NUMEROS.docx　ver. 2016.4.7

言語研究のための数値データ分析法

２０１６

これは 1 学期間の授業用テキストです。

随時更新します。

上田博人（東京大学）Hiroto Ueda (University of Tokyo)

0. はじめに

このテキストには言語の歴史的な変化や現代語の変異(バリエーション)を数量的に観察するときに有用だと思われる方法を取り上げました。数値が並ぶデータ行列を前にして、目視だけではよくわからない状況を、さまざまな分析手法を用いて明らかにしていきます。

私たちの文系の課程では、高校で行列・ベクトルと確率・統計、大学で線形代数と数理統計学を履修していないことが多いのですが、その初歩的な部分だけでも学習しておくと、数値データ分析法の数理の理解と、プログラミングの作業が容易になります。さらに、このテキストでは一般に定義されていないような行列演算や統計処理をあえて導入しました。そのような演算を各所で活用しますので確認してください。このテキストで扱う内容は基本的なことばかりで、難易度はそれほど高くはありません。

ここで説明する方法は言語学の各分野に限らず、実務や研究でよく使われているものばかりですが、各所で私たちが独自に開発してきた方法も紹介していきます。おおまかに「～とよばれています」という受動文であれば周知の方法を指します。一方「～を提案します」「～とよびます」のような能動文で紹介する方法は、私たち独自の方法・名称だと思いますが、すでに開発されている方法や使われている呼称であるかもしれません。一応、各種の統計学書で確認していますが、すべてを見渡すことは不可能なので既存の同じ方法・名称をご存知の方はぜひご教示ください。

学部の前期・後期課程では基礎的な内容を取り上げながらプログラム操作を練習します。大学院では発展的内容を扱いプログラム開発を練習をします。

●の箇所は数理・統計に関する補足です。私の専門は■で言語研究（スペイン語研究）での応用例を示しましたが、とくにスペイン語の専門的な知識を前提としません。

学期期間中は、このテキストとプログラムのコード NUMEROS.xlsm と次のウェブプログラムサイトを随時更新しています。いつも最新のファイルをダウンロードしてください。

＊ファイルダウンロードサイト：

http://lecture.ecc.u-tokyo.ac.jp/~cueda/gengo/index.html

＊ウェブプログラムサイト

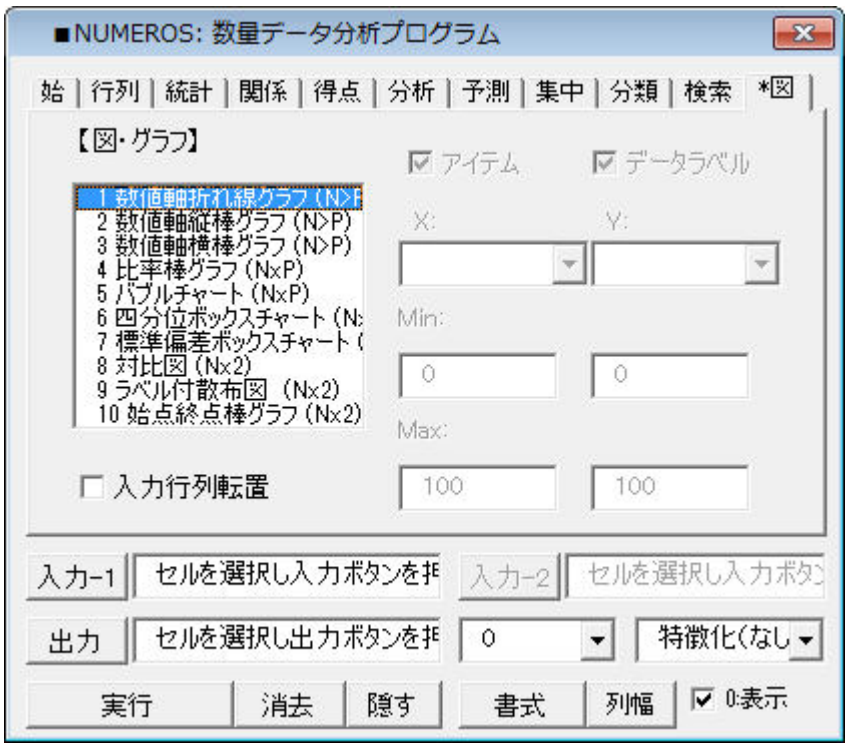
LETRAS: http://lecture.ecc.u-tokyo.ac.jp/~cueda/letras/

NUMEROS: http://lecture.ecc.u-tokyo.ac.jp/~cueda/letras/

1. グラフ

データ行列には多くの情報が含まれていますが、縦と横に並んだ数値の連続のままでは、その情報を読み取ることが困難です。そこで、さまざまなグラフを使って数値の情報を視覚化します。

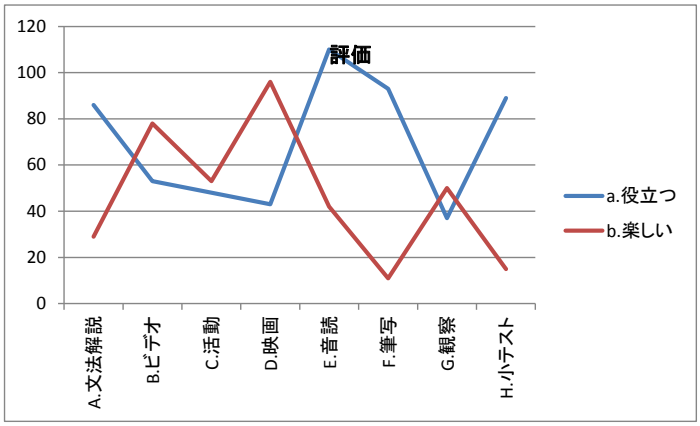
ここでは変数の関係を示す図を描くためのプログラムを扱います。Excel の標準的なグラフにないものをマクロで作成しました。



1.1. 数値軸折れ線グラフ

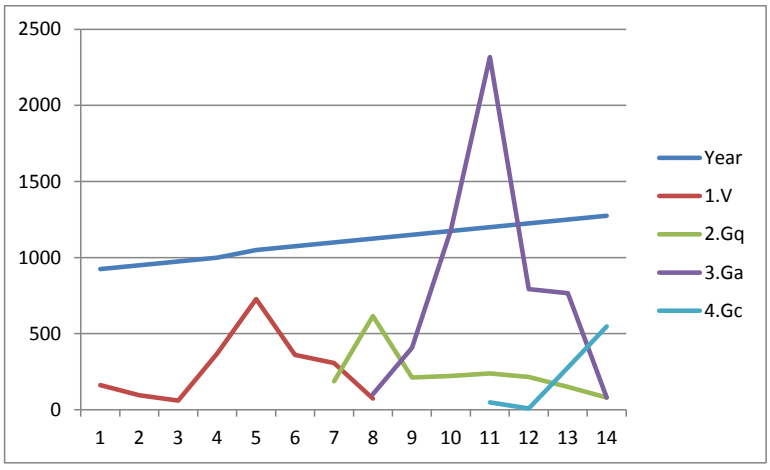
Excel では次のような複数列のデータの折れ線グラフは簡単に表示できます。

評価	a.役立つ	b.楽しい
A.文法解説	86	29
B.ビデオ	53	78
C.活動	48	53
D.映画	43	96
E.音読	110	42
F.筆写	93	11
G.観察	37	50
H.小テスト	89	15

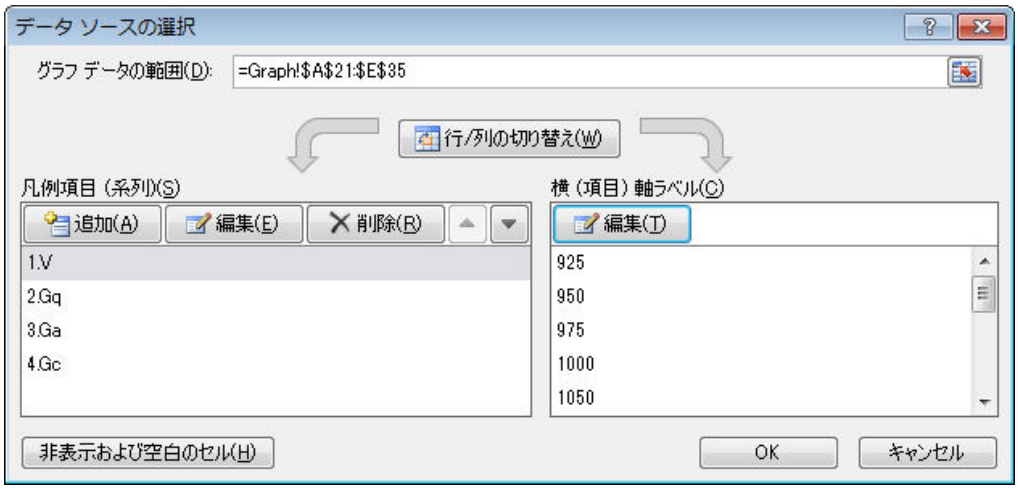


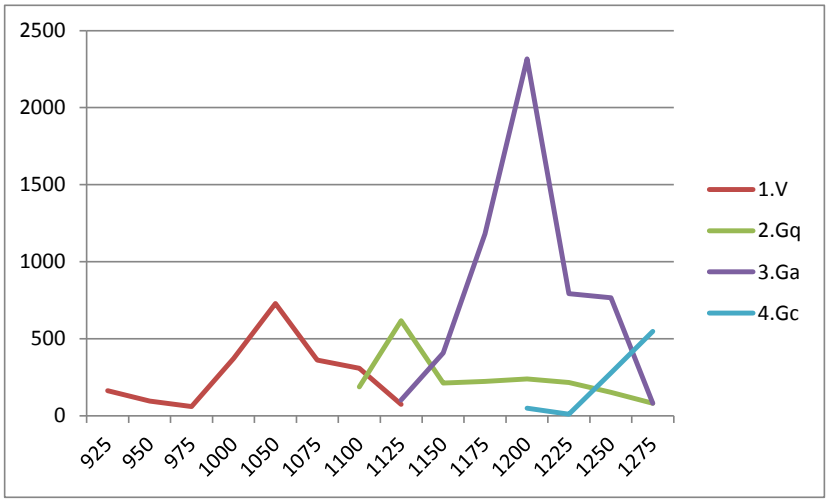
一方、次表のようにタイトル列（表側とよばれます）が数値の場合、そのまま表示するとうまく表示できません。表側も数値データと見なされてしまうからです。

Year	1.V	2.Gq	3.Ga	4.Gc
925	162			
950	95			
975	60			
1000	371			
1050	728			
1075	361			
1100	308	188		
1125	73	616	102	
1150		212	407	
1175		223	1180	
1200		239	2317	49
1225		215	792	9
1250		151	766	277
1275		81	80	548

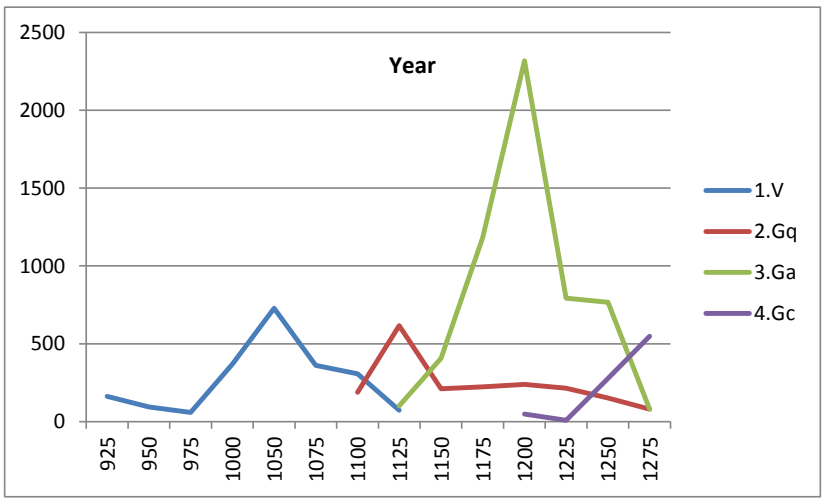


そのときは、プロットエリア（グラフの中央の領域）を右クリックして「グラフの選択」から、次の「データソースの選択」画面で、「凡例項目」で不要な項目を削除し、次に「横（項目ラベル）」の「編集」ボタンで、表側の数値列を選択して OK ボタンを押します。



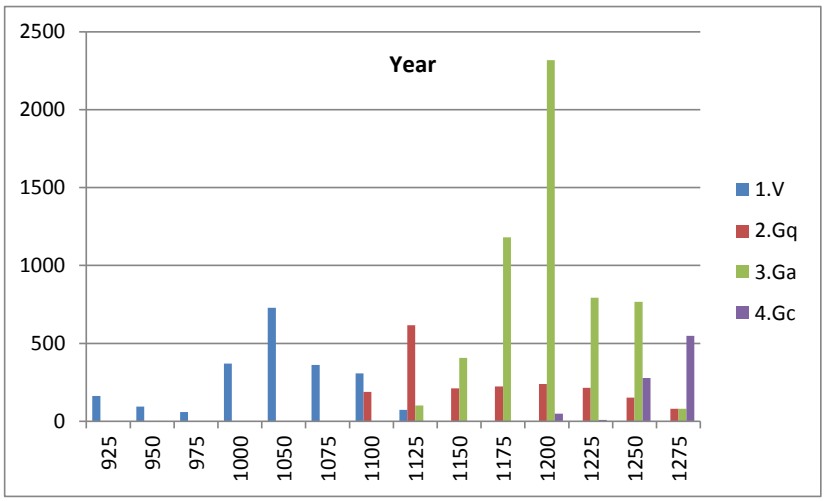


この一連の操作を回避するために、ExcelVBA でプログラム NUMEROS の「図」のタブに「数値軸折れ線グラフ」を用意しました。

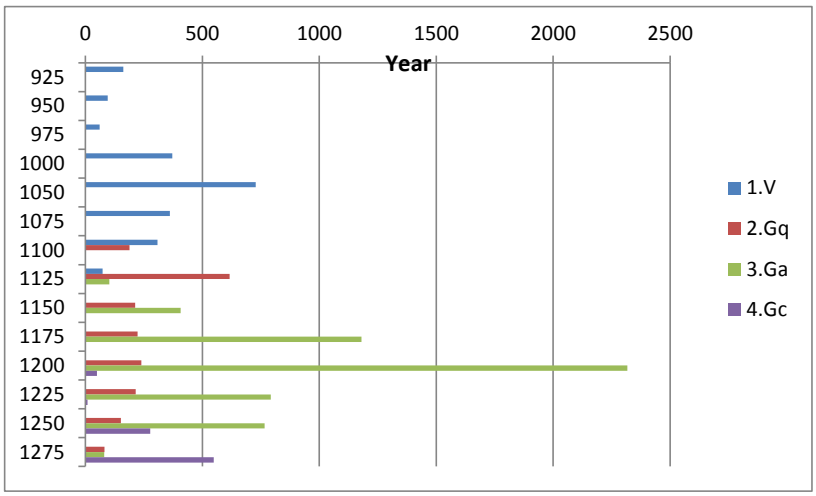


●数値棒グラフ

同様に、数値縦棒グラフを作成します。



同様に、数値横棒グラフを作成します。



1.2. 比率棒グラフ

次の入力データ「評価」は、スペイン語の授業についてのアンケートの結果です。たとえば、第 1 行の「文法解説」について、それがスペイン語の学習上「効果がある」と思う人の総数は 86 名でした（総数 124 名）。そして、同じ項目が「楽しい」と思う人の総数は 29 名であることを示しています。

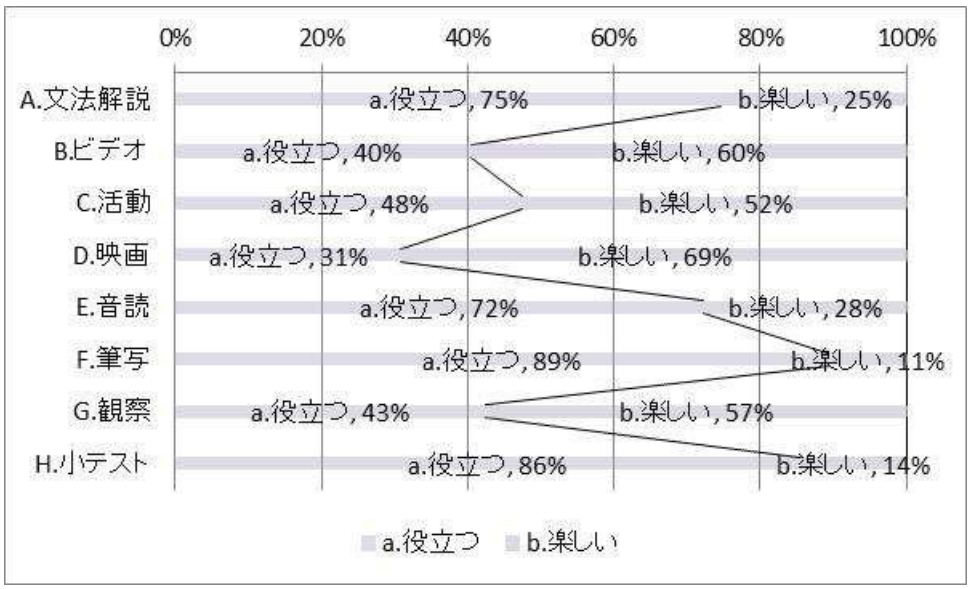
はじめに項目間(A, B, …, H)のパーセントの比較をします。

データと結果

評価	a.役立つ	b.楽しい
A.文法解説	86	29
B.ビデオ	53	78
C.活動	48	53
D.映画	43	96
E.音読	110	42
F.筆写	93	11
G.観察	37	50
H.小テスト	89	15

項目	a.役立つ	b.楽しい
A.文法解説	0.75	0.25
B.ビデオ	0.40	0.60
C.活動	0.48	0.52
D.映画	0.31	0.69
E.音読	0.72	0.28
F.筆写	0.89	0.11
G.観察	0.43	0.57
H.小テスト	0.86	0.14

＊プログラムははじめに上右表を作成し、これを参照して次のグラフを表の上に出力します。グラフをドラッグして他の場所に移動すると、表の内容を確認することができます。書式を「0%」にすると次のようにパーセント表示になります。



1.3. バブルチャート

次表の値をバブルの大きさで表示した散布図を作成します。

はじめに行と列に連番をつけ、これを標準化した値を X と Y の座標として使います。それぞれの座標に位置するデータの値を第 3 列に用意します。

項目	行	列	値
1	-1.00	-1.53	86.00
2	1.00	-1.53	29.00
3	-1.00	-1.09	53.00
4	1.00	-1.09	78.00
5	-1.00	-0.65	48.00
6	1.00	-0.65	53.00
7	-1.00	-0.22	43.00
8	1.00	-0.22	96.00
9	-1.00	0.22	110.00
10	1.00	0.22	42.00
11	-1.00	0.65	93.00
12	1.00	0.65	11.00
13	-1.00	1.09	37.00
14	1.00	1.09	50.00
15	-1.00	1.53	89.00
16	1.00	1.53	15.00

次にこれを参照してバブルチャートを出力します。

＊この図は Excel の「条件付き書式」の「データバー」（下図）とほとんど同じ情報を示しますが、列と行の参照値を座標としていることが異なります。

項目	a.役立つ	b.楽しい
A.文法解説	86	29
B.ビデオ	53	78
C.活動	48	53
D.映画	43	96
E.音読	110	42
F.筆写	93	11
G.観察	37	50
H.小テスト	89	15

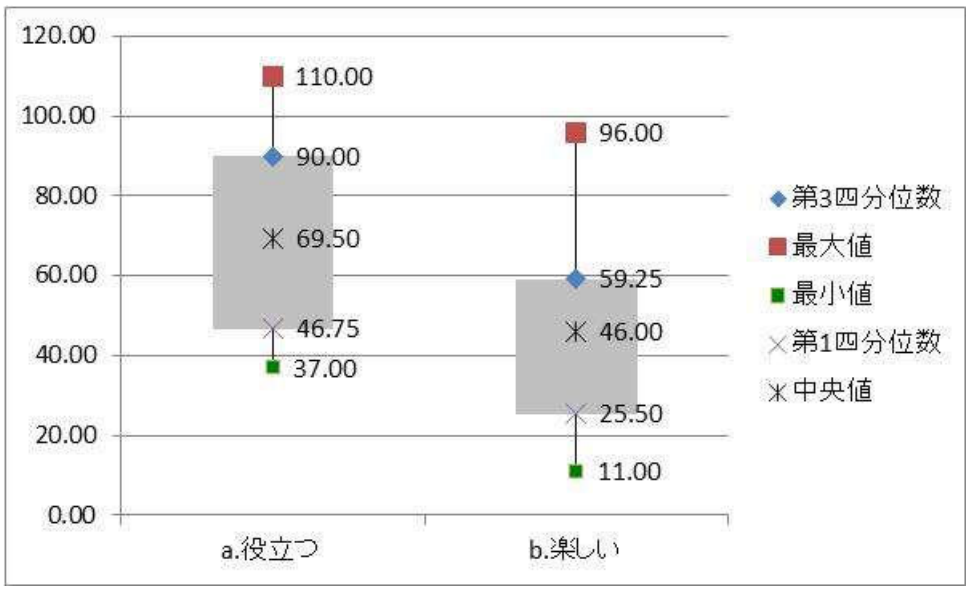
1.4. ボックスチャート

「QT ボックスチャート」は最大値、最小値、第 1 四分点、第 3 四分点、中央値を使ってデータの分布の様子を示します。四分点と中央値については→『基礎』(p.**)。ボックスチャートはデータの分布の様子（拡がりと中央値の位置）を観察するときに役立ちます。プログラムはデータ行列から次の表を作成します。

結果

要約値	a. 役立つ	b. 楽しい
第 3 四分位数	90.00	59.25
最大値	110.00	96.00
最小値	37.00	11.00
第 1 四分位数	46.75	25.50
中央値	69.50	46.00

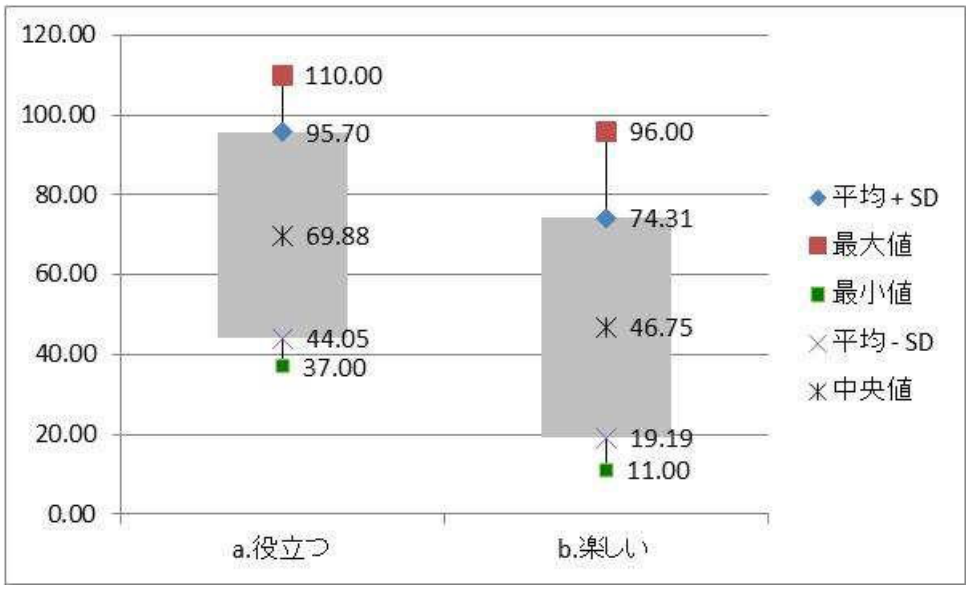
次にこれを参照してボックスチャートを出力します。



●SD ボックスチャート

「SD ボックスチャート」を選択すると要約値として平均と標準偏差(Sd)を使います。

要約値	a.役立つ	b.楽しい
平均 + Sd	95.70	74.31
最大値	110.00	96.00
最小値	37.00	11.00
平均 - Sd	44.05	19.19
平均	69.88	46.75



1.5. 二変数対比図

棒グラフの一種である対比図は棒が左右に伸びていくので、それぞれの量を比べながら観察するときに便利です。「最大値」は、セル内の最大値を超える値で切りのよい数を設定をします。ここではセルの最大値が 110 なので、グラフの最大値を 120 とします。

プログラムははじめに次のような行を反転した表を作成します。

項目	a.役立つ	b.楽しい
H.小テスト	89	15
G.観察	37	50
F.筆写	93	11
E.音読	110	42
D.映画	43	96
C.活動	48	53
B.ビデオ	53	78

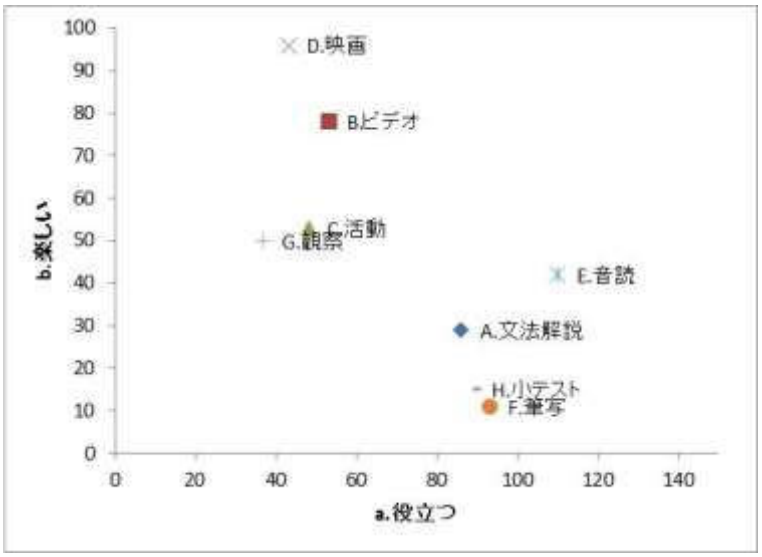
A. 文法解説	86	29

次にこれを参照して次のような二変数対比図を出力します。

結果

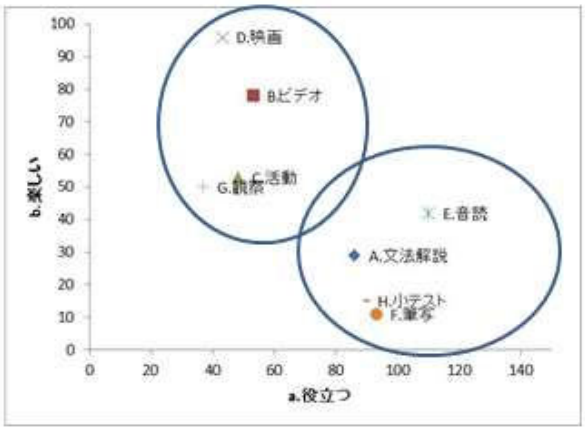
1.6. ラベル付き散布図

散布図は 2 つの変数をもつデータが 2 次元の平面上にどのような配置されるのかを見るために使います（→『基礎』p.**）。ここでは散布図の中にデータのラベル（項目名）を表示する**ラベル付き散布図**を作ります。



次は Word の「挿入」→「図」を使ってそれぞれのグループを○で囲みました。この図を見ると、それぞれの項目が「＋楽しい・－役立つ」のグループと、「－楽しい・＋役立つ」のグループに分類できることがわかり

ます。



1.7. 始点・終点棒グラフ

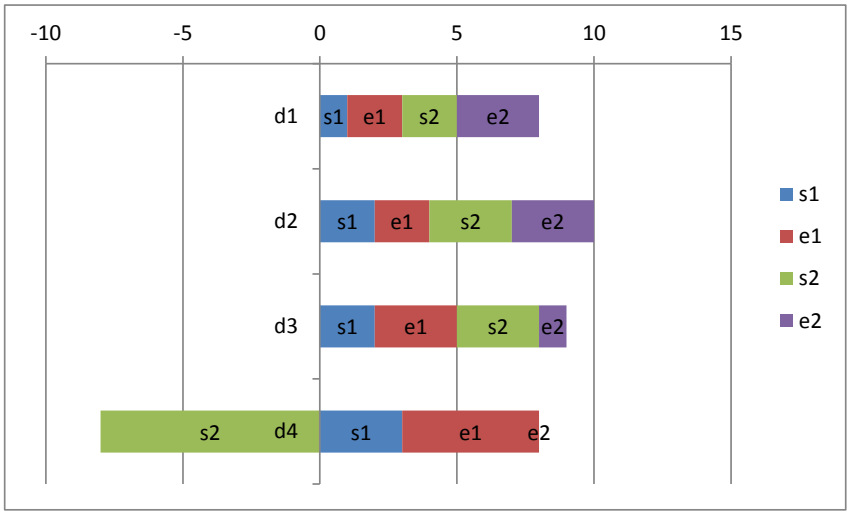
Excel の積み上げ棒グラフの奇数番号の成分の色と枠線を消すことによって、始点(s1, s2, ...)と終点(e1, e2, ...)の範囲を示すグラフを作成します[1]。プログラムで次の下左表（入力行列）から下右表（作業行列）を作成し、この作業行列を使ってグラフを出力します。下右表は、入力行列の行方向の増加分だけにした行列です。

X	s1	e1	s2	e2
d1	1	3	5	8
d2	2	4	7	10
d3	2	5	8	9
d4	3	8		

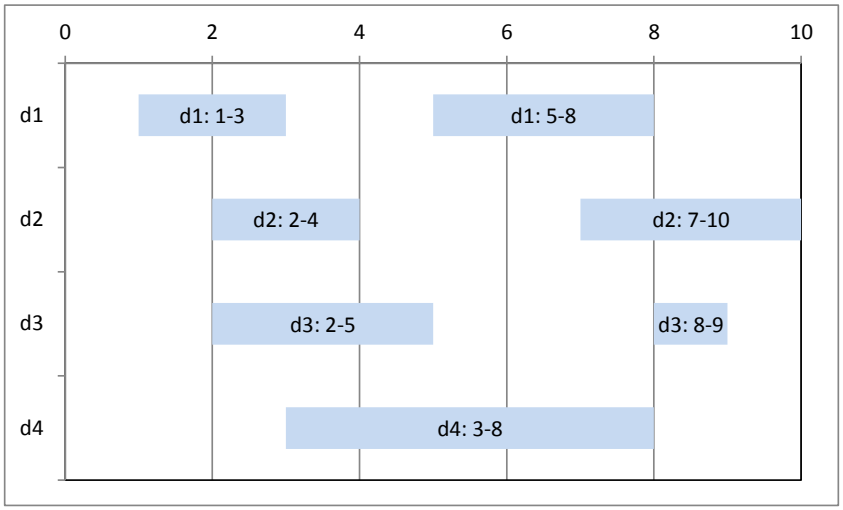
X	s1	e1	s2	e2
d1	1	2	2	3
d2	2	2	3	3
d3	2	3	3	1
d4	3	5	-8	0

Excel の棒グラフは、これを積み上げて、連続する 4 つの部分棒からなる次のような棒グラフを作ります。このとき、データの行／列を切り替え、軸を反転し、データラベルを記入します。

[1] 堀川遼太さんの創案です(2015)。



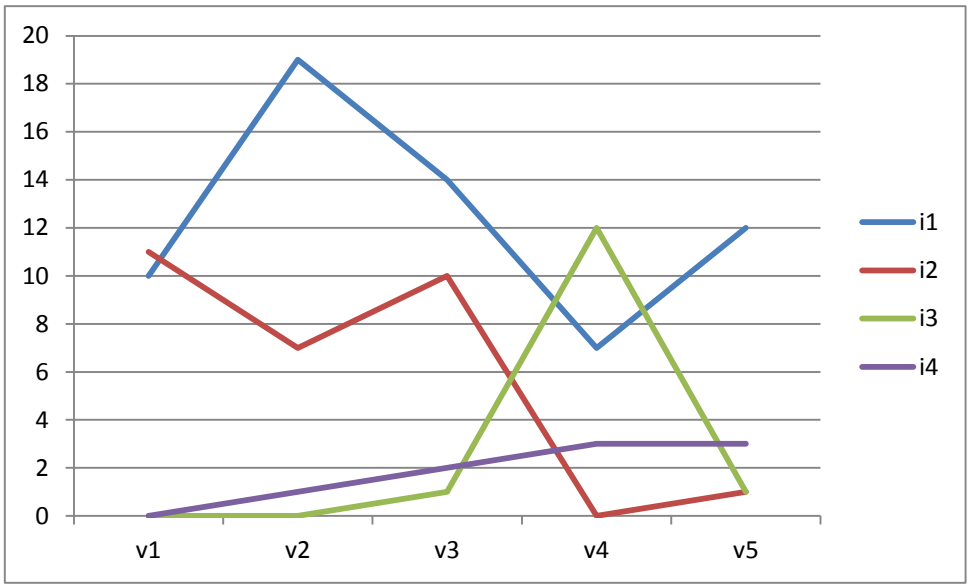
プログラムは上の s1, s2 の背景色をなくし、e1, e2 の色を統一し、さらに入力行列がゼロの領域を消して(d4:e2)、次のグラフを出力します。



1.8. 系列名付折れ線グラフ

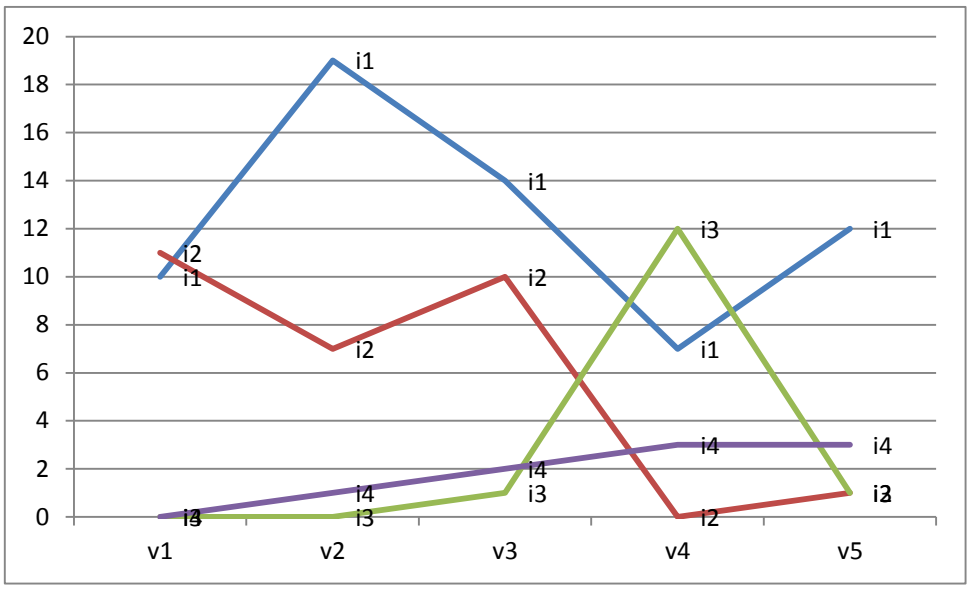
Excel 標準の折れ線グラフは、凡例に系列名を示しているため、系列が多くなると、その対応関係の読み取りが困難になります。

X	v1	v2	v3	v4	v5
i1	10	19	14	7	12
i2	11	7	10	0	1
i3	0	0	1	12	1
i4	0	1	2	3	3



そこで、折れ線を 1 つずつ選択して、「レイアウト」→「データラベル」→「その他のデータラベルオプション」→「ラベルオプション」→「系列名」をチェックし「値」のチェックを外す、という操作をしなければなりません。

VBA プログラム NUMEROS の「図」→「系列名付折れ線グラフ」は、この一連の作業を行います。



上田 (2016/4/7)