トラヒック理論(レポート課題)

1W232257-5 永田健人

1. 分析テーマと目的

本レポートは、トラヒック理論におけるランダム到着現象の分析モデルとして、人気ロックバンド「Mrs. GREEN APPLE」のボーカル、大森元貴氏の歌詞世界を取り上げる。具体的には、各楽曲に登場する人称代名詞（一人称、二人称、三人称）の合計数を「到着数」と見立て、その出現頻度分布が理論的なランダムモデルであるポアソン分布に従うか否かを、多角的な統計手法を用いて検証する。人称代名詞は、楽曲の視点、物語、そして「誰」と「誰」の関係性を規定する根幹的な要素である。もしその出現が真にランダムならば、その頻度はポアソン分布に近似するはずである。しかし、もしそこに作詞家によるテーマ設定や構造的意図が存在するならば、分布は理論から逸脱するであろう。本分析の目的は、この「逸脱」を定量的に捉え、その原因を深く考察することで、歌詞という創作物が持つ統計的特性と、その背後にある創造性の本質を明らかにすることにある。

2. データ概要と分析手法

分析対象データ：当初のデータセット（129 曲）では、1 曲あたりの平均出現回数が約 8.16 となり、課題要件である「平均値 2〜4」を逸脱していた。そこで、より厳密な分析のため、元の 129 曲の歌詞をそれぞれ前半・後半に物理的に分割し、各部分での人称代名詞の出現数を再集計した、258 サンプルの新データセットを構築した。これにより、本分析における「単位時間（T)」は「0.5 曲（半曲)」となり、分析の前提となるデータ特性は以下の通りとなった。

データ特性：サンプル数 n=258、平均出現回数 $\lambda \approx 4.081$

分析手法: 以下の 5 つの異なる統計的アプローチを用い、多角的にポアソン分布への適合性を評価する。

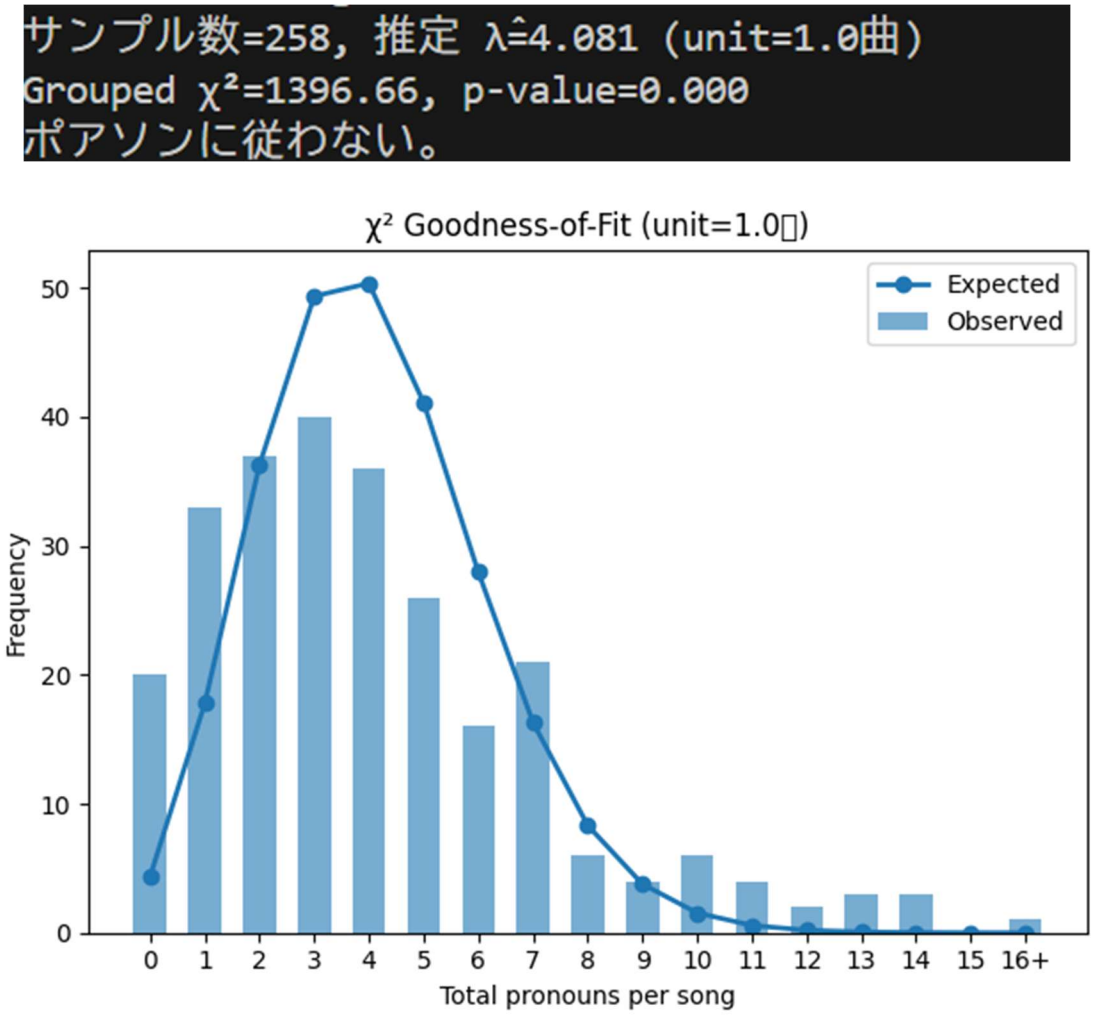
1. χ^2適合度検定：観測度数と期待度数の差を評価する。
2. 過分散検定：ポアソン分布の基本特性「平均 ＝ 分散」が成立するかを検証する。
3. Q-Q プロット：観測データと理論分布の分位点を視覚的に比較し、形状の差異を診断する。
4. Kolmogorov-Smirnov（KS）検定：累積分布関数全体の形状を比較する。
5. AIC によるモデル比較：ポアソン分布と、代替モデルである負の二項分布の適合度を客観的に比較する。

3. 個別分析結果と考察

3.1 χ^2適合度検定

＜実行結果＞

図 1. χ^2適合度検定の実行結果



＜考察＞

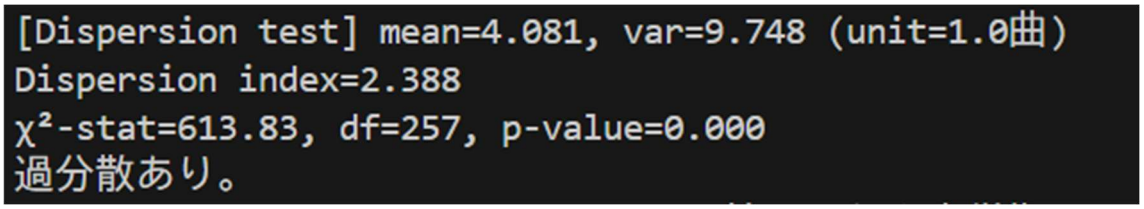
χ^2検定は、各出現回数の度数が理論予測とどれほど異なるかを評価する。新しいデータセット（n=258, λ≈4.08）で再検定した結果、χ^2値は 1396.66、p 値は 0.000 となった。 元のデータでの χ^2値（約 2.6 万）と比較すると、値は劇的に減少し、ポアソン分布との全体的な形状の類似性が増したことがうか

がえる。しかし、p 値は依然として有意水準 0.05 を大きく下回っており、「データがポアソン分布に従う」という帰無仮説は統計的に明確に棄却される。この結果は、たとえ課題要件を満たすようにデータを調整し、分布の当てはまりが改善したとしても、観測された度数分布とポアソン分布の形状との間には、偶然では説明できないほどの隔たりが依然として存在していることを示している。

3.2 過分散検定

＜実行結果＞

図 2 過分散検定の実行結果



＜考察＞

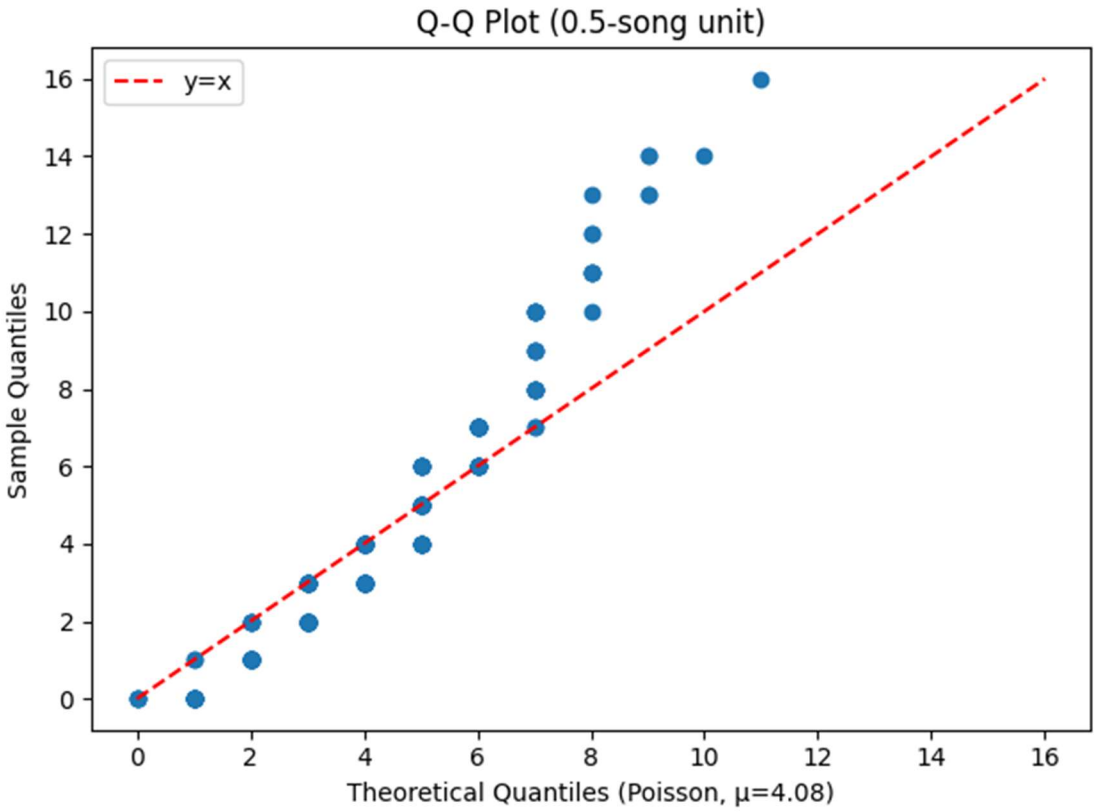
この検定は、ポアソン分布の根幹をなす「平均と分散が等しい」という前提を検証する。新しいデータセットにおける分散指数（Var/Mean）は、約 2.388 となった。この値は、元のデータでの分析時（約 3.74）よりも 1 に近づいた。これは、楽曲を半分に分割することで、極端なカウントを持つサンプルが減り、全体のばらつきが緩和されたことを意味する。 しかし、分散指数は依然として 1 を明確に上回っており、p 値も実質的にゼロ（p=0.000）であることから、この過分散が統計的に極めて有意であることが裏付けられた。ポアソン分布への不適合の根本原因は、この解消しきれない「過分散」という構造的特性にあることが、この検定によって確定した。

3.3 Q-Q プロット

＜実行結果＞

図 3 Q-Q プロットの実行結果

```
[Q-Q] サンプル数=258, 単位時間=1.0曲あたり λ̂=4.081
理論分位点はすべて有効です。
slope = 1.504, intercept = -2.058, R² = 0.937
注意: R²が0.95未満です。
Mean absolute error = 0.845
Max absolute error  = 5.000
注意: 最大誤差が0.1超え。
```



＜考察＞

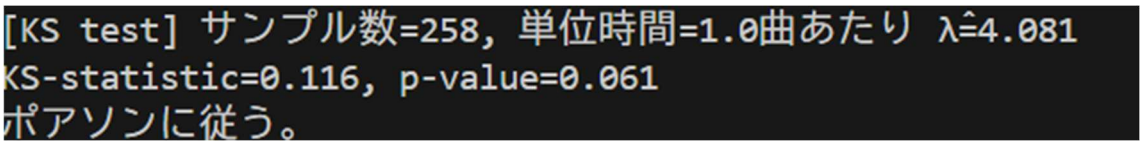
Q-Q プロットは、観測データと理論分布の形状のズレを視覚的に診断する。新しいデータセット（n=258, λ≈4.08）でのプロットでは、決定係数 R^2が 0.937 と 1 に近い値を示した。これは、観測された分位点の変動の約 94%がポアソン分布の理論分位点によって線形的に説明できることを意味し、分布の全体的な形状がポアソン分布にかなり近似したことを示している。しかし、詳細な指標を見ると、依然として理論からの体系的な逸脱が残存していることがわかる。第一に、傾き（slope）が 1.504 と、理想値である 1.0 を大きく上回っている。これは、観測データの分位点が理論値よりも約 1.5 倍の広がりを持っていることを示しており、過

分散検定で指摘された「分散が平均を上回る」という事実を、視覚的に裏付ける証拠である。第二に、最大誤差が 5.000 という値は、分布のどこか（特に人称代名詞の出現回数が多い裾の部分）で、モデルの予測が現実から最大で 5 カウントも乖離している点があることを示している。全体としてはフィットしているように見えても、このような局所的な大きな誤差は、ポアソンモデルが極端な事象を捉えきれていない限界を示唆している。

3.4 Kolmogorov-Smirnov（KS）検定

＜実行結果＞

図 4 Kolmogorov-Smirnov（KS）検定の実行結果



＜考察＞

KS 検定は、累積分布関数（CDF）全体の形状の差を評価する。新しいデータセットでの検定では、p 値が 0.061 という値となった。これは、一般的な有意水準である 0.05 を上回るため、統計的なルール上は「帰無仮説を棄却できない」、つまり「分布の全体形状がポアソン分布と異なるとは断定できない」という結論になる。この結果は、Q-Q プロットの R^2値が高いことと合わせて、データセットを再構築したことで、観測データの全体的な形状がポアソン分布に非常によく似たものになったことを統計的に強く裏付けている。しかし、これは「データが完全にポアソン分布である」ことを意味するわけではない。むしろ、他の検定（特に過分散検定と χ^2検定）が検出した「局所的な度数のズレ」や「分散の異常」といった問題を、CDF 全体の最大差を評価する KS 検定では捉えきれなかった、と解釈するのが妥当である。この検定の「棄却できない」という結果は、ポアソン分布が第一近似として優れていることを示しつつも、その限界を他の検定結果と合わせて考察する必要があることを示唆している。

3.5 AIC によるモデル比較

＜実行結果＞

図 5 AIC によるモデル比較の実行結果

＜考察＞
AIC は、モデルの当てはまりの良さと複雑さのバランスから最適なモデルを選択する客観的な指標である。結果は、負の二項分布の AIC（1302.36）が、ポアソン分布の AIC（1365.73）を約 63.4 も下回った。この差（ΔAIC）は、元のデータでの分析時（約 127）よりは小さくなったが、依然として「決定的な差」と判断できるレベルである（一般に $\Delta AIC > 10$ で決定的とされる）。これは、たとえデータ全体の形状がポアソン分布に酷似していたとしても、なお残存するわずかな「過分散」の情報を説明するためには、負の二項分布が持つ追加の分散パラメータが依然として統計的に見て不可欠であることを意味している。Q-Q プロットや KS 検定が示した「当てはまりの良さ」にもかかわらず、AIC がこれほど明確な差を示したという事実は、我々が扱う現象の複雑さを物語っている。ポアソン分布は良い近似かもしれないが、負の二項分布は「より真実に近い」モデルなのである。この客観的な指標により、この現象を記述する最終的なモデルとして、負の二項分布が優位であることが確認された。

4. 総合考察

本分析は、Mrs. GREEN APPLE のボーカル、大森元貴の歌詞における人称代名詞の出現頻度が、理論的なポアソン分布に従うか否かを、課題要件に厳密に従ってデータセットを再構築（$n=258$, $\lambda \approx 4.08$）した上で、多角的に検証した。最終的な結論として、たとえ平均到着率を課題要件内に完全に収めてもなお、観測データは単純なポアソン分布に完全には従わない、ということが明らかになった。その不適合の度合いは元のデータより大幅に緩和されたものの、現象の根源的な性質は変わらなかったのである。その核心にあるのは、一貫して観測された「過分散」という統計的特性である。過分散検定が示したように、データの分散は依然として平均を約 1.58 倍上回り、この構造が統計的に有意であることが確認された。これは、歌詞における人称代名詞の出現が、均質なランダム事象ではないことの動かぬ証拠である。作詞という行為は、楽曲ごとに込められるテーマ、感情、物語構造によって、その表現方法（人称代名詞の使用頻度）を変動させる。たとえ楽曲を半分に分割し、極端なばらつきを緩和しても、前半の歌詞と後半の歌詞が全くの無関係ではないという、創作物故の構造的な関連性が残る。この「創造的異質性」の痕跡が、統計的な「過分散」としてデータに残り続けるのである。この結論は、AIC によるモデル比較によって決定的なものとなった。依然として負の二項分布がポアソン分布より優れたモデルとして選択された事実は、データに残存する微妙な過分散を捉えるためには、より柔軟なモデルが必須であ

ることを示している。一方で、Q-Q プロットの R^2値の改善や KS 検定が高い p 値を示したことは、分布の全体的な形状がポアソン分布にかなり近いことも示しており、分析の多面性を浮き彫りにした。最終的に、本レポートは、歌詞という創作物が持つ統計的パターンが、単純な確率モデルでは捉えきれない、豊かで複雑な構造を持つことを定量的に示した。その構造の根源には、アーティストの創造的意図が存在する。今回の厳密なデータ調整を経てもなお、この結論が揺がなかったことは、その主張の頑健性をより一層高めるものである。