福山平成大学経営学部紀要
第 18 号（2022），117－128 頁

College Analysis を使い易くする追加機能４

福井 正康[*1]・奥田 由紀恵[*2]・細川 光浩[*2]

[*1] 福山平成大学経営学部経営学科
[*2] 福山平成大学大学教育センター

要旨：社会システム分析ソフト College Analysis の分析メニュー「対応のある質的データの検定」にコクランの Q 検定、「量的データの集計」にレコードごとの横集計機能を加えた。またユーティリティとして、エディタなどのヘッダー入力を補助する表示枠移動機能や群分けされたデータで多変量解析が行える機能、入力ミスなどを調べるデータチェック機能などを追加した。

キーワード：College Analysis、C.Analysis、コクランの Q 検定、ユーティリティ

1. はじめに

社会システム分析ソフト College Analysis（C.Analysis と略す）に追加した単独の分析やユーティリティを紹介することがこの報告の目的である[1]-[4]。最初に、対応のある質的データの検定で、マクネマー検定の 3 変数以上への拡張であるコクランの Q 検定について説明する[5]。次に、量的データについて各個体の部分的な小計などを計算する横集計、エディタなどのヘッダー部分の編集用に作った表示枠移動機能について解説する。

C.Analysis の懸案はデータの一部を使った分析であった。例えば、男女別にみた多変量解析などである。これについて、重回帰分析には群分け機能を付けたが、一般の分析については、まずデータを群に分けて、その後群別のデータを用いてそれぞれの分析を行う必要があった。しかし、これはデータを書き換えたり、取り除いたりする作業を含んでおり、ひと手間が必要であった。これを簡単にしかも自由な分類で行えるのがツールの「文字列結合」と「データ形式変換」というメニューである。

複数のファイルにあるデータを 1 つのキーデータでつなげることも考えておく必要がある。但し、C.Analysis では複数ファイルは異なるページに読み込まない限り扱えないので、今回のプログラムでは 1 頁内にそれらのデータを貼り付けて使うようにした。

最後に、データの入力ミスへの対応も考えてみた。アンケート調査のデータで手入力した場合、入力ミスは付きものである。例えば、分類にないデータを入力したり、半角を全角と間違えて入力したりすることはよくある事例である。卒業論文などで利用するにはこれらを簡単に検出する機能が必要となる。

ここで報告する追加機能は、ネットを通じて要望があったものや、開発者が卒業論文の指導で不自由さを感じた部分を改良したものである。

2. コクランの Q 検定

質的データの検定メニュー中の「対応のある比率の検定」または、メニュー［分析－基本統計－質的データの検定－対応のある比率の検定］を選択すると、図 1 の対応のある比率の検定の実行画面が表示される。

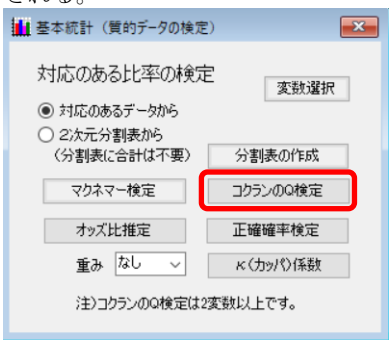


図 1　対応のある比率の検定実行画面

ここではこの画面に新たに加えたコクランの Q 検定について説明する。図 2 のデータの形式（テキスト 3.txt, p14）は、マクネマー検定の変数の数が増えたデータである。但し、分割表からというのは次元数の関係から不可能である。

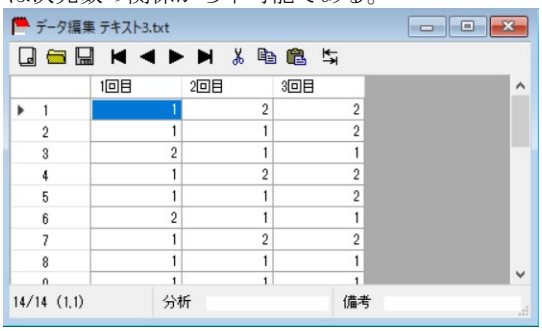


図 2　コクランの Q 検定のデータ

すべての変数を選択し、「コクランの Q 検定」ボタンをクリックすると図 3 のような分析結果が表示される。

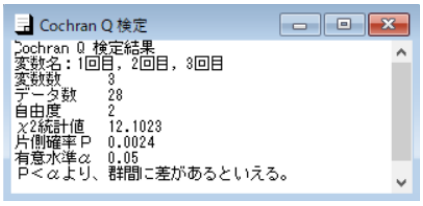


図 3　コクランの Q 検定の結果

以下にこの検定の簡単な理論を与えておく。但し、図 4 のデータを一度 {0,1} データに変換して（例えば、1→0、2→1）利用する。

データを $x_{ij} = \{0, 1\}$ とすると、データは以下の表 1 のように与えられる。

表 1　対応のある質的データ

	1	2	⋯	k	計
1	x_{11}	x_{12}	⋯	x_{1k}	$x_{1\cdot}$
2	x_{21}	x_{22}	⋯	x_{2k}	$x_{2\cdot}$
⋮	⋮	⋮	⋮	⋮	⋮
n	x_{n1}	x_{n2}	⋯	x_{nk}	$x_{n\cdot}$
計	$x_{\cdot 1}$	$x_{\cdot 2}$	⋯	$x_{\cdot k}$	

このデータを利用して、以下の Q 統計量を計算する。

$$Q=\frac{k(k-1)\sum_{j=1}^{k}(x_{\cdot j}-\bar{x})^2}{k\sum_{i=1}^{n}x_{i\cdot}-\sum_{i=1}^{n}x_{i\cdot}^2}\sim\chi_{k-1}^2$$

ここに、$x_{i\cdot}=\sum_{j=1}^{k}x_{ij}$，$x_{\cdot j}=\sum_{i=1}^{n}x_{ij}$，$\bar{x}=\frac{1}{k}\sum_{j=1}^{k}x_{\cdot j}$

特に、$k=2$ の場合、表 2 の分割表を使うと、

表 2　$k=2$ の対応のあるでーたの分割表

変数 1＼変数 2	0	1
0	a	b
1	c	d

データの組み合わせの度数が、(0,0)→ a，(0,1)→ b，(1,0)→ c，(1,1)→ d　となり、$x_{\cdot 1}=c+d$，$x_{\cdot 2}=b+d$，$\bar{x}=(b+c+2d)/2$　であるから、

$$Q=\frac{2\left[(x_{\cdot 1}-\bar{x})^2+(x_{\cdot 2}-\bar{x})^2\right]}{2\sum_{i=1}^{n}(x_{i1}+x_{i2})-\sum_{i=1}^{n}(x_{i1}+x_{i2})^2}$$

$$=\frac{2\left[(x_{\cdot 1}-\bar{x})^2+(x_{\cdot 2}-\bar{x})^2\right]}{2(x_{\cdot 1}+x_{\cdot 2})-\sum_{i=1}^{n}(x_{i1}^2+x_{i2}^2+2x_{i1}x_{i2})}$$

$$=\frac{2\left[(c-b)^2/4+(b-c)^2/4\right]}{2(b+c+2d)-(b+c+2d+2d)}=\frac{(b-c)^2}{b+c}\sim\chi_1^2$$

のように、マクネマー検定の結果と一致する。

また、イェーツ補正の入れ方について、以下のように考えると、

$$Q=\frac{k(k-1)\sum_{j=1}^{k}(|x_{\cdot j}-\bar{x}|-\frac{1}{2})^2}{k\sum_{i=1}^{n}x_{i\cdot}-\sum_{i=1}^{n}x_{i\cdot}^2}\sim\chi_{k-1}^2$$

$k=2$ の場合、

$$Q=\frac{2\left[(x_{\cdot 1}-\bar{x})^2+(x_{\cdot 2}-\bar{x})^2-|x_{\cdot 1}-\bar{x}|-|x_{\cdot 2}-\bar{x}|+1/2\right]}{2\sum_{i=1}^{n}(x_{i1}+x_{i2})-\sum_{i=1}^{n}(x_{i1}+x_{i2})^2}$$

$$= \frac{2\left[(b-c)^2/2 - |b-c| + 1/2\right]}{2(b+c+2d) - (b+c+2d+2d)}$$

$$= \frac{(b-c)^2 - 2|b-c| + 1}{b+c} = \frac{(|b-c|-1)^2}{b+c} \sim \chi_1^2$$

となり、マクネマー検定のイェーツ補正と一致する。

3. 量的データの横集計

量的データの集計の実行画面は、メニュー［分析－基本統計－量的データの集計］を選択すると図 1 のように表示される。

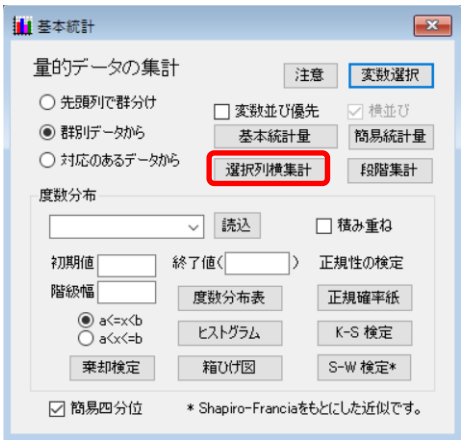


図 1　量的データ集計実行画面

アンケートの質問票などで各人の部分合計などが必要になる場合がある。Excel などでは簡単に実行できる計算であるので、これまでは元のデータの段階で実行しておくものと考えていたが、学生の指導上 C.Analysis の上で実行する必要もある。そのような場合に利用するのが「選択列横集計」である。部分集計したい変数を選択し、ボタンをクリックすると、個体別のデータの個数、合計、平均、（分散からの）標準偏差がグリッド出力に図 2 のように出力される。応急処置のような機能である。

	カウント	合計	平均	標準偏差P
1	5	15	3.000	0.894
2	5	22	4.400	1.200
3	5	15	3.000	0.894
4	5	17	3.400	1.020
5	5	15	3.000	1.095
6	5	18	3.600	1.200
7	5	17	3.400	1.020

図 2　選択列横集計

必要な項目を選んでグリッドエディタに貼り付けることで、このデータを分析データとして利用することができる。なお、さらに複雑な集計や計算はエディタメニューの［ツー

ルー計算］を利用する。

4. エディタとグリッド出力の表示枠移動

エディタを利用する際、ヘッダー部分に変更を加えたいときがある。例えば、変数を追加して名前を付けたいときや、範囲を間違い行名部分にデータが入ってしまったときなどである。これまで前者の場合はエディタの「列名入力」機能を利用し、後者の場合は範囲を選び直して再度貼り付けを行っていた。今回これらを統一的に扱うことができないかと考え、グリッドの表示枠の移動機能を追加した。これまで列名や行名であった部分をデータ側に動かしたり、その逆を行ったりする機能である。例えば、図 1 のようなデータ枠を図 2 のように動かす機能である。

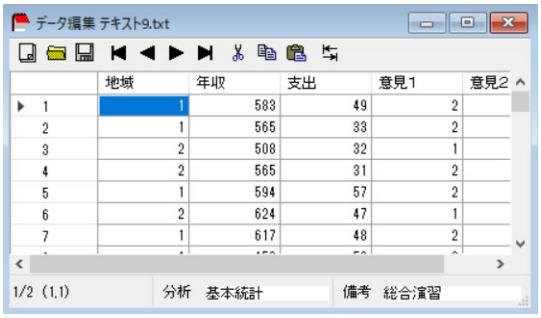


図 1　移動前元データ

図 2　移動後データ

ここでは元データを右と下に移動しているが、メニュー［編集ー枠移動］では、「右下へ」、「左上へ」、「右へ」、「左へ」、「下へ」、「上へ」がある。実際に利用してみると気軽に使えるので、この機能をグリッド出力へも追加している。

5. データツール

5.1 列並び替え

C.Analysis では、ツール「データ生成」、「計算」、「文字列結合」で新しい列を作る場合、一度列を挿入して「範囲指定」して出力するか、「新規追加」にして最後の列に追加する方法をとっていた。また、分析の出力結果、例えば重回帰分析の予測値や因子分析の因子得点などは一度グリッド出力に出力した後、グリッド出力の機能を用いて、エディタの最後の列に追加していた。いずれの場合もデータを見易くするため、列の並びについて順

番を自由に変えられる機能が必要であった。

エディタのメニュー[ツールー列並び替え]を選択すると図1のメニューが表示される。

図1　列並び替え

表示された時点でグリッドエディタの変数名が表示されている。このメニューが表示された後で変数名を変えた場合は「再設定」ボタンをクリックする必要がある。変数名の並び替えは変数選択と同じである。並び替えた後に「実行」ボタンをクリックするとエディタの列が変更される。エディタの変数名と列並び替えの変数名が異なる場合はその旨のエラーメッセージが表示されるので、「再設定」ボタンをクリックしてそろえる。

このような方法で自由に列の並び替えができるので、列追加の際の列並びには気を使う必要はなくなった。

5.2　データ形式変換（多変量解析の群分け法）

このツールは文字列結合のツールと並んで重要である。これは元々列の並び替え用に作ったツールであったが、先頭列で群分けのデータを特殊な形式の群別データに変換する機能を追加することで非常に重要なツールとなった。サンプルデータを用いて機能を見てみよう。メニュー［ツールーデータ形式変換］を選択すると図1のような実行画面が表示される。

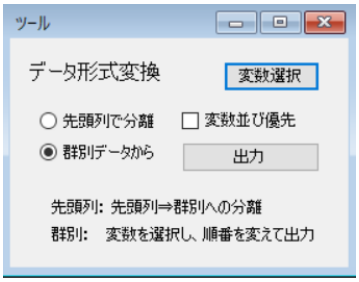


図1　データ形式変換実行画面

例として利用するデータを図2に示す。

図 2　データ（テキスト 9.txt）

ラジオボタン「群別データから」を選択した場合は、変数選択の際に変数とその順番を選ぶと、例えば図 3 のような出力結果となる。

図 3　「群別データから」の出力

この出力データのレコード並びは、欠損値も含めて図 2 と同じものになっている。このデータはエディタの新しいページに貼り付けて使用することもできるし、そのままエディタの現在のページに上書きすることもできる。

「先頭列で分離」を用いて、地域、年収、支出を選択して、出力すると図 4 のような結果になる。ここでは、実行画面の「変数並び優先」チェックボックスにチェックを入れている。

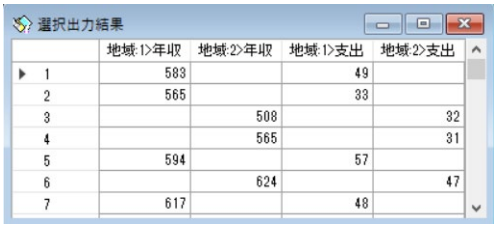


図 4　「先頭列で分離」の出力

この結果は、地域 1 と地域 2 でデータが分離されている。この形式の利点は、これを群別データとして集計すると地域 1 と地域 2 の集計結果が求められることである。さらに、多変量解析などでは、レコード単位のデータ除去を行うことから、このデータを含めて分析を実行すると地域 1 か地域 2 のデータだけの分析結果が得られる。

簡単な例として、地域別の支出と年収の回帰分析を求めてみよう（すでに機能としてはあるが）。図 4 で求めたグリッド出力の 1 列目と 2 列目をグリッド出力の「エディタ指定

列追加」メニューを用いてエディタに追加出力し、図 5 のようなデータを得る。

図 5　分離データの追加

ここで例えば、支出と地域:1>年収を選んで回帰分析にかければ、図 6 のように先頭列で群分けした結果と同じ結果を得る。

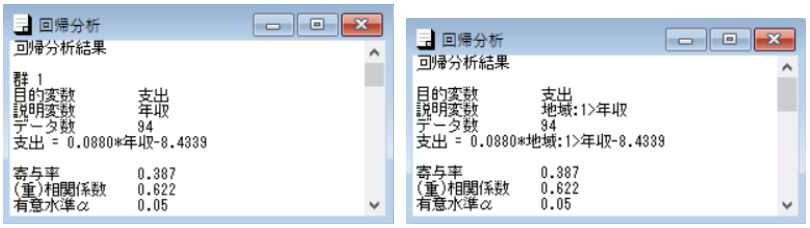


図 6　先頭列で群分けの結果（左）と分離データを用いた結果（右）

さらに、地域と意見 1 のデータを図 7 の文字列結合メニューで図 8 のように結合する。

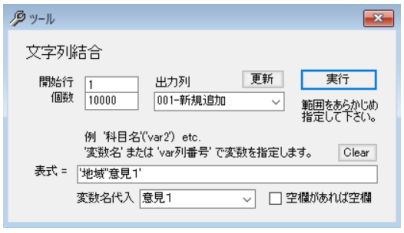


図 7　文字列結合

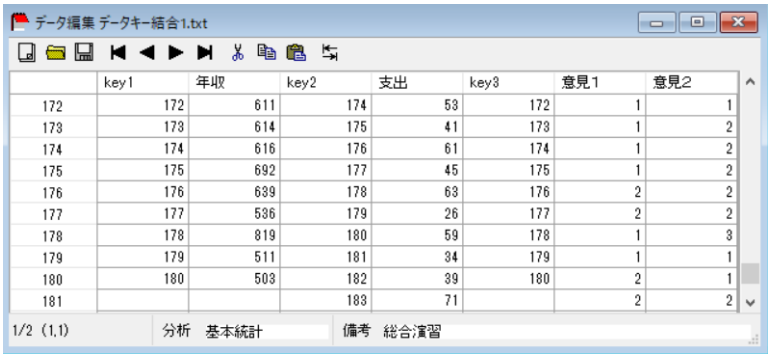
図 8　文字列結合後

この地域・意見 1 のデータ使って、図 4 の「先頭列で分離」処理を行うとさらに細かい分類で分析を実行することができる。

5.3　データキー結合

複数のファイルなどに含まれるキーでつながったデータを、キーを元に 1 つのデータにつなげる作業を考える。しかし、C.Analysis では複数のファイルは扱えないので、データを、キー、複数データ、キー、複数データ、…の形に集めてから処理を行うこととする。これは R や Python などのコマンドを入力するタイプの（コマンド系とする）統計ソフトに比べて弱いところである。

データの形式を図 1 に示す。



	key1	年収	key2	支出	key3	意見1	意見2
172	172	611	174	53	172	1	1
173	173	614	175	41	173	1	2
174	174	616	176	61	174	1	2
175	175	692	177	45	175	1	2
176	176	639	178	63	176	2	2
177	177	536	179	26	177	2	2
178	178	819	180	59	178	1	3
179	179	511	181	34	179	1	1
180	180	503	182	39	180	2	1
181			183	71		2	2

図 1　データ形式

これは、3 つの群からなるデータで、それぞれの key に対して、変数が 1 つ、1 つ、2 つと含まれている。図の下の方には欠損値が含まれている。これらの変数をすべて利用してもよいし、変数選択で選んで利用してもよい。

グリッドエディタから、メニュー［ツールーデータキー結合］を選択すると、図 2 のような実行画面が表示される。

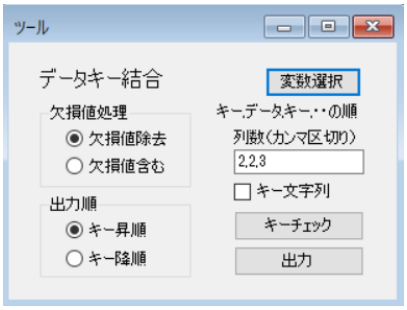
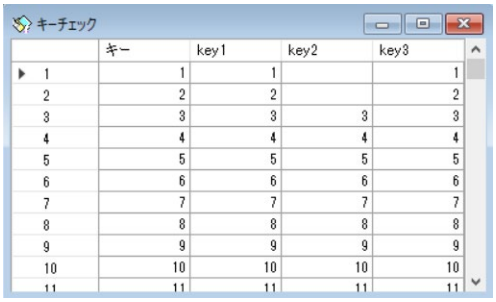


図 2　実行画面

すべての変数を選択すると、列数の指定はキー列も含めて、「2,2,3」となる。「キーチェック」ボタンは、キーの漏れや 2 重登録を調べるボタンである。クリックすると図 3 の結果を得る。



	キー	key1	key2	key3
1	1	1		1
2	2	2		2
3	3	3	3	3
4	4	4	4	4
5	5	5	5	5
6	6	6	6	6
7	7	7	7	7
8	8	8	8	8
9	9	9	9	9
10	10	10	10	10
11	11	11	11	11

図 3　キーチェック結果

左端の「キー」列は、選択した範囲で現れたすべてのキーである。それに対して 2 列目以降が個々のキーの対応である。

「出力」ボタンをクリックすると、設定に合わせて結果をグリッド出力で表示する。ここでは例として、「欠損値除去」の場合と「欠損値含む」の場合の結果をそれぞれ図 4a と図 4b に示す。

	キー	年収	支出	意見1	意見2
1	3	508	49	2	3
2	4	565	33	2	1
3	5	594	32	2	3
4	6	624	31	1	1
5	7	617	57	2	1
6	8	458	47	2	3
7	9	754	48	2	1
8	10	667	53	2	1
9	11	470	62	1	1
10	12	578	53	2	3
11	13	592	37	2	3

図 4a　出力結果（欠損値除去）

	キー	年収	支出	意見1	意見2
1	1	583		1	3
2	2	565		2	3
3	3	508	49	2	3
4	4	565	33	2	1
5	5	594	32	2	3
6	6	624	31	1	1
7	7	617	57	2	1
8	8	458	47	2	3
9	9	754	48	2	1
10	10	667	53	2	1
11	11	470	62	1	1

図 4b　出力結果（欠損値含む）

この場合、キーについては数字として取り扱っているが、文字列として取り扱うこともできる。「キー文字列」チェックボックスにチェックを入れ、欠損値除去の場合の結果を図 5 に示す。並びの違いが明らかである。

	キー	年収	支出	意見1	意見2
1	10	667	53	2	1
2	100	582	41	2	3
3	101	664	39	1	3
4	102	580	51	1	3
5	103	1168	50	2	3
6	104	685	39	2	2
7	105	787	78	2	1
8	106	618	45	1	3
9	107	482	57	2	3
10	108	746	47	2	2
11	109	557	40	2	2

図 5　出力結果（キー文字列）

5.4　データチェック

卒論データなどを見ていると様々な入力ミスがある。これを素早く見つけるために作られたのがこのプログラムである。グリッドエディタのメニュー[ツールーデータチェック]を選択すると、図 1 のような実行画面が表示される。

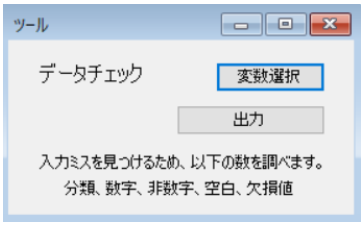
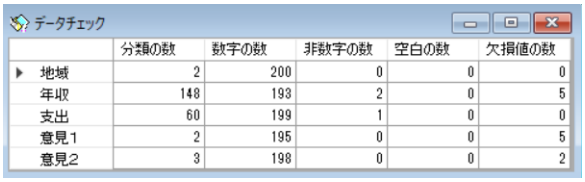


図 1　データチェック実行画面

使い方は変数選択をして、「出力」ボタンをクリックするだけである。結果は図 2 のように表示される。



	分類の数	数字の数	非数字の数	空白の数	欠損値の数
▶ 地域	2	200	0	0	0
年収	148	193	2	0	5
支出	60	199	1	0	0
意見1	2	195	0	0	5
意見2	3	198	0	0	2

図 2　出力結果（データチェック.txt）

卒論データなどの入力ミスで最も多いのは、分類の打ち間違えである。それは分類の数で確認する。入力ミスが発見されたら、分割表などで確認する。また、見えない半角や全角の空白もときどき存在する。これらは、同じツールの「空白除去」機能で範囲を選択して一括消去する。また、欠損値に「*」、「-」、「0」などを入力する場合もあるが、これらの修正は置換処理で対応する。

6.　おわりに

今回我々が追加した機能は、コクランの Q 検定を除いて、処理の手間や時間を短縮するものが多かった。特に、データを群分けして分析を実行できる機能はその代表である。このような機能はコマンド処理を基本とする R や Python などを意識して作成した。近年、これらのコマンド系の分析ソフトの書籍が多く出版されるようになって、データを表形式に読み込んでメニューのボタンを操作するような分析ソフトの勢いが衰えている。

世の中ではデータサイエンスという言葉がもてはやされ、大学でも多くの講座が開講されるようになった。大きなデータを扱うデータサイエンスには確かにコマンド系の分析ソフトが向いているように思える。なぜならほとんどデータを視覚化することなく、ファイルから直接結果が求められるからである。機械を使って自動的に入力され更新されて行く巨大なデータではデータ自身に目を通すことにあまり意味があるようには思えない。また、表形式でそれを見ようと思うとコンピュータにかなりの負荷がかかる。そのためコマンド系の分析ソフトは近年の巨大データの分析には不可欠である。

しかしこのような分析に携わる人がどれだけ必要であろうか。多くの人は分析に関わるのではなく、それを評価、利用する側である。そういう意味でデータを見てそれにより分析を理解し、結果を判断する能力が重要となる。データを大量に扱えるソフト、データを細かくチェックできるソフト、それぞれに意味があるように思う。

参考文献

［１］ 福井正康、「College Analysis 総合マニュアル　－ツールー」、http://www.heisei-u.ac.jp/ba/fukui/gmanual/gmanual01.pdf

［２］ 福井正康・細川光浩、「College Analysis を使い易くする追加機能」、福山平成大学経営研究、第 14 号（2018）107-118

［３］ 福井正康・細川光浩・奥田由紀恵、「College Analysis を使い易くする追加機能２」、福山平成大学経営研究, 第 15 号（2019）177-187

［４］ 福井正康・細川光浩、「College Analysis を使い易くする追加機能３」、福山平成大学経営研究、第 17 号（2021）177-187

［５］ 奥田由紀恵・細川光浩・福井正康、「College Analysis への機能追加 －多変量分散分析, コクランの Q 検定他－」、日本教育情報学会第 37 回年会論文集（2021）350-351、(岐阜女子大学, 2021/8/28-29)