УДК 004.624

Корреляция Пирсона на языке программирования Python

Кизянов Антон Олегович
Приамурский государственный университет имени Шолом-Алейхема
Студент

Аннотация
В данной статье будет представлено, что такое корреляция Пирсона и как она может использоваться при анализе данных.
Ключевые слова: Python, numpy, matplotlib

Pearson Correlation in the Python Programming Language

Kizyanov Anton Olegovich
Sholom-Aleichem Priamursky State University
student

Abstract
This article will describe what Pearson's correlation is and how it can be used in data analysis.
Keywords: Python, numpy, matplotlib

Метод был придуман Карлом Пирсоном в 1896 году, он измеряет линейные корреляция между двумя переменными. Основные формулы представлены ниже:

$$r = r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}} \quad 1)$$

$$F(r) = \frac{1}{2} ln \frac{1+r}{1-r} = \operatorname{arctanh}(r) \quad 2)$$

$$SE = \frac{1}{\sqrt{n-3}} \quad 3)$$

$$z = \frac{x - mean}{SE} = [F(r) - F(\rho_0)]\sqrt{n-3} \quad 4)$$

1) определяет коэффициент; 2) описывает преобразование Фишера, используемое для вычисления доверительных интервалов; 3) дает стандартную ошибку корреляции; 4) относится к z-оценке преобразованной Фишером корреляции. Если применим нормальное распределение, можем использовать z-оценку для вычисления доверительных интервалов. В качестве альтернативы, можем выполнить загрузку путем повторной выборки пар значений с заменой. Кроме того, scipy.stats.pearsonr() функция возвращает значение p, которое (согласно документации) неточно для

образцов с менее чем 500 значениями. Будем сопоставить данные по выбросам углекислого газа из Всемирного банка с соответствующими данными о температуре для Нидерландов [1].

Цель исследования – демонстрация корреляции Пирсона на примере анализа выброса углеводорода и температуры Нидерландов.

Ранее этим вопросом интересовалась Л. Шишлянникова развивала тему «Применение корреляционного анализа в психологии» [2] в которой обсуждается применение корреляционного анализа Пирсона в психологии. Вводится понятие корреляционной связи и ее характеристик. Описываются возможности для применения коэффициентов корреляции по типу шкалы. И.М. Янников, Н.В. Козловская, А.Д. Емшанов с темой «Применение корреляционного анализа для обработки результатов биомониторинга мест размещения отходов» [3], а подробнее про нахождения корреляции Пирсона при анализе данных биомониторинга. Были получены результаты корреляции Пирсона в рамках НИР и применение технологии определения критериев экологической безопасности. В.С. Попукайло опубликовал статью «Исследование линейной корреляционной связи в парных выборках малого объема» [4] рассказал про методы нахождения линейной корреляционной модели в выборках малого объема. Исследовано влияние виртуального увеличения объема выборки на значение коэффициента корреляции Пирсона и модифицированного индекса Фехнера.

Сначала нужно импортировать все нужные библиотеки.

```
import dautil as dtl
import pandas as pnd
from scipy import stats
import numpy as nmp
import math
from sklearn.utils import check_random_state
import matplotlib.pyplot as mplt
from IPython.display import HTML
from IPython.display import dsply
```

Загрузить данные и настроить соответствующие структуры данных:

```
wb = dtl.data.Worldbank()
indicator = wb.get_name('co2')
co2 = wb.download(country='NL', indicator=indicator, start=1900,
                  end=2014)
co2.index = [int(year) for year in co2.index.get_level_values(1)]
temp = pnd.DataFrame(dtl.data.Weather.load()['TEMP'].resample('A'))
temp.index = temp.index.year
temp.index.name = 'year'
df = pnd.merge(co2, temp, left_index=True, right_index=True).dropna()
```

Вычислить корреляция следующим образом:

```
stats_corr = stats.pearsonr(df[indicator].values, df['TEMP'].values)
```

```
print('Correlation={0:.4g}, p-value={1:.4g}'.format(stats_corr[0], stats_corr[1]))
```

Вычислить доверительный интервал с преобразованием Фишера:

```
z = nmp.arctanh(stats_corr[0])
n = len(df.index)
se = 1/(math.sqrt(n - 3))
ci = z + nmp.array([-1, 1]) * se * stats.norm.ppf((1 + 0.95)/2)
ci = nmp.tanh(ci)
dtl.options.set_pnd_options()
ci_table = dtl.report.DFBuilder(['Low', 'High'])
ci_table.row([ci[0], ci[1]])
```

Повторная выборка пар с заменой:

```
rs = check_random_state(34)
ranges = []
for j in range(200):
    corrs = []
    for i in range(100):
        indices = rs.choice(n, size=n)
        pairs = df.values
        gen_pairs = pairs[indices]
        corrs.append(stats.pearsonr(gen_pairs.T[0], gen_pairs.T[1])[0])
    ranges.append(dtl.stats.ci(corrs))
ranges = nmp.array(ranges)
bootstrap_ci = dtl.stats.ci(corrs)
ci_table.row([bootstrap_ci[0], bootstrap_ci[1]])
ci_table = ci_table.build(index=['Formula', 'Bootstrap'])
```

Вычислить результаты и отобразить отчет:

```
x = nmp.arange(len(ranges)) * 100
mplt.plot(x, ranges.T[0], label='Low')
mplt.plot(x, ranges.T[1], label='High')
mplt.plot(x, stats_corr[0] * nmp.ones_like(x), label='SciPy estimate')
mplt.ylabel('Pearson Correlation')
mplt.xlabel('Number of bootstraps')
mplt.title('Bootstrapped Pearson Correlation')
mplt.legend(loc='best')
rslt = dtl.report.HTMLBuilder()
rslt.h1('Pearson Correlation Confidence intervals')
rslt.h2('Confidence Intervals')
rslt.add(ci_table.to_html())
HTML(rslt.html)
```

Результат работы представлен на рисунке 1.

Pearson Correlation Confidence Intervals

Tested with scipy=0.16.0, pandas=0.16.2, numpy=1.9.2, IPython=3.2.1, numexpr=2.3.1

Confidence Intervals

	High	Low
Formula	0.658	0.225
Bootstrap	0.668	0.260

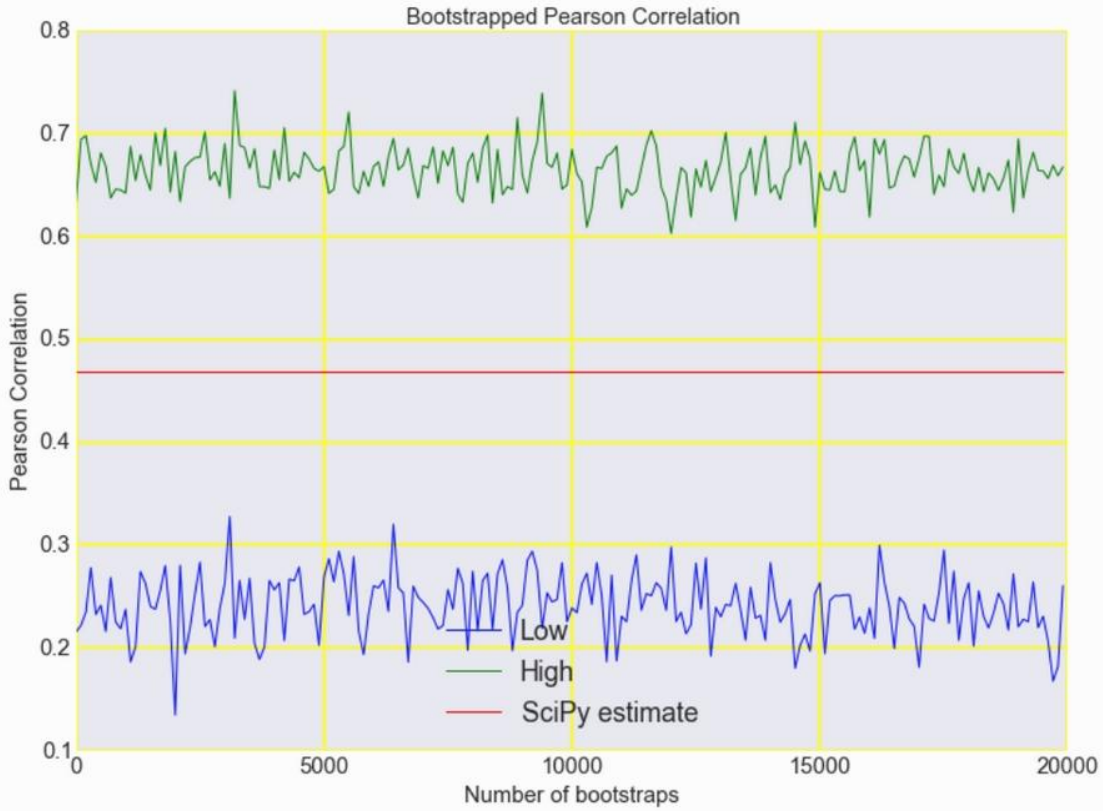

Рис. 1 График корреляции Пирсона

Вывод

Таким образом, коэффициент Пирсона получился чуть меньше 0.5, что говорит о том, что прямой связи между количеством выброса углеводорода и температуры в Нидерландах нет.

Библиографический список

1. Центр анализа информации о двуокиси углерода URL: https://data.worldbank.org/indicator/EN.ATM.CO2E.PC (Дата обращения: 11.08.2018
2. Шишлянникова Л. Применение корреляционного анализа в психологии // Психологическая наука и образование. 2009. №1. С. 98-107. URL: https://elibrary.ru/item.asp?id=12242651 (Дата обращения: 11.08.2018)
3. Янников И.М., Козловская Н.В., Емшанов А.Д. Применение

корреляционного анализа для обработки результатов биомониторинга мест размещения отходов // Научно-исследовательские публикации. 2014. № 7(11). С. 38-43. URL: https://elibrary.ru/item.asp?id=21500977 (Дата обращения: 11.08.2018)

4. Попукайло В.С. Исследование линейной корреляционной связи в парных выборках малого объема // Технология и конструирование в электронной аппаратуре. 2016. №1. С. 27-32. URL: https://elibrary.ru/item.asp?id=26099962 (Дата обращения: 11.08.2018)