Автор: Анна Смелова

Решение домашнего задания к уроку 6 “Взаимосвязь величин. Параметрические и непараметрические показатели корреляции. Корреляционный анализ.”

In [1]:
```
import numpy as np
import pandas as pd
```

1. Задача

Даны значения величины заработной платы заемщиков банка (zp) и значения их поведенческого кредитного скоринга (ks):

$zp = [35, 45, 190, 200, 40, 70, 54, 150, 120, 110]$, $ks = [401, 574, 874, 919, 459, 739, 653, 902, 746, 832]$.

Найдите ковариацию этих двух величин с помощью элементарных действий, а затем с помощью функции cov из numpy

Полученные значения должны быть равны.

Найдите коэффициент корреляции Пирсона с помощью ковариации и среднеквадратичных отклонений двух признаков, а затем с использованием функций из библиотек numpy и pandas.

Решение:

In [2]:
```
zp=np.array([35,45,190,200,40,70,54,150,120,110])
ks=np.array([401,574,874,919,459,739,653,902,746,832])
```

$$cov_{XY} = M[(X - M(X))(Y - M(Y))] = M(XY) - M(X)M(Y) = \overline{X \cdot Y} - \overline{X} \cdot \overline{Y}$$

In [3]:
```
cov_1 = np.mean(zp * ks) - np.mean(zp) * np.mean(ks)
print(f'Ковариация zp и ks, вычисленная с помощью элементарных действий: {cov_1}')
```

Ковариация zp и ks, вычисленная с помощью элементарных действий: 9157.839999999997

In [4]:
```
cov_2 = np.cov(zp,ks,ddof=0)
print(f'Ковариация, вычисленная с помощью функции cov: \n{cov_2}')
```

Ковариация, вычисленная с помощью функции cov:
```
[[ 3494.64  9157.84]
 [ 9157.84 30468.89]]
```

$$r_{XY} = \frac{cov_{XY}}{\sigma_X \sigma_Y}$$

In [5]:
```
sigma_zp = np.std(zp, ddof=0)
sigma_ks = np.std(ks, ddof=0)
r_xy_1 = cov_1 / (sigma_zp * sigma_ks)
print(f'Коэффициент корреляции Пирсона, вычисленный с помощью ковариации и среднеквадратичных отклонений: {r_xy_1}')
```

Коэффициент корреляции Пирсона, вычисленный с помощью ковариации и среднеквадратичных отклонений: 0.8874900920739158

In [6]:
```
r_xy_2 = np.corrcoef(zp, ks)
print(f'Коэффициент корреляции Пирсона, вычисленный с помощью библиотеки numpy:\n {r_xy_2}')
```

Коэффициент корреляции Пирсона, вычисленный с помощью библиотеки numpy:
```
 [[1.         0.88749009]
 [0.88749009 1.        ]]
```

In [7]:
```
a = [[35,45,190,200,40,70,54,150,120,110],
     [401,574,874,919,459,739,653,902,746,832]]
df = pd.DataFrame(data=a).T
r_xy_3 = df.corr(method="pearson")
print(f'Коэффициент корреляции Пирсона, вычисленный с помощью библиотеки pandas:\n {r_xy_3}')
```

Коэффициент корреляции Пирсона, вычисленный с помощью библиотеки pandas:
```
          0         1
0  1.00000  0.88749
1  0.88749  1.00000
```

2. Задача

Измерены значения IQ выборки студентов, обучающихся в местных технических вузах:

131, 125, 115, 122, 131, 115, 107, 99, 125, 111.

Известно, что в генеральной совокупности IQ распределен нормально. Найдите доверительный интервал для математического ожидания с надежностью 0.95.

Решение:

In [8]:
```
a = [131,125,115,122,131,115,107,99,125,111]
x = np.mean(a)
print(f'Выборочное среднее: {x}')
```

```
Выборочное среднее: 118.1
```

In [9]:
```
n = len(a)
print(f'n: {n}')
```

```
n: 10
```

In [10]:
```
sigma = np.std(a, ddof=1)
print(f'Несмещенное среднеквадратическое отклонение: {sigma}')
```

```
Несмещенное среднеквадратическое отклонение: 10.54566788359614
```

Среднеквадратическое отклонение генеральной совокупности неизвестно, поэтому используем t-критерий Стьюдента.
$T_{1,2} = \overline{x} \pm t_{\frac{\alpha}{2}} \cdot \frac{\sigma_H}{\sqrt{n}}$

In [11]:
```
T_1 = x - 2.262 * sigma / n**0.5
T_2 = x + 2.262 * sigma / n**0.5
print(f'Искомый доверительный интервал [{T_1},{T_2}]')
```

```
Искомый доверительный интервал [110.55660776308164,125.64339223691834]
```

3. Задача

Известно, что рост футболистов в сборной распределен нормально с дисперсией генеральной совокупности, равной 25 кв.см.

Объем выборки равен 27, среднее выборочное составляет 174.2.

Найдите доверительный интервал для математическогоожидания с надежностью 0.95.

Решение:

In [12]:
```
x = 174.2
n = 27
sigma = 25**0.5
```

Среднеквадратическое отклонение генеральной совокупности известно, поэтому используем z-критерий.
$T_{1,2} = \overline{x} \pm z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$

In [13]:
```
T_1 = x - 1.96 * sigma / n**0.5
T_2 = x + 1.96 * sigma / n**0.5
print(f'Искомый доверительный интервал [{T_1},{T_2}]')
```

```
Искомый доверительный интервал [172.31398912064722,176.08601087935276]
```

4. Задача

Выберите тему для проектной работы по курсу Теории вероятностей и математической статистики и напишите ее в комментарии к Практическому заданию. (ПРОЕКТ ДОБРОВОЛЬНЫЙ)

Описание курсового проекта по предмету Теория вероятностей и математическая статистика

Во второй половине курса студентам даётся проект с темой на выбор. Студентам предлагается общее направление для проекта - исследовать данные с imdb. Для этих данных можно сделать EDA (Exploratory Data Analysis – разведочный анализ данных) либо проверить предложенную самим студентом статистическую гипотезу. При наличии опыта работы с данными и статистикой некоторые студенты также могут взять другую тему - например, исследовать курс биткойна.

В качестве данных студенты могут брать датасеты с информацией с сайта imdb (либо с Kaggle, либо откуда-то из Интернета, эти данные нужно искать самим). Также есть возможность взять любые другие данные, если выбрана другая тема. Работы выполняются с использованием любых методов из вебинаров и методичек по предмету Теория вероятностей и математическая статистика. Это может быть постановка и проверка статистических гипотез, EDA, дисперсионный анализ и т.д.

Готовый проект должен быть выложен на github в аккаунте студента в виде файла с расширением ipynb (Jupyter Notebook), ссылка на проект прикладывается в раздел Задание. Срок выполнения задания - неделя после 8-го вебинара. Начинать можно после 6-го вебинара, после того, как преподаватель одобрит тему (студент указывает тему работы в Задание к вебинару 6).

Ссылка на проект должна быть добавлена студентом к Заданию вебинара 8.

Примеры постановки темы проекта:
Проверить гипотезу: Комедийные фильмы, снятые до 1990 года, в среднем имеют более высокую оценку, чем комедийные фильмы, снятые позже.
Статистическое сопоставление цены биткоина с курсами фиатных активов (USD,EUR,RUR), ценами на нефть, драг металлы или с ценами других криптовалют по отношению к биткойну. Также студент может до выполнения работы поставить широкий круг задач для проекта и уже в процессе выполнения оставить только то, что вошло в проект.

Пример:

"В проектной работе хотела бы сделать анализ по Странам по данным imbd.com. По странам рассчитать основные статистические показатели: Мода, медиана, квартили, дисперсию. Проверить, работает ли нормальное распределение и Паретто. Проверить есть ли зависимость бюджетов, рейтингов, жанров от страны выпуска фильмов. Посмотреть поведение стран во времени. В идеале, хотела бы сделать прогноз на 2020 год: какие страны с какой вероятностью выпустят определенное количество фильмов с определенным бюджетом определенного жанра."