

Лабораторная работа 1

ОСНОВНЫЕ КОМПОНЕНТЫ СТАТИСТИЧЕСКОЙ СРЕДЫ R и RStudio

Система статистического анализа и визуализации данных R состоит из следующих основных частей:

- языка программирования высокого уровня R, позволяющего одной строкой реализовать различные операции с объектами, векторами, матрицами, списками и т.д.;
- большого набора функций обработки данных, собранных в отдельные пакеты (package);
- развитой системой поддержки, включающей обновление компонентов среды, интерактивную помощь и различные образовательные ресурсы, предназначенные как для начального изучения R, так и последующих консультаций по возникающим затруднениям.

Скачать дистрибутив системы вместе с базовым набором из 29 пакетов (54 мегабайта) можно совершенно бесплатно с основного сайта проекта <http://cran.r-project.org>. Процесс инсталляции системы из скачанного дистрибутива затруднений не вызывает и не требует никаких особых комментариев.

R обладает встроенными обширными справочными материалами, которые можно получить непосредственно в RGui. Если подать с консоли команду **help.start()**, то в вашем интернет-браузере откроется страница, открывающая доступ ко всем справочным ресурсам: основным руководствам, авторским материалам, ответам на вероятные вопросы, спискам изменений, ссылкам на справки по другим объектам R и т.д.*

Также в R можно использовать различные библиотеки, которые включают наборы данных, функции, реализованные методы и алгоритмы, а также другие возможности.

Установка пакетов в RStudio

1. Зайти на вкладку *Packages* (рис.1)
2. Щелкнуть на вкладку *Install packages*
3. В появившемся окне введите пакеты, которые необходимо установить, например, *ggplot2*
4. Далее необходимо нажать на кнопку *Install*

ggplot2 — популярный графический пакет, полноценная и законченная система, наследующая идеи “Графической грамматики” (Grammar of Graphics, отсюда в названии gg).

* Мастицкий С.Э., Шитиков В.К. (2014) Статистический анализ и визуализация данных с помощью R. – Электронная книга, адрес доступа: <http://r-analytics.blogspot.com>

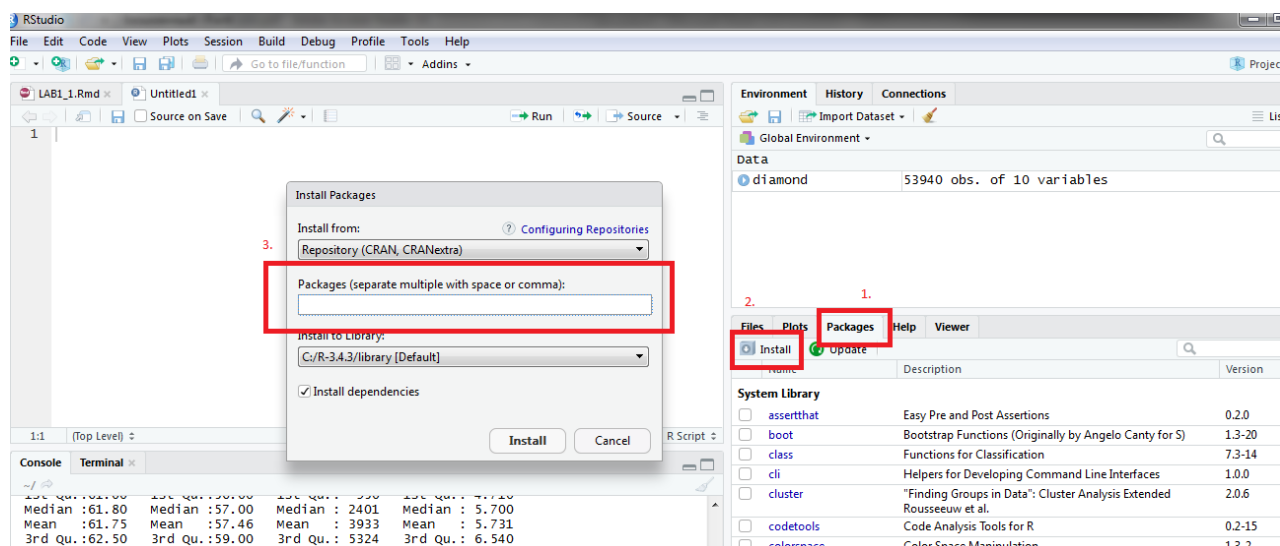


Рисунок 1 – Установка пакетов в RStudio

Оформление отчетов в RStudio

Для оформления отчетов в RStudio необходимо установить следующие пакеты: *knitr* и *rmarkdown*.

Knitr - пакет для генерации динамического отчета с R. Данный пакет на языке статистического программирования R, который позволяет интегрировать R-код в документы LaTeX, LyX, HTML, Markdown, AsciiDoc и reStructuredText.

Markdown позволяет создать интерактивный документ. **Markdown** это облегченный язык разметки, созданный с целью написания максимально читаемого и удобного для правки текста. Markdown является и лёгким для понимания, и легким для чтения даже без каких-либо трансформаций.

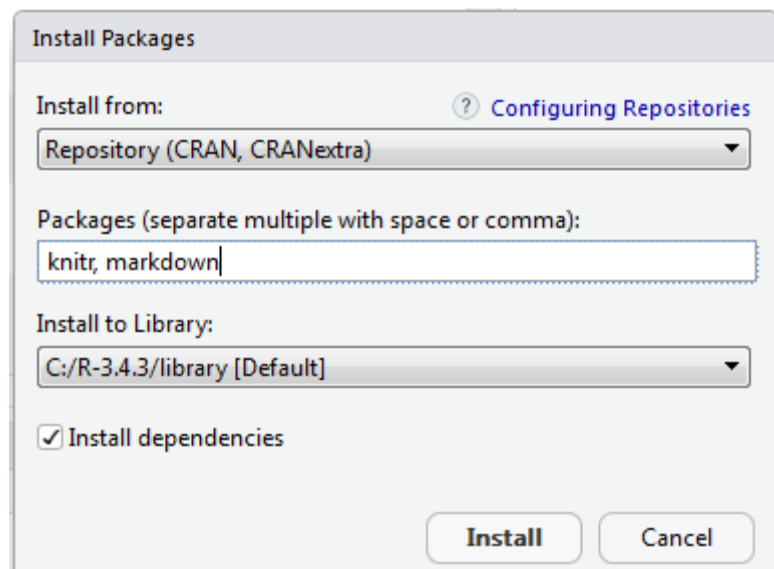


Рисунок 2 – Установка пакетов для оформления отчетов
Для создания отчета необходимо зайти new file – new R Markdown (рис.3). На рисунках 3-5 показано поэтапное создание отчета в формате html.

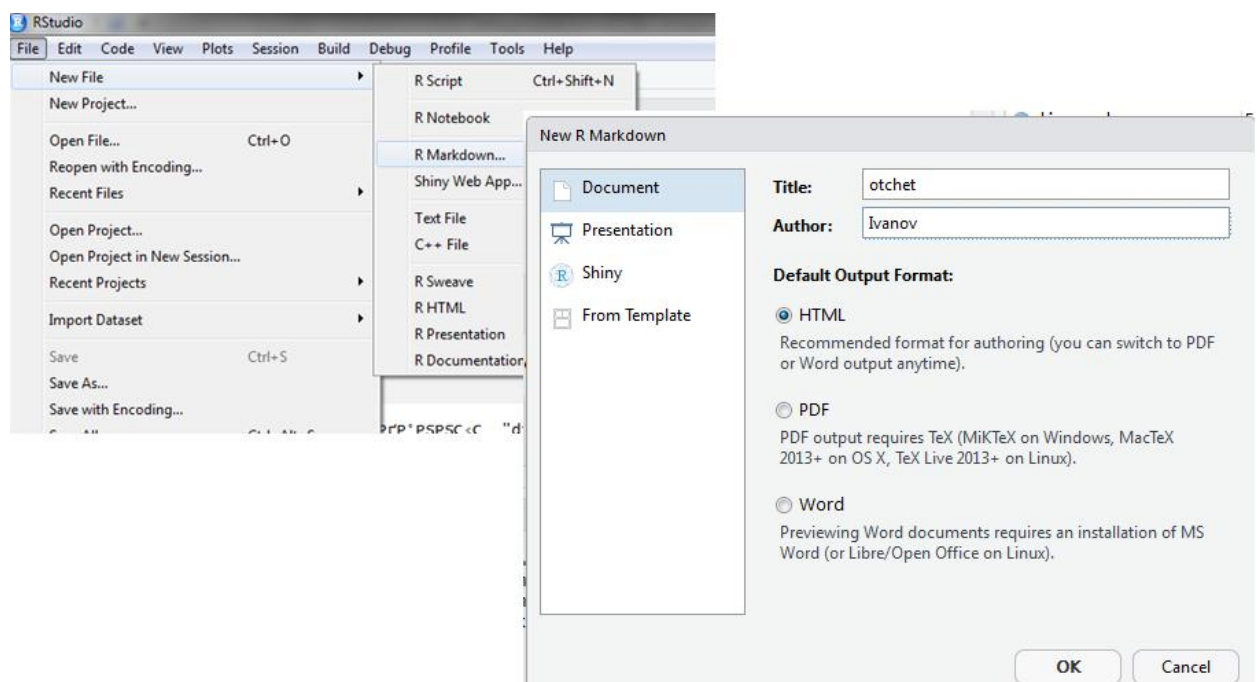


Рисунок 3 – Создание файла формата R Markdown

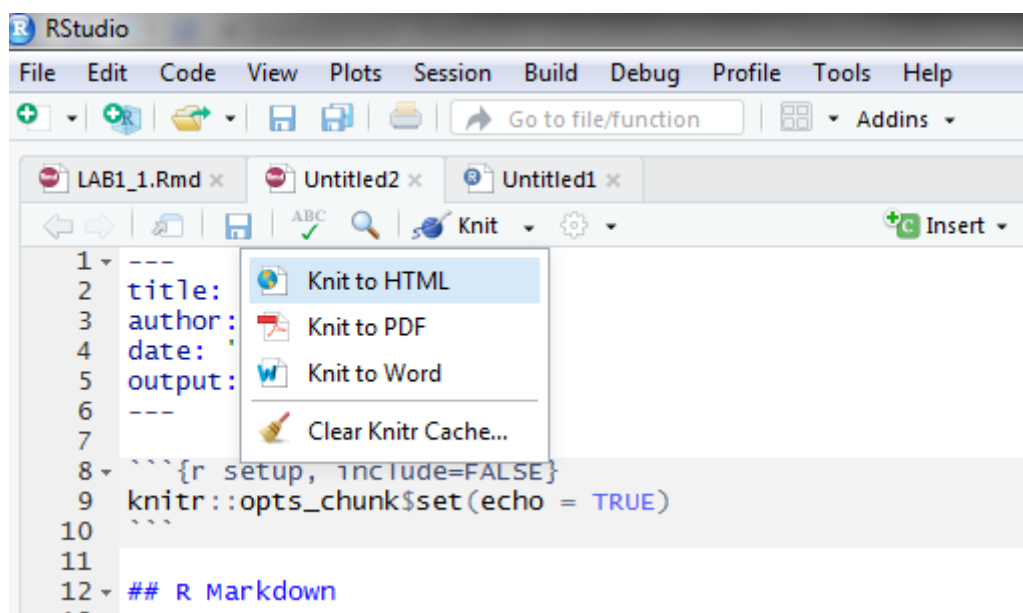


Рисунок 4 – Формирование отчета в RStudio

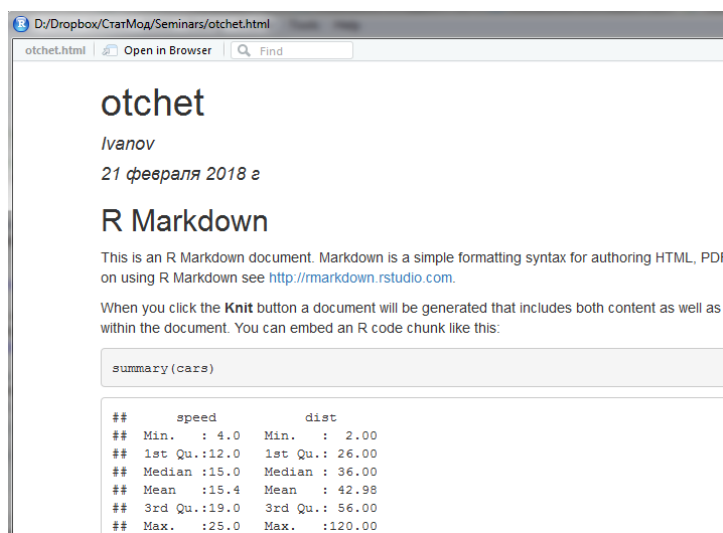


Рисунок 5 – Отчет в формате html

Для вставки кода в файл R Markdown необходимо зайти в Code-Insert Chunk (рис.6)

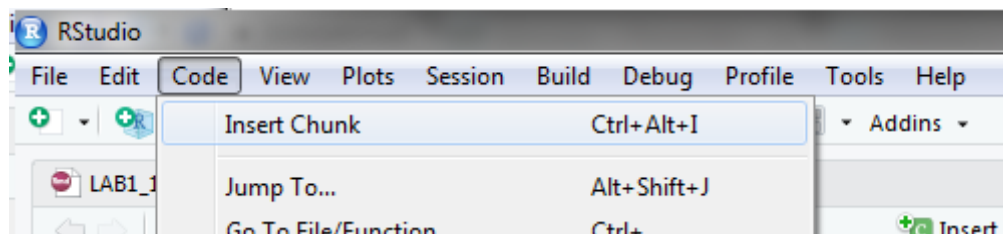


Рисунок 6 – Вставка кода

Ввод исходных данных*

Для создания векторов небольшой длины в R используется функция конкатенации `c()` (от "*concatenate*" – объединять, связывать). В качестве аргументов этой функции через запятую перечисляются объединяемые в вектор значения, например:

```
```{r}
X <- c(0.71, 0.17, 1.06, 3.21, 7.26, 0.24, 3.84, 1.96, 0.17, 7.83, 0.02, 0.99, 1.62,
1.15, 0.08, 1.09, 4.56, 0.14, 0.25, 0.53)
```
```

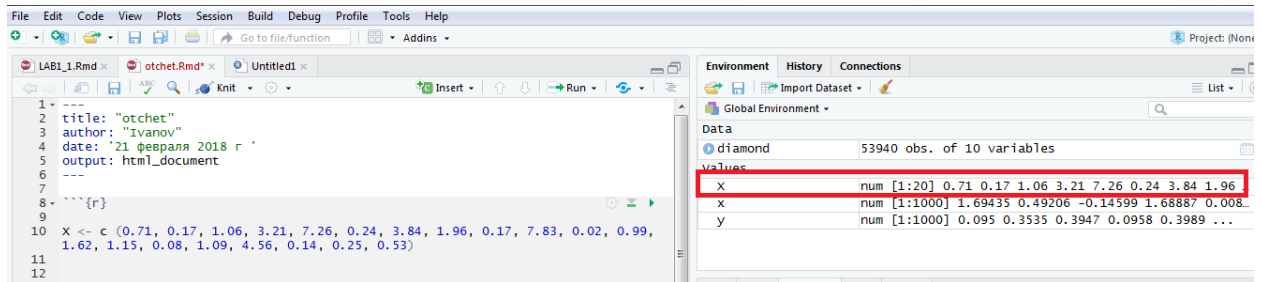


Рисунок 7 – Вектор небольшой длины

Импортирование данных в R*

Подлежащий импортированию файл рекомендуется поместить в рабочую папку программы, т.е. папку, в которой R по умолчанию будет "пытаться найти" этот файл. Чтобы выяснить путь к рабочей папке R на своем компьютере необходимо использовать команду `getwd()` (*get working directory* – узнать рабочую директорию); например:

`getwd()`

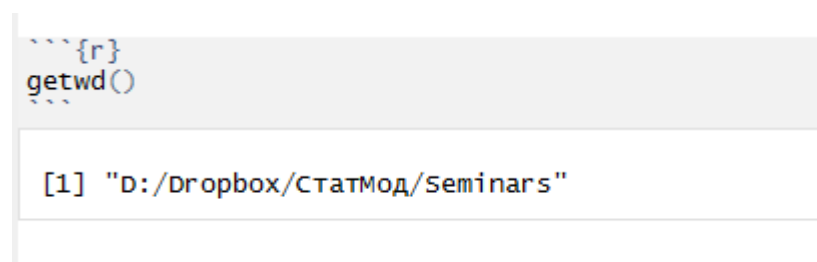
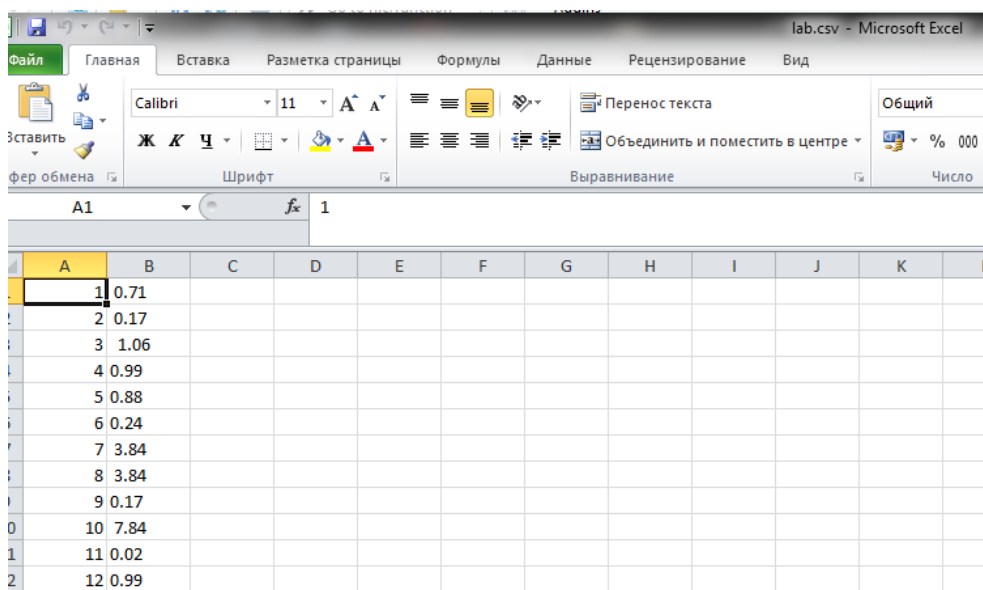


Рисунок 8 - Путь к рабочей директории

Основной функцией для импортирования данных в рабочую среду R является `read.table()`.

* Мастицкий С.Э., Шитиков В.К. (2014) Статистический анализ и визуализация данных с помощью R. – Электронная книга, адрес доступа: <http://r-analytics.blogspot.com>

Аналогом `read.table()` для считывания csv-файлов является функция `read.csv()`. Создадим файл MS Excel, заполним 20 данных, сохраним в формате *csv (рис.9).



| | A | B | C | D | E | F | G | H | I | J | K | L |
|----|----|------|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 0.71 | | | | | | | | | | |
| 2 | 2 | 0.17 | | | | | | | | | | |
| 3 | 3 | 1.06 | | | | | | | | | | |
| 4 | 4 | 0.99 | | | | | | | | | | |
| 5 | 5 | 0.88 | | | | | | | | | | |
| 6 | 6 | 0.24 | | | | | | | | | | |
| 7 | 7 | 3.84 | | | | | | | | | | |
| 8 | 8 | 3.84 | | | | | | | | | | |
| 9 | 9 | 0.17 | | | | | | | | | | |
| 10 | 10 | 7.84 | | | | | | | | | | |
| 11 | 11 | 0.02 | | | | | | | | | | |
| 12 | 12 | 0.99 | | | | | | | | | | |

Рисунок 9 - Данные в формате *csv.

Для импортирования применим следующий код:

```
``{r}
Z <- read.csv(file = "lab.csv", header = TRUE)
``
```

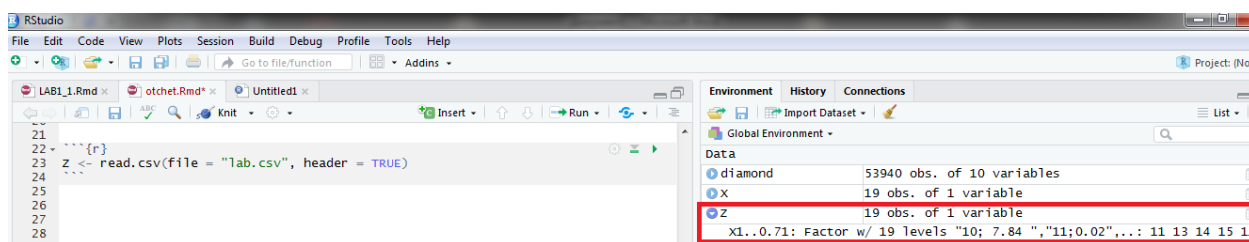


Рисунок 10 – Импортирование данных в формате *csv.

Основные описательные статистики*

Для нахождения различных описательных статистик возьмем набор данных *diamond* из пакета *ggplot2* (рис.11).

* Мастицкий С.Э., Шитиков В.К. (2014) Статистический анализ и визуализация данных с помощью R. – Электронная книга, адрес доступа: <http://r-analytics.blogspot.com>

```

#Подключим пакет "ggplot2"
```{r}
library(ggplot2)

#Загрузим набор "diamond"
```{r}
diamond <- diamonds
```

```

Рисунок 11 – Подключение *ggplot2* и загрузка набора данных *diamond*

В системе R имеется возможность и более быстрого расчета основных параметров описательной статистики. Для этого, в частности, служит функция общего назначения *summary()* (рис.12). Структуру документа можно посмотреть с помощью функции *str ()*(рис.13).

```

#Выведем статистику набора
```{r}
summary(diamond)
```

```

| carat   |         | cut       |        | color |       | clarity  |        |
|---------|---------|-----------|--------|-------|-------|----------|--------|
| Min.    | :0.2000 | Fair      | : 1610 | D:    | 6775  | SI1      | :13065 |
| 1st Qu. | :0.4000 | Good      | : 4906 | E:    | 9797  | VS2      | :12258 |
| Median  | :0.7000 | Very Good | :12082 | F:    | 9542  | SI2      | : 9194 |
| Mean    | :0.7979 | Premium   | :13791 | G:    | 11292 | VS1      | : 8171 |
| 3rd Qu. | :1.0400 | Ideal     | :21551 | H:    | 8304  | VVS2     | : 5066 |
| Max.    | :5.0100 |           |        | I:    | 5422  | VVS1     | : 3655 |
|         |         |           |        | J:    | 2808  | (other): | :2531  |

| depth   |        | table   |        | price   |        | x       |         |
|---------|--------|---------|--------|---------|--------|---------|---------|
| Min.    | :43.00 | Min.    | :43.00 | Min.    | : 326  | Min.    | : 0.000 |
| 1st Qu. | :61.00 | 1st Qu. | :56.00 | 1st Qu. | : 950  | 1st Qu. | : 4.710 |
| Median  | :61.80 | Median  | :57.00 | Median  | : 2401 | Median  | : 5.700 |
| Mean    | :61.75 | Mean    | :57.46 | Mean    | : 3933 | Mean    | : 5.731 |
| 3rd Qu. | :62.50 | 3rd Qu. | :59.00 | 3rd Qu. | : 5324 | 3rd Qu. | : 6.540 |
| Max.    | :79.00 | Max.    | :95.00 | Max.    | :18823 | Max.    | :10.740 |

| y       |         | z       |         |
|---------|---------|---------|---------|
| Min.    | : 0.000 | Min.    | : 0.000 |
| 1st Qu. | : 4.720 | 1st Qu. | : 2.910 |
| Median  | : 5.710 | Median  | : 3.530 |
| Mean    | : 5.735 | Mean    | : 3.539 |
| 3rd Qu. | : 6.540 | 3rd Qu. | : 4.040 |
| Max.    | :58.900 | Max.    | :31.800 |

Рисунок 12 – Суммарная оценка набора *diamond*

```

str(diamond)

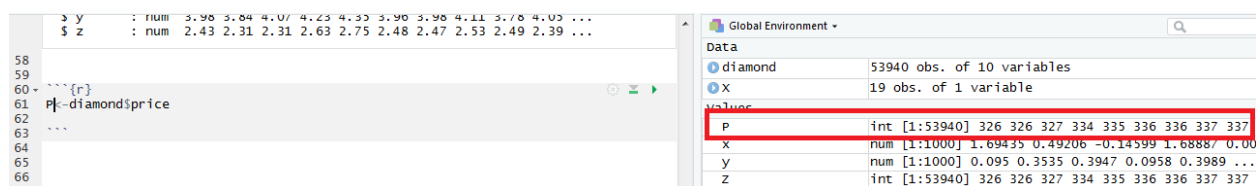
classes 'tbl_df', 'tbl' and 'data.frame': 53940 obs. of 10 variables:
 $ carat : num 0.23 0.21 0.23 0.29 0.31 0.24 0.24 0.26 0.22 0.23 ...
 $ cut : Ord.factor w/ 5 levels "Fair"<"Good"<...: 5 4 2 4 2 3 3 1 3 ...
 $ color : Ord.factor w/ 7 levels "D"<"E"<"F"<"G"<...: 2 2 2 6 7 7 6 5 2 5 ...
 $ clarity: Ord.factor w/ 8 levels "I1"<"SI2"<"SI1"<...: 2 3 5 4 2 6 7 3 4 5 ...
 $ depth : num 61.5 59.8 56.9 62.4 63.3 62.8 62.3 61.9 65.1 59.4 ...
 $ table : num 55 61 65 58 58 57 57 55 61 61 ...
 $ price : int 326 326 327 334 335 336 336 337 337 338 ...
 $ x : num 3.95 3.89 4.05 4.2 4.34 3.94 3.95 4.07 3.87 4 ...
 $ y : num 3.98 3.84 4.07 4.23 4.35 3.96 3.98 4.11 3.78 4.05 ...
 $ z : num 2.43 2.31 2.31 2.63 2.75 2.48 2.47 2.53 2.49 2.39 ...

```

Рисунок 13 – Структура набора *diamond*

Для расчета среднего арифметического, медианы, дисперсии, стандартного отклонения, а также минимального и максимального значений в R служат функции `mean()`, `median()`, `var()`, `sd()`, `min()` и `max()` соответственно.

Рассмотрим расчет для переменной в наборе *diamond*, для этого необходимо ввести новую переменную *P*, присвоим ей значения переменной *price* в наборе *diamond*, выполнив команду `P <- diamond$price` (рис.14)



The screenshot shows the R console with the command `P <- diamond$price` being executed. The Global Environment pane on the right shows the updated data structure: *diamond* (53940 obs. of 10 variables), *X* (19 obs. of 1 variable), and *P* (int [1:53940] 326 326 327 334 335 336 336 337 337 338 ...).

Рисунок 14 – Присвоение переменной *P* значений переменной *price*

Найдем значение *mean* для переменной *price*:



The screenshot shows the R console with the command `mean(P)` being executed. The output is `[1] 3932.8`.

Рисунок 15 – Значение *mean* () для переменной *P*

Квантили рассчитываются в R при помощи функции `quantile()` (рис.16):



| #квантили |        |         |         |          |
|-----------|--------|---------|---------|----------|
| 0%        | 25%    | 50%     | 75%     | 100%     |
| 326.00    | 950.00 | 2401.00 | 5324.25 | 18823.00 |

Рисунок 16 – Квантили в R

При настройках, заданных по умолчанию, выполнение указанной команды приведет к расчету минимального (326.00) и максимального (18823.00) значений, а также трех *квартелей*, т.е. значений, которые делят совокупность на четыре равные части – 950.00, 2401.00 и 5324.25.

Для расчета коэффициентов *эксцесса* (англ. *kurtosis*) и *асимметрии* (*skewness*) необходимо установить дополнительный пакет **moments**. Если этот пакет не установлен на Вашем компьютере, выполните следующую команду (при этом компьютер должен быть при этом подключен к сети Интернет):

Рассчитаем коэффициенты эксцесса и асимметрии (рис.17):

**library(moments)** *#загрузка пакета moments*

**kurtosis(X, na.rm = TRUE)**

**skewness(X, na.rm = TRUE)**

|                                                                                  |
|----------------------------------------------------------------------------------|
| <code>library(moments)</code>                                                    |
| <code>kurtosis(P, na.rm = TRUE)</code><br><code>skewness(P, na.rm = TRUE)</code> |
| [1] 5.177383<br>[1] 1.61835                                                      |

Рисунок 17 - Коэффициенты эксцесса и асимметрии

### Законы распределения вероятностей, реализованные в R\*

В базовой версии R имеются функции для работы с целым рядом распространенных законов распределения вероятностей. В зависимости от

\* Мастицкий С.Э., Шитиков В.К. (2014) Статистический анализ и визуализация данных с помощью R. – Электронная книга, адрес доступа: <http://r-analytics.blogspot.com>

назначения, имена этих функций начинаются с одной из следующих четырех букв:

- ° d (от "*density*", плотность): функции плотности вероятности ("функция распределения масс" для дискретных величин);
- ° p (от "*probability*", *вероятность*): кумулятивные функции распределения вероятностей;
- ° q (от "*quantile*", квантиль): функции для нахождения квантилей того или иного распределения;
- ° r (от "*random*", случайный): функции для генерации случайных чисел в соответствии с параметрами того или иного закона распределения вероятностей.

В частности, в базовой версии R реализованы следующие законы распределения вероятностей:

- Бета-распределение (см. dbeta)
- Биномиальное распределение (включая распределение Бернулли) (dbinom)
- Распределение Коши (dcauchy)
- Распределение хи-квадрат (dchisq)
- Экспоненциальное распределение (dexp)
- Распределение Фишера (df)
- Гамма-распределение (dgamma)
- Геометрическое распределение (как частный случай отрицательного биномиального распределения) (dgeom)
- Гипергеометрическое распределение (dhyper)
- Логнормальное распределение (dlnorm)
- Полиномиальное (или мультиномиальное) распределение (dmultinom)
- Отрицательное биномиальное распределение (dnbinom)
- Нормальное распределение (dnorm)
- Распределение Пуассона (dpois)
- Распределение Стьюдента (dt)
- Равномерное распределение (dunif)
- Распределение Вейбулла (dweibull)

Рассмотрим графики функции и плотности нормального закона распределения. Рассмотрим пример графика функции нормального закона распределения:

## 1 способ

```
x <- seq(-2, 2, by=0.1)
```

\* Мастицкий С.Э., Шитиков В.К. (2014) Статистический анализ и визуализация данных с помощью R. – Электронная книга, адрес доступа: <http://r-analytics.blogspot.com>

```
y <- pnorm(x, mean=-1, sd=1)
```

```
plot(x, y, type='l')
```

ИЛИ

```
x <- seq(-2, 2, 0.01)
```

```
y <- pnorm(x, -1, 1)
```

```
plot(x, y, type='l')
```

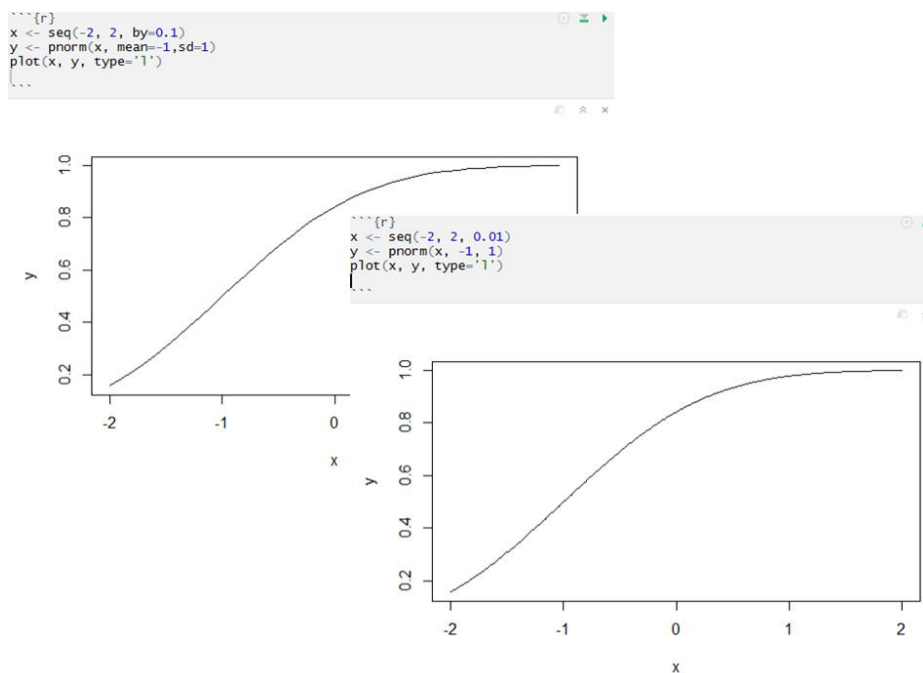


Рисунок 18 - Пример графика функции нормального закона распределения  
(1 способ)

## 2 способ

```
```{r}
```

```
x <- rnorm(1000, mean = 0, sd = 1)
```

```
y <- pnorm(x)
```

```
plot(x, y, type='p')
```

```
```
```

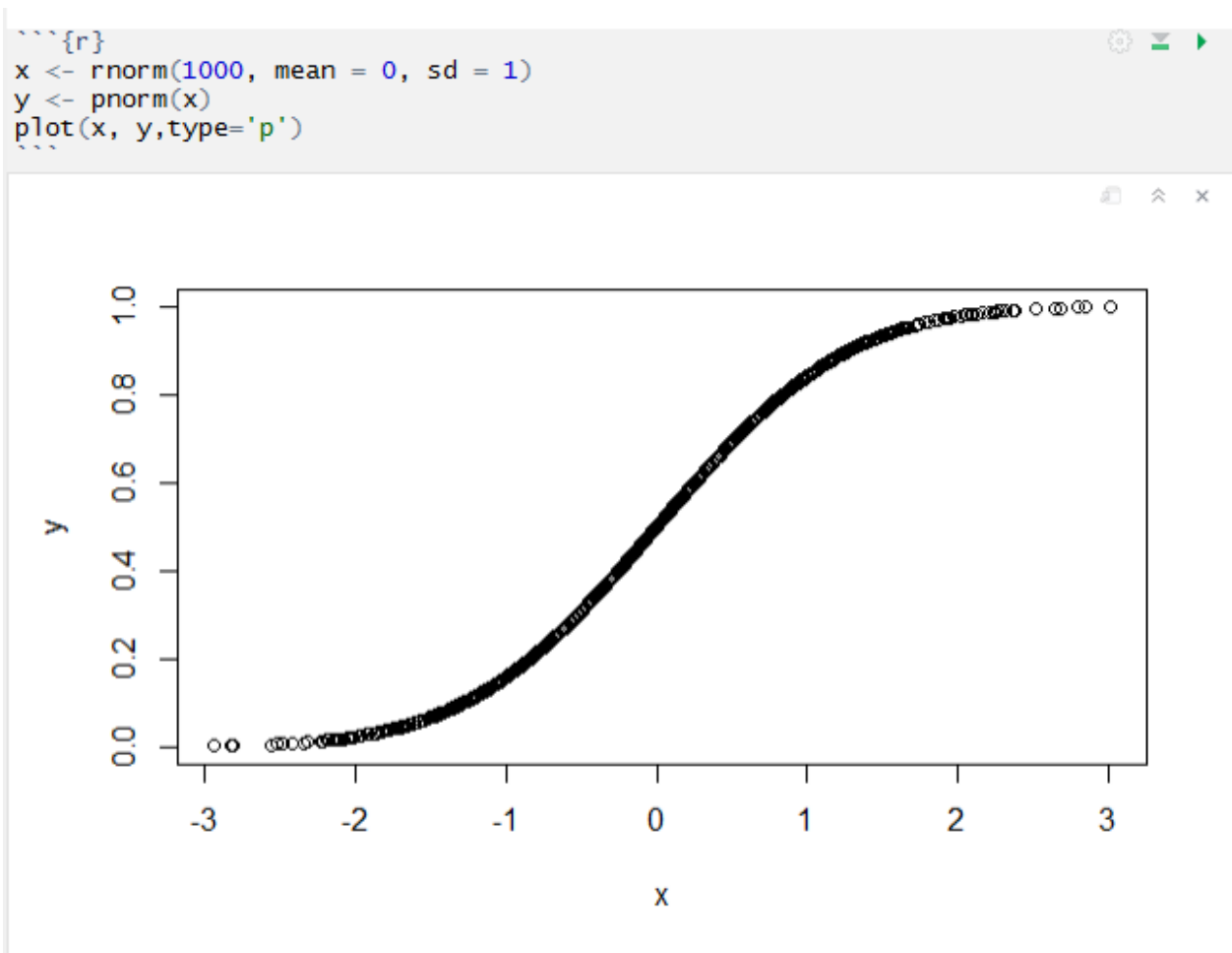


Рисунок 19 - Пример графика функции нормального закона распределения  
(2 способ)

Рассмотрим пример графика плотности нормального закона распределения:

### 1 способ (рис.20)

```
x <- seq(-2, 2, by=0.1)
```

```
y <- dnorm(x, mean=-1, sd=1)
```

```
plot(x, y, type='l')
```

или

```
x <- seq(-2, 2, 0.01)
```

\* Мастицкий С.Э., Шитиков В.К. (2014) Статистический анализ и визуализация данных с помощью R. – Электронная книга, адрес доступа: <http://r-analytics.blogspot.com>

```
y <- dnorm(x, -1, 1)
```

```
plot(x, y, type='l')
```

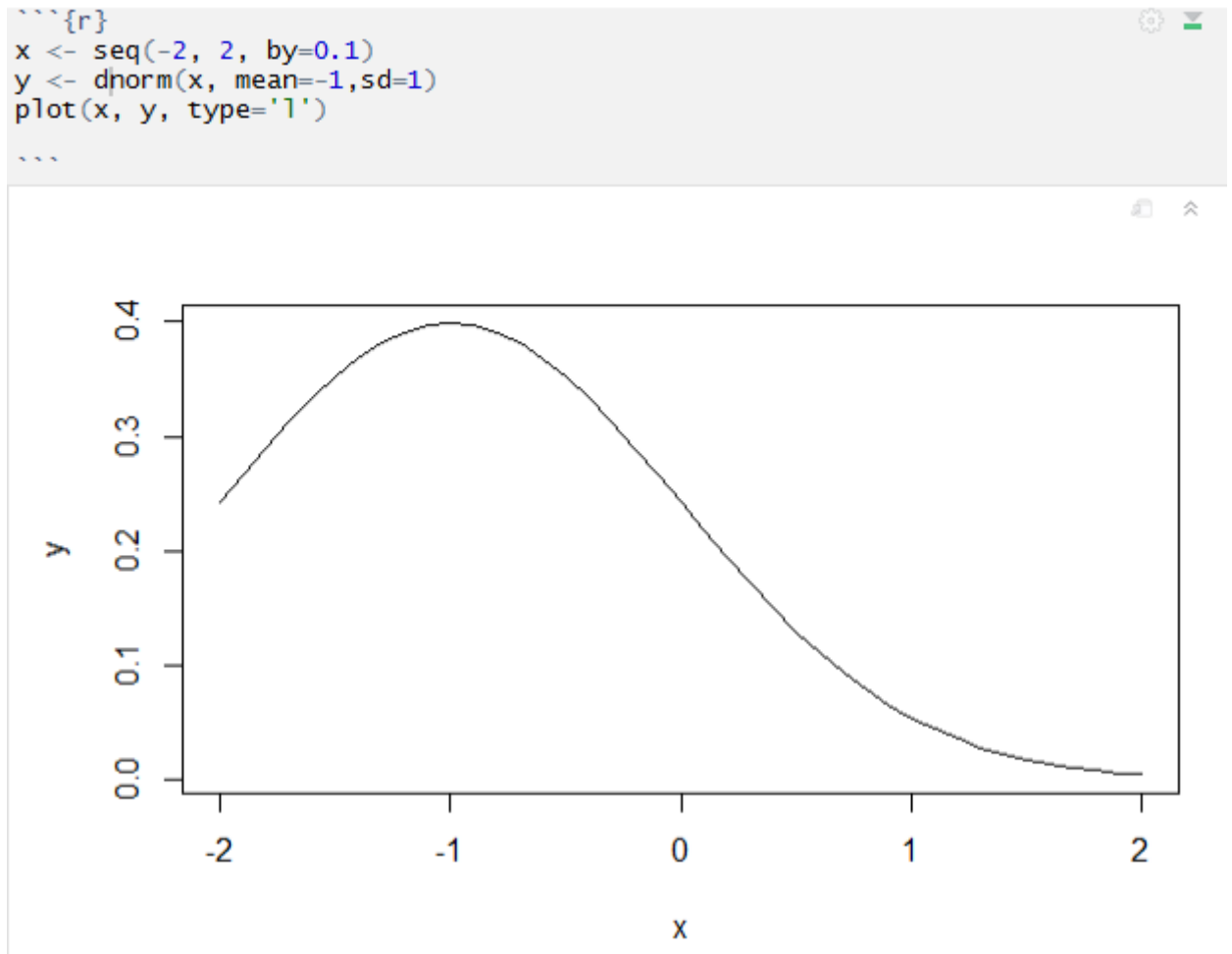


Рисунок 20 - Пример графика плотности нормального закона распределения (1 способ)

## 2 способ (рис.21)

```
```{r}
```

```
x <- rnorm(1000, mean = 0, sd = 1)
```

```
y <- dnorm(x)
```

```
plot(x, y, type='p')
```

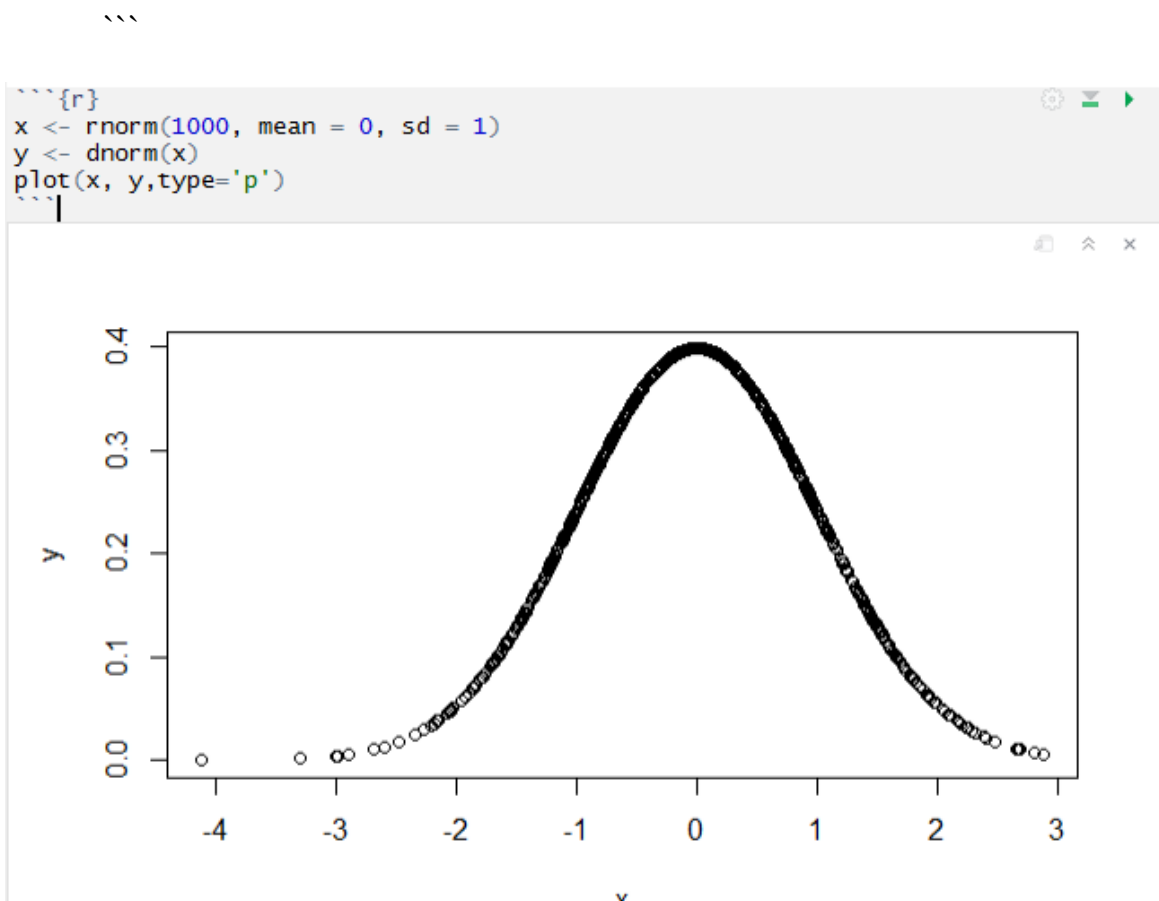


Рисунок 21 - Пример графика плотности нормального закона распределения
(2 способ)

Недостаток 1 способа состоит, в том, что необходимо выбирать границы последовательности, и может быть рассмотрен случай, где график функции или плотности не будут полностью видны на области, именно такой случай мы наблюдаем с плотностью и функцией распределения случайной величины.

Для определения критических значений статистик по заданному p -уровню (например, $p=0.99$) и параметрам распределения ($mean$, sd) для нормального закона распределения:

`qnorm(p=0.95, mean=50, sd=10)`

или

`qnorm(0.95, 50, 10)`

```
{r}
qnorm(p=0.95, mean=50, sd=10)

[1] 66.44854
```

Рисунок 22 - Определение критических значений статистик нормального закона распределения при $p=0.95$, $mean=50$, $sd=10$

Для определения уровня доверия p по заданному значению критической статистики и соответствующим параметрам распределения для нормального закона распределения необходимо применить следующую команду:

`pnorm(66, mean=50, sd=10)`

или

`pnorm(66, 50, 10)`

```
{r}
pnorm(66, mean=50, sd=10) |

[1] 0.9452007
```

Рисунок 23 – Определение уровня доверия p при известных $q=66$, $mean=50$, $sd=10$

Задание

1. Изучить ввод данных в R.

Импортировать данные в формате *xlsx в R. Результаты представить в виде скрипта и скриншота.

2. Изучить оформление отчета в R.

2.1 Экспортировать документ в формате Word.

2.2 Экспортировать документ в формате PDF.

2.3 Разработать «стиль» оформления в MS Word и загрузить в R.

2.4 Использовать загруженный стиль при оформлении отчета к лабораторной работе.

3. Изучить описательные статистики в R **.

* Мастицкий С.Э., Шитиков В.К. (2014) Статистический анализ и визуализация данных с помощью R. – Электронная книга, адрес доступа: <http://r-analytics.blogspot.com>

3.1 Выбрать самостоятельно исходные данные в виде переменной, состоящей из 20 и более наблюдений (данные должны быть реального сектора, например экономические показатели - ВВП, объем экспорта, уровень безработицы). В отчете привести ссылку на исходные данные.

3.2 Вычислить основные описательные статистики (среднее, стандартное отклонение, дисперсия, максимальное и минимальное значение, медиана, коэффициенты асимметрии и эксцесса, нижнюю и верхнюю квартили).

**** - Примечание: временные ряды не рассматривать.**

4. Изучить законы распределения в R

4.1. Построить графики функций распределения и плотностей распределения следующих распределений:

- а) Фишера (при степенях свободы $df_1=10$ и $df_2=10$; $df_1=2$ и $df_2=50$; $df_1=10$ и $df_2=50$; $df_1=10$ и $df_2=500$, $df_1=30$ и $df_2=1000$);
- б) Стьюдента (при степенях свободы $df=10$; $df=50$; $df=200$);
- в) показательного (при параметре $\lambda=0,5$; $\lambda=5$; $\lambda=20$);
- г) χ^2 -распределения (при степенях свободы $df=10$; $df=50$; $df=200$);
- д) логнормального (при $\mu=0$, $\sigma=1$; $\mu=1$, $\sigma=2$);
- е) нормального (при $\mu=0$, $\sigma=1$; $\mu=1$, $\sigma=2$).

4.2. Проанализировать изменение графиков функций и плотности рассмотренных распределений при изменении параметров распределений (степеней свободы).

4.3. Определить критическое значение статистик по заданному p -уровню ($p=0,99$, $p=0,95$, $p=0,9$) и параметрам распределения (например, степеням свободы) для следующих распределений:

- а) Фишера (при степенях свободы $df_1=10$ и $df_2=10$; $df_1=2$ и $df_2=50$; $df_1=10$ и $df_2=50$; $df_1=10$ и $df_2=500$, $df_1=30$ и $df_2=1000$);
- б) Стьюдента (при степенях свободы $df=10$; $df=50$; $df=200$);
- в) показательного (при параметре $\lambda=0,5$; $\lambda=5$; $\lambda=20$);
- г) χ^2 -распределения (при степенях свободы $df=10$; $df=50$; $df=200$);
- д) логнормального (при $\mu=0$, $\sigma=1$; $\mu=1$, $\sigma=2$);
- е) нормального (при $\mu=0$, $\sigma=1$; $\mu=1$, $\sigma=2$).

4.4. Определить уровень доверия p по заданному значению критической статистики и соответствующим параметрам распределения:

- a) Фишера (при степенях свободы $df_1=10$ и $df_2=10$ и значении $F=1,55$; $df_1=2$ и $df_2=50$ и значении $F=2,33$; $df_1=10$ и $df_2=50$ и значении $F=4,8$; $df_1=10$ и $df_2=500$ и значении $F=1,72$, $df_1=30$ и $df_2=1000$ и значении $F=1,35$);
 - b) Стьюдента при односторонней проверке (при степенях свободы $df=10$ и значении $t=1,37$; $df=50$ и значении $t=2,11$; $df=200$ и значения $t=0,55$);
 - c) Показательного (при параметре $\lambda=0,5$ и значении $\exp=1,38$; $\lambda=5$ и значении $\exp=0,6$; $\lambda=20$ и значении $\exp=0,23$);
 - d) χ^2 -распределения (при степенях свободы $df=10$ и значении $\chi^2=12,54$; $df=50$ и значении $\chi^2=67,54$; $df=200$ и значении $\chi^2=220$);
 - e) логнормального (при $\mu=0$, $\sigma=1$ и значении $L=1,96$; $\mu=1$, $\sigma=2$ и значении $L=72,96$);
 - f) нормального (при $\mu=0$, $\sigma=1$ и значении $Z=1,96$; $\mu=1$, $\sigma=2$ и значении $Z=2,67$).
5. Оформить результаты лабораторной работы в виде отчета. Отчет должен содержать комментарии к выполняемым командам, результаты необходимо представить наглядном виде.