姓の多様性

BV19081　鈴木 美喜

2020 年 6 月 12 日

目次

1　はじめに

今回の研究テーマを探しているとき, ベトナムの姓が約 100 種類しかなく, 人口の 40% がグエンさんであることを知った. 日本も結婚による姓の変更などにより姓の種類は減少し続け, いつかはベトナムのように佐藤さんが人口の半数を占めてしまうのではないかと考えた. そこで, まず現在の日本の姓はどのくらい多様なのかを調べるために, 多様度を表す指標を調べてみた.

2　Simpson の多様度指数

2.1　定義

総数 N で K 種類の個体からなる群集の中から, ランダムに二つの個体を取り出したとき, 二つの個体が同じ種類に属するという確率 $\sum \pi^2$ は,

$$\begin{aligned}\sum \pi^2 &= \sum_{i=1}^{K} \frac{n_i(n_i-1)}{N(N-1)} \\ &= \frac{1}{N(N-1)} \left(\sum_{i=1}^{K} n_i^2 - \sum_{i=1}^{K} n_i \right) \\ &= \frac{1}{N(N-1)} \left\{ N^2 \sum_{i=1}^{K} \left(\frac{n_i}{N} \right)^2 - N \sum_{i=1}^{K} \left(\frac{n_i}{N} \right) \right\} \end{aligned} \tag{1}$$

K : 種の総数
n_i : 種 i の個体数
N : 総個体数

と定義される.

定義 2.1. Simpson の多様度指数 D は $1 - \sum \pi^2$, すなわち, ランダムに取り出した 2 個体が異なる種類のものである確率を多様度指数とするものであり, 次のように定義される.

$$D = 1 - \sum_{i=1}^{K} \frac{n_i(n_i-1)}{N(N-1)} \tag{2}$$

Simpson の多様度指数は, 少数の種による独占的傾向が強いほど小さくなる傾向がある. つまり, Simpson の多様度指数が大きいほど「複雑な」群集であり, 小さいほど「単純な」群集であると判断される.

姓を基準とした社会集団に対しては, 次のように考えることができる:集団全体の総人口が N であるとき, 姓の種類に番号をつけ, 姓と番号を 1 対 1 に対応させる. K が姓の種類の総数, n_i が i 番目の姓を持つ人口を表すとすると, ある社会集団において, 現れた姓の種類と,

各姓ごとの人口, 集団全体の総人口を調べれば, その社会集団における姓の多様性を計量化することが可能になる.

2.2 種の均等性

種の均等性, すなわち, ごく少数の種の個体数が群集の中の全個体数の中の大きな割合を占めて, 他の多数種の個体数がそれぞれ大変少ないか, あるいは各種に属する個体数が比較的近いのかの目安は,

$$E = \frac{D}{D_{\max}} \tag{3}$$

で表せる. ここで $D_{\max}$ はすべての種の頻度が等しい場合の多様度指数である. この場合, $\frac{n_i}{N} = \frac{1}{K}$ であるから,(1) より, $D_{\max}$ は

$$D_{\max} = \frac{N}{N-1}\left(1 - \frac{1}{K}\right) \tag{4}$$

である. E が大きいほど群集の均等性は高い.

それぞれの姓を持つ人口が同程度である場合, 非常に均等性は高く, 均等度 E は大きくなる. 逆に E が小さいほど均等性は小さい.

3 Shannon-Weaver 関数

3.1 定義

定義 3.1. 情報理論に基づいて, 不確実さの程度を測るものとして提案された多様度指数が, Shannon-Weaver 関数 H' であり, 次のように定義される:

$$H' = -\sum_{i=1}^{K}\left(\frac{n_i}{N}\right)\left(\log_2 \frac{n_i}{N}\right) \tag{5}$$

Shannon-Weaver 関数 H' が大きい場合, その群集の多様度は大きく, 無作為に選び出す個体がどの種に属するか予測することは困難になる. 逆に, H' の値が小さい場合, ある特定の種を無作為に選び出す確率は高く, その群集の多様度は小さいと判断できる.

姓を基準にした社会集団で考えると, H' が大きい場合, その社会集団における姓の多様度は大きく, 無作為に 1 人を取り出したとき, その人の姓を予測することが難しくなる.

種の豊富さは, 群集の中に現れる種の数 K で示す. つまり, 姓を基準にした社会集団では, 姓の種類 K が大きいほど種が豊富であると判断される. 一方, 種の均等度 J は, H' の最大値 $H'_{\max}$ に対する H' の割合で測ることができる. H' は, すべての種が等しい個体数を持つ場合に最大値 $H'_{\max}$ をとる. 姓を基準にした社会集団の場合では, すべての姓についてそれぞれの姓を持つ人口が相等しい場合である. この最大値 $H'_{\max}$ は次式で与えられる:

$$H'_{\max} = \log_2 K \tag{6}$$

このとき, 均等度 J は次式で与えられるものとする:

$$J = \frac{H'}{H'_{\max}} = \frac{-\sum_{i=1}^{K} \frac{n_i}{N} \log_2 \frac{n_i}{N}}{\log_2 K} \tag{7}$$

$H'_{\max}$ は, サンプルあたりの最大の情報量を表し, 情報容量と呼ばれるものである. 関係式 (6) より, 種の数 K が増加するほど, つまり, 姓の種類が増えるほど対象となる集団のサンプルあたりの情報容量 $H'_{\max}$ は増加する.

4 多様度指数と Zipf の法則

4.1 Zipf の法則

複数の種類からなる群集において, 種類ごとの頻度 (サイズ) と, 頻度から見た順位 (ランク) を考える. このような, 順位データにおける順位と頻度の関係を, ランク-サイズ関係と呼ぶ. 自然現象について観測された様々なランク-サイズ関係で, ある実数 θ が存在して, $(\text{ランク})^{\theta} * \text{サイズ} = \text{定数}$ という関係が成り立つことが報告されてきた. $\theta = 1$ の場合に特に注目して研究した G.K.Zipf にちなんでこの関係はしばしば Zipf の法則と呼ばれる.

姓の分布についてのランク-サイズ関係についても, Zipf の法則が当てはまる場合が報告されている.

姓の種類数と, その姓を名乗る人口が Zipf の法則にしたがうと仮定し, Simpson の多様度指数 D, 均等度 E, Shannon-Weaver 関数 H', 均等度 J を考えてみる. まず, Zipf の法則が成り立つという仮定から, サイズ n_i とランク i の間に次の関係が成り立つ:

$$\frac{n_i}{N} = \frac{C}{i^{\theta}}, C : \text{正定数} \tag{8}$$

関係式 (8) の両辺について, $i = 1, \ldots, K$ で和をとると, 左辺は, 全頻度の和であるから 1 となり, 正定数 C は次のように求められる:

$$C = \frac{1}{\sum_{i=1}^{K} \frac{1}{i^{\theta}}} \tag{9}$$

Simpson の多様度指数 D の定義式 (2) に, (8) を用いると,

$$\begin{aligned} D &= 1 - \frac{N}{N-1} \left\{ \sum_{i=1}^{K} \left(\frac{n_i}{N} \right)^2 - \frac{1}{N} \sum_{i=1}^{K} \frac{n_i}{N} \right\} \\ &= \frac{N}{N-1} \left\{ 1 - \sum_{i=1}^{K} \left(\frac{n_i}{N} \right)^2 \right\} \\ &= \frac{N}{N-1} \left[1 - \frac{\sum_{i=1}^{K} (\frac{1}{i^{\theta}})^2}{\{\sum_{i=1}^{K} \frac{1}{i^{\theta}}\}^2} \right] \end{aligned} \tag{10}$$

となる.

式 (10) において, $\theta \to 0$ とすると,

$$\lim_{\theta \to 0} \sum_{i=1}^{K} \frac{1}{i^{\theta}} = K \tag{11}$$

より,

$$\lim_{\theta \to 0} D = \frac{N}{N-1} \left(1 - \frac{1}{K} \right) \tag{12}$$

を得る. 一方, 式 (10) において, $\theta \to \infty$ とすると,

$$\begin{aligned} \lim_{\theta \to \infty} D &= \frac{N}{N-1} \left[1 - \lim_{\theta \to \infty} \frac{\sum_{i=1}^{K} (\frac{1}{i^{\theta}})^2}{\{\sum_{i=1}^{K} \frac{1}{i^{\theta}}\}^2} \right] \\ &= \frac{N}{N-1} \left[1 - \lim_{\theta \to \infty} \frac{1 + \sum_{i=2}^{K} (\frac{1}{i^{\theta}})^2}{\{1 + \sum_{i=2}^{K} \frac{1}{i^{\theta}}\}^2} \right] \\ &= 0 \end{aligned} \tag{13}$$

を得る.

均等度 E は, 定義式 (3) より, それぞれ

$$\begin{aligned} \lim_{\theta \to 0} E &= 1 \\ \lim_{\theta \to \infty} E &= 0 \end{aligned} \tag{14}$$

となる.

Shannon-Weaver 関数については, Zipf の法則を仮定すると, (5) および (9) より,

$$\begin{aligned} H' &= -\sum_{i=1}^{K} \left(\frac{C}{i^{\theta}} \right) \left(\log_2 \frac{C}{i^{\theta}} \right) \\ &= -\log_2 C + \frac{\theta \sum_{i=1}^{K} \frac{\log_2 i}{i^{\theta}}}{\sum_{i=1}^{K} \frac{1}{i^{\theta}}} \end{aligned} \tag{15}$$

を得る. 先ほどと同じく, $\theta \to 0$ および, $\theta \to \infty$ とした場合の H' を考えると, (15) より,

$$\lim_{\theta \to 0} H' = -\lim_{\theta \to 0} \log_2 \frac{1}{\sum_{i=1}^{K} \frac{1}{i^{\theta}}} = \log_2 K \tag{16}$$

および,

$$\begin{aligned} \lim_{\theta \to \infty} H' = &- \lim_{\theta \to \infty} \log_2 \frac{1}{1 + \sum_{i=2}^{K} (\frac{1}{i^{\theta}})} \\ &+ \lim_{\theta \to \infty} \frac{\frac{\theta \log_2 i}{i^{\theta}} + \sum_{i=2}^{K} \frac{\theta \log_2 i}{i^{\theta}}}{1 + \sum_{i=2}^{K} (\frac{1}{i^{\theta}})} \\ = &0 \end{aligned} \tag{17}$$

となる. さらに, 均等度 J は関係式 (6) および (7) より,

$$\lim_{\theta \to 0} J = \frac{\log_2 K}{\log_2 K} = 1 \tag{18}$$

$$\lim_{\theta \to \infty} J = \frac{0}{\log_2 K} = 0 \tag{19}$$

となる.

4.2 まとめ

以上のことをまとめると, 表 1 が得られる. 姓の種類数とそれぞれの姓を名乗る人口が Zipf の法則に従う場合において, $\theta \to 0$ となるのは, どの姓を名乗る人口も同程度の場合, すなわち, 発現頻度がほぼ等しいような場合である. このとき発現頻度は姓の種類数 K を用いてほぼ $\frac{1}{K}$ となる. どの姓の発現頻度も等しいので , そのような社会集団において姓の均等度は最大である.

一方, 同じ場合において, $\theta \to \infty$ となるのは, ある姓を名乗る人口が圧倒的に多く, その他の姓を名乗る人口が非常に少ない場合である . このような社会集団においては, 1 位の姓の発現頻度は大変大きいが, その他の姓の発現頻度が非常に低く, 姓の均等性が小さくなる.

表 1　$\theta \to 0, \infty$ の場合の多様度指数および均等度

指数 \ θ	0	∞
D	$\frac{N}{N-1}(1-\frac{1}{K})$	0
E	1	0
H'	$\log_2 K$	0
J	1	0

5 今後の目標

以上のような多様性の指標を使って日本の姓の多様性を評価するためには, 姓の種類数とそれぞれの姓を名乗る人口を調べる必要がある. しかし, 現在の日本ではそれらの正確なデータが存在しない. そのため, 存在するどのデータを使って推定するかを検討する必要がある.

また, 参考文献では, 姓の多様性の評価を行った後, Galton-Watson 過程という数理モデルを使って, 姓が絶滅していく中でどのように多様性を保つかなどの考察もされていた.

次の研究ではこれらを検討, 考察し, 日本の姓の多様性について深く考えたい.

参考文献

[1] 佐藤葉子・瀬野裕美著, 姓の継承と絶滅の数理生態学-Galton-Watson 分枝過程によるモデル解析, 京都大学学術出版会, 2003 年