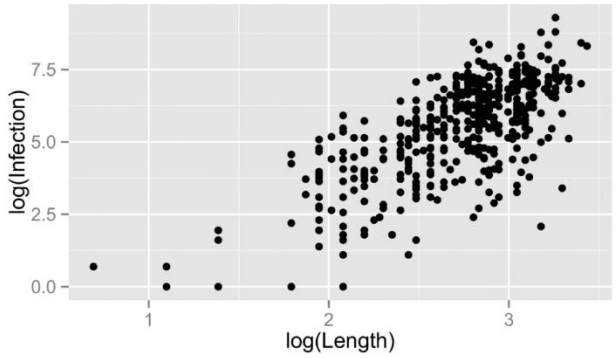
- stat — текстовый вектор, определяющий тип статистического преобразования данных;
- xlim и ylim — задают границы значений переменных X и Y соответственно (в виде c(нижняя граница, верхняя граница));
- log -- позволяет логарифмически «растянуть» ось X (log = "x"), ось Y (log = "y"), или обе оси одновременно (log = "xy");
- main — текстовый вектор и (или) математическое выражение, образующие заголовок графика;
- xlab и ylab — текстовые векторы и (или) математические выражения, образующие подписи осей X и Y соответственно;
- asp — число, задающее отношение длины X к длине оси Y.

Источник данных

X и Y

dataFrame

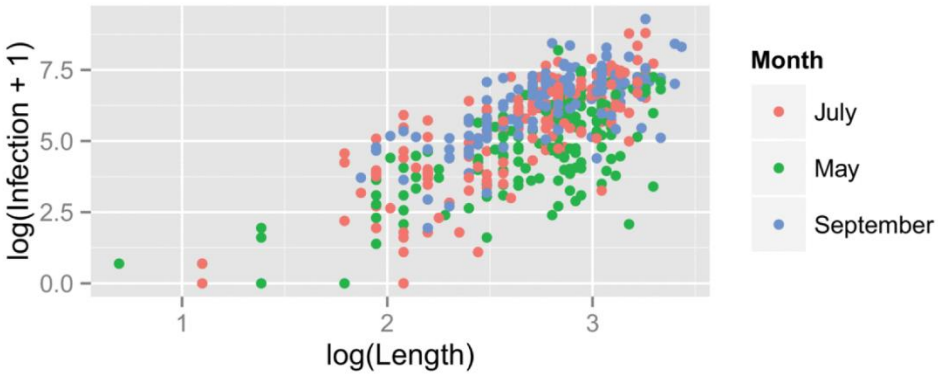
Попытка объединить векторы и преобразовать в таблицу

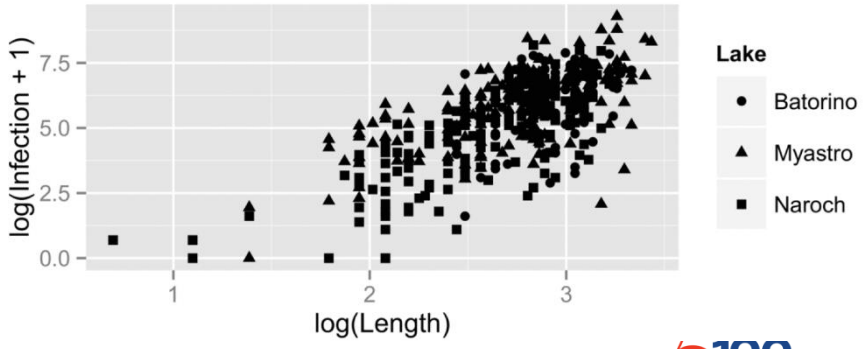


Отличие qplot(){ggplot2} от plot(){base}

Метод присвоения эстетических атрибутов, т. е. цвет, размер и форма. В случае с plot () пользователь должен самостоятельно конвертировать уровни интересующей его качественной переменной (например, «зима», «весна», «лето», «осень») в соответствующие значения эстетических атрибутов (например, цвет для разных сезонов года: «белый», «голубой», «зеленый», «оранжевый»). Функция же qplot() выполняет такие преобразования автоматически, одновременно создавая легенду с цветовой шкалой, которую пользователь может изменить в соответствии со своими требованиями.

```
qplot(log(Length), log(Infection + 1), data = dreissena,
      colour = Month)
qplot(log(Length), log(Infection + 1), data = dreissena,
      shape = Lake)
```





Функция I()

```
qplot(log(Length), log(Infection + 1), data = dreissena,
      colour = Month)
qplot(log(Length), log(Infection + 1), data = dreissena,
      shape = Lake)
```

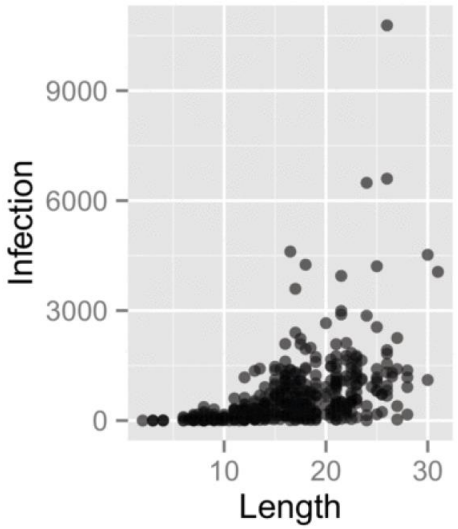
Автоматическое присвоение эстетических атрибутов можно отменить с помощью функции I()

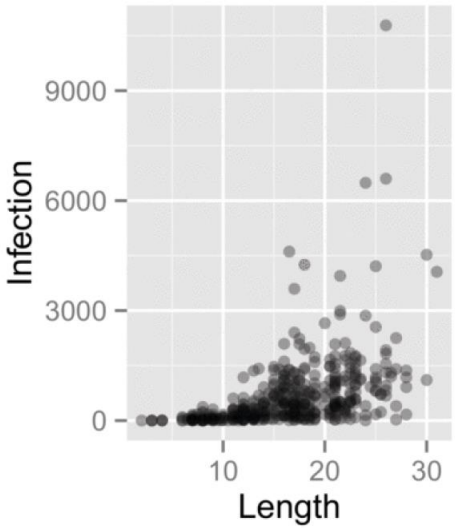
I() подавляет любые преобразования аргументов, сохраняя их исходный класс

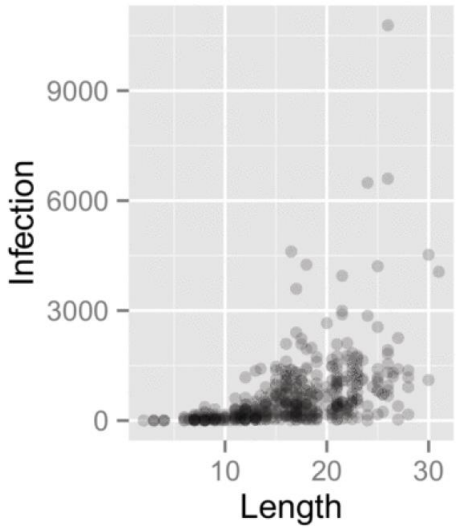
```
colour = I("red")
shape = I(2))
```

Прозрачность

```
qplot(Length, Infection, alpha = I(1/2), data = dreissena)
qplot(Length, Infection, alpha = I(1/4), data = dreissena)
qplot(Length, Infection, alpha = I(1/8), data = dreissena)
```







geom определяет тип геометрических объектов

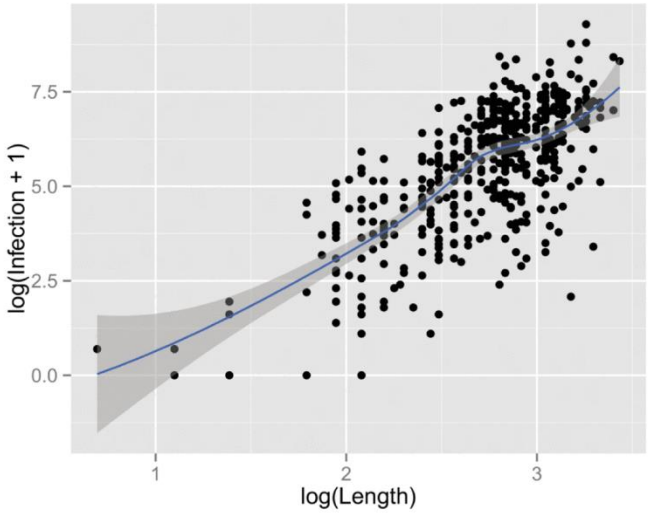
- geom = "point" — изображает данные в виде точек
- geom = "smooth" — подгоняет сглаживающую кривую к данным и одновременно изображает ее 95%-ную доверительную область;
- geom = "jitter" — создает одномерные диаграммы рассеяния;
- geom = "boxplot" — создает диаграммы размахов;
- geom = "path" и geom = "line" — соединяют точки линиями. Традиционно используются для изображения временных изменений количественных переменных (geom = "line"). Однако точки могут соединяться не только в соответствии с ходом времени, т. е. слева направо, но и любым другим образом (geom = "path").
- При анализе свойств только одной переменной выбор возможных значений аргумента geom будет определяться типом этой переменной:
- количественные переменные: значение geom = "histogram" приведет к созданию гистограммы, geom = "freqpoly" - полигона распределения частот, а geom = "density" — кривой плотности вероятности;
- качественные переменные: значение geom = "bar" приведет к созданию столбиковой диаграммы.

Линии тренда

```
qplot(log(Length), log(Infection + 1),
      geom = c("point", "smooth"), data = dreissena)
```

smooth – сглаживающая линия

добавление производится в параметре geom путем объединения двух типов в векторе



Одномерная диаграмма рассеяния

Инструмент для визуализации значений какой-либо количественной переменной в соответствии с уровнями качественной переменной. Для создания в ggplot2 служит геометрический объект типа "jitter".

```
qplot(Lake, log(Infection + 1), data = dreissena,
      geom = "jitter", alpha = I(0.6))

qplot(Lake, log(Infection + 1), data = dreissena,
      geom = "jitter", alpha = I(0.6), colour = Month)
qplot(Lake, log(Infection + 1), data = dreissena,
      geom = "jitter", alpha = I(0.6), colour = Month,
      size = Length)
```

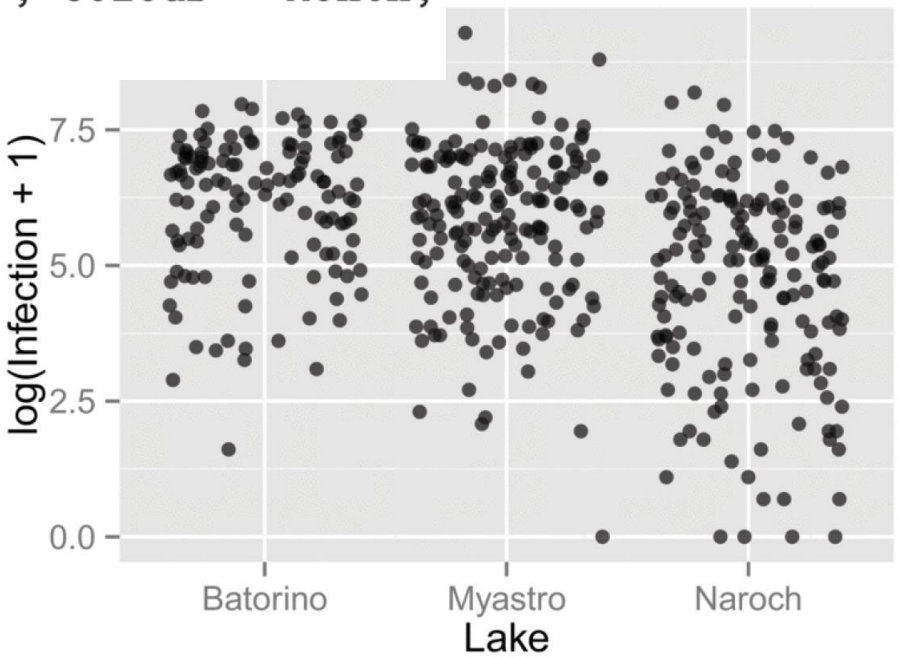
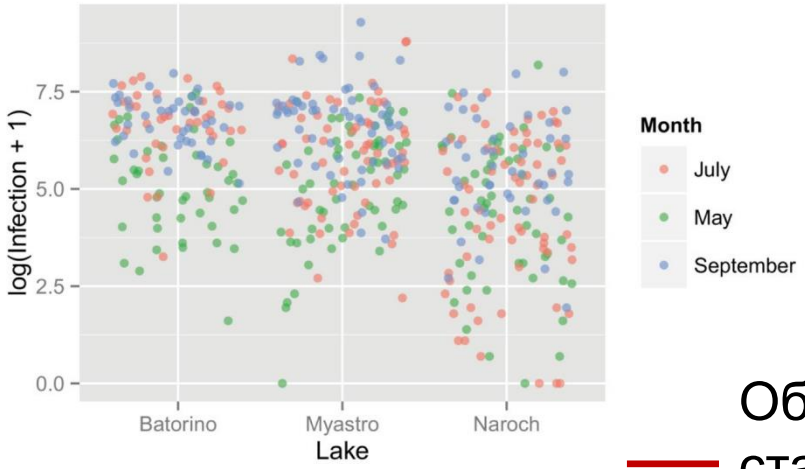
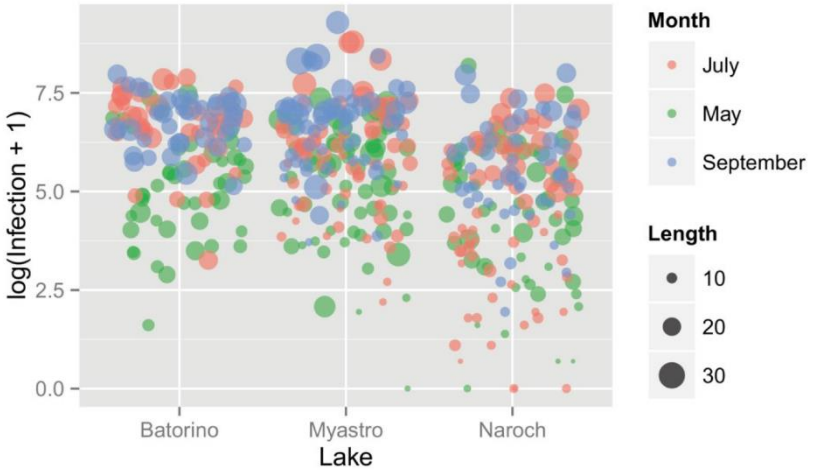


Диаграмма размахов





Обобщенная статистическая информация

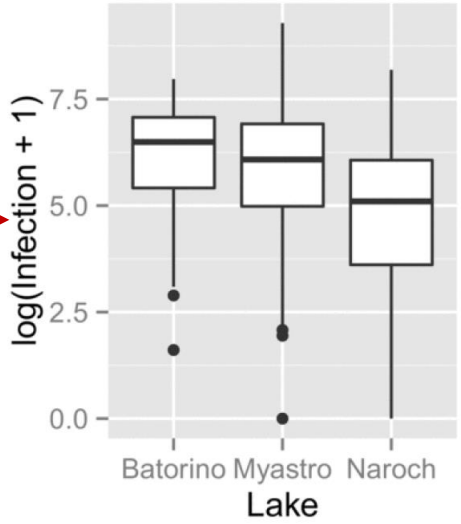


Диаграмма размахов

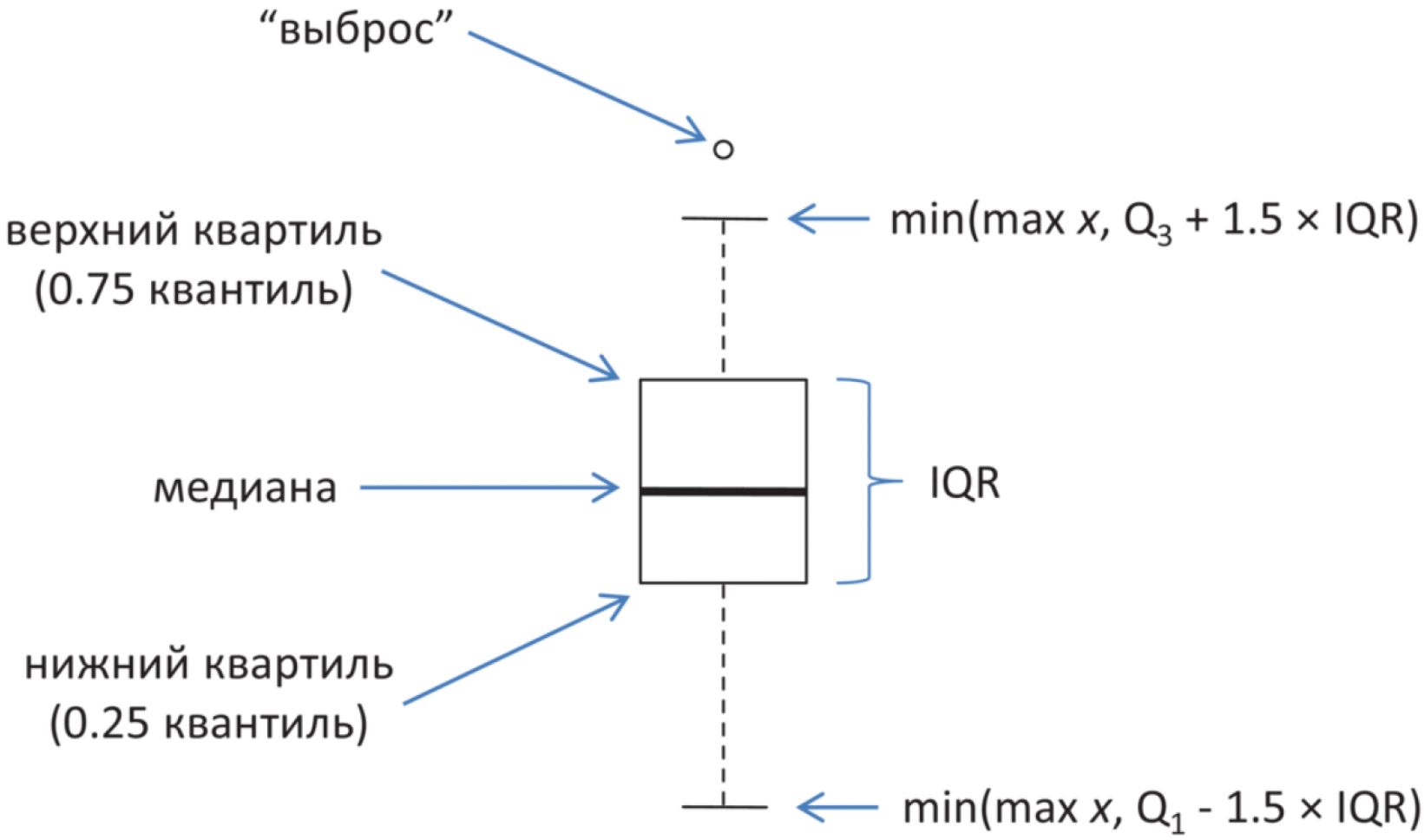
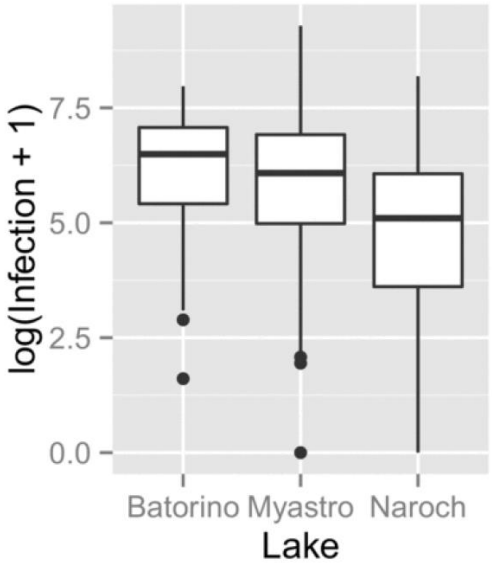
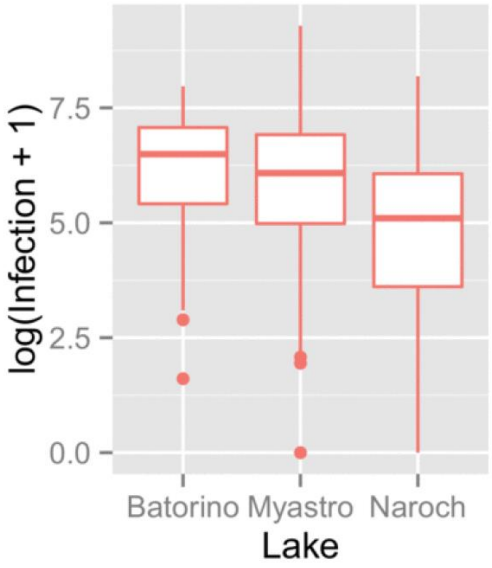
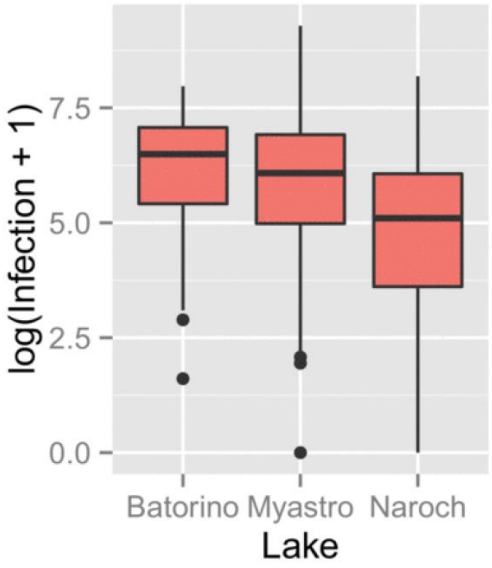


Диаграмма размахов

```
qplot(Lake, log(Infection + 1), data = dreissena,
      geom = "boxplot")
qplot(Lake, log(Infection + 1), data = dreissena,
      geom = "boxplot", colour = "red")
qplot(Lake, log(Infection + 1), data = dreissena,
      geom = "boxplot", fill = "coral")
```



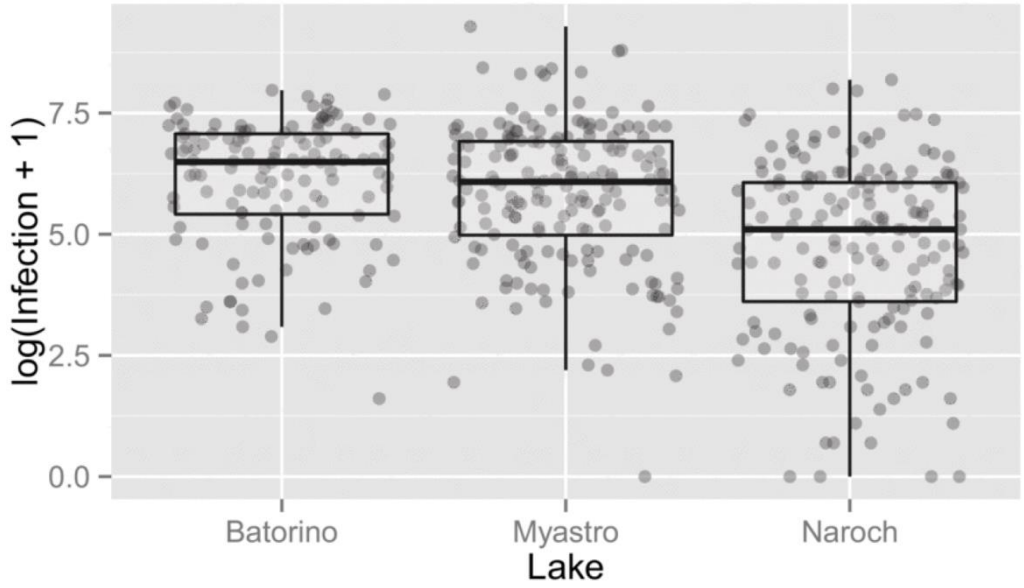




Совмещение диаграмм

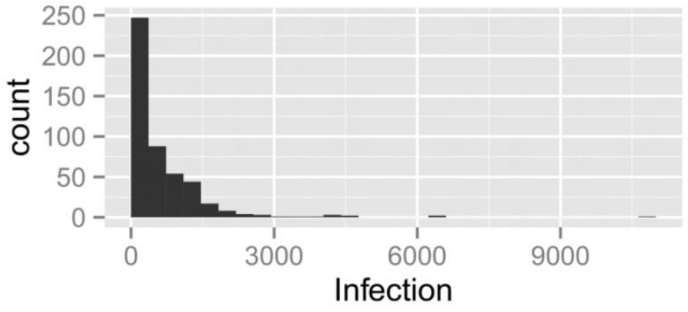
Совмещение диаграммы размахов с одномерной диаграммой рассеивания
Объединение производится в векторе

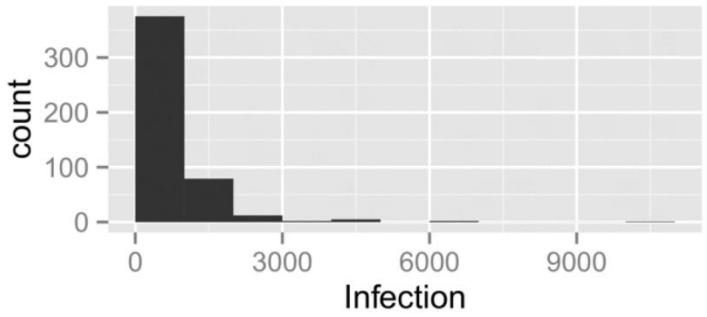
```
qplot(Lake, log(Infection + 1), data = dreissena,
      geom = c("jitter", "boxplot"), alpha = I(1/5),
      outlier.colour = NA)
# отображение "выбросов" отключено
# во избежание дублирования точек, являющихся
# одновременно частью диаграммы рассеяния
```

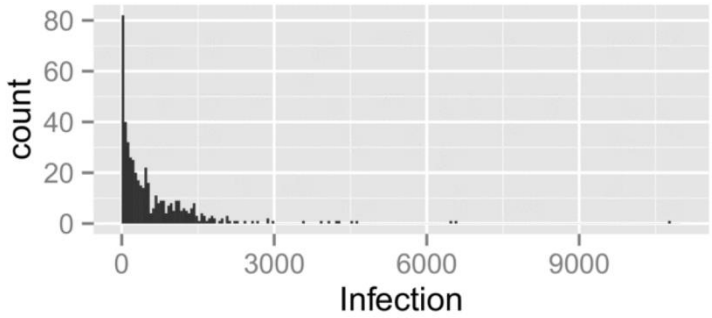


Гистограммы

```
qplot(Infection, data = dreissena,
      geom = "histogram", xlim = c(0, 11000))
qplot(Infection, data = dreissena,
      geom = "histogram", binwidth = 1000,
      xlim = c(0, 11000))
qplot(Infection, data = dreissena,
      geom = "histogram", binwidth = 50, xlim = c(0, 11000))
```



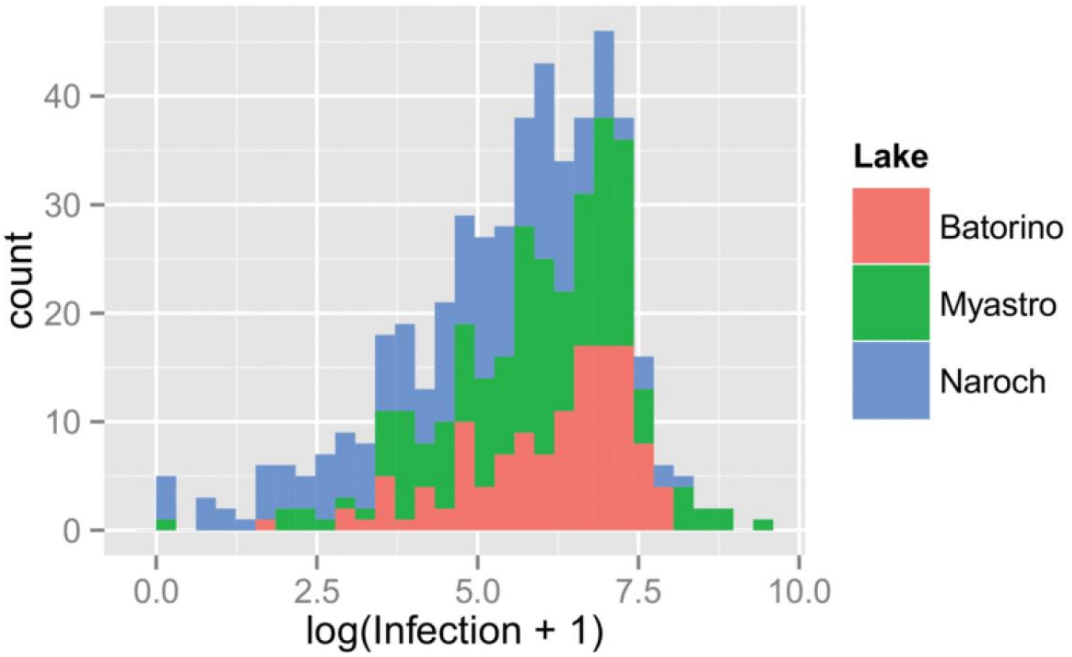




binwidth - подбор оптимального классового промежутка

Гистограммы

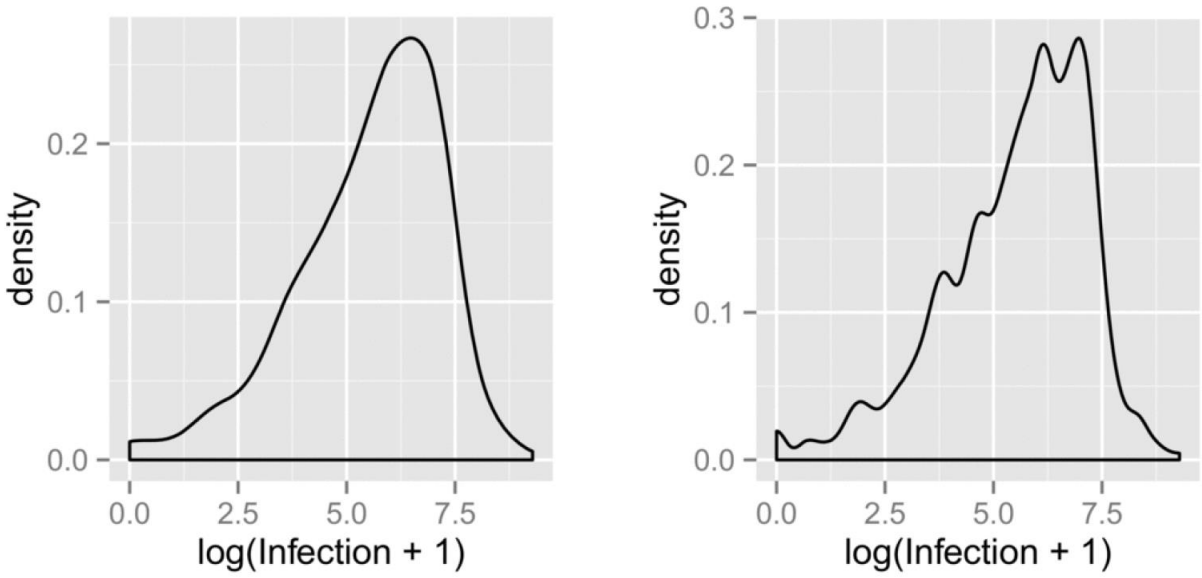
```
qplot(log(Infection + 1), data = dreissena,
      geom = "histogram", fill = Lake)
```



Кривые плотности вероятности

Предназначены для работы с непрерывными количественными переменными
Задается геометрическим объектом типа density
вызывается базовая функция R density()

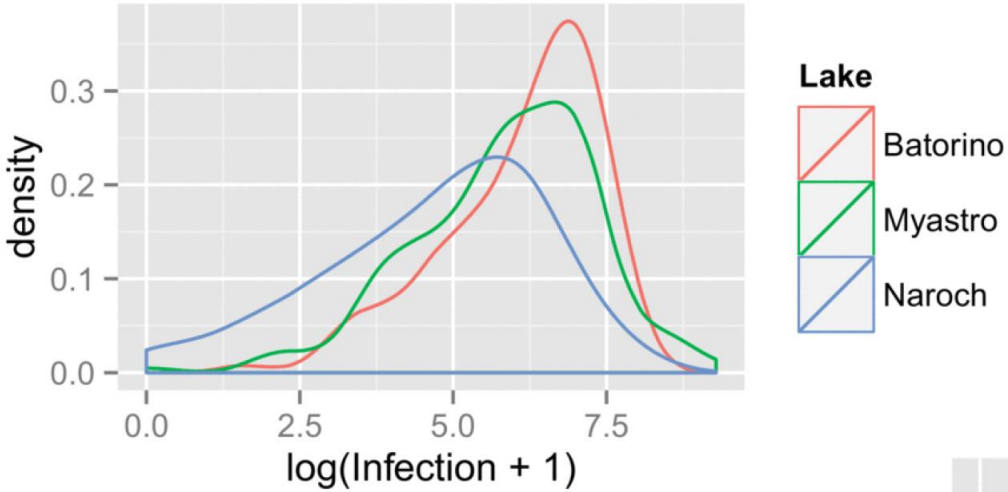
```
qplot(log(Infection + 1), data = dreissena, geom = "density")
qplot(log(Infection + 1), data = dreissena,
      geom = "density", adjust = 0.5)
```

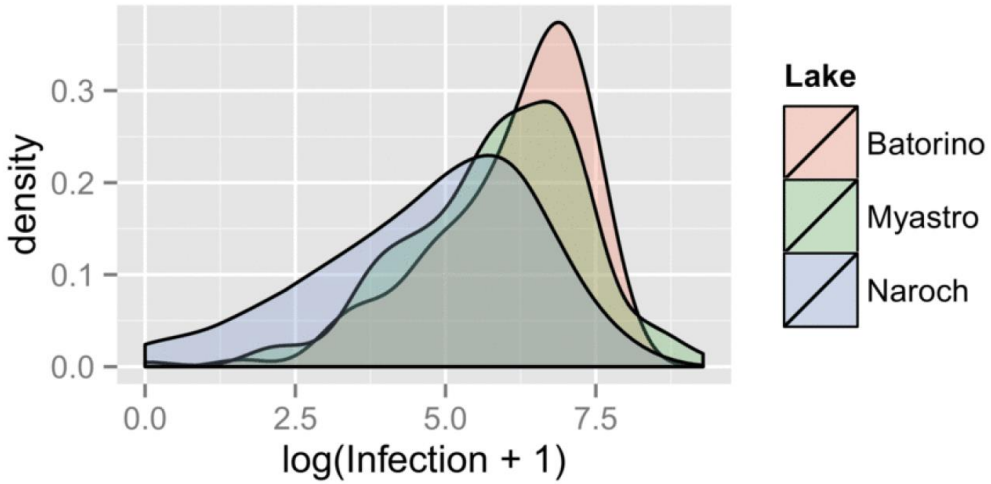


adjust – аргумент сглаживания

Кривые плотности вероятности

```
qplot(log(Infection + 1), data = dreissena, geom = "density",
      colour = Lake)
qplot(log(Infection + 1), data = dreissena, geom = "density",
      fill = Lake, alpha = I(1/4))
```

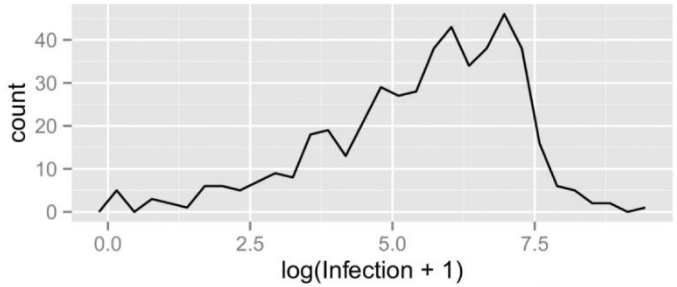


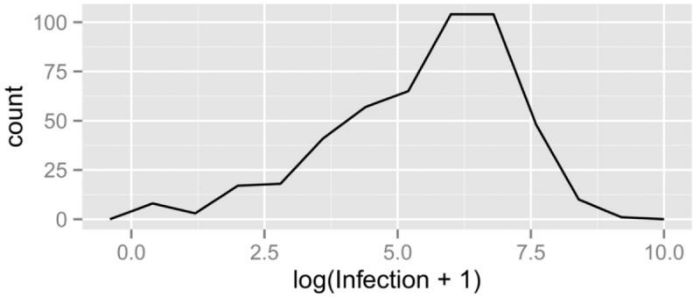


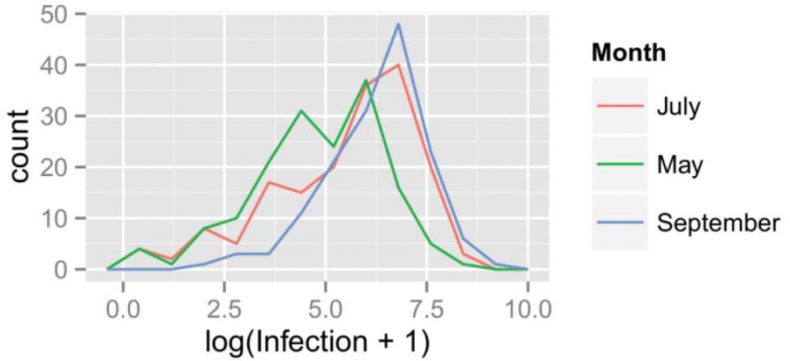
Полигоны частот

Визуализации распределения количественной переменной
Ось абсцисс – значение переменной
Ось ординат – частоты встречаемости

```
qplot(log(Infection+1), data = dreissena, geom = "freqpoly")
qplot(log(Infection+1), data = dreissena,
      geom = "freqpoly", binwidth = 0.8)
qplot(log(Infection+1), data = dreissena,
      geom = "freqpoly", binwidth = 0.7, colour = Month)
```



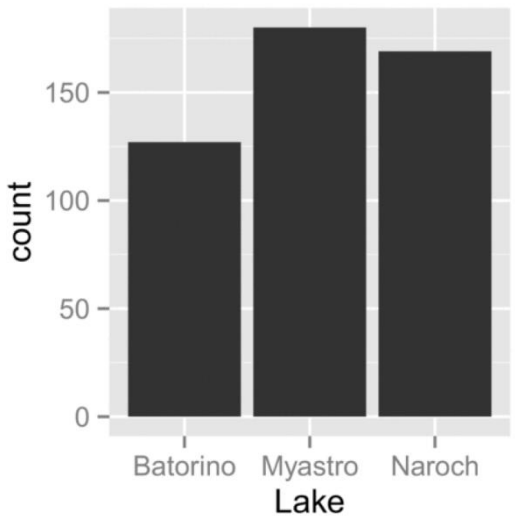




Столбиковые диаграммы

Показывают число наблюдений в группах, образованных качественными переменными
Создается геометрическим объектом barplot()
Для изменения принципа подсчета количества наблюдений на суммарное значение количественной переменной задается параметр weight

```
qplot(Lake, data = dreissena, geom = "bar")
qplot(Lake, data = dreissena, geom = "bar",
      weight = Infection)
```



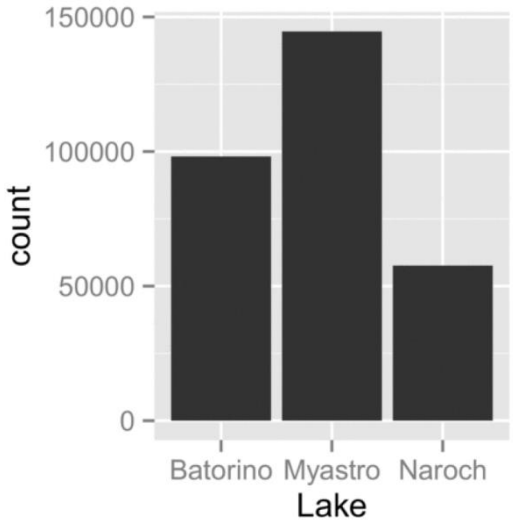
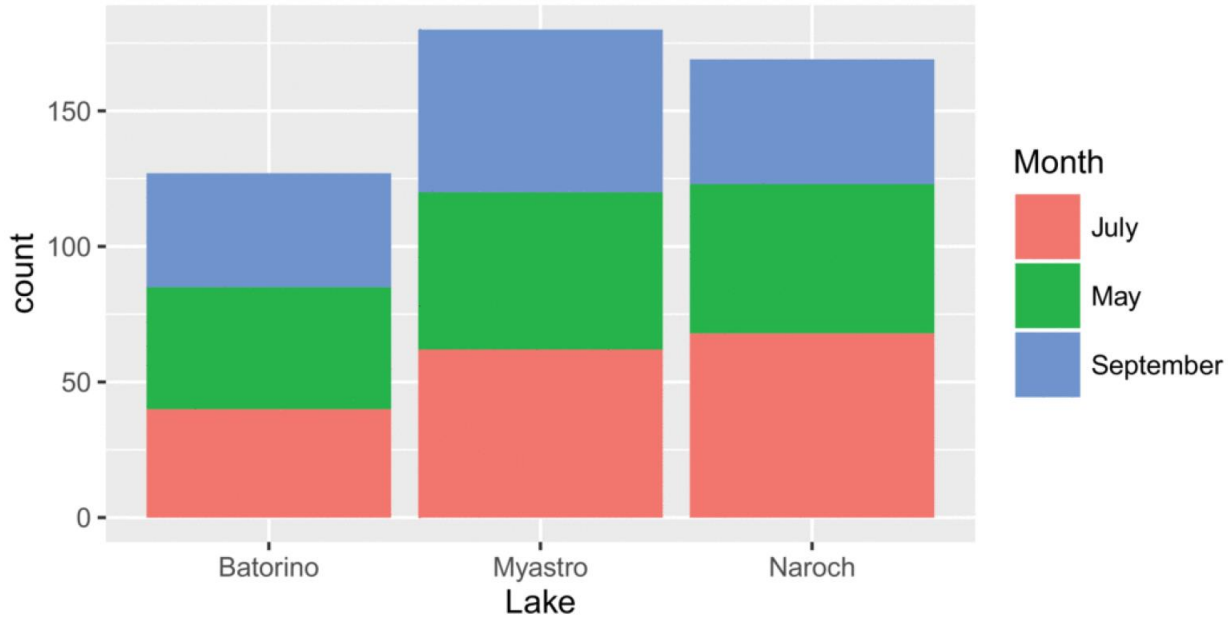


Диаграмма с накоплением (столбиковая составная диаграмма)

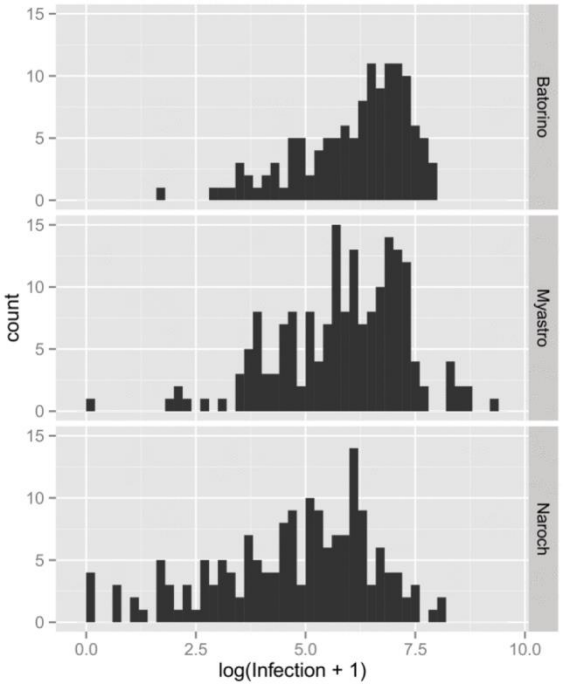
```
qplot(Lake, data = dreissena, geom = "bar", fill = Month)
```



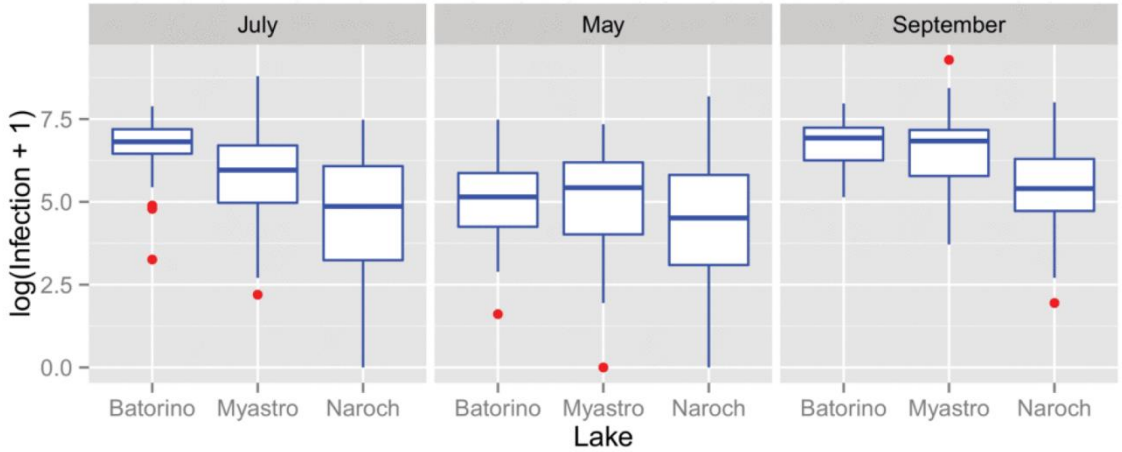
Категоризованные графики

Для каждой группы строится график определенного типа, потом графики компонуются в одной графической области.
Задается аргумент facet = rowvar ~ colvar
При отсутствии одного аргумента ставится точка
rowvar ~ . ; . ~ colvar

```
qplot(log(Infection + 1), data = dreissena,
      facets = Lake ~ ., geom = "histogram", binwidth = 0.2)
```

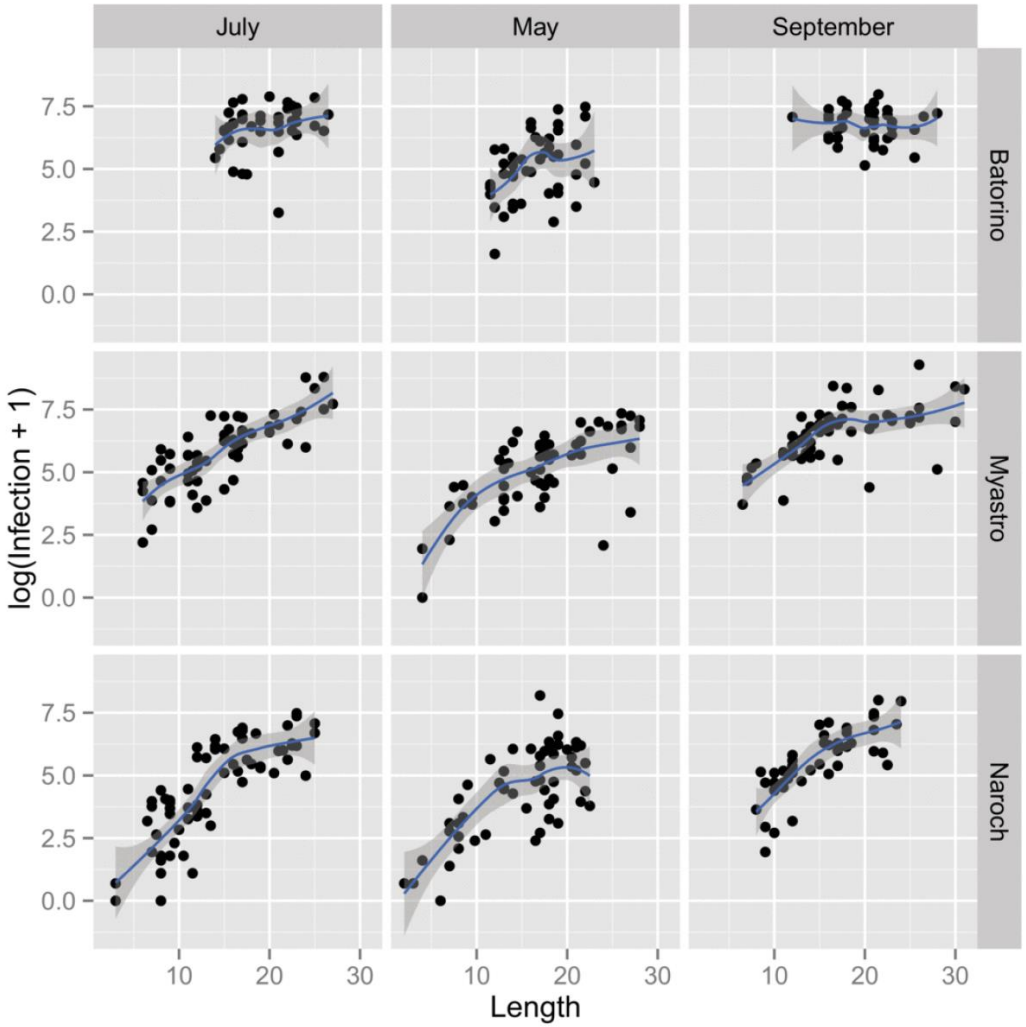



```
qplot(Lake, log(Infection + 1), data = dreissena,
      facets = . ~ Month, geom = "boxplot",
      colour = I("blue"), outlier.colour = "red")
```



Категоризованные графики

```
qplot(Length, log(Infection + 1), data = dreissena,
      facets = Lake ~ Month, geom = c("point","smooth"))
```



функция ggplot()

qplot() и ggplot()

qplot() автоматически инициировала новый график, добавляла к нему слои с геометрическими объектами и выводила результат на экран. Однако для более детальной настройки графика следует использовать функцию ggplot ()

Аргументы - данные

- data — имя таблицы с данными, на основе которых строится график;
- aes (от англ. aesthetics, что значит «эстетика») — функция, которая присваивает эстетические атрибуты геометрическим объектам, используемым для изображения данных на графике. У разных типов геометрических объектов эти атрибуты будут разными.

Слои

```
p <- ggplot(data = dreissena, aes(x = Infection))
p + layer(geom = "bar", stat = "bin",
          position = "identity",
          params = list(fill = "blue", binwidth = 100))
```

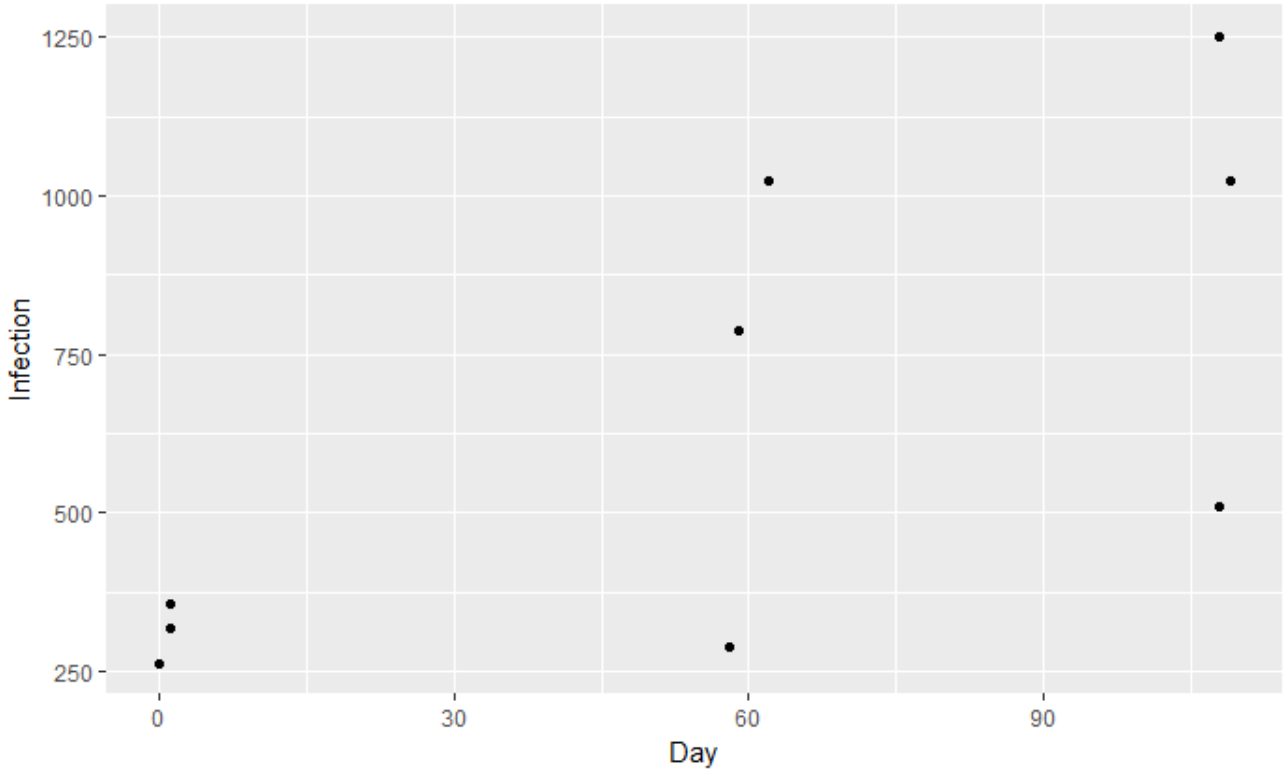
Объекты, создаваемые ggplot и layer являются самостоятельными объектами типа список (list).
Созданные объекты можно повторно использовать.

Группировка данных

```
> df2 <- aggregate(Infection~Lake+Day,dreissena,mean)
> df2
      Lake Day Infection
1   Naroch   0  262.7455
2 Batorino   1  316.3556
3  Myastro   1  356.6034
4   Naroch  58  289.4265
5  Myastro  59  786.5323
6 Batorino  62 1022.5250
7  Myastro 108 1251.8000
8   Naroch 108  510.7826
9 Batorino 109 1024.0238
```

Группировка данных

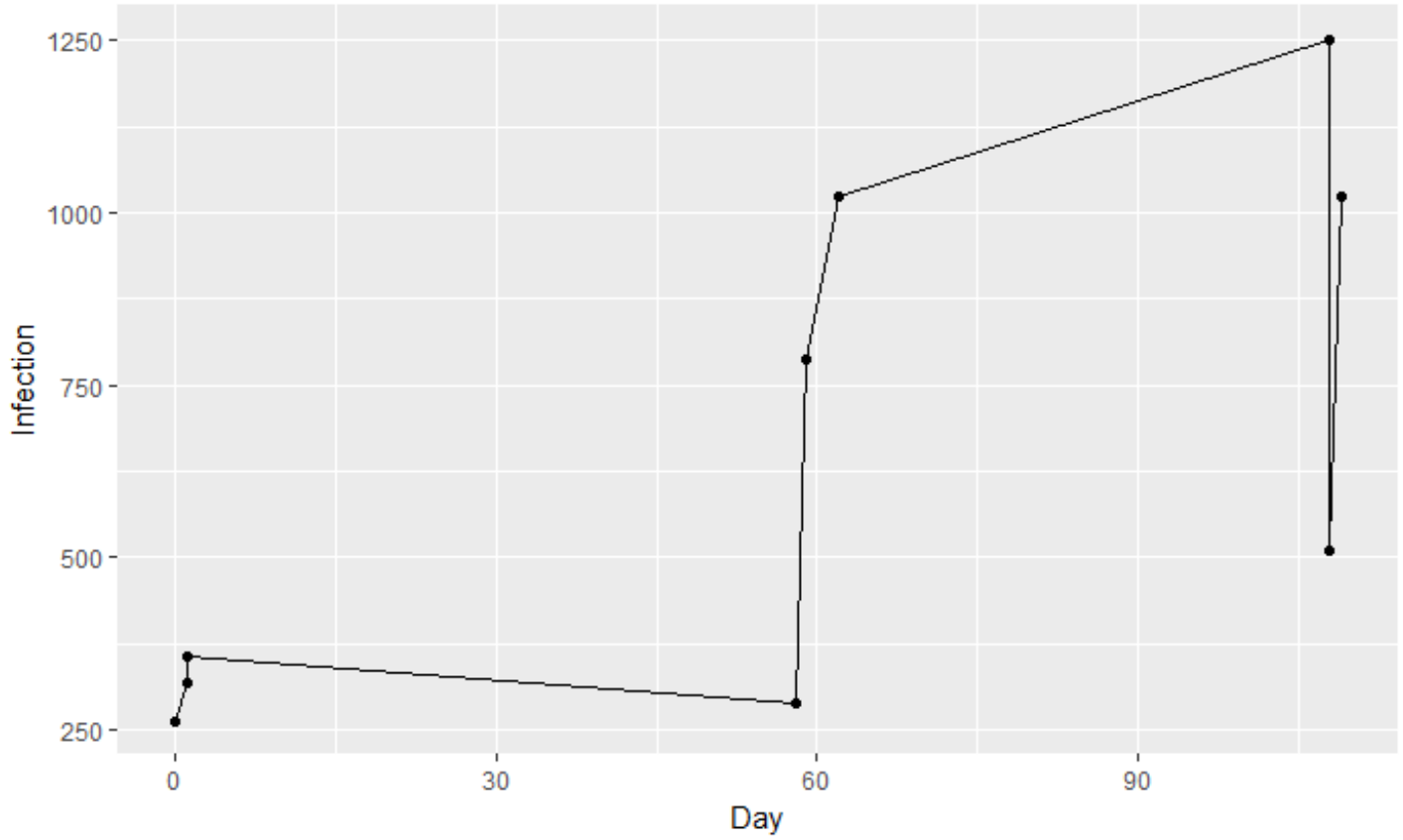
ggplot(df2,aes(Day,Infection))+geom_point()



- Нет принадлежности к озеру
- Непонятна тенденция (нужен график, а не точки)

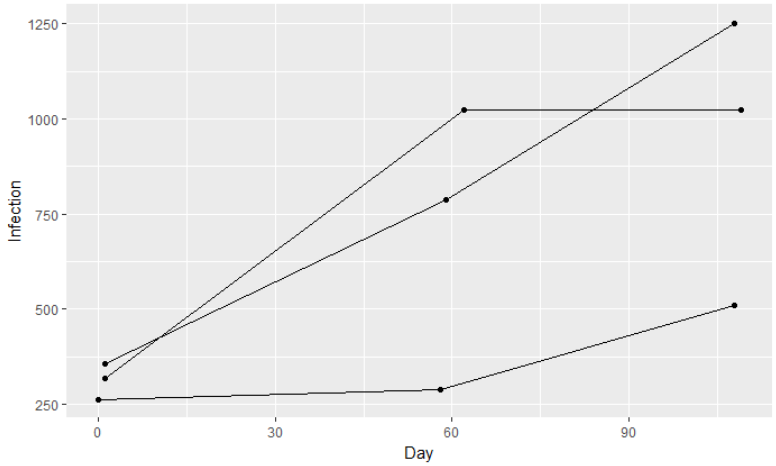
Группировка данных

ggplot(df2,aes(Day,Infection))+geom_point()+geom_line()

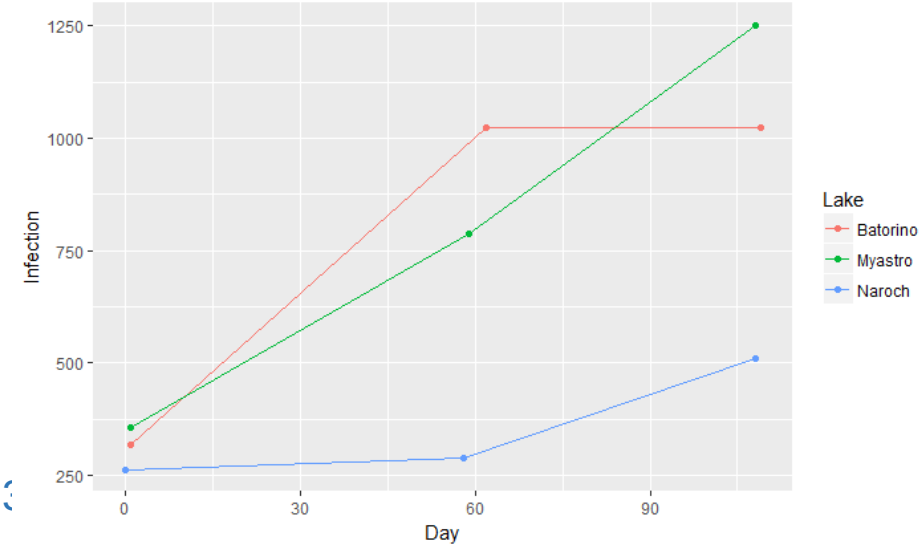


Группировка данных

ggplot(df2,aes(Day,Infection,group=Lake))+geom_point()+geom_line()



ggplot(df2,aes(Day,Infection,group=Lake,colour=Lake))+geom_point()
+geom_line()



Геометрические объекты ggplot2

Объект	Результат применения объекта
abline	Линия, заданная уравнением y = a + bx
area	Площадь под кривой
bar	Столбиковая диаграмма
bind2d	Тепловая карта для двух переменных, значения которых разбиты на классовые промежутки
blank	Пустой слой
boxplot	Диаграмма размахов
contour	Контурная диаграмма
crossbar	Прямоугольник, внутри которого изображена линия, параллельная его торцам; линия соответствует медиане или среднему значению
density	Диаграмма плотности вероятности
density2d	2D-диаграмма плотности вероятности
dotplot	Точечная диаграмма Уилкинсона
errorbar	Диаграмма диапазонов (с использованием вертикальных отрезков)
errorbarh	Диаграмма диапазонов (с использованием горизонтальных отрезков)
freqpoly	Полигон частот
hex	«Сотовая диаграмма»: координатная плоскость разбита на гексагоны, цвет заливки которых соответствует плотности расположения точек
histogram	Гистограмма
hline	Горизонтальная линия
jitter	Точечная диаграмма, на которой к координатам точек добавлен небольшой «шум»

Геометрические объекты ggplot2

Объект	Результат применения объекта
line	Линия, соединяющая упорядоченные по оси X наблюдения
linerange	Один из вариантов диаграммы диапазонов (с использованием верт. отрезков)
map	Географическая карта (и другие похожие многоугольники)
path	Линия, соединяющая наблюдения в порядке, который задан одной из переменных в таблице с данными
point	Диаграмма рассеяния
pointrange	Точка с исходящими из нее отрезками
polygon	Многоугольник
quantile	Линии квантильной регрессии
raster	Растровое изображение
rect	Прямоугольник
ribbon	Ленточная диаграмма
rug	Небольшие перпендикулярные координатной оси отрезки, обозначающие отдельные наблюдения
segment	Линии, координаты начала и конца которых заданы пользователем
smooth	Сглаживающая линия (линия тренда)
step	Эмпирическая кумулятивная функция плотности вероятности
text	Текстовые аннотации
tile	Плоскость, разбитая на прямоугольники
violin	«Скрипичная диаграмма»: смесь диаграммы размахов с диаграммой плотности вероятности
vline	Вертикальная линия

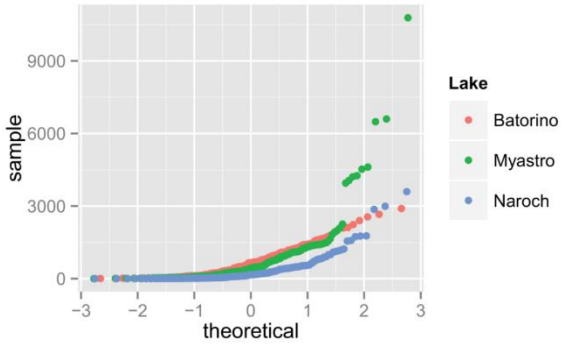
Статистические преобразования

Статистические преобразования представляют собой функции, которые выполняют определенные математические операции над исходными данными.

Эти функции принимают таблицу с исходными данными и возвращают таблицу с новыми переменными, содержащими результаты вычислений. Часто новые переменные добавляются непосредственно в исходную таблицу, и им можно присвоить эстетические атрибуты.

Так, stat_qq() — функция, применяемая для построении графиков нормальной вероятности, — возвращает таблицу со следующими дополнительными переменными: sample (выборочные квантили) и theoretical (теоретически ожидаемые квантили нормального распределения).

```
ggplot(dreissena, aes(sample = Infection, colour = Lake)) +
       stat_qq()
```



Статистические преобразования

Объект	Описание
bin	Разбиение данных на классы
bin2d	Разбиение данных на классы для построения двухмерных диаграмм плотности вероятности
bindot	Сортировка данных для построения точечных диаграмм
binhex	Разбиение данных на классы для построения «сотовых диаграмм»
boxplot	Вычисления, необходимые для построения диаграмм размахов
contour	Вычисления, необходимые для построения контурных диаграмм
density	Одномерное ядерное оценивание плотности вероятности
density2d	Двухмерное ядерное оценивание плотности вероятности
ecdf	Сортировка данных для построения эмпирической кумулятивной функции плотности вероятности
function	Любая пользовательская функция для преобразования данных

Статистические преобразования

Объект	Описание
identity	Возвращает данные в неизменном виде
qq	Вычисления, необходимые для построения квантильных графиков
quantile	Расчет линий квантильной регрессии
smooth	Подгонка сглаживающей линии
summary	Позволяет создавать пользовательские функции для расчета сводных статистических показателей по значениям переменной Y для каждого уникального значения переменной X.
summary_hex	Позволяет создавать пользовательские функции для расчета показателей, отражаемых на «сотовых диаграммах»
summary2d	Позволяет создавать пользовательские функции для расчета показателей, отражаемых на двухмерных диаграммах плотности вероятности
unique	Удаление повторяющихся значений из выборки
ydensity	Одномерное ядерное оценивание плотности вероятности, необходимое для построения диаграмм типа «виолончель»

Статистические преобразования

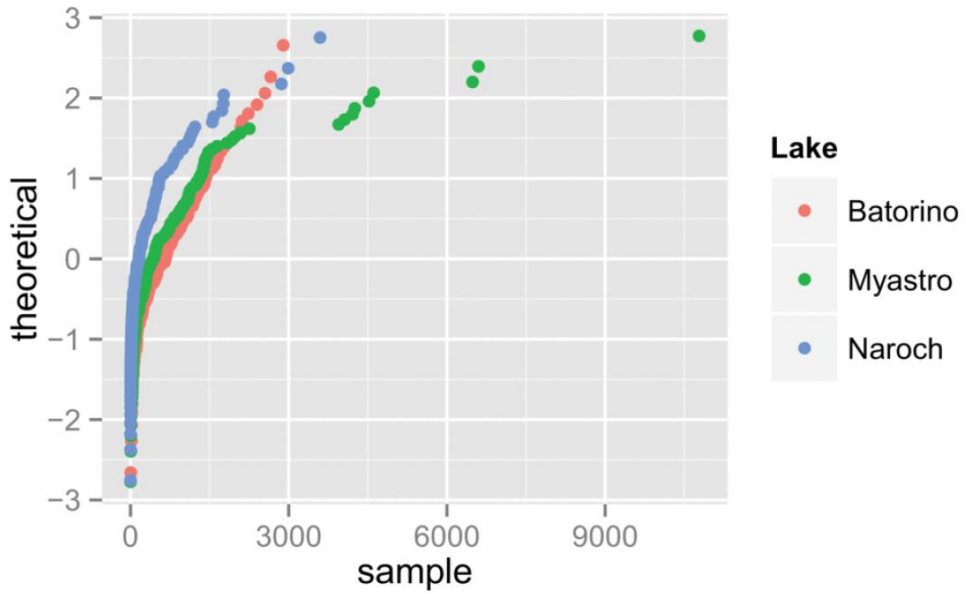
К дополнительным переменным, которые создаются функциями статистических преобразований, можно обращаться, окружая их имена двойными точками (например, . .sample. . и . .theoretical. . в случае с stat_-qq().

Наличие двойных точек позволяет избежать возникновения ошибок при совпадении имен новых переменных с именами переменных, которые могли присутствовать в данных изначально. Кроме того, наличие двойных точек помогает идентифицировать новые переменные при написании кода и его чтении.

С перечнем имен переменных, создаваемых функциями статистических преобразований, можно ознакомиться в соответствующих справочных файлах

Статистические преобразования

```
ggplot(data = dreissena, aes(sample = Infection,
       colour = Lake)) +
       stat_qq(aes(x = ..sample.., y = ..theoretical..))
```



Переназначены данные для X и Y

Основные типы статистических графиков

Общие аргументы geom- и stat- функций

- data — имя таблицы данных, специфичной для конкретного слоя. Применяется, когда необходимо изобразить на слое данные, отличные от тех, которые были заданы при инициализации графика функцией ggplotO;
- mapping — служит для присваивания эстетических атрибутов (обычно при помощи функции aes()). Используется только при необходимости отменить глобальные настройки графика и задать определенные эстетические атрибуты на уровне слоя;
- stat — задает статистическое преобразование, которое применяется к изображаемым на слое данным.
- position - определяет взаимное расположение перекрывающихся геометрических объектов. Значения этого аргумента будут разными у разных функций;
- na.rm — логический аргумент, определяющий действия в отношении пропущенных данных. При па. rm = FALSE (значение, принятое по умолчанию) пропущенные наблюдения будут удалены и пользователь увидит на экране соответствующее предупреждение. При na.rm = TRUE пропущенные значения также будут удалены, но без вывода предупреждающего сообщения.

Точечные диаграммы Уилкинсона geom_dotplot()

Точечные диаграммы Уилкинсона (англ. Wilkinson dot plots) служат для визуализации распределений непрерывных количественных переменных. Это один из простейших статистических графиков, хорошо подходящий для работы с малыми выборками (20-30 наблюдений). Данные на диаграмме Уилкинсона изображаются в виде точек, определенным образом упорядоченных друг над другом вдоль координатной оси.

Графики этого типа хорошо подходят для выявления выбросов и кластеров наблюдений.

Точечные диаграммы Уилкинсона geom_dotplot()

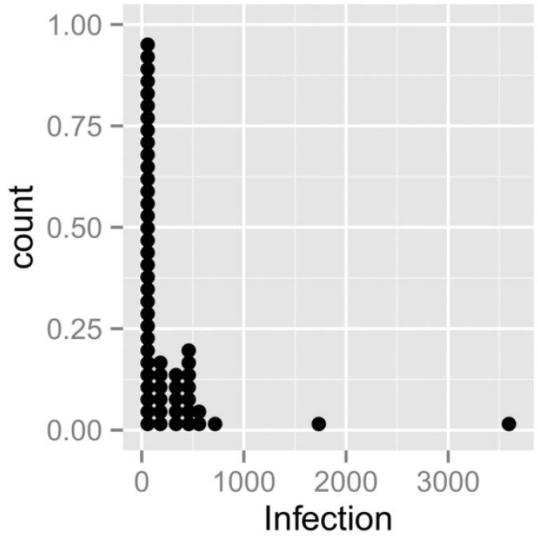
- binaxis — переменная, значения которой подлежат разбиению на классы (по умолчанию binaxis = "x").
- method — название алгоритма, используемого для разбиения наблюдений на классы: method = "dotdensity" для разбиения с учетом плотности вероятности и method = "histodot" для разбиения с применением фиксированного классового промежутка (как при построении гистограмм).
- binwidth — при method = "dotdensity" задает степень сглаживания для алгоритма оценивания плотности вероятности; при method = "histodot" задает фиксированный размер классового промежутка. По умолчанию этот параметр равен (max — min) /30.
- binpositions — при method = "dotdensity" значение binpositions = "bygroup" указывает на необходимость расчета плотности вероятности в пределах каждой из групп данных. При binpositions = "all" оценивание плотности вероятности выполняется для всех имеющихся данных.
- stackdir — задает направление, вдоль которого точки на графике должны укладываться в «стопки». Возможные значения: "up" (по умолчанию), "down", "center" и "centerwhole" (см. примеры ниже).
- stackratio — определяет степень перекрытия точек. По умолчанию stackratio = 1 (точки почти касаются друг друга). При меньших значениях точки частично перекрывают друг друга.
- dotsize — относительный размер точек (по умолчанию dotsize =1)
- stackgroups — логический аргумент, включающий тот же эффект для точечных диаграмм, что и position = "stack" в случае со столбиковыми диаграммами.

Точечные диаграммы Уилкинсона geom_dotplot()

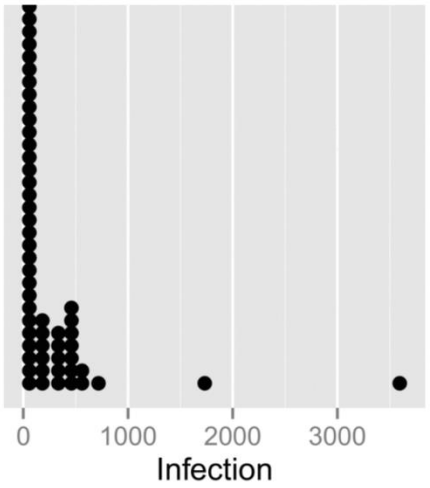
Эстетические атрибуты

- x* и y — переменные X и Y соответственно .
- alpha — степень прозрачности цвета.
- colour — цвет линии, окаймляющей точки.
- fill — цвет точек.

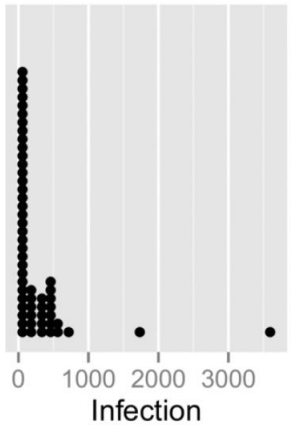
```
ggplot(data = subset(dreissena,
       Lake == "Naroch" & Month == "May"),
       aes(x = Infection)) + geom_dotplot()
```

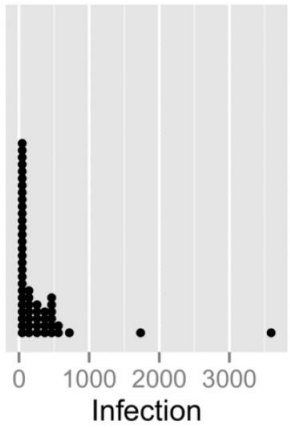


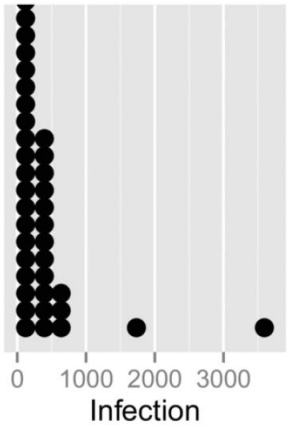

```
ggplot(data = subset(dreissena,
       Lake == "Naroch" & Month == "May"),
       aes(x = Infection)) + geom_dotplot() +
       scale_y_continuous(name = "", breaks = NULL)
```



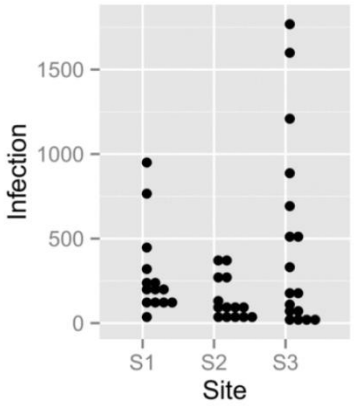
Точечные диаграммы Уилкинсона geom_dotplot()

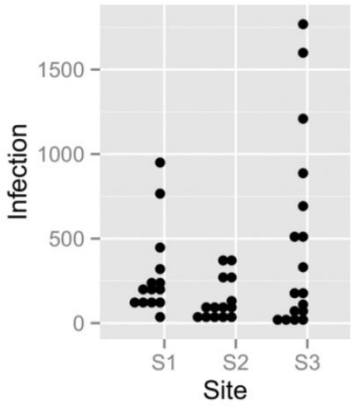


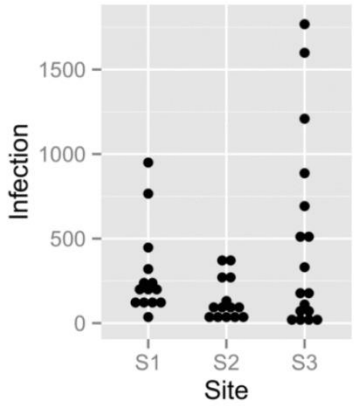




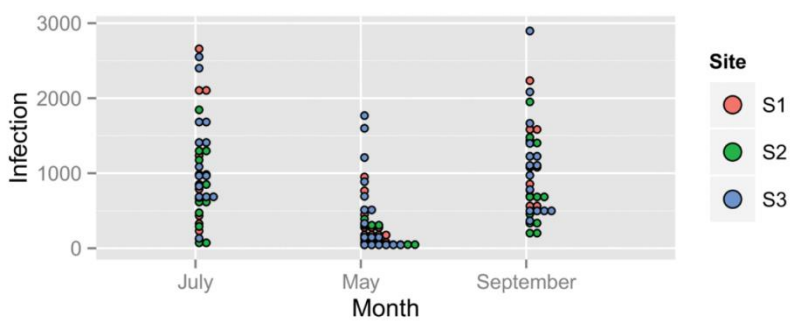
binwidth=def
binwidth=100
binwidth=250

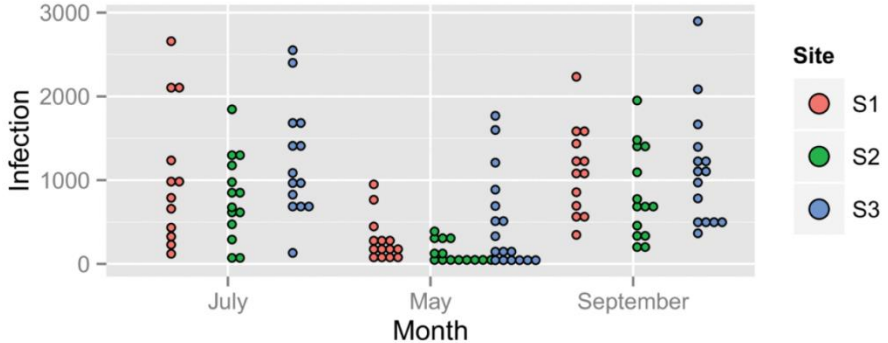






stackdir="up"
stackdir="down"
stackdir="center"





Столбиковые диаграммы geom_bar()

Аргументы

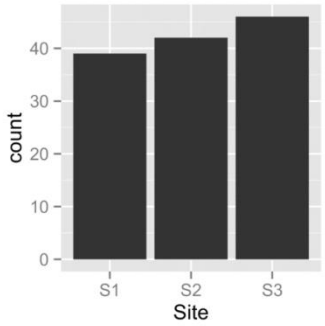
- weight — переменная, по значениям которой выполняется «взвешивание» значений переменной X

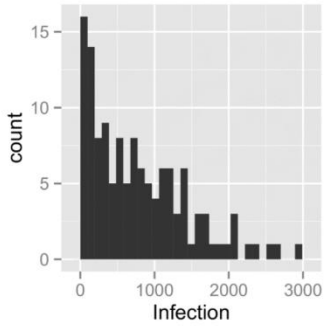
Эстетические атрибуты

- x* — переменная X.
- alpha — степень прозрачности цвета.
- colour — цвет линии, окаймляющей столбики.
- linetype — тип линии, окаймляющей столбики.
- size — толщина линии, окаймляющей столбики.
- fill — цвет заливки столбиков.

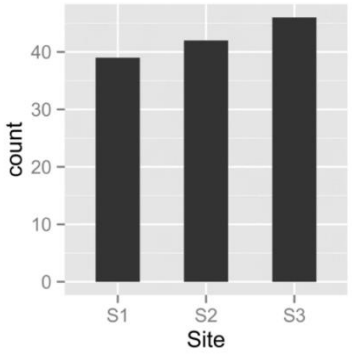
Столбиковые диаграммы geom_bar()

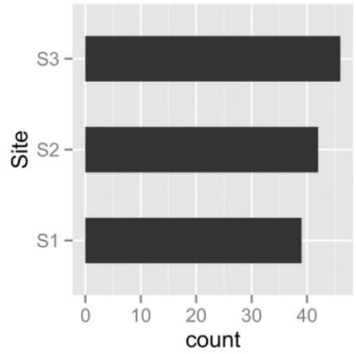
```
ggplot(data = subset(dreissena, Lake == "Batorino"),
       aes(x = Site)) + geom_bar()
ggplot(data = subset(dreissena, Lake == "Batorino"),
       aes(x = Infection)) + geom_bar()
```





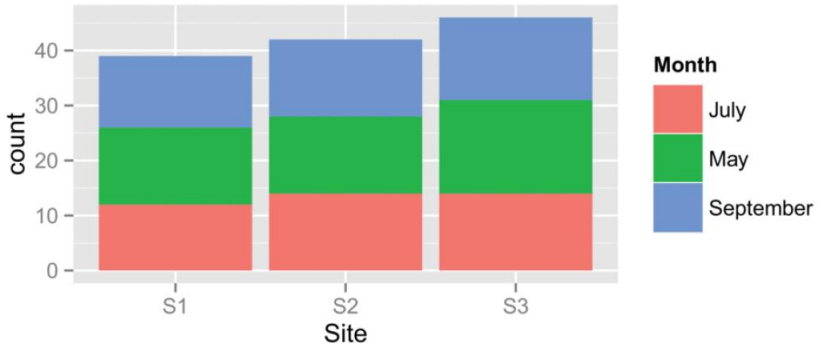

```
p <- ggplot(data = subset(dreissena,
     Lake == "Batorino"), aes(x = Site))
p + geom_bar(width = 0.5)
p + geom_bar(width = 0.5) + coord_flip()
```

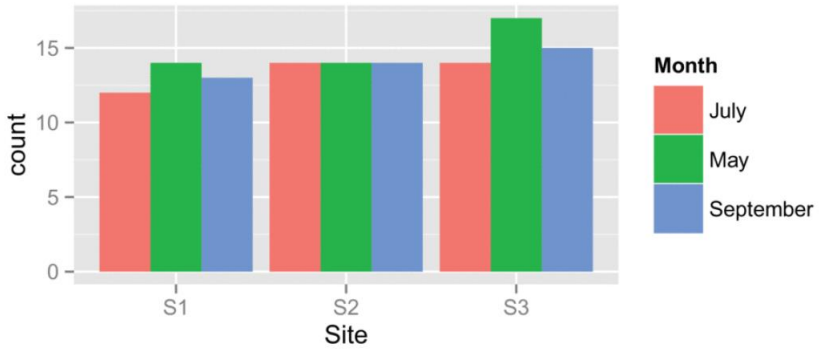


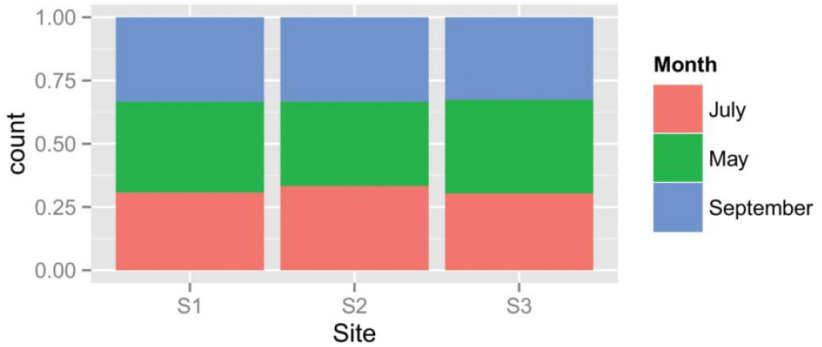


Столбиковые диаграммы geom_bar()

```
p <- ggplot(data = subset(dreissena,
            Lake == "Batorino"), aes(x = Site))
p + geom_bar(aes(fill = Month), position = "stack")
p + geom_bar(aes(fill = Month), position = "fill")
p + geom_bar(aes(fill = Month), position = "dodge")
```



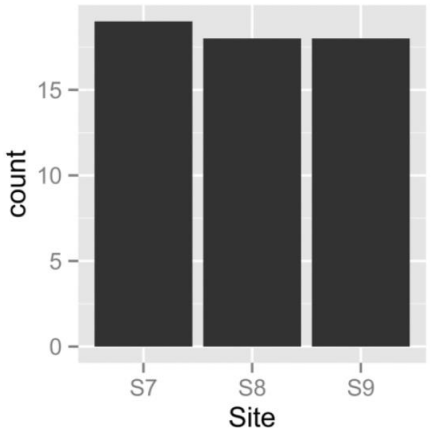


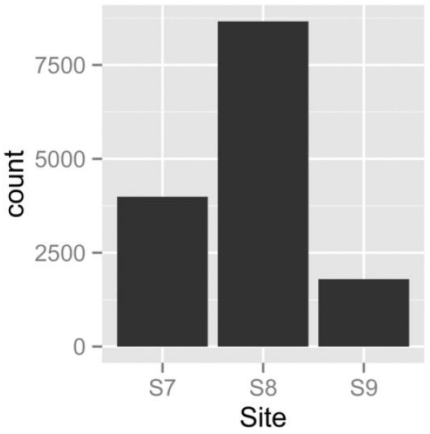


Столбиковые диаграммы geom_bar()

```
p <- ggplot(data = subset(dreissena,
            Lake == "Naroch" & Month == "May"),
            aes(x = Site))
p + geom_bar()
p + geom_bar(aes(weight = Infection))
```

Слева: столбиковая диаграмма, на которой ось Y соответствует числу особей дрейссены из озера Нарочь, обследованных в мае (атрибут weight не задействован). Справа: значения переменной Station «взвешены» по значениям переменной Infection (aes(weight = Infection)), что привело к подсчету суммарного количества инфузорий C. acuminatus, обнаруженных во всех исследованных особях дрейссены на каждой станции





Гистограммы: geom_histogram()

Гистограмма представляет собой вариант столбиковой диаграммы, применяемый для визуализации распределений количественных переменных с относительно большим размахом значений.

При построении гистограммы значения анализируемой переменной упорядочиваются по возрастанию, а затем разбиваются на классы в соответствии с некоторым классовым промежутком.

Получаемый в итоге график выглядит как совокупность из нескольких столбиков, ширина которых соответствует величине классового промежутка, а высота -- частоте встречаемости соответствующего класса.

Гистограммы: geom_histogram()

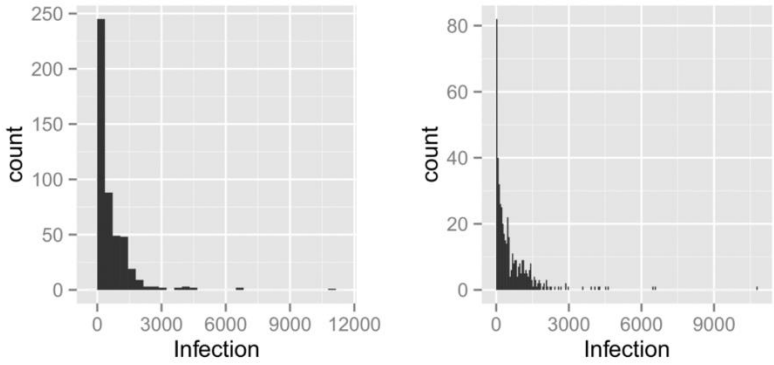
Аргументы

- binwidth — размер классового промежутка.
- weight — переменная, по значениям которой происходит «взвешивание» значений переменной X

Эстетические атрибуты

- x* — переменная X.
- alpha — степень прозрачности цвета.
- colour — цвет линии, окаймляющей столбики
- fill — цвет заливки столбиков
- linetype — тип линии, окаймляющей столбики
- size — толщина линии, окаймляющей столбики

```
p <- ggplot(data = dreissena, aes(x = Infection))
p + geom_histogram()
p + geom_histogram(binwidth = 50)
```



Полигоны частот: geom_freqpoly()

Подобно гистограмме, полигон частот представляет собой приближенный вариант распределения плотности вероятности количественной переменной. По сути, единственное различие между этими двумя типами графиков состоит в том, что полигон частот изображают в виде сплошной ломаной линии.

Координаты узловых точек этой линии соответствуют вершинам столбиков гистограммы.

Кроме того, ломаная касается оси X по обеим сторонам распределения в точках, соответствующих ближайшим классам с нулевыми значениями частот. В результате этого образуется замкнутая фигура («полигон»)

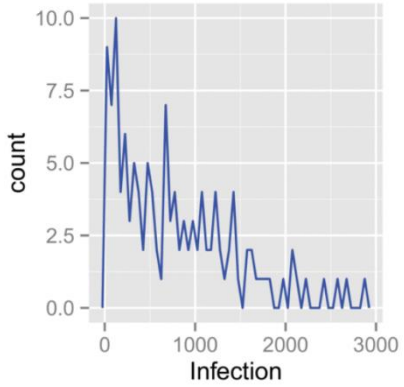
Полигоны частот: geom_freqpoly()

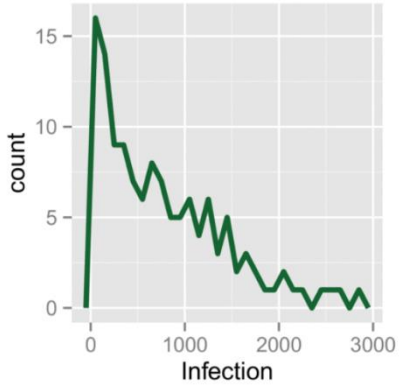
Аргументы

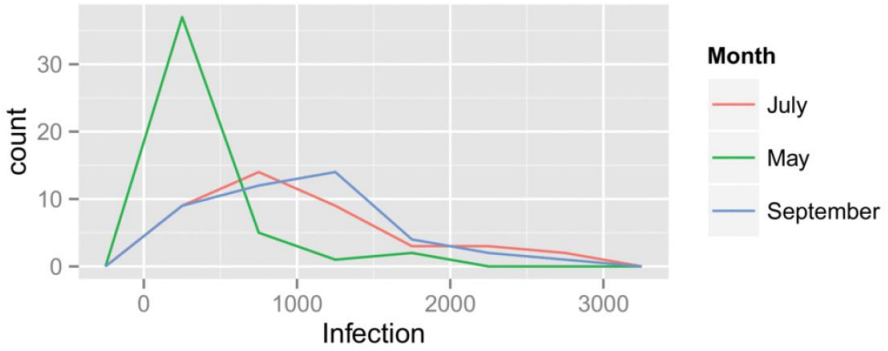
- binwidth — размер классового промежутка.

Эстетические атрибуты

- x* — переменная X.
- alpha — степень прозрачности цвета.







Кривые плотности вероятности: geom_density()

Аргументы

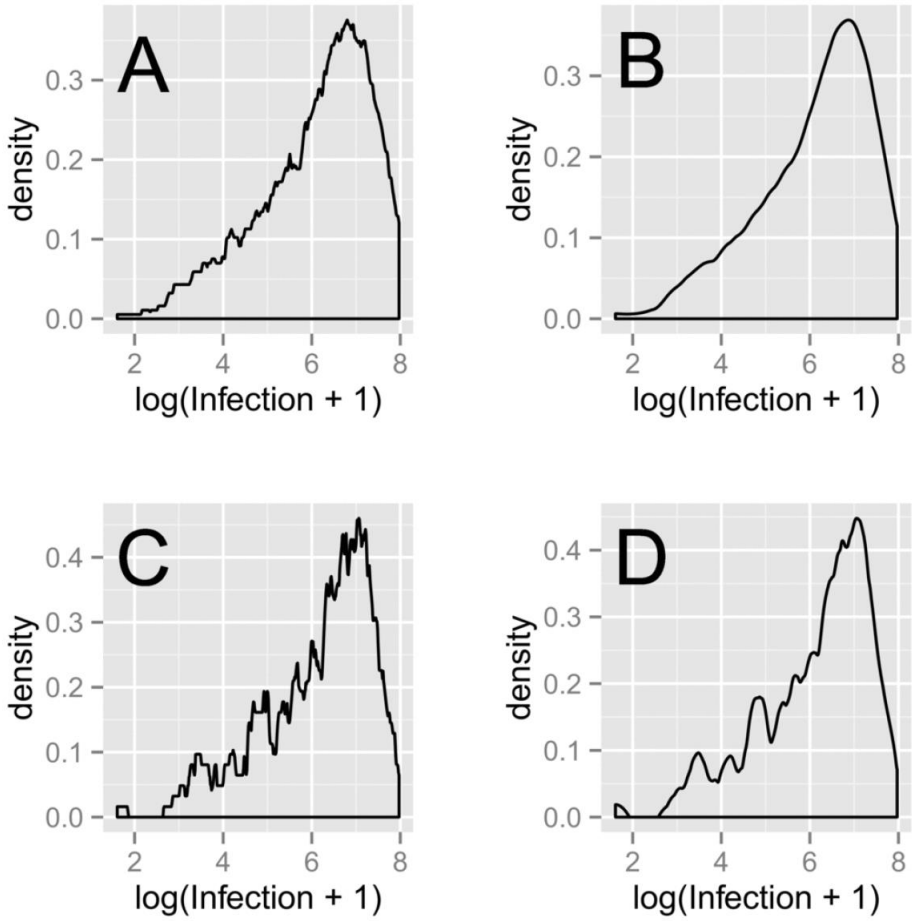
- kernel - - задает алгоритм ядерного оценивания плотности вероятности; возможные значения: "gaussian" (принято по умолчанию), "rectangular", "biweight", "epanechnikov", "triangular", "cosine" и "optcosine".
- adjust — определяет степень сглаживания кривой плотности вероятности (более высокие значения соответствуют большей степени сглаживания
- trim — логический аргумент. При trim = TRUE (значение, принятое по умолчанию) рассчитанные вероятности ограничиваются размахом выборочных значений анализируемой переменной. При trim = FALSE диапазон оцениваемых вероятностей расширяется на некоторую небольшую величину
- weight — переменная, по значениям которой происходит «взвешивание» значений переменной X.

Эстетические атрибуты

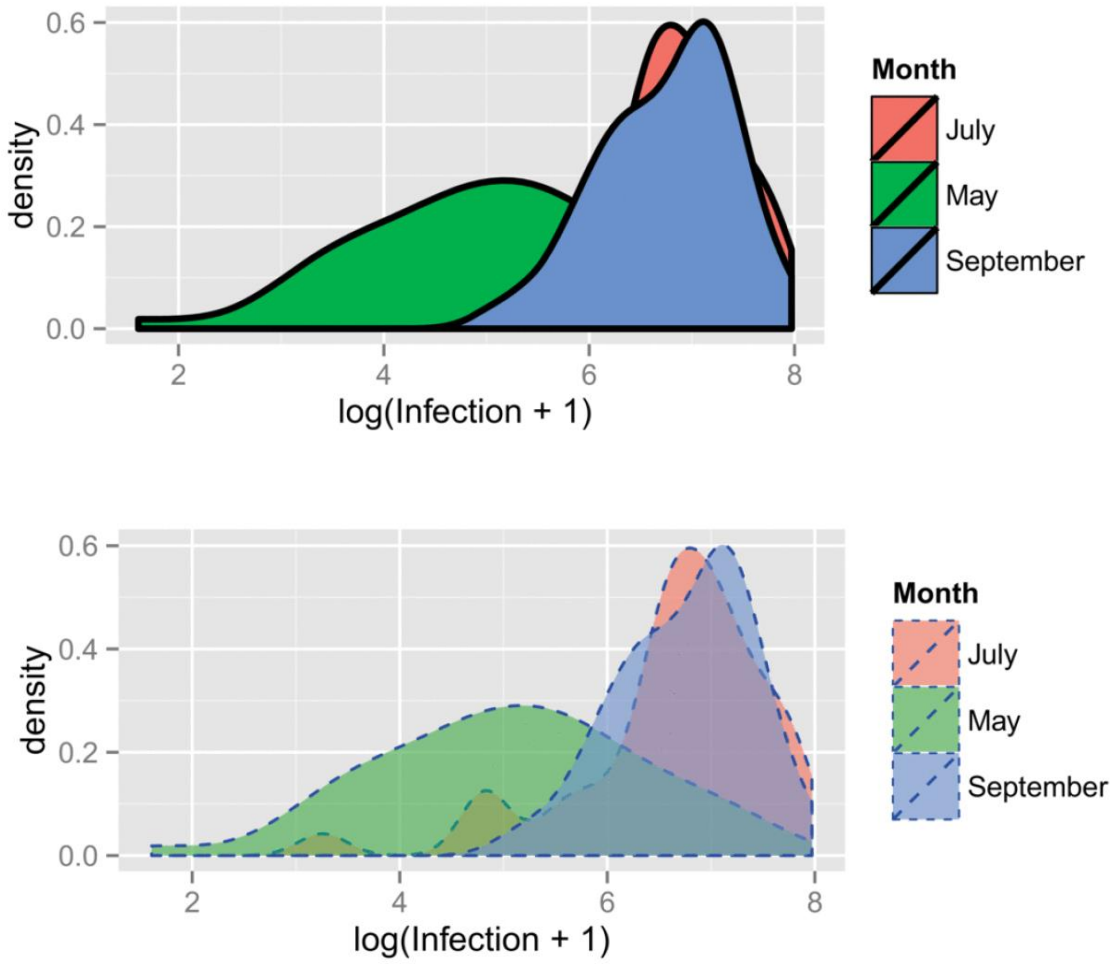
- x* и y — переменные X и Y соответственно.
- alpha — степень прозрачности цвета.
- colour — цвет линии.
- fill — цвет заливки площади под кривой.
- linetype — тип линии.
- size — толщина линии.

Кривые плотности вероятности: geom_density()

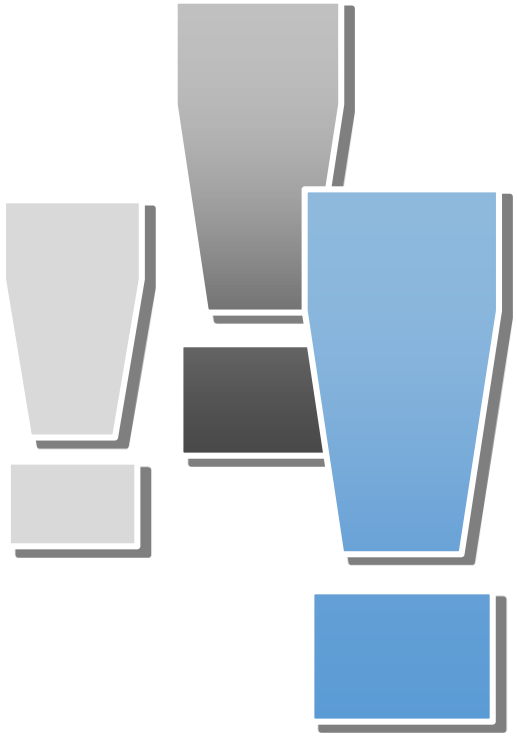
```
p <- ggplot(data = subset(dreissena, Lake == "Batorino"),
            aes(x = log(Infection + 1)))
p + geom_density(kernel = "rectangular")
p + geom_density(kernel = "epanechnikov")
p + geom_density(kernel = "rectangular", adjust = 1/3)
p + geom_density(kernel = "epanechnikov", adjust = 1/3)
```



Кривые плотности вероятности: geom_density()



Спасибо за внимание!



Шевцов Василий Викторович

shevtsov_vv@rudn.university
+7(903)144-53-57