МИНОБРНАУКИ РОССИИ
САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ
ЭЛЕКТРОТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ
«ЛЭТИ» ИМ. В.И. УЛЬЯНОВА (ЛЕНИНА)
Кафедра Информационных систем

ОТЧЕТ

по практической работе №1

по дисциплине «Статический анализ»

Тема: Разведочный анализ данных

Вариант: Дрезден

Отчёт подготовил: Кошеляев А.С

Отчёт сдан: 15.03.2024

Студент гр.1323 _____ **Кошеляев А.С 90%**

Студент гр.1323 _____ **Русских В.Д 10%**

Преподаватель _____ **к.т.н. Бурков Е.А**

Санкт-Петербург

2024

Цель работы: анализ предложенного набора данных с помощью базовых статистических характеристик и методов.

Задание:

1. Взять с сайта ***www.pogodaiklimat.ru/history.php*** данные о среднемесячной температуре воздуха в городе, Дрезден.
2. Привести в отчете таблицу входных данных (только ее начальный и конечный фрагменты).
3. Для каждого месяца рассчитать и представить в отчете в виде удобочитаемой таблицы:
   - объем выборки
   - минимальное значение,
   - максимальное значение,
   - первый квартиль,
   - медиану
   - третий квартиль,
   - межквартильный размах,
   - среднее,
   - стандартное отклонение,
   - стандартную ошибку среднего,
   - коэффициент вариации,
   - коэффициент асимметрии.

4.1. Построить и привести в отчете *на одном графике* блочные диаграммы (с визуализацией выбросов) для всех двенадцати месяцев.

4.2. Идентифицировать все выбросы и составить таблицу выбросов, в которой следует указать значение, год и месяц каждого выброса.

5. Для каждого месяца привести гистограмму (график распределения) и сделать обоснованный вывод о наличии и степени выраженности свойства симметричности данных (положительная или отрицательная,

сильная или слабая; при необходимости можно дополнительно воспользоваться коэффициентом асимметрии).

6.1. Для каждого месяца рассмотреть гипотезу о том, подчиняется ли распределение данных нормальному закону на основе:

1) визуального анализа гистограммы (наличие симметрии и колоколообразности);

2) визуального анализа графика квантилей;

3) анализа численных характеристик набора данных (совпадение моды, медианы и среднего; соотношение между межквартильным размахом и станд.отклонением; попадание всех значений в диапазон «шести сигм»);

4) выбранного статистического критерия (например, Шапиро-Уилка) при $\alpha = 0{,}05$.

6.2. Сформировать итоговый вывод о нормальности данных (для каждого месяца), используя совокупные результаты пп. 6.1.1 – 6.1.4.

Выполнение работы:

Для выполнения задания используется язык обработки данных R (R Studio) и (IDE) RStudio Desktop свободно распространяемый в рамках лицензии. Доступная для разных ОС в рамках работы использованы варианты для windows и linux. Отличие в разных методах установки пакетов среды RStudio (install.packages(" "). В рамках данного отчет считаю возможным ограничится стандартной строчкой (# Загружаем пакет) и упустить процедуры установки.

1. Взять с сайта ***www.pogodaiklimat.ru/history.php*** данные о среднемесячной температуре воздуха в городе, Дрезден.

```
# Загружаем пакет XML
library(XML)
url <- "http://www.pogodaiklimat.ru/history/10488.htm"
table <- readHTMLTable(url, which = 2)
```

Преобразуем данные в числовой формат с фильтрами:

```
# Преобразование данных в таблице в числовой формат и замена 999.9 на NA
table[table == 999.9] <- NA
table[, -13] <- sapply(table[, -13], function(x) as.numeric(as.character(x)))
```

2. Привести в отчете таблицу входных данных (только ее начальный и конечный фрагменты).

	янв	фев	мар	апр	май	июн	июл	авг	сен	окт	ноя	дек	за год
1	-1.5	-0.9	4.2	9.2	13.0	16.8	18.8	15.6	12.8	8.1	4.3	2.5	

Рис. 1 Начальный фрагмент таблицы.

196	4.0	3.0	5.8	7.6	13.5	18.7	20.6	19.8	18.4	13.1	5.9	4.3	11.2
197	1.3	7.0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA

Showing 177 to 197 of 197 entries, 13 total columns

Рис. 2 Конечный фрагмент таблицы.

3. Для каждого месяца рассчитать и представить в отчете в виде удобочитаемой таблицы: объем выборки, минимальное значение, максимальное значение, первый квартиль, медиану, третий квартиль, межквартильный размах, среднее, стандартное отклонение, стандартную ошибку среднего, коэффициент вариации, коэффициент асимметрии.

Для выполнения этого пункта используем функцию describe слегка её дополняя необходимыми и отсутствующими в ней, но требующимися в отчете данными. В процессе многократного выполнения появилась необходимость задать фильтры для исключения ошибок связанных с появлением в расчетах нечисловых данных. Вывод таблицы формируем в таблицу со столбцами согласно заданию. (Решил добавить в расчет не только для каждого месяца, но и за год так как имеется соответствующая колонка хотя и меньшим объёмом выборки)

```
# Загрузка библиотеки
library(psych)
# Модифицированная функция describe
```

```
describe_with_quantiles <- function(x) {
  # Фильтрация данных: исключаем NA и нечисловые значения
  x <- as.numeric(x[!is.na(x) & grepl("^-?\\d+\\.?\\d*$", x)])
  desc <- describe(x)
  q1 <- quantile(x, 0.25)
  q3 <- quantile(x, 0.75)
  iqr <- q3 - q1
  cv <- sd(x) / mean(x)  # Коэффициент вариации
  result <- c(
    desc$n,
    desc$min,
    desc$max,
    q1,
    desc$median,
    q3,
    iqr,
    desc$mean,
    desc$sd,
    desc$se,
    cv,
    desc$skew )
  names(result) <- c(
    "n", "min", "max", "Q1", "median", "Q3", "IQR","mean", "sd", "se", "cv", "skew"  )
  return(result)}
# Применение модифицированной функции describe к каждому столбцу таблицы
month_descriptions_with_quantiles <- lapply(table[, 1:13], describe_with_quantiles)
# Преобразование списка в датафрейм
month_descriptions_with_quantiles_df <- do.call(rbind, month_descriptions_with_quantiles)
# Вывод таблицы
month_descriptions_with_quantiles_df
```

```
> month_descriptions_with_quantiles_df
         n   min  max     Q1 median     Q3   IQR       mean        sd         se          cv        skew
янв    197 -10.0  5.0 -2.300  -0.10  1.600 3.900 -0.3994924 3.0190128 0.21509576 -7.55712235 -0.65165916
фев    197 -11.1  7.0 -0.700   1.10  3.000 3.700  0.6883249 3.1606180 0.22518471  4.59175322 -0.91321054
мар    196  -3.2  8.2  2.200   4.25  5.625 3.425  3.9137755 2.2945340 0.16389528  0.58627123 -0.37302924
апр    196   4.2 13.8  7.275   8.55  9.925 2.650  8.5198980 1.7938819 0.12813442  0.21055204  0.03341960
май    196   9.1 18.8 12.300  13.50 14.700 2.400 13.4500000 1.7800461 0.12714615  0.13234543  0.09908209
июн    196  11.2 22.1 15.875  16.90 17.925 2.050 16.9173469 1.6261659 0.11615471  0.09612417  0.05702737
июл    196  15.1 23.5 17.400  18.65 19.500 2.100 18.5479592 1.5141879 0.10815628  0.08163636  0.17688369
авг    196  13.4 22.5 16.900  17.70 19.000 2.100 17.9969388 1.5392941 0.10994958  0.08553089  0.44673260
сен    196  10.0 18.5 13.300  14.30 15.200 1.900 14.3377551 1.4894856 0.10639183  0.10388555  0.13171885
окт    196   5.0 13.6  8.400   9.50 10.700 2.300  9.5448980 1.6500270 0.11785907  0.17287006 -0.10441954
ноя    196  -2.0  8.4  3.075   4.45  5.625 2.550  4.3035714 1.8454483 0.13181774  0.42881787 -0.30022740
дек    196  -7.8  6.9 -0.225   1.40  2.700 2.925  1.0173469 2.5618713 0.18299081  2.51818847 -0.70558583
за год 173   6.6 11.2  8.600   9.20  9.800 1.200  9.1277457 0.9266934 0.07045519  0.10152489 -0.17317895
> par(mfrow = c(3, 4))  # Устанавливаем макет графиков 3x4
```

Рис. 3 Отчет в виде таблицы.

4.1. Построить и привести в отчете *на одном графике* блочные диаграммы (с визуализацией выбросов) для всех двенадцати месяцев.

```
# Устанавливаем макет графиков 3x4
par(mfrow = c(3, 4))
for (i in 1:12) {  # Начинаем с 1-го столбца
        boxplot(table[, i], main = names(table)[i], outline = TRUE, na.rm = TRUE)}
```

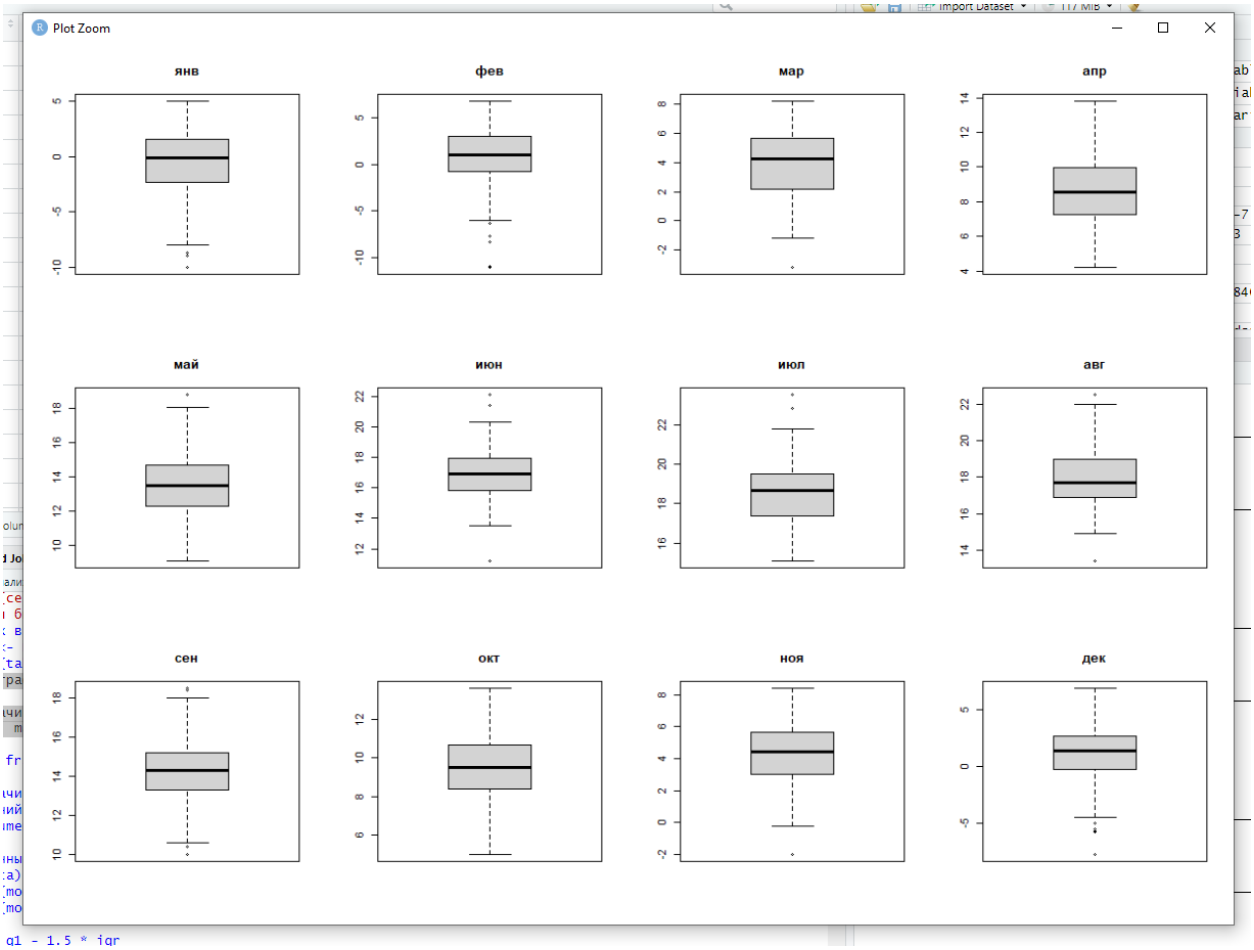

Рис. 4 Блочные диаграммы 12 месяцев.

4.2. Идентифицировать все выбросы и составить таблицу выбросов, в которой следует указать значение, год и месяц каждого выброса. Данный пункт создал больше всего проблем с написанием кода. Фактически с сайта выгружается 2 таблицы первая приведена на рисунках 1 и 2, а вторая содержит 1 столбик с годами. В рамках данной работы и этого пункта сопоставить их не получилось. Поэтому был придуман громоздкий механизм сопоставления 1 строки с годом 1827. Также чтобы

в таблицу не попадали данные с не числовыми данными NA был создан фильтр. Вывод табличных данных сформирован согласно заданию.

```
outliers_table <- data.frame(month = character(), Value = numeric(), Year = integer())

for (i in 1:12) {  # Начинаем с 1-го столбца и заканчиваем на 12-м
  # Фильтрация значений, исключая NA
  month_data <- as.numeric(table[!is.na(table[[i]]), i])

  # Проверка, что данные являются числами
  if(length(month_data) > 0) {
    q1 <- quantile(month_data, 0.25)
    q3 <- quantile(month_data, 0.75)
    iqr <- q3 - q1
    lower_bound <- q1 - 1.5 * iqr
    upper_bound <- q3 + 1.5 * iqr

    outliers_indices <- which(month_data < lower_bound | month_data > upper_bound)

    # Проверяем, есть ли выбросы в данном столбце, чтобы создать сопоставление строк
    if (length(outliers_indices) > 0) {
      row_numbers <- outliers_indices + 1827

      outliers <- data.frame(month = rep(names(table)[i], length(outliers_indices)),
                             Value = table[[i]][outliers_indices],
                             Year = row_numbers)
      outliers_table <- rbind(outliers_table, outliers)  }
  } }
# Выводим таблицу выбросов
print(outliers_table)
```

```
> # Выводим таблицу выбросов
> print(outliers_table)
   month Value Year
1    янв  -8.9 1838
2    янв -10.0 1940
3    янв  -8.9 1942
4    янв  -8.7 1963
5    фев  -6.4 1855
6    фев -11.1 1929
7    фев  -8.4 1947
8    фев -11.0 1956
9    фев  -7.7 1986
10   мар  -3.2 1845
11   май  18.8 1868
12   май  18.8 1889
13   июн  21.4 1889
14   июн  11.2 1923
15   июн  22.1 2019
16   июл  22.8 1994
17   июл  23.5 2006
18   авг  13.4 1833
19   авг  22.5 1842
20   сен  10.0 1912
21   сен  18.5 1947
22   сен  10.4 1996
23   сен  18.4 2023
24   ноя  -2.0 1858
25   дек  -7.8 1829
26   дек  -5.7 1840
27   дек  -5.8 1870
28   дек  -5.5 1879
29   дек  -5.0 1933
30   дек  -5.7 1969
> 
```

Рис. 5 Таблица выбросов.

5. Для каждого месяца привести гистограмму (график распределения) и сделать обоснованный вывод о наличии и степени выраженности свойства симметричности данных (положительная или отрицательная, сильная или слабая; при необходимости можно дополнительно воспользоваться коэффициентом асимметрии).

Для исключения визуальных ошибок и неточностей был вычислен коэффициент асимметрии и выведен на графиках месяцев.

```
# Загружаем пакет moments
library(moments)
```

```
# Устанавливаем макет графиков 3x4 для отображения гистограмм для каждого месяца
par(mfrow = c(3, 4))

# Создаем цикл для построения гистограммы для каждого месяца
for (i in 1:12) {
  # Получаем данные для текущего месяца, исключая NA
  month_data <- as.numeric(table[[i]][!is.na(table[[i]])])

  # Рисуем гистограмму
  hist(month_data, main = names(table)[i], xlab = "Value", ylab = "Frequency", col = "lightblue", border = "white")

  # Вычисляем коэффициент асимметрии
  skewness <- skewness(month_data)

  # Выводим значение коэффициента асимметрии
  text <- paste("Skewness:", round(skewness, 2))
  mtext(text, side = 1, line = -2, cex = 0.7)
}

# Сбросим макет графиков на значение по умолчанию
par(mfrow = c(1, 1))
```

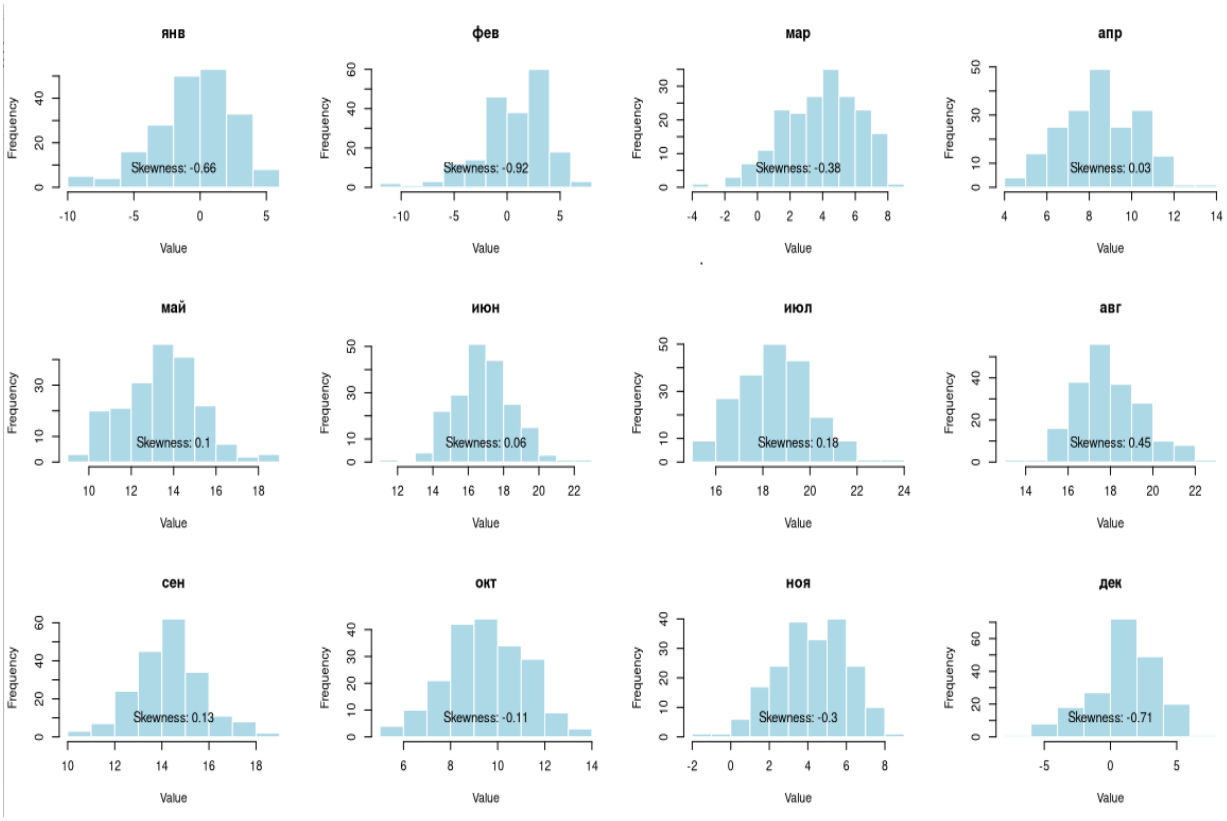

Рис. 6 График распределения.

Исходя из приведённых на рисунке 6 месячных гистограмм можно сделать следующие выводы.

Январь: отрицательная асимметрия так как левый хвост гистограммы длиннее правого также коэффициентом асимметрии отрицательный. Сильно выраженный на рисунке 5 приведена таблица с 4 выбросами в январе в левую (минусовую) сторону.

Февраль: отрицательная асимметрия так как левый хвост гистограммы длиннее правого также коэффициентом асимметрии отрицательный. Сильно выраженный на рисунке 5 приведена таблица с 5 выбросами в феврале в левую (минусовую) сторону.

Март: отрицательная асимметрия так как левый хвост гистограммы длиннее правого также коэффициентом асимметрии отрицательный. Слабо выраженный на рисунке 5 приведена таблица с 1 выбросами в марте в левую (минусовую) сторону.

Апрель: Данные симметричные. Выбросов нет. Коэффициент асимметрии 0.03.

Май: Данные симметричные. Коэффициент асимметрии 0.1.

Июнь: Данные симметричные. Коэффициент асимметрии 0.06.

Июль: положительная асимметрия так как правый хвост гектограммы длиннее левого. Слабо выраженная на рисунке 5 приведена таблица с 2 выбросами в правую (плюсовую) сторону. Коэффициент асимметрии 0.18.

Август: положительная асимметрия так как правый хвост гектограммы длиннее левого. Слабо выраженная на рисунке 5 приведена таблица с 2 выбросами в правую (плюсовую) сторону. Коэффициент асимметрии 0.45.

Сентябрь: Данные симметричные. Коэффициент асимметрии 0.13.

Октябрь: Данные симметричные. Коэффициент асимметрии -0.11.

Ноябрь: отрицательная асимметрия так как левый хвост гистограммы длиннее правого также коэффициентом асимметрии отрицательный. Слабо выраженный на рисунке 5 приведена таблица с 1 выбросами в ноябре в левую (минусовую) сторону.

Декабрь: отрицательная асимметрия так как левый хвост гистограммы длиннее правого также коэффициентом асимметрии отрицательный. Сильно выраженный на рисунке 5 приведена таблица с 6 выбросами в декабре в левую (минусовую) сторону. Коэффициент асимметрии -0.71.

6.1. Для каждого месяца рассмотреть гипотезу о том, подчиняется ли распределение данных нормальному закону на основе:

1) визуального анализа гистограммы (наличие симметрии и колоколообразности);

2) визуального анализа графика квантилей;

3) анализа численных характеристик набора данных (совпадение моды, медианы и среднего; соотношение между межквартильным размахом и станд.отклонением; попадание всех значений в диапазон «шести сигм»);

4) выбранного статистического критерия (например, Шапиро-Уилка) при $\alpha = 0{,}05$.

```
# Устанавливаем макет графиков 3x4 для отображения гистограмм и QQ-графиков для каждого месяца
par(mfrow = c(3, 4))
# Создаем вектор для хранения результатов теста Шапиро-Уилка
shapiro_p_values <- numeric(length = 12)
# Создаем цикл для анализа каждого месяца
for (i in 1:12) {
  # Получаем данные для текущего месяца, исключая NA
  month_data <- as.numeric(table[[i]][!is.na(table[[i]])])
  # Рисуем гистограмму
  hist(month_data, main = paste("Histogram for", names(table)[i]), xlab = "Value", ylab = "Frequency", col = "lightblue", border = "white")
  # Рисуем QQ-график с использованием базовой функции qqnorm
  qqnorm(month_data, main = paste("QQ-plot for", names(table)[i]))
  qqline(month_data)
```

```
    # Применяем тест Шапиро-Уилка
    shapiro_test <- shapiro.test(month_data)
    shapiro_p_values[i] <- shapiro_test$p.value
    # Выводим результат теста
    cat("Month:", names(table)[i], "\n")
    cat("Shapiro-Wilk test p-value:", shapiro_test$p.value, "\n")
    # Вычисляем моду, медиану и среднее
    mode_value <- names(sort(table(month_data), decreasing = TRUE))[1]
    median_value <- median(month_data)
    mean_value <- mean(month_data)
    # Вычисляем межквартильный размах и стандартное отклонение
    iqr_value <- IQR(month_data)
    sd_value <- sd(month_data)
    # Выводим численные характеристики
    cat("Mode:", mode_value, "\n")
    cat("Median:", median_value, "\n")
    cat("Mean:", mean_value, "\n")
    cat("Interquartile Range:", iqr_value, "\n")
    cat("Standard Deviation:", sd_value, "\n")
    # Проверяем, попадают ли все значения в диапазон "шести сигм"
    within_six_sigma <- sum(month_data > mean_value - 3 * sd_value & month_data < mean_value + 3 * sd_value) ==
length(month_data)
    cat("All values within six sigma range:", within_six_sigma, "\n\n")
}
# Сбросим макет графиков на значение по умолчанию
par(mfrow = c(1, 1))
```

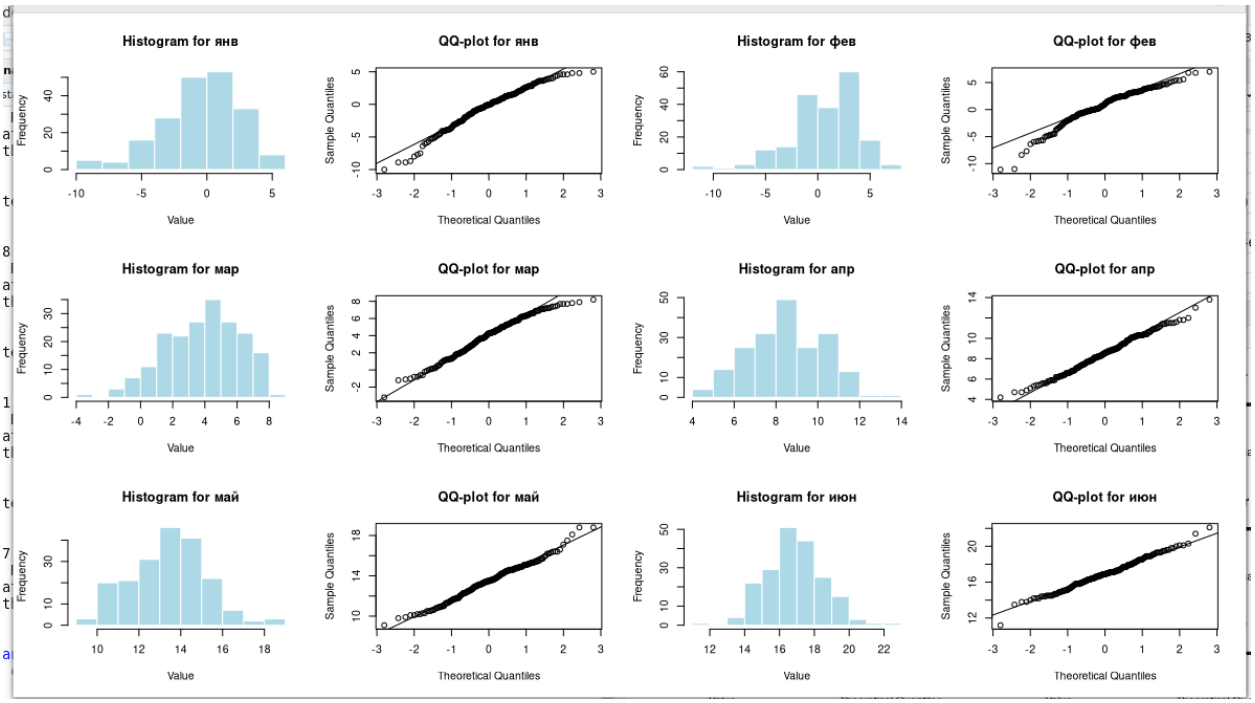

Рис.7 Месячное распределение данных январь-июнь.

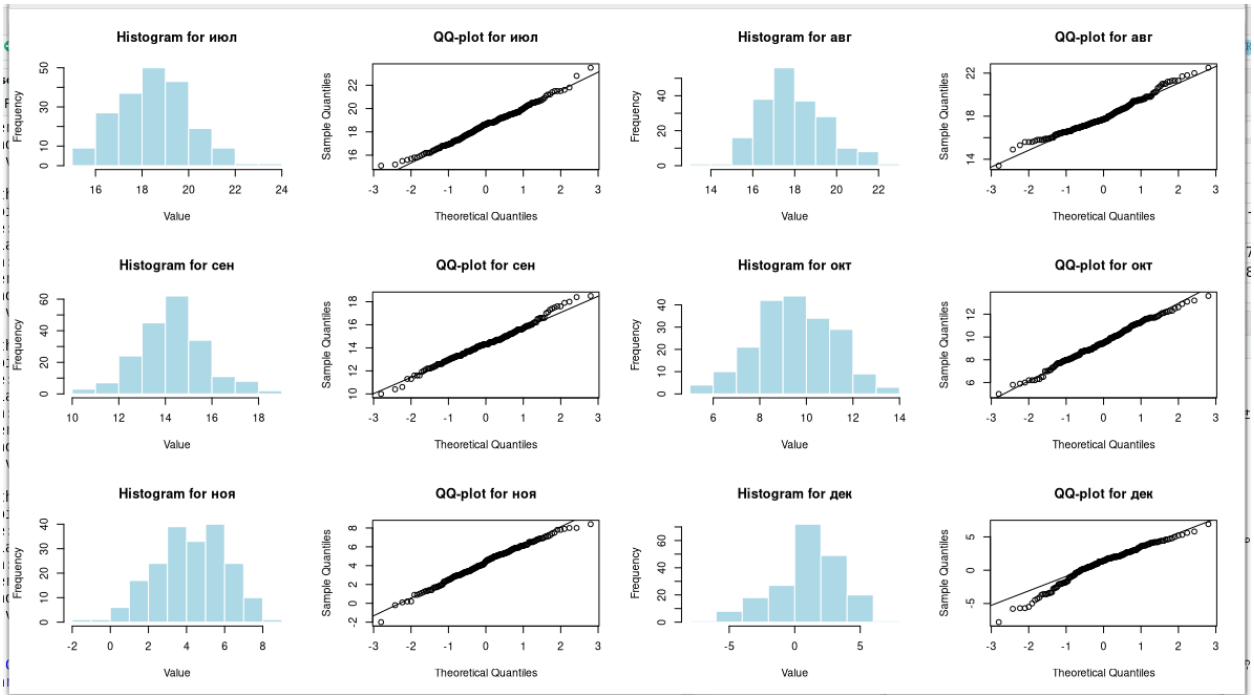

Рис. 8 Месячное распределение данных июль-декабрь.

```
Month: янв
Shapiro-Wilk test p-value: 0.0002212034
Mode: -0.7
Median: -0.1
Mean: -0.3994924
Interquartile Range: 3.9
Standard Deviation: 3.019013
All values within six sigma range: FALSE

Month: фев
Shapiro-Wilk test p-value: 2.332164e-06
Mode: 2.5
Median: 1.1
Mean: 0.6883249
Interquartile Range: 3.7
Standard Deviation: 3.160618
All values within six sigma range: FALSE

Month: мар
Shapiro-Wilk test p-value: 0.005081855
Mode: 3.7
Median: 4.25
Mean: 3.913776
Interquartile Range: 3.425
Standard Deviation: 2.294534
All values within six sigma range: FALSE

Month: апр
Shapiro-Wilk test p-value: 0.5469258
Mode: 10.3
Median: 8.55
Mean: 8.519898
Interquartile Range: 2.65
Standard Deviation: 1.793882
All values within six sigma range: TRUE

Month: май
Shapiro-Wilk test p-value: 0.1538455
Mode: 14.6
Median: 13.5
Mean: 13.45
Interquartile Range: 2.4
Standard Deviation: 1.780046
All values within six sigma range: FALSE

Month: июн
Shapiro-Wilk test p-value: 0.4344176
Mode: 17.1
Median: 16.9
Mean: 16.91735
Interquartile Range: 2.05
Standard Deviation: 1.626166
All values within six sigma range: FALSE
```

Рис.9 табличные данные.

```
Month: июл
Shapiro-Wilk test p-value: 0.4992383
Mode: 18.8
Median: 18.65
Mean: 18.54796
Interquartile Range: 2.1
Standard Deviation: 1.514188
All values within six sigma range: FALSE

Month: авг
Shapiro-Wilk test p-value: 0.003282923
Mode: 17.5
Median: 17.7
Mean: 17.99694
Interquartile Range: 2.1
Standard Deviation: 1.539294
All values within six sigma range: TRUE

Month: сен
Shapiro-Wilk test p-value: 0.3078189
Mode: 14.3
Median: 14.3
Mean: 14.33776
Interquartile Range: 1.9
Standard Deviation: 1.489486
All values within six sigma range: TRUE

Month: окт
Shapiro-Wilk test p-value: 0.6335555
Mode: 8.8
Median: 9.5
Mean: 9.544898
Interquartile Range: 2.3
Standard Deviation: 1.650027
All values within six sigma range: TRUE

Month: ноя
Shapiro-Wilk test p-value: 0.2408281
Mode: 5.2
Median: 4.45
Mean: 4.303571
Interquartile Range: 2.55
Standard Deviation: 1.845448
All values within six sigma range: FALSE

Month: дек
Shapiro-Wilk test p-value: 0.0001042269
Mode: 1.3
Median: 1.4
Mean: 1.017347
Interquartile Range: 2.925
Standard Deviation: 2.561871
All values within six sigma range: FALSE
```

Рис.10 Табличные данные.

Согласно приведённым данным из рисунков 7-10 очевидно, что нормальному закону распределения подчиняются данные месяцев апрель, август, сентябрь, октябрь.

6.2. Сформировать итоговый вывод о нормальности данных (для каждого месяца), используя совокупные результаты пп. 6.1.1 – 6.1.4.

Приведённые результаты на рисунках 9 и 10 полностью помесячно выводят характеристики для оценки нормальности данных, описанных мной в п 6.1