Тема 5. Различные виды диаграмм в R. Гистограммы, диаграммы размахов. Столбчатые и круговые диаграммы

Помимо функции *plot()* в R есть и другие функции высокого уровня. Они позволяют строить диаграммы размахов, гистограммы, столбчатые и круговые диаграммы.

Диаграммы размахов, гистограммы

Диаграмма размахов («ящик с усами») иллюстрирует распределение непрерывной величины, отображая описательную статистику: максимальное и минимальное значение, медиану, верхний и нижний квартили. Диаграмма также иллюстрирует наличие выбросов – объектов, значения которых выходят за пределы в ±1,5 межквартильных размаха. Межквартильный размах –это разность между верхним и нижним квартилями. Сравнение плотности распределения и «ящика с усами» показано на рисунке 6.

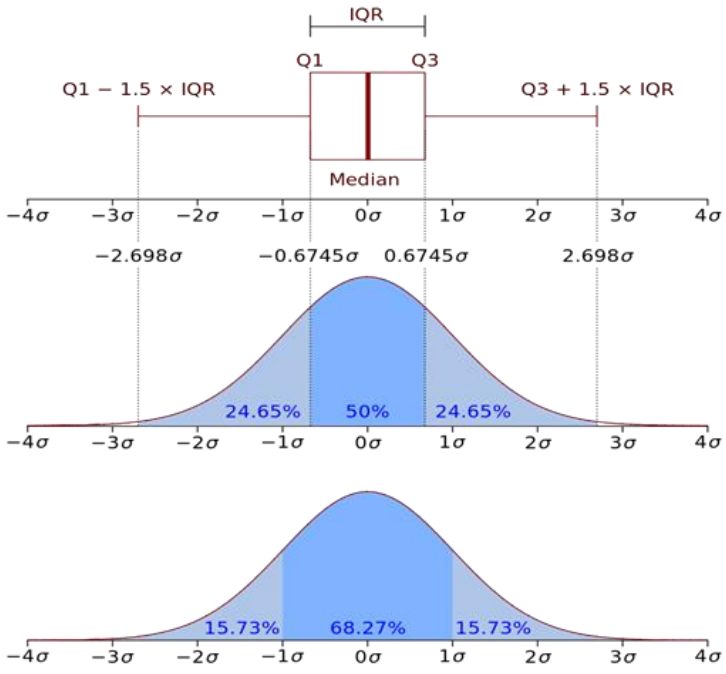

Рисунок 6. Сравнение плотности распределения и «ящика с усами».

Построить диаграмму размахов позволяет функция ***boxplot***. Рассмотрим особенности её использования на конкретном примере.

Загрузим существующий в R набор данных ToothGrowth при помощи команды ***data()***. Это данные о длине зубов морских свинок, которые получали апельсиновый сок или витамин С. Построим диаграмму размахов для длины зубов:

```
data(ToothGrowth)
head(ToothGrowth)

   len supp dose
1  4.2   VC  0.5
2 11.5   VC  0.5
3  7.3   VC  0.5
4  5.8   VC  0.5
5  6.4   VC  0.5
6 10.0   VC  0.5

boxplot(ToothGrowth$len, main="Распределение длины зубов у морских
свинок")
```

В результате мы получим диаграмму, представленную на рисунке 7.

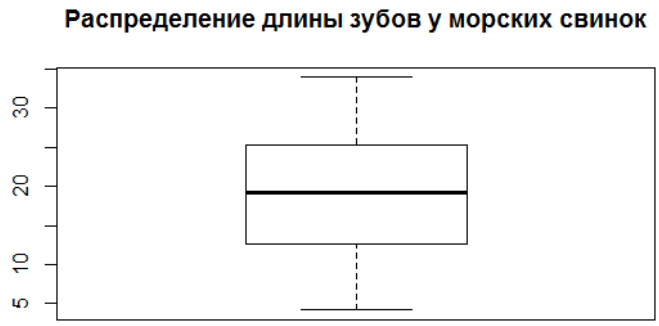

Рисунок 7. Распределение длины зубов морских свинок.

Для того чтобы построить два «ящика с усами» на одной диаграмме, которые будут описывать распределение длины зубов для свинок, получавших апельсиновый сок и витамин С, в качестве первого аргумента функции нужно указать формулу ***«название столбца 1» ~ «название столбца 2»***. В качестве значения параметра ***data*** необходимо указать название таблицы, где содержатся эти столбцы. В примере ниже значение «истина» выставлено еще для двух аргументов. Параметр ***varwidth = TRUE*** позволяет получить диаграмму, на которой ширина «ящиков» будет пропорциональна квадратному корню из объема выборки. Параметр ***notch = TRUE*** добавляет к «ящикам» насечки, ширина которых пропорциональна

ширине 95%-го доверительного интервала для медианы. Если насечки не перекрываются, то велика вероятность того, что медианы соответствующих генеральных совокупностей также различаются. Кроме того, используя вектор ***c("green","red")***, мы раскрасили «ящики» в разные цвета (рисунок 8).

```
boxplot(len~supp, data=ToothGrowth, varwidth=TRUE, notch=TRUE,
col=c("green","red"), main="Распределение длины зубов у морских
свинок")
```

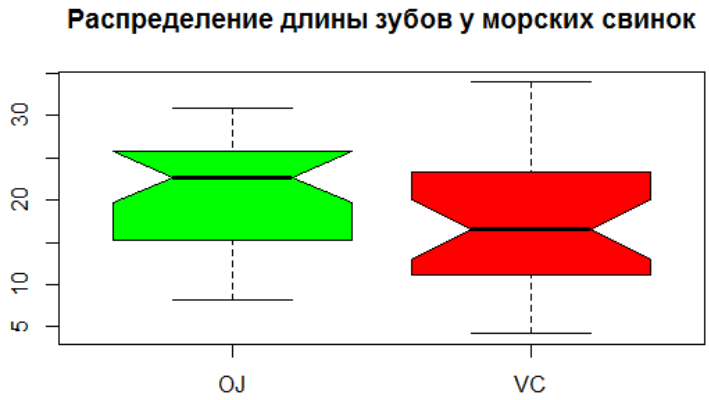

Рисунок 8. Распределение длины зубов морских свинок. Ширина насечек на «ящиках», пропорциональна ширине 95%-го доверительного интервала для медианы.

Из рисунка 8 видно, что насечки не перекрываются, следовательно, можно сделать вывод, что апельсиновый сок (OJ) сильнее влияет на рост зубов, по сравнению с витамином С (VC).

Гистограммы также описывают распределение случайной величины. Построить гистограмму можно при помощи функции ***hist(x)***, где x – числовой вектор. Построим гистограмму для длины зубов морских свинок (рисунок 9).

```
hist(ToothGrowth$len, col="lightblue")
```

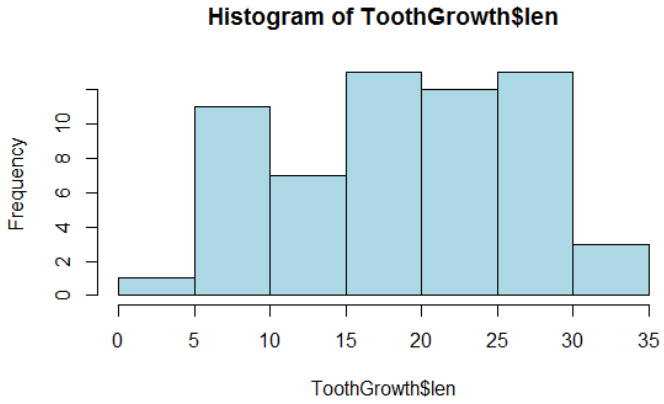

Рисунок 9. Гистограмма распределения длины зубов морских свинок.

На гистограмме по оси ординат указано количество свинок, длина зубов которых попадает в данный интервал. Если установить опцию ***freq = FALSE***, то на оси ординат будет указана плотность вероятности. Добавим к гистограмме график плотности вероятности при помощи функции низкого уровня ***lines***, вычислив плотность вероятности при помощи функции ***density***.

```
hist(ToothGrowth$len,col="lightblue", freq=FALSE, ylim=c(0,0.05))
lines(density(ToothGrowth$len), col="red", lwd=2)
```

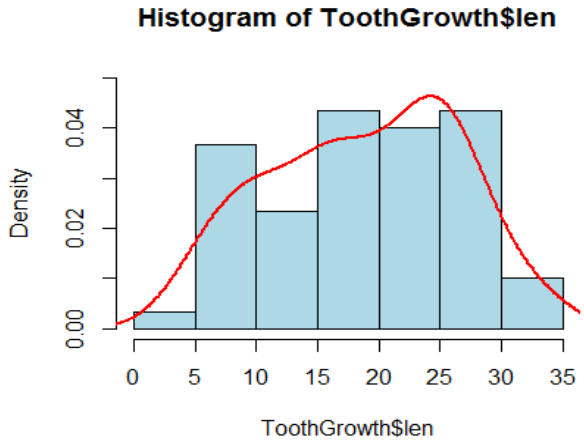

Рисунок 9. Гистограмма распределения длины зубов морских свинок. Красная линия показывает плотность вероятности.

При оформлении результатов анализа данных часто бывает необходимо объединить несколько диаграмм в одну. Этого можно добиться, воспользовавшись функцией ***par***. Функция ***par*** позволяет устанавливать графические параметры для всех рисунков, которые строит пользователь во

время R сессии, например *col, pch, cex* и др. Эти же параметры можно установить для каждого рисунка по отдельности, как это было показано во всех предыдущих примерах.

Разместим все четыре рисунка на одной диаграмме, заполнив её построчно, указав параметр *mfrow = c(2, 2)* функции ***par***. Однако вначале сохраним текущие значения графических параметров, чтобы потом можно было легко вернуться к значениям по умолчанию.

```
pr<- par(no.readonly=TRUE)
par(mfrow=c(2,2))

boxplot(ToothGrowth$len, main="Длина зубов")

boxplot(len~supp,    data=ToothGrowth,    varwidth=TRUE,    notch=TRUE,
col=c("green","red"), main=" Длина зубов")

hist(ToothGrowth$len, col="lightblue", main=" Длина зубов")

hist(ToothGrowth$len,   col="lightblue",   freq=FALSE,   main="   Длина
зубов")
lines(density(ToothGrowth$len),col="red",lwd=2)

par(pr)
```

В результате получится следующий график (рисунок 10):

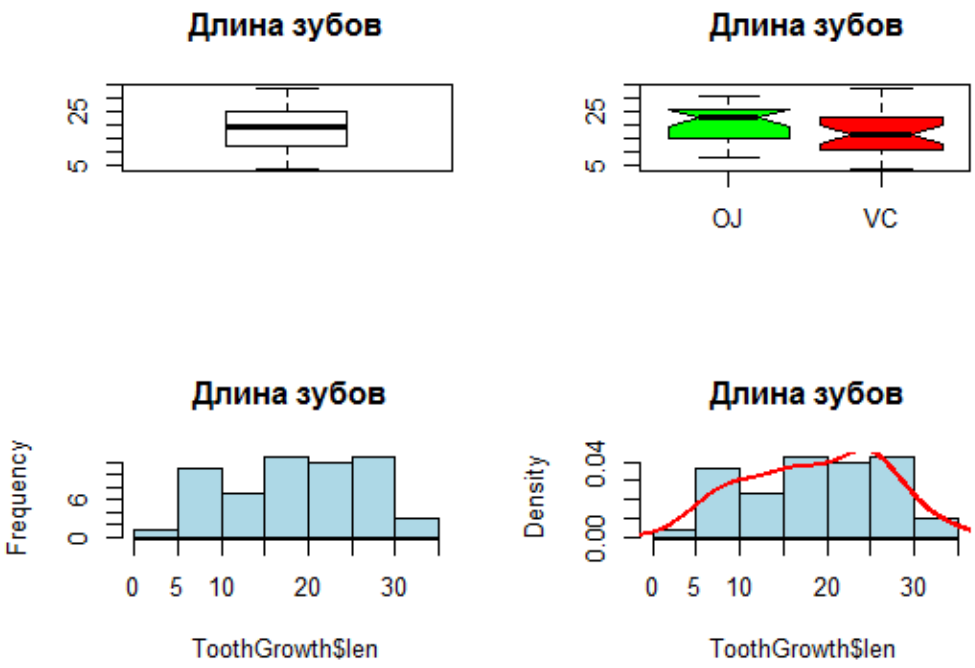

Рисунок 10. Гистограммы распределения и диаграммы размахов длины зубов морских свинок, расположенные на одном графике. Красная линия показывает плотность вероятности.

Для того чтобы построить график, где на верхней строке будет *одна* диаграмма, а на нижней строке *две*, следует воспользоваться функцией ***layout***. Вначале создадим матрицу, описывающую то, как будут располагаться диаграммы:

```
> mm<-matrix(c(1,1,2,3),ncol=2,nrow=2,byrow=T)
> mm
     [,1] [,2]
[1,]    1    1
[2,]    2    3
```

Далее подставляем её в функцию ***layout()*** и строим графики (рисунок 11):

```
layout(mm)

hist(ToothGrowth$len, col="lightblue", main="Длина зубов")

boxplot(len~supp, data=ToothGrowth, varwidth=TRUE, notch=TRUE, col=c("green", "red"), main="Длина зубов")

hist(ToothGrowth$len, col="lightblue", freq=FALSE, main="Длина зубов")

lines(density(ToothGrowth$len), col="red", lwd=2)
```

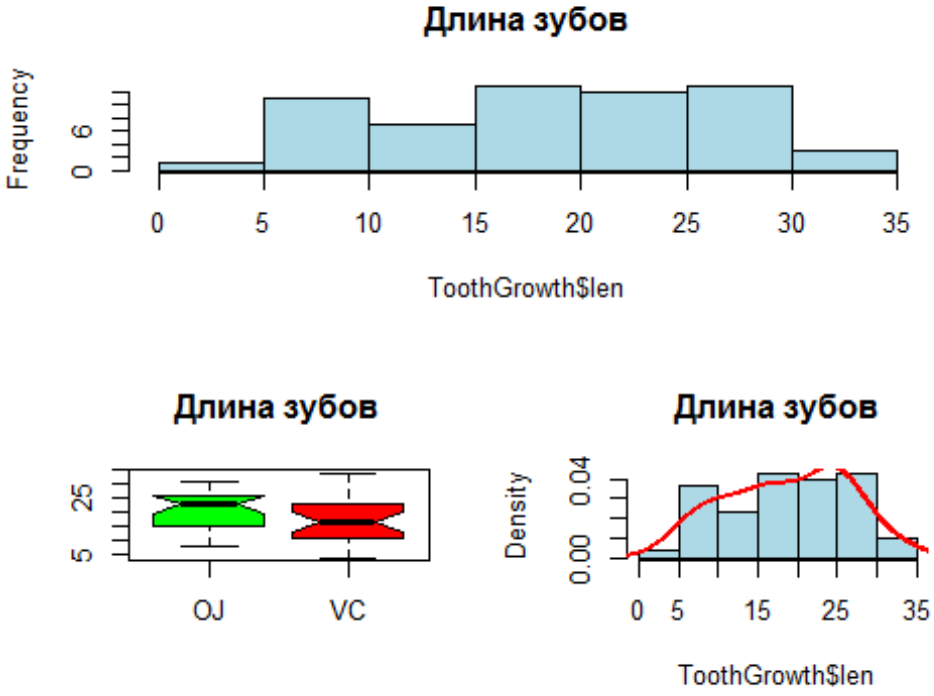

Рисунок 11. Гистограммы распределения и диаграмма размахов длины зубов морских свинок, расположенные на одном графике (1:2). Красная линия показывает плотность вероятности.

Столбчатые и круговые диаграммы

Столбчатые диаграммы отражают распределение значений категориальной переменной. Различают *простые* столбчатые диаграммы, когда анализу подвергается одна категориальная переменная, а также *составные* диаграммы и диаграммы *с группировкой*, когда проводится анализ связи двух категориальных переменных. Построить столбчатые диаграммы можно при помощи функции ***barplot***. Рассмотрим особенности использования этой функции на примере данных по степени гепатотоксичности лекарственных веществ. Загрузим соответствующие данные из файла «**hepatotoxicity.txt**»:

```
> tt<-read.delim("Hepatotoxicity.txt",as.is=T)
> head(tt)
        Generic.name    Class        Dosage
1           Abacavir Moderate          High
2       Acetaminophen Moderate          High
3           Aciclovir  Without          High
4 Adefovir Dipivoxil  Without           Low
5           Alfuzosin Moderate           Low
6         Allopurinol   Severe          High
```

Соответствующая таблица содержит несколько столбцов: название лекарства, класс гепатотоксичности (выраженная, умеренная, отсутствие) и доза (низкая, средняя, высокая). Построим простую столбчатую диаграмму для класса гепатотоксичности. Посчитать количество лекарств каждого класса гепатотоксичности можно при помощи функции ***table():***

```
> counts<-table(tt$Class)
> counts

Moderate   Severe  Without
      66       47       38
```

Если подставить полученную таблицу в качестве аргумента функции ***barplot()***, то можно получить следующий график (рисунок 12):

```
barplot(counts, col=c("red","blue","green"), cex.names=2, cex.main=2,
main="Распределение классов гепатотоксичности")
```

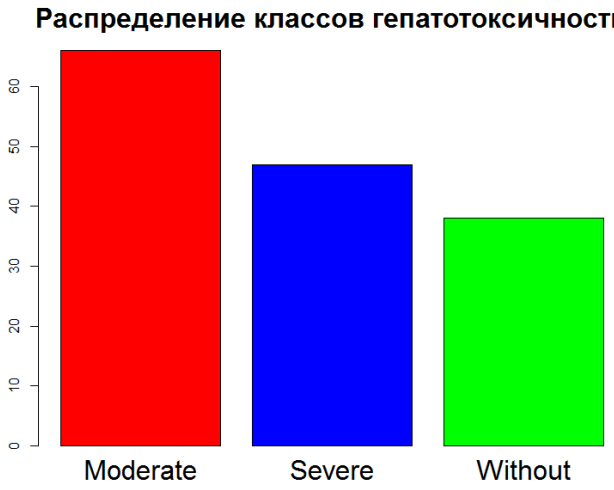

Рисунок 12. Столбчатая диаграмма распределения класса гепатотоксичности лекарственных веществ.

Для того чтобы исследовать связь между степенью гепатотоксичности и дозой лекарства, необходимо построить *составную* столбчатую диаграмму или столбчатую диаграмму *с группировкой*. Сделать это можно, рассчитав таблицу сопряженности для этих двух переменных и подставив ее в качестве аргумента функции ***barplot()***. *Таблица сопряженности* рассчитывается при помощи функции ***table()***:

```
> counts<-table(tt$Class,tt$Dosage)
> counts

          High Low Medium
  Moderate   35   9     22
  Severe     36   4      7
  Without     8  19     11
> counts<-counts[c(2,1,3),c(1,3,2)]
> counts

          High Medium Low
  Severe    36      7   4
  Moderate  35     22   9
  Without    8     11  19
```

Помещаем таблицу сопряженности в переменную *count*, а потом меняем местами строки и столбцы, чтобы степени гепатотоксичности и дозы шли по порядку. Далее строим составную столбчатую диаграмму (рисунок 13):

```
barplot(counts, col=c("red","blue","green"), cex.names=2, cex.main=2,
legend=rownames(counts), main="Зависимость степени гепатотоксичности
от дозы")
```

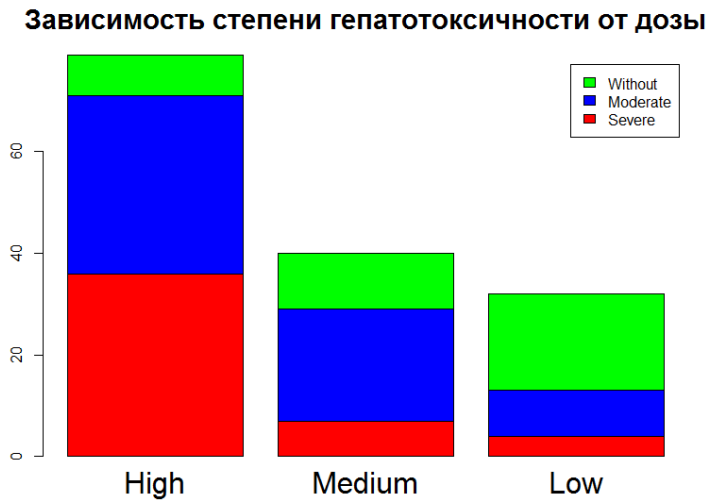

Рисунок 13. Составная столбчатая диаграмма зависимости степени гепатотоксичности лекарственных веществ от дозы.

Из диаграммы, представленной на рисунке 13, видно, что с увеличением дозы степень гепатотоксичности увеличивается.

Для того чтобы построить *столбчатую диаграмму с группировкой*, нужно установить опцию ***besides=TRUE***. Диаграмма представлена на рисунке 14.

```
barplot(counts, col=c("red","blue","green"), cex.names=2, cex.main=2, legend=rownames(counts), main="Зависимость степени гепатотоксичности от дозы", beside=TRUE)
```

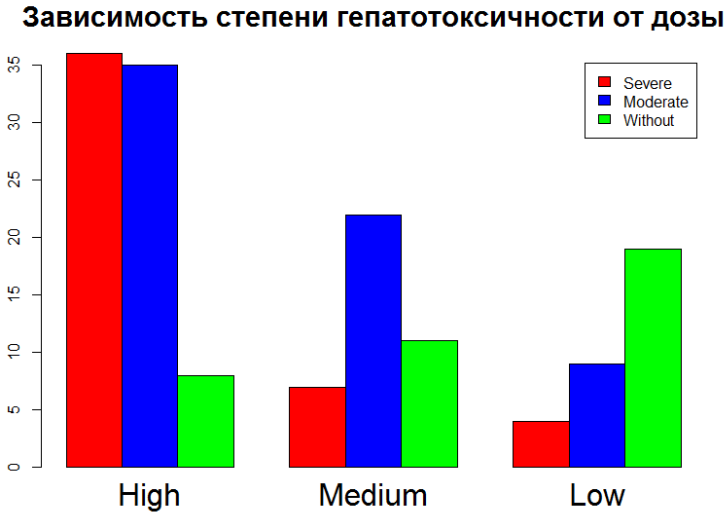

Рисунок 14. Столбчатая диаграмма с группировкой зависимости степени гепатотоксичности лекарственных веществ от дозы.

Круговую диаграмму можно построить с помощью функции ***pie().*** Построим соответствующую диаграмму для степеней гепатотоксичности (рисунок 15):

```
counts<-table(tt$Class)
pie(counts, labels=names(counts), col=c("red","blue","green"), cex=2)
```

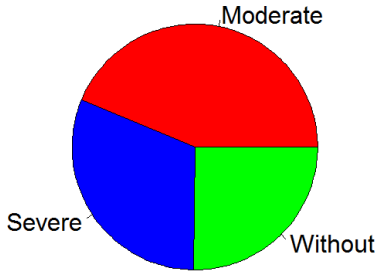

Рисунок 15. Круговая диаграмма степеней гепатотоксичности лекарственных веществ.

Практическое задание

1) В таблице из файла «**Arthritis.txt**» представлены данные по эффекту нового лекарства для лечения ревматоидного артрита. Терапевтический эффект подразделяется на три категории: market – состояние пациента значительно улучшилось, some – улучшилось незначительно, none – состояние не изменилось. Создайте простую столбчатую диаграмму для терапевтического эффекта. Добавьте заголовок, подписи осей. Раскрасьте каждый столбец в свой цвет. Подберите подходящий размер значений на осях, подписей, заголовка, их шрифт и цвет.

Создайте соответствующую круговую диаграмму. Сделайте так, чтобы подписи сегментов содержали процент от числа пациентов. Для этого сначала посчитайте проценты, а потом используйте функцию paste(), для того чтобы склеить проценты с названиями сегментов, например: ***paste("названия сегментов", " " , "проценты", "%", sep=" ")***.

Постройте составную столбчатую диаграмму и диаграмму с группировкой, чтобы на них была показана группа пациентов (пациенты, принимающие лекарство или плацебо) и степень терапевтического эффекта. Подберите для этих диаграмм подходящие параметры (заголовки, подписи осей, цвет и т.д.). Добавьте легенды. Объедините две диаграммы в одну. Есть ли связь между приемом лекарства и наблюдаемым эффектом?

Сохраните все построенные диаграммы в виде файлов с расширениями jpeg или tiff. Сделать это можно при помощи Export/Save as Image в нижнем правом окне RStudio.

2) В таблице из файла «**mRNA-protein correlation.txt**" представлены концентрации матричной РНК и белка для 4962 генов мыши. Концентрации измерены в мышиных фибробластах и представлены в виде числа молекул на клетку. Импортируйте и сохраните таблицу в переменную. Переименуйте названия столбцов, сделав их более короткими, чтобы потом можно было обращаться к столбцам по имени. Проверьте, есть ли пропущенные значения концентраций РНК или белка. Если есть, то удалите соответствующие строки. Сколько генов (строк) осталось в таблице?

Постройте гистограммы и диаграммы размахов для РНК и белка. Информативны ли соответствующие графики?

Преобразуйте исходные значения концентраций в логарифмы. Постройте снова те же самые диаграммы. Изменились ли они? Какие выводы можно сделать исходя из них? Объедините их в одну диаграмму, так чтобы на верхней строке были две гистограммы, а на нижней – диаграмма размахов. Для этого используйте функцию ***layout()***. Не забудьте сначала сохранить в переменную текущие значения параметров, например ***opar<-par(no.readonly=T)***, чтобы потом можно было восстановить исходные значения.

Постройте диаграмму зависимости логарифмов концентраций белка от мРНК, используя функцию ***plot()***. Подберите подходящие параметры (заголовки, цвет, размер точек и др). Зависит ли концентрация белка в клетке от концентрации соответствующей мРНК? Посчитайте коэффициент корреляции Пирсона с помощью функции ***cor()***. О чем говорят полученные значения? Можете ли Вы объяснить полученный результат? Добавьте значения коэффициента корреляции Пирсона на график (выражение вида «R = значение»). Используйте функцию ***text()*** для добавления текста и ***locator(1)***, чтобы определить координаты вставки.

Сохраните все построенные диаграммы в виде файлов с расширениями jpeg или tiff.

Вопросы

1. Какие функции высокого уровня есть в R?
2. Какие функции высокого уровня используются для анализа числовых данных?
3. Какие функции высокого уровня используются для анализа категориальных данных?