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

ОТЧЕТ

по практической работе №2

по дисциплине «Статический анализ»

Тема: Проверка статических гипотез

Вариант: Дрезден

Отчёт подготовил: Русских В.Д

Отчёт сдан: 10.04.2024

Студент гр.1323 _____ **Кошеляев А.С 50%**

Студент гр.1323 _____ **Русских В.Д 50%**

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

Санкт-Петербург

2024

Цель работы: анализ имеющегося набора данных с помощью статистических критериев и доверительных интервалов.

Задание:

1. Построить доверительный интервал уровня 1–α для среднегодовой температуры в выбранном для анализа городе, принимая здесь и далее значение α=0.06

```
library(ggplot2)

# Фильтр для проверки данных на корректность
correct_data <- complete.cases(table[, 13])  # Индексы строк с корректными данными

# Выбираем только корректные данные
temperatures <- table[correct_data, 13]

# Рассчитываем доверительный интервал
confidence_interval <- t.test(temperatures, conf.level = 0.94)$conf.int

# Выводим результаты
print(confidence_interval)

# Строим график
ggplot() +
  geom_point(aes(x = 1, y = mean(temperatures)), color = "blue") +
  geom_errorbar(aes(x = 1, ymin = confidence_interval[1], ymax = confidence_interval[2]), width = 0.1, color = "red") +
  annotate("text", x = 1, y = mean(temperatures), label = round(mean(temperatures), 2), vjust = -1, hjust = -0.5, color = "blue") +
  annotate("text", x = 1, y = confidence_interval[1], label = round(confidence_interval[1], 2), vjust = -1, hjust = -0.5, color = "red") +
  annotate("text", x = 1, y = confidence_interval[2], label = round(confidence_interval[2], 2), vjust = -1, hjust = -0.5, color = "red") +
  labs(x = "", y = "Среднегодовая температура", title = "Доверительный интервал для среднегодовой температуры") +
  theme_minimal()
```

результат выполнения кода приведён на рисунке 1.

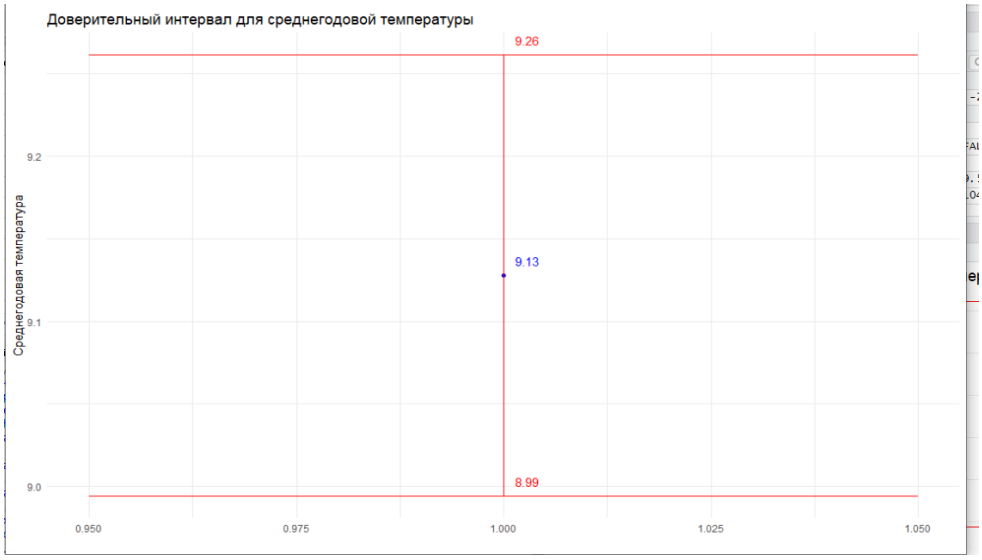


Рис.1 Доверительный интервал для среднегодовой температуры.

2. С помощью построенного доверительного интервала оценить *выполнение гипотезы* о равенстве среднегодовой температуры 0, 10 и 50 градусам по Цельсию

   Чтобы оценить выполнение гипотезы о равенстве среднегодовой температуры 0, 10 и 50 градусам по Цельсию, нужно проверить, попадают ли эти значения в построенный доверительный интервал. Если значение лежит внутри интервала, это может свидетельствовать о том, что гипотеза о равенстве среднегодовой температуры этому значению не противоречит.

```
# Значения для проверки
values_to_check <- c(0, 10, 50)

# Проверяем каждое значение
for (value in values_to_check) {
  if (value >= confidence_interval[1] && value <= confidence_interval[2]) {
    cat("Среднегодовая температура", value, "градусов по Цельсию содержится в доверительном интервале.\n")
  } else {
    cat("Среднегодовая температура", value, "градусов по Цельсию НЕ содержится в доверительном интервале.\n")
  }
}
```

```
Среднегодовая температура 0 градусов по Цельсию НЕ содержится в доверительном интервале.
Среднегодовая температура 10 градусов по Цельсию НЕ содержится в доверительном интервале.
Среднегодовая температура 50 градусов по Цельсию НЕ содержится в доверительном интервале.
>
```

   Рис.2 Вывод

   Гипотеза отвергается.

3. С помощью критерия хи-квадрат проверить гипотезу о том, что стандартное отклонение среднегодовой температуры не превышает 1, 2 и 3 градусов по Цельсию.

   Для этого нам нужно сначала вычислить выборочное стандартное отклонение среднегодовой температуры из имеющихся данных. Затем мы сравним это значение с заданными значениями (1, 2 и 3 градуса по Цельсию) с помощью критерия хи-квадрат.

```
# Вычисляем выборочное стандартное отклонение среднегодовой температуры
sample_sd <- sd(temperatures)

# Значения стандартного отклонения для проверки
values_to_check <- c(1, 2, 3)

# Проводим тест хи-квадрат для каждого значения
for (value in values_to_check) {
 # Вычисляем статистику хи-квадрат
 chi_sq_stat <- ((length(temperatures) - 1) * sample_sd^2) / value^2

 # Вычисляем p-value
 p_value <- 1 - pchisq(chi_sq_stat, df = length(temperatures) - 1)

 # Выводим результаты
 cat("Значение стандартного отклонения", sample_sd, "не превышает", value, "градусов по Цельсию:\n")
 cat("Статистика хи-квадрат:", chi_sq_stat, "\n")
 cat("p-value:", p_value, "\n")
 cat("\n")
}
```

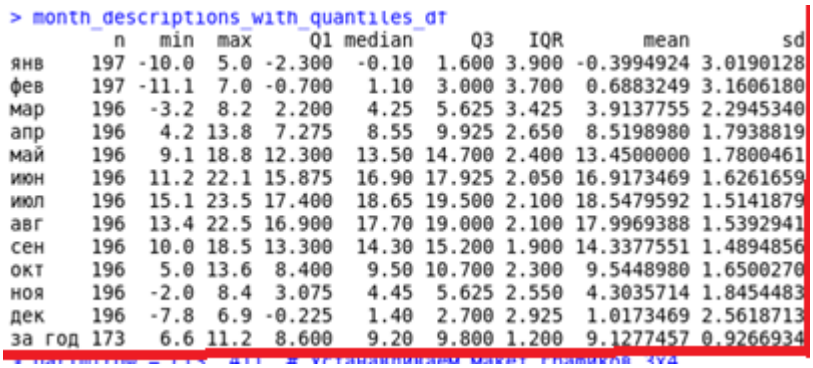
Результаты теста показывают, что значение выборочного стандартного отклонения, 0.9266934, не превышает заданные значения (1, 2 и 3 градуса по Цельсию).

При проверке гипотезы о том, что стандартное отклонение не превышает 1 градус по Цельсию, статистика хи-квадрат равна 147.7068, а p-value составляет 0.9099925. Поскольку p-value значительно больше обычного уровня значимости 0.06, нет оснований отклонять нулевую гипотезу о том, что стандартное отклонение не превышает 1 градус по Цельсию.

Точно так же при проверке гипотез о том, что стандартное отклонение не превышает 2 и 3 градуса по Цельсию, p-value = 1. Поэтому мы не можем отклонить нулевую гипотезу в этих случаях.

Это означает что нет достаточных доказательств для отклонения нулевой гипотезы. Нельзя считать, что стандартное отклонение превышает указанные значения. На основе проведенного теста гипотеза принимается.

Соотнести результаты проверки гипотез с вычисленной ранее выборочной оценкой станд.отклонения среднегод. температуры (1-ая практ.работа) рисунок 3.



```
> month_descriptions_with_quantiles_df
         n   min  max     Q1 median     Q3   IQR       mean        sd
янв    197 -10.0  5.0 -2.300  -0.10  1.600 3.900 -0.3994924 3.0190128
фев    197 -11.1  7.0 -0.700   1.10  3.000 3.700  0.6883249 3.1606180
мар    196  -3.2  8.2  2.200   4.25  5.625 3.425  3.9137755 2.2945340
апр    196   4.2 13.8  7.275   8.55  9.925 2.650  8.5198980 1.7938819
май    196   9.1 18.8 12.300  13.50 14.700 2.400 13.4500000 1.7800461
июн    196  11.2 22.1 15.875  16.90 17.925 2.050 16.9173469 1.6261659
июл    196  15.1 23.5 17.400  18.65 19.500 2.100 18.5479592 1.5141879
авг    196  13.4 22.5 16.900  17.70 19.000 2.100 17.9969388 1.5392941
сен    196  10.0 18.5 13.300  14.30 15.200 1.900 14.3377551 1.4894856
окт    196   5.0 13.6  8.400   9.50 10.700 2.300  9.5448980 1.6500270
ноя    196  -2.0  8.4  3.075   4.45  5.625 2.550  4.3035714 1.8454483
дек    196  -7.8  6.9 -0.225   1.40  2.700 2.925  1.0173469 2.5618713
за год 173   6.6 11.2  8.600   9.20  9.800 1.200  9.1277457 0.9266934
```

Рис.3 результат вычисления работа 1

4. На основании результатов 1-й практической работы выбрать для анализа один из четырех сезонов, среднемесячная температура в месяцах которого распределена достаточно близко к нормальному закону (обосновать сделанный выбор, приведя ту часть результатов предыдущей работы, где были сделаны выводы о нормальности выборок).

   Существует несколько способов оценить, насколько близко распределение среднемесячной температуры к нормальному закону. Далее приводится сезон, удовлетворяющий 3-м методам оценки, выполненным в работе 1.

   1.

   Если распределение данных близко к нормальному, гистограмма будет иметь форму колокола («колокольчатую форму»). Это означает, что большинство значений сосредоточены вокруг среднего значения, и они симметрично распределены относительно среднего. Возьмем график распределения работа 1 рисунок 4.

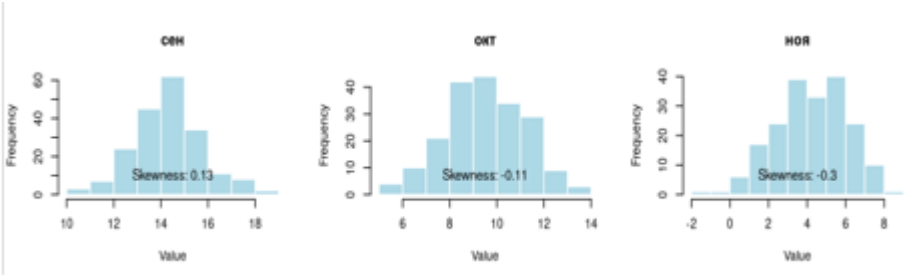


Рис.4 График распределения форма близкой к колоколу. Сентябрь, Октябрь, Ноябрь

2.

График квантилей-квантилей (Q-Q plot): Этот график сравнивает квантили набора данных с квантилями нормального распределения. Если данные близки к нормальному распределению, точки на графике будут следовать прямой линии. Отклонения от прямой линии могут указывать на отклонения от нормальности. Приведён на рисунке 5

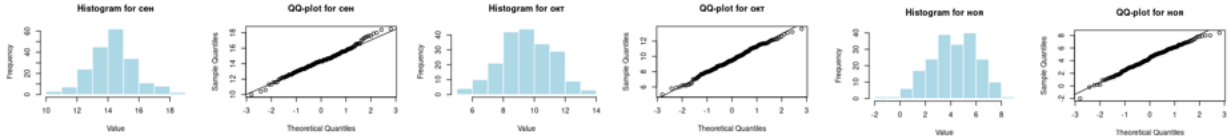


Рис.5 График квантилей-квантилей Сентябрь, Октябрь, Ноябрь

3.

Тест Шапиро-Уилка, который может проверить гипотезу о нормальности данных. Рисунок 6.

```
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
```

Рис.6 табличные данные.

Исходя из значений p-value теста Шапиро-Уилка:

Для сентября p-value составляет 0,3078189, что больше стандартной альфа ошибки 0,05. Это означает, что у нас недостаточно оснований для отклонения нулевой гипотезы о нормальности данных.

Для октября p-value равно 0,6335555, что также больше 0,05. Значит, у нас недостаточно оснований для отклонения нулевой гипотезы о нормальности данных.

Для ноября p-value составляет 0,2408281, что также больше 0,05. Следовательно, у нас недостаточно оснований для отклонения нулевой гипотезы о нормальности данных.

Таким образом, данные сентября, октября и ноября можно считать достаточно близкими к нормальному распределению.

Рассчитать средневзвешенную температуру выбранного сезона, получив значение среднесезонной температуры (ast).

```
# Задаем вектор количества дней в каждом месяце сезона
days_in_month <- c(30, 31, 30)  # В моём случае в каждом месяце 30 или 31 день

# Создаем пустой вектор для хранения  температур по годам
ast_yearly <- numeric(nrow(table))

# Рассчитываем среднесезонную температуру для каждого года
for (i in 1:nrow(table)) {
  # Берем данные из 9, 10, 11 столбцов (сен, окт, ноя) для текущего года
  season_temperatures <- table[i, c(9, 10, 11)]
  # Фильтруем значения NA
  season_temperatures <- season_temperatures[!is.na(season_temperatures)]
  # Если в году остались значения
  if (length(season_temperatures) > 0) {
    # Рассчитываем средневзвешенную температуру для текущего года
    ast_yearly[i] <- sum(season_temperatures * days_in_month) / sum(days_in_month)
  } else {
    # Если все значения NA, присваиваем NA
    ast_yearly[i] <- NA
  }
}
# Рассчитываем среднюю сезонную температуру за весь период наблюдений игнорим  NA
ast_season_total <- mean(ast_yearly, na.rm = TRUE)

# Выводим результат округлив до 2-х знаков
cat("Полученное значение среднесезонной температуры за весь период наблюдений:", round(ast_season_total,
2))
```

5. С помощью критерия Стьюдента проверить следующие статистические гипотезы:

   1) Средняя температура первого месяца сезона *не меньше* ast.

   2) Средняя температура второго месяца сезона *равна* ast.

   3) Средняя температура третьего месяца сезона *не больше* ast.

   Для проверки статистических гипотез с помощью критерия Стьюдента, необходимо:

   Выполним t-тест для каждой из трех гипотез.

   Используем критические значения t-статистики для уровня значимости alpha и степеней свободы, чтобы принять или отвергнуть каждую гипотезу.

```
# Извлекаем данные о температуре за сентябрь
september_temperatures <- table[, 9]

# Фильтруем значения NA
september_temperatures <- september_temperatures[!is.na(september_temperatures)]

# Считаем среднюю температуру за сентябрь
mean_september_temperature <- mean(september_temperatures)

# Выводим результат
cat("Средняя температура за сентябрь:", mean_september_temperature, "\n")

# Аналогично для остальных месяцев

# Извлекаем данные о температуре за октябрь
october_temperatures <- table[, 10]
october_temperatures <- october_temperatures[!is.na(october_temperatures)]
mean_october_temperature <- mean(october_temperatures)
cat("Средняя температура за октябрь:", mean_october_temperature, "\n")

# Извлекаем данные о температуре за ноябрь
november_temperatures <- table[, 11]
november_temperatures <- november_temperatures[!is.na(november_temperatures)]
mean_november_temperature <- mean(november_temperatures)
cat("Средняя температура за ноябрь:", mean_november_temperature, "\n")
```

1) Средняя температура первого месяца сезона *не меньше* ast.

```
t_test_september <- t.test(september_temperatures, mu = 9.4, alternative = "less")
print(t_test_september)
```

Результаты для проверки первой гипотезы (Средняя температура первого месяца сезона не меньше 9.4) следующие:

t-статистика: 46.411

p-значение: 1

Уровень значимости: 0.06

Так как p-значение больше уровня значимости, у нас нет оснований отклонить нулевую гипотезу. Следовательно, на основании имеющихся данных средняя температура в сентябре не меньше 9.4.

2) Средняя температура второго месяца сезона *равна* ast.

```
t_test_october <- t.test(october_temperatures, mu = 9.4, alternative = "two.sided")
print(t_test_october)
```

Результаты для проверки второй гипотезы (Средняя температура второго месяца сезона равна 9.4) следующие:

t-статистика: 1.2294

p-значение: 0.2204

Уровень значимости: 0.06

Так как p-значение больше уровня значимости, у нас нет оснований отклонить нулевую гипотезу. Следовательно, на основании имеющихся данных средняя температура в октябре не отличается от 9.4.

3) Средняя температура третьего месяца сезона *не больше* ast.

```
t_test_november <- t.test(november_temperatures, mu = 9.4, alternative = "less")
print(t_test_november)
```

Результаты для проверки третьей гипотезы (Средняя температура третьего месяца сезона не больше 9.4):

t-статистика: -38.663

p-значение: < 2.2e-16

Уровень значимости: 0.06

Так как p-значение крайне мало, мы можем отклонить нулевую гипотезу о том, что средняя температура равна или больше 9.4 в пользу альтернативной гипотезы о том, что средняя температура меньше 9.4.

6. Используя критерий Стьюдента и поправку Бонферрони проверить для выбранного сезона гипотезу о том, что все три выборки значений среднемесячной температуры извлечены из одной

генеральной совокупности, т.е. *что их средние значения статистически одинаковы*.

Для оцени сезона из 3-х месяцев оптимально применить дисперсионный анализ (ANOVA).

Хотя оба метода используются для статистического анализа данных, критерий Стьюдента и дисперсионный анализ (ANOVA) различаются по своим основным целям и применению.

Критерий Стьюдента:

- Используется для сравнения средних значений двух групп.
- Подходит для случаев, когда мы хотим определить, есть ли статистически значимые различия между средними значениями двух независимых выборок.
- Может быть одновыборочным (для сравнения среднего значения выборки с известным значением) или двухвыборочным (для сравнения средних значений двух независимых выборок).

ANOVA:

- Используется для сравнения средних значений трех или более групп.
- Позволяет определить, есть ли статистически значимые различия между средними значениями двух или более групп.
- Предполагает, что наблюдения в каждой группе независимы и имеют одинаковую дисперсию.
- Может быть однофакторным (один фактор с несколькими уровнями) или многофакторным (два и более факторов, каждый с несколькими уровнями).

Таким образом, хотя оба метода связаны с оценкой средних значений, они применяются в разных ситуациях и для разного числа групп данных.

```
# Загружаем библиотеку для работы с ANOVA
library(stats)

# Создаем датафрейм с данными о температуре за каждый месяц
temperatures <- data.frame(
  Month = rep(c("September", "October", "November"), each = length(september_temperatures)),
  Temperature = c(september_temperatures, october_temperatures, november_temperatures)
)

# Выполняем дисперсионный анализ
anova_result <- aov(Temperature ~ Month, data = temperatures)

# Выводим результаты анализа
print(summary(anova_result))

# Применяем поправку Бонферрони к уровню значимости
alpha <- 0.06
alpha_bonferroni <- alpha / 3  # 3 - количество сравнений (месяцев)
print(alpha_bonferroni)
```

Вывод представлен на рисунке 7

```
> print(summary(anova_result))
             Df Sum Sq Mean Sq F value Pr(>F)
Month         2   9874    4937    1774 <2e-16 ***
Residuals   585   1628       3
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
> # Применяем поправку Бонферрони к уровню значимости
> alpha <- 0.06
> alpha_bonferroni <- alpha / 3  # 3 - количество сравнений (месяцев)
> print(alpha_bonferroni)
[1] 0.02
>
```

Рис.7 Анализ гипотезы температуры извлечены из одной генеральной совокупности.

"Pr(>F)" - значение, которое показывает вероятность получить такие или более экстремальные результаты при условии верности нулевой гипотезы (т.е., что средние значения всех групп одинаковы). В данном случае p-значение очень мало (<2e-16). Результаты ANOVA с поправкой Бонферрони показывают, что p-значение меньше уровня значимости α=0.02. Это означает, что мы отвергаем нулевую гипотезу о равенстве средних значений среднемесячной температуры для всех трех месяцев. Таким

образом, мы имеем статистически значимые различия между средними температурами за разные месяцы и не можем сказать, что они извлечены из одной и той же генеральной совокупности.