Практическая работа №8

Некоторые сведения о возможностях статистического анализа

Затабулированные распределения

Пакет R позволяет проводить различные виды статистической обработки данных, информация о некоторых из них приведена в таблице.

Таблица 1

Затабулированные распределения в R

Название	Обозначение в R	Параметры
Нормальное	norm	mean, sd
t-распределение	t	df
Равномерное	unif	min, max
Хи-квадрат	chisq	df
F-распределение	f	dfl, df2
Гамма	gamma	shape, scale

Стандартное нормальное распределение

Многие естественные процессы соответствуют закону нормального распределения. В R есть группа взаимосвязанных функций для работы с такими данными.

Функция rnorm()

Служит для генерации совокупности нормально распределенных случайных чисел:

rnorm(n, mean = 0, sd = 1)

Параметр n представляет собой объем создаваемой случайной выборки.

Параметр mean – математическое ожидание.

Параметр sd – среднее квадратическое отклонение.

Функция qnorm()

Предназначена для вычисления квантилей нормального распределения:

qnorm (p, mean = 0, sd = 1, lower.tail = TRUE, log.p = FALSE)

Параметр p представляет значение вероятности.

Параметры mean и sd аналогичны параметрам функции rnorm().

Параметр lower.tail – логическая величина: для значения TRUE (по умолчанию) вероятность рассчитывается как P[X < x], иначе – как P[X > x].

Параметр log.p – логическая величина; если его значение TRUE, вероятности p задаются как log (p).

Функция pnorm()

Является функцией распределения и описывает вероятность того, что случайная переменная X примет какое-либо значение, не превышающее либо равное x:

Pnorm(q, mean = 0, sd = 1, lower.tail = TRUE, log.p = FALSE)

Параметр q – квантиль случайной величины.

Параметры mean, sd, lower.tail, log,p соответствуют аналогичным параметрам функций rnorm() и qnorm().

Функция dnorm()

Рассчитывает значения функции плотности вероятности для заданных значений вектора x:

dnorm(x, mean = 0, sd = 1, log = FALSE)

Параметр x – вектор значений случайной величины. Параметр log – логическая величина; если его значение TRUE, вероятности p задаются как log (p).

Иллюстрацию возможностей указанных функций приведем в примере.

Пример

С помощью функций rnorm(), qnorm(), pnorm() и dnorm() сформировать массив случайных чисел.

```
2  n<-rnorm(100,mean = 15,sd=5)
3  n
4  hist(n)
```

На рисунке 54 показано графическое представление выборки.

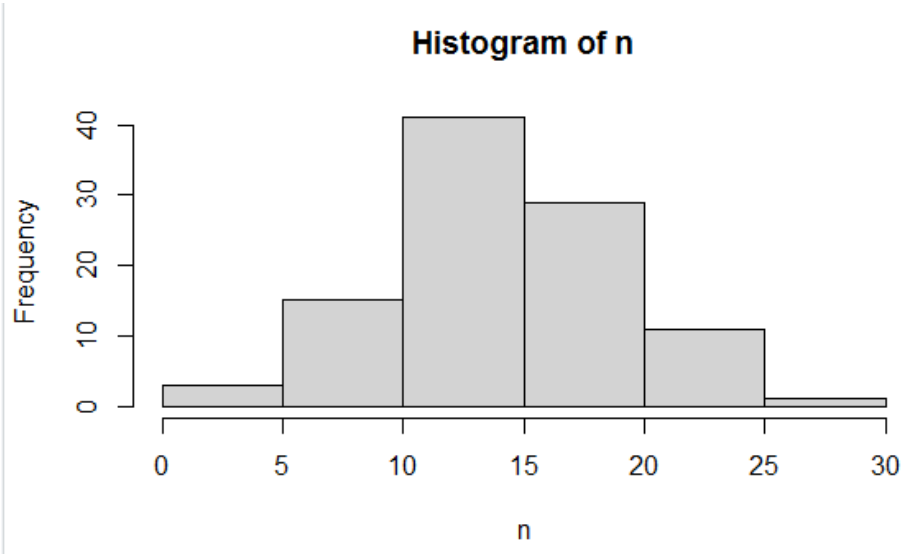

Рисунок 1 Графическое представление выборки

Функция qnorm() принимает значение вероятности и дает число, совокупное значение которого соответствует значению вероятности.

Например, предположим, что мы хотим найти 85-й процентиль нормального распределения, среднее значение которого равно 70 и стандартное отклонение которого равно 3.

```
6  x <- seq(0, 1, by = 0.02)
7  y <- qnorm(x, mean = 2, sd = 1)
8  plot(x,y)
```

График функции распределения представлен на рисунке 55.

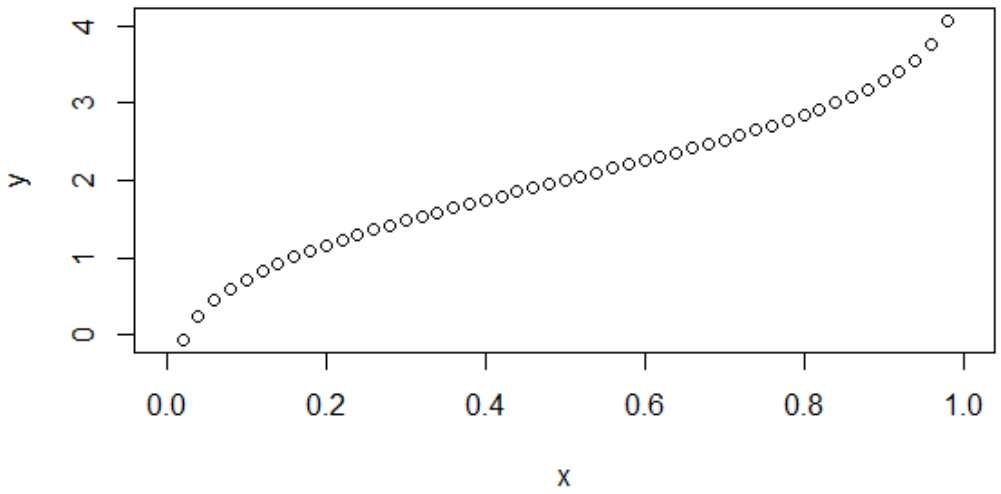

Рисунок 2 График функции распределения

Функция pnorm() дает вероятность того, что обычно распределенное случайное число меньше значения данного числа. Она также называется «кумулятивная функция распределения».

```
x <- seq(-10,10,by = .2)
x
y <- pnorm(x, mean = 2.5, sd = 2)
y
plot(x,y)
```

График функции распределения представлен на рисунке 56.

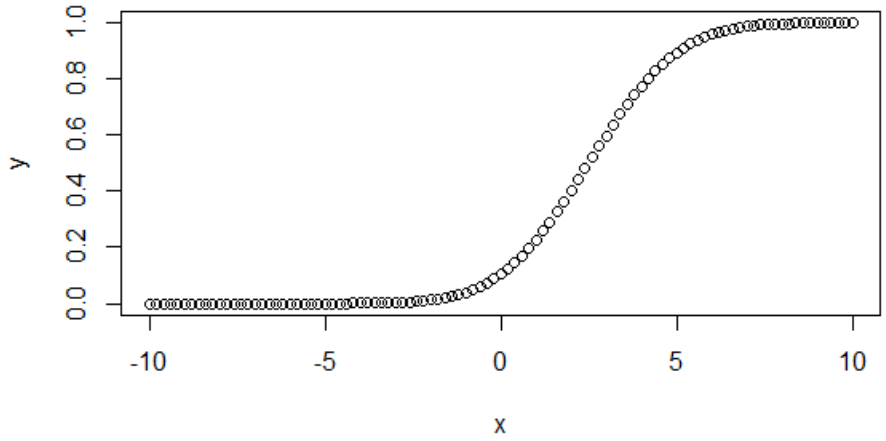

Рисунок 3 Графическая иллюстрация функции распределения нормально распределенных чисел

Функция dnorm () дает высоту распределения вероятностей в каждой точке для данного среднего значения и стандартного отклонения.

```
x <- seq(-10, 10, by = .1)
y <- dnorm(x, mean = 2.5, sd = 0.5)
plot(x,y)
```

Графическая иллюстрация функции плотности нормально распределенных чисел представлена на рисунке 57.

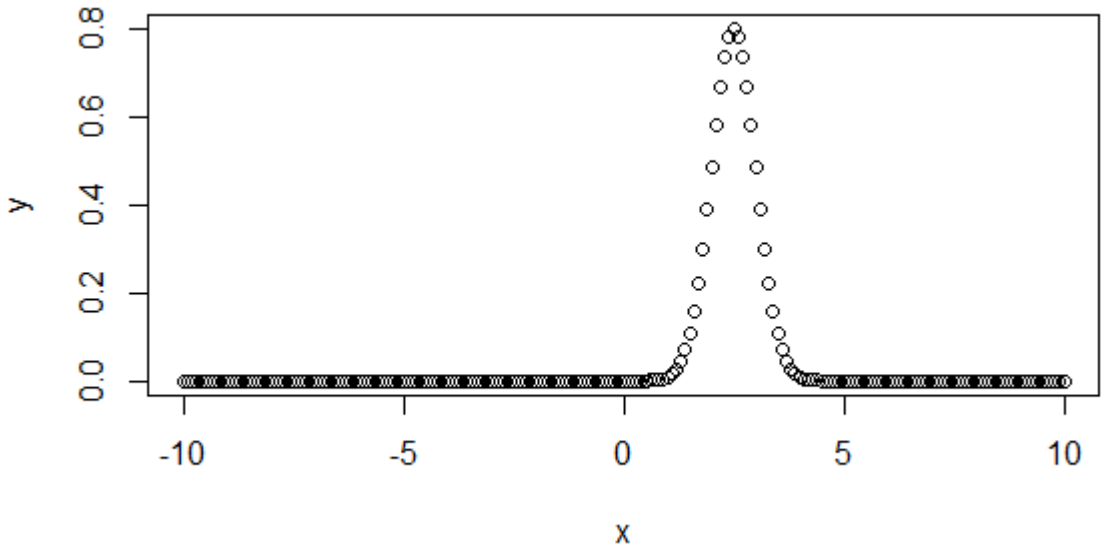

Рисунок 4 Графическая иллюстрация функции плотности нормально распределенных чисел

Гистограмма и эмпирическая плотность

При работе с пакетом R исследователю предоставляется возможность доступа к многочисленным массивам данных. Задав необходимые дополнительные значения, можно исследовать, например, понятие эмпирической плотности, нанесенной на гистограмму.

Не вдаваясь в теоретические подробности, приведем пример скрипта с подробными комментариями, решающего указанную задачу, и результаты его выполнения. На рисунке 58 показан текст скрипта построения гистограммы с нанесенной эмпирической плотностью.

```
# Эмпирическая плотность
y<-faithful$eruptions # исходные данные
y
hist(y,breaks=seq(1.6,5.2,by=0.2),prob=T,
     main="Гистограмма",
     ylab="Плотность",
     xlab="Диапазон данных")
# Расчет значений эмпирической плотности
y.pdf<-density(y,bw="ucv")
# Вывод кривой эмпирической плотности
lines(y.pdf,col="red")
text(3.1,0.5,"Эмпирическая плотность")
# Добавление исходных данны на ось OX:
rug(y)
```

Рисунок 5 Текст скрипта построения гистограммы с нанесенной эмпирической плотностью

На рисунке 59 показана гистограмма с нанесенной эмпирической плотностью.

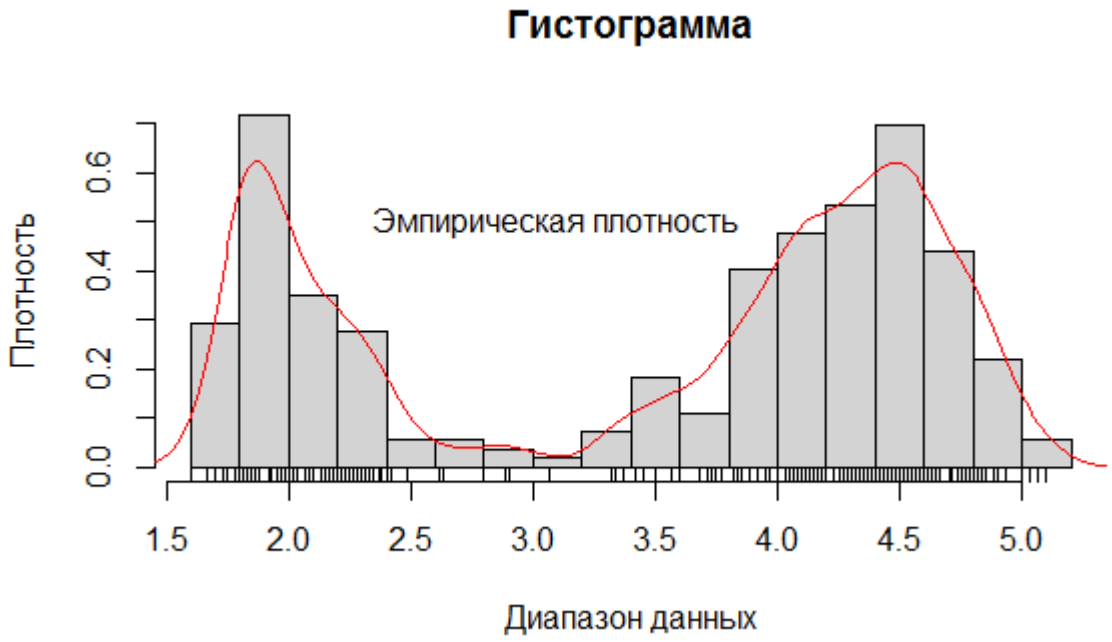

Рисунок 6 Гистограмма с нанесенной эмпирической плотностью

Эмпирическая функция распределения

Эмпирической функцией распределения (функцией распределения выборки) называется функция F*(x), определяющую для каждого значения x частоту события X<x.

Приведем текст скрипта, демонстрирующего возможность исследования эмпирических функций распределения, а также графический результат его выполнения.

Текст скрипта:

```
# Эмпирическая плотность
y<-faithful$eruptions # исходные данные
y.cdf<-ecdf(y)
# y.cdf - функция, подставляя в которую квантили,
# дает значения эмпирической функции распределения
y.cdf(3) # значение функции распределения для квантили 3

plot(y.cdf,do.points=F,verticals=T,
     main="График эмпирической функции распределения",
     ylab="Значения функции распределения",
     xlab="Диапазон данных")
points(3,y.cdf(3),col="red",pch=10)
# вывод на график значения эмпирической функции
# распределения для квантиля 3
text(3,0.5,y.cdf(3))
```

Результат выполнения скрипта представлен на рисунке 60.

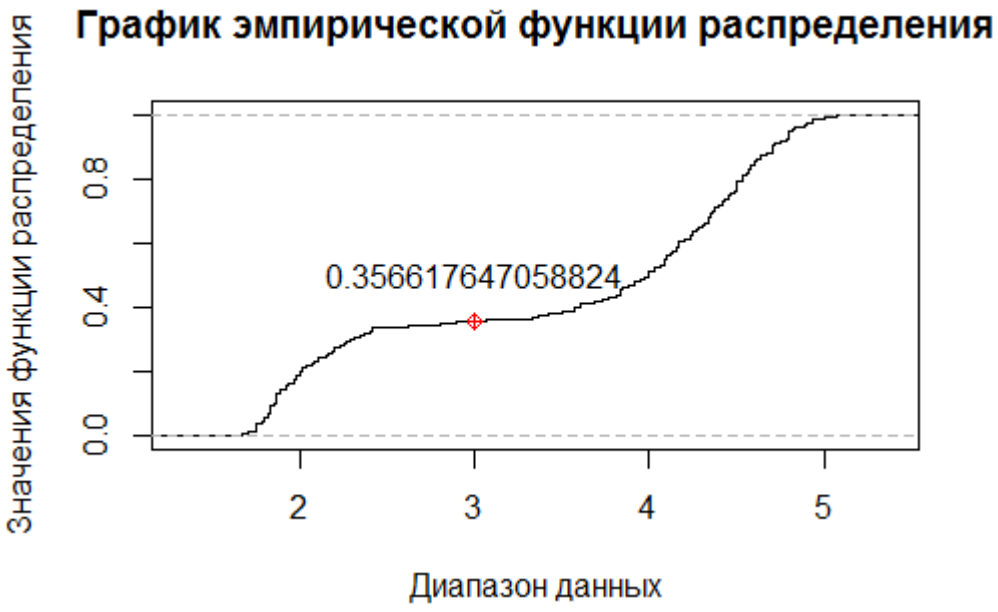

Рисунок 7 Результат выполнения скрипта исследования эмпирической функции распределения

Самостоятельная работа №8

Исследование статических и вероятностных характеристик временных рядов

В работе рассмотрены данные временного ряда, полученные в результате эксперимента снятия физиологических показателей, связанных с электрической активностью организма в покое и при различных внешних психических воздействиях. Вес подопытного в момент эксперимента – 90 кг, рост подопытного в момент эксперимента – 195см.

Данные можно получить по следующей ссылке: https://raw.githubusercontent.com/qwerty29544/RpracticeBook/master/2Data/01FlatTables/ECG_yurchenkov.txt.

Работа является больше исследовательской с точки зрения отработки навыков предобработки данных в R, анализа статистических закономерностей в данных, графического анализа.

Функциональное требование: вся обработка текстового файла должна производиться с помощью средств R. Необходимо, чтобы все переменные анализа:

- частота дискретизации,
- единицы измерения,
- названия характеристик,
- сами данные,

были внесены в программный скрипт только из файла с данными и только при помощи R.

Задание 1.

1. Скачать данные по указанной ссылке при помощи функции download.file().
2. Произвести импорт данных в среду R различными методами функционалом языка R.
3. Все метаданные таблицы измерений обработать при помощи R различными способами, результатом должны являться переменные, содержащие метаданные каналов снятия измерений. Если для обработки данных понадобятся метаданные по росту и весу, разрешается внести эти данные как переменные среды внутри скрипта.
4. Функционально установить количество стадий эксперимента, т.е. сколько опытов в процессе эксперимента проводилось над подопытным.
5. Установить точное время в мс точек снятия показателей в связи с полученной частотой дискретизации каналов снятия измерений и номерами строк. Считать время с 0минут:00секунд:00миллисекуд.
6. Построить гистограммы распределения каждого из временных рядов в целом и по выявленным стадиям эксперимента. Предоставить сводку по

описательным статистикам данных как полностью, так и по стадиям эксперимента.

7. Предложить свои варианты методов обработки данных исходя из ранее полученных навыков и разработанных функций в рамках выполнения самостоятельной работы №6.

Важно. Для выполнения такого рода задания вам, возможно, понадобится знание организации численного вектора гистограммы, т.е. независимого получения количества элементов в столбцах данных:

table(cut(x = ts, breaks = seq(min_ts, max_ts, (max_ts – min_ts)/n))),

где ts – временной ряд или просто числовой вектор данных, min_ts и max_ts — естественные границы данных по оси ординат, т.е. естественный максимум и минимум данных, n – количество столбцов данных.

Функционально разбиения можно организовать и другими способами, важным является критерий – это должен быть вектор чисел.

Контрольные вопросы:

1. Приведите примеры функций обработки стандартных нормальных распределений.

Каким образом могут быть построены гистограммы