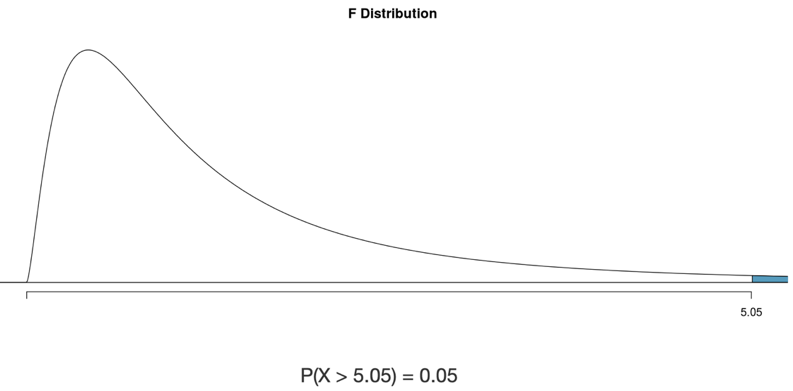
## Дисперсионный анализ

Регрессия и дисперсионный анализ (ANOVA) – два статистических метода, использующих общую линейную модель (GLM), в основе которых лежит предположение о том, что зависимая переменная представляет собой функцию от одной или более независимых переменных.

**Дисперсия** – характеристика рассеивания данных вокруг их среднего значения.

**Дисперсионный анализ (ANOVA) –** статистическая процедура, используемая для сравнения средних значений определенной переменной в двух и более независимых группах.

Основная статистика в дисперсионном анализе – F-отношение, используемое для выявления статистической значимости различий между группами.



F-распределение для числа групп – 5 и числа степеней свободы – 5

В дисперсионном анализе теоретическое распределение F-отношения не является нормальным, оно подчиняется распределению Фишера.

F-значение всегда является положительным, потому вероятность отклонения рассчитывается только в правую сторону.

В дисперсионном анализе рассматривается отношение двух дисперсий: **межгрупповой** и **внутригрупповой**.

**Общая сумма квадратов** SST (общая изменчивость данных) **–** показатель, характеризующий степень изменчивости данных без учета разделения их на группы. Вычисляется общая сумма квадратов следующим образом:

* для каждого наблюдения рассчитывается насколько оно отклонится от среднего значения,
* складывается сумма квадратов полученных отклонений.

Общая сумма квадратов SST получена из двух источников: **межгрупповая сумма квадратов** SSB (характеристика, показывающая насколько групповые средние отклоняются от общего среднего) и **внутригрупповая сумма квадратов** SSW (сумма квадратов отклонений от среднего внутри каждой из групп).

**Межгрупповая дисперсия MSB**, объяснённая влиянием фактора, характеризует рассеивание значений между градациями (группами) вокруг средней всех данных.

**Внутригрупповая дисперсия MSW**, необъяснённая, характеризует рассеивание данных внутри градаций фактора (групп) вокруг средних значений этих групп.

Отношение межгрупповой и внутригрупповой дисперсий – **фактическое отношение Фишера**. Его сравнивают с **критическим значением отношения Фишера**. В случае, когда фактическое отношение Фишера превышает критическое, то средние классов градации различны, а исследуемый фактор оказывает существенное влияние на изменение данных. В обратном случае: средние классов градации друг от друга не отличаются, а фактор не оказывает существенного влияния на изменение данных.

**Целью дисперсионного анализа** является исследование наличия/отсутствия существенного влияния некоторого количественного/качественного фактора на изменения исследуемого признака.

Фактор, предположительно имеющий/не имеющий существенное влияние, делят на группы и на основе исследования значимости средних в наборах данных, соответствующих группам фактора, выясняют одинаково ли влияние фактора.

**Пример 1.** Исследование зависимости прибыли предприятия от типа используемого сырья. В данном случае группы – типы сырья.

**Пример 2.** Исследование зависимости себестоимости выпуска единицы продукции от размера предприятия. Здесь группы – величины предприятий (малое, среднее, большое).

Минимальное число групп в дисперсионном анализе – две. Группы могут быть количественные и качественные.

В дисперсионном анализе вычисляется удельный вес суммарного воздействия одного/нескольких факторов. Насколько влияние фактора существенно, исследуется с помощью гипотез:

*Нулевая гипотеза* H0 утверждает, что все *a* классов градации имеют одинаковые значения средних: *μ*1 = *μ*2 = ... = *μa*.

*Альтернативная гипотеза* H1: не все классы градации имеют одно значение средних.

## Однофакторный дисперсионный анализ

При формировании групп для сравнения в однофакторном дисперсионном анализе используется только одна переменная (фактор).

**Пример.** Исследуется эффективность работы нового станка по обработке металлов с помощью дисперсионного анализа. Сравнение проводится с работой старого станка, который уже используется в производстве. В данном исследовании фактор – используемый станок. У него два уровня: новый, старый станки.

В дисперсионном анализе фактор может иметь более двух уровней.

Однофакторный дисперсионный анализ с двумя уровнями аналогичен t-критерию. Нулевая гипотеза обычно говорит о равенстве средних двух групп, альтернативная – о различии средних (двусторонний тест) или различии в определенном направлении (односторонний тест).

Основные условия проведения дисперсионного анализа:

1. Зависимая переменная должна быть непрерывной, неограниченной/изменяющейся в широком интервале и представлена интервальными/характеризующими отношения данными; факторы должны быть дихотомическими/категориальными.
2. Каждое значение зависимой переменной не должно зависеть от других ее значений.

Исключения: рассматривается временная зависимость или значения были измерены у объектов, которые объединены в группы (члены одной семьи, учащиеся в одном классе) и это повлияло на зависимую переменную.

1. В каждой группе непрерывная переменная имеет приблизительно нормальное распределение. Нормальность распределения можно проверить, используя гистограмму («на глаз») или статистические тесты на нормальность.
2. Дисперсии изучаемых групп должны быть приблизительно одинаковыми. Проверить похожесть дисперсий можно с помощью теста Левина, в котором нулевая гипотеза гласит, что дисперсия однородна, и если результат теста Левина статистически не значим (при применении критерия α < 0,05), то дисперсии достаточно похожи.

Некоторые условия проведения дисперсионного анализа могут нарушаться, например, F-статистика надежна в случае, когда распределение непрерывной переменной отлично от нормального, а размеры групп одинаковы. Одинаковый размер обеспечивает и устойчивость F-статистики к нарушениям однородности дисперсии. А нарушение условия независимости может сильно исказить результаты.

***Однофакторный дисперсионный анализ*** основан на том, что общая сумма квадратов SST получена из двух компонент: межгрупповой суммы квадратов SSB и внутригрупповой суммы квадратов SSW:

*SST*= *SSB* + *SSW*

**Пример 3.**

|  |  |  |
| --- | --- | --- |
| **Группа 1** | **Группа 2** | **Группа 3** |
| 1 | 3 | 5 |
| 2 | 4 | 6 |
| 3 | 5 | 7 |

Сравниваем 3 группы, в каждой из которых по 3 значения.

Нулевая гипотеза: в генеральной совокупности нет значимых различий между средними, все средние трёх групп равны друг другу. Альтернативная гипотеза: хотя бы пара средних значимо различается между собой.

**H0**:

**H1**: или или

Вычислим среднее значение всех наблюдений:

Вычислим общую сумму квадратов:

Степени свободы для общей суммы квадратов:

Вычислим средние значения внутри каждой из групп:

Внутригрупповая сумма квадратов:

Степени свободы для внутригрупповой суммы квадратов:

Межгрупповая сумма квадратов:

Степени свободы для межгрупповой суммы квадратов:

SST =30

SSB=24 SSW=6

Получили, что большая часть общей изменчивости обеспечивается благодаря межгрупповой сумме квадратов, значит группы значительно различаются между собой.

Межгрупповая дисперсия:

Внутригрупповая дисперсия:

Вычислим F-значение:

Критическое значение отношения Фишера:

F0,05; 2;6 =5,14

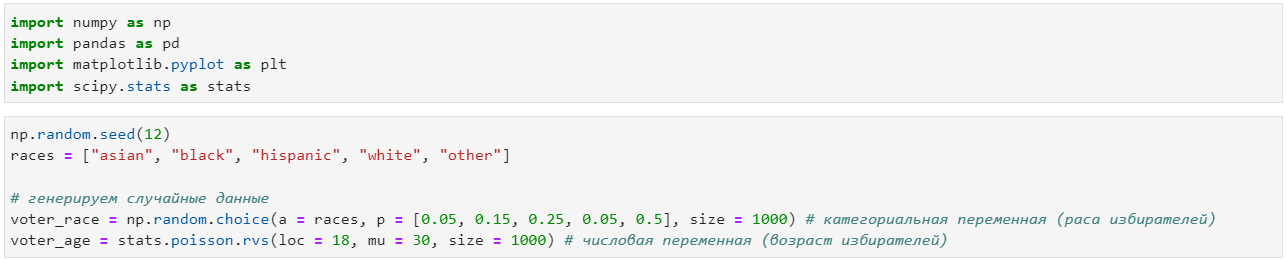
Так как фактическое отношение Фишера меньше критического:

можно сделать вывод, что есть существенные различия между группами.

**Пример 4.**

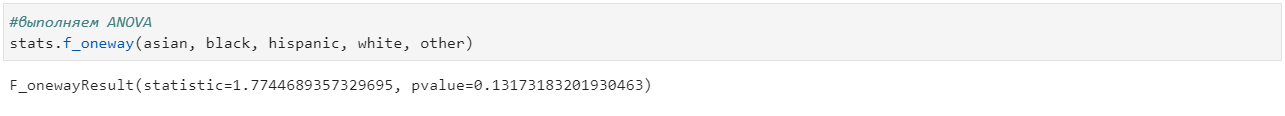
Мы хотим проверить, отличается ли возраст избирателей на основе какой-либо категориальной переменной, например от расы избирателя. Для этого сгенерируем данные с различными параметрами, что позволит продемонстрировать выполнения дисперсионного анализа в Python.

Сгенерируем выборку из 1000 элементов, в которую включены следующие расы: asian, black, hispanic, white, other и возраста избирателей.

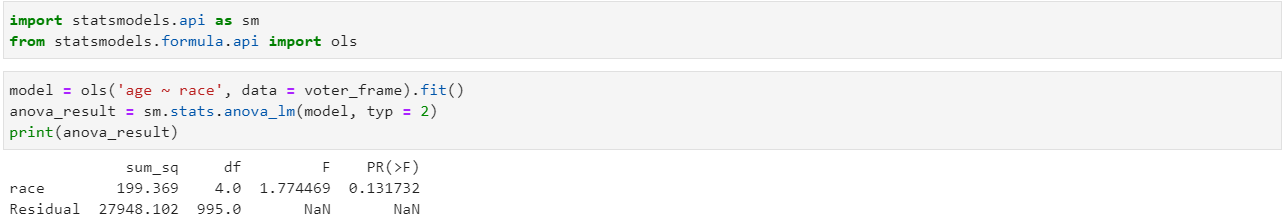
****

Так как все возраста генерируются одинаково, то это говорит нам о том, что они все из одной генеральной совокупности, поэтому ANOVA должна дать результат, что существенной разницы нет.



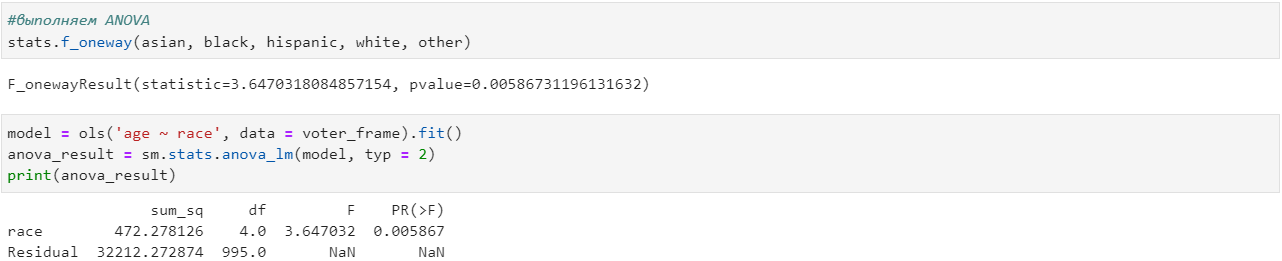


Рассмотрим альтернативный способ, используем функцию anova\_lm() из библиотеки statsmodels:



Попробуем сгенерировать чуть измененные данные. Сгенерируем возраста для белых людей отдельно. В качестве среднего возраста возьмем 32 года. Это изменение должно отразиться на результатах ANOVA. Он должен показать различность возрастов белых людей и остальных.





ANOVA нашел различие, поскольку p-значение меньше 0,05. Это означает, что фактор раса оказывает статистически значимое влияние на возраст избирателей, но было бы интересно узнать в каких именно группах есть влияние. Для этого необходимо вернуться на шаг назад. Можно использовать t критерий Стьюдента для всех пар рас, но такой метод при большом разнообразии групп может дать слишком большую ошибку.

Метод Бонферрони является одним из наиболее простых и известных способов контроля над групповой вероятностью ошибки.

Предположим, что мы применили определенный статистический критерий 3 раза (например, сравнили при помощи критерия Стьюдента средние значения групп А и В, А и С, и В и С) и получили следующие три Р-значения: 0.01, 0.02 и 0.005. Если мы хотим, чтобы групповая вероятность ошибки при этом не превышала определенный уровень значимости α = 0.05, то, согласно методу Бонферрони, мы должны сравнить каждое из полученных Р-значений не с α, а с , где m – число проверяемых гипотез. Деление исходного уровня значимости α на m – это и есть поправка Бонферрони. В рассматриваемом примере каждое из полученных Р-значений необходимо было бы сравнить с . В результате мы выяснили бы, что Р-значение для второй гипотезы (0.02) превышает 0.017 и, соответственно, у нас не было бы оснований отвергнуть эту гипотезу.

Вместо деления изначально принятого уровня значимости на число проверяемых гипотез, мы могли бы умножить каждое из исходных Р-значений на это число. Сравнив такие скорректированные Р-значения (англ. adjusted P-values; обычно обозначаются буквой q) с α, мы пришли бы к точно тем же выводам, что и при использовании поправки Бонферрони.

* 0.01\*3 = 0.03 < 0.05: гипотеза отклоняется;
* 0.02\*3 = 0.06 > 0.05: гипотеза принимается;
* 0.005\*3 = 0.015 < 0.05: гипотеза отклоняется.

Вернемся к нашему примеру с избирателями. Выполним попарные сравнения.



Мы имеем 10 сравнений, поэтому m = 10. Поэтому pvalue необходимо уменьшить в 10 раз. То есть можно сказать, что критическое значение α у нас становится 0.005.

Сделаем выводы.

|  |  |  |
| --- | --- | --- |
| Группы | p-значение | гипотеза |
| asian - black | 0.4027281369339345 | принимается |
| asian - hispanic | 0.6704669004240726 | принимается |
| asian - white | 0.027828801627453537 | принимается |
| asian - other | 0.712474249112879 | принимается |
| black - hispanic | 0.05156197171952594 | принимается |
| black - white | 0.0006893463707824467 | отклоняется |
| black - other | 0.3555931499524523 | принимается |
| hispanic - white | 0.033930889763891824 | принимается |
| hispanic - other | 0.10037925272137736 | принимается |
| white - other | 0.0010062493632570478 | отклоняется |

Видим, что различия в возрастах есть у следующих пар рас: black-white и white-other. Так же у asian-white и hispanic-white p-значение достаточно небольшое, но всё же больше 0.005. Эти результаты говорят о том что возраст светлокожих имеет отличие от остальных, что собственно мы и реализовали при генерации данных.

Также для такого рода анализа можно использовать пост-хок тесты.

## Пост-хок тесты

После проведения дисперсионного анализа получаем данные о том, значимо ли влияние изучаемого фактора на данные: различаются ли между группами средние значения зависимой переменной. Однако результаты анализа не дают ответа на вопрос: благодаря каким различиям это влияние оказалось значимым?

Для решения данной задачи предназначены пост-хок тесты.

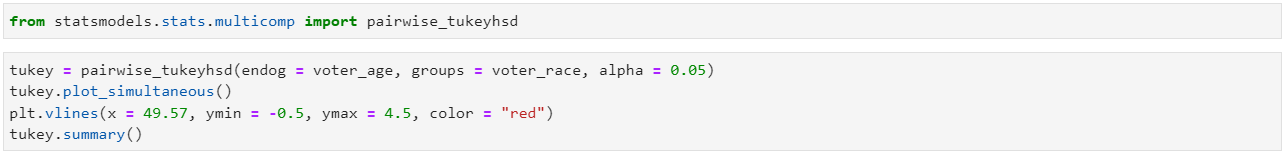
**Свойства пост-хок тестов**

* post-hoc тесты применяются когда влияние фактора значимо;
* тесты делают поправку для снижения вероятности ошибки I рода;
* они учитывают величину различий между средними значениями и количество сравниваемых между собой пар;
* тесты отличаются по степени консервативности (разумный компромисс – пост-хок тест Тьюки).

**Пост-хок тест Тьюки**

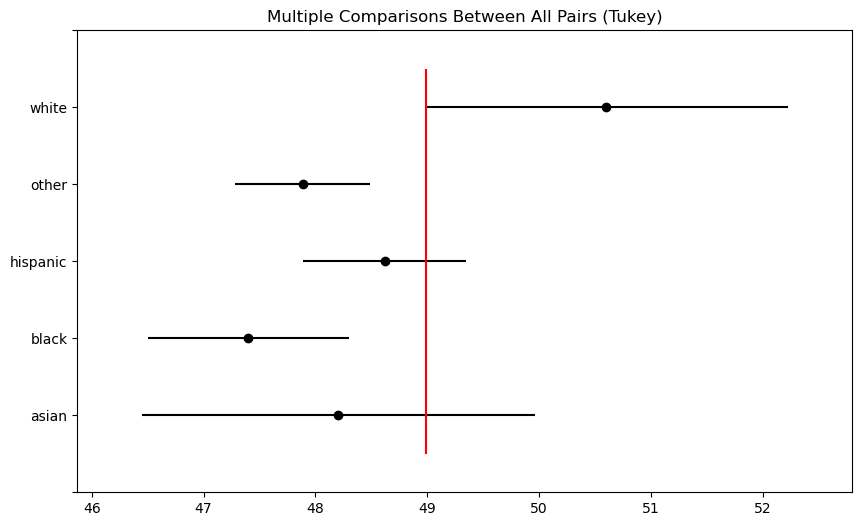
* строго контролирует значимость критерия α (0.05)
* одновременно проверяет все парные гипотезы;
* чувствителен к неравенству дисперсий;
* если размер групп имеет сильные различия, работает плохо.

Выполним пост-хок тест Тьюки для нашего примера с данными об избирателях, а также построим график с доверительными интервалами.

****

****

Заметим, что, как и в прошлом случае различие в возрастах есть у black-white и other-white. И можно сделать аналогичные выводы. Посмотрим на график со средними значениями и их доверительными интервалами.

****

Видим, что доверительные интервалы white-hispanic и white-asian перекрываются, поэтому пост-хок тесты показали что различия между ними не существенные.

## Двухфакторный дисперсионный анализ без повторений

Двухфакторный дисперсионный анализ применяется для проверки возможной зависимости результативного признака от двух факторов.

Пусть *m* – число градаций первого фактора и k – число градаций второго фактора.

Двухфакторный дисперсионный анализ основан на том, что общая сумма квадратов SST получена из трёх компонент: объяснённой влиянием фактора *A* суммы квадратов отклонений *SSBA*, объяснённой влиянием фактора *B* суммы квадратов отклонений *SSBB* и необъяснённой суммы квадратов отклонений (суммы квадратов отклонений ошибки):

*SST*= *SSBA+SSBB* + *SSW*

Рассмотрим следующий пример:

Ботаник хочет знать, влияет ли на рост растений воздействие солнечного света и частота полива. Она сажает 30 семян и позволяет им расти в течение двух месяцев при различных условиях солнечного света и частоты полива. Через два месяца она записывает высоту каждого растения в дюймах.

Используйте следующие шаги, чтобы выполнить двусторонний дисперсионный анализ, чтобы определить, оказывают ли частота полива и воздействие солнечного света существенное влияние на рост растений, а также определить, есть ли какой-либо эффект взаимодействия между частотой полива и воздействием солнечного света.

* вода: как часто поливалось каждое растение: ежедневно или еженедельно
* солнце: сколько солнечного света получило каждое растение: низкое, среднее или высокое
* высота: высота каждого растения (в дюймах) через два месяца





Мы можем видеть следующие p-значения для каждого из факторов в таблице:

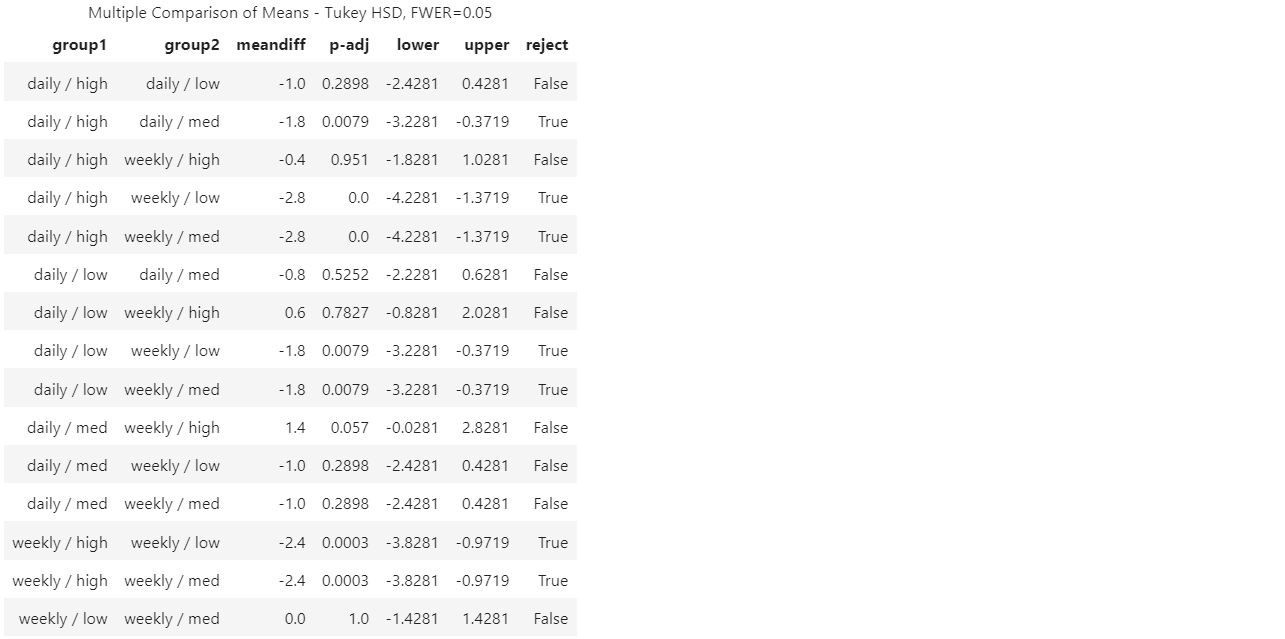
* вода: p-значение = 0,000527
* солнце: p-значение = 0,0000002
* вода \* солнце: p-значение = 0,120667

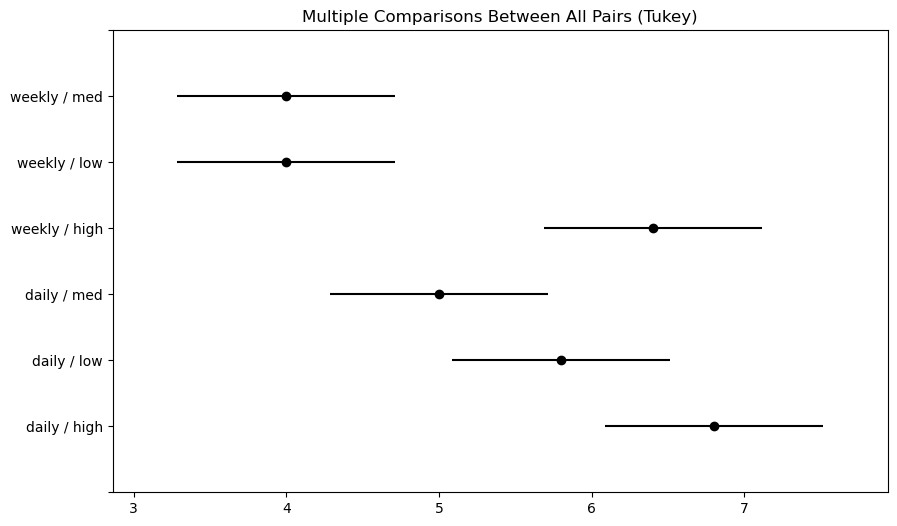
Поскольку p-значения для воды и солнца меньше 0,05, это означает, что оба фактора оказывают статистически значимое влияние на высоту растений.

А поскольку p-значение для эффекта взаимодействия (0,120667) составляет не менее 0,05, это говорит нам об отсутствии значительного эффекта взаимодействия между воздействием солнечного света и частотой полива.

Выполним пост-хок тест Тьюки для нашего примера, а также построим график с доверительными интервалами.







## Двухфакторный дисперсионный анализ с повторениями

Двухфакторный дисперсионный анализ с повторениями применяется для проверки не только возможной зависимости результативного признака от двух факторов – *A* и *B*, но и возможного взаимодействия факторов *A* и *B*.

Пусть *m* – число градаций фактора *A*,  *k* – число градаций фактора *B*, *r* – число повторений.

В данном статистическом комплексе общая сумма квадратов SST получена из четырех компонент:

*SST*= *SSBA+SSBB* + *SSBAB* +*SSW*

## Практическая работа

1. Загрузить данные: 'insurance.csv'. Вывести и провести предобработку. Вывести список уникальных регионов.
2. Выполнить однофакторный ANOVA тест, чтобы проверить влияние региона на индекс массы тела (BMI), используя первый способ, через библиотеку Scipy.
3. Выполнить однофакторный ANOVA тест, чтобы проверить влияние региона на индекс массы тела (BMI), используя второй способ, с помощью функции anova\_lm() из библиотеки statsmodels.
4. С помощью t критерия Стьюдента перебрать все пары. Определить поправку Бонферрони. Сделать выводы.
5. Выполнить пост-хок тесты Тьюки и построить график.
6. Выполнить двухфакторный ANOVA тест, чтобы проверить влияние региона и пола на индекс массы тела (BMI), используя функцию anova\_lm() из библиотеки statsmodels.
7. Выполнить пост-хок тесты Тьюки и построить график.
8. Оформить отчет о проделанной работе, написать выводы.