論　文　*Article*

インターンシップおよびオープンキャンパステキストデータの統計的分析

原稿受付　2024 年 7 月 31 日
ものつくり大学紀要　第 14 号　(2024)　11～16

佐久田茂[*1]、会田将貴[*2]

*1 ものつくり大学　技能工芸学部　情報メカトロニクス学科
*2 ものつくり大学　技能工芸学部　総合機械学科　卒業生

概要　インターンシップ，およびオープンキャンパスに関するアンケートデータの統計的分析を行なった．インターンシップ評価票中の「研修学生の実習態度や仕事（課題）への取り組み方などお気付きの点をお聞かせください」に関して非構造テキスト文を英訳してデータを作成後、統計分析ソフト JMP のテキストエクスプローラ機能を用いて感情分析を中心に統計分析を実施した。オープンキャンパスアンケート（邦文）は，来場者のオープンキャンパス体験に関する感想を求めたテキスト文の特異値分解・トピック分析を行ない，来場者の満足度が低い傾向にある意見を客観的・定量的に抽出する見通しを得ることができた．

キーワード：情報，統計，分析，テキスト文，非構造化，アンケート

Statistical analysis of unstructured text data in internship and open campus

Shigeru SAKUTA and Masaki AIDA

Dept. of Information Science and Mechatronics Technologists, Institute of Technologists

Abstract　There are many questionnaires collected from corporations regarding internship or high school students regarding open campus in the Institute of Technologists. Unfortunately, they have not been analyzed objectively and quantitatively and actually they have not been made the most use of. JMP, statistical discovery software, deals with text data by “Text Explorer” function based on a kind of statistics principal component analysis. In this paper, four kinds of visual statistical data analyses, that is, term/phrase analysis, sentiment analysis, singular value decomposition and topic analysis have been discussed from viewpoint of its feasibility. The useful information on internship and open campus has been shown for more meaningful internship as well as open campus with more students.

Key Words : statistical analysis, text, questionnaire, information science

1. はじめに

現状，オープンキャンパス，インターンシップ評価票やインターンシップ報告会，企業説明会などでの来場者・企業様からの各種アンケート，特に自由記述欄データが単年毎での主観的・定性的総括にしか活かされていない．統計ソフトウエア JMP（JMP16, SAS Institute Japan 株式会社）では非構造化テキストデータ，すなわちアンケートの自由記述欄などの自由フォーマットテキスト文の統計的分析が可能である（一部の機能は英文に限られる）．前報 [1)]を受けて今回，インターンシップアンケートの層別分析 [2)]，および少子化の時勢の中，現在の大学の喫緊の課題である新入生募集に密接な関連があると思われるオープンキャンパス来場者アンケートの分析を試みた．

2. 分析方法

2.1 インターンシップアンケート

自由フォーマットテキスト文を JMP16 のテキストエクスプローラ機能にかけて分析した．テキストエクスプローラは自然言語処理（NLP）中の特異値分解アルゴリズムを使ってテキストを数値データに変換，そのデータを統計分析にかけている [3]．分析対象は，インターンシップ評価票での「研修学生の実習態度や仕事（課題）への取り組み方などお気付きの点をお聞かせください」における，企業様からの回答テキスト文とした．理由は，甘口・辛口様々な忌憚のない意見が一番期待できると考えたからである．本報では以下の年度・学科について分析を行なった．

・2021 年度製造系(2 年生，3 年生)
・2021 年度建設系(2 年生，3 年生)
・2022 年度製造系(2 年生)

企業様からのコメントから 1 行に 1 文としたデータテーブルを作成した．JMP データテーブルにおけるテキスト列の 1 セルが 1 文書に対応している．分析データ数は 1797 である．

なおテキストエクスプローラ中の一部の分析の対象言語が英語のみでの対応となっているので(2024/6 時点)，今回はアンケート文を英訳して分析にかけた．表 1 にデータテーブル（抜粋）を示す．なおアンケート文の英文への翻訳は全て著者が行なった．

2.2 オープンキャンパスアンケート

少子化時勢の中，学生募集は喫緊の課題である．

Table. 1 Data sample

No.	Year	Department	Grade	Comments
1	2021	Manufactur-ing	Sopho-more	OK for manners as trainee. Positive thinking for Internship and for oneself were seen sometimes.
2	2021	Manufactur-ing	Sopho-more	We can make him feel the principle of manufacturing through all processes in our plant.
3	2021	Manufactur-ing	Sopho-more	We hope this experience makes him grow up considering his future dream as a employee.
4	2021	Manufactur-ing	Sopho-more	Positive thinking towards Internship and for oneself positive could be seen to complete his work.
5	2021	Manufactur-ing	Sopho-more	Thanks to his curiosity he could understand immersion technique, a kind of special one for engineer.
6	2021	Manufactur-ing	Sopho-more	We believe he will be a good engineer If he finds out his goal and dream.

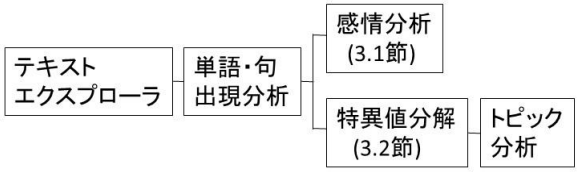


Fig 1 Layer diagram of analysis

2023 年度オープンキャンパス来場者アンケートを客観的かつ定量的に分析し，入学者増に結び付ける．オープンキャンパスアンケート分析は日本語版で行なった．これはデータ整理のワーク量を含め英語版との比較を図ったためである．分析データ数は 907 である．本報で使用する各種分析手法の階層図を図 1 に示す．感情分析と特異値分析・トピック分析は，単語・句出現分析後に階層が分かれる形になる．

3. 分析結果

3.1 インターンシップアンケートの傾向分析

図 2 はインターンシップアンケートの単語と句の出現結果である．度数分布表の形でアンケート文中での出現単語・出現句の傾向を見ることができる．出現単語・出現句の傾向を見ながら，ケースに応じて文書間の違いを読み取るのに役立たない単語を分析から除いていく．具体的には単語を選択して右クリックして，“ストップワード”に追加する．例えば，旅客機の墜落原因アンケートでは，「pilot」を含む単語は他の単語より頻繁に出現してくるが，文書間の違いを読み取るためには役立たないため，ストップワードに指定して分析から除く [4]．本報ではストップワード無しでの分析で不具合が見られなかったので，ストップワードを特に指定せずに分析を進めた．インターンシップアンケート分析は前報 [1]で基本的分析手順は確認できていたので，以下に示す層別に感情分析を行なった．

テキストエクスプローラでアンケート文中での出現単語・出現句の傾向を把握した後，感情分析に移ることができる．感情分析は，辞書に基づいて文書内の感情語を特定し，それらの語に対して，肯定的・否定的・全体的なスコアを付けることが

単語と句のリスト

単語	度数	句	度数
work	211	good engineer	24
good	121	think he should	21
think	88	asked questions	18
ask	68	very much	18
Internship	66	so i think	10
so	64	didn't	9
best	53	good engineer if	8
posit	51	little by little	8
question	47	engineer if	8
very	47	given problems	8
thing	45	many things	8
greet	44	quick learner	7
use-	43	communicate with others	6
communic	42	couldn't	6
employe-	42	too much	6
engin	42	unclear things	6
seem-	42	seemed to be quiet	5

Fig 2　Term and phrase analysis (Internship)

できる．簡単に言うと感情分析で総体的感情の度数分布を見ることができる．なお現状，JMP の感情分析は，英語のみしか扱えない [5].

(1) 学年次間の差異分析

2021 年度は 2020 年のコロナ禍によるインターンシップ中止の影響で２年次・３年次のインターンシップが同時に実施された．データ数の異なる学年次間を評価するために, 良感情と 悪感情の比(=[良感情 / 悪感情])を指標としてデータの正規化を試みた．図 3 に示すように感情分析の結果，２年次と３年次学生の[良感情 / 悪感情]の値は 9.6 と 7.8 というように大きな差は見られなかった．なお[良感情 / 悪感情]は，前報同様 [1]，例えば図 3(2 年生 : 良感情)の場合, スコア 95, 85, 75, 65…25, 15，5 の度数をそれぞれ 31，9，17，77…，1，0，0 として，良感情総計と悪感情総計を簡易的に[スコア中心値]×[度数]より近似計算し，その比より求めている。 例えば，２年次の[良感情 / 悪感情]割合は,

良感情総計=95×31＋85×9＋75×17+65×77
+…+25×1+15×0+5×0=11060

同様に,

悪感情総計=(-95)×0＋(-85)×2＋(-75)×0+(-65)
×0+…+(-25)×0+(-15)×1+(-5)×0= -1155

よって，[良感情 / 悪感情]=|11060/(-1155)|≒9.6

なお，感情分析でスコアの低いアンケートを一例して抽出してみると,

・時々居眠りしていて，われわれの教え方が悪いのかなとも思った

となっていた [1]．スコアがマイナスのアンケート内容は，学内でのインターンシップ事前教育の場などで活用できると考える．

インターンシップ実施学年を２年次/３年次にするかは議論のあるところであるが，感情分析結果の学年間の差異については，今後 N 数を増やしての詳細な分析が必要と考えられる．

(2) 学科間の差異分析

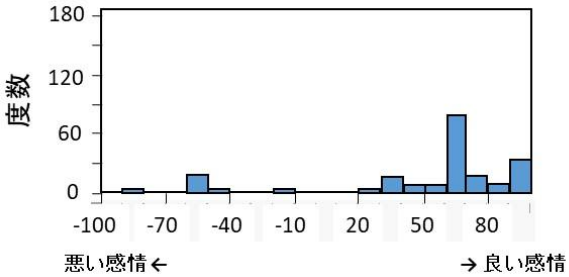
図 4 は、製造系・建設系それぞれのアンケート感情分析結果である．3.1 (1)同様に，例えば製造系の割合は,

良感情総計=95×82＋85×20＋75×35+65×
170+…+5×1=25190

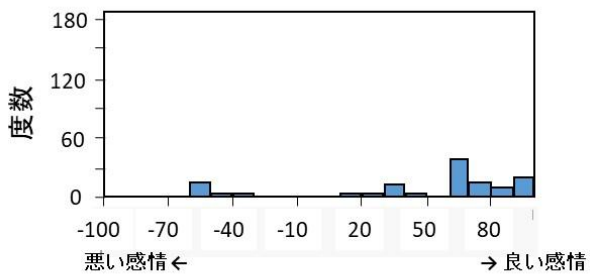
同様に，悪感情総計=-2300

よって，[良感情 / 悪感情]=|25190/(-2300)|≒10.6

製造系と建設系の[良感情 / 悪感情]値は 10.6, 17.1 であった．良感情総計値は製造系が建設系を大きく上回っているものの(25190 と 9830)，それ以上に製造系の悪感情総計値が建設系と比較して大きい (2380 と 575) ことが製造系と建設系の[良感情 / 悪感情]値の大きな差を招いている．製造



(a) Sophomores



(b) Juniors

Fig 3　Sentiment analysis (1)

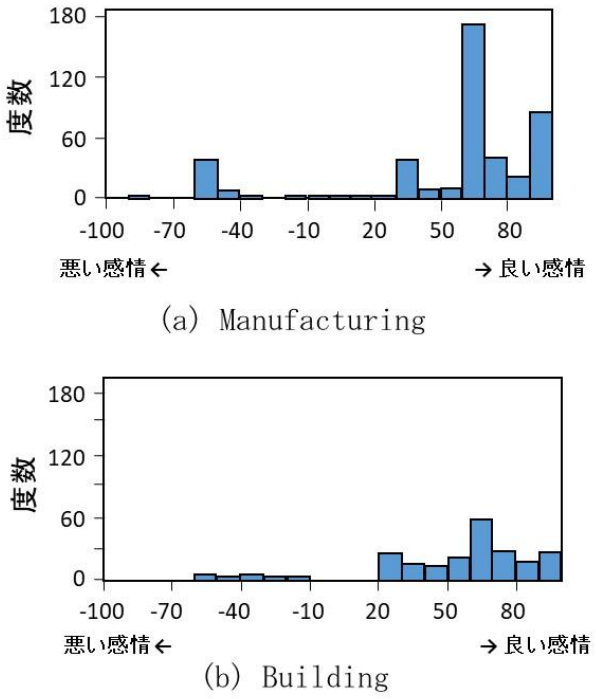


Fig 4　Sentiment analysis (2)

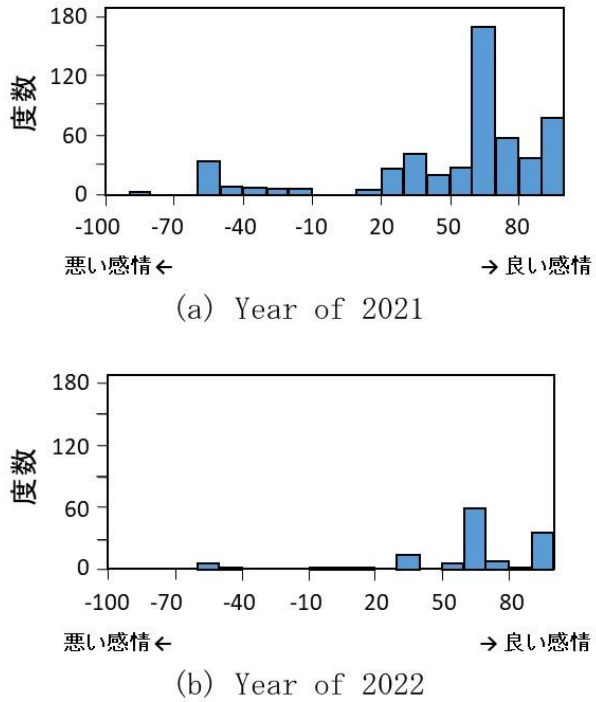


Fig 5　Sentiment analysis (3)

系・建設系の経年的傾向，および製造系の悪感情詳細分析が必要と考える．

(3) 年度間の差異分析

図 5 は、2021 年度/2022 年度のそれぞれのアンケート感情分析結果である．3.1 (1)同様に，例えば 2021 年度の割合は，

良感情総計=95×82＋85×20＋75×35+65×170+…+15×1=28655

同様に，悪感情総計=-2110

よって, [良感情 / 悪感情]=|28655/(-2110)|≒13.6

感情分析では，2021 年度と 2022 年度の[良感情 / 悪感情]値は 13.6，22.5 であった．

2021 年度と 2022 年度ではデータ数が異なる（1492：2021 年度，305：2022 年度）．良感情総計はデータ数にほぼ比例しているのに対し（28655：2021 年度，7300：2022 年度），悪感情の傾向は異なる（2110：2021 年度，325：2022 年度）．3.1(2)の学科間の差異分析同様，2021 年度の悪感情総計値(=2110)が 2022 年度のそれ(=325)と比較して大きいことが 2021 年度と 2022 年度の[良感情 / 悪感情]値の大きな差を招いていると考えられる．

以上，インターンシップアンケートの感情分析結果より，今後は感情分析を長いレンジでみたときの変化の定量化，および受け入れ企業様からの評価の客観的・定量的実態のインターンシップ事前教育へのフィードバック等が課題と考えられる．

3.2　オープンキャンパスアンケートの傾向分析

図 6 はオープンキャンパスアンケートの単語と句の出現結果である [6]．単語抽出時は JMP 保有の辞書を参照している．単語抽出の際は，構文や品詞の解析は行なわず，辞書にある単語とのマッチングのみで抽出を行っている．短い単語の合体によって長い単語が成り立ち，三者ともに辞書に含まれる時は，JMP 内のアルゴリズムによって長い単語を優先するか，あるいは短い単語を優先するかを判断している．

出現傾向を確認後、特異値分解 [7]によってテキストデータを数値データに変換する．

その変換された数値データに基づいて統計分析が実施される．特異ベクトルの分析でその次元（特異値分解プロット上の X 軸・Y 軸）が何を表しているかがわかる．特異値分解プロットの各点は，文書または単語と紐付けされていて，プロット上の点を選択（クリック）することで該当する文書や単語を表示することができる．図 7 は 2023 年度

オープンキャンパスアンケートを季節毎(春・夏・秋)にマーカー分けした特異値分解である.季節毎(春・夏・秋)にマーカー分けしたが、季節による偏りは無かった。文書プロットの右側の点群（図7：丸部A）には、例えば以下のようなオープンキャンパスや学生スタッフに対する改善を求める意見があった（原文を掲載）。

・キャンパスツアーの際に説明してくれる学生がいまひとつだった。説明、態度、姿勢等。学校を代表する立場なのでもっと訓練するか，適任のある人間に任せた方が良いと感じた。

・食堂のゴミの多さに来客を迎える準備ができていないと感じた。キャンパスツアー中も自販機のゴミがいっぱいになっていたり、各所にゴミが目立つ。

これに対して、中央と左側の点群には、オープンキャンパスや学生スタッフに対する肯定的な意見があった。

アンケート分析の特異値分解分析では，肯定的意見と否定的意見に分かれていることが分かる．否定的意見では，キャンパスツアーを行う学生スタッフの改善を求める意見の他，食堂の机にゴミが落ちていたり，自販機のゴミがいっぱいになっていたりなどキャンパス内の整備環境を指摘する意見があった．より良いオープンキャンパスを実施できるようにするためにも受け入れ態勢を見直す必要があると考えられる．

アンケート文中での出現単語･出現句の傾向(図6）を把握し，特異値分解（図7）後にトピック分析[8]に移る．テキストの内容を解釈しやすくするため，文書単語行列の特異値分解した結果をVarimax 回転し．単語を「トピック」と呼ばれるグループにまとめる．アンケートの代表的トピック上位(例えば上位5トピックや上位10トピックなど：トピックの数は変更可能）を示唆できる．

図8はトピック分析結果である[6]．図8から，トピック1では負荷量の高い順で「ツアー」・「キャンパス」・「説明し」という単語があり，オープンキャンパスの説明に関する感想を示唆してい

単語と句のリスト

単語	度数
分かり	51
だった	41
説明し	31
をして	29
詳しく	28
ｷｬﾝﾊﾟｽ	26
実際に	26
ありがと	25
雰囲気	25
皆さん	23
ｵｰﾌﾟﾝｷｬﾝﾊﾟｽ	22
ﾂｱｰ	21
ｽﾀｯﾌ	19
さった	14
ですが	14
そうだ	13
に対して	12
どんな	11
初めて	11
ｶﾘｷｭﾗﾑ	10
ｺﾝｸﾘｰﾄ	10
魅力的	10
したい	9
実践的	9
ても良い	8

句	度数	N
ｷｬﾝﾊﾟｽﾂｱｰ	17	2
詳しく説明し	3	2
ｽﾀｯﾌや先生方が良い	2	4
ﾂｱｰの際に説明し	2	4
一つ一つ丁寧に説明し	2	4
自分でもやってみたい	2	4
表する立場なのでもっと	2	4
に関して詳しく説明し	2	3
ﾛﾎﾞｯﾄを動かす	2	3
をしてくださった	2	3
持ちの良い対応をして	2	3
表する立場なので	2	3
なのでもっと	2	2
に関して詳しく	2	2
魅力的だった	2	2

Fig 6　Term and phrase analysis (Open campus)

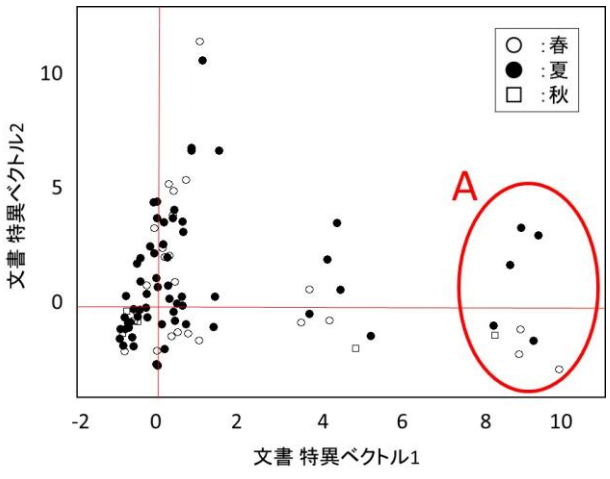


Fig 7　Singular value decomposition

トピック別上位負荷量

トピック1		トピック2		トピック3		トピック4		トピック5	
単語	負荷量	単語	負荷量	単語	負荷量	単語	負荷量	単語	負荷量
ﾂｱｰ	0.91277	ありがと	0.6460	に対して	0.7022	だった	0.6144	詳しく	0.6698
キャンパス	0.90584	さった	0.5940	をして	0.6720	初めて	0.5965	説明し	0.4545
説明し	0.23776	ですが	0.5036	初めて	0.2880	魅力的	0.4162	実際に	-0.4145
だった	0.15828	をして	0.2779	さった	0.2560	そうだ	-0.2569	どんな	-0.2631
		説明し	-0.2083	だった	-0.1734	ｵｰﾌﾟﾝｷｬﾝﾊﾟｽ	0.2531	そうだ	-0.2564
		皆さん	0.2020	ですが	-0.1435	実際に	-0.2254	スタッフ	-0.2326
		に対して	-0.1874	詳しく	0.1205	スタッフ	-0.1897	雰囲気	0.1876
								ｶﾘｷｭﾗﾑ	0.1408

トピック6		トピック7		トピック8		トピック9		トピック10	
単語	負荷量	単語	負荷量	単語	負荷量	単語	負荷量	単語	負荷量
ｶﾘｷｭﾗﾑ	0.6824	ｵｰﾌﾟﾝｷｬﾝﾊﾟｽ	0.6806	皆さん	0.6858	どんな	0.6749	ｺﾝｸﾘｰﾄ	0.7879
魅力的	0.5993	スタッフ	0.6619	分かり	0.5407	雰囲気	0.6193	実際に	0.3065
実際に	0.4632	雰囲気	-0.1895	ですが	-0.4439	そうだ	-0.3552	分かり	-0.2833
詳しく	0.1367	実際に	0.1804	説明し	-0.1513	さった	0.1570	ですが	--0.2385
に対して	-0.1343	そうだ	-0.1652	実際に	-0.1511	ですが	-0.1475	詳しく	0.2206
雰囲気	-0.1253	初めて	0.1471	そうだ	-0.1493	に対して	-0.1210	説明し	-0.1958
		説明し	0.1326					だった	0.1822
		さった	-0.1299					ｶﾘｷｭﾗﾑ	-0.1380
		ありがと	0.1222					雰囲気	-0.1324

Fig 8　Topic analysis

ることが分かる．トピック 2 と 3 では正確にトピックスを示唆することができないがどちらも説明を行ったスタッフへの感想だと考えられる．また，「オープンキャンパス」「スタッフ」・「カリキラム」「魅力的」といった単語も高い負荷量を記録しており，大学の特色やスタッフに対する感想を示唆していることが分かる．

今回邦文テキストエクスプローラでのトピック分析では，単語当りの最小文字数の設定が難しいことが分かった．すなわち最小文字数が小さいと断片的な日本語が羅列されるだけでトピック文章を示唆できず，一方最小文字数が大きいとトピック候補ワードの数が少なくなり過ぎて，これもまたまたトピック文章が示唆し辛くなってしまう．図 8 では単語当りの最小文字数は，試行錯誤の結果，「3」とした．

以上，オープンキャンパス来場者アンケート分析より，特異値分解を通してマイナス意見を抽出してオープンキャンパスの改善につなげること，および各年度の感情分析と翌年の入学者数間の相関評価を行なって来場者の良感情が大きくなるようにオープンキャンパスのコンテンツを見直していくこと等が課題と考える．

4. まとめ

1) インターンシップ評価票中の研修学生に関して受入れ企業様からの気付きに関する自由フォーマットテキスト文に対して，層別の感情分析を通して評価を定量化した．
2) オープンキャンパス来場者アンケートについては特異値分解による否定的意見の抽出，およびトピック分析に対する考察を行なった．

今後の課題は以下である．

1) インターンシップ：感情分析経年変化の定量とインターンシップ事前教育への受け入れ企業様評価実態のフィードバック
2) オープンキャンパス：特異値分解によるマイナス意見の抽出とそれに対する改善，および感情分析と入学者数間の相関評価を踏まえたオープンキャンパスコンテンツの改善

謝辞

本研究は，「2023 年度ものつくり大学教育力・研究力強化プロジェクト」の採択を受け支援を頂きました．荒木邦成氏，石川一樹氏，鴇崎正和氏にはデータ収集・分析検討に助言頂きました．感謝申し上げます．

文献

1) 佐久田茂，ものつくり大学紀要 No.13，pp7-11，(2023)
2) https://bellcurve.jp/statistics/blog/14333.html (参照　2024-07-30)
3) https://www.jmp.com/support/help/ja/16.2/index.shtml#page/jmp/text-explorer.shtml (参照 2024-06-10)
4) SAS Institute Inc., JMP16 ドキュメンテーションライブラリ,2021, p1409.
5) https://www.jmp.com/support/help/ja/17.2/index.shtml#page/jmp/sentiment-analysis.shtml (参照 2024-06-10)
6) 会田将貴，オープンキャンパスデータの統計的分析，ものつくり大学総合機械学科卒業論文，2023
7) SAS Institute Inc., JMP16 ドキュメンテーションライブラリ,2021, p1392.
8) https://www.jmp.com/support/help/ja/16.2/index.shtml#page/jmp/topic-analysis.shtml (参照 2024-06-16)