**提出日：2022年1月4日(火)**

# コレスポンデンス分析（レポート課題6）

- 環境情報学部１年　学籍番号：72145163　中村蒼 -

**2021年秋学期[DS2]【学期後半】**
**「ビジネスのためのデータサイエンス」**
**【火曜1,2限】**

## 目次



## １．はじめに

本レポートでは、授業で事前に配布された「BJData_univ2.xls」をcsvファイルに変換し、R言語で読み込みデータにおける相関比最大や相関係数、情報量などの数値と散布図などグラフなどでデータを可視化させ、比較することでコレスポンデンス分析を行う。各大学の分析を行なったのち、相関比最大、相関係数や情報量などの数値がどのようにして散布図などのグラフと対応しているか、その関係性についても今回のレポートを通して分析し、最終レポートに向けて理解を進めたい。

## ２．統計とデータ

・「BJData_univ2.xls」- 各大学のイメージ調査

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **京都大学** | 331 | 194 | 105 | 26 | 7 | 18 | 72 | 2 | 25 | 66 | 55 | 133 | 41 | 56 | 13 | 13 |
| **慶應義塾大学** | 346 | 183 | 66 | 38 | 10 | 28 | 40 | 2 | 18 | 55 | 35 | 111 | 54 | 42 | 13 | 7 |
| **東京大学** | 347 | 172 | 66 | 34 | 5 | 9 | 37 | 4 | 20 | 46 | 49 | 122 | 21 | 37 | 11 | 8 |
| **一橋大学** | 320 | 127 | 49 | 19 | 3 | 10 | 25 | 2 | 13 | 21 | 21 | 81 | 20 | 10 | 2 | 1 |
| **早稲田大学** | 351 | 169 | 55 | 43 | 15 | 28 | 22 | 1 | 14 | 39 | 31 | 88 | 23 | 21 | 14 | 8 |

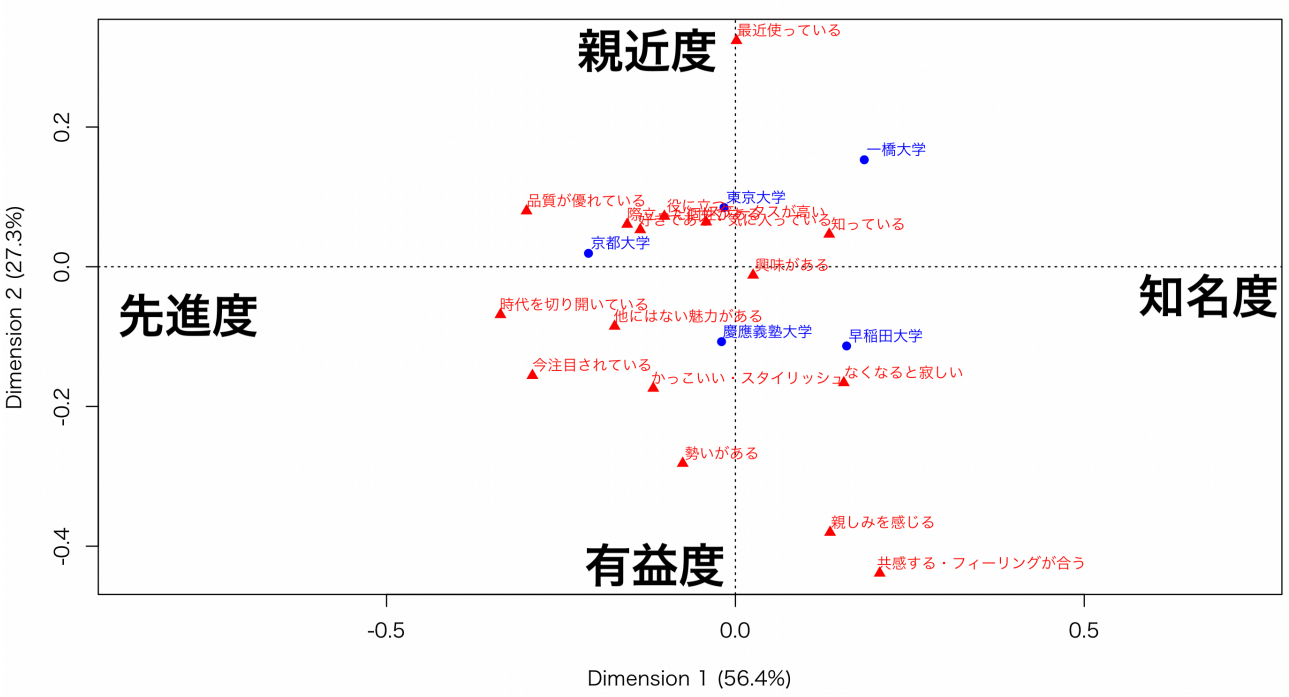
1. 知っている
2. 興味がある
3. 好きである・気に入っている
4. なくなると寂しい
5. 共感する・フィーリングが合う
6. 親しみを感じる
7. 品質が優れている
8. 最近使っている
9. 役に立つ

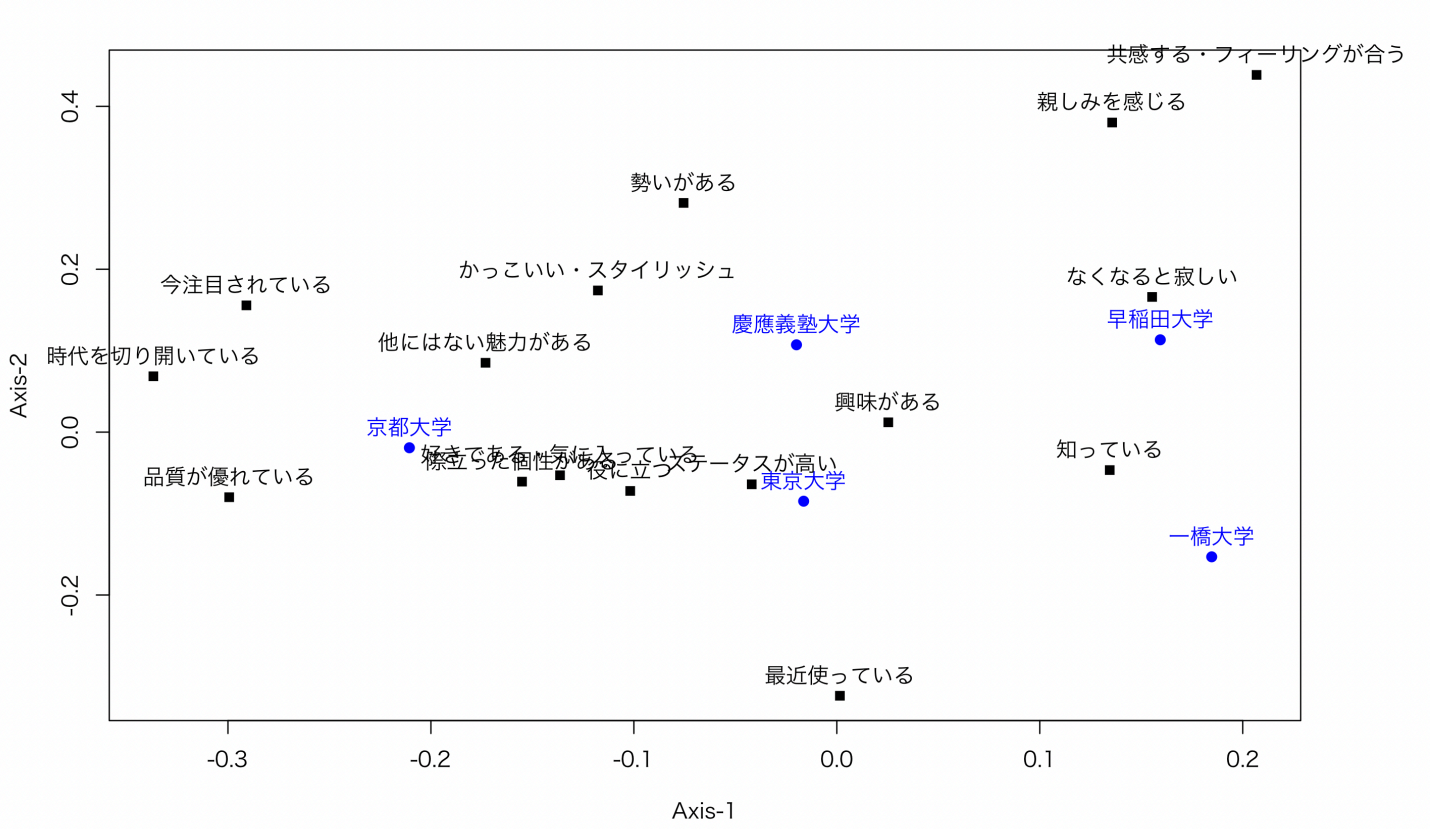
10. 他にはない魅力がある
11. 際立った個性がある
12. ステータスが高い
13. かっこいい・スタイリッシュ
14. 時代を切り開いている
15. 勢いがある
16. 今注目されている

・データにおける相関関係や情報量

| | 固有値1 | 固有値2 | 固有値3 | 固有値4 |
|---|---|---|---|---|
| **相関比最大** | 0.021 | 0.010 | 0.004 | 0.002 |
| **相関係数** | 0.144 | 0.100 | 0.065 | 0.042 |
| **寄与度（情報量）** | 56.375 | 27.263 | 11.465 | 4.898 |
| **累積寄与率** | 56.375 | 83.638 | 95.102 | 100 |

・コレスポンデンス分析






## 3．コレスポンデンス分析における縦軸と横軸の解釈について

※第一象限と第二象限の間の線を上線、第二象限と第三象限の間の線を左線、第三章限と第四章限の間の線を下線、第四象限と第一象限の間の線を右線とする。

- 親近度：固有値2において最も数値が高い選択肢が「共感する・フィーリングが合う」であり、次に数値の高い選択肢が「親しみを感じる」であることから、上線は親近度の度合いと定義することとする。
- 先進度：固有値1において最も数値の低い選択肢は順番に「時代を切り開いている」、「品質が優れている」、「今注目されている」の三つである。大学の「品質が優れている」とは少し選択肢として不明瞭であるため、今回は大学の研究における結果や論文の「品質が優れている」と定義することとする。また、同様に研究の領域において「時代を切り開いている」「今注目されている」と定義するなら、左線は先進度の度合いと定義することができる。
- 有益度：固有値2において最も数値の少ない選択肢が「最近使っている」である。大学において「最近使っている」とは広義であるためなかなか定義し辛いが、今回は論文など研究において「最近使っている」と定義することにする。また、固有値2が次に低い選択肢は「役に立つ」であり、こちらも学習などで大学の資料や研究が「役に立つ」と定義することで下線を有益度の度合いと解釈することができる。
- 知名度：固有値1において最も数値の高い選択肢は順番に「共感する・フィーリングが合う」「なくなると寂しい」「親しみを感じる」「知っている」である。4番目の「知っている」は他の三つの選択肢と関連性が高いことから、右線を知名度の度合いと定義することとする。

## ４．結果と考察

### ４.１．全体の考察

まず、コレスポンデンス分析のプロットよりどの大学にも共通している点として「共感する・フィーリングが合う」と感じる回答者は少なかったとわかる。また、同様に「親しみを感じる」や「最近使っている」、「時代を切り開いている」なども回答者が少なかった。一方、どの大学にも共通して「興味がある」と回答する人は多く、特に慶應義塾大学と東京大学が高いということが読み取れる。他にも、慶應義塾大学や早稲田大学は他の国立大学と比較すると「かっこいい・スタイリッシュ」と相関性が高い。また、近年ノーベル賞受賞者を多数輩出している京都大学では「時代を切り開いている」、「際立った個性がある」などの選択肢との関連性が高いことが伺える。次に、各大学について考察をしていきたい。

### ４.２．各大学における考察

京都大学

京都大学において最も特出すべき点は「品質が優れている」、「際立った個性がある」、「好きである、興味がある」、「他にはない魅力がある」、「時代を切り開いている」など多くの選択肢において高い数値が算出されるという点である。前述した通り、近年京都大学は他の大学と比較するとノーベル賞受賞者が多く、研究のレベルも非常に高い。研究機関や施設が整っていることや優秀な教授や研究者、学生が会合することからこういった数値が出ることが考えられる。一方で、「共感する・フィーリングが合う」や「親しみを感じる」などの質問に対しては数値が比較的低かったことがわかる。同様に、東京大学と一橋大学も二つの質問に対して相関性が低いことから回答者は私立大学に通う大学が多い傾向にあることがわかる。

慶應義塾大学

他の四つの大学と比較すると、慶應義塾大学は「かっこいい・スタイリッシュ」において最も高い相関性があることがわかる。また、「知っている」は東京大学と、「親しみを感じる」は早稲田大学と同程度の数値を持っている。しかし、「かっこいい・スタイリッシュ」の質問を除いて、特出して高い数値を持つ質問がなく特徴がコレスポンデンス分析を通してあまりないということが読み取れる。京都大学と比較すると研究などは劣ってしまうことから「品質が優れている」や「際立った個性がある」なども低く、特にこれと言ったニュースがないことから「勢いがある」や「今注目されている」も高くない。よって、散布図から慶應義塾大学は特徴をもっと増やすことが求められていることが考えられる。

東京大学

東京大学は京都大学と「際立った魅力がある」、「好きである・興味がある」や「役に立つ」、「ステータスが高い」など数値が類似している点が多い。実際、東京大学と京都大学はともに日本国内の偏差値がもっとも高い国立大学であることから、類似点は多いことは予想できる。一方で、東京大学と京都大学の相違点は「時代を切り開いている」や「今注目されている」の数値が京都大学の方が高いというところである。その理由として、直近のノーベル賞受賞者が京都大学の方が多く輩出しているからであると私は考える。実際、東京大学のノーベル賞受賞者は2016年のオートファジーの研究をした大隈教授が最後であり、ノーベル賞受賞者を輩出して5年経過している。一方で、京都大学では2018年に免疫チェックポイント阻害因子を発見した本庶教授、2019年にリチウムイオン二次電池の開発を行った吉野教授など3年以内に二人のノーベル賞受賞者がいるため、話題性は京都大学の方が高いと考える。「際立った魅力がある」や「ステータスが高い」などの印象があることから、より特徴を伸ばし研究に尽力することが求められていることが分析の結果より読み取ることができる。

一橋大学

一橋大学は他の四つの大学と比較すると多くの質問において数値が低いことがわかる。特に、「親しみを感じる」、「勢いがある」、「今注目されている」などの質問に対してはコレスポンデンス分析を見ると対局の場所に位置していることが読み取れる。また、差が激しくはないものの、「知っている」や「興味がある」も五つの大学のうち最も低い数値である。このことから、他の大学と比べると知名度が低いように感じる。総合大学ではないため学生数の母数も少なく、研究も他の大学と比較すると有名度が低いことが知名度の低さの要因ではないかと考える。

早稲田大学

散布図を見ると、早稲田大学は他の大学と比較したとき相関関係の高い選択肢は「なくなると寂しい」である。また離れてはいるが、他の大学と比較すると「共感する・フィーリングが合う」が比較的相関性が高かった。一方で、「品質が優れている」は他の大学と比較すると圧倒的に数値が低く、また「好きである・気に入っている」や「ステータスが高い」などの選択肢も数値が低い。このことから、回答者は全体として早稲田大学に対して良いイメージを持ってない。「他にはない魅力がある」や「時代を切り開いている」の数値の低さから一橋大学と傾向が似ているとも読み取れるが、二つの大学の大きな違いは「なくなったら寂しい」や「親しみを感じる」の二つの質問による数値の違いである。したがって、早稲田大学がなくなったら寂しく、親しみを感じるという点から回答者の大多数が慶應義塾大学の学生であると私は予想する。

## 5．総括

今回のコレスポンデンス分析を通して、数値だけでは読み取ることのできない各大学の特徴や特出すべき点、また改善点など数多くの要素を解析することができた。また、コレスポンデンス分析における散布図があることで特徴の似ている大学やそれぞれの性質、また各大学の差なども瞬時に解釈することができるため、各項目の特徴を捉えることに特化したツールであることを今回のレポートを通して感じることができた。

しかし、一方で今回の大学の収集された情報において散布図の情報自体少し読み取りにくさも感じた。まず、「時代を切り開いている」と似た意味を持つ「勢いがある」はグラフ上で似た場所に位置していないことが不思議である。集計の意味が明白でないことから、大学のどういった分野に着目して回答者が「勢いがある」や「今注目されている」の評価をすれば良いか理解していないことが統計にばらつきが激しい理由であると思われる。例えば、大学野球に着目すると慶應義塾大学や早稲田大学は一橋大学や東京大学、京都大学比較すると注目されているが、ノーベル賞受賞者という分野に注目すると慶應義塾大学や早稲田大学と比べると京都大学や東京大学の方が注目されている。このように、データを集計する際、どういった分野に着目し、何を主目的としてデータを収集しているかを明確化しなければ集まったデータは目的に沿った結果とならないケースとなってしまう可能性がある。また、「最近使っている」や「役に立つ」の定義が曖昧であるなど、選択肢の文言を見直す必要があるように思えた。大学を「使う」という選択肢は、人によって大学が出している論文を「最近使ってる」や大学の最寄り駅を「最近使っている」など様々な解釈ができるため、もう少し具体的な選択肢を設定しなければ有意義なデータを収集することができないことを今回のデータのコレスポンデンス分析を元に理解することができた。

## 6．プログラム

report6.R*

Source on Save　Run

```r
source("http://aoki2.si.gunma-u.ac.jp/R/src/all.R", encoding="euc-jp")

setwd("~/Desktop/授業/1年　秋学期/ビジネスのためのデータサイエンス/4/課題/レポート課題6")
univ <- as.matrix(read.table("BJdata_univ.csv", header=FALSE, sep=",", nrows=-1))



# ラベルづけ
colnames(univ) <-c("知っている", "興味がある", "好きである・気に入っている", "なくなると寂しい", "共感する・フィーリングが合う", "親しみを感じる",
                   "品質が優れている", "最近使っている", "役に立つ", "他にはない魅力がある", "際立った個性がある",
                   "ステータスが高い", "かっこいい・スタイリッシュ", "時代を切り開いている", "勢いがある", "今注目されている")
rownames(univ) <- c("京都大学", "慶應義塾大学", "東京大学", "一橋大学", "早稲田大学")

univ

# コレスポンデンス分析（双対尺度法）の実行
ans <- dual(univ)
summary(ans)
summary(ans, weighted=TRUE)
plot(ans,1,2)


# コレスポンデンス分析には、関数 ca もスグレモノ。　ただし、パッケージ ca と rgl をインストールする必要あり。
install.packages("ca", dependencies=TRUE)
library(ca)
ca(univ)
plot(ca(univ))
```