**Лекция №10 Дисперсионный анализ**

**Цель лекции:**

* Научиться выполнять однофакторный дисперсионный анализ
* Выполнить post hoc тест
* Провести двухфакторный дисперсионный анализ
* Познакомиться с условиями применимости дисперсионного анализа

**Материал прошлого урока:**

На прошлых вебинарах мы рассмотрели корреляционный и регрессионный анализы, с помощью которых мы научились выявлять линейную зависимость и определять, как быстро изменяется зависимая переменная y при изменении независимой величины X. Сегодня мы научимся отвечать на вопрос, а влияет ли качественный признак на количественный признак. И поможет в этом дисперсионный анализ.

**План урока:**

1. Однофакторный дисперсионный анализ
2. Post hoc test
3. Двухфакторный дисперсионный анализ
4. Двухфакторный дисперсионный анализ в Python
5. Условия применимости дисперсионного анализа

**Однофакторный дисперсионный анализ**

Когда нам нужно понять, а влияет ли качественный показатель на количественный показатель, мы прибегаем к дисперсионному анализу. Т.е. зависимая переменная должна быть количественной случайной величиной, а (фактор) - качественной случайной величиной.

Рассмотрим идею применения дисперсионного анализа сначала на однофакторном дисперсионном анализе. Кстати, иногда вы можете встретить название ANOVA (от английского Analysis of variance, что в переводе значит анализ дисперсий).

В однофакторном дисперсионном анализе на одну количественную переменную влияет один фактор (один качественный показатель), который наблюдается на уровнях или имеет выборок для переменной . Поясним это. Предположим, есть сообщество людей, среди которых есть представители трех разных профессий: юристы, программисты и бухгалтера. Нам известна заработная плата каждого человека из этого сообщества. Т.о. мы имеем количественную переменную – заработная плата.

*юристы*

*бухгалтера*

*программисты*

И мы хотим понять, а влияет ли профессия (качественный показатель) на заработную плату (количественную переменную). Т.е. профессия наблюдается на 3-х уровнях или есть 3 выборки для переменной : зарплаты юристов, зарплаты программистов и зарплаты бухгалтеров.

**Каким образом мы можем понять, оказывает ли профессия статистически значимый эффект на заработную плату?**

Мы можем сделать следующее. Посчитаем средние заработные платы по каждой выборке и сравним их средние между собой. И если мы обнаружим статистически значимые различия хотя бы между одной парой, то приходим к выводу, что профессия оказывает статистически значимый эффект на заработную плату.

Мы научились сравнивать два средних арифметических, используя параметрические критерии Z и t , когда изучали тестирование гипотез. Здесь тоже нам придется сравнивать средние арифметические, но мы имеем дело с несколькими средними, когда их число больше двух. И здесь появляется неприятный момент: эффект множественных сравнений. Дело в том, что с увеличением числа сравнений растет вероятность ошибки 1 рода, т.е. растет вероятность найти различия там, где их нет, иными словами, растет вероятность принять гипотезу , когда на самом деле верна гипотеза . И вместо привычного, заранее выбранного уровня статистической значимости α, появляется которая является истинным уровнем значимости многократно примененного критерия. И вычисляется по формуле

Если есть 2 средних арифметических, то мы имеем одно парное сравнение

, ,

Если три средних арифметических, то число парных сравнений равно 3

, и т.д.

Как вы видите, уровень статистической значимости растет, а и есть вероятность ошибки 1 рода.

Чтобы этого избежать, лучше сразу использовать дисперсионный анализ. По сути, это то же самое тестирование гипотезы, где нужно следовать определенному алгоритму. И мы с ним уже знакомы.

Сразу будем разбирать дисперсионный анализ на конкретной задаче. У нас даны заработные платы бухгалтеров, юристов и программистов. Рекомендуется брать выборки одинакового объема, поясним это, когда будем обсуждать условия применимости.

Бухгалтера: 70, 50, 65, 60, 75, 67, 74

Юристы: 80, 74, 90, 70, 75, 65, 85

Программисты: 148, 142, 140, 150, 160, 170, 155

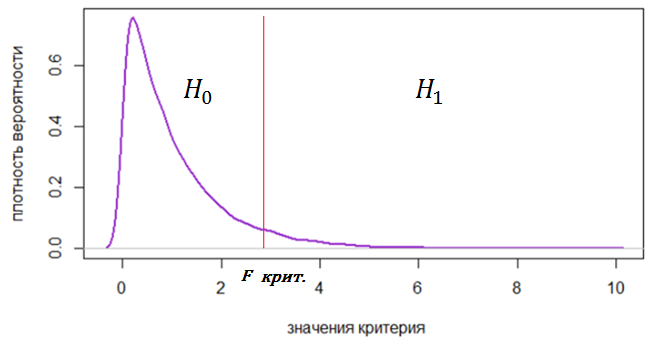
1. Формулируем нулевую и альтернативную гипотезы.

Нулевая гипотеза

Альтернативная гипотеза состоит из нескольких частей. Сравнивать будем попарно средние арифметические.

Дисперсионный анализ не даст ответа на вопрос, между какими именно двумя средними арифметическими найдены статистически значимые различия. Он лишь ответит на вопрос, влияет ли профессия на заработную плату или не влияет. Дисперсионный анализ покажет влияние фактора в случае, если хотя бы между одной парой средних арифметических будут найдены статистически значимые различия и покажет, что не влияет, если будут отвергнуты все три маленькие альтернативные гипотезы .

1. Теперь заранее устанавливаем уровень статистической значимости . Мы выберем, как обычно, 0,05
2. На этом этапе тестирования гипотезы идет выбор критерия. В дисперсионном анализе используется критерий Фишера. Мы уже с ним немного познакомились, когда применяли его для проверки статистической значимости подобранной регрессионной модели. Напомню, как выглядит распределение Фишера:



Нам нужно будет найти расчетное значение критерия и сравнить с табличным значением.

1. А на этом этапе мы делаем вывод.

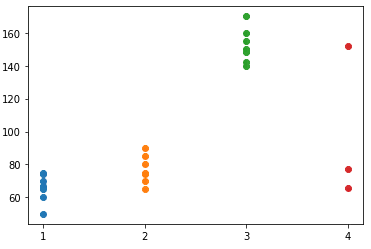
С критерием мы определились, уровень статистической значимости выбрали. Давайте теперь найдем расчетное значение критерия Фишера и табличное значение критерия Фишера. И уже исходя из этого, сделаем вывод.

Критерий Фишера находится по формуле:

,

где – факторная дисперсия, а – остаточная дисперсия.

Чтобы эти понятия не были абстрактными для вас, давайте посмотрим на рисунок ниже.

**

На рисунке выше показаны заработные платы 3х групп (синие, желтые, зеленые точки). В 4-ом столбике три красные точки – это средние заработные платы для каждой группы. Мы видим разброс среди этих трех точек. Это и есть факторная дисперсия. Ее ещё называют объясненной, ведь мы ее объясняем влиянием фактора. Т.е. различия между средними арифметическими заработных плат объясняем тем, что эти средние посчитаны для разных профессий, т.е. профессия оказывает эффект. Есть еще один вариант названия этой дисперсии – межгрупповая.

А теперь посмотрим на каждую отдельную группу. В каждой из них тоже есть разброс. Это внутригрупповая дисперсия, или необъясненная, или остаточная дисперсия. Т.е. осталось необъясненной влиянием фактора.

Разобравшись с дисперсиями, можно приступать к расчету наблюдаемого критерия Фишера *F*. Величины (факторная и остаточная дисперсии соответственно), мы найдем из формул:

, где – число групп

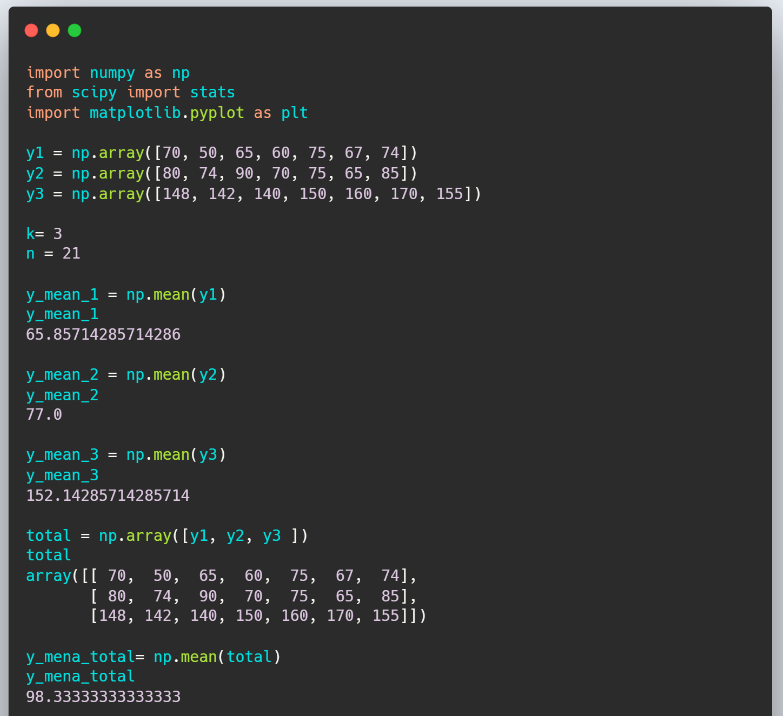
, где - это общее количество значений заработных плат во всех трех подгруппах (7\*3 = 21)

- факторная сумма квадратов отклонений средних групповых значений от общего среднего, - объем выборки.

*-* остаточная сумма квадратов отклонений значений от группового среднего.

Сейчас мы посчитаем наблюдаемый критерий Фишера, а затем применим готовую функцию для расчета этого критерия. Так мы поймем, какие вычисления спрятаны внутри функции и заодно убедимся в правильности наших расчетов, ведь значения критерия должны совпасть.

Создадим выборки зарплат бухгалтеров, юристов, программистов в Python и посчитаем средние арифметические для каждой выборки и среднее по всем значениям заработных плат, а всего значений 21.

**

Теперь рассчитаем сумму квадратов отклонений от общего среднего.

32400 . Ниже эти расчеты будут приведены в Python.

*-* значение заработной платы

Затем рассчитаем факторную сумму квадратов отклонений

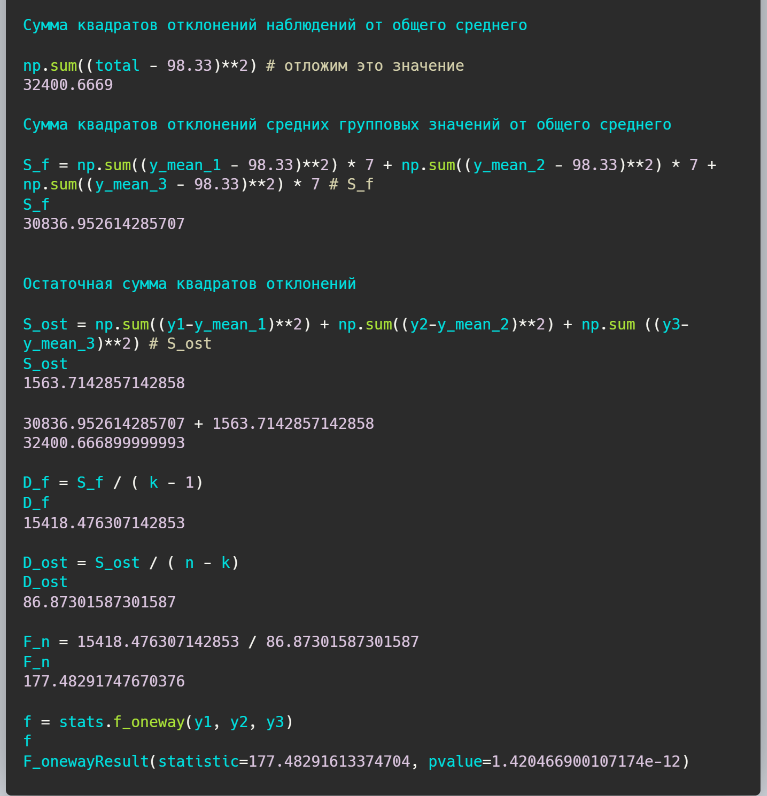
30836. 95

*-* среднее группы

и рассчитаем остаточную сумму квадратов отклонений значений от среднего группового

.

А теперь заметьте, что т.е.

Затем рассчитаем факторную дисперсию (в Python обозначим D\_f) и остаточную дисперсию (в Python - D\_ost). Зная их, получим наблюдаемый критерий Фишера 177,48. После расчетов воспользуемся готовой функцией для выполнения однофакторного дисперсионного анализа stats.f\_oneway() и убедимся, что расчетный критерий Фишера имеет тоже самое значение, что и при ручных расчетах. В работе пользуйтесь функцией, не надо считать вручную. **

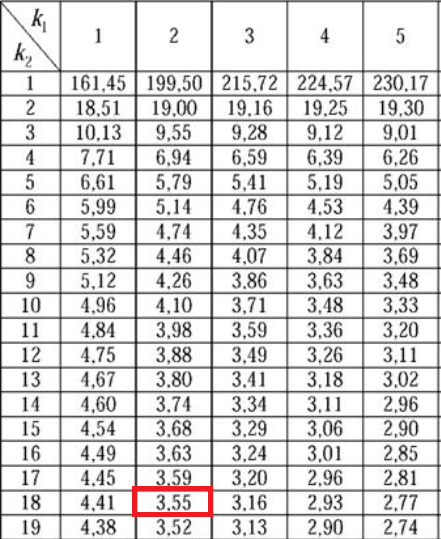
Теперь, чтобы сделать вывод, о том, оказывает ли профессия эффект на заработную плату, найдем по таблице Фишера табличное значение критерия F. Этот критерий зависит от α и степеней свободы. Уровень статистической значимости мы выбрали заранее α = 0.05 , а степени свободы рассчитаем ниже и получим 2 и 18.

Рассчитали эти значения следующим образом:

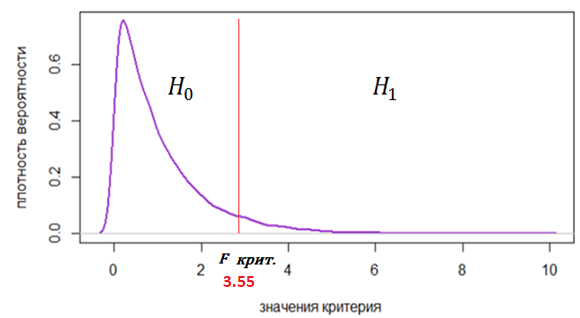
– степени свободы числителя и

– степени свободы знаменателя.

− число групп, n –общее число зарплат



Т. о. табличное значение критерия Фишера равно 3.55. Отметим его на схематично изображенном распределении Фишера.



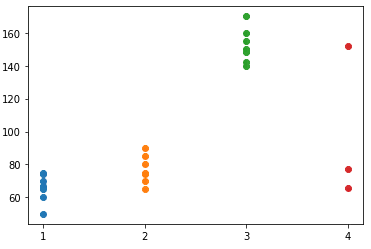
Расчетное значение 177,48 попадает в область принятия альтернативной гипотезы . Т.е. мы делаем вывод, что профессия оказывает статистически значимый эффект на заработную плату.

Вы уже знаете, что функция не показывает табличное значение критерия, а показывает pvalue, на графике – это та площадь, которая лежит за наблюдаемым (расчетным) критерием. Мы видим, что функция показала pvalue . Это значение намного меньше, чем 0.05. А когда pvalue < , то делаем вывод в пользу альтернативной гипотезы.

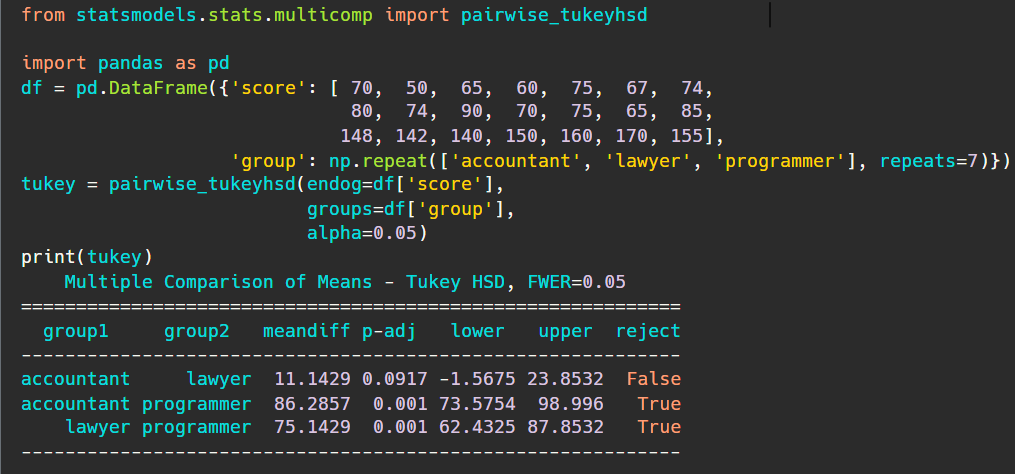
**Post hoc test**

Но, как упоминалось выше, дисперсионный анализ не покажет, между какими именно средними были найдены различия. Если по каким-то причинам нам нужно выяснить, а между какими именно группами найдены статистически значимые различия, то воспользуемся post hoc тестом.

Если мы предварительно сделаем небольшой разведочный анализ (например, график выше с синими, зелеными, желтыми и красными точками), то увидим, что средняя заработная плата программистов лежит намного выше, чем средняя зарплата юристов и бухгалтеров. Т.е. мы можем предположить, что между зарплатами программистов и юристов и зарплатами программистов и бухгалтеров есть статистически значимые различия. А вот между заработными платами юристов и бухгалтеров - нет, т.к. на графике нет такой большой разницы.



Чтобы подтвердить или опровергнуть наше предположение, воспользуемся post hoc test’ом Тьюки. Проведем его сразу в Python. Для этого сначала создадим датафрейм df из заработных плат, а затем применим функцию . Выведем результат с помощью функции



Давайте проанализируем результат. Для этого смотрим на столбец p-adj и сравниваем с уровнем статистической значимости, что мы заранее установили 0.05. Те значения, которые меньше 0.05 говорят нам о наличии статистически значимых различий. Мы видим, что между бухгалтерами (accountant) и программистами (programmer), а также между юристами (lawer) и программистами (programmer) обнаружены статистически значимые различия, поскольку p-adj = 0.001 и это меньше 0.05. За счет этих результатов дисперсионный анализ и сообщил нам о статистически значимом эффекте профессии на заработную плату.

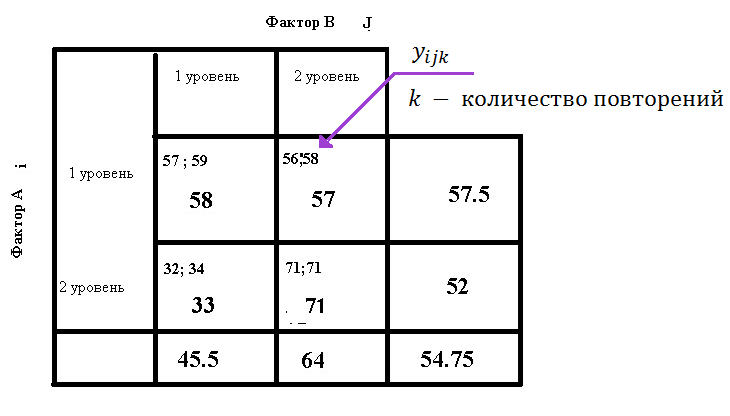
**Двухфакторный дисперсионный анализ**

В двухфакторном дисперсионном анализе на одну количественную переменную Y влияют два фактора (два качественных показателя), наблюдаемых соответственно на и уровнях. Разберем двухфакторный анализ на следующем примере.

Предположим, есть компания, которая выпускает кондиционеры. В них есть две детали, которые конструктора определили ключевыми факторами, влияющими на производительности продукции. Предположим, мы будем измерять количество квадратных метров, охлажденных за какой-то определенный промежуток времени. Первая и вторая деталь - это фактор A и фактор B соответственно. Те размеры деталей, с которыми сейчас изготавливают кондиционер, называются первым уровнем, а следующий размер для каждой детали – это 2й уровень. Для простоты мы рассмотрим дизайн эксперимента 2 Х 2. Есть 2 детали и для каждой из них есть 2 уровня. Мы будем комбинировать эти уровни и замерять производительность. Для каждой комбинации проводим эксперимент не менее 2х раз.

Первая комбинация – это 1й уровень детали A и 1й уровень детали B. Помним, что измерения проводятся минимум 2 раза. Мы для каждой комбинации проведем опыт по 2 раза. Первый опыт при уровнях 1-1 дал значение 57 , а при повторном измерении на этих же уровнях 59 . Среднее арифметическое для них 58.

Давайте оформим эти комбинации уровней и значений производительности в виде таблице. Значения 57, 59 и их среднее 58 – это 1я ячейка со значениями. Значения 57.5 и 52 – это средние по срокам, а 45.5 и 64 – это средние по столбцам. 54.75 – это среднее по всем значениям.



Мы каждое значение представим в виде математической модели, где величина складывается из какой- то истинной величины M, на которую оказывает влияние фактор А ( эффект влияния фактора А – это ), фактор B (эффект ), взаимодействие факторов () и случайная ошибка ().

Т.е. – это уровень фактора , – уровень фактора , принимает значения 1 или 2 в нашем примере – номер измерения на каждой комбинации уровней.

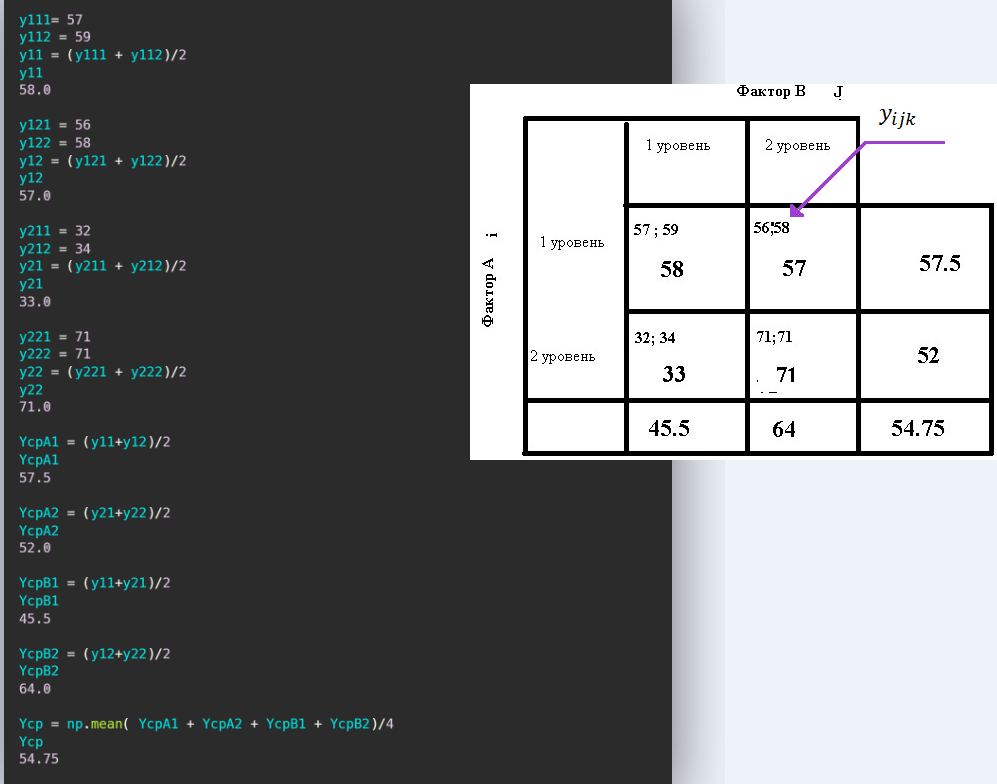
= M +

Если мы вычтем из истинную величину M, то справа останутся влияния факторов и случайной ошибки.

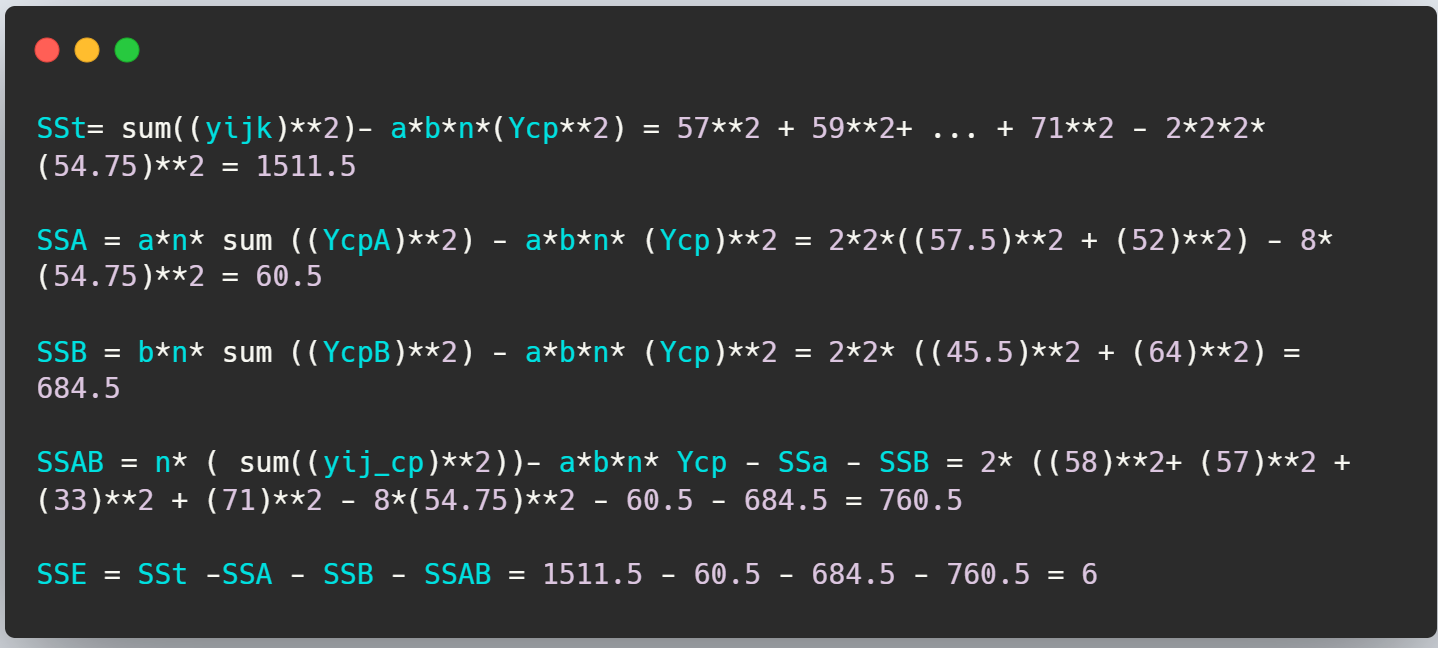
Не буду утомлять вас сложными математическими выкладками, но в итоге мы получаем, что общая сумма квадратов отклонений равна сумме квадратов отклонений фактора A , фактора B , взаимодействия факторов и случайной ошибки .

По сути, здесь тоже все сводится к сравнению табличного критерия и расчетного критерия Фишера. Давайте сначала посмотрим, как это все считается вручную, чтобы было понятно, какие вычисления спрятаны внутри функции, ну а потом все посчитаем с помощью функций и сравним результаты. Они должны быть одинаковыми.

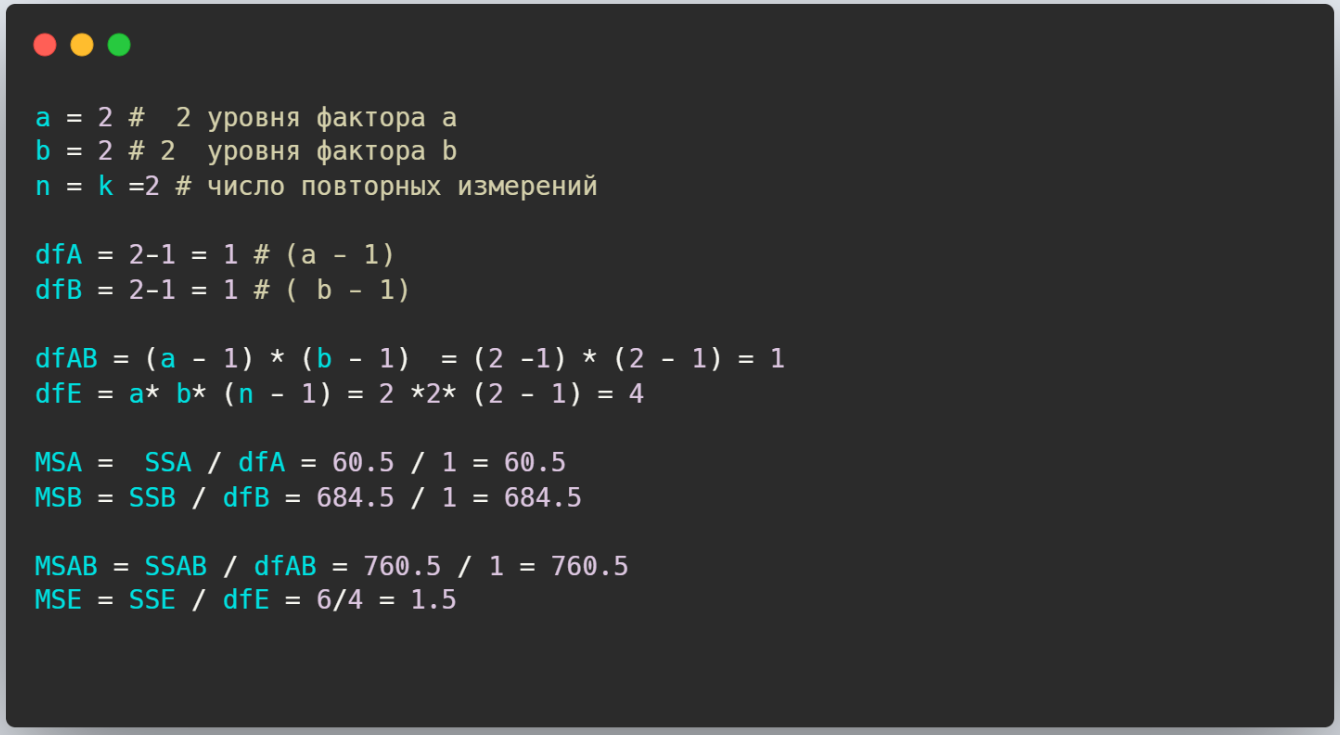
Представим данные из таблицы выше в Python, потому что расчеты будут достаточно объемными.



Теперь будем делать расчеты и заносить их результаты в ANOVA таблицу.



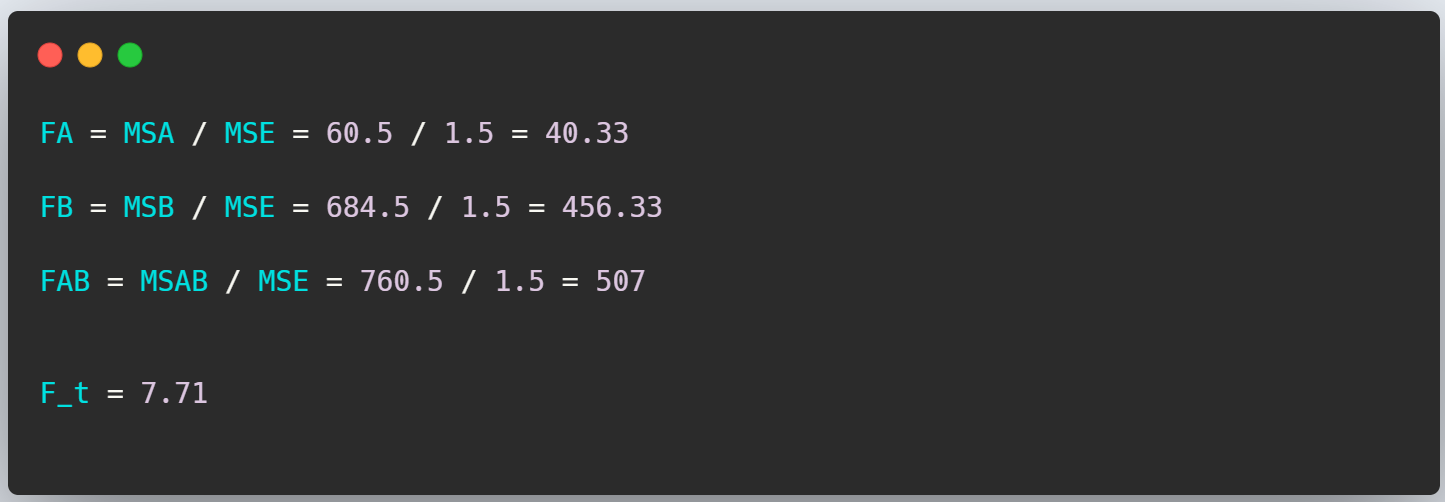
Рассчитаем теперь степени свободыи сумму квадратов отклонений в расчете на одну степень свободы.



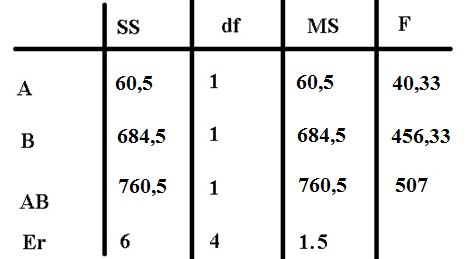
Теперь осталось рассчитать для каждого фактора и взаимодействия факторов критерий Фишера. Расчеты критериев мы сделаем ниже. Значения получим следующие И также для каждого нужно найти табличный критерий Фишера для α = 0.05 и степеней свободы 1 и 4. Как видите, степени свободы совпадают для факторов и их взаимодействия (ниже разберем, откуда 1 и 4), но так происходит именно в нашем случае, т.к. эксперимент 2 Х 2.

**Откуда 1 и 4 степеней свободы?** Например, критерий Фишера для фактора рассчитываем следующим образом:

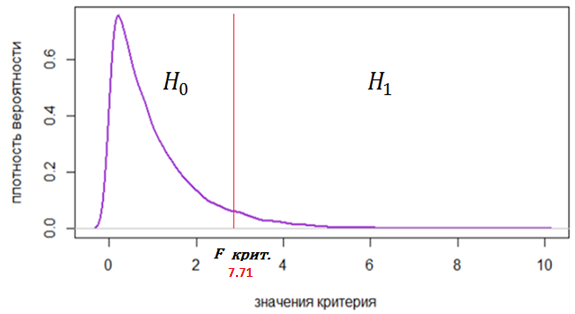
. Когда мы искали – мы делили на 1 степень свободы . 1 – это степень свободы числителя и в таблице находятся в горизонтальной строке. А, когда находили делили на 4 степени свободы. 4 –это степени свободы знаменателя и находятся в вертикальной строке таблицы. По тому же принципу степени свободы критерия Фишера для фактора B и взаимодействия факторов будут 1 и 4. Т.о. в нашем примере для всех факторов и их взаимодействия равен 7.71



Теперь строим ANOVA таблицу, где последний столбец - это расчетный критерий Фишера.



Исходя из полученных значений, мы видим, что мы нашли статистически значимый эффект взаимодействия факторов и также отдельные факторы оказывают статистически значимый эффект. Потому что все три критерия Фишера упали дальше 7.71, т.е. попали в область принятия альтернативной гипотезы.



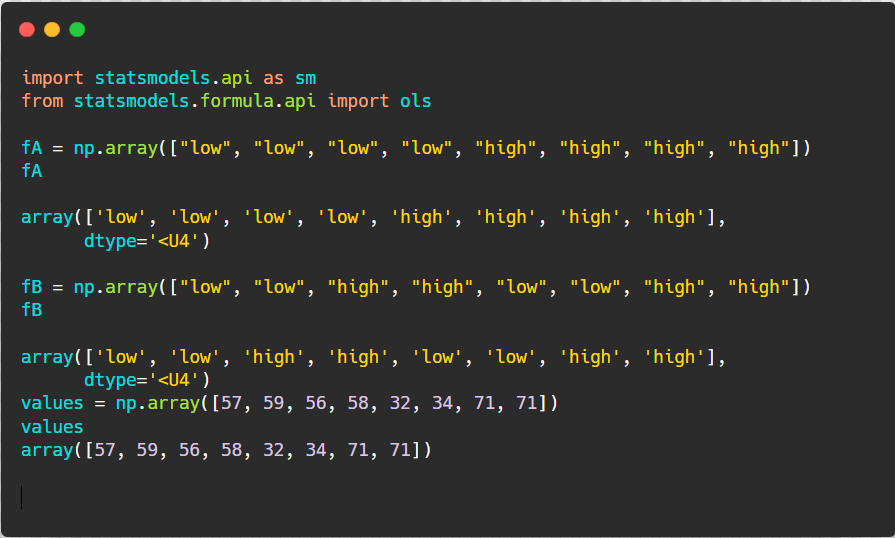
А теперь выполним те же действия, но в Python.

**Двухфакторный дисперсионный анализ в Python**

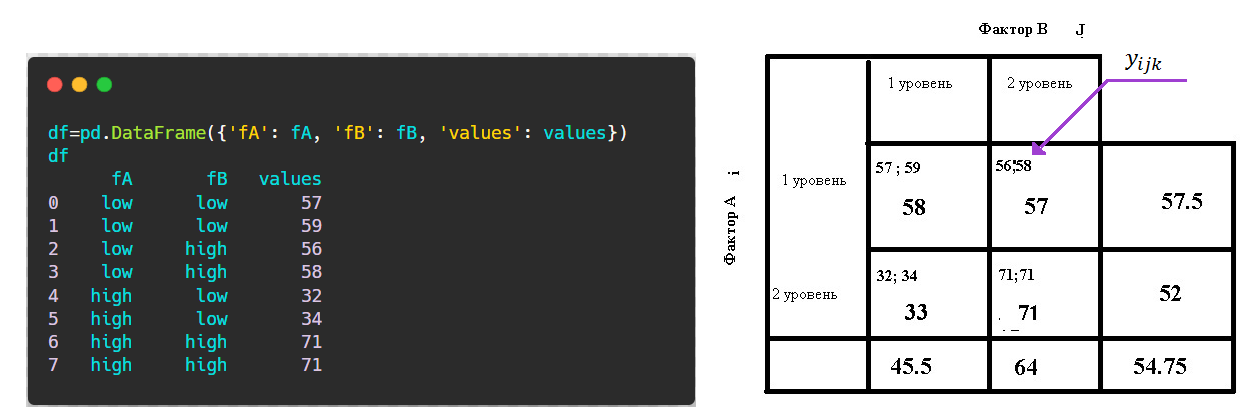
Для понимания того, что мы делаем, я буду выводить рядом две картинки: ту, которую мы рисовали, делая расчеты вручную и ту, которую мы получим в Python.

Сейчас мы создадим уровни фактора A, уровни фактора B –массивы fA и fB.

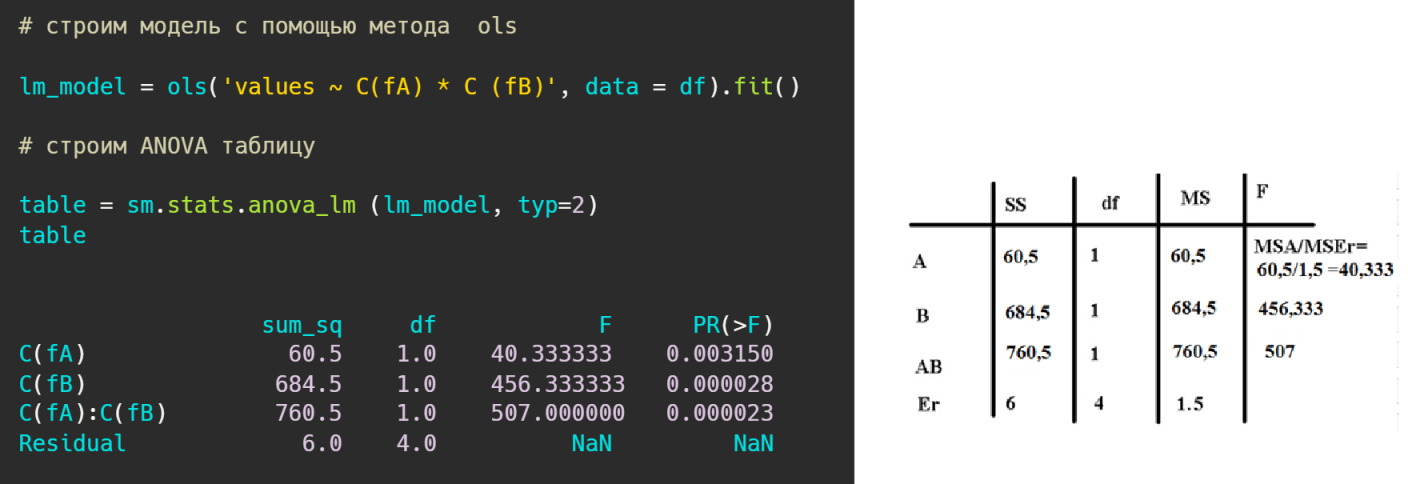
Массив values - это значения площади, которые соответствуют комбинации уровней.



Построим таблицу эксперимента 2 Х 2 в Python. Справа для сравнения приведена таблица, построенная от руки.



Теперь пришло время строить ANOVA таблицу, но сначала построим математическую модель

.

Как видим, мы получили идентичные значения критерия Фишера, но опять, вместо табличного значения, функция дает нам значения pvalue. Все pvalue меньше 0.05, значит, взаимодействие факторов и сами факторы оказывают статистически значимый эффект. Самое маленькое значение pvalue = 0.000023 для взаимодействия факторов, следовательно наибольшее влияние на производительность оказывает эффект взаимодействия факторов.

**Условия применимости дисперсионного анализа**

Для дисперсионного анализа есть следующие условия применимости:

1. Значения групп должны следовать нормальному распределению
2. Однородность дисперсий
3. Независимость измерений

Проверку по первому пункту вы можете выполнить с помощью теста Шапиро - Уилка или с помощью QQ – графика, Однородность (равенство) дисперсий определяете с помощью Барлетт теста или с помощью теста Левене. Независимость измерений вы должны обеспечить во время планирования эксперимента, чтобы исключить влияния какие-то внешних факторов, избежать предвзятых оценок.

Также рекомендуется делать выборки для каждой подгруппы одинакового объема, тогда нарушения нормальности не сильно влияет на результат.

На этом методе мы завершаем курс по статистическому анализу. Мы познакомились с различными распределениями, научились проводить тестирование гипотез, работать с качественными, количественными и порядковыми данными. Проводить тестирование гипотез, получать интервальные оценки параметров генеральной совокупности, научились прогнозировать зависимую переменную с помощью регрессионных моделей.