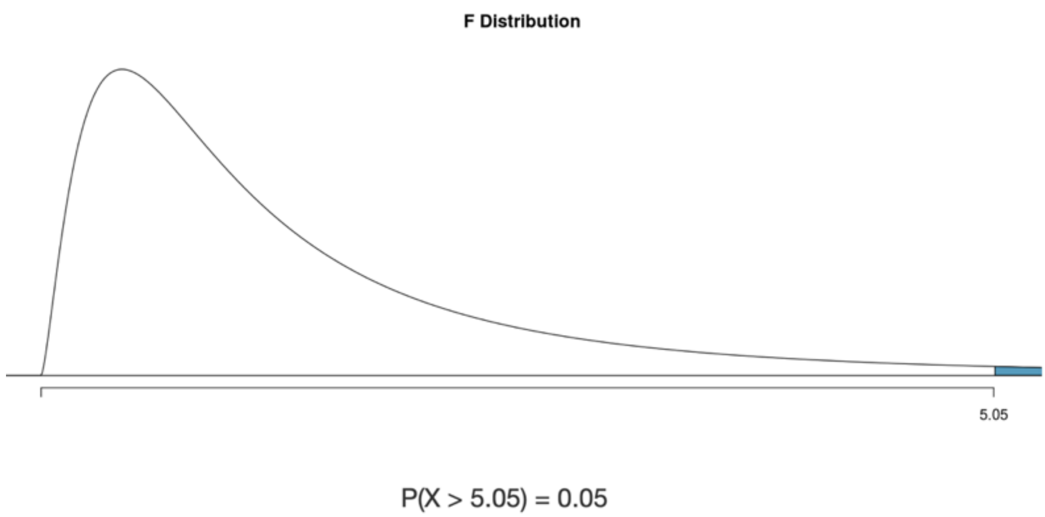
Дисперсионный анализ

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 H_0 утверждает, что все a классов градации имеют одинаковые значения средних: $\mu_1 = \mu_2 = ... = \mu_a$.

Альтернативная гипотеза H_1: 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.

3. В каждой группе непрерывная переменная имеет приблизительно нормальное распределение. Нормальность распределения можно проверить, используя гистограмму («на глаз») или статистические тесты на нормальность.
4. 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 $\alpha < 0{,}05$), 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: $\mu_1 = \mu_2 = \mu_3$

H₁: $\mu_1 \neq \mu_2 = \mu_3$ или $\mu_1 = \mu_2 \neq \mu_3$ или $\mu_1 \neq \mu_2 \neq \mu_3$

Вычислим среднее значение всех наблюдений:

$$\bar{\bar{x}} == \frac{\sum_{i=1}^{m}\sum_{j=1}^{n_i} x_{ij}}{n} = \frac{1+2+3+3+4+5+5+6+7}{9} = 4$$

Вычислим общую сумму квадратов:

$$SST = \sum_{i=1}^{m}\sum_{j=1}^{n_i}\left(x_{ij} - \bar{\bar{x}}\right)^2 (1-4)^2 + (2-4)^2 + (3-4)^2 +$$

$$+(3-4)^2 + (4-4)^2 + (5-4)^2 + (5-4)^2 + (6-4)^2 + (7-4)^2 = 30$$

Степени свободы для общей суммы квадратов:

$$dF_{SST} = n - 1 = 9 - 1 = 8$$

Вычислим средние значения внутри каждой из групп:

$$\bar{x}_i = \frac{\sum_{j=1}^{n_i} x_j}{n_i}, i = \overline{1, m}$$

$$\bar{x}_1 = \frac{1+2+3}{3} = 2$$

$$\bar{x}_2 = \frac{3+4+5}{3} = 4$$

$$\bar{x}_3 = \frac{5+6+7}{3} = 6$$

Внутригрупповая сумма квадратов:

$$SSW = \sum_{i=1}^{m}\sum_{j=1}^{n_i}\left(x_{ij} - \bar{x}_i\right)^2 = (1-2)^2 + (2-2)^2 + (3-2)^2 +$$

$$+(3-4)^2 + (4-4)^2 + (5-4)^2 +$$

$$+(5-6)^2 + (6-6)^2 + (7-6)^2 = 6$$

Степени свободы для внутригрупповой суммы квадратов:

$$dF_{SSW} = n - m = 9 - 3 = 6$$

Межгрупповая сумма квадратов:

$$SSB = \sum_{i=1}^{m} n_i(\bar{x}_i - \bar{\bar{x}})^2 = 3(2-4)^2 + 3(4-4)^2 + 3(6-4)^2 = 24$$

Степени свободы для межгрупповой суммы квадратов:

$$dF_{SSB} = m - 1 = 3 - 1 = 2$$

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:

$$MS_B = \frac{SSB}{dF_{SSB}} = \frac{24}{2} = 12$$

Внутригрупповая дисперсия:

$$MS_W = \frac{SSW}{dF_{SSW}} = \frac{6}{6} = 1$$

Вычислим F-значение:

$$F = \frac{MS_B}{MS_W} = \frac{12}{1} = 12$$

Критическое значение отношения Фишера:

$F_{0,05;\,2;6} = 5,14$

Так как фактическое отношение Фишера меньше критического:

$$F = 12 > 5{,}14 = F_{\,0,05;\,2;6}$$

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.

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import scipy.stats as stats
```

```
np.random.seed(12)
races = ["asian", "black", "hispanic", "white", "other"]

# генерируем случайные данные
voter_race = np.random.choice(a = races, p = [0.05, 0.15, 0.25, 0.05, 0.5], size = 1000) # категориальная переменная (раса избирателей)
voter_age = stats.poisson.rvs(loc = 18, mu = 30, size = 1000) # числовая переменная (возраст 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.

```python
# группируем данные возраста по рассе
voter_frame = pd.DataFrame({"race":voter_race, "age":voter_age})
groups = voter_frame.groupby("race").groups

# добавляем конкретные группы
asian = voter_age[groups["asian"]]
black = voter_age[groups["black"]]
hispanic = voter_age[groups["hispanic"]]
white = voter_age[groups["white"]]
other = voter_age[groups["other"]]
voter_frame.head()
```

	race	age
0	black	51
1	other	49
2	hispanic	51
3	other	48
4	asian	56

```python
#выполняем ANOVA
stats.f_oneway(asian, black, hispanic, white, other)
```

```
F_onewayResult(statistic=1.7744689357329695, pvalue=0.13173183201930463)
```

Рассмотрим альтернативный способ, используем функцию anova_lm() из библиотеки statsmodels:

```python
import statsmodels.api as sm
from statsmodels.formula.api import ols
```

```python
model = ols('age ~ race', data = voter_frame).fit()
anova_result = sm.stats.anova_lm(model, typ = 2)
print(anova_result)
```

```
             sum_sq     df         F    PR(>F)
race        199.369    4.0  1.774469  0.131732
Residual  27948.102  995.0       NaN       NaN
```

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.

```python
np.random.seed(12)

# генерируем случайные данные
voter_race = np.random.choice(a = races, p = [0.05, 0.15, 0.25, 0.05, 0.5], size = 1000)

white_ages = stats.poisson.rvs(loc = 18, mu = 32, size = 1000)
voter_age = stats.poisson.rvs(loc = 18, mu = 30, size = 1000)
voter_age = np.where(voter_race == "white", white_ages, voter_age)

# группируем данные возраста по расе
voter_frame = pd.DataFrame({"race":voter_race, "age":voter_age})
groups = voter_frame.groupby("race").groups

# добавляем конкретные группы
asian = voter_age[groups["asian"]]
black = voter_age[groups["black"]]
hispanic = voter_age[groups["hispanic"]]
white = voter_age[groups["white"]]
other = voter_age[groups["other"]]
```

```python
#выполняем ANOVA
stats.f_oneway(asian, black, hispanic, white, other)
```

```
F_onewayResult(statistic=3.6470318084857154, pvalue=0.00586731196131632)
```

```python
model = ols('age ~ race', data = voter_frame).fit()
anova_result = sm.stats.anova_lm(model, typ = 2)
print(anova_result)
```

```
                sum_sq     df         F    PR(>F)
race        472.278126    4.0  3.647032  0.005867
Residual  32212.272874  995.0       NaN       NaN
```

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 A и B, A и C, и B и C) и получили следующие три P-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 P-значений не с α, а с $\frac{\alpha}{\mathrm{m}}$, где m – число проверяемых гипотез. Деление исходного уровня значимости α на m – это и есть поправка Бонферрони. В рассматриваемом примере каждое из полученных P-значений необходимо было бы сравнить с $\frac{0.05}{3} = 0.017$. В результате мы выяснили бы, что P-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 P-значений на это число. Сравнив такие скорректированные P-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.

```python
# перебираем все пары
race_pairs = []

for race1 in range(4):
  for race2 in range (race1 + 1, 5):
    race_pairs.append((races[race1], races[race2]))

# t-test
for race1, race2 in race_pairs:
  print(race1, race2)
  print(stats.ttest_ind(voter_age[groups[race1]], voter_age[groups[race2]]))
```

```
asian black
Ttest_indResult(statistic=0.8386446909747979, pvalue=0.4027281369339345)
asian hispanic
Ttest_indResult(statistic=-0.42594691924932293, pvalue=0.6704669004240726)
asian white
Ttest_indResult(statistic=-2.235132300024921, pvalue=0.027828801627453537)
asian other
Ttest_indResult(statistic=0.3687230802619566, pvalue=0.712474249112879)
black hispanic
Ttest_indResult(statistic=-1.9527839210712925, pvalue=0.05156197171952594)
black white
Ttest_indResult(statistic=-3.4490459390086468, pvalue=0.0006893463707824467)
black other
Ttest_indResult(statistic=-0.9244381856060086, pvalue=0.3555931499524523)
hispanic white
Ttest_indResult(statistic=-2.1309216040170442, pvalue=0.033930889763891824)
hispanic other
Ttest_indResult(statistic=1.6450276425039192, pvalue=0.10037925272137736)
white other
Ttest_indResult(statistic=3.306112013211683, pvalue=0.0010062493632570478)
```

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.

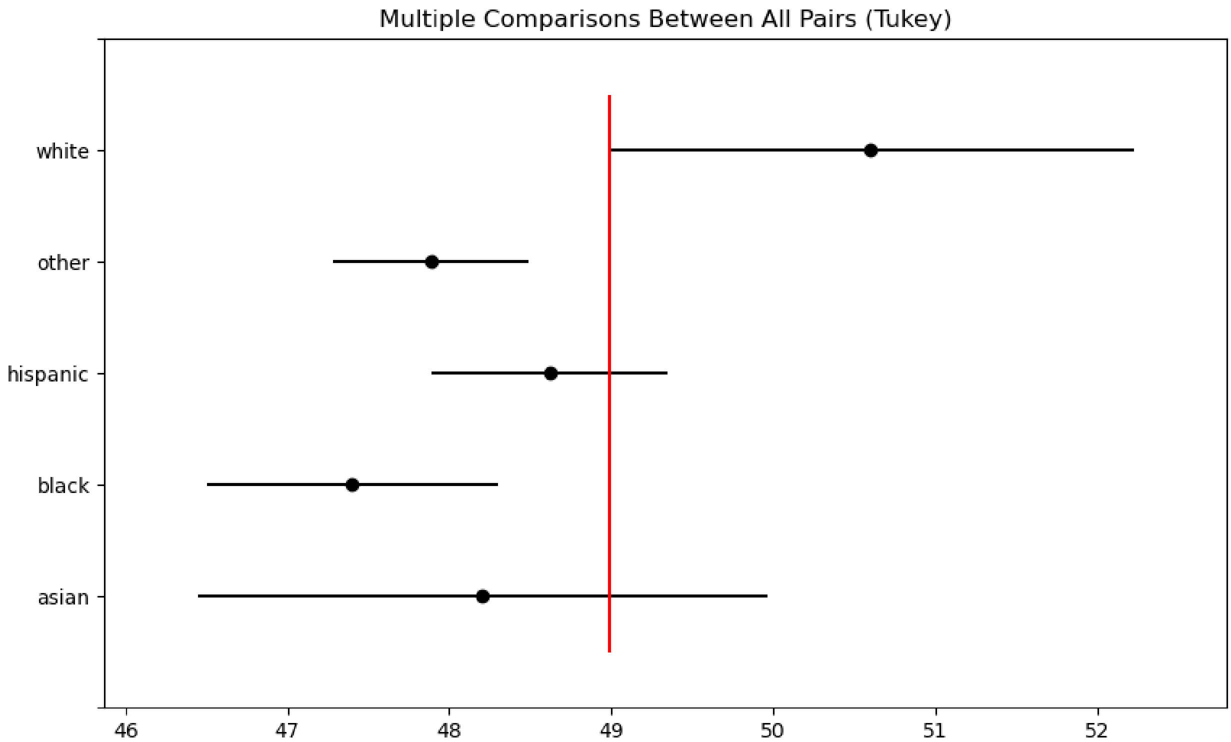
```
from statsmodels.stats.multicomp import pairwise_tukeyhsd
```

```
tukey = pairwise_tukeyhsd(endog = voter_age, groups = voter_race, alpha = 0.05)
tukey.plot_simultaneous()
plt.vlines(x = 49.57, ymin = -0.5, ymax = 4.5, color = "red")
tukey.summary()
```

Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
asian	black	-0.8032	0.924	-3.4752	1.8688	False
asian	hispanic	0.4143	0.9919	-2.1324	2.961	False
asian	other	-0.3191	0.9965	-2.7613	2.1231	False
asian	white	2.3955	0.2492	-0.8186	5.6095	False
black	hispanic	1.2175	0.2433	-0.406	2.8409	False
black	other	0.4841	0.8932	-0.9699	1.9381	False
black	white	3.1986	0.0056	0.653	5.7443	True
hispanic	other	-0.7334	0.4603	-1.9419	0.475	False
hispanic	white	1.9811	0.1649	-0.4326	4.3949	False
other	white	2.7146	0.0115	0.4113	5.0178	True

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 SSB_A, объяснённой влиянием фактора *B* суммы квадратов отклонений SSB_B и необъяснённой суммы квадратов отклонений (суммы квадратов отклонений ошибки):

$$SST = SSB_A + SSB_B + SSW$$

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

```
import numpy as np
import pandas as pd
```

```
# создаем данные
df = pd.DataFrame({'water': np.repeat(['daily', 'weekly'], 15),
 'sun': np.tile(np.repeat(['low', 'med', 'high'], 5), 2),
 'height': [6, 6, 6, 5, 6, 5, 5, 6, 4, 5,
 6, 6, 7, 8, 7, 3, 4, 4, 4, 5,
 4, 4, 4, 4, 4, 5, 6, 6, 7, 8]})
```

```
df.head()
```

	water	sun	height
0	daily	low	6
1	daily	low	6
2	daily	low	6
3	daily	low	5
4	daily	low	6

```python
import statsmodels.api as sm
from statsmodels.formula.api import ols
```

```python
# Выполняем ANOVA
model = ols('height ~ C(water) + C(sun) + C(water):C(sun)', data=df).fit()
sm.stats.anova_lm(model, typ=2)
```

	sum_sq	df	F	PR(>F)
C(water)	8.533333	1.0	16.0000	0.000527
C(sun)	24.866667	2.0	23.3125	0.000002
C(water):C(sun)	2.466667	2.0	2.3125	0.120667
Residual	12.800000	24.0	NaN	NaN

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.

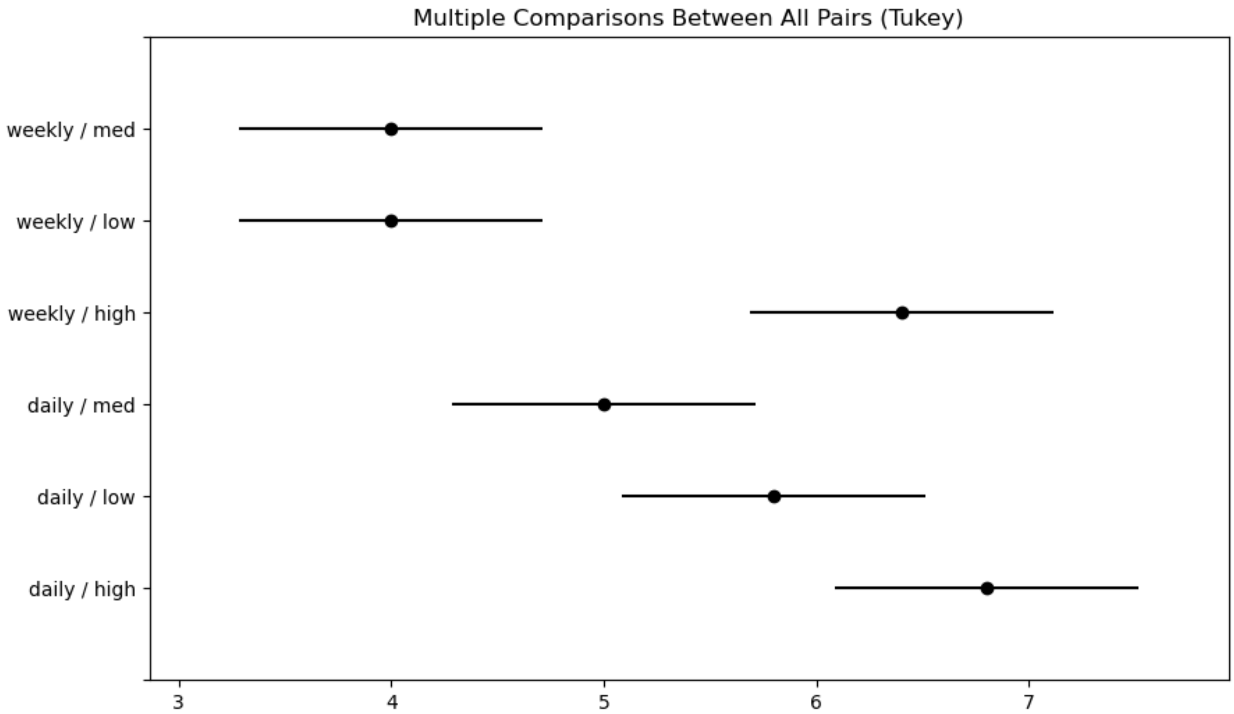
Выполним пост-хок тест Тьюки для нашего примера, а также построим график с доверительными интервалами.

```python
df['combination'] = df.water + " / " + df.sun
```

```python
tukey = pairwise_tukeyhsd(endog = df['height'], groups = df['combination'], alpha = 0.05)
tukey.plot_simultaneous()
#plt.vlines(x = 49.57, ymin = -0.5, ymax = 4.5, color = "red")
tukey.summary()
```

Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
daily / high	daily / low	-1.0	0.2898	-2.4281	0.4281	False
daily / high	daily / med	-1.8	0.0079	-3.2281	-0.3719	True
daily / high	weekly / high	-0.4	0.951	-1.8281	1.0281	False
daily / high	weekly / low	-2.8	0.0	-4.2281	-1.3719	True
daily / high	weekly / med	-2.8	0.0	-4.2281	-1.3719	True
daily / low	daily / med	-0.8	0.5252	-2.2281	0.6281	False
daily / low	weekly / high	0.6	0.7827	-0.8281	2.0281	False
daily / low	weekly / low	-1.8	0.0079	-3.2281	-0.3719	True
daily / low	weekly / med	-1.8	0.0079	-3.2281	-0.3719	True
daily / med	weekly / high	1.4	0.057	-0.0281	2.8281	False
daily / med	weekly / low	-1.0	0.2898	-2.4281	0.4281	False
daily / med	weekly / med	-1.0	0.2898	-2.4281	0.4281	False
weekly / high	weekly / low	-2.4	0.0003	-3.8281	-0.9719	True
weekly / high	weekly / med	-2.4	0.0003	-3.8281	-0.9719	True
weekly / low	weekly / med	0.0	1.0	-1.4281	1.4281	False



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:

$$SST = SSB_A + SSB_B + SSB_{AB} + SSW$$

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.