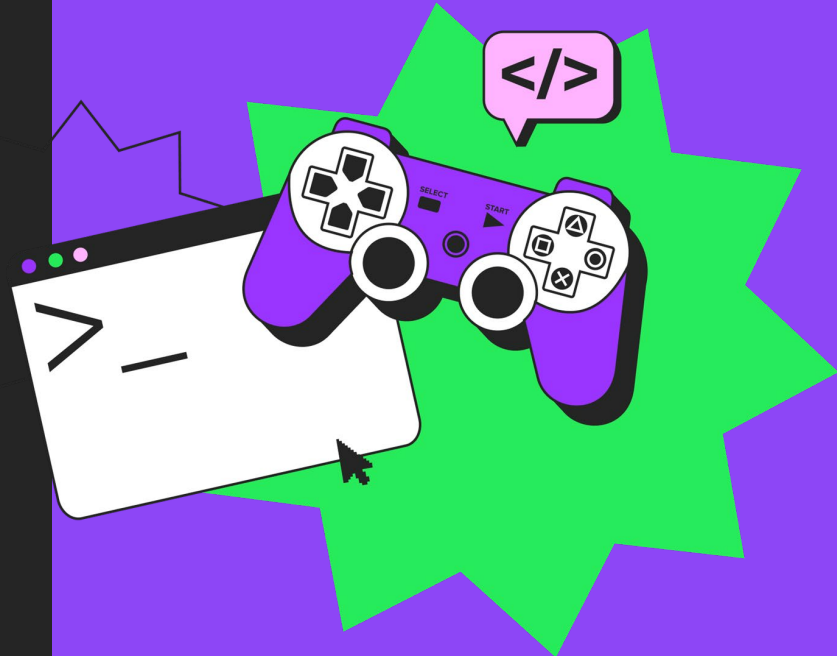




Дисперсионный анализ

Однофакторный и двухфакторный дисперсионный анализ.
Post hoc тест





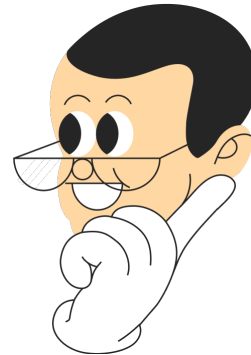
План курса





Что будет на уроке сегодня

- ✚ Однофакторный дисперсионный анализ
- ✚ Двухфакторный дисперсионный анализ
- ✚ Post hoc тест Тьюки
- ✚ Условия применимости дисперсионного анализа





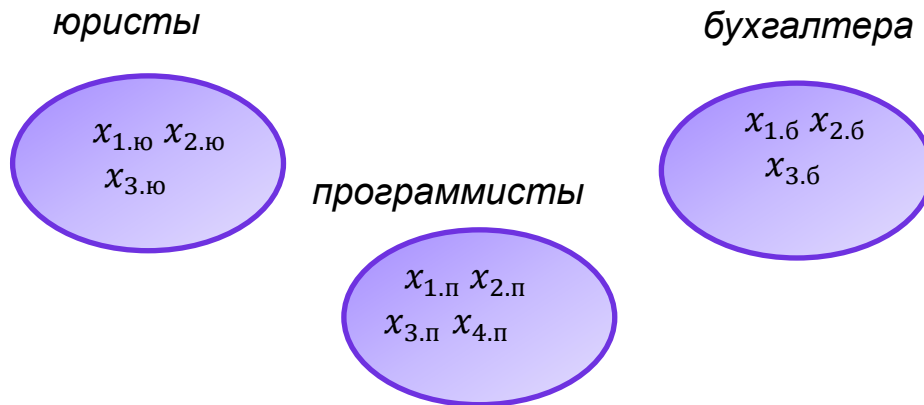
Дисперсионный анализ

Дисперсионный анализ используется для исследования влияния одного или нескольких качественных показателей на количественный показатель.



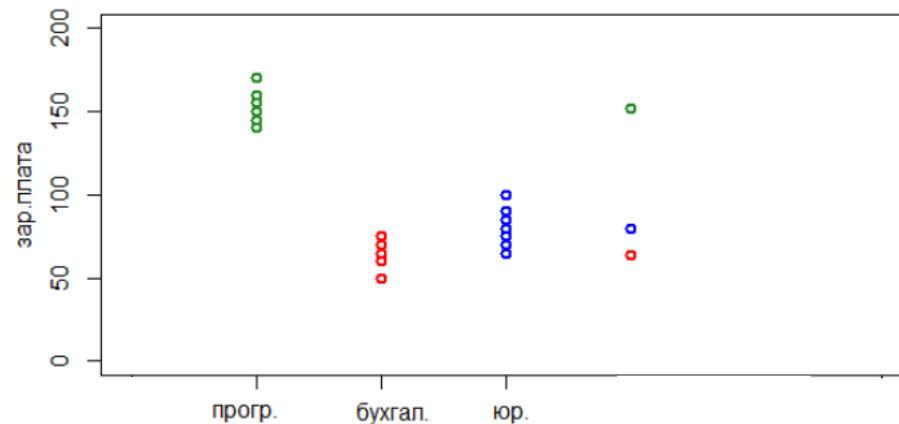
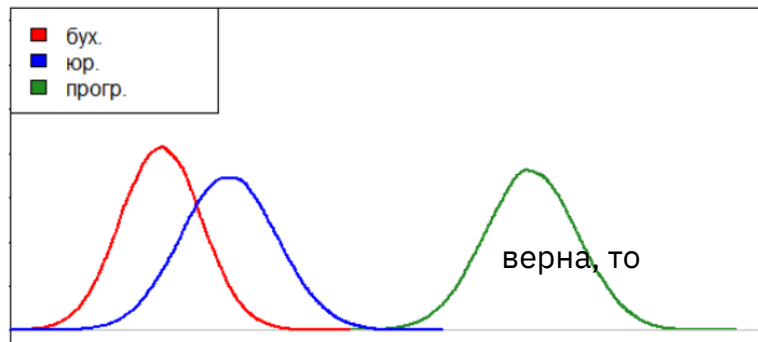
Однофакторный дисперсионный анализ

В однофакторном дисперсионном анализе на одну количественную переменную Y влияет один фактор (один качественный показатель), наблюдаемый на k уровнях, т.е. имеет k выборок для переменной Y .



Идея дисперсионного анализа

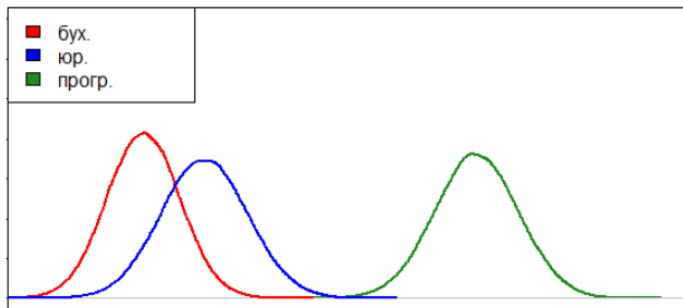
$$H_1 \begin{cases} H_0: \mu_1 = \mu_2 = \mu_3 \\ H_{1.1}: \mu_1 = \mu_2 \\ H_{1.2}: \mu_1 = \mu_3 \\ H_{1.3}: \mu_2 = \mu_3 \end{cases}$$



Если одна из альтернативных гипотез верна, то обнаружено влияние профессии на заработную плату.

Проблема множественных сравнений

Более 2 групп – критерий Фишера F



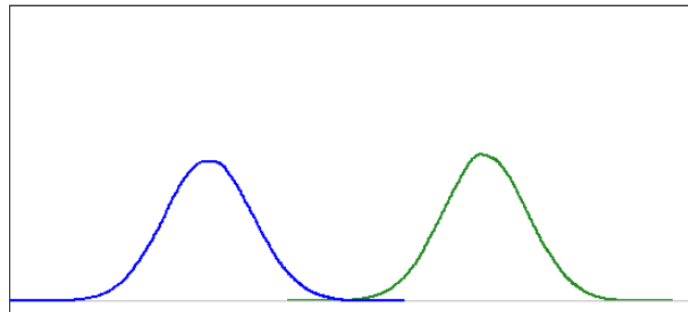
$$\bar{\alpha} = 1 - (1 - \alpha)^m$$

С увеличением числа сравнений m растет вероятность ошибки I рода для множественных сравнений ($\bar{\alpha}$).
Т.е. $\bar{\alpha}$ является истинным уровнем значимости многократно примененного критерия

$$m = 1, \bar{\alpha} = 1 - (1 - 0.05)^1 = 0.05,$$

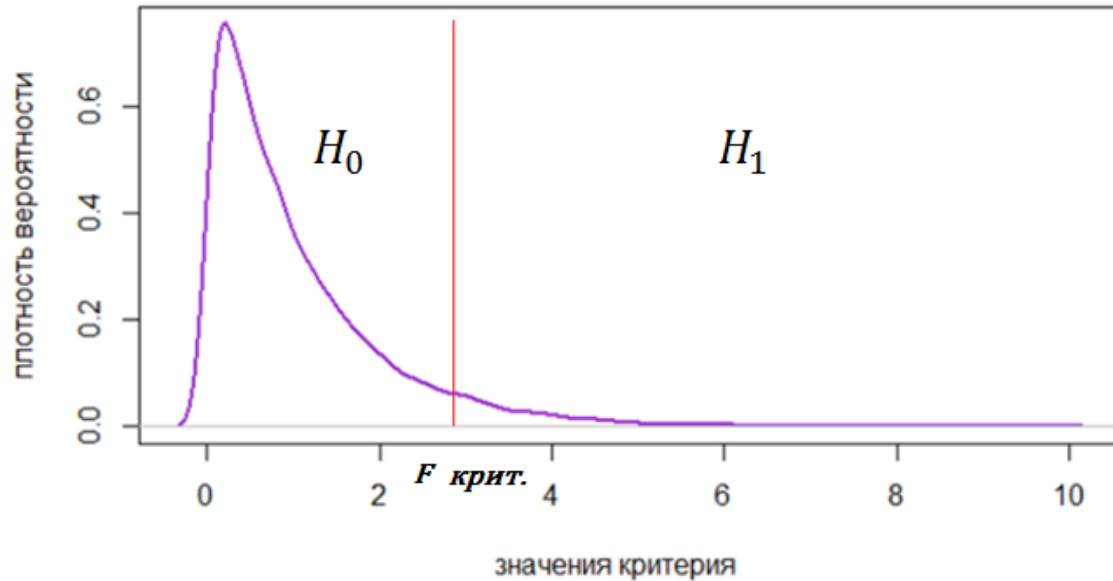
$$m = 3, \bar{\alpha} = 1 - (1 - 0.05)^3 = 0.14$$

2 группы – критерий Стьюдента t



Распределение Фишера

$$F_H = \frac{\sigma_{\Phi}^2}{\sigma_{\text{ост}}^2},$$

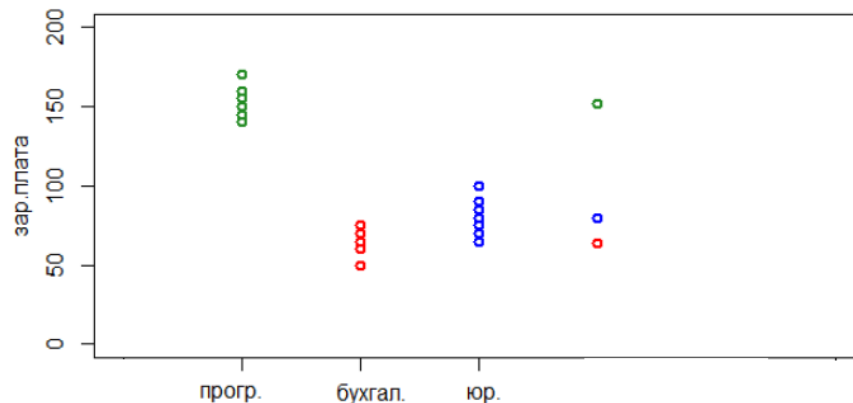




Факторная и остаточная дисперсия

Если бы все значения были взяты из одной генеральной совокупности, в которой профессия не оказывала бы влияния на заработную плату, то разброс внутри группы и межгрупповой были бы приблизительно одинаковыми. И в этом случае H_0 не отвергалась бы.

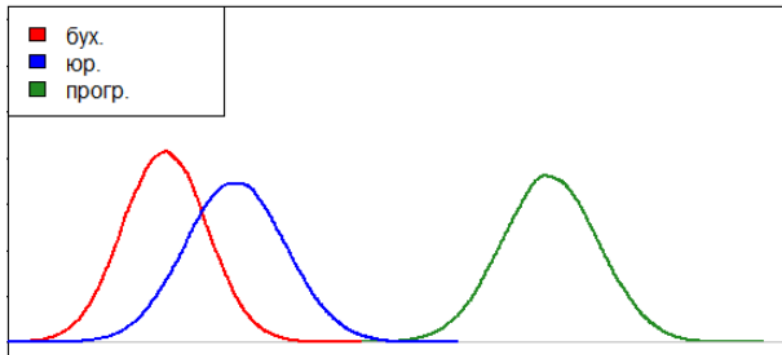
$$F_H = \frac{\sigma_{\phi}^2}{\sigma_{\text{ост}}^2},$$



post hoc tests

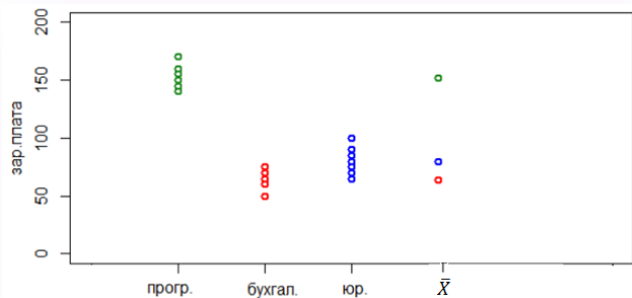
Дисперсионный анализ не отвечает на вопрос, между какими именно группами найдены статистически значимые различия. Если влияние фактора обнаружено и есть необходимость определить между какими группами есть статистически значимые различия, используют post hoc тесты для парных сравнений.

- ✓ Ньюмена-Кейлса
- ✓ Тест Тьюки
- ✓ Поправка Бонферрони (не использовать, когда более 8 сравнений)



Задача

Даны заработные платы юристов, программистов и бухгалтеров. Определить, влияет ли профессия на заработную плату.



$$F_H = \frac{\sigma_{\Phi}^2}{\sigma_{\text{ост}}^2}$$
$$\sigma_{\Phi}^2 = \frac{S_{\Phi}^2}{k-1}, \text{ где } k = 3$$
$$\sigma_{\text{ост}}^2 = \frac{S_{\text{ост}}^2}{n-k}, \text{ где } n = 21$$

```
import numpy as np
from scipy import stats
import matplotlib.pyplot as plt

y1 = np.array([70, 50, 65, 60, 75, 67, 74])
y2 = np.array([80, 74, 90, 70, 75, 65, 85])
y3 = np.array([148, 142, 140, 150, 160, 170, 155])

k = 3
n = 21

y_mean_1 = np.mean(y1)
y_mean_1
65.85714285714286

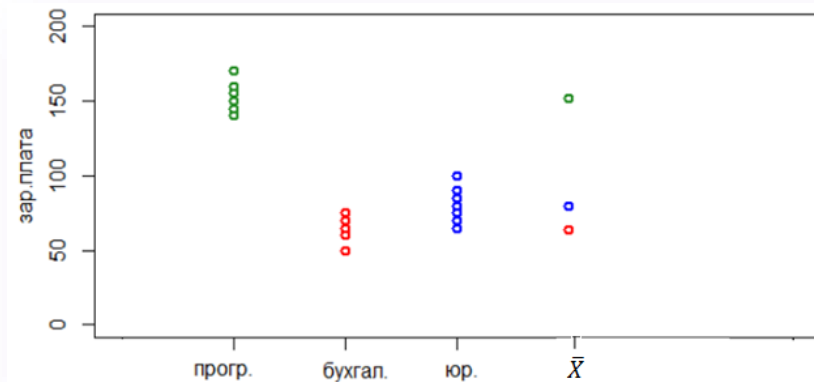
y_mean_2 = np.mean(y2)
y_mean_2
77.0
```



```
y_mean_3 = np.mean(y3)
y_mean_3
152.14285714285714

total = np.array([y1, y2, y3 ])
total
array([[ 70,  50,  65,  60,  75,  67,  74],
       [ 80,  74,  90,  70,  75,  65,  85],
       [148, 142, 140, 150, 160, 170, 155]])

y_mena_total= np.mean(total)
y_mena_total
98.33333333333333
```





$$S_{\text{общ}}^2 = \sum (y_{ij} - \bar{Y})^2 \approx 32400.$$

$$S_{\phi}^2 = \sum_{i=1}^k (\bar{y}_i - \bar{Y})^2 * n_i \approx 30836.95$$

$$S_{\text{ост}}^2 = \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 \approx 1563,71.$$

Сумма квадратов отклонений наблюдений от общего среднего

```
np.sum((total - 98.33)**2) # отложим это значение
32400.6669
```

Сумма квадратов отклонений средних групповых значений от общего среднего

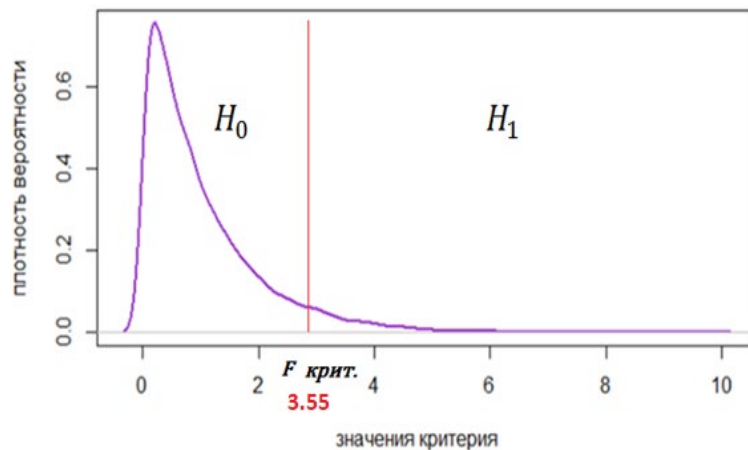
```
S_f = np.sum((y_mean_1 - 98.33)**2) * 7 + np.sum((y_mean_2 - 98.33)**2) * 7 +
np.sum((y_mean_3 - 98.33)**2) * 7 # S_f
S_f
30836.952614285707
```

Остаточная сумма квадратов отклонений

```
S_ost = np.sum((y1-y_mean_1)**2) + np.sum((y2-y_mean_2)**2) + np.sum ((y3-
y_mean_3)**2) # S_ost
S_ost
1563.7142857142858
```

```
30836.952614285707 + 1563.7142857142858
32400.666899999993
```

$$F_n = \frac{\sigma_{\Phi}^2}{\sigma_{\text{ост}}^2}$$
$$\sigma_{\Phi}^2 = \frac{S_{\Phi}^2}{k-1}, \text{ где } k = 3$$
$$\sigma_{\text{ост}}^2 = \frac{S_{\text{ост}}^2}{n-k}, \text{ где } n = 21$$



```
D_f = S_f / ( k - 1)
```

```
D_f
```

```
15418.476307142853
```

```
D_ost = S_ost / ( n - k)
```

```
D_ost
```

```
86.87301587301587
```

```
F_n = 15418.476307142853 / 86.87301587301587
```

```
F_n
```

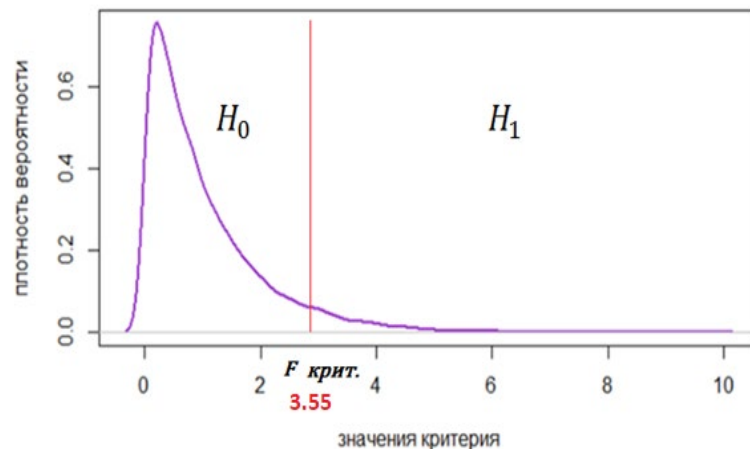
```
177.48291747670376
```

```
f = stats.f_oneway(y1, y2, y3)
```

```
f
```

```
F_onewayResult(statistic=177.48291613374704,  
                pvalue=1.420466900107174e-12)
```

$F_n = 15418.476307142853 / 86.87301587301587$
 F_n
 177.48291747670376



$k_1 \backslash k_2$	1	2	3	4	5
1	161,45	199,50	215,72	224,57	230,17
2	18,51	19,00	19,16	19,25	19,30
3	10,13	9,55	9,28	9,12	9,01
4	7,71	6,94	6,59	6,39	6,26
5	6,61	5,79	5,41	5,19	5,05
6	5,99	5,14	4,76	4,53	4,39
7	5,59	4,74	4,35	4,12	3,97
8	5,32	4,46	4,07	3,84	3,69
9	5,12	4,26	3,86	3,63	3,48
10	4,96	4,10	3,71	3,48	3,33
11	4,84	3,98	3,59	3,36	3,20
12	4,75	3,88	3,49	3,26	3,11
13	4,67	3,80	3,41	3,18	3,02
14	4,60	3,74	3,34	3,11	2,96
15	4,54	3,68	3,29	3,06	2,90
16	4,49	3,63	3,24	3,01	2,85
17	4,45	3,59	3,20	2,96	2,81
18	4,41	3,55	3,16	2,93	2,77
19	4,38	3,52	3,13	2,90	2,74



Post hoc test Tukey

```
from statsmodels.stats.multicomp import pairwise_tukeyhsd

import pandas as pd
df = pd.DataFrame({'score': [ 70,  50,  65,  60,  75,  67,  74,
                             80,  74,  90,  70,  75,  65,  85,
                             148, 142, 140, 150, 160, 170, 155],
                  'group': np.repeat(['accountant', 'lawyer', 'programmer'], repeats=7)})
tukey = pairwise_tukeyhsd(endog=df['score'],
                          groups=df['group'],
                          alpha=0.05)

print(tukey)
Multiple Comparison of Means - Tukey HSD, FWER=0.05
=====
  group1    group2  meandiff p-adj   lower   upper  reject
-----
accountant    lawyer  11.1429  0.0917 -1.5675  23.8532   False
accountant programmer  86.2857  0.001  73.5754  98.996    True
  lawyer programmer  75.1429  0.001  62.4325  87.8532    True
=====
```


Двухфакторный дисперсионный анализ

$$y_{ijk} = M + A_i + B_j + AB + E_{ijk}$$

$$y_{ijk} - M = A_i + B_j + AB + E_{ijk}$$

$$SS_T = SS_A + SS_B + SS_{AB} + SS_E$$

$$F_{HA} \text{ — } F_{кр.A}$$

$$F_{HB} \text{ — } F_{кр.B}$$

$$F_{HAB} \text{ — } F_{кр.AB}$$

		Фактор В j		
		1 уровень	2 уровень	
Фактор А i	1 уровень	57; 59 58	56; 58 57	57.5
	2 уровень	32; 34 33	71; 71 71	52
		45.5	64	54.75

Создаем данные в Python

```

y111= 57
y112 = 59
y11 = (y111 + y112)/2
y11
58.0

y121 = 56
y122 = 58
y12 = (y121 + y122)/2
y12
57.0

y211 = 32
y212 = 34
y21 = (y211 + y212)/2
y21
33.0

y221 = 71
y222 = 71
y22 = (y221 + y222)/2
y22
71.0

YcpA1 = (y11+y12)/2
YcpA1
57.5

YcpA2 = (y21+y22)/2
YcpA2
52.0

YcpB1 = (y11+y21)/2
YcpB1
45.5

YcpB2 = (y12+y22)/2
YcpB2
64.0


Ycp = np.mean( YcpA1 + YcpA2 + YcpB1 + YcpB2)/4
Ycp
54.75

```

		Фактор В J		
		1 уровень	2 уровень	
Фактор А i	1 уровень	57 ; 59 58	56;58 57	57.5
	2 уровень	32; 34 33	71;71 71	52
		45.5	64	54.75



Теперь будем производить расчеты и заносить их в ANOVA таблицу.


$$SSt = \sum (y_{ijk})^2 - \frac{a \cdot b \cdot n}{n} (Y_{cp})^2 = 57^2 + 59^2 + \dots + 71^2 - \frac{2 \cdot 2 \cdot 2}{2} (54.75)^2 = 1511.5$$

$$SSA = \frac{a \cdot n}{n} \sum (Y_{cpA})^2 - \frac{a \cdot b \cdot n}{n} (Y_{cp})^2 = \frac{2 \cdot 2}{2} ((57.5)^2 + (52)^2) - \frac{8}{2} (54.75)^2 = 60.5$$

$$SSB = \frac{b \cdot n}{n} \sum (Y_{cpB})^2 - \frac{a \cdot b \cdot n}{n} (Y_{cp})^2 = \frac{2 \cdot 2}{2} ((45.5)^2 + (64)^2) - \frac{8}{2} (54.75)^2 = 684.5$$

$$SSAB = \frac{n}{n} (\sum (y_{ij_cp})^2) - \frac{a \cdot b \cdot n}{n} Y_{cp}^2 - SSA - SSB = \frac{2}{2} ((58)^2 + (57)^2 + (33)^2 + (71)^2) - \frac{8}{2} (54.75)^2 - 60.5 - 684.5 = 760.5$$

$$SSE = SSt - SSA - SSB - SSAB = 1511.5 - 60.5 - 684.5 - 760.5 = 6$$



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



```
a = 2 # 2 уровня фактора a
b = 2 # 2 уровня фактора b
n = k = 2 # число повторных измерений

dfA = 2-1 = 1 # (a - 1)
dfB = 2-1 = 1 # (b - 1)

dfAB = (a - 1) * (b - 1) = (2 - 1) * (2 - 1) = 1
dfE = a * b * (n - 1) = 2 * 2 * (2 - 1) = 4

MSA = SSA / dfA = 60.5 / 1 = 60.5
MSB = SSB / dfB = 684.5 / 1 = 684.5

MSAB = SSAB / dfAB = 760.5 / 1 = 760.5
MSE = SSE / dfE = 6/4 = 1.5
```



Рассчитаем критерий Фишера и построим ANOVA таблицу, где последний столбец - это расчетный критерий Фишера.



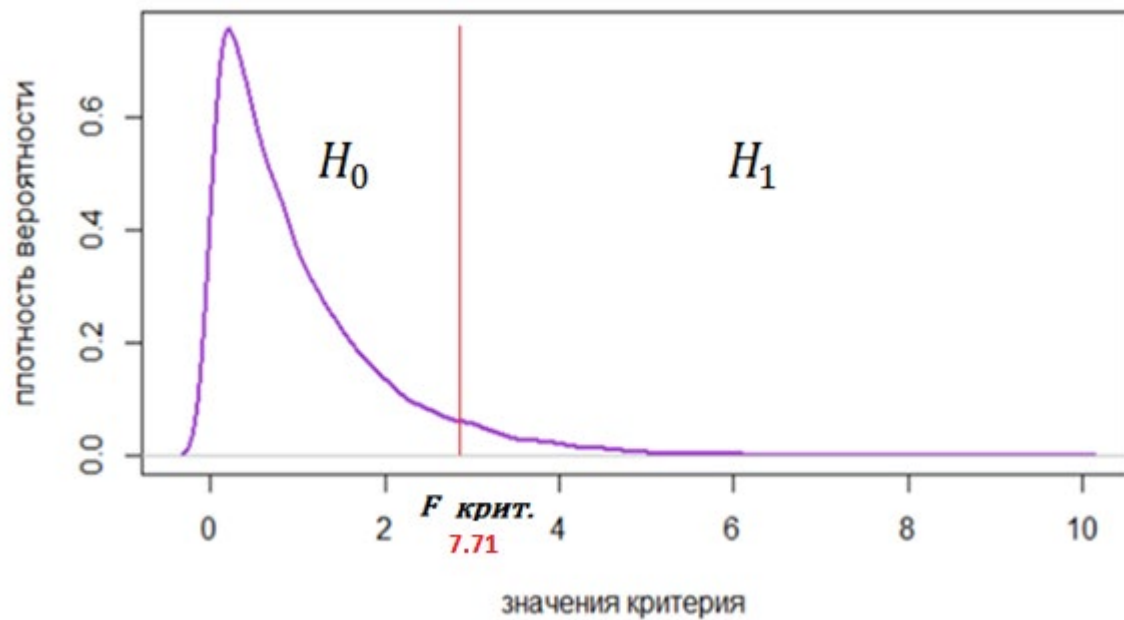
$$F_A = MSA / MSE = 60.5 / 1.5 = 40.33$$

$$F_B = MSB / MSE = 684.5 / 1.5 = 456.33$$

$$F_{AB} = MSAB / MSE = 760.5 / 1.5 = 507$$

$$F_{t} = 7.71$$

	SS	df	MS	F
A	60,5	1	60,5	40,33
B	684,5	1	684,5	456,33
AB	760,5	1	760,5	507
Er	6	4	1.5	





Двухфакторный дисперсионный анализ в Python

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

fA = np.array(["low", "low", "low", "low", "high", "high", "high", "high"])
fA

array(['low', 'low', 'low', 'low', 'high', 'high', 'high', 'high'],
      dtype='<U4')

fB = np.array(["low", "low", "high", "high", "low", "low", "high", "high"])
fB

array(['low', 'low', 'high', 'high', 'low', 'low', 'high', 'high'],
      dtype='<U4')
values = np.array([57, 59, 56, 58, 32, 34, 71, 71])
values
array([57, 59, 56, 58, 32, 34, 71, 71])
```

Создадим данные в Python



```
df=pd.DataFrame({'fA': fA, 'fB': fB, 'values': values})  
df
```

	fA	fB	values
0	low	low	57
1	low	low	59
2	low	high	56
3	low	high	58
4	high	low	32
5	high	low	34
6	high	high	71
7	high	high	71

		Фактор В J		
		1 уровень	2 уровень	
Фактор А i	1 уровень	57; 59 58	56; 58 57	57.5
	2 уровень	32; 34 33	71; 71 71	52
		45.5	64	54.75

y_{ijk}

Строим ANOVA- таблицу в Python

```
# строим модель с помощью метода ols

lm_model = ols('values ~ C(fA) * C(fB)', data = df).fit()

# строим ANOVA таблицу

table = sm.stats.anova_lm(lm_model, typ=2)
table
```

	sum_sq	df	F	PR(>F)
C(fA)	60.5	1.0	40.333333	0.003150
C(fB)	684.5	1.0	456.333333	0.000028
C(fA):C(fB)	760.5	1.0	507.000000	0.000023
Residual	6.0	4.0	NaN	NaN

	SS	df	MS	F
A	60,5	1	60,5	MSA/MSEr= 60,5/1,5 =40,333
B	684,5	1	684,5	456,333
AB	760,5	1	760,5	507
Er	6	4	1.5	



Условия применимости дисперсионного анализа

- ✓ Независимость измерений
- ✓ Значения групп должны следовать нормальному распределению
- ✓ Однородность (равенство) дисперсий

Если размеры выборок одинаковые, то неоднородность дисперсий слабо влияет на результат.



Что изучили в этом курсе?

- ✓ Случайные события. Формула Байеса
- ✓ Дискретные распределения
- ✓ Описательная статистика. EDA
- ✓ Нормальное распределение. ЦПТ
- ✓ Тестирование гипотез. Z и t критерии
- ✓ Доверительные интервалы
- ✓ Работа с долями
- ✓ Непараметрические тесты
- ✓ Корреляционный анализ
- ✓ Линейная регрессия
- ✓ Дисперсионный анализ



Конец