語彙数の推定問題

長谷川　　環（指導教員　杉浦成昭）

【序】

語彙統計学は作者・時代の識別，著者の真偽判定などにおいて統計分野に限られずより広域な研究分野である。Shakespeare，紫式部など語彙数の多いと思われる作者に注目し作品外にさらに出現する可能性のある単語数の推定を試みた。また，長編小説James Joyceの作品“Ulysses”一作品についても単語数の推定を行った。次に，作者がShakespeareか否か不明な詩（＝Taylor poem）に対して作者がShakespeareかどうか判定する検定を行った。さらに，語彙統計学の方法で大鱗翅類昆虫の度数を推定することを考えた。

【方法】

(1) Efron and Thisted (1976) の論文の模型を修正し，先験分布の特異点を除去し，先験分布が発散しないような閾値模型を考え，点推定を行った。閾値模型以外にも，負二項模型，対数級数分布模型と比較し赤池情報量基準（AIC）によるあてはまりのよさの比較，また各模型の特徴をつかみたくピアソン図形を描いた。(2) 源氏物語，“Ulysses”のデータベースを元に (1) 同様，閾値模型をあてはめ点推定を行った。(3) Thisted and Efron (1987) の論文を元にTaylor poemに対して作者がShakespeareかどうか，Shakespeareの作品，同時代に書かれた作者の異なる作品と比較し仮説検定を行った。(4) 捕獲された大鱗翅類昆虫の度数に上記にある3つの模型をあてはめ，AICによりあてはまりのよさを判定した。

【結果】

(1) AICで比較した結果，閾値模型，負二項模型のあてはまりが対数級数分布模型に比べて極めてよく，ともにほぼ同じ値となったが，先験分布の形状母数の最尤推定値が負になることを考慮すると閾値模型のみ点推定が可能であり，Shakespeareの知っていたけれども作品中に使用されなかった単語は約143,000語，漸近的95％信頼区間は（73,000, 213,000）であることがわかった。(2) 紫式部がさらに知っていた単語数は約31,000語，95％信頼区間は（21,000, 41,000），Joyceがさらに知っていた単語数は約80,000語，95％信頼区間は（66,000, 94,000）であることがわかった。(3) Taylor poemに対して行った仮説検定においては，Taylor poemがShakespeareの作品と似た特徴を示し，統計的には作者がShakespeareであると考えてほぼ間違いがないという結論に達した。(4) 大鱗翅類昆虫の度数に対しては3つの模型ともにAICの値にほとんど差がなくどの模型もあてはまりが良いことがわかった。昆虫のデータでは先験分布の形状母数の最尤推定値が正になり先験分布は発散しない。対象が語彙と昆虫ではデータが質的に異なると思われる。

【考察】

ガンマーポアソン分布からなる負二項模型，閾値模型に対して逆ガウシアンーポアソン模型のあてはまりを検討する必要がある。

【参考文献】

[1] Efron, B. and Thisted, R. (1976) Estimating the number of unseen species: How many words did Shakespeare know？ *Biometrika*, **63**, 433-447.

[2] Thisted, R. and Efron, B. (1987) Did Shakespeare write a newly-discovered poem？ *Biometrika*, **74**, 445-455.

[3] 杉浦成昭・長谷川環 (2001) Revisit to “How many words did Shakespeare know”, 第69回日本統計学会講演論文集, 110-111.

[4] N. Sugiura and T. Hasegawa (2002) Estimating the number of unseen species that Shakespeare knew, Compstat 2002.

[5] Sichel, H.S. (1975) On a distribution law for word frequencies *J. of American Statistical Association* **70**, 542-547.