**Федеральное агентство по образованию Российской Федерации**

**Государственное образовательное учреждение**

**высшего профессионального образования**

**«КУБАНСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»**

**Кафедра прикладной математики**

Преподаватель,

д.т.н. А.А. Халафян

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**ОТЧЕТ**

**СТАТИСТИЧЕСКИЙ АНАЛИЗ ЛУЧШИХ ДЕЙСТВУЮЩИХ ИГРОКОВ НБА**

Работу выполнил студент 4 курса

факультета компьютерных технологий и прикладной математики  
спец. 01.03.02 – Прикладная математика и информатика

Сторчак Вадим, группа 4ИТ

Краснодар 2021

**СОДЕРЖАНИЕ**

[**1 Исходные данные** 4](#_Toc87056809)

[**2 Графический анализ** 6](#_Toc87056810)

[**2.1 2D Graphs** 6](#_Toc87056811)

[**2.1.1 2D Histogramms** 6](#_Toc87056812)

[**2.1.2 2D Scatterplots** 8](#_Toc87056813)

[**2.1.3 2D Box Plots** 10](#_Toc87056814)

[**2.2 Средство «закрашивание»** 11](#_Toc87056815)

[**3 Основные статистики** 14](#_Toc87056816)

[**3.1 Описательные статистики** 14](#_Toc87056817)

[**3.2 Корреляционная матрица** 14](#_Toc87056818)

[**3.3** **Критерий Стьюдента сравнения средних** 15](#_Toc87056819)

[**3.3.1 t-test, independent, by groups** 15](#_Toc87056820)

[**3.3.2 t-test, independent, by variables** 16](#_Toc87056821)

[**3.3.3 t-test, dependent samples** 16](#_Toc87056822)

[**3.3.4 t-test, single samples** 16](#_Toc87056823)

[**4 Группировка и однофакторная ANOVA** 17](#_Toc87056824)

[**5 Дисперсионный анализ** 22](#_Toc87056825)

[**6** **Непараметрическая** **статистика** 25](#_Toc87056826)

**1 Исходные данные**

Изображение выглядит как стол

Автоматически созданное описание

В исходной таблице представлены данные по действующим лучшим игрокам баскетбольной лиги NBA.

Показатели:

1. Player – имя и фамилия игрока;
2. Age – возраст игрока на текущий момент;
3. Цвет кожи – цвет кожи игрока;
4. Национальность – национальность игрока, за какую сборную играет;
5. TM – команда, за которую играет игрок;
6. Pos – позиция, под которой играет игрок;
7. G – количество игр, сыгранных игроком на момент сбора статистики;
8. GS – количество игр, в которых игрок выходит в стартовой пятерке;
9. MP – количество сыгранных минут;
10. FG – количество забитый бросков;
11. FGA – количество совершённых попыток;
12. FG% – общий процент попаданий игровых бросков;
13. 3P – количество реализованных трёх очковых бросков;
14. 3PA – количество совершённых трёх очковых попыток;
15. 3P% – процент реализованных трёх очковых бросков;
16. 2P – количество забитых двух очковых бросков;
17. 2PA – количество совершённых двух очковых попыток;
18. 2P% - процент реализованных двух очковых бросков;
19. FT – количество забитых штрафных бросков;
20. FTA – количество совершённых попыток штрафного броска;
21. FT% - процент попадания со штрафной линии;
22. ORB – количество подборов совершённых на щите соперника;
23. DRB – количество подборов совершённых на своё щите;
24. TRB – общее количество подборов;
25. AST – количество отданных результативных передач;
26. STL – количество отборов мяча у соперника;
27. BLK – количество заблокированных бросков;
28. TOV – количество потерь игрока;
29. PF – количество персональных замечаний игрока;
30. PTS – количество набранных очков;
31. Trp Dbl – количество триплдаблов (когда игрок набирает 10 и более в трёх статистически полезных показателях: PT, AST, REB, BLK, STL. В рамках одного матча)

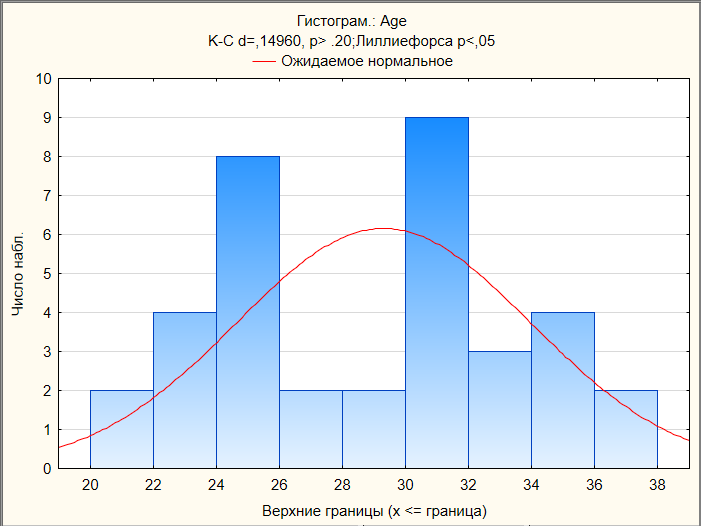
**2 Графический анализ**

**2.1 2D Graphs**

**2.1.1 2D Histogramms**

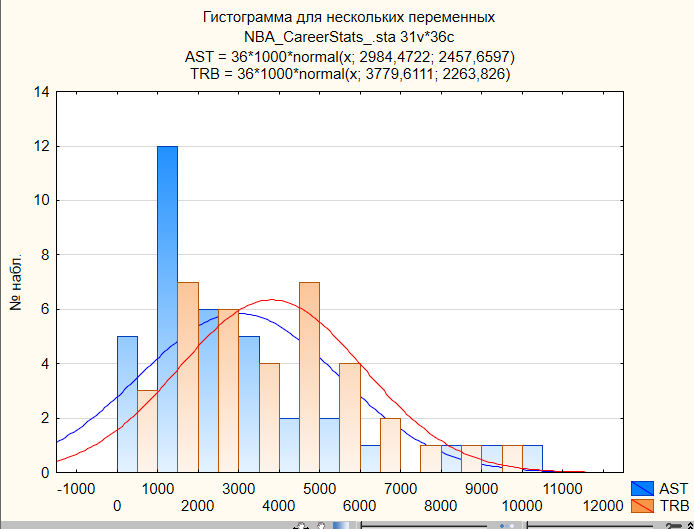
2D Histogramms являются графическими представлениями распределения частот выбранных переменных.

2D Histogramms Regular (простые) – столбчатая диаграмма распределения частот для выбранной переменной.



Гистограмма была построена по количественной переменной Age (возраст) и показывает, что в генеральной совокупности чаще встречаются игроки возраста 30-32 лет (9 игроков) и возраста 24-26 (8 игроков), что составляет 25% и 22,22% соответственно.

2D Histogramms Multiple (составные) – изображают распределение частот для нескольких переменных на одном графике.



Данная гистограмма построена по общему числу результативных передач и общему числу подборов у игроков. Из гистограммы видно, что большинство игроков имеют от 1000 до 2000 общих подборов и передач. Обусловлено это двумя причинами:

* формирование выборки;
* кол-во игроков, которые отыграли большое число сезонов в лиге крайне мало, так как такие игроки являются звездами и их по определению не может быть много.

2D Histogramms Double-Y (с двойной осью Y) – Комбинация двух по-разному масштабированных составных гистограмм.



Данная гистограмма построена по двум разномасштабным переменным: общее количество бросков и кол-во набранных очков. Из диаграммы видно, что чаще всего встречаются игроки набравшие от 5к до 10к очков, а так же совершившие от 10к до 15к, и от 0 до 5к попыток броска.

**2.1.2 2D Scatterplots**

2D Scatterplots (диаграммы рассеяния) визуализируют зависимость между двумя переменными.

2D Scatterplots Regular (протые) - визуализируют зависимость между двумя переменными X и Y.



Данная диаграмма отображает зависимости между количеством выброшенных бросков и набранными очками за карьеру с линейной подгонкой.

2D Scatterplots Multiple (составные) – состоит из нескольких зависимостей и изображает несколько корреляций.



Данная диаграмма отображает зависимости между кол-вом сыгранных матчей и числом забитых очков, а так же между кол-вом сыгранных матчей и кол-вом совершенных попыток атаки кольца.

2D Scatterplots Double-Y (с двойной осью Y) – комбинация двух составных диаграмм рассеяния для одной переменной X и двух различных наборов переменных Y.



Данная диаграмма отображает комбинацию двух составных диаграмм: первая зависимость между количеством реализованных трёхочковых бросков и количеством сыгранных игр, вторая зависимость между количеством реализованных двухочковых бросков и количеством сыгранных игр.

**2.1.3 2D Box Plots**

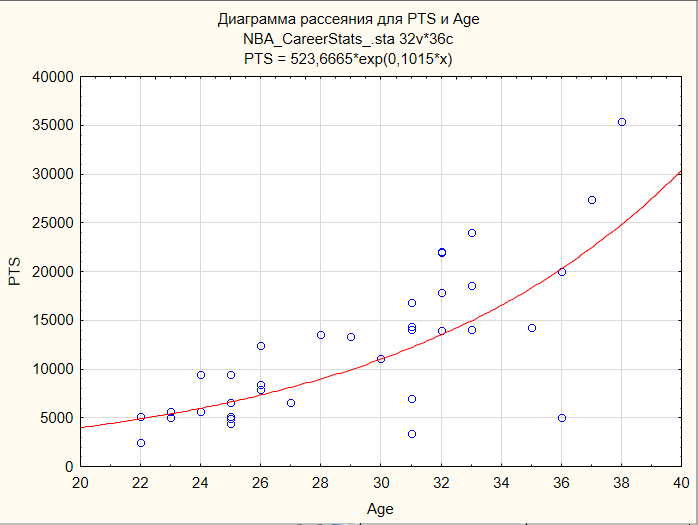
2D Box Plots (графики ящика – диаграммы размаха) – на этих диаграммах изображаются диапазоны или характеристики распределения значений выбранной переменной отдельно по группам, заданным категориальной переменной.



На диаграмме представлены характеристики распределения значений общих набранных очков по позициям игроков.

**2.2 Средство «закрашивание»**

Построим диаграмму рассеивания для переменных количества набранных очков за карьеру и возраста игрока.



Из диаграммы видно, что есть точки, которые располагаются далеко от линии регрессии, поэтому аппроксимация может быть некачественной. Для исключения этой ситуации воспользуемся средством «закрашивание», чтобы исключить некоторые точки.

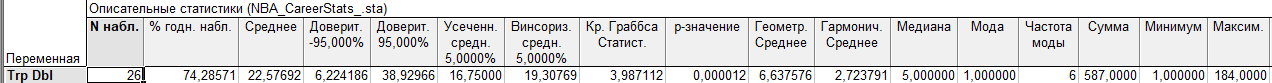


Сравнивая r до и после исключения точки, можно сказать, что качество аппроксимации увеличилось (доля объясненной дисперсии).

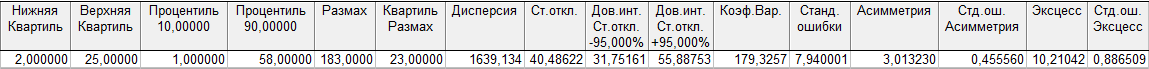
**3 Основные статистики**

**3.1 Описательные статистики**

Таблица описательных статистик для количества сделанных трипл даблов:



*продолжение*



В таблице представлены следующие статистики: число наблюдений 26, среднее 22.57692, сумма 587, медиана 5, геометрическое среднее 6.637576, гармоническое среднее 2.723791, стандартное отклонение 2.423481, дисперсия 1639.134, стандартная ошибка среднего 7.940001, доверительные пределы для среднего 6.224186 и 38.92966, ассиметрия 3.013230, стандартная ошибка ассиметрии 0.4555560, эксцесс 10.21042, стандартная ошибка эксцесса 0.886509, минимум 1, максимум 184, нижний квартиль 2, верхний квартиль 25, 10-я процентиль (квантиль 0,1) 1, 90- процентиль (квантиль 0,9) 58, размах 183, квартильный размах 23.

**3.2 Корреляционная матрица**

Изображение выглядит как текст, кроссворд

Автоматически созданное описание

Из таблицы видно, что сильная корреляция r>0.75 присутствует между средними количеством сыгранных игр и всем показателями кроме результативных передач, сыгранными минутами и всеми показателями, количеством реализованных бросков и всеми показателями кроме результативных передач, между общим числом подборов и сыгранных игр, сыгранных минут, совершенных бросков, а так же между возрастом игрока и общим числом подборов и сыгранных игр, сыгранных минут; средняя корреляция между количеством сыгранных матчей и результативными передачами, реализованных бросков и результативными передачами, результативными передачами и всеми другими показателями. Слабой корреляции не наблюдается. Красным отмечены значения, где уровень значимости <0.05.

## 3.3 Критерий Стьюдента сравнения средних

**3.3.1 t-test, independent, by groups**

Проверим равенство средних для количества сыгранных игр, числу сыгранных минут и количеству набранных типл даблов, по группам игроков на позиции разыгрывающего и малого форварда на основе t-критерия (t-test, independent, by groups).

Изображение выглядит как стол

Автоматически созданное описание

Как видно из таблицы все p>0.05, принимаем гипотезу о равенстве средних, то есть отличие средних статистически не значимо. Это связано с тем, что количество сыгранных минут и проведённых матчей зависит в первую очередь от таланта и здоровья игрока, но никак не от позиции, а вот для параметра количества трипл даблов недостаточная выборка (так как опираясь на собственный опыт, могу сказать, что количество триплдаблов должно быть сильно больше у разыгрывающих и атакующих защитников, если брать большую выборку). Для корректности вывода отличие дисперсий не должно быть статистически значимым (p дисп > 0.05). Соответственно, относительно всех выбранных переменных в рамках выбранных групп вывод корректен.

**3.3.2 t-test, independent, by variables**

Предположим, что процент попаданий с трёхочковых бросков и процент попадания двух очковых бросков статистически независимы. Тогда для них проверим равенство средних на основе t-критерия (t-test, independent, by variables).

Изображение выглядит как стол

Автоматически созданное описание

По данным таблицы можно сделать вывод, что отличие средних статистически значимо, так как p>0.05.

**3.3.3 t-test, dependent samples**

Проверим равенство средних для общего количества сыгранных матчей и количества матчей в стартовом составе на основе t-критерия (t-test, dependent samples).

Изображение выглядит как стол

Автоматически созданное описание

По данным таблицы уровень значимости меньше 0.05, поэтому гипотезу о равенстве средних отвергаем, отличие средних статистически значимо. Также, так как отличие статистически значимо, мы можем перенести вывод на генеральную совокупность, то есть в среднем общее количество матчей больше, чем количество матчей в старте.

**3.3.4 t-test, single samples**

Предположим, что в среднем за карьеру игроки совершают 8000 попыток броска за карьеру. Проверим гипотезу, что среднее количество попыток броска за карьеру из выборки не отличается статистически значимо от указанного среднего.

Изображение выглядит как текст, стол

Автоматически созданное описание

По данным таблицы можно сделать вывод, что отличие статистически не значимо, так как p>0.05.

**4 Группировка и однофакторная ANOVA**

Для сравнения средних в более чем 2 группах необходимо воспользоваться модулем дисперсионного анализа ANOVA. Метод тоже параметрический и требуется выполнение тех же условий, что и для критерия Стьюдента (нормальный закон распределения переменной, гипотеза о равенстве дисперсий).

Для анализа выберем зависимые переменные: сыгранные минуты, трёхочковые броски, двухочковые броски, штрафные броски, а группирующая переменная – национальность.

Изображение выглядит как текст, кроссворд

Автоматически созданное описание

Итоговая таблица средних показывает среднее, количество наблюдений и стандартное отклонение сыгранных минут, трёхочковых бросков, двухочковых бросков и штрафных бросков для представителей каждой национальности. Необходимо проверить верна ли гипотеза о равенстве средних в генеральной совокупности.

Изображение выглядит как стол

Автоматически созданное описание

На вкладке дисперсионный анализ проверяется гипотеза о равенстве дисперсий для сыгранных минут, трёхочковых бросков, двухочковых бросков и штрафных бросков. Как видно из таблицы, уровень значимости p везде больше 0.05, поэтому верна гипотеза о равенстве дисперсий. Поэтому выполняется условие применимости однофакторного дисперсионного анализа.

Также можем рассмотреть критерий Левена и критерий Брауна и Форсайта.

Изображение выглядит как стол

Автоматически созданное описание

Изображение выглядит как стол

Автоматически созданное описание

Опять-таки p больше 0.05 говорим, что гипотеза о равенстве дисперсий верна.

Изображение выглядит как стол

Автоматически созданное описание

Изображение выглядит как стол

Автоматически созданное описание

По национальности USA посмотрим коэффициенты корреляции и уровни значимости. p < 0.05 для всех парамтеров.

Для проверки есть ли статистически значимые отличия средних генеральной совокупности в группирующих переменных возьмем критерий НЗР, Шеффе и Тьюки.

Изображение выглядит как стол

Автоматически созданное описание

Изображение выглядит как текст, кроссворд

Автоматически созданное описание

Изображение выглядит как стол

Автоматически созданное описание

Все критерии Шеффе, Тьюки и НЗР на всех сравнениях выдали значения p > 0.05, поэтому можем сказать, что верна гипотеза о равенстве средних всех генеральных совокупностей. То есть национальность не влияет на основные статистические показатели игрока

Выберем теперь 2 группирующие переменные – национальность и цвет кожи.

Изображение выглядит как стол

Автоматически созданное описание

Итоговая таблица средних содержит информация о среднем, количестве наблюдений и стандартном отклонении для сыгранных минут, трёхочковых бросков, двухочковых бросков и штрафных бросков в зависимости от национальности и цвета кожи.

Изображение выглядит как текст, стол

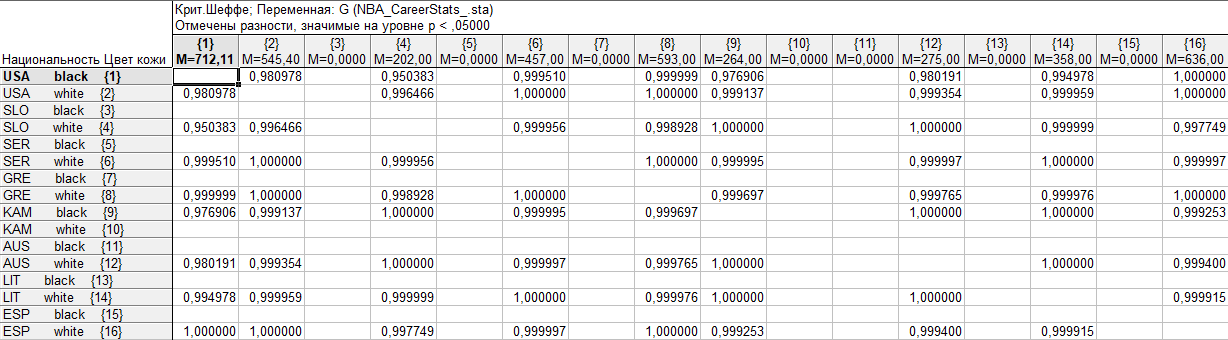
Автоматически созданное описание

Как видно из таблицы дисперсионного анализа, уровень значимости p везде больше 0.05, поэтому верна гипотеза о равенстве дисперсий. Поэтому выполняется условие применимости однофакторного дисперсионного анализа.

Рассмотрим теперь критерии НЗР, Шеффе и Тьюки.

Изображение выглядит как стол

Автоматически созданное описание



Изображение выглядит как текст, снимок экрана, полный, несколько

Автоматически созданное описание

Все критерии Шеффе, Тьюки и НЗР на всех сравнениях выдали значения p > 0.05, поэтому можем сказать, что верна гипотеза о равенстве средних всех генеральных совокупностей.

**5 Дисперсионный анализ**

Дисперсионный анализ является наиболее общим методом сравнения средних. В дисперсионном анализе можно исследовать зависимость количественного признака (зависимой переменной) от одного или нескольких качественных признаков (факторов).

В качестве зависимых переменных выбраны количество подборов и результативных передач, а в качестве категориальных позиция игрока.

Изображение выглядит как стол

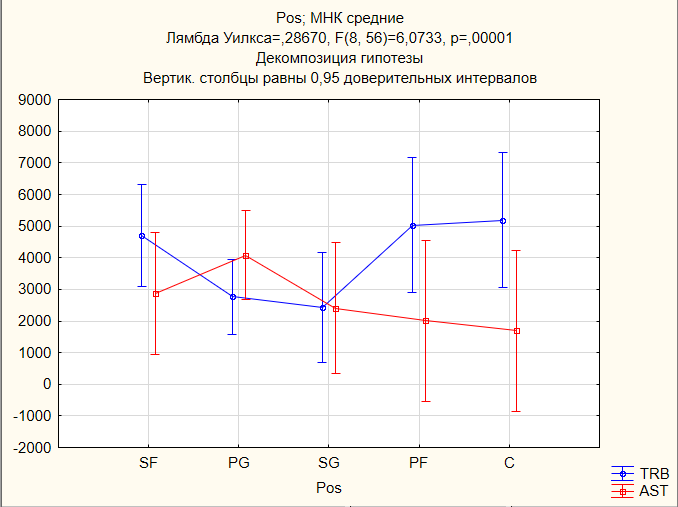
Автоматически созданное описание

В последнем столбце отображены уровни значимости критерия Фишера. Так как уровень значимости p < 0.05, то при разбиении на группы по позиции наверняка будут группы со статистически значимыми отличиями средних.

Изображение выглядит как текст, кроссворд

Автоматически созданное описание

На таблице МНК средние можем увидеть среднее, стандартную ошибку и доверительный интервал для подборов и результативных передач. Также эти данные можем посмотреть на графике.



Для применения дисперсионного анализа необходимо нормальное распределение зависимой переменной и равенство дисперсий. Посмотрим критерий Левена для игровых позиций.

Изображение выглядит как стол

Автоматически созданное описание

Как оказалось гипотеза о равенстве дисперсий верна для подборов и результативных передач. Посмотрим также диаграммы рассеяния.





Точки не находятся на одной прямой, а соответственно средние и стандартные отклонения коррелируют незначительно.

# 6 Непараметрическая статистика

Если условия применения параметрических критериев (нормальность распределения, равенство дисперсий) не выполнены, необходимо воспользоваться непараметрическими критериями.

Коэффициент Спирмена используется, если переменные количественные (закон распределения переменных неизвестен или не является нормальным) и (или) качественные (порядковые). Рассмотрим переменные подборы в нападении, потери, блокшоты и персональные замечания.

Изображение выглядит как стол

Автоматически созданное описание

Корреляция для всех пар переменных умеренная (Спирмен R в промежутке 0.25–0.75), кроме подборов в нападении и персональных замечаний, там сильная >0.75. Так как p<0.05 во всех наблюдениях, то корреляция между переменными статистически значима.

Коэффициент Кендалла используется, если хотя бы одна переменная качественная (порядковая). Рассмотрим возраст игрока и количество набранных триплдаблов.

Изображение выглядит как стол

Автоматически созданное описание

У возраста игрока и числу набранных трипдаблов слабая корреляция (Кендалл тау <0.25). Так как p>0.05, то корреляция между переменными статистически не значима.

Коэффициент Гамма используют, если переменные содержат много повторяющихся значений.

Изображение выглядит как текст, стена, снимок экрана

Автоматически созданное описание

Для возраста игрока и числа набранных трипдаблов корреляция слабая (Гамма <0.25).

Полагая, что зависимы количество реализованных трёхочковых бросков, количество выброшенных трёхочковых бросков и процент реализации трёхочковых бросков, проверим сравнением нескольких зависимых переменных.

Изображение выглядит как стол

Автоматически созданное описание

Коэффициент p < 0.05, значит верна альтернативная гипотеза о неравенстве средних в 3 группах. Посмотрим попарно в каких группах средние неравны.

Изображение выглядит как стол

Автоматически созданное описание

Изображение выглядит как стол

Автоматически созданное описание

Изображение выглядит как стол

Автоматически созданное описание

При попарном сравнении всюду p < 0.05, значит верна альтернативная гипотеза о неравенстве средних.

Теперь проверим, есть ли различие в цвете кожи, в зависимости от подборов, передач и набранных очков. Посмотрим сравнение нескольких независимых групп.

Изображение выглядит как текст, стол

Автоматически созданное описание

Изображение выглядит как текст, кроссворд

Автоматически созданное описание

Из таблиц мы делаем вывод, что верна гипотеза о равенстве средних в 2 группах (уровни значимости p больше 0.05).

Проверим, есть ли различие в возрасте в зависимости цвета кожи. Для этого проведем сравнение 2 независимых групп.

Изображение выглядит как стол

Автоматически созданное описание

Как результат, гипотеза о равенстве средних подтверждается, так как коэффициент p > 0.05.