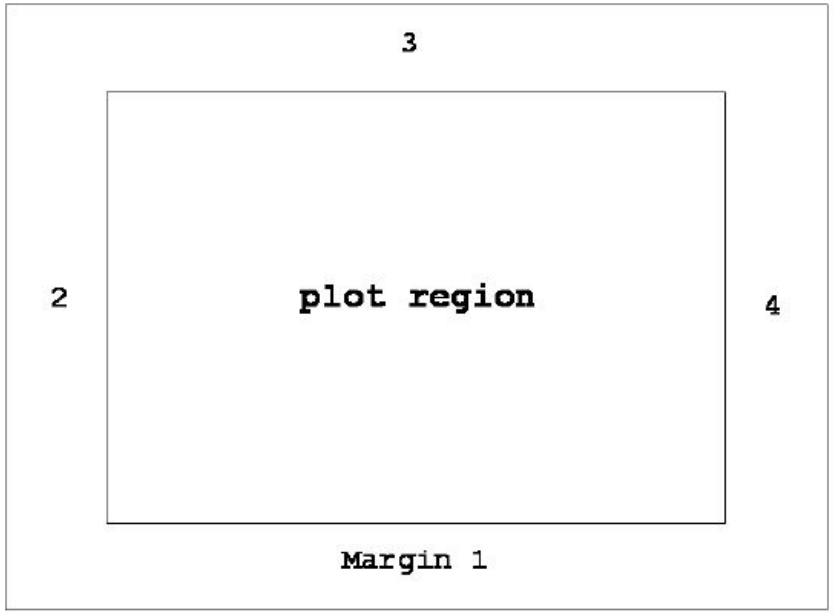
Графика в R

Анатомия графика



- График состоит из plot region, окруженного полями margins. За размер полей отвечает аргумент *mai*. Значение *mai* – вектор c(bottom,left,top,right), задающий ширину соответствующих полей в дюймах. С помощью функции par можно задавать графические параметры:
- par(mai=c(5,5,8,5)/10)

Анатомия графика

- Оси, метки осей и подписи находятся на полях рисунка. На каждом из полей находится несколько текстовых строк. Строки, определенные в 0 позиции, находятся на границе plot region, там, где рисуются оси. Большие значения позиций строк соответствуют строкам, удаленным от графика. Графический параметр *mar* определяет, сколько строк находится в каждом из полей, это альтернативный способ определения полей. За шрифт отвечает параметр *font*.

Анатомия графика

- Функция *axis()* рисует оси на текущем графике. Аргумент *side* определяет, на какой стороне появятся оси. Обычно они рисуются на строке 0, но это можно поменять. Также оси можно рисовать внутри графика с помощью аргумента *pos*. Для настройки отметок на осях используется аргумент *at*.
- Метки на осях задаются с помощью графических параметров xlab и ylab, переданных, например, в функцию plot(). После создания графика такие метки можно добавить с помощью функции title() или функции mtext():
- > mtext("Label text",side=1,line=2) – добавление текста сразу под осью x. По умолчанию текст центрируется.

Анатомия графика

- За стиль осей и меток на них отвечают следующие параметры:
- axes: рисовать ли оси? (TRUE/FALSE)
- bty: тип рамки вокруг графика
  - bty="o": есть рамка вокруг графика (default)
  - bty="l": оси в виде буквы L
  - bty="7": частичные оси слева и снизу.
  - bty="c": оси в виде буквы C
  - bty="u": оси в виде буквы U
  - bty="]": оси в виде] с частью осей слева от графика
  - bty="n": нет рамки вокруг графика

Анатомия графика

- lab=c(nx,ny,len): определяет способ, которым помечаются оси. Определяет количество интервалов отметок на графике и длину меток (в символах).
- las: стиль меток на осях
  - las=0: всегда параллельны осям (default)
  - las=1: всегда горизонтальны
  - las=2: всегда перпендикулярны осям
  - las=3: всегда вертикальны
- • tck: длина отметок на осях как доля области графика. Отрицательные значения соответствуют позициям, которые находятся вне области графика. Положительные значения указывают на отметки на осях, находящиеся внутри области графика.

Несколько графиков

- Для расположения нескольких графиков на одной поверхности есть два основных способа. Графический параметр fig позволяет расположить несколько графиков, даже беспорядочно, на одной области рисунка.
- Также возможно нарисовать несколько графиков в виде массива из n × m рисунков. Это определяется графическими параметрами mfrow или mfcol Например,
- > par(mfrow=c(3,2))
- даст область графика с 3 строками и 2 столбцами.
- Каждая высокоуровневая графическая команда начинается с нового рисунка (figure). Когда все рисунки исчерпаны, создается новая страница. Графический параметр mfg следит за строками и столбцами текущего рисунка.

Другие графические параметры

- ask=T: R спрашивает перед выводом графика.
- new=T: декларирует, что текущий график не используется. Это значит, что R не будет стирать его перед переходом к другому графику. Т.о. можно рисовать несколько графиков на одном рисунке.
- • fin: дает ширину и высоту текущего рисунка в дюймах.
- • din: параметр только для чтения, который возвращает ширину и высоту используемого устройства в дюймах.

Обзор графических функций

- В R есть ряд графических функций. Они обычно делятся на высокоуровневые – для построения графиков, и низкоуровневые – для добавления элементов в существующие графики. У каждой функции есть свой набор аргументовhas a variety of graphics functions. Основные из них
- xlim,ylim: диапазон значений по осям x и y соотвтетственно
- pch, col, lty: символы графика, цвет и тип линии
- xlab, ylab: названия осей x и y соответственно
- main, sub: заголовок и подзаголовок графика

Обзор графических функций

- Главные графические параметры могут быть заданы с помощью функции par(). Например, чтобы посмотреть настройки типа линии:

```
> par()$lty
```

- Чтобы задать тип линии:

```
>par(lty=2)
```

- Чтобы нарисовать несколько графиков на рисунке:

```
# 2x2 plotting region where plots
# appear by row
> par(mfrow=c(2,2))
# 2x2 plotting region where plots
# appear by column
> par(mfcol=c(2,2))
```

Изображение одномерных данных

- Графические методы для исследования свойств распределений векторов включают:
- hist (гистограмма)
- Boxplot
- density
- qqnorm
- qqline

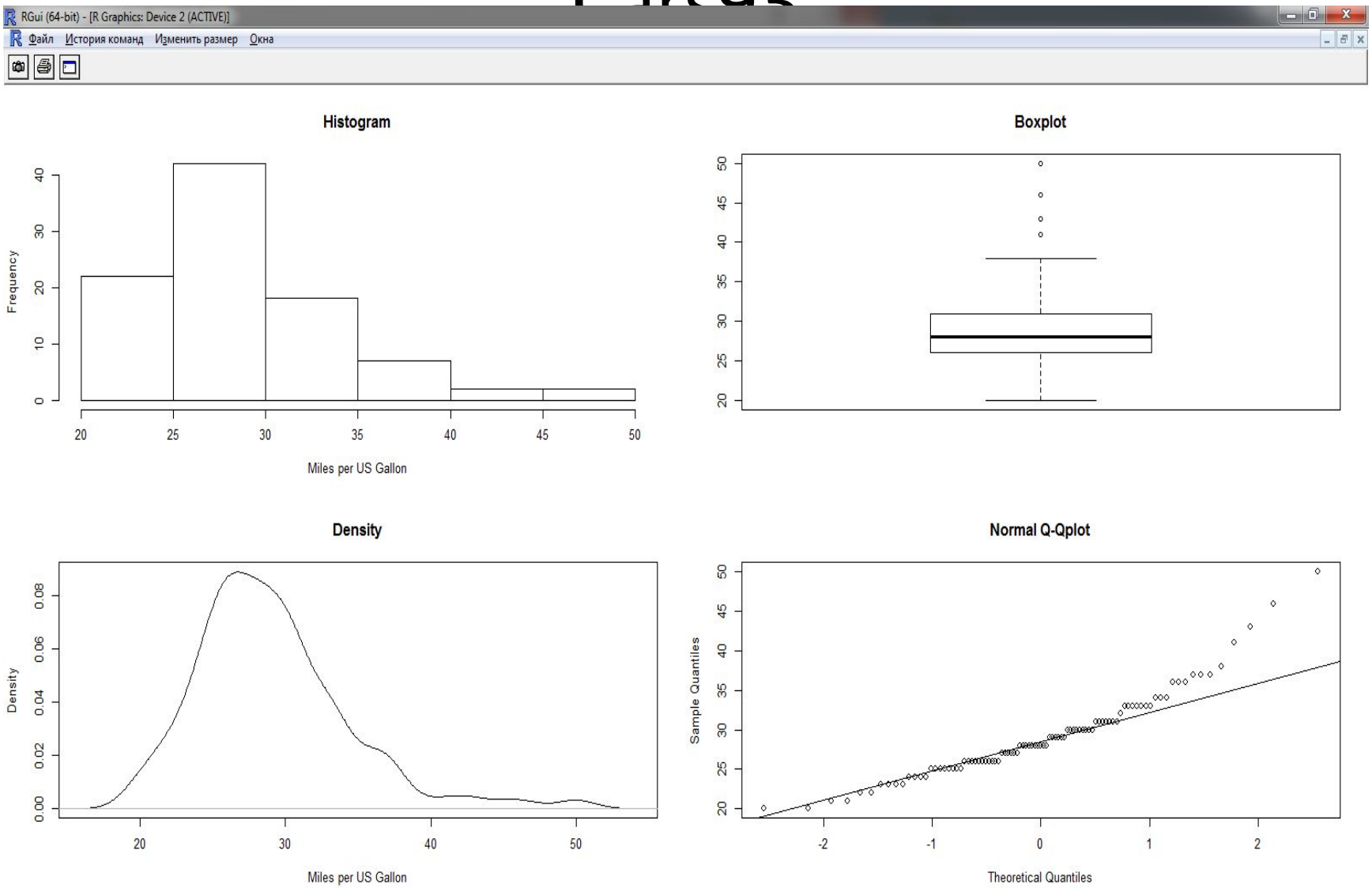
Набор данных Cars93

- Manufacturer - производитель
- Model - модель
- Type –
  тип: "Small", "Sporty", "Compact", "Midsize", "Large" и "Van".
- Min.Price – минимальная цена (в $1,000)
- PriceMidrange – средняя цена (в $1,000).
- Max.Price – максимальная цена (в $1,000).
- MPG.city – MPG в городе(количество миль за галлон топлива)
- MPG.highway – MPG на трассе
- и т.д.

Графики для набора данных Cars93

```
> attach(Cars93)
> par(mfrow=c(2,2))
# Histogram
> hist(MPG.highway,xlab="Miles per US Gallon",
main="Histogram")
# Boxplot
> boxplot(MPG.highway,main="Boxplot")
# Density
> plot(density(MPG.highway),type="l",
xlab="Miles per US Gallon",main="Density")
# Q-Q Plot
> qqnorm(MPG.highway,main="Normal Q-Qplot")
> qqline(MPG.highway)
```

Графики для набора данных Cars93



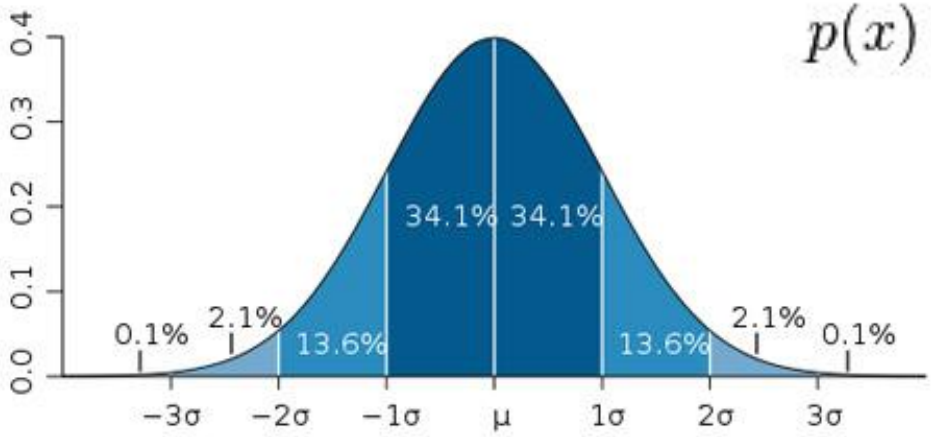
Нормальное распределение

- Нормальное распределение, также называемое распределением Гаусса, — распределение вероятностей, которое играет важнейшую роль во многих областях знаний, особенно в физике. Физическая величина подчиняется нормальному распределению, когда она подвержена влиянию огромного числа случайных помех. Ясно, что такая ситуация крайне распространена, поэтому можно сказать, что из всех распределений, в природе чаще всего встречается именно нормальное распределение — отсюда и произошло одно из его названий.
- Нормальное распределение зависит от двух параметров —*смещения* и *масштаба*, то есть, является, с математической точки зрения, не одним распределением, а целым их семейством. Значения параметров соответствуют значениям среднего (математического ожидания) и разброса (стандартного отклонения).

Нормальное распределение

- Стандартным нормальным распределени-ем называется нормальное распределение с математическим ожиданием 0 и стандартным отклонением 1.
- Плотность вероятности нормально распределённой случайной величины с параметром смещения μ и масштаба σ^2 (или, что тоже самое, дисперсией) имеет следующий вид:

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$



Нормальное распределение

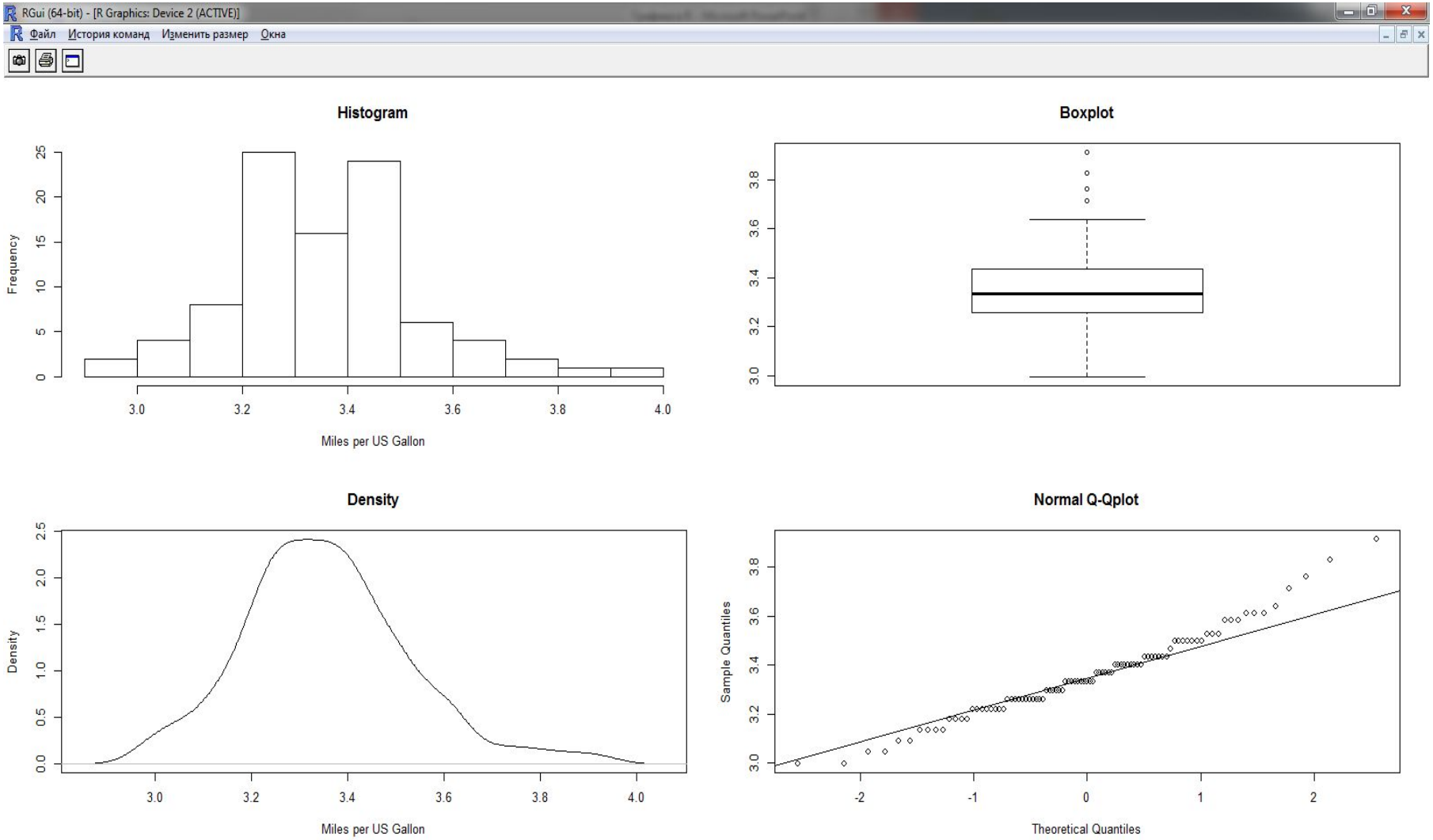
- Важное значение нормального распределения во многих областях науки (например, в математической статистике и статистической физике) вытекает из центральной предельной теоремы теории вероятностей. Если результат наблюдения является суммой многих случайных слабо взаимозависимых величин, каждая из которых вносит малый вклад относительно общей суммы, то при увеличении числа слагаемых распределение центрированного и нормированного результата стремится к нормальному. Этот закон теории вероятностей имеет следствием широкое распространение нормального распределения, что и стало одной из причин его наименования.

Анализ графиков

- Гистограмма и график плотности асимметричны. Это говорит о том, что распределение переменной MPG.highway отличается от нормального. Попробуем нормализовать его с помощью преобразования:

```
> log(MPG.highway)
```

Результат нормализации

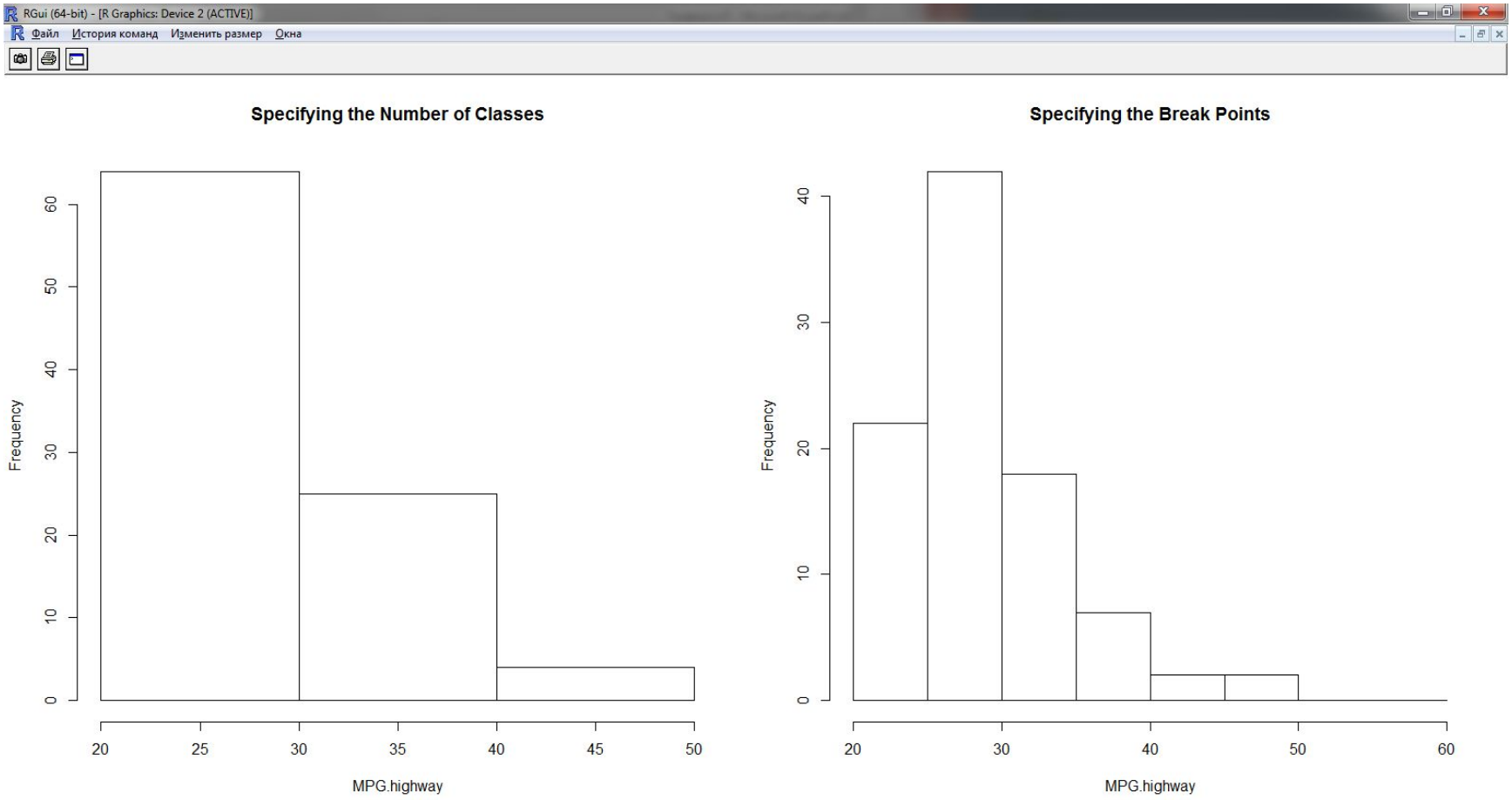


Гистограммы

- В гистограмме данные разбиты на интервалы, и каждый интервал изображен в виде столбика, высота которого пропорциональна количеству точек данных, попавших в этот интервал. Количество интервалов (классов) или их границы можно задавать.

```
> par(mfrow=c(1,2))
> hist(MPG.highway,nclass=4,main="Specifying the
  Number of Classes")
> hist(MPG.highway,breaks=seq(from=20,to=60,by=5),
main="Specifying the Break Points")
> par(mfrow=c(1,1))
```

Гистограммы



Квантиль, квартиль и медиана

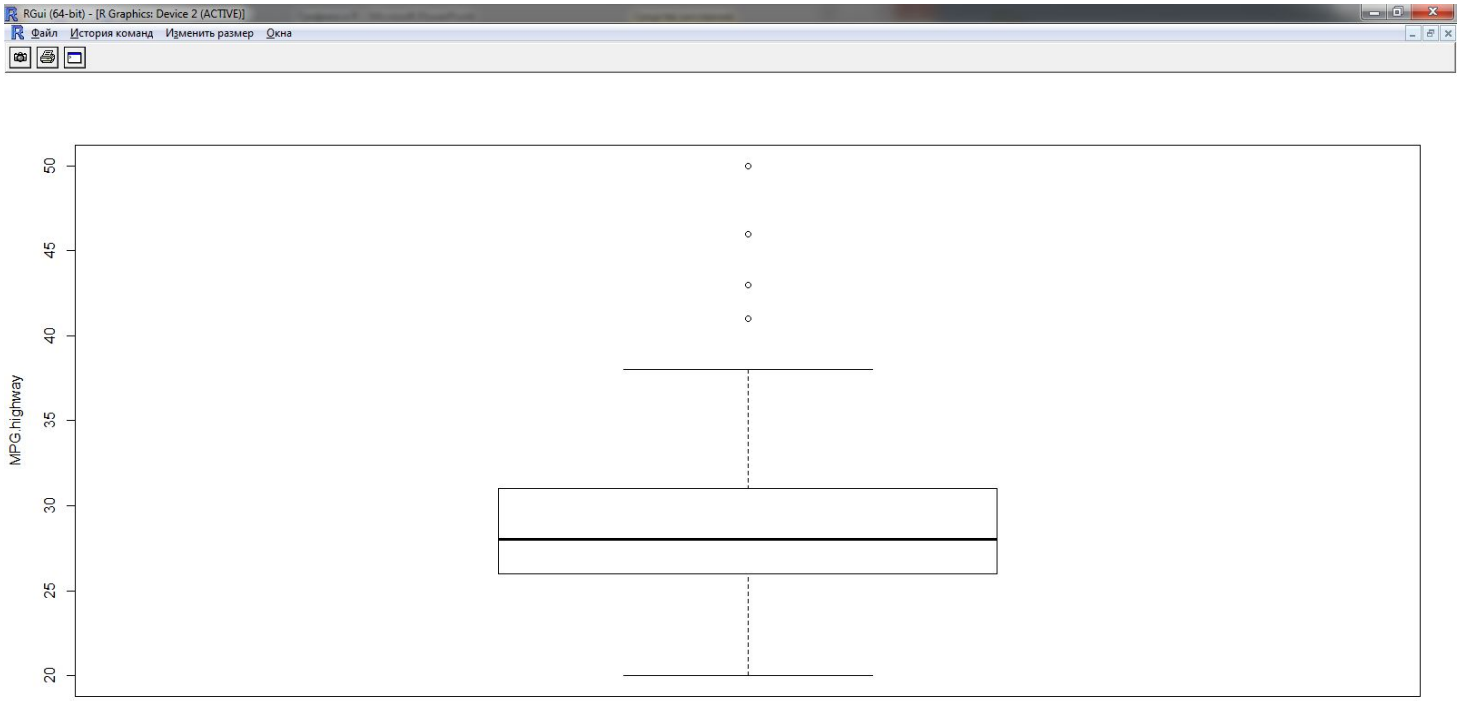
- Квантиль в математической статистике — значение, которое заданная случайная величина не превышает с фиксированной вероятностью.
- 0,25-квантиль называется первым (или нижним) квартилем ;
- 0,5-квантиль называется медианой (или вторым) квартилем;
- 0,75-квантиль называется третьим (или верхним) квартилем.
- Интерквартильным размахом называется разность между третьим и первым квартилями. Интерквартильный размах является характеристикой разброса распределения величины и является робастным (устойчивым к выбросам) аналогом дисперсии. Вместе, медиана и интерквартильный размах могут быть использованы вместо математического ожидания и дисперсии в случае распределений с большими выбросами, либо при невозможности вычисления последних.

Ящичковые диаграммы

- Ящичковые диаграммы обобщают данные и выводят в определенном виде (box and whisker formation).
- Прямоугольник представляет интервартильный размах (IQR) и показывает медиану (линия), первый (нижний край прямоугольника) и третий квартили (верхний край прямоугольника) распределения. Минимальное и максимальное значения показаны усиками (линиями, которые выходят за пределы прямоугольника к минимальной и максимальной точкам).
- Если расстояние между минимальным значением и первым квартилем превышает 1.5 × IQR, то усики продолжаются от нижнего квартиля к наименьшему значению в пределах 1.5 × IQR. Крайние точки, выходящие за этот предел, изображаются точками. Похожая процедура проводится и для расстояний между максимальным значением и третьим квартилем.

Ящичковая диаграмма

```
> summary(MPG.highway}
Min. 1st Qu. Median Mean 3rd Qu. Max.
20.00 26.00 28.00 29.09 31.00 50.00
> boxplot(MPG.highway,ylab="MPG.highway")
```



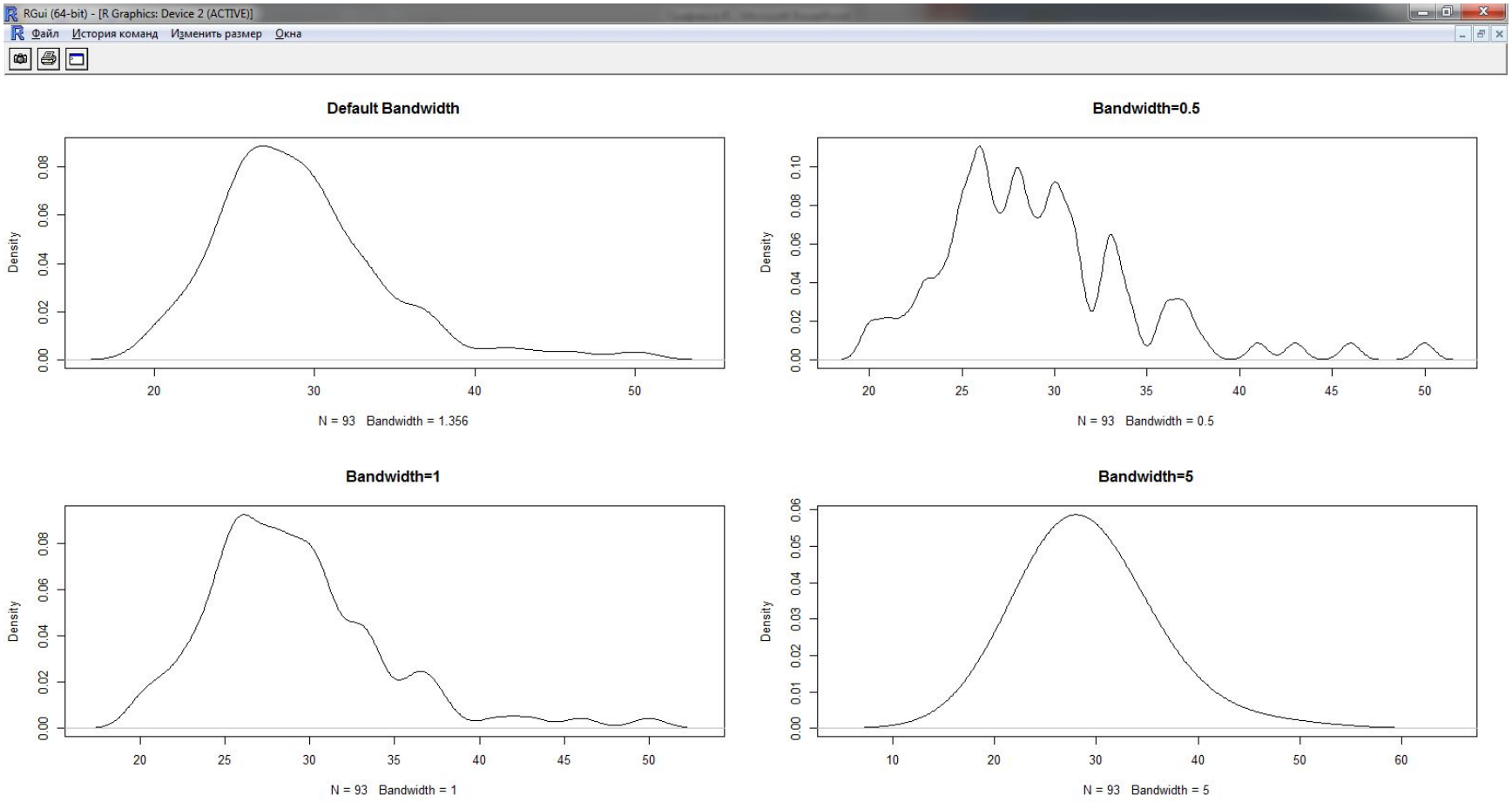
Плотность

- Плотность вероятности – характеристика ряда распределения, показывающая, сколько единиц совокупности приходится на единицу интервала.
- Плотность используется для вычисления сглаженных представлений наблюдаемых данных. Функция плотности в R дает оценку плотности для заданного ядра (распределения) и пропускной способности (bandwidth).
- По умолчанию используется Гауссово ядро, но можно использовать и другое. Пропускная способность отвечает за степень сглаживания. По умолчанию, пропускная способность принимается равной СКО ядра, но ее также можно менять.

Сглаженные плотности для Cars93

```
> par(mfrow=c(2,2))
> plot(density(MPG.highway),type="l",
main="Default Bandwidth)
> plot(density(MPG.highway,bw=0.5),type="l",
main="Bandwidth=0.5")
> plot(density(MPG.highway,bw=1),type="l",
main="Bandwidth=1")
> plot(density(MPG.highway,bw=5),type="l",
main="Bandwidth=5")
> par(mfrow=c(1,1))
```

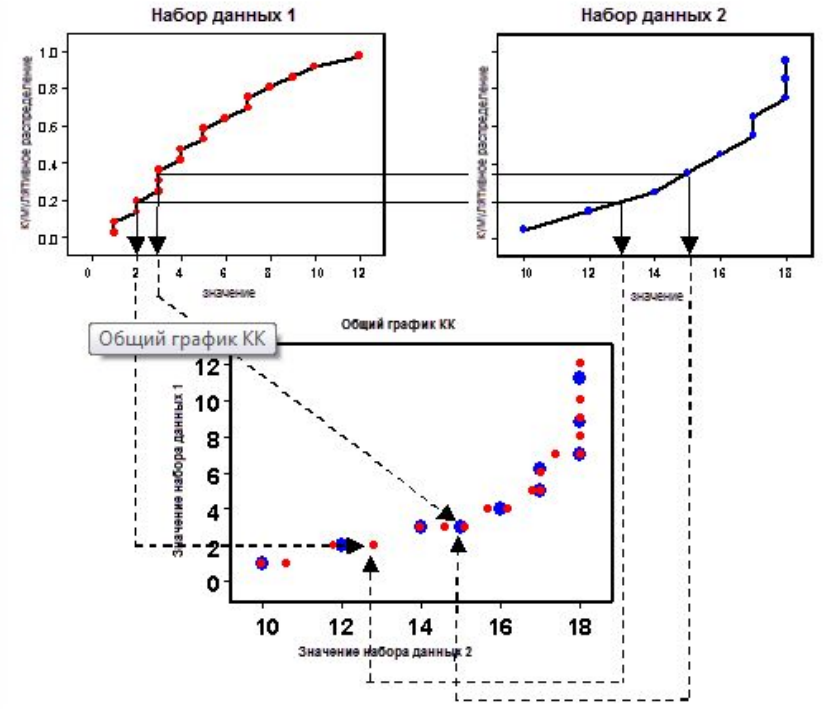
Сглаженные плотности для Cars93



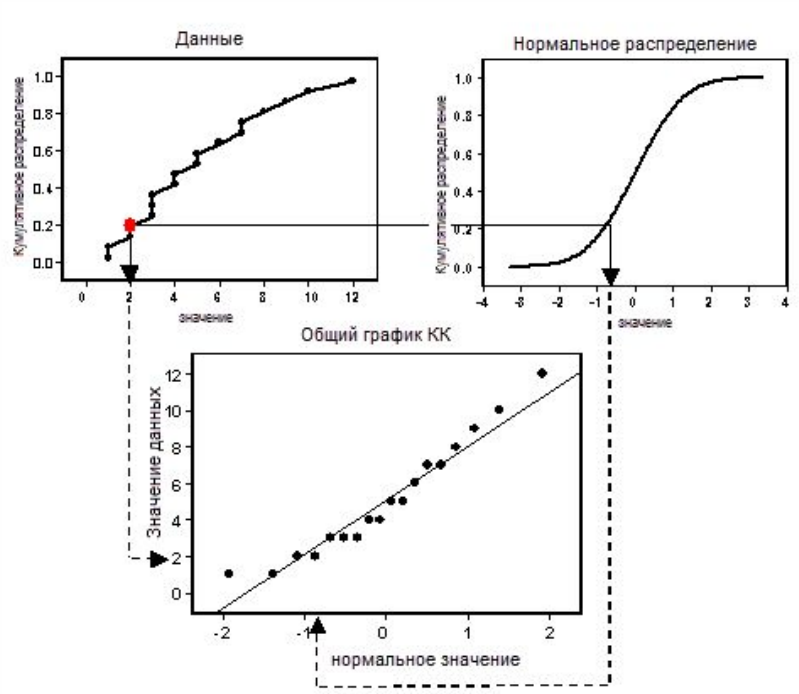
Нормальный график КК (квантиль-квантиль)

- Графики квантиль-квантиль (КК) — это графики, на которых квантили из двух распределений расположены относительно друг друга. Такими графиками удобно пользоваться для проверки предположений относительно свойств распределений данных. На них изображен график квантилей одного распределения в сравнении с другим и, возможно, добавлена линия, изображающая теоретические квантили интересующего распределения. Если распределения одной формы, то точки примерно попадут на прямую линию.
- Крайние точки отличаются большей вариабельностью, чем точки в центре, следовательно, можно ожидать, что верхняя и нижняя части графика будут отклоняться от этой линии.

Примеры построения общего и нормального графиков КК



Пример общего графика КК

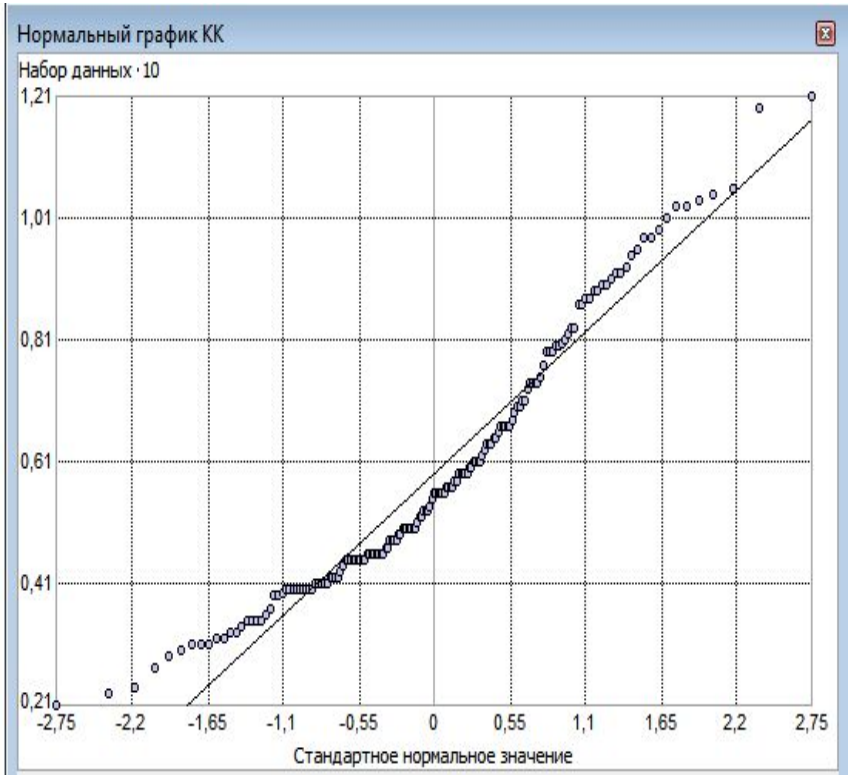


Пример нормального графика КК

Значения кумулятивного распределения вычисляются как $(i — 0{,}5)/n$ для i-го упорядоченного значения из n общих значений.

Проверка распределения данных с помощью графиков КК

- Точки нормального графика КК дают представление об одномерной нормальности набора данных. Если данные распределены нормально, точки выстроятся на базовой линии, проходящей под углом 45 градусов. Если данные не распределены нормально, точки отклоняться от базовой линии.

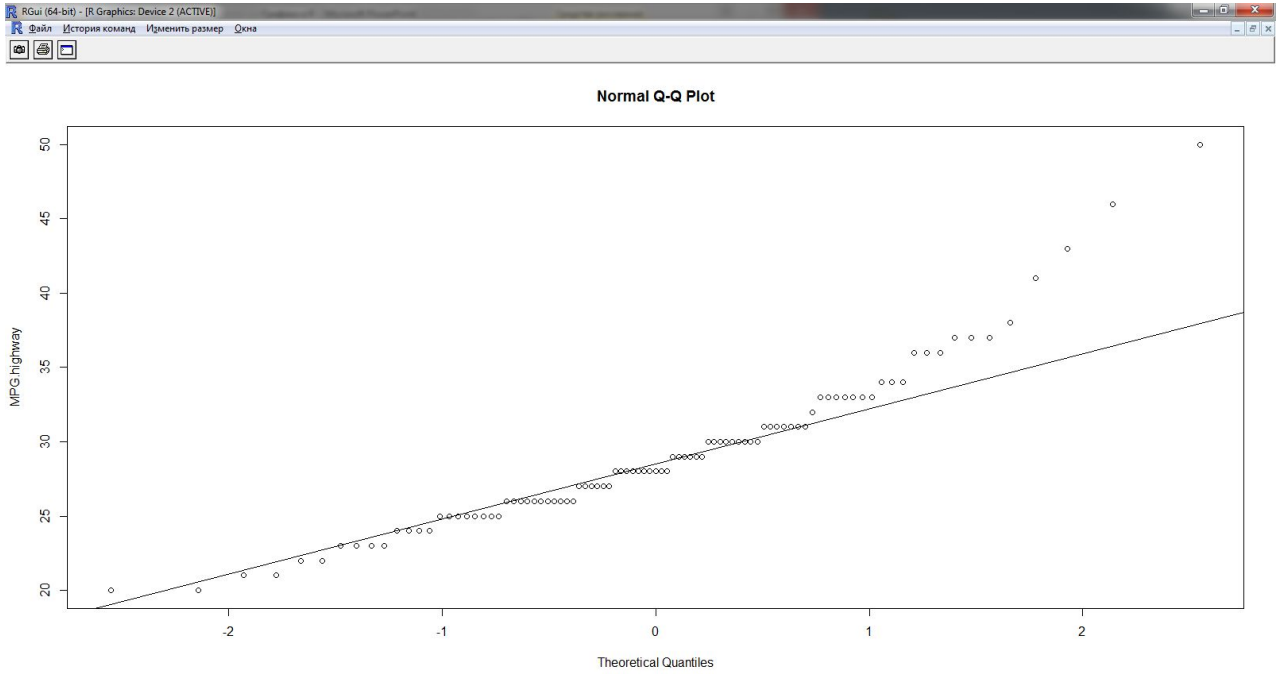


Функции qqnorm и qqline

- Функция qqnorm сравнивает квантили наблюдаемых данных с квантилями нормального распределения. Функция qqline добавит на график квантилей линию, основанную на квантилях теоретического нормального распределения.

```
> qqnorm(MPG.highway,ylab="MPG.highway")
> qqline(MPG.highway)
```

- Видно отклонение от нормальности

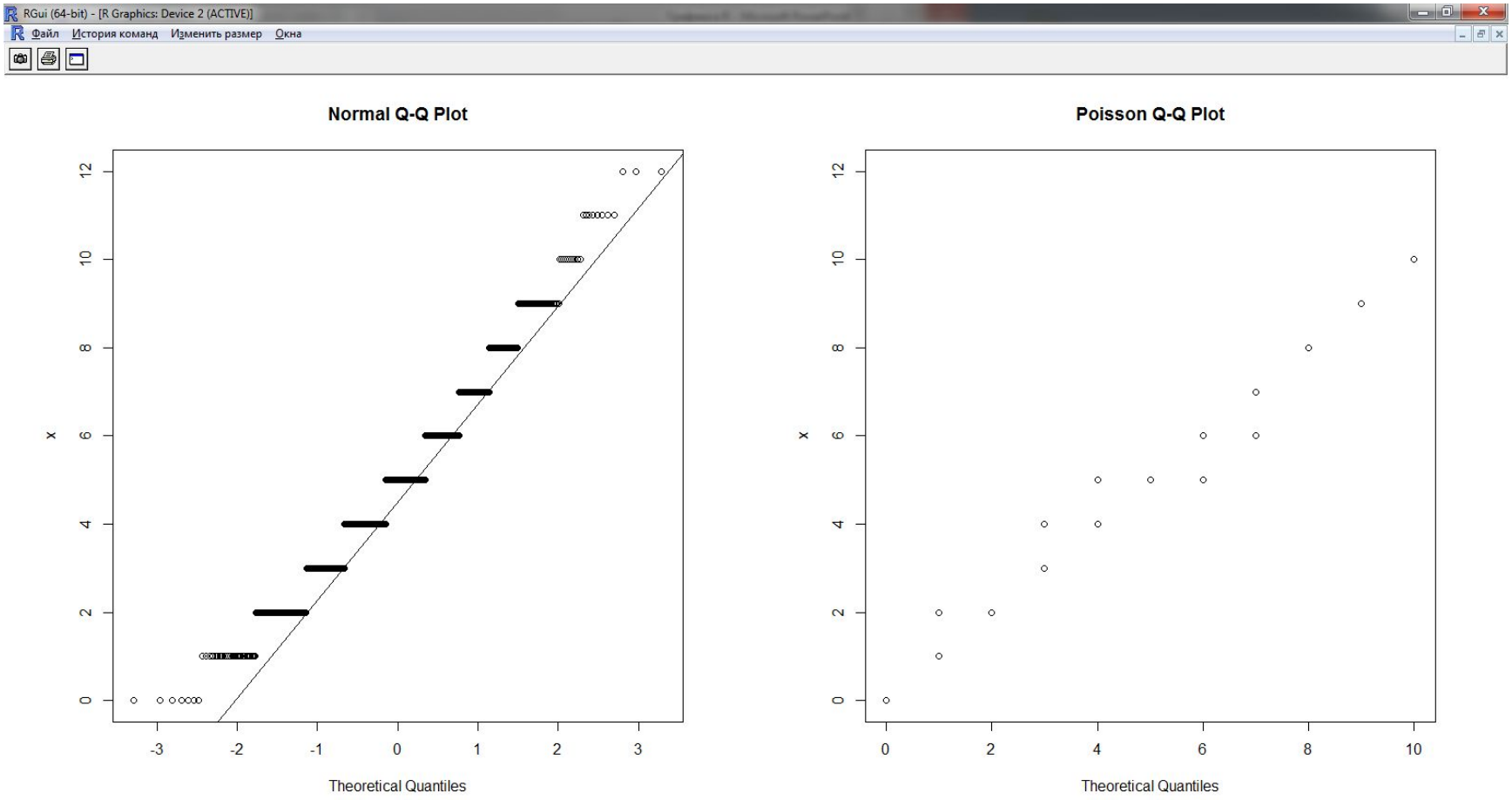


Функция qqplot

- Для сравнения данных выборки с данными других распределений, используется функция qqplot.

```
# Generating Data from Poisson Distribution
> x <- rpois(1000,lambda=5)
> par(mfrow=c(1,2),pty="s")
# Comparing against a Normal
> qqnorm(x,ylab="x")
> qqline(x)
# Comparing against a Poisson
> qqplot(qpois(seq(0,1,length=50),
lambda=mean(x)),x,
xlab="Theoretical Quantiles",ylab="x")
> title(main="Poisson Q-Q Plot")
```

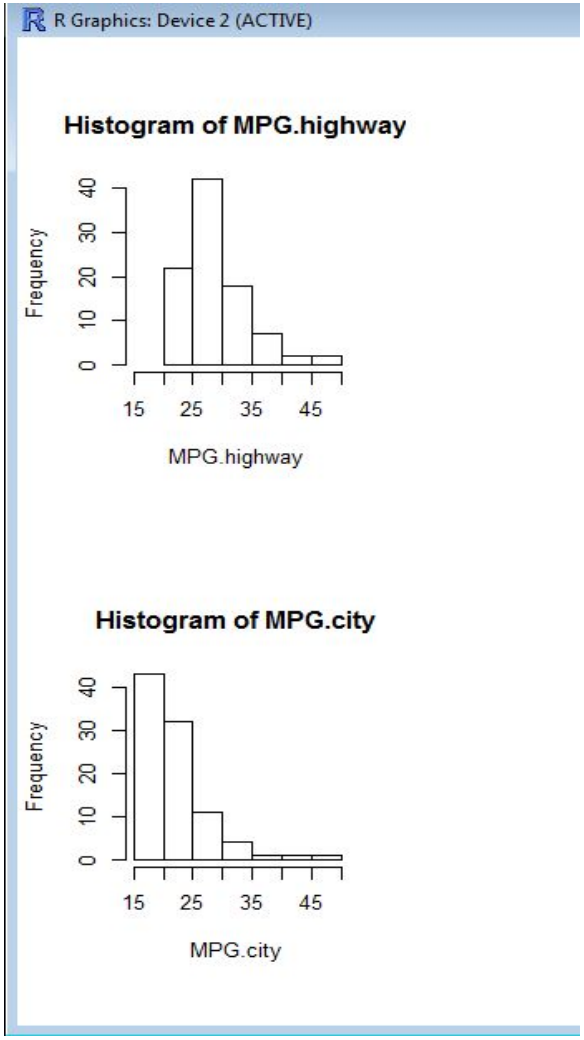
Сравнение с распределением Пуассона



Сравнение групп

- Для сравнения свойств распределений различных наборов данных можно использовать следующие графические средства:
- Множественные диаграммы, изображенные на одной шкале
- Ящичковые диаграммы, разбитые по группам
- Графики квантиль-квантиль

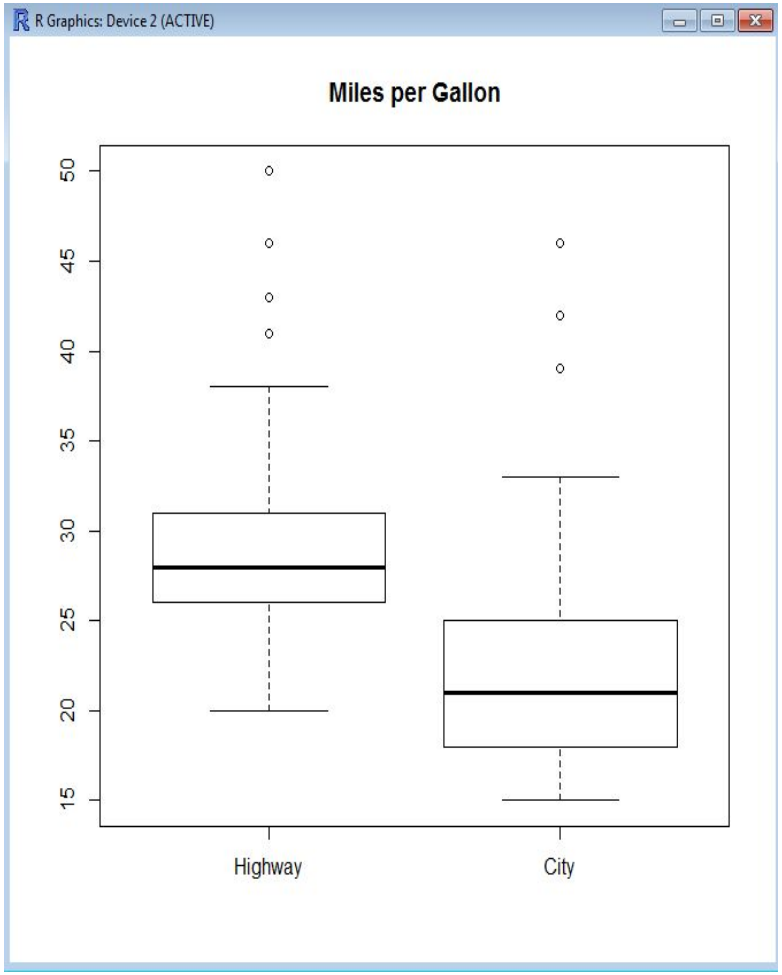
```
# Set up plotting region
> par(mfcol=c(2,2))
# Produce histograms to compare each dataset
> hist(MPG.highway,
     xlim=range(MPG.highway,MPG.city))
>hist(MPG.city,xlim=range(MPG.highway,MPG.city))
```



Сравнение групп

Produce boxplot split by type of driving

```
>boxplot(list(MPG.highway,MPG.city),names=c("Highway", "City"),main="Miles per Gallon")
```

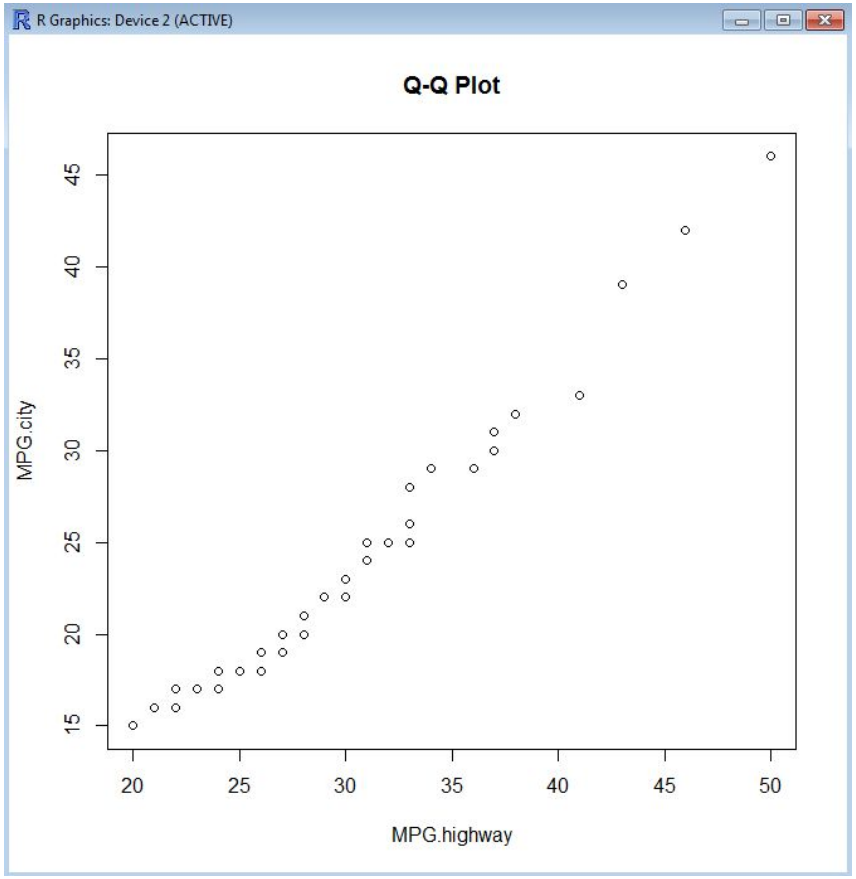


Сравнение групп

Q-Q plot to check distribution shape and scale

>qplot(MPG.highway, MPG.city, main="Q-Q Plot")

> par(mfrow=c(1,1))



Отображение двумерных данных

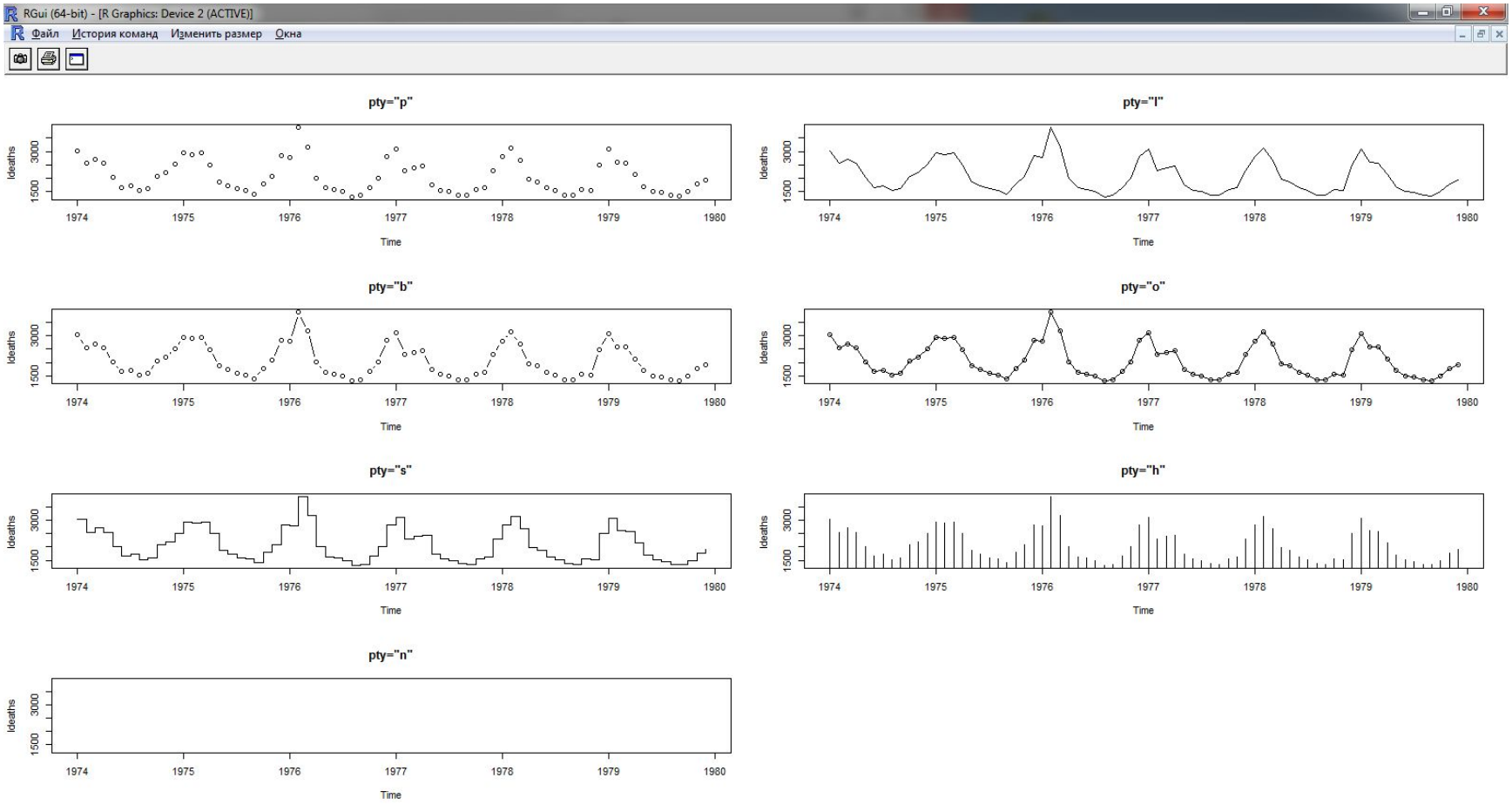
- Самый простой способ – построение диаграммы рассеяния с помощью функции plot. Аргумент type позволяет строить различные типы графиков.
- type="p": рисует символ в каждой точке
- type="l": рисует линию, соединяющую точки
- type="b": рисует и символы и линии
- type="o": рисует линии и сверху символы
- type="s": рисует ступеньки
- type="h": рисует гистограммо-подобные вертикальные линии
- type="n": не рисует ни точек ни линий

Изображение набора данных ldeath

- Набор данных ldeaths – временной ряд, содержащий данные о ежемесячном количестве смертей от бронхита, эмфиземы и астмы для мужчин и женщин в Великобритании в период с 1974 по 1979.

```
> par(mfrow=c(4,2))
> plot(ldeaths,type="p",main='pty="p"')
> plot(ldeaths,type="l",main='pty="l"')
> plot(ldeaths,type="b",main='pty="b"')
> plot(ldeaths,type="o",main='pty="o"')
> plot(ldeaths,type="s",main='pty="s"')
> plot(ldeaths,type="h",main='pty="h"')
> plot(ldeaths,type="n",main='pty="n"')
```

Изображение набора данных ldeath

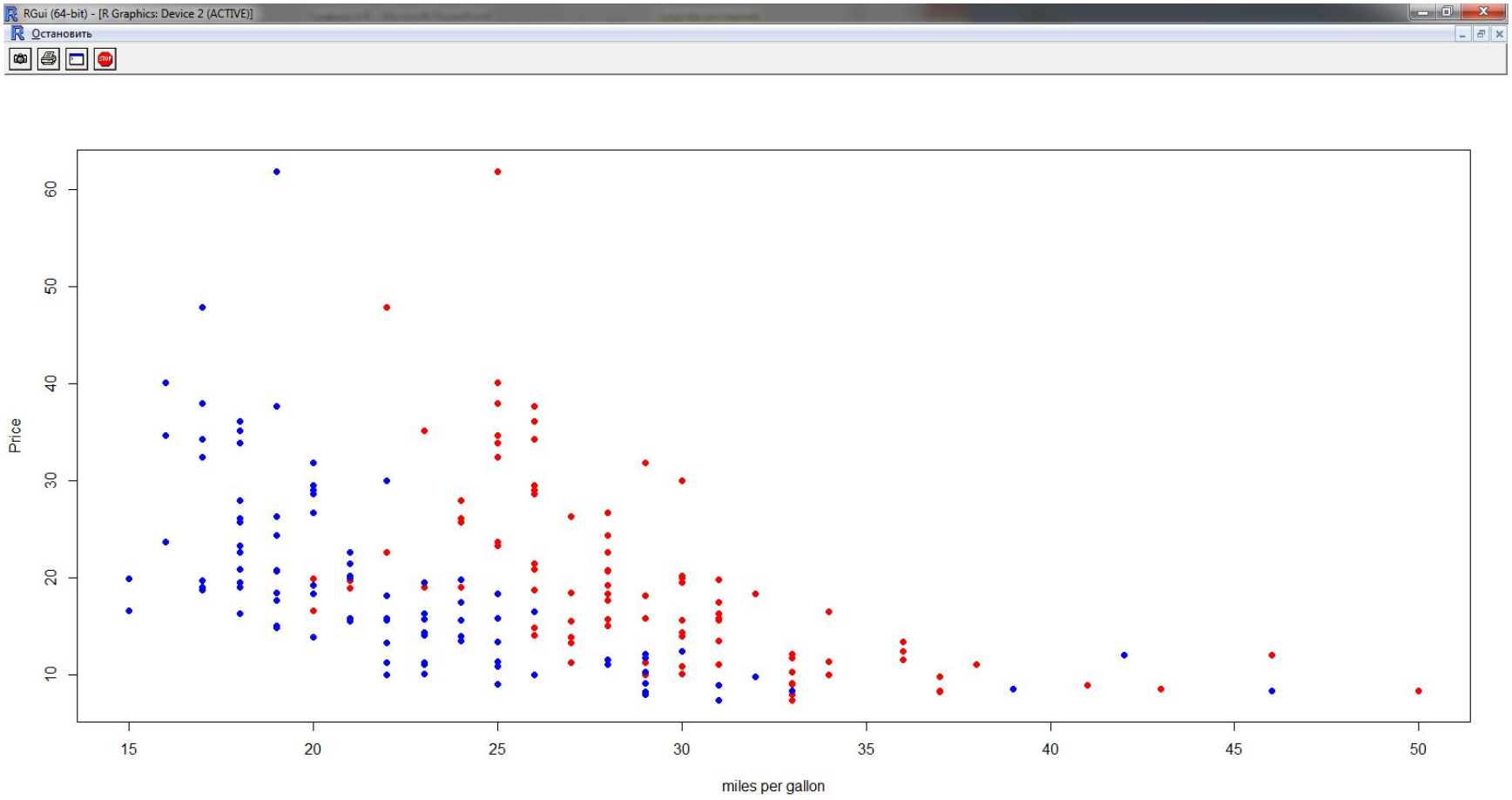


Добавление точек

- На существующий график точки добавляются с помощью функции point.

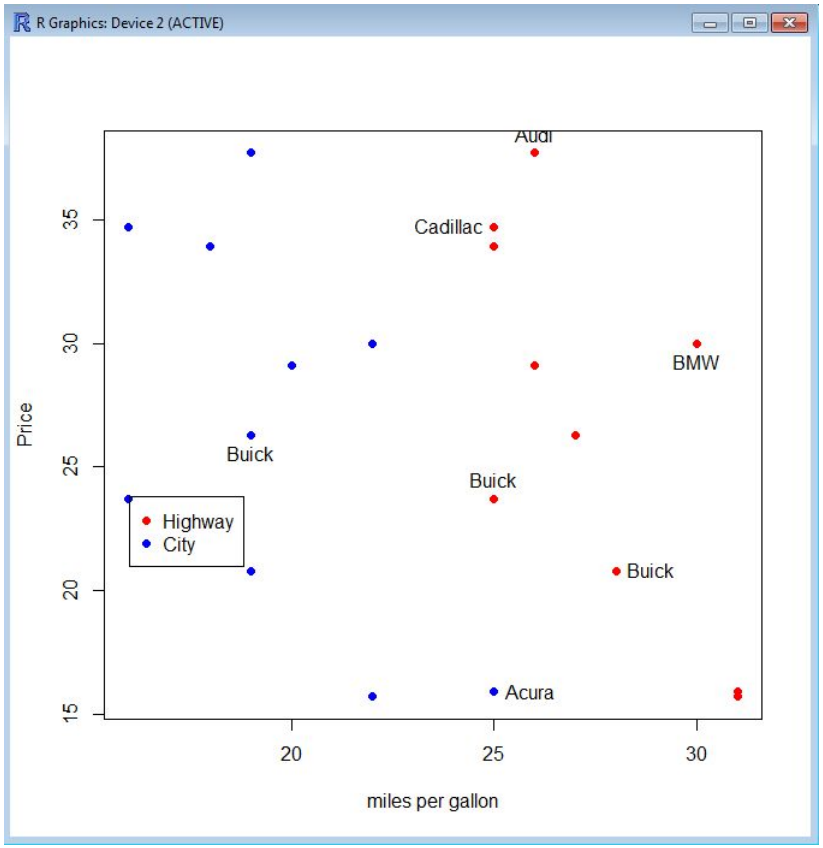
```
# Set up plotting region
> plot(MPG.highway,Price,type="n",
xlim=range(MPG.highway,MPG.city),
xlab="miles per gallon")
> points(MPG.highway,Price,col="red",pch=16)
>points(MPG.city,Price,col="blue",pch=16)
> legend(locator(1),c("Highway","City"),
   col=c("red","blue"),pch=16,bty="n")
```

Добавление точек



Интерактивный выбор точек

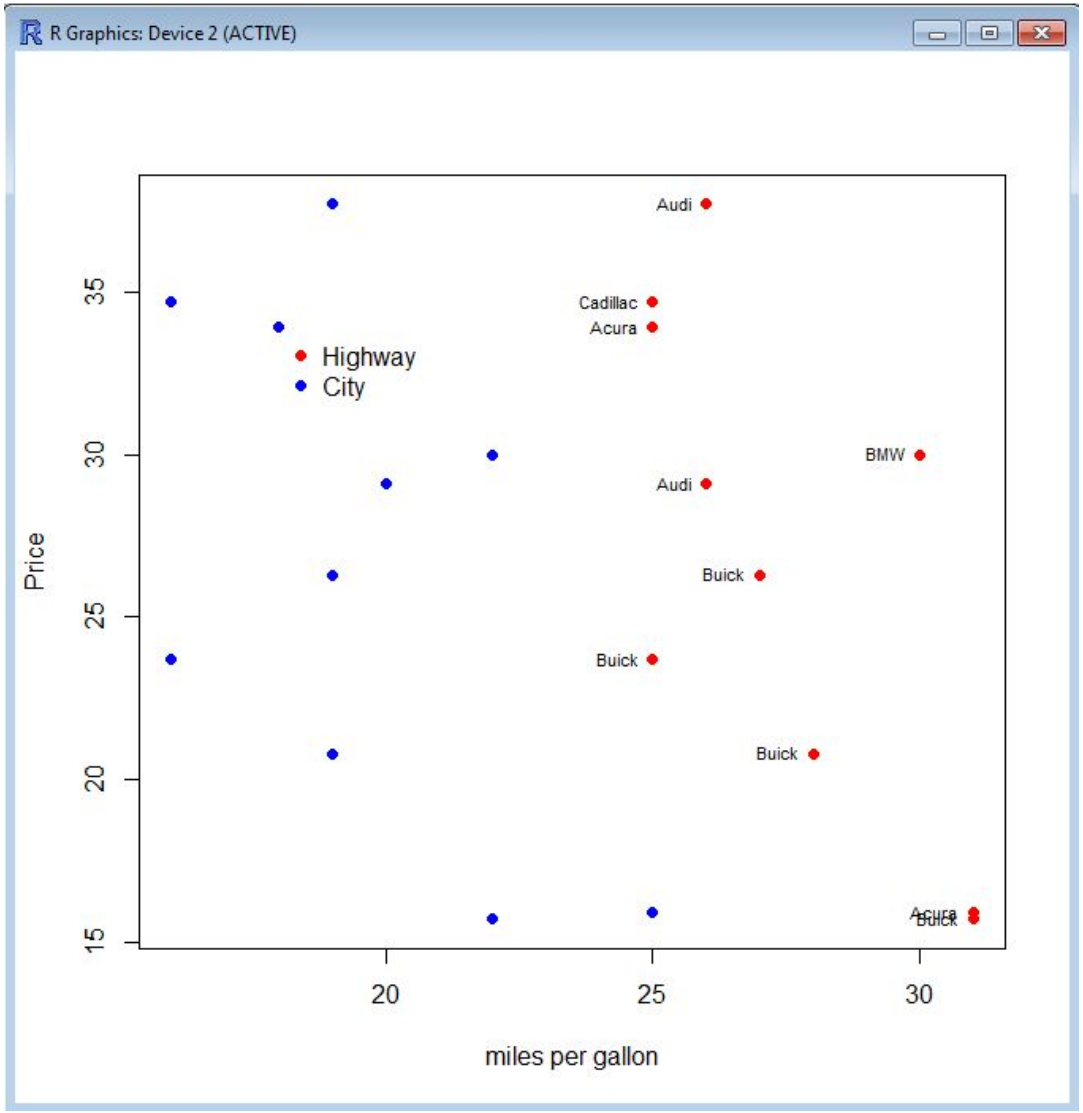
```
>identify(c(MPG.city[1:10],MPG.highway[1:10]),
   rep(Price[1:10],2),rep(Manufacturer[1:10],2),
   pos=2)
```



Добавление текста

```
> plot(MPG.highway[1:10],Price[1:10],type="n",
ylab="Price",xlim=range(MPG.highway[1:10],
MPG.city[1:10]),xlab="miles per gallon")
> points(MPG.highway[1:10],Price[1:10],col="red",pch=16)
> points(MPG.city[1:10],Price[1:10],col="blue",pch=16)
> legend(locator(1),c("Highway","City"),
col=c("red","blue"),pch=16,bty="n")
# label highway data
> text(MPG.highway[1:10],Price[1:10],Manufacturer[1:10],
cex=0.7,pos=2)
```

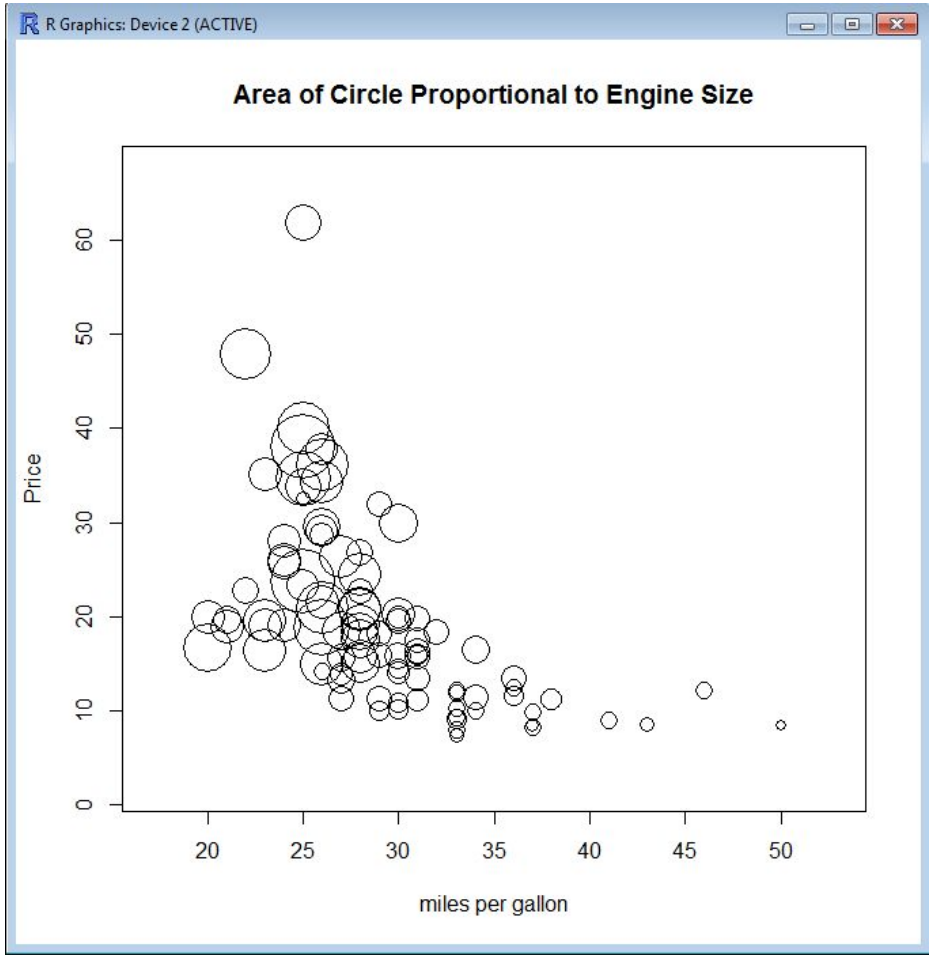
Добавление текста



Добавление символов

- Добавим на график символы, которые представляют размер двигателя. Получим график, который показывает, что машины с большим двигателем имеют тенденцию стоить дороже и иметь меньший MPG, чем другие машины.
  >symbols(MPG.highway,Price,circles=EngineSize, xlab="miles per gallon",ylab="Price",inches=0.25, main="Area of Circle Proportional to Engine Size")

Добавление символов



Добавление линий

- Добавим на график линии. Это можно сделать с помощью функции lines, которая добавляет линию, соединяющую определенные точки, или функцию abline, которая добавляет вертикальную, горизонтальную или прямую линию с определенными углом наклона и пересечением с осью x.
- Построим график MPG в сравнении с весом, с добавленными линиями : (1) наименьшего сглаживания, (2) линия регрессии наименьших квадратов с помощью lines и (3) линия регрессии наименьших квадратов с помощью abline.

Добавление линий

```
# Adding Lines
with(Cars93, {
    plot(Weight,100/MPG.city,pch=16)
    lines(lowess(Weight,100/MPG.city),col="red")
    lines(lsfit(Weight,100/MPG.city),col="blue")
    abline(coef(lsfit(Weight,100/MPG.city)),col="blue")
    xy <- par("usr")[c(1,4)]
    legend(xy[1], xy[2],c("Lowess Smoother","Least Squares"),
   col=c("red","blue"),lty=1,bty="n")
})
```

Добавление линий