Теория вероятностей и математическая статистика ¶

Урок 8. Дисперсионный анализ. Логистическая регрессия

Урок 8

1. Провести дисперсионный анализ для определения того, есть ли различия среднего роста среди взрослых футболистов, хоккеистов и штангистов.

Даны значения роста в трех группах случайно выбранных спортсменов:

- Футболисты: 173, 175, 180, 178, 177, 185, 183, 182.
- Хокеисты: 177, 179, 180, 188, 177, 172, 171, 184, 180.
- Штангисты: 172, 173, 169, 177, 166, 180, 178, 177, 172, 166, 170.

В [1]:
```
import numpy as np
```

Рост футбалистов (footballers)

В [2]:
```
y1 = np.array([173, 175, 180, 178, 177, 185, 183, 182], dtype=np.float64)
```

Рост хокеистов (hockey_players)

В [3]:
```
y2 = np.array([177, 179, 180, 188, 177, 172, 171, 184, 180], dtype=np.float64)
```

Рост штангистов (weightlifters)

В [4]:
```
y3 = np.array([172, 173, 169, 177, 166, 180, 178, 177, 172, 166, 170], dtype=np.float64)
```

В [5]:
```
np.sort(y1)
```

Out[5]:
```
array([173., 175., 177., 178., 180., 182., 183., 185.])
```

В [6]:
```
np.sort(y2)
```

Out[6]:
```
array([171., 172., 177., 177., 179., 180., 180., 184., 188.])
```

В [7]:
```
np.sort(y3)
```

Out[7]:
```
array([166., 166., 169., 170., 172., 172., 173., 177., 177., 178., 180.])
```

В [8]:
```
n1 = len(y1)
n2 = len(y2)
n3 = len(y3)
n = n1 + n2 + n3
print(f'n1 = {n1}')
print(f'n2 = {n2}')
print(f'n3 = {n3}')
print(f'n = n1 + n2 +n3 = {n}')
```

```
n1 = 8
n2 = 9
n3 = 11
n = n1 + n2 +n3 = 28
```

Всего 3 группы

В [9]:
```
k = 3
```

Проведем однофакторный дисперсионный анализ (однофакторная ANOVA). Сначала найдем средний рост для каждой группы:

Средние арифметические по подгруппам

В [10]:
```
y1_mean = y1.mean()
print(y1_mean)
```

```
179.125
```

В [11]:
```
y2_mean = y2.mean()
print(y2_mean)
```

```
178.66666666666666
```

В [12]:
```
y3_mean = y3.mean()
print(y3_mean)
```

```
172.72727272727272
```

Видно, что средний рост в каждой из групп отличается от остальных. Установим, что это отличие статистически значимо. Для этого сначала соберем

все значения заработных плат в один массив:

В [13]:
```
y_all = np.concatenate([y1, y2, y3])
y_all
```

Out[13]:
```
array([173., 175., 180., 178., 177., 185., 183., 182., 177., 179., 180.,
       188., 177., 172., 171., 184., 180., 172., 173., 169., 177., 166.,
       180., 178., 177., 172., 166., 170.])
```

Найдем среднее значение роста по всем значениям:

В [14]:
```
y_mean = np.mean(y_all)
print(y_mean)
```

```
176.46428571428572
```

Найдем S^2 — сумму квадратов отклонений наблюдений от общего среднего:

В [15]:
```
s2 = np.sum((y_all - y_mean)**2)
s2
```

Out[15]:
```
830.9642857142854
```

Найдем S_F^2 - сумму квадратов отклонений средних групповых значений от общего среднего:

$$S_F^2 = \sum_{i=1}^{k} (\overline{y}_i - \overline{Y})^2 n_i$$

В [16]:
```
s2_f = ((y1_mean - y_mean)**2) * n1 + ((y2_mean - y_mean)**2) * n2 + ((y3_mean - y_mean)**2) * n3
s2_f
```

Out[16]:
```
253.9074675324678
```

Найдем $S_{\text{ост}}^2$ — остаточную сумму квадратов отклонений:

$$S_{\text{ост}}^2 = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \overline{y}_i)^2$$

В [17]:
```
# s2_ost, s2_residual
s2_residual  = np.sum((y1 - y1_mean) ** 2) + np.sum((y2 - y2_mean) ** 2) + np.sum((y3 - y3_mean) ** 2)

print(s2_residual )
```

```
577.0568181818182
```

Удостоверимся, что соблюдается равенство $S^2 = S_F^2 + S_{\text{ост}}^2$:

В [18]:
```
print(s2)
print(s2_f + s2_residual )
```

```
830.9642857142854
830.964285714286
```

Найдем общую дисперсию:

В [19]:
```
sigma2_general = s2 / (n - 1)
sigma2_general
```

Out[19]:
```
30.776455026455015
```

Найдем факторную дисперсию:

В [20]:
```
sigma2_f = s2_f / (k - 1)
sigma2_f
```

Out[20]:
```
126.9537337662339
```

Найдем остаточную дисперсию:

В [21]:
```
sigma2_residual = s2_residual / (n - k)
sigma2_residual
```

Out[21]:
```
23.08227272727273
```

Вычислим критерий Фишера F_H:

В [22]:
```
F_h = sigma2_f / sigma2_residual
F_h
```

Out[22]:
```
5.500053450812598
```

Найдем значение $F_{\text{крит}}$ в таблице критических точек распределения Фишера-Снедекора для заданного уровня значимости $\alpha = 0.05$ и двух степеней свободы:

$df_{\text{межд}} = k - 1 = 3 - 1 = 2$ и $df_{\text{внутр}} = n - k = 28 - 3 = 25$.

Для данных значений $F_{\text{крит}} = 3.4928$. Так как $F_H > F_{\text{крит}}$, различие среднего роста в трех группах статистически значимо.

В [23]:
```
alpha = 0.05
d_f1 = k - 1
d_f2 = n - k
n, k, d_f1, d_f2
```

Out[23]:
```
(28, 3, 2, 25)
```

Вычислим эмпирическое корреляционное отношение η^2:

В [24]:
```
eta2 = s2_f / s2
eta2
```

Out[24]:
```
0.30555761769498
```

Значение η^2 = 0.3056 — значит, связь в величине средних по выделенным группам колеблемость слабая.

В [25]:
```
from scipy import stats
```

Однофакторная ANOVA

В [26]:
```
#stats.f_oneway?
```

В [27]:
```
stats.f_oneway(y1, y2, y3)
```

Out[27]:
```
F_onewayResult(statistic=5.500053450812596, pvalue=0.010482206918698694)
```

- $F_H = \sigma_F^2/\sigma_{ost}^2$ = 5.500053450812596
- p_{value} = 0.010482206918698694
- $\alpha > p_{value}$ - различие среднего роста в трех группах статистически значимо на уровне значимости $\alpha = 5\%$.

В [ ]: