

# pythonを用いたe-Statのデータとの相関係数を求めるプログラムの作成

群馬県立太田女子高等学校 2年 青木響加

## 概要

APIからe-Statの全ての都道府県別統計データ呼び出し、任意のデータとの相関係数を出すプログラムを作成・公開した。ここでの「任意のデータ」とは、プログラム使用者が用意した47都道府県順の変量である数値データのことであり、これにより相関係数を短時間で大量に求めることができる。またプログラム・統計学初心者や勉強中の人の参考にしてもらえると考えている。

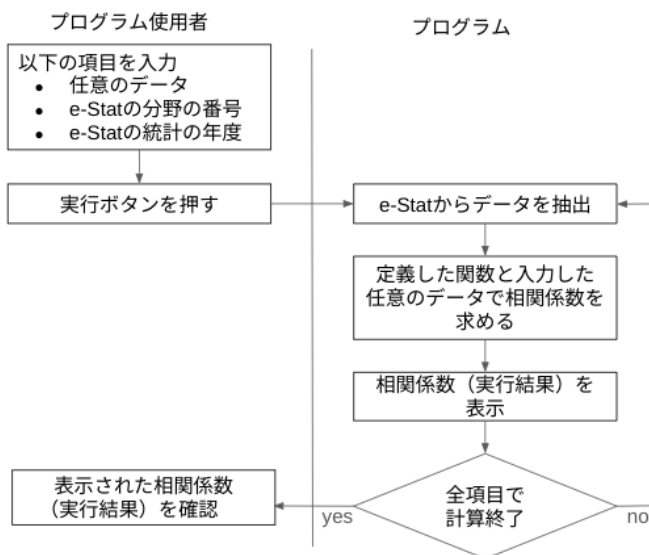
## きっかけ

e-Stat [1] では国内の統計情報が誰でも利用できるように公開されている。しかし任意のデータとの相関を求めたい場合、表計算ソフト等でデータごとに手作業で行う必要がある。そのため人為的なミスもあり時間もかかってしまう。そこでAPIからe-Statのデータ呼び出しして相関係数を出すpythonプログラム [2] を作成しようと考えた。

## 方針

APIからe-Statのデータ呼び出しして任意のデータとの相関係数を出すプログラムをGoogle Colaboratoryで作成した。

Google Colaboratoryではセルと呼ばれるソースコードを処理するためのまとまりを使用する。



(図1)アクティビティ図

このプログラムは3つのセルから構成されている。

1. パラメーターのセル(セル1)
  - 行政番号順の変量となる数値データの配列
  - e-Statの10桁の分野の文字列
  - e-Statの10桁の統計の年度の文字列
2. 関数定義のセル(セル2)  
e-Statから全データを都道府県別に抽出し、そのデータを引数として関数を定義するセル
3. 関数呼び出しのセル(セル3)  
その関数を使用して相関係数をpandasライブラリ [3] を用いて求めるセル(セル3)

## プログラムを公開するねらい

- プログラミングや統計学の初心者参考にもらえる。
- e-Statのデータと任意のデータで相関係数を求める際一からプログラムを作らなくて済む。

## プログラムの使用方法

例として、e-Statの2021年度都道府県別データの人口・世帯 分野のデータと公立高校の都道府県別女子スラックス採用率 [4] との相関係数を出す。

1. QRコードよりプログラムにアクセスする  
<https://colab.research.google.com/drive/1tTZ4UFd2RW4JzsphqROzLtoNqusXu6bT>
2. 一番最初のセルの3つの括弧の中に行政番号順の変量となる実数、e-Statの10桁の分野の文字列、e-Statの10桁の統計の年度の文字列を入力する。



```
3 # 任意のデータを入力。
4 target_array = (59.4,0.7,0.5,40,10,30,20,20,60,40,50,77.3,60,84.3,70,10)
5 # e-Statの分野の番号を入力。13の分野で0000010101~0000010113までである。
6 target_category = ("0000010113")
7 # e-Statのデータの年度を入力。上4桁が西暦。
8 target_year = ("2021100000")
```

(図2)任意のデータを入力するセル(一部)

3. セル1、セル2、セル3の順に実行する。

## 結果・考察

人口・世帯分野の569項目の都道府県別データとの相関係数を約7分で全て求めることができた。他の分野の実行時間は下の表の通りである。

1項目あたりの平均実行時間は0.95秒となった。このように手作業よりも短時間で相関係数を出すことができた。

分野の文字列	分野	項目数	実行時間(秒)
"0000010101"	人口・世帯	569	420
"0000010102"	自然環境	44	40
"0000010103"	経済基盤	502	360
"0000010104"	行政基盤	194	120
"0000010105"	教育	486	360
"0000010106"	労働	530	360
"0000010107"	文化・スポーツ	248	180
"0000010108"	居住	451	300
"0000010109"	健康・医療	397	240
"0000010110"	福祉・社会保障	366	240
"0000010111"	安全	167	120
"0000010112"	家計	265	180
"0000010113"	生活時間	214	120
合計		3189	3040

(表1)項目ごとの実行時間一覧

## 展望

- 任意のデータとe-Statの年度別都道府県別データで相関を出しているため、それをe-Statの月別都道府県別データでも可能にすることができるの幅を広げたい。
- 実行結果を相関係数が大きい順に並べられるプログラムを作成すれば実行結果からの読み取りが今より容易になると思った。

## 参考文献

1. 政府統計の総合窓口 (e-Stat) のAPI仕様 3.0版  
<https://api.e-stat.go.jp/swagger-ui/e-statapi3.0.html>
2. python公式リファレンス  
<https://www.python.org/>
3. pandas公式リファレンス  
<https://pandas.pydata.org/>
4. Culture Convenience Club  
[https://www.ccc.co.jp/news/2021/20210909\\_002181.html](https://www.ccc.co.jp/news/2021/20210909_002181.html)