python を用いたジニ係数計算プログラムの作成

○吉村文孝、山崎絹世、安藤洋

生物・生体技術支援室　動植物育成管理技術グループ

概要

汎用プログラミング言語 python のプログラミング技術向上を目的とし、データの整形から拡張ジニ係数計算までを自動で行う python スクリプトの作成を行った。本スクリプトには csv ファイルの読み込みからデータの加工のほか、繰り返し処理、条件分岐などプログラミングの基礎となる事項を含むため、基礎の習得と確認に有益と考え作成を企図した。

完成した python スクリプトの動作確認のため、1 グループのサンプルサイズ 70～90、全 96 グループのテストデータセットを用意し本スクリプトに計算を行わせた。その結果、Microsoft Excel で手動計算した結果と同一の計算結果を算出することができたため、スクリプトは正常に動作していると考えられる。

スクリプトの拡張ジニ係数に関する部分を他の計算に置き換えることで様々な計算を自動化できるため、今後も担当業務の自動化を継続したい。

1　初めに

python はオープンソースの汎用プログラミング言語であり数値計算だけでなく作業の自動化、アプリの開発、ゲーム作成など多様な応用範囲を持っている。筆者はこれまでに python を用いて各種統計計算、機械学習、図表作成などを行っているが、用いるデータの整列や分類といった下準備に関しては Microsoft Excel を利用し手動で行っている。大量のデータを定期的、高頻度に分析する際、Excel によるデータの手動加工にかかる時間と労力は大きなものとなる。また、データを特定の条件で分類する場合も肉眼分類では実施者に対して大きな労力を要求する。こういった反復や条件分岐などを繰り返すような処理はプログラムの得手とする分野である。本報告はデータの加工から分析、計算までを行える python スクリプトの作成を通じたプログラミング技術向上を企図して実施するものである。

ジニ係数は標本間格差を評価する代表的な指標の一つで「任意の 2 標本の差が全標本の平均に対して取る比率の期待値」と定義される（伊藤ほか、2012）。標本値が負の値を取る場合には拡張ジニ係数（伊藤ほか、2012）を利用することができる。

通常のジニ係数の算出を行うには、まず標本値を昇順に並び替え、横軸に累積標本数を取り、縦軸に累積標本値を取りこれを図示することでローレンツ曲線と均等分布直線を作成する。次に、ローレンツ曲線と均等分布直線に囲まれた面積と均等分布直線以下の三角形の面積との比を求めることでジニ係数の算出となる。発表者が Microsoft Excel で拡張ジニ係数の計算を試みたところ、累積標本値の状態で拡張ジニ係数の計算式に場合分けを要することがわかった。拡張ジニ係数の算出には作業者によるデータの確認、場合分けを必要とすることから、多量のデータを頻回に計算する労力は大きなものとなる。

拡張ジニ係数の計算には上記のようにデータの加工処理、繰り返しや条件分岐を要することから、自動化プログラム作成の練習に適していると考えられる。以上より、本報告ではプログラミング技術の向上を目的とし、データの読み込みから拡張ジニ係数算出までを行う python スクリプトの作成を行う。

2　方法

2.1　python スクリプトの作成

Web ブラウザ上から python を開発できる IDE（総合開発環境）である Google Colaboratory を主に利用し python スクリプトを作成した。作成したスクリプト中の「A、B、C、D、E」のアルファベットは伊藤ほか（2012）による拡張ジニ係数の計算方法に倣い設定した累積標本値数グラフの各領域名である（図 1）。図 1 中の (0 , 0) はグラフの原点を示す。B と C はそれぞれ正と負の累積標本値数の総和となる。他の領域は拡頂点の値から三角形、四角形として求めた面積から B や C の値を除して算出した。各領域の面積をそれぞれ求め、拡張ジニ係数 G** = (A + C) / (A +B + C + D + E) により拡張ジニ係数を算出した。

2.2　python スクリプトで実行する事柄

スクリプトを作成するにあたり、満たすべき条件を以下のように設定した。

・現場で取得したデータの csv ファイルの読み込めること

・データの量として、1 グループのサンプルサイズを最大 200 程度とし、200 グループ程度を想定

・データの昇順配列、累積標本値計算ののち、ローレンツ曲線と均等分布直線に囲まれた面積と均等分布直線以下の三角形の面積の計算などを行い、拡張ジニ係数の計算を行えること

・上記計算をデータのグループごとに行い、グループ数まで順番に自動計算できること

・計算した拡張ジニ係数をすべて出力できること

2.3　python スクリプトの機能確認

python スクリプトの機能確認のため、python スクリプトが算出したジニ係数を Microsoft Excel により手動で行った計算の結果と比較した。

一群が 70～100 個ほどの増減数値を持つテストデータセットを用意し、利用した。一群を一列とし総群数 96 グループを一つの csv ファイルに収容した。本データセットについて Microsoft Excel (office 365) と作成した python スクリプトのそれぞれで拡張ジニ係数を 1 群ずつ計算し、結果の比較を行った。正確に計算できていれば両者の計算結果は完全に一致するはずなので、両者の計算結果を散布図に図示するとともに相関係数を算出し一致しているかの確認を行った。

3　結果と考察

表 1 に作成したスクリプトを示した。以下にスクリプトの詳細な説明を述べる。

1～3 行目で Google drive をマウントして Google drive 内のファイルを読み込めるようにし､9 行目で Google drive 内に入れておいた csv ファイルを読み込んでいる。

13 行目で取得したデータセットの列数を利用して 15～48 行の拡張ジニ係数を列数まで繰り返し実行させている。これにより列数がどれだけ増えても自動でその列数まで計算を繰り返すことができる。

16～22 行目でデータセット全体から 1 列ずつデータを取り出し､昇順整列､累積標本値の計算をしている。このとき、データのない部分（NaN）を 0 に置換している。この処理を行わずにデータに NaN を含むと最終的な計算結果をすべて NaN にしてしまう。

25～29 行目では if – else 文による条件分岐と for 文による繰り返しを用いて累積標本値の値を正と負に分類し、正の値の累積標本値の合計（領域 B）（31 行目）と負の値の累積標本値の絶対値合計（領域 C）（33 行目）に振り分けて値を累積させている。

35～43 行目で累積標本値の合計から算出した面積の値によって拡張ジニ係数の計算式を分岐させている。

領域 B の面積の有無を分岐条件としている。これは、B の面積の有無によって他の部分の面積を求めるために参照する頂点を変更する必要があるためである。

領域 A の面積を計算する際、三角形の面積として計算したことにより、領域 A 左肩の斜辺（完全平等線）が実際には棒グラフの集合体である A の面積をわずかに削いでしまい、そのままでは計算結果に影響する。そこで欠けた部分の面積を数式で補正（43、45 行）することでこの問題を解決している。別のアイディアとして領域 A の棒グラフをサンプルサイズまで生成しその和によって領域 A の面積とする方法も考えられる。本報告ではより単純な計算で済む前述の方式を採用している。

python スクリプトによる計算結果が正しいことを確認するために Microsoft Excel で手動計算した結果との比較を行った。両計算結果を散布図による図示（図 2）と相関係数によって比較した。両者の計算結果が一致していれば x 軸、y 軸に両者の値を取った散布図のプロットは一直線上に並び、相関係数はちょうど 1.0 になるはずである。図 2 に示したようにプロットは一直線上に並んでおり、両者の計算結果は完全に一致していると見られる。相関係数も 1.0 となったことから、本報告で作成した python スクリプトはジニ係数を誤りなく計算できていると考えられる。

計算に要した時間は Microsoft Excel による手動計算では数時間に及んだが、python スクリプトでは数秒であり効率の差は大きい。スクリプトの記述上はサンプルサイズ、グループ数ともに無制限に増やすことが可能であるが、Google Colaboratory の処理能力に上限を依存すると考えられる。本報告でテストに利用したデータ（データ 70〜100 個を 1 グループとし全 96 グループ）程度では数秒でデータ量、計算量的に問題ないと考えられる。計算速度、処理能力ともに実用上快適な水準にある。

今回作成したスクリプトではライブラリーpandas でデータを読み込んでいるが計算自体はライブラリー numpy で行っている。ここにプログラム的な無駄を生んでいると考えられる。numpy ならばそもそも見出し行をデータに含めることができないため、今回のスクリプトのように pandas で header = none（図 1，9 行目）と記述してデータに見出し行を含ませないよう処理する必要がない。また、半角数字のみで構成されたデータセットを読み込んでいることも、numpy で読み込ませることに適している条件である。数値や文字列の混じったデータを読み込む場合であれば pandas が有効なので柔軟に対応したい。コード作成時の発表者の知識水準ではこの非効率の残る形になってしまったが今後はより無駄のないコード作成を心掛けたい。

スクリプトの拡張ジニ係数に関する部分を他の計算に置き換えることで様々な計算を自動化できるため、今後も自動化の可能性を検討しながら担当業務を進めたい。

参考文献

[1] 伊藤尚・前田義信・谷賢太郎・林豊彦・宮川道夫． 2012．標本合計が負の場合へ拡張されたジニ係数の評価．理論と方法，7 : 117 – 130.

表1. データの加工から拡張ジニ係数計算、出力までを行うpythonスクリプト

行	コード
1	# google driveをマウントしてファイルを読み込めるようにする
2	from google.colab import drive
3	drive.mount (' / content / drive ')
4	#モジュールの読み込み
5	import numpy as np
6	import matplotlib.pyplot as plt
7	import pandas as pd
8	# csvファイル（ data.csv ）をデータフレーム形式で読み込んむ。ヘッダーは無しとする。
9	df = pd.read_csv ('drive / My Drive / data.csv' , header = None)
10	#行数の取得（結局使わないが勉強のため書いておく）
11	a = len (df)
12	#列数の取得
13	b = len (df.columns)
14	#データ処理～ジニ係数の計算までをデータの列数まで繰り返す
15	for i in range (0, b) :
16	df_i = df.iloc [: , i] #1列ずつ順番に取り出す。列の数まで繰り返す。
17	df_i2 = df_i.sort_values () #取り出した列のデータを小さい順（昇順）に並べる
18	df2 = df_i2.cumsum () #累積和の計算
19	sum_n = 0 #累積和の負の値部分の和（になる入れ物）を初期化
20	sum_p = 0 #累積和の正の値部分の和（になる入れ物）を初期化
21	df2.dropna (inplace = True) #データなし (NaN) の部分の置き換えを許可
22	df2.fillna (0) # NaNを0に置換（しないと計算結果がNaNになる）
23	#面積の計算として実行しているので結果に影響はない
24	#サンプルサイズまでデータの値を一つずつ確認して正(positive)と負 (negative) に振り分け
25	for j in range (0 , len (df2)) :
26	if df2 [j] <= 0:
27	sum_n = sum_n + df2 [j] #負値の合計
28	else:
29	sum_p = sum_p + df2 [j] #正値の合計
30	# B (面積なので絶対値になおす)
31	B = abs (sum_p)
32	# C (面積なので絶対値になおす)
33	C = abs (sum_n)
34	# Bの面積がゼロか正かで条件分岐 (実質B = 0（ 累積値の最大が負なので正の累積値 ＝ 0 ）の場合の想定)
35	if B <= 0 :
36	mi = abs (min (df2))
37	mx = 0
38	A = 0
39	B = 0
40	else: # Aの面積を (A + B) - Bによって求める、#完全平等線で欠ける分の面積を補填済
41	mi = abs (min (df2))
42	mx = abs (max (df2))
43	A = ((len (df2) * max (df2) * 0.5 + (mx / len (df2) / 2 * len (df2))) - B)
44	# all = A + B + C + D + E
45	all = A + B + (mi * len (df2)) #AとBの面積＋CDEによる四角形の面積（欠ける分の面積を補填済）
46	#拡張ジニ係数の計算、面積の計算なので絶対値 (abs)
47	Gini = (A + C) / all
48	print (' { : . 03f } ' . format (Gini)) #ジニ係数の計算結果を小数点第3位まで出力

スクリプト中の″#″で始まる文章はコメントと見なされ実行されない

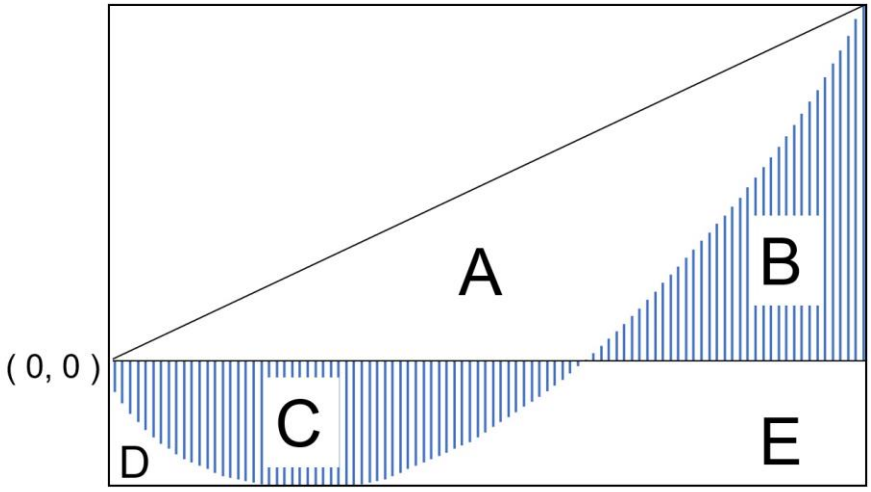

図 1. 拡張ジニ係数計算のために設定した領域名（伊藤ほか（2012）を参考に作成）

*図中の (0 , 0) は原点

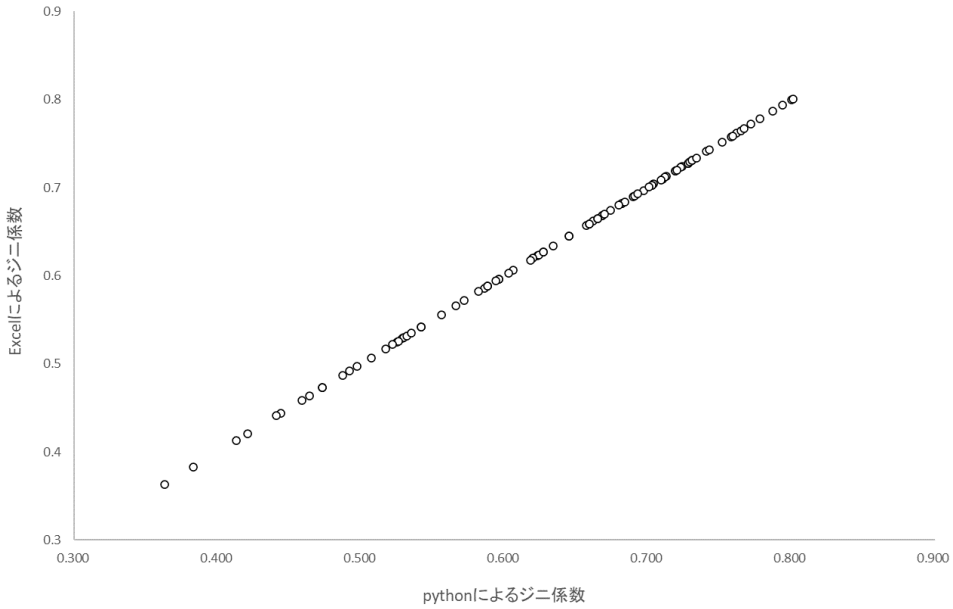

図 2. 同一データから算出した Excel と python による拡張ジニ係数の比較

*縦軸と横軸にそれぞれ Excel と python で算出した拡張ジニ係数を取ってプロットした