ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ

ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ

«БАШКИРСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»

Отчет по лабораторной работе №4

«Однофакторный дисперсионный анализ»

Выполнил: студент гр. 1ФКЦ

Ахметов Руслан Олегович

Проверил: кандидат технических наук, доцент

Прудников Вадим Борисович

Уфа – 2023

1. Проведем однофакторный дисперсионный анализ. Подключаем пакет ggplot2 (пакет ggplot2 предназначен для построения двумерных графиков и диаграмм) и отбираем только две требуемые переменные в набор данных df (рис.1).

Напомним, в статистике любая сумма квадратов отклонений связана с числом степеней свободы df (degress of freedom), которое показывает, сколько независимых отклонений из n возможных участвует в образовании данной суммы квадратов.

library(ggplot2)

df = var3[, c('wage', 'educ')]





Рисунок 1.

Отбрасываем ненужные уровни фактора и отсутствие образования устанавливаем базовым уровнем (рис.2): df$educ <- factor(df$educ, levels=c('нет', 'Учитесь', 'Учились'))

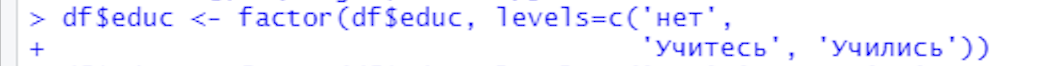


Рисунок 2.

Для визуального анализа наличия потенциальных различий в групповых средних построим график «боксплот» (рис. 3, 4):

ggplot(df, aes(x = educ, y = wage)) +

geom\_boxplot(fill = "grey80", colour = "blue") +

scale\_x\_discrete() + xlab("Education") + ylab("Wage")

fill – заполнение

grey - серый

scale – масштаб

discrete – дискретный

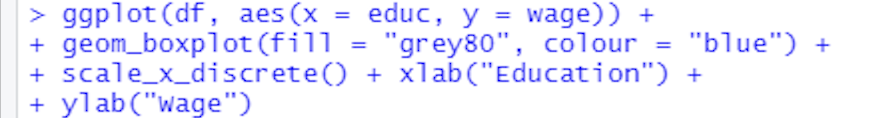


Рисунок 3.

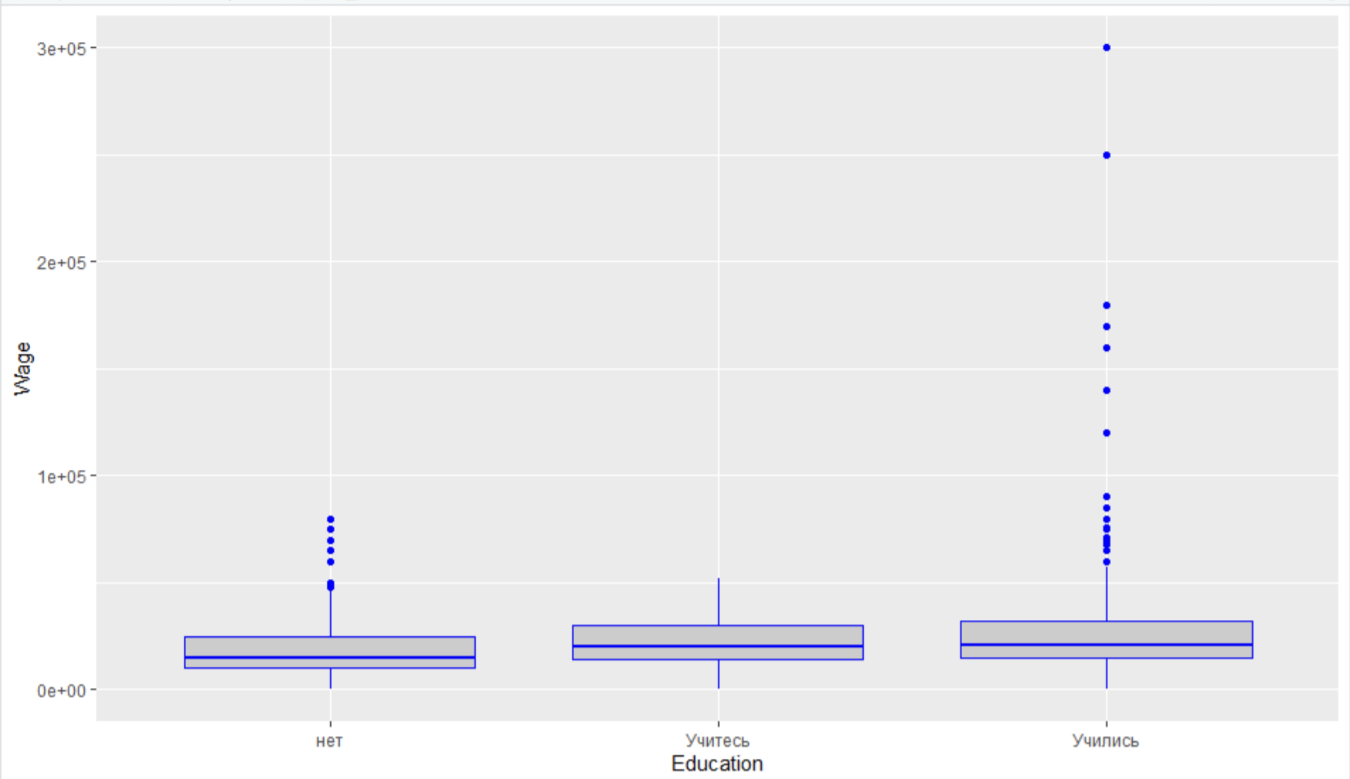


Рисунок 4 - График «boxplot» значений заработной платы для трех уровней фактора образования

На основе визуального анализа (рис. 4) можно предположить, что имеются различия в средних как минимум между группами «Учились» и «Нет».

Построим линейную модель (что позволяет выявить возможные различий групповых средних) и на ее основе проведем дисперсионный анализ с помощью функции anova (рис.5).

anova в статистике - это мощный инструмент для определения влияния различных групп наблюдений между собой.   
anova - это сокращение от analysis of variance (дисперсионный анализ).

Линейная модель: m1 = lm(wage ~ educ, data = df) summary(m1)

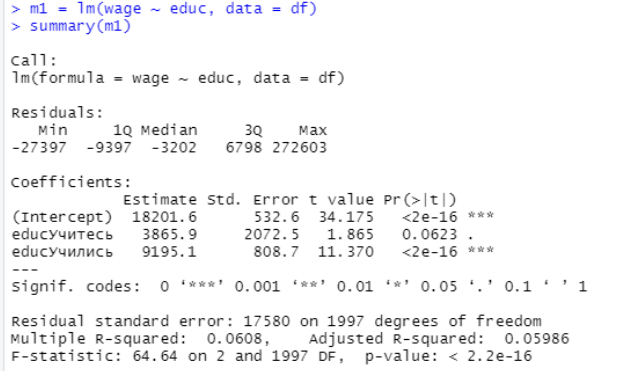


Рисунок 5 – Анализ линейной модели m1

Residuals – остаточное количество

Estimate – оценка

residual standard – остаточный стандарт

multiple r-squared – кратный р-квадрат

adjusted r-squared – скорректированный р-квадрат

p-value — это вероятность получить значение статистики критерия равное наблюдаемому или более нетипичное по сравнению с наблюдаемым при условии, что нулевая гипотеза верна; получается, что р-значение теста

Результаты оценивания (рис. 5) линейной модели свидетельствуют о возможном наличии различий между средними в группах «Учитесь» и «Нет» (p-значение равно 0.0623), а также между средними в группах «Учились» и «Нет» (p-значение <2е-16).

Проведем однофакторный дисперсионный анализ (рис. 6).

anova(m1)

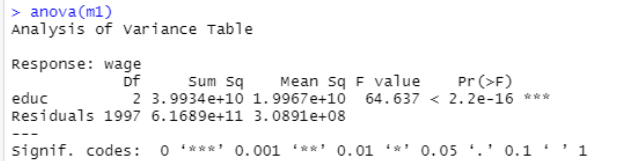


Рисунок 6 – Результаты однофакторного дисперсионного анализа

Response - ответ

Residuals – остаточное количество

Из рис. 6 видим, что p-значение, соответствующее наблюдаемому значению тестовой 𝐹-статистики, <2e-16, что значительно меньше любого разумного уровня значимости, а значит нулевую гипотезу о равенстве групповых средних следует отвергнуть.

Выполнение дисперсионного анализа тоже осуществляется при помощи **функции** **aov**(). Заметим, что при использовании функции aov, получаем тот же результат, что и при последовательном применении функций lm и anova: m0 <- aov(data=df, wage ~ educ); summary(m0).

Построим доверительные интервалы для различий (рис. 7, 8): confint(m1)

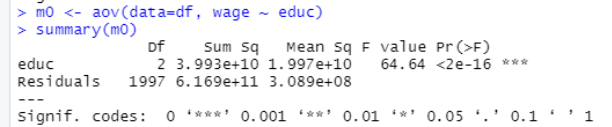


Рисунок 7.

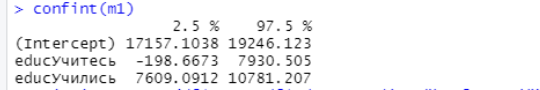
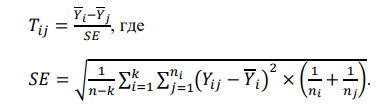


Рисунок 8 – Доверительные интервалы для различий межгрупповых средних

2. Поправка Бонферрони заключается в делении уровня значимости (по умолчанию – 0.05) на число тестов (парных сравнений). Такой подход считается более консервативным.

: (групповые средние равны) против альтернативной гипотезы : .

Тестовая статистика:

**

При проверке используется скорректированное значение уровня значимости 𝛼`= 𝛼/ , где 𝑘 – число сравниваемых групп.

В R синтаксис следующий: pairwise.t.test(df$wage, df$educ, p.adj = "bonferroni") – с учетом поправки и pairwise.t.test(df$wage, df$educ, p.adj = "none") – без нее. Результаты тестирования попарных сравнений представлены на рис. 9

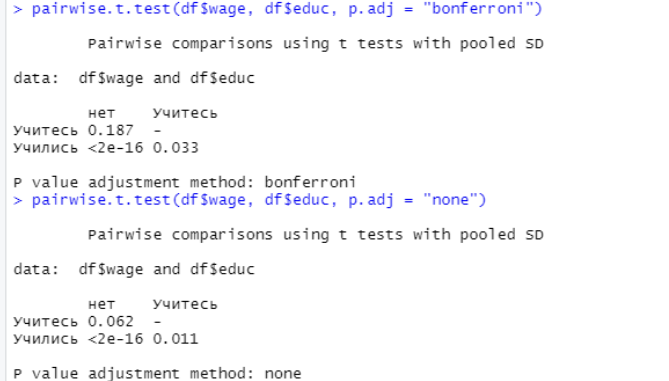


Рисунок 9 - Парное тестирование равенства групповых средних

Можно сделать вывод о том, что различия в средних по группам «Нет» и «Учились» значимы в любом случае, а различия между группами «Нет» и «Учитесь», а также «Учились» и «Учитесь» остаются незначимыми, если делать поправку Бонферрони.

Для применения метода Фишера LSD требуется подключение пакета agricolae:

library('agricolae');

MS.error <- deviance(m1)/df.residual(m1);

(LSD.test(df$wage, df$educ, df.residual(m1), MS.error, group=FALSE)).

Deviance – отклонение от нормы

Форма представления результатов оценивания немного другая и представлена на рис. 10.

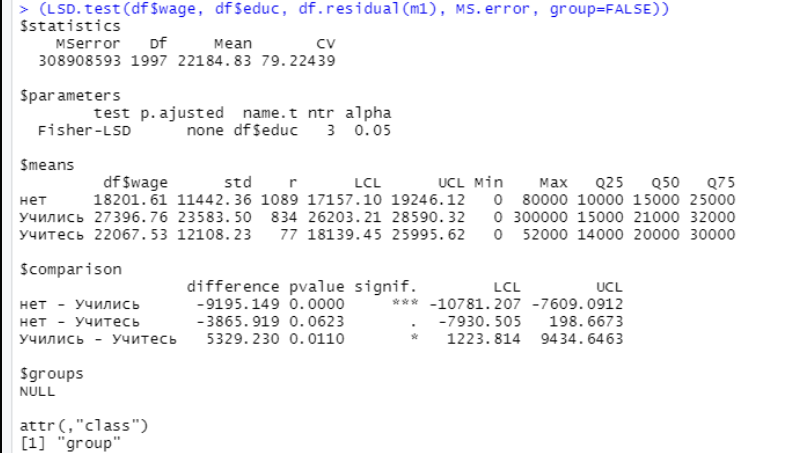


Рисунок 10 – Результаты теста LSD Фишера

Обратим внимание на часть, посвященную результатам сравнения групповых средних (**comparison** - сравнение). Видим, что p-значения соответствуют значениям, полученным для теста pairwise.t.test (попарный тест) без учета поправки. Кроме этого, функция рассчитывает границы доверительного интервала для различия групповых средних (LCL, UCL).

Применим теперь HSD метод Тьюки и построим график для доверительных интервалов различий:

hsd <- TukeyHSD(aov(df$wage ~ df$educ));

hsd par(mfrow=c(1,1))

par(mar=c(8.1, 8.1, 4.1, 2.1))

plot(hsd, las=2).

Результаты тестирования и построения доверительных интервалов приведены на рис. 11 и рис. 12 соответственно.

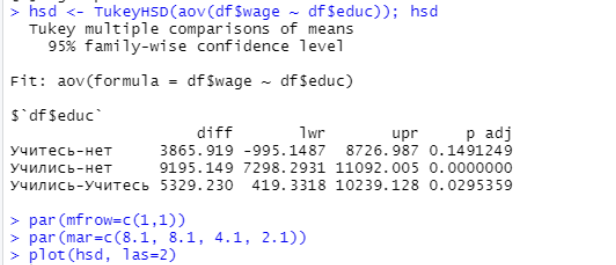


Рисунок 11 Результаты теста HSD Тьюки

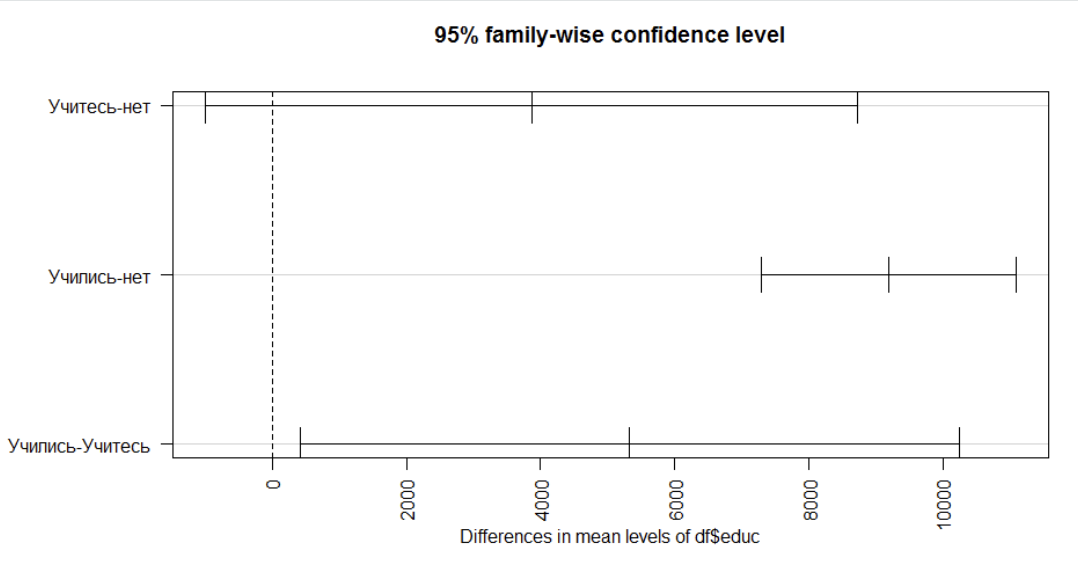


Рисунок 12 – График 95% доверительных интервалов для различий

Из рисунков 11 и 12 делаем вывод о том, что метод Тьюки, как и учет поправки Бонферрони рекомендует считать различными средние лишь для групп «Учились» и «Нет». Различий между другими группами метод на уровне значимости 5% не выявляет.

Вывод: при построении графика " boxplot " значений заработной платы для трех уровней фактора образования предположили, что имеются различия в средних как минимум между группами "Учились" и "Нет"

Результат оценивания линейной модели свидетельствует о возможном наличии различий между средними в группах "Учитесь" и "Нет", а также между средними в группах "Учились" и "Нет".

В результате проведения однофакторного дисперсионного анализа сделали вывод о том, что p-значение значительно меньше любого разумного уровня значимости, значит нулевую гипотезу о равенство групповых средних отвергаем.

В результате проведения парного тестирования равенства групповых средних можно сделать вывод о том, что различия в средних по группам "Нет" и "Учились" значимы в любом случае, а различия между группами "Нет" и "Учитесь", а также "Учились" и "Учитесь" остаются неизменными, если делать поправку Бонферрони

В результате проведения теста LSD Фишера видим, что р-значения соответствуют значениям, полученным для парного тестирования равенства групповых средних без учёта поправки

Метод Тьюки, как и учёт поправки Бонферрони рекомендует считать различными средние лишь для групп "Учились" и "Нет".