Теория вероятностей и математическая статистика[¶](#gjdgxs)

## Урок 8[¶](#30j0zll)

## Дисперсионный анализ. Метод главных компонент. Логистическая регрессия[¶](#1fob9te)

### Разбор домашнего задания[¶](#3znysh7)

**Задача 1**

Используя данные о хоккеистах (скачать можно [здесь](https://yadi.sk/d/qPyKnmlLwl4SBA)), проверьте, является ли среди хоккеистов из Финляндии, Норвегии и Дании значимым отличие: а) роста, б) BMI.

**Решение**

Указанные гипотезы можно проверить с помощью однофакторного дисперсионного анализа. Реализуем для этого готовую функцию.

In [1]:

**import** numpy **as** np  
**import** pandas **as** pd

In [2]:

**def** anova(**\***ys: np**.**ndarray) **->** float:  
 """Функция возвращает значение F-статистики  
 для дисперсионного анализа.  
 """  
   
 y **=** np**.**concatenate(ys)  
  
 ss\_b **=** sum((yi**.**mean() **-** y**.**mean()) **\*\*** 2 **\*** yi**.**size **for** yi **in** ys)  
 ss\_w **=** sum(((yi **-** yi**.**mean()) **\*\*** 2)**.**sum() **for** yi **in** ys)  
   
 sigma\_b **=** ss\_b **/** (len(ys) **-** 1)  
 sigma\_w **=** ss\_w **/** (y**.**size **-** len(ys))  
   
 **return** sigma\_b **/** sigma\_w

In [3]:

df **=** pd**.**read\_csv('../../practice/hockey\_players.csv')  
  
df**.**head(2)

Out[3]:

|  | **year** | **country** | **no** | **name** | **position** | **side** | **height** | **weight** | **birth** | **club** | **age** | **cohort** | **bmi** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **0** | 2001 | RUS | 10 | tverdovsky oleg | D | L | 185 | 84.0 | 1976-05-18 | anaheim mighty ducks | 24.952772 | 1976 | 24.543462 |
| **1** | 2001 | RUS | 2 | vichnevsky vitali | D | L | 188 | 86.0 | 1980-03-18 | anaheim mighty ducks | 21.119781 | 1980 | 24.332277 |

In [4]:

df['country']**.**unique()

Out[4]:

array(['RUS', 'AUT', 'BLR', 'CAN', 'CZE', 'FIN', 'GER', 'ITA', 'JPN',  
 'LAT', 'NOR', 'SUI', 'SVK', 'SWE', 'UKR', 'USA', 'POL', 'SLO',  
 'DEN', 'FRA', 'KAZ', 'HUN'], dtype=object)

In [5]:

countries **=** ['FIN', 'NOR', 'DEN']

Выборки из роста хоккеистов выбранных стран:

In [6]:

ys **=** [df**.**loc[df['country'] **==** country, 'height'] **for** country **in** countries]  
  
F **=** anova(**\***ys)  
F

Out[6]:

3.59345750724064

Для построения критической области нам нужны квантили распределения Фишера с параметрами $k\_1 = k - 1$, $k\_2 = n - k$, где $k$ — число уровней, $n$ — общее число элементов во всех выборках.

In [7]:

k **=** len(countries)  
n **=** sum(y**.**size **for** y **in** ys)  
  
k1 **=** k **-** 1  
k2 **=** n **-** k  
  
k1, k2

Out[7]:

(2, 430)

In [8]:

**from** scipy **import** stats

In [9]:

alpha **=** 0.05  
  
t **=** stats**.**f**.**ppf(1 **-** alpha, k1, k2)  
t

Out[9]:

3.016700272452327

Итак, критическая область: $$\Omega\_\alpha = \left( 3.017, \infty \right)$$

In [10]:

**from** matplotlib **import** pyplot **as** plt  
  
plt**.**style**.**use('seaborn-whitegrid')  
**%config** InlineBackend.figure\_formats = ['svg']

In [11]:

ox **=** np**.**linspace(0, 5, 500)  
oy **=** stats**.**f**.**pdf(ox, k1, k2)  
  
ox\_right **=** np**.**linspace(t, 5, 100)  
oy\_right **=** stats**.**f**.**pdf(ox\_right, k1, k2)  
  
plt**.**plot(ox, oy)  
plt**.**fill\_between(ox\_right, oy\_right, alpha**=**0.5)

Out[11]:

<matplotlib.collections.PolyCollection at 0x122efb580>

2021-01-13T16:32:18.903626 image/svg+xml Matplotlib v3.3.3, https://matplotlib.org/

Поскольку статистика попала в критическую область, заключаем, что между ростами хоккеистов выбранных стран действительно есть значимое отличие.

Аналогично для BMI. Пересчитать нужно только F-статистику, поскольку остальные параметры такие же, как раньше.

In [12]:

ys **=** [df**.**loc[df['country'] **==** country, 'bmi'] **for** country **in** countries]  
  
F **=** anova(**\***ys)  
F

Out[12]:

8.682513420531967

Разница между BMI хоккеистов также признаётся значимой.

In [ ]: