6. 関係

データ行列の変数間の関係を各種の係数を使って示します。そして、データ行列の個体間の関係を各種の距離行列を使って測ります。また、一般に行列の成分は連続変数や 1-0 という二値変数（または「v」など 1 文字の表示）になりますが、言語データ分析に欠かせない文字行列を分析する方法も考えます。

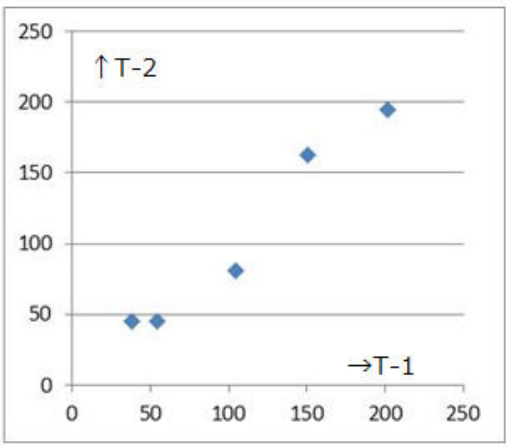
6.1. 相関

6.1.1. 相関係数

次のデータはスペイン語の T-1 (Madrid)と T-2(Sevilla)に関して主要な前置詞の頻度を集計したものです。

前置詞	T-1 Madrid	T-2 Sevilla
a	151	163
con	38	45
de	202	195
en	105	81
por	54	45

この 2 つの文は前置詞の観点からみると、どの程度連関しているのでしょうか。本節ではこのような 2 つのデータの連関の強度を計算する方法を見ていきます。はじめに 2 つのデータの関係性を捉えるために散布図にして視覚化してみましょう。



一見したところ、T-1 と T-2 は比例関係があるようです。T-1 の数値が上昇すると、それに合わせて T-2 の数値が上昇しているからです。この 2 つのテキストの連関の強度を数値化するためには、前章で見た「縦標準得

点」(Standardized Score.vartical: SSv)が使われます。これは次のようにして計算されます。

$M_{1p} = (I_{1n}\ X_{np}) / N$ ←縦平均行

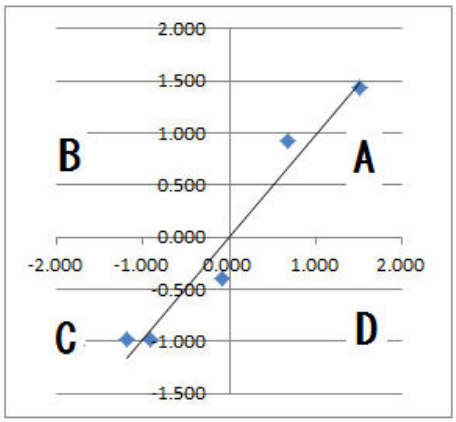
$SD_{1p} = [I_{1n}\ (X_{np} - M_{1p})^2 / N]^{1/2}$ ←縦標準偏差行

$SS = (X_{np} - M_{1p}) / SD_{1p}$ ←縦標準得点行列

次の表が縦標準得点行列です。この行列の縦平均が 0、縦標準偏差が 1 になります。

SSv	T-1 Madrid	T-2 Sevilla
a	.674	.922
con	-1.184	-.980
de	1.513	1.438
en	-.082	-.400
por	-.921	-.980

この標準得点に変換したデータで、もう一度散布図を作成すると次のようになります。



このように行の原点（ゼロの位置）を文 1 の平均までずらし、列の原点を文 2 の平均までずらした散布図になります。この図を見れば、すべてのデータがすべて A と C の領域に入っていることがはっきりと分かります。A と C の領域は、T-1 軸の値と T-2 の軸の値の標準得点を掛け合わせると、その 2 つとも正（＋）、または 2 つとも負（－）であるので、その積は正になります。一方、B と D の領域は 2 つの正負が異なるため積は負となることがわかります。

よって、X の標準得点と Y の標準得点を掛けた値の総和を求めれば X と Y の連関する度合いが数値化できます。共に正（＋）、または共に負（－）であれば、それらの積は正になりますから、この積の数が多ければ多いほど相関が強くなります。そしてすべてのデータが図中の斜線に近づけば相

関の程度はますます高くなり、全部が斜めの線に完全に一致すれば相関は最大になります。

逆に、B と D の領域にあるデータは正の相関を減少させます。それが多くなればなるほど相関の程度は弱まります。それらのデータは X と Y の値の積が負になるからです。もし、負ばかりのデータであれば、逆の相関が強くなります[1]。また、A, B, C, D に平均して分布しているとXとYの間には相関関係がない、と考えられるでしょう。

このような積の合計（積和）はデータの量に左右されます。つまり、データ量が多くなればなるほど値はどんどん大きくなり、スケールが一定になりません。そこで、積和を全体の個数(N)で割って積和の平均を出したものが「相関係数」(Coefficient of Correlation: CC)です。相関係数の求め方を一般化した公式に変えましょう。

$$CC = \Sigma_i [(X_i - Mx)/SDx]*[(Y_i - My)/SDy] / N \quad \leftarrow \text{定義}$$

$$\Sigma_i (X_i - Mx)(Y_i - My) / (N\ SDx\ SDy) \quad \leftarrow SDx, SDy \text{ を外へ}$$

$$CC = SSx_{n1}{}^{T}\ SSy_{n1} / N \quad \leftarrow \text{行列式 SS:標準得点}$$

SSc^T	a	con	de	en	por
1 Madrid	.674	-1.184	1.513	-.082	-.921

X

SSc	2 Sevilla
a	.922
con	-.980
de	1.438
en	-.400
por	-.980

/ 5

次が、その計算の過程と結果です[2]。

CC =
[(.674*.922)+(-1.184*-.980)+(1.513*1.438)+(-.082*-.400)+ (-.921*-.980) / 5
= .979

[1] 中心の点(0, 0)に近い位置のデータは、相関にあまり影響しません。逆に中心から離れた位置のデータは相関に強く影響します。

[2] ここでは例として、データ数が 5 つだけで計算しています。実際には、後述するように、このような少数のデータの分布は偶然による可能性が高いので相関係数を出す意味がありません。

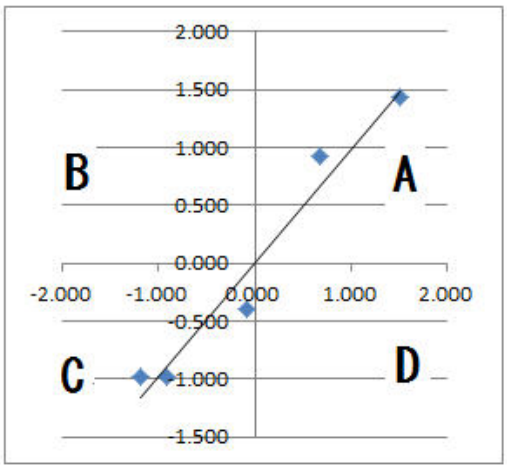
前置詞	T1 Madrid	T2 Sevilla	V1:T1-M1	V2:T2-M2	V1*V2
a	151	163	41.0	57.2	2345.2
con	38	45	-72.0	-60.8	4377.6
de	202	195	92.0	89.2	8206.4
en	105	81	-5.0	-24.8	124
por	54	45	-56.0	-60.8	3404.8
M:平均	M1: 110.0	M2: 105.8	0.0	0.0	3691.6
標準偏差	60.8	62.0			
R:	0.979				
R:Excel	0.979				

ここでは簡単のために 5 個の前置詞の頻度を使って相関係数の計算法を説明しましたが、後述するように（→「相関係数の注意」）このように少数のデータで相関係数を求めても、ほとんど意味がありません。次の「相関係数の範囲」のデータ例についても同様です。

●相関係数の範囲

相関係数の範囲は[-1 ~ 1]です。その理由を簡単に説明します。2 つの標準得点が次の図の斜線のように 1 直線に並んだときが、最大の相関係数を示します。この値は、一方の値 X_{n1} に一定の値 a を掛け、一定の値 b を足したような Y_{n1} との間の相関係数となります。

$$Y_{n1} = a\, X_{n1} + b$$



先の「標準得点の性質」で見たように、データに一定の一定の値 a を掛け、一定の値 b を足したデータの標準得点は、もとのデータの標準得点と同じ値になります。そこで、両者の相関係数は、X_{n1} と X_{n1} の間の相関係数と同じです。これは「自己相関」とよばれます。自己相関 CC(X, X)は

$$CC(X, X) = SSv_{n1}{}^{T}\, SSv_{n1} / N$$ ←相関係数の定義

$= [(X_{n1} - M) / Sd]^T [(X_{n1} - M) / Sd] / N$ ←標準得点の定義
$= \{\Sigma [(X_i - M) / Sd]^2\} / N$ ←2 乗和
$= \{\Sigma [(X_i - M)^2 / SD^2] / N$ ←乗数を分配
$= \{\Sigma [(X_i - M)^2 / V] / N$ ←分散(V) = SD^2
$= \Sigma [(X_i - M)^2 / N / V$ ←V を外へ
$= V / V = 1$ ←分散(V)の定義

先の「標準得点の性質」で見たように、a が-a のときは標準得点にすべて-1 がつくので

$$CC(X, -X) = SSx_{n1}^T (-SSvx_1) / N = -1$$

これは上の図の斜線の傾き(a)が右下がりになることを示し、このような関係は「逆相関」と呼ばれます。よって、相関係数(CC)の最小値は-1 になります。相関係数の範囲は-1 ≦ CC ≦ 1 です。

●相関係数の解釈

計算された相関係数は目安として次のような解釈されます[3]。

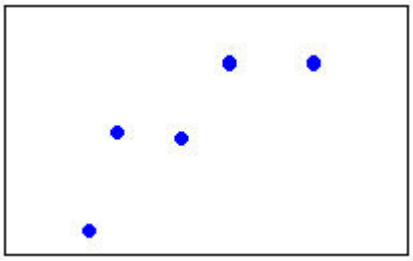
$\|r\| = 0.0$	XとYの間に相関がない
$0.0 < \|r\| \leqq 0.2$	XとYの間にほとんど相関がない
$0.2 < \|r\| \leqq 0.4$	XとYの間に弱い相関がある
$0.4 < \|r\| \leqq 0.7$	XとYの間にやや強い相関がある
$0.7 < \|r\| \leqq 1.0$	XとYの間に強い相関がある

●相関係数の注意

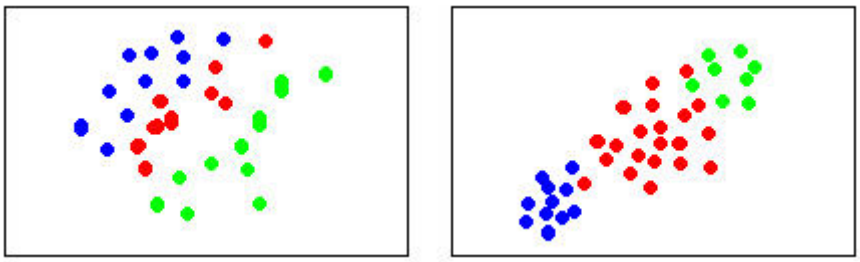
相関係数を計算することによってあらゆる数値データの間の相関関係が一応わかります。しかし、これはデータの本質については何も知らないコンピュータが、入力された数値だけをもとに出した結果にすぎないので注意が必要です。いろいろなケースが考えられますが、たとえば次のような場合に単に相関係数だけを求めて、それを現象の解釈の結論にしてしまうのは危険です。

(1) データの数が極端に少ない場合。たとえば次のように 5 つのデータだけで相関係数を出してもあまり意味はないでしょう。このような分布は偶然に生まれたのかも知れません。

[3] 相関係数の範囲は$-1 \leqq r \leqq 1$ になるので、ここではマイナスとなる逆相関も含めて絶対値|r|で示します。

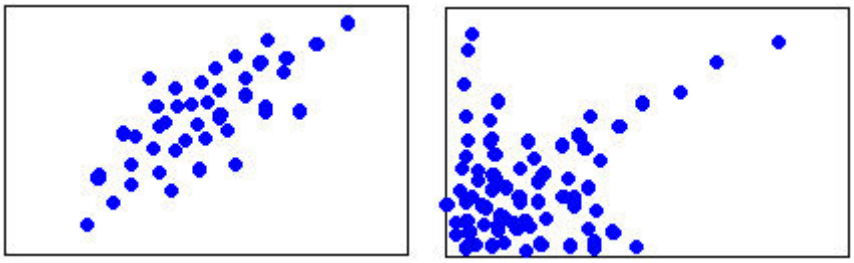


(2) 異質なデータが混在している場合。全く異なるデータを寄せ集めて相関係数を求めると、現象の正しい解釈ができないことがあります。



上左図は異質のグループを総合して判断したために、個々のグループの中では強い相関がありながら、全体としてはそれが弱くなるケースです。上右図は異質のグループの間には相関がないのに総合させると、相関らしきものが見えてしまうケースです。

(3) 大きな偏りを持つデータの場合。データの分布に大きな偏りがあるときは注意が必要です。相関係数を計算するには、一般に下左図のように平均のそばに多く分布していて、周辺に少なくなるタイプのデータが適しています。



ところが、たとえば大量のテキスト内の語彙の分布は上右図のようになるので一般に高い相関係数を示します。

下左図と下右表のデータはどちらも v1, v2 の相関係数は非常に低くなります。しかもマイナスになっているので，逆相関になっています。しかし，どちらもほとんどの値(h1-h5)は一致しているので，この結果は変だと思います。

h4a	v1	v2
h1	1	1
h2	1	1
h3	1	1
h4	1	1
h5	1	1
h6	0	1
h7	1	0

Correl. -0.1667

h4b	v1	v2
h1	0	0
h2	0	0
h3	0	0
h4	0	0
h5	0	0
h6	0	1
h7	1	0

Correl. -0.167

M	0.857	0.857
SD	0.350	0.350

M	0.143	0.143
SD	0.350	0.350

平均(M)と標準偏差(SD)を見ると,どちらも大多数の値に平均が近くなり,標準偏差はかなり大きな値になっています。このことが影響して，相関係数が低くなったことが考えられます。このような歪んだ分布（正規的でない分布）を示すデータの変数間の関係を調べるときに相関係数を使うことはできません。（ひとつの解決策として後述する距離係数を使うことが考えられます。）

このようなさまざまなケースについて正しく分析するするためには散布図をしっかり観察することが大切です。

また、相関関係が必ずしも因果関係を示しているわけではないことに注意しましょう。たとえば勉強時間と試験の成績の間に相関関係があったとても、それが必ずしも、勉強時間を増やせば試験の成績向上につながる、という「原因→結果」の関係を示していることにはならないでしょう。そこには、たとえば「教科への関心・興味」のような隠れた要素があって、それが勉強時間と試験成績のどちらにも影響していることが考えられます[4]。

相関係数の算出はあくまでも数学的な操作に過ぎません。資料の本質を知らずに計算すると意味のない分析結果を示すことにもなりかねないのです。分析者が散布図を提示せず相関係数だけを示すときはとくに注意すべきです。私たちは言語データを扱うとき、ただやみくもにデータを分析するのではなく、そのデータをしっかりと見つめること、できれば全部読むことが必要です。そうすれば、データについての理解が深まるので、変な分析結果が出てきたときには直感で気がつくはずです。しっかりとデータを読みこんでおくと、そのデータについて自分がよくわかっている、という自信につながります。自分の経験に基づいた直感と、数学的に得られたデータ分析の結果を比較しながら、一致しているかどうか、一致していな

[4] 勉強時間と試験成績というように、単位が異なっていても、また、実技テストと筆記試験のように規模（満点）が異なっていても、どちらも、標準化された値（標準得点）を比べるので、そのまま相関係数を計算することができます。

いときは何の要因がありうるか考えなければなりません。

■スペイン語の que 節と de que 節

スペイン語ではしばしば de que 節の de が省略されたり(queísmo と呼ばれる)、逆に他動詞の que 節に de が付加されたりすることがあります(de queísmo)。次は VARIGRAMA 研究計画のアンケート調査資料から、使用される queísmo の例(*estoy seguro que*「私は…が確かだと思う」, *está contenta que*「彼女は…であることに満足している」)と dequeísmo の例(*sospecho de que*「私は…を疑う」)について、スペインの各地での使用率(%)を示す表です。

España (%)	SAL	HUE	ALC	SEV	PAM	TEN	MAD	BAR	LPA	OVI	Total
estoy seguro que	15.8	72.2	45.5	41.7	38.1	35.0	33.3	30.0	25.0	22.2	34.6
está contenta que	15.8	38.9	18.2	12.5	23.8	5.0	14.3	25.0	4.2	7.4	15.6
sospecho de que	10.5	11.1	9.1	4.2	9.5	0.0	9.5	0.0	16.7	0.0	6.8

SAL: Salamanca, HUE: Huelva, ALC: Alcalá de Henares, SEV: Sevilla, PAM: Pamplona, TEN: Tenerife, MAD: Madrid, BAR: Barcelona, LPA: Las Palmas, OVI: Oviedo

これを *estoy seguro* の使用率をキーにして降順に並べると次のようになります。

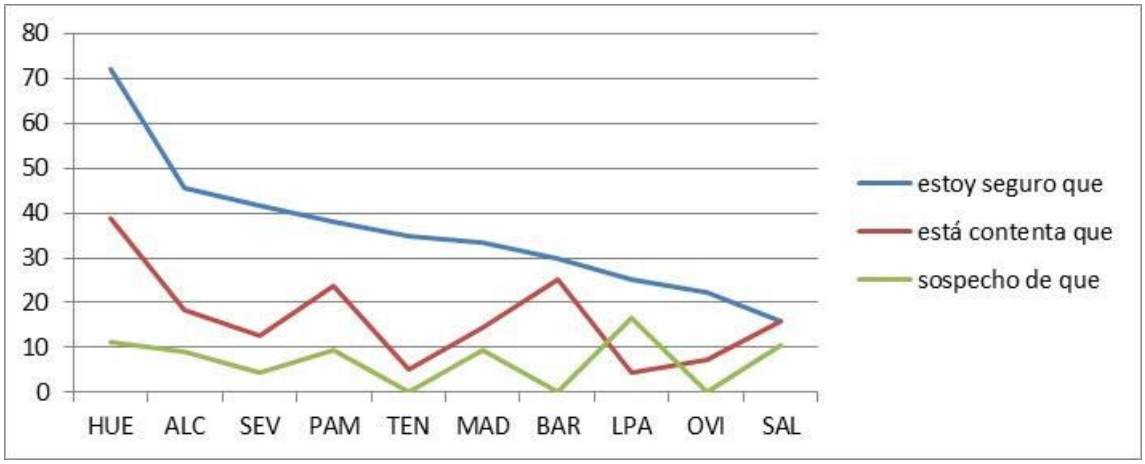


Fig. 4. Porcentaje. Respuestas afirmativas en ciudades españolas

上図から、queísmo の 2 例(*estoy seguro que, está contenta que*)の間には相関があるように見えますが、それらと de queísmo (*sospecho de que*)の間にはあまり相関がないように見えます。

次は同じアンケート調査をラテンアメリカの各都市で行った結果です。

América % (1)	PN	PR	CO	PE	BO	MX	PA	UR
estoy seguro que	97.1	90.9	88.0	82.6	80.0	76.2	75.0	75.0
está. contenta que	80.0	68.2	64.0	65.2	55.0	52.4	60.0	55.0
sospecho de que	37.1	27.3	8.0	30.4	42.5	19.0	45.0	5.0

América % (2)	CU	CH	AR	RD	VE	EC	CR	Total
estoy seguro que	73.7	72.0	70.0	68.6	68.0	65.2	39.1	75.5%
está. contenta que	42.1	44.0	60.0	51.4	40.0	52.2	21.7	54.8%
sospecho de que	5.3	20.0	10.0	40.0	12.0	17.4	13.0	24.2%

PN: Panamá (Panamá), PR: San Juan (Puerto Rico), CO: Bogotá (Colombia), PE: Lima (Perú), BO: La Paz (Bolivia), MX: Ciudad de México (México), PA: Asunción (Paraguay), UR: Montevideo (Uruguay), CU: La Habana (Cuba), CH: Santiago (Chile), AR: Buenos Aires (Argentina)., RD: Santo Domingo (República Dominicana), VE: Caracas (Venezuela), EC: Quito (Ecuador), CR: San José (Costa Rica)

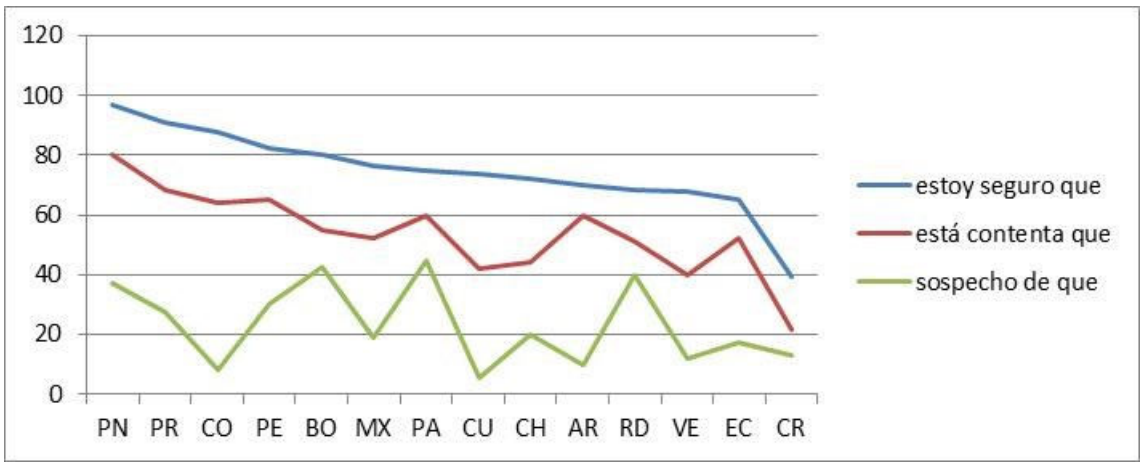


Fig. 5. Porcentaje. Respuestas afirmativas en América

やはり queísmo の 2 例(*estoy seguro que, está contenta que*)の間には相関があるように見えますが、それらと de queísmo (*sospecho de que*)の間にはあまり相関がないようです。先のスペインの調査よりもラテンアメリカの調査のほうが、その傾向が鮮明に表れています。

従来の文法研究では queísmo と dequeísmo は前置詞が省略されたり、付加されたりする、という同レベルの文法の誤用の問題として扱われてきましたが、このデータを見ると、2 つの文法現象の発生は地理的に相関していないので、両者の要因は異なる、という可能性が高いと思います。

次の表は 3 者間の相関係数を求めた結果を示します。

España	*seguro*	*contenta*	*sospecho*	América	*seguro*	*contenta*	*sospecho*
seguro	1.000	.709	.148	***seguro***	1.000	.900	.332
contenta	.709	1.000	.150	***contenta***	.900	1.000	.443
sospecho	.148	.150	1.000	***sospecho***	.332	.443	1.000

このように、少数の変数（queísmo と dequeísmo の 3 例）であれば折れ線グラフを使って相関を視覚化することができます。

3 者間であれば、個別に変数のペアを作って、それぞれの相関係数を求めることができますが、多数の変数を扱うデータでは次に説明する「相関行列」を作成しなければなりません。

6.1.2. 相関行列

多変数間の相関係数を一度に示す「相関行列」(R_{pp}：下右表)を出力します。

D_{np}	v1	v2	v3
d1	45	48	66
d2	56	59	54
d3	58	51	78
d4	77	72	20
d5	43	44	32
d6	58	34	90
d7	50	53	100

Z_{np}	v1	v2	v3
d1	-.980	-.323	.115
d2	.068	.673	-.324
d3	.259	-.052	.554
d4	2.068	1.850	-1.569
d5	-1.170	-.686	-1.130
d6	.259	-1.591	.994
d7	-.504	.129	1.360

R_{pp}	v1	v2	v3
v1	1.000	.643	-.335
v2	.643	1.000	-.545
v3	-.335	-.545	1.000

はじめに、データ行列(D_{np}：下左表)から標準測度行列(Z_{np}：下中表)を作成します。

$$Z_{np} = (D_{np} - M_{1p}) / S_{1p}$$

ここで、M_{1p} は D_{np} の縦平均行を示し、S_{1p} は D_{np} の縦標準偏差行を示します。この標準測度行列(Z_{np})を掛け合わせて積和の正方対称行列を作り、個数(N)で割って平均を出したものが相関行列(R_{pp})です。

$$R_{pp} = Z_{np}{}^{T} Z_{np} / N$$

この式は重要なので上の例で行列の成分を確かめておきましょう。

$$Z_{np}{}^{T} Z_{np} = \begin{bmatrix} -0.98 & 0.07 & \dots & -0.50 \\ -0.32 & 0.67 & \dots & 0.13 \\ 0.12 & -0.32 & \dots & 1.36 \end{bmatrix} \begin{bmatrix} -0.98 & -0.32 & 0.12 \\ 0.07 & 0.67 & -0.32 \\ \dots & \dots & \dots \\ -0.50 & 0.13 & 1.36 \end{bmatrix}$$

$$= \begin{bmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \end{bmatrix}$$

行列積の演算により

r_{11} = -0.98*-0.98 + 0.07*0.07 + … + -0.50*0.50 ≒ 7.00
r_{12} = -0.98*-0.32 + 0.07*0.67 + … + -0.50*0.13 ≒ 4.50
r_{13} = -0.98*0.12 + 0.07*-0.32 + … + -0.50*1.36 ≒ -2.34
r_{21} = -0.32*-0.98 + 0.67*0.07 + … + 0.13*0.50 ≒ 4.50
r_{22} = -0.32*-0.32 + 0.67*0.67 + … + 0.13*0.13 ≒ 7.00
r_{23} = -0.32*0.12 + 0.67*-0.32 + … + 0.13*1.36 ≒ -3.82
r_{31} = 0.12*-0.98 + -0.32*0.07 + … + 1.36*0.50 ≒ -2.34
r_{32} = 0.12*-0.32 + -0.32*0.67 + … + 1.36*0.13 ≒ -3.82
r_{33} = 0.12*0.12 + -0.32*-0.32 + … + 1.36*1.36 ≒ 7.00

このように R_{pp} がすべての成分が積の和になること、対角成分がそれぞれの列の 2 乗和になること、非対角成分が該当する 2 つの列の成分の積の和になること、全体の行列の形が対称行列であること、そして行列の大きさが[3 行 7 列]x[7 行 3 列]の積なので[3 行 3 列]になることを確認しましょう。

■語頭の無強勢 e-と語末の無強勢-e

ラテン語の語頭の「s＋子音」(sC-)はスペイン語になると、たとえば stare > estar, scribere > escribir のように「es＋子音」(esC-)となって、語頭に e を付加しました。しかし、この現象は中世スペイン語でとくにスペイン東部のナバーラ・アラゴン地方では比較的少数でした(sC-)。一方、中世スペイン語の語末母音が 2 子音の後で脱落した現象(-CC)も、とくにナバーラ・アラゴン地方に多く見つかります。次表の左 3 列は、旧カスティーリャ(CV)・ナバーラ(NA)・アラゴン(AR)で発行された公証文書に現れた(e)star と(e)scribir とその派生形の出現数を示します。右 3 列は-CC の後で-e が脱落した語数(present(e), veint(e), adelant(e), part(e), est(e), end(e))です。

年 : Año	CV:sC-	NA:sC-	AR:sC-	CV:-CC	NA:-CC	AR:-CC
1200						
1220			4	1		1
1240			4	8		7
1260		5	9	5	22	13
1280		8	1	5	27	8
1300		8	3	2	34	8
1320	3	2	2		10	4

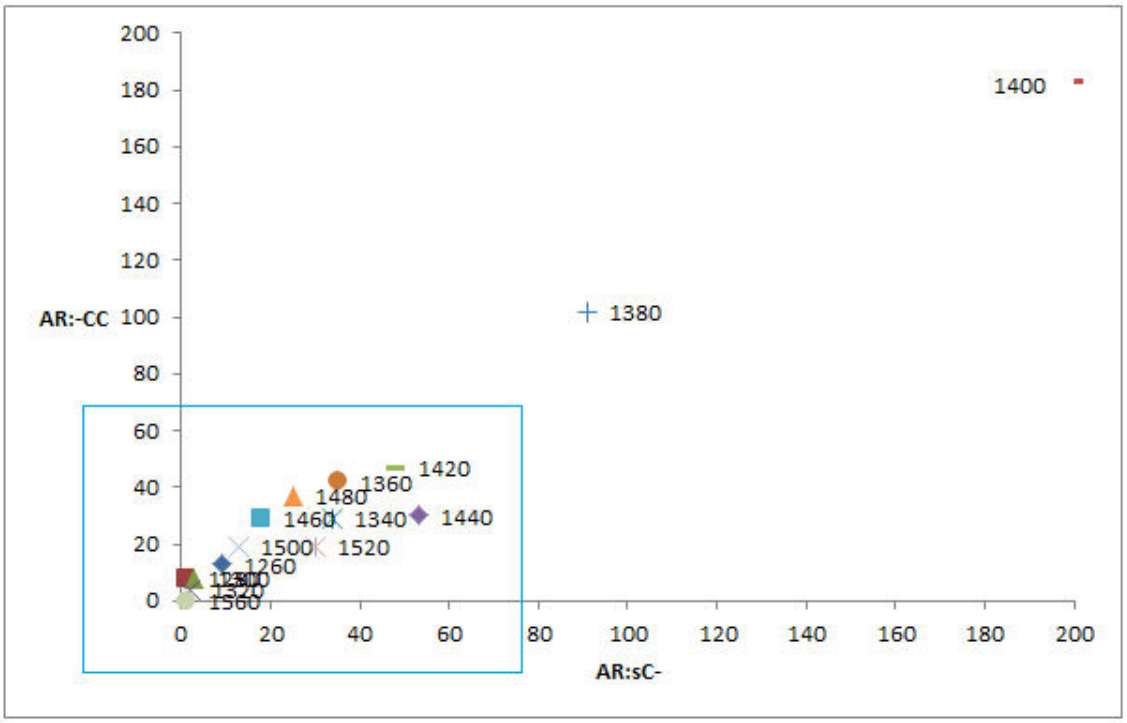
1340	1	1	34	1	25	29
1360		1	35		25	42
1380	4	2	91	2	2	102
1400		12	200	3	35	183
1420	4		48			47
1440			53	1	15	30
1460			18			29
1480	5	1	25	3	4	37
1500	3		13	1		19
1520	5		30			19
1540	5					
1560	19		1			
1580	35					
1600	1					
1620	9					
1640	4					
1660						
1680	4					

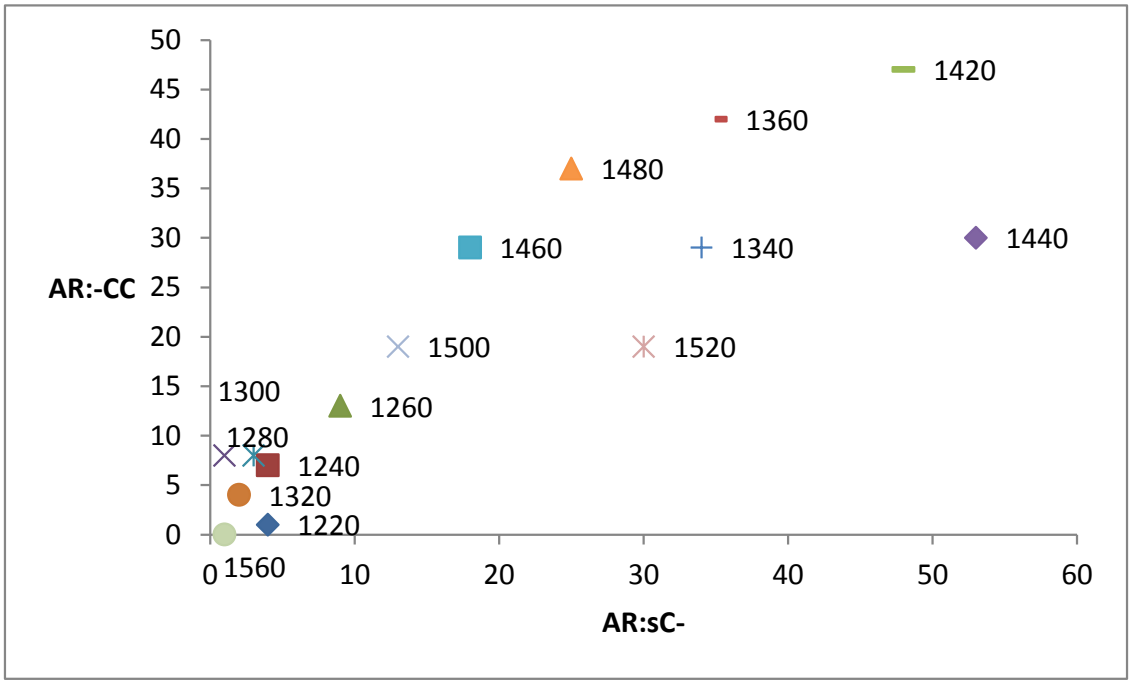
この 2 つはどちらも無強勢の母音 e に関わる現象ですが、両者間に通時的な相関関係があるのでしょうか？次は上表から計算した相関行列です。

CC	CV:sC-	NA:sC-	AR:sC-	CV:-CC	NA:-CC	AR:-CC
CV:sC-	1.000	-.244	-.176	-.272	-.323	-.185
NA:sC-	-.244	1.000	.557	.465	.829	.574
AR:sC-	-.176	.557	1.000	.148	.441	.984
CV:-CC	-.272	.465	.148	1.000	.360	.188
NA:-CC	-.323	.829	.441	.360	1.000	.435
AR:-CC	-.185	.574	.984	.188	.435	1.000

はたして、NA でも AR でも sC-と-CC の間には強い相関があるようです。CV では相関しません。次の 2 図はアラゴン地方の sC-と-CC の散布図です。最初の図を見ると、1380, 1400 のデータが強く作用して、大きな相関係数(.984)を生んでいることがわかります。しかし、これらの外れ値を除いてもやはり相関が高いことが 2 番目の図からも、相関係数(外れ値を除いた相関係数は.863)からもわかりました。

従来の説では、極端な語末母音の脱落は当時の 13 世紀はじめのフランス人越境者がカスティーリャに多かったことの影響によるものである、と説明されていましたが、年代的にも(14-15 世紀に多い)、地理的にも(CV よりも NA, AR に多い)、そして言語現象の相関関係からも(無強勢の e)、再考の余地があると思います。





■中世カスティーリャ語の 2 連続子音文字

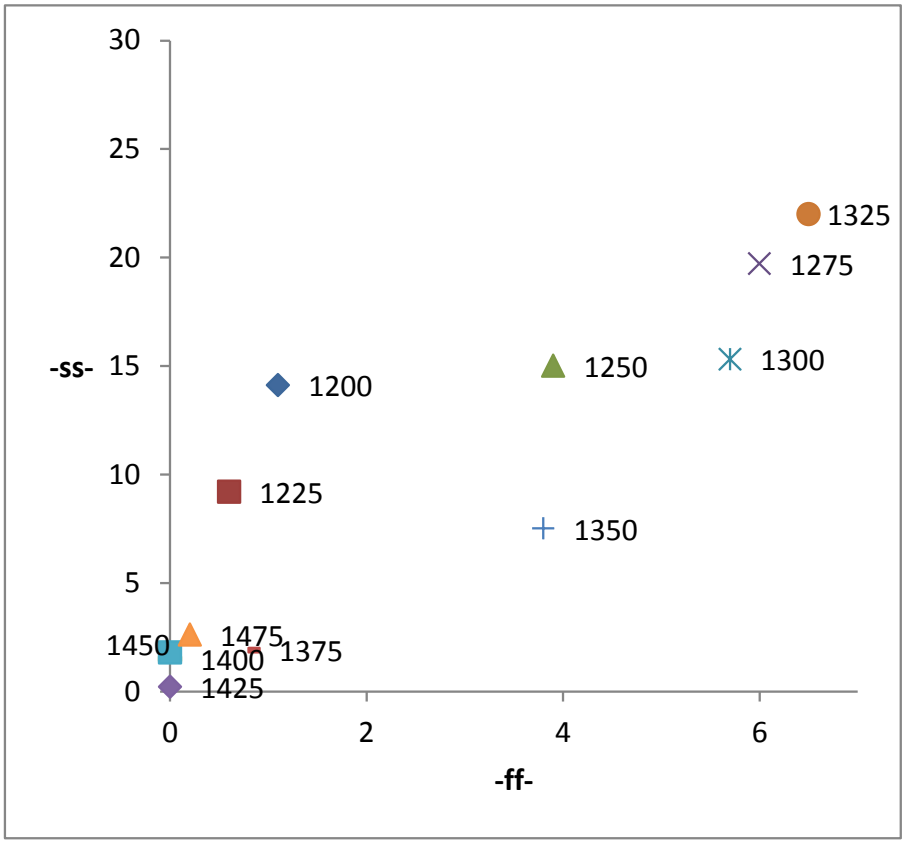
現代スペイン語では語中に-ll-, -rr-という 2 連続子音文字(CC)が用いられ、語頭では#ll-が用いられます。それが中世スペイン語では、さらに ff, ss, ll, rr が語頭でも語中でも使われていました。次の表は、中世公証文書において CC を含む語の頻度を千語率(1000 語あたりの相対頻度)にして計算した結果です。

CC	-ff-	#ff-	-ss-	#ss-	-ll-	#ll-	-rr-	#rr-
1200	1.1		14.1		31.0		16.9	
1225	.6	4.6	9.2		28.9		13.7	
1250	3.9	11.2	15.0	1.4	31.1	.2	9.5	.4
1275	6.0	16.2	19.7	15.9	33.8	.8	10.0	3.5
1300	5.7	20.9	15.3	19.7	28.0	.8	9.6	9.3
1325	6.5	37.2	22.0	45.5	30.5	1.5	16.3	8.2
1350	3.8	17.2	7.5	25.2	24.4	.3	11.4	12.2
1375	.8	5.6	1.9	5.1	27.5	1.7	14.5	12.9
1400		.9	1.5	1.1	27.5	1.1	7.4	4.2
1425		.2	.2		23.4	2.0	10.1	8.3
1450		.1	1.8		25.1	1.2	7.7	2.6
1475	.2	.3	2.6		24.5	2.5	7.2	2.1

上表のそれぞれの CC 間の相関行列が次の表です。

相関係数	-ff-	#ff-	-ss-	#ss-	-ll-	#ll-	-rr-	#rr-
-ff-	1.000	**.913#**	**.874#**	**.816#**	.573+	-.256^	.214^	.255^
#ff-	**.913#**	1.000	**.763#**	**.956#**	.393^	-.104^	.341^	.407^
-ss-	**.874#**	**.763#**	1.000	.636+	**.817#**	-.455^	.442^	-.141^
#ss-	**.816#**	**.956#**	.636+	1.000	.233^	-.003^	.367^	.519+
-ll-	.573+	.393^	**.817#**	.233^	1.000	-.506+	.404^	-.373^
#ll-	-.256^	-.104^	-.455^	-.003^	-.506+	1.000	-.322^	.325^
-rr-	.214^	.341^	.442^	.367^	.404^	-.322^	1.000	.143^
#rr-	.255^	.407^	-.141^	.519+	-.373^	.325^	.143^	1.000

上表を見ると、ff と ss の相関が高いことがわかります。次の散布図は、語中の-ff-と-ss-の相関の様子を示しています。



6.1.3. 共分散行列

次の右上表(V.Cov)は**共分散行列**とよばれる行列で、その対角成分にそれぞれの列の分散が配置され、非対角成分に該当する変数どうしの**共分散**(Covariance: Cov)が配置されています。

D_{np}	v1	v2	v3
d1	45	48	66
d2	56	59	54
d3	58	51	78
d4	77	72	20
d5	43	44	32
d6	58	34	90
d7	50	53	100

V.Cov	v1	v2	v3
v1	110.204	74.551	-95.959
v2	74.551	121.959	-164.490
v3	-95.959	-164.490	746.122

値	v1	v2	v3
分散	110.204	121.959	746.122

共分散は次の式で計算されます。

$$Cov = \Sigma_i [(X_i - AveX) (Y_i - AveY)] / N$$

ここで X_i は X 列のデータ、AveX は X 列の平均、Y_i は Y 列のデータ、AveY は Y 列の平均、N はデータ数を示します。次は共分散行列(Rpp)を返すプログラムの主要部です。

```
Wnp = S(Xnp, AveV(Xnp))          '縦偏差行列
Rpp = D(X(Tr(Wnp), Wnp), N)      '共分散行列
```

相関係数の分子に使われています。共分散行列は実際のデータ分析であまり使われる機会がありませんが、多変量解析の導出過程の確認で必要になることがあります。

6.2. 連関

6.2.1. 連関係数

言語データとして、数値データ（量的データ）ではなく、+/－や「v」印で示されるような二値データ（質的データ）を扱うことがあります。たとえば、次の表では「手紙」と「演劇」で共にプラスになっている語は *abajo*, *abandonar*, *abeja*, *abogado* の 4 語です[5]。これは「共起回数」(Cooccurrence) とよばれます。共起回数はデータの規模に左右されるので、これを標準的な値にするためにいろいろな方法が提案されてきました。ここでは、2 つのデータ（たとえば、「手紙」と「演劇」）が連関している度合いを数値化するための 7 つの係数を紹介します。

語	手紙	演劇		手紙	演劇		a (+/+)	b (+/-)	c (-/+)	d (-/-)
abajo	5	10		+	+		1	0	0	0
abandonar	9	6		+	+		1	0	0	0
abandono	0	0		-	-		0	0	0	1
abarcar	1	0		+	-		0	1	0	0
abastecimiento	2	0		+	-		0	1	0	0
abatir	0	1		-	+		0	0	1	0
abeja	2	3		+	+		1	0	0	0
abertura	0	0		-	-		0	0	0	1
abismo	0	0		-	-		0	0	0	1
abnegación	0	0		-	-		0	0	0	1
abogado	3	6		+	+		1	0	0	0
abonar	5	0		+	-		0	1	0	0
abono	0	0		-	-		0	0	0	1
abordar	0	0		-	-		0	0	0	1
aborrecer	0	6		-	+		0	0	1	0

次のような 2 × 2 の表を作り、それぞれ a, b, c, d の 4 つのマス目の値を考慮します。a は x も y も「有」(=1)の個数です。b は x が「有」(=1)か

[5] データは次を参照しました。A. Juilland y E. Chang Rodríguez en su *Frequency dictionary of Spanish words,* (The Hague: Mouton, 1964).

つ y が「無」(=0)のとき、c は x が「無」(=0)かつ y が「有」(=1)のとき、そして d は *x* も *y* も「無」(=0)の個数です。たとえば先の図のデータではとなります。

X / Y	Y (X)	Y (-)
X (+)	a (X+, Y+) 4	b (X+, Y-) 3
X (-)	c (X-, Y+) 2	d (X-, Y-) 6

「連関係数」(Coefficient of Association: CA)はこれらの数値(a, b, c, d)を利用します。d を使わない係数もあります。連関係数全体についてほぼ共通していることは、どちらにも共通する肯定的要素(a)と、どちらにも共通している否定的要素(d)の数が多ければ多いほど、連関係数は大きくなる、ということです。逆に一方だけにある要素の数(b, c)が大きくなればなるほど、連関係数は小さくなります。以下の 7 つは、その連関度を正規化した数値として求めるために考案された係数です。

(1) **「単純一致係数」**(Simple Matching coefficient: S)では、対象 X と対象 Y に共通して「+」がある回数(a)と、それが共に存在しない回数(d)の和を全体の数で割ります。a = d = 0 のとき最小値 0 になり、b = c = 0 のとき最大値 1 になります。

$$S = (a + d) / (a + b + c + d) \quad 0 \leqq SM \leqq 1$$

(2) **「Jaccard 係数」**(J)は分子にも分母にも d を使いません。a = 0 のとき最小値 0 になり、b = c = 0 のとき最大値 1.0 になります。

$$J = a / (a + b + c) \quad 0 \leqq J \leqq 1$$

(3) **「Dice 係数」**(D)は Jaccard 係数の a を 2 倍にしたものです。a = 0 のとき最小値 0 になり、b = c = 0 のとき最大値 1 になります。(→後述)

$$D = \quad 2a / (2a + b + c) \quad 0 \leqq D \leqq 1$$

(4) **「Yule 係数」**(Y)は a*d と b*c の差を扱います。(1)の単純一致係数では a と d を足していますが、Yule 係数では掛けます。それから分子は a*d と b*c の差なので、それがマイナスになることもあります。a*d = 0 のとき最小値 -1 になり、b*c = 0 のとき最大値 1 になります。a*d = b*c のときは中間値 0 になります。a, b, c, d のいずれかが 0 のとき、結果に大きく影響します。

$$Y = (ad - bc) / (ad + bc) \quad -1 \leqq Y \leqq 1$$

(5) **「Hamann 係数」**(H)は a + d と b + c の差を問題にします。Yule 係数では a と d, b と c の関係を積で示しますが、Hamann 係数ではそれを和で示

しています。a = d = 0 のとき最小値 -1 になり、b = c = 0 のとき最大値 1 になります。a + d = b + c のときは中間値 0 になります。

$$H = [(a+d) - (b+c)] / [(a+d) + (b+c)]$$

$$-1 \leqq H \leqq 1$$

(6) 「Phi **係数**」(P)は少し複雑な式です。これは積率相関係数と一致します。(→後述)

$$P = (ad - bc) / [(a + b)(c + d)(a + b)(c + d)]^{1/2} \qquad -1 \leqq Ph \leqq 1$$

(7) 「Ochiai **係数**」(O)は、a / (a + b)と a / (a + c) の幾何平均です。それぞれの a の比率に注目しています。

$$O = a / [(a + b)(a + c)]^{1/2} \qquad 0 \leqq O \leqq 1$$

(8) 最後に次の係数(Ueda: U)を提案します。U は後述するように他の係数と比較して利点が多いからです。

$$U = [2a - (b + c)] /[2a + (b + c)] \qquad [-1\ (a=0),\ 1\ (b=c=0)]$$

●連関係数の比較

実際の分析でこれらの連関度係数のうちどれを使えばよいのか迷うことがあります。そのとき、いくつかの選択の方法が考えられるでしょう。その選択の基準もさまざまです。たとえば、これらの係数を利用して誰かの前で発表することを考えてみましょう。発表の目的が係数の数値自体によって裏づける根拠よりも、その先にある連関性を主張することであるならば、S や J のように係数の説明に多くの時間を割かずに済む、わかりやすい係数を選択するという決定も考えられます。連関係数が強い裏付けの根拠として重要な意味を持つならば、Y や H を選択し、その数値の性質について丁寧な説明が必要になります。そして、統計に慣れている人に発表するならば、よく知られている P を使えばその説明は必要なくなります。P にわずかな説明を加えることで O を使うこともできるでしょう。(→後述)
1 つだけでなく複数の係数を選択して、それぞれを比較し、考察することも考えられます。

しかし、このような決定は本質的ではなく、実際的な条件に従っています。本質を追究するには、それぞれの係数の性質と分析対象のデータの性質をよく理解して、本質的な条件と実際的な条件のどちらも考慮に入れた上で決定しなくてはいけません。そうすれば自分でも納得ができますし、自信をもって説明できます。

それぞれの係数の性質を比べると、共通する性質があることがわかります。「両者に存在しない特徴(d)」の扱いのほかに、逆方向を検知するかど

うか(マイナスになるか)、完全に等質な分布のときゼロになるかどうか、などについて、しっかり理解しておく必要があります。次の表はそれぞれの特徴の分布を比較したものです。

性質	S	J	Y	H	P	O	U
d (-/-)を扱う	+	-	+	+	+	-	-
逆方向(-)を検知	-	-	+	+	+	-	+
積算がある	-	-	+	-	+	+	+
振幅	-	-	++	-	-	+	+

ここで、たとえばd値(-/-)を扱わない(-)、逆方向を検知する(v)、積算がない(-)、という条件をつけるならばUを選択するとよいでしょう。

データの性質として、方向性があるものならば、d(-/-)を探知する係数を選択すべきです。たとえば「賛成」と「反対」で回答したアンケート調査などは、「賛成」の数だけでなく「反対」の数も考慮に入れるべきです。一方、2 つの文献の語彙比較調査などは、ある単語が使われている、と、使われていない、という数値を同等に扱うよりも、使われているケースだけで計算したほうがよいと思われます。どちらにも使われていない、という語彙は無限に存在するからです。しかし、一定の語彙範疇(たとえば「指示詞」「関係代名詞」など)で複数の文献を調査するときは、否定的な反応も考慮に入れるべきでしょう。

逆方向(-)を検知する係数(Y, H, P, U)は範囲が[-1, 1]で、完全に等質な分布のとき中間値のゼロ(0)になります。他の係数では、そのとき、0.5 (S, O), 0.33 (J) になる、ということを心得ておかなければなりません。たとえば、相関係数が 0.5 ならば「中度の相関がある」と判断しますが、それが SM や O の値ならばまったく相関がないことを示しています。

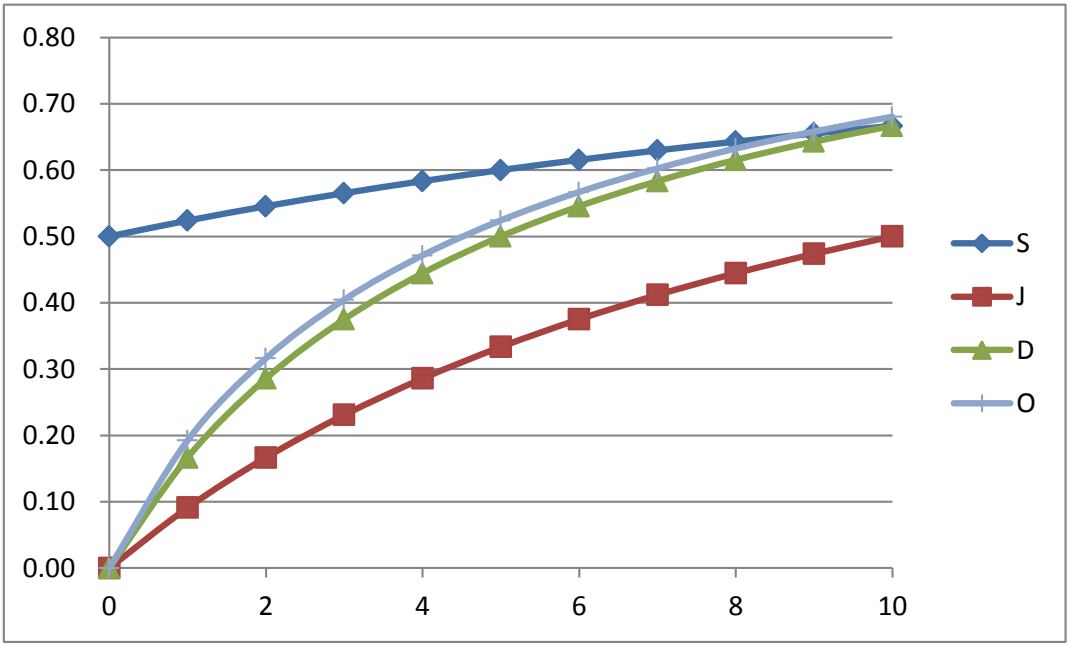
数値を積算している係数(Y, P, U)は、それぞれの項目の増減がそれを構成する要素の増減に比例しているので、考えてみると納得できますが、問題点として積算の片方がゼロになると他方にどのような数値があっても、ゼロになってしまうことがあげられます。また、分母で積算されているとそれがゼロになったとき計算できなくなります。たとえばOで(a+b)がゼロになった場合です。このときcに値があっても計算されません。一方、数値を積算していない係数は、結局「割合」に過ぎないので、ほとんど考えなくてもわかります。これが実際的な選択の条件となることもあるかもしれません。

次の表と図は b=5, c=4, d=10 で固定し、共起回数(a)を 0 から 10 に上げていったときのそれぞれの係数の変化を示しています。

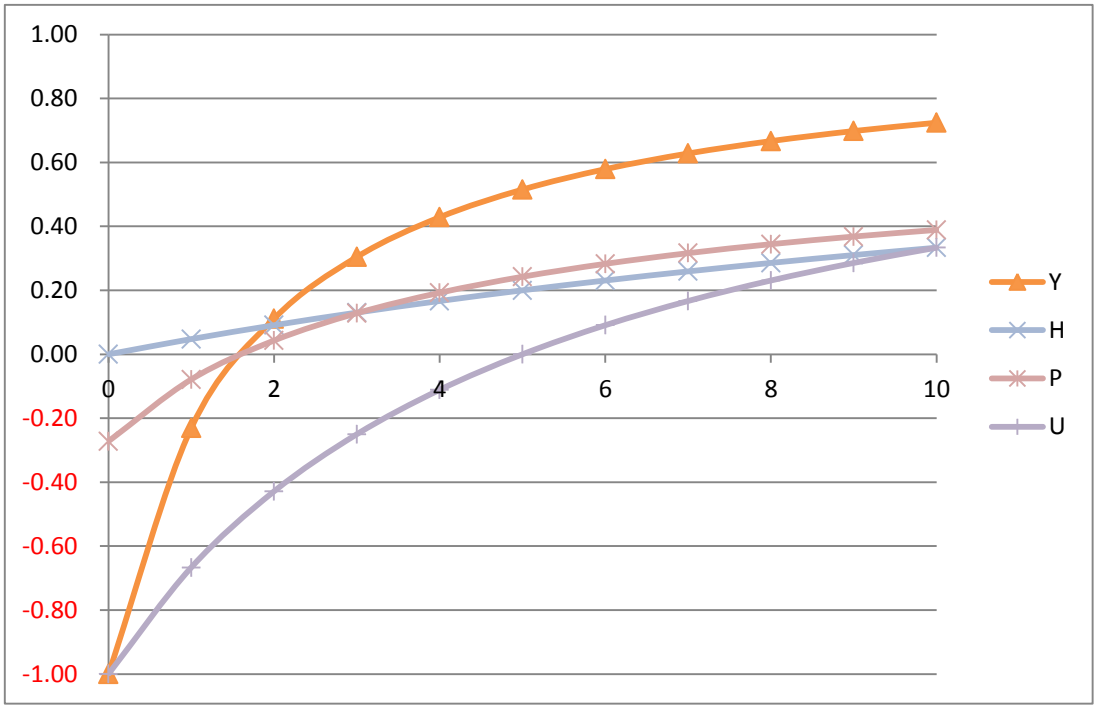
a(+/+)	0	1	2	3	4	5	6	7	8	9	10
b (+/-)	8	8	8	8	8	8	8	8	8	8	8
c(-/+)	2	2	2	2	2	2	2	2	2	2	2

d(-/-)	10	10	10	10	10	10	10	10	10	10	10
S	0.50	0.52	0.55	0.57	0.58	0.60	0.62	0.63	0.64	0.66	0.67
J	0.00	0.09	0.17	0.23	0.29	0.33	0.38	0.41	0.44	0.47	0.50
D	0.00	0.17	0.29	0.38	0.44	0.50	0.55	0.58	0.62	0.64	0.67
Y	-1.00	-0.23	0.11	0.30	0.43	0.52	0.58	0.63	0.67	0.70	0.72
H	0.00	0.05	0.09	0.13	0.17	0.20	0.23	0.26	0.29	0.31	0.33
P	-0.27	-0.08	0.04	0.13	0.19	0.24	0.28	0.32	0.34	0.37	0.39
O	0.00	0.19	0.32	0.40	0.47	0.52	0.57	0.60	0.63	0.66	0.68
U.	-1.00	-0.67	-0.43	-0.25	-0.11	0.00	0.09	0.17	0.23	0.29	0.33

次の図は相対値[0, 1]の係数の比較です。これを見ると、S と J の振幅が比較的小さく、とくに S の振幅が小さいことが確認できます。一方、Ochiai の振幅は大きいことがわかります。係数の振幅が大きいことは弁別する力が強いことを示します。



両側相対値[-1, 1]の係数については Y と U の振幅が大きいことが特徴的です。Y の上昇は急ですが、U は比較的緩やかに上昇します。このことは a[++]の値が大きい場合の Y の弁別力が弱くなりますが、U は比較的直線に近いので一定した高い弁別性を保証します。U は d を入れないので d が大きくても影響されません。Y の高い上昇値は d の影響によるものです。



●相関係数と Phi 係数

Phi 係数は「有(＋)」を 1,「無(-)」をゼロ(0)とすれば、一般の連続量を扱う相関係数から導出できます。

X:Y	Y = 1	Y = 0	和
X = 1	a (1,1)	b (1,0)	a + b
X = 0	c (0,1)	d (0,0)	c + d
和	a + c	b + d	N: a + b + c + d

はじめに総データ数を N とします。

[1]　　N = a + b + c + d

先に見たように相関係数(CC)の式は次の通りです。

$$CC = \Sigma i\,(Xi - Mx)(Yi - My) / N\,[SDx\ SDy]$$

ここで、Mx は X の平均、My は Y の平均、SDx は X の標準偏差、SDy は Y の標準偏差です。最初に、この分子だけを取り上げましょう。

CC の分子= $\Sigma i\,(Xi - Mx)(Yi - My)$

$= \Sigma i\,(Xi\,Yi - Xi\,My - Mx\,Yi + Mx\,My)$　←展開

$= \Sigma i\,Xi\,Yi - \Sigma i\,Xi\,My - \Sigma i\,Mx\,Yi + \Sigma i\,Mx\,My$　←Σ を分配

$= \Sigma i\,Xi\,Yi - My\ \Sigma i\,Xi - Mx\ \Sigma i\,Yi + N\,Mx\,My$

←非 i 項を外へ

ここで、Xi Yi のうち、b(1, 0), c(0, 1), d(0, 0)にあたる部分では X と Y の

少なくとも 1 つがゼロなので、その積もゼロになります。よって

[2]　　$\Sigma i\ Xi\ Yi = a$　　←積 Xi Yi が 1 のケースの合計

となります。また

[3]　　$\Sigma i\ Xi = a + b$　←X の和←上表(X:Y)
[4]　　$\Sigma i\ Yi = a + c$　←Y の和←上表(X:Y)
[5]　　$Mx = \Sigma i\ Xi / N = (a + b) / N$　←X の平均←[3]
[6]　　$My = \Sigma i\ Yi / N = (a + c) / N$　←Y の平均←[4]

となるので、分子は

CC の分子　$= \Sigma i\ Xi\ Yi - My\ \Sigma i\ Xi - Mx\ \Sigma i\ Yi + N\ Mx\ My$
$= a - (a+c)(a+b)/N - (a+b)(a+c)/N + N\ (a+b)/N\ (a+c)/N$　[2-6]
$= a - (a+c)(a+b)/N - (a+b)(a+c)/N + (a+b)(a+c)/N$
$= a - (a + b)(a + c) / N$
$= [Na - (a + b)(a + c)] / N$
$= [(a + b + c + d)a - (aa + ac + ba + bc)] / N$　←[1]
$= (aa + ab + ac + ad - aa - ac - ab - bc) / N$
[7]　　$= (ad - bc) / N$

次に CC の分母の 1 つ SDx を見ます。

SDx　$= \{[\Sigma i\ (Xi - Mx)^2]^{1/2} / N\}^{1/2}$　←X の標準偏差
$= \{[\Sigma i\ (Xi^2 - 2\ Xi\ Mx + Mx^2)]^{1/2} / N\}^{1/2}$　←展開
$= \{[\Sigma i\ Xi^2 - \Sigma i\ 2\ Xi\ Mx + \Sigma i\ Mx^2)] / N\}^{1/2}$　←Σ を分配
$= \{[\Sigma i\ Xi^2 - 2\ Mx\ \Sigma i\ Xi + N\ Mx^2)] / N\}^{1/2}$　←非 i 項を外へ

Xi はすべて 1 または 0 なので Xi^2 の和は

[8]　$\Sigma i\ Xi^2 = a + b$　←X^2 の和←上表(X:Y)

SDx　$= \{[(a + b) - 2\ (a + b)^2 / N + (a + b)^2 / N] / N\}^{1/2}$　←[8], [3], [5]
$= \{[a + b - (a + b)^2 / N] / N\}^{1/2}$　←$(a + b)^2 / N$ が共通
$= \{[(a + b)N - (a + b)^2] / N^2\}^{1/2}$　←N を分母へ
$= \{[(a + b)(a + b + c + d) - (a + b)^2] / N^2\}^{1/2}$　←[1]
$= \{(a + b)[(a + b + c + d) - (a + b)] / N^2\}^{1/2}$　←(a + b) が共通
$= [(a + b)(c + d) / N^2]^{1/2}$　←(a + b) が共通
[9]　$= [(a + b)(c + d)]^{1/2} / N$　←N を外へ

同様にして、CC の分母の 1 つ SDy は

[10] SDy　$= (a + c)(b + d)^{1/2} / N$　←$\Sigma i\ Yi^2 = a + c$ に注意

よって

CC の分母　= N [SDx SDy]
　　= N {[(a + b)(c + d)]$^{1/2}$ / N}*{[(a + b)(c + d)]$^{1/2}$ / N}　←[9, 10]
　　= [(a + b)(c + d)]$^{1/2}$}*{[(a + b)(c + d)]$^{1/2}$ / N　←N を整理
[11]　= [(a + b)(c + d)(a + b)(c + d)]$^{1/2}$ / N　←乗数 1/2 を整理

よって、相関係数(CC)は

CC = Σ i (Xi – Mx)(Yi – My) / N [SDx SDy]
　　= [(ad - bc) / N] / {[(a + b)(c + d)(a + c)(b + d)]$^{1/2}$ / N}　←[7, 11]
　　= (ad - bc) / [(a + b)(c + d)(a + c)(b + d)]$^{1/2}$　←/N が共通
　　= Phi　←定義

●Phi 係数と Ochiai 係数

Phi 係数を実際に適用してみると不都合なときがあります。次のデータ A, B を比べてみましょう。

A	Y (+)	Y (-)	和
X (+)	100	10	110
X (-)	20	2	22
和	120	12	132

B	Y (+)	Y (-)	和
X (+)	4	10	14
X (-)	20	50	70
和	24	60	84

どちらも Phi 係数の分子の ad－bc がゼロとなるので（100*2 － 10*20 = 0; 4*50 － 10*20 = 0）、Phi 係数はゼロになります。しかし、データ A とデータ B を比べれば A のほうがずっと連関度が高いように思えます。プラス(+)を共有するケースが 100 もあるからです。これは全体 132 の 75.8%にあたります。それに対して B はどうでしょうか。わずか 4 回の共起回数で計算すると 4.8%になります。

この原因は d(-/-)の数値の扱い方にあります。X にも Y にもない要素は与えられたデータに限れば有限ですが、X、Y 以外のデータに存在して、X にも Y にもなかったものです。そうした d の値は、X と Y の内容にかかわらず、一般にいくらでも増やすことができます。つまり、理論的には d の数は無限(∞)であると考えられます。たとえば、X と Y という二人が読んだことがある本を数えるとき、どちらも読んだことのない本の数は無限と考えられます（本が無限に出版されるとして）。

そこで、先の Phi の式で d が無限になると仮定してみましょう。phi 係数で d が無限大になるものを Phi'とします。

Phi　= (ad - bc) / [(a + b)(c + d)(a + b)(c + d)]$^{1/2}$

Phi'　= lim(d→∞) (ad - bc) / [(a + b)(c + d)(a + c)(b + d)]$^{1/2}$

$= \lim(d \to \infty)\ [(ad - bc)/d] / \{[(a + b)(c + d)(a + c)(b + d)]^{1/2} /d\}$

←分子と分母を d で割る

$= \lim(d \to \infty)\ (a - bc/d) / [(a + b)(c + d)(a + c)(b + d) / d^2]^{1/2}$

←d を移動

$= \lim(d \to \infty)\ (a - bc/d) / [(a + b)(c/d + 1)(a + c)(b/d + 1)]^{1/2}$

←/d を分配

$= a / [(a + b)(a + c)]^{1/2}$ ←分母 d を無限大に

これが Phi 係数の修正版（Ochiai 係数）です。とてもシンプルになりました。先のデータ A, B で計算してみましょう。

$Phi'(A) = 100 / [(199+10)(100+20)]^{1/2} = .870$

$Phi'(B) = 4 / [(4+10)(4+20)]^{1/2} = .218$

このように、Phi 係数で区別できなかった両者も Ochiai 係数(Phi')を利用すればデータ(A)の方がデータ(B)よりも連関性が高いという直感を裏づけることができます。

●両者に存在しない特徴

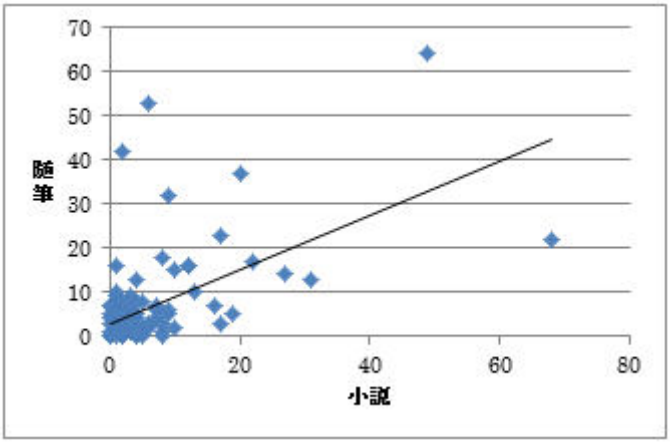
かつて印欧言語学の分野では Phi 係数を使った Kroeber (1937, 1969)と Ochiai 係数を使った Ellegard (1959)の間に論争がありました。これを安本(1995)が簡単に解説しています。この問題については、一般に連関係数のどちらかが正しいということではなくて、データの種類や性格によって係数の選択を考えるべきでしょう。たとえば、アンケート調査などで「賛成」と「反対」という回答があるとすれば、単に両者が一致して「賛成」と答えた場合の数(a)だけでなく、一致して「反対」と答えた場合の数(d)も同時に考慮されるべきです。

2 つのデータだけでなく、多数のデータ間の連関度を見る場合には、問題の両者に存在しない特徴であっても他のデータに存在する特徴であるならば、どちらもその特徴を持たないという否定的な一致はそれなりの意味をもつと考えられます。

一方、a « d、つまり先述の「Phi 係数と Ochiai 係数」で見たような d(-, -)が a(+, +)を大きく上回るデータを扱うときは、d のない係数を選択するとよいでしょう。

●量的データと質的データ

先に見たように、単語の頻度数は非常に偏った分布を示すので相関係数による分析には適しません。次の散布図には一応「線形近似曲線」が描かれていますが、データは左下に固まっていて、右上になるとほとんどデータがありません。頻度の高い単語の数は少なく、一方あまり使われない単語の数は非常に多いのです。



ここで、単語の頻度を単語使用の「有無」に変えて分析する方法を採ります。そうすれば、すべてのデータの分布は「有」と「無」の 2 種類の値になります。頻度数などのような連続的なデータは「量的なデータ」(quantitative data)とよばれ、このように単に有・無を示すようなデータは「質的なデータ」(qualitative data)とよばれます。

言語研究では、たった一度だけ出現するデータ(hapax)を特別に扱うことがあります。偶然に現れたケースかもしれないからです。2 度の偶然はほとんどあり得ないので、2 を基準値として、それ以上を「有」(1)のデータとして基準化する場合があります。データが巨大になったときは、さらにこの基準を上げることも考えられます。いずれにしても結果はこの基準値に左右されますから、それをしっかりと認識しておくことが必要です。

●拡大連関係数

相関係数は、たとえば勉強時間（x 分:範囲[0, 600]）と試験の得点（y 点：範囲[0, 100]）などのように単位や範囲が異なる変数間の関係を調べるときに使用できますが、連関係数は存在するか(1 / "+")、存在しないか(0 / "-")、という質的なデータの変数だけを扱います。

たとえば{A: 2, 3, 5, 7, 9}と{B: 22, 23, 25, 27, 29}などのように、定数(20)の差があるデータどうしは完全に直線になるので相関係数は最大値の 1.000 になります。データ{A: 2, 3, 5, 7, 9}を定数倍(x 10)したデータ {C: 20, 30, 50, 70, 90}の間でも同様に相関係数は最大値の 1.000 になります。どちらも 2 データが完全に相関すると考えれば当然でしょう。しかし、一方で、{A: 2, 3, 5, 7, 9}と{B: 22, 23, 25, 27, 29}（または{C: 20, 30, 50, 70, 90}）よりも、{A: 2, 3, 5, 7, 9}と{D: 3, 2, 4, 4, 8}のほうが「近い」関係にある、とも考えられます。次は、現代スペイン語の 5 つのテキスト(T1~5)に現れた 2 つの語形(X, Y)の千語率を示します。

D	X	Y	a=min(X,Y)	b = X - min	c = Y - min	J
T-1	44	43	43	1	0	0.936

T-2	41	48	41	0	7	
T-3	40	41	40	0	1	
T-4	41	36	36	5	0	
T-5	44	44	44	0	0	
和	210	212	204	6	8	

連関係数の扱う数値を拡大して、0/1 (+/-) に限らず一般の非負数として、次の a, b, c を計算します。

$$a = \Sigma (i) \min[x(i), y(i)]$$
$$b = \Sigma (i) \{x(i) - \min[x(i), y(i)]\}$$
$$c = \Sigma (i) \{y(i) - \min[x(i), y(i)]\}$$

上の式の min[x(i), y(i)]は x(i), y(i)の小さい方の値（2 数の最小値）を示します（表中の下線部）。その最小値を足し上げた和 Σ (i)を a(+/+)とします。b(+/-)は x にだけ存在する値なので、x - min(x, y)とします。x と min (x, y)が同じならば b(+/-) = 0 になります。同様にして c(-/+)の値は y にだけ存在する値 y - min(x, y)です。上表の例では

$$a = 43 + 41 + 40 + 36 + 44 = 204$$
$$b = (44-43)+ (41-41) + (40-40) + (41-36) +(44-44) = 6$$
$$c = (43-43)+ (48-41) + (41-40) + (36-36) +(44-44) = 8$$

たとえば T-1 では X=44, Y=43 となっていますが、これは T-1 で X が 44 回、Y が 43 回出現したことを意味しています。よって、T-1 というデータ内で X と Y が共起した回数 a (+/+)は min(44, 43) = 43 回になります。それに加えて、X は Y と共起しなかった回数 b (+/-)が 1 回ある、と考えます。

このデータでは、X, Y の共通性を示す a(+/+)の値(204)が、X, Y の差異性を示す b(+/-), c(-/+)の値(6, 8)と比べてかなり大きいため、たとえば Jaccard 係数(J)を計算すると

$$J = a / (a + b + c) = 204 / (204 + 6 + 8) = .936$$

のように高い数値を示します。

このように対象を 0/1 データから一般の非負データ（小数を含む）に拡大して求めた a, b, c を使って計算した連関係数を「拡大連関係数」(Expanded Association Coefficient: EAC)とよびます[6]。

相関係数は 2 つの変数の動きの傾向を見るのに対して、拡大連関係数は 2 つの変数が共通する度合いを測っています[7]。プログラムで連関係数の入

[6] 拡大連関係数の計算では x(i)と y(i)のどちらにもないケースの数 d(-/-)を求めることができません。よって d を使わない連関係数だけを適用します。
[7] この点で後述する距離係数に似ています。

力データが 0/1 型でないときに拡大連関係数を計算するようにします。

●順序連関行列

データ行列の数値そのものの相関ではなく、大小関係の順序の連関から相互の関係を見るために、グッドマンとクラスカルの順序連関係数を使います（→後述「分析」）。たとえば、次のデータの v1 と v2 の順序連関係数(GK)を計算しましょう。

X_{np}	v1	v2
d1	10	19
d2	11	7
d3	0	0
d4	0	1

はじめに、その肯定値(Positive: P)と否定値(Negative: N)を次のように計算します。

P(v1, v2) = 10 * (7+1) + 11 * 1 = 91
N(v1, v2) = 11 * 19 = 209

よって

GK (v1, v2) = (91 – 209) / (91 + 209) = -.393

とくに順位得点（→「得点」）の連関を見るときに順序連関係数が役立ちます。

■外国語学習・獲得と「価値」の優先度

語彙学習、さらに外国語学習一般において、学習者が認識する「価値」の優先度が高い、という仮説を立てます。語彙についていうと、単語の意味に学習者が「価値」を見出すと、それが優先的に獲得される、という仮説です。ここでいう「価値」は、いわゆる「重要単語」のことではありません。なぜなら、重要単語で示されている「重要性」は学習者の認める価値とは異なる場合があるからです。

この仮説を検証するために次のような実験をしてみました。一定の量のスペイン語の単語リストについて、はじめに「自分にとって価値の優先度の高い」単語にマークし、その後単語リスト全体の記憶練習をして、その結果をそれぞれの単語数について集計しました。この実験に「スペイン語学習・教育法」の履修者 12 人が参加し、毎回語数と出席人数が異なる実験を数回行いました。

個人	a (+/+)	b (+/-)	c (-/+)	d (- /-)	Yule	Hamann
1	4	1	0	1	1.000	0.667
2	7	3	5	5	0.400	0.200
3	6	2	3	4	0.600	0.333
4	23	13	7	17	0.622	0.333
5	18	13	12	17	0.325	0.167
6	8	3	2	7	0.806	0.500
7	7	3	3	7	0.690	0.400
8	15	15	0	11	1.000	0.268
9	17	13	1	5	0.735	0.222
10	10	3	4	9	0.765	0.462
11	11	5	4	10	0.692	0.400
12	14	1	6	9	0.909	0.533

(a) +/+: 「比較的価値が高い単語(+)」/「学習成功(+)」
(b) +/-: 「比較的価値が高い単語(+)」/「学習失敗(-)」
(c) -/+: 「比較的価値が低い単語(-)」/「学習成功(+)」
(d) -/-: 「比較的価値が低い単語(-)」/「学習失敗(-)」

参加した 12 人の結果は Yule も Hamann もプラスになっていますから、先の仮説に沿うものでした。

敷衍して考えてみると、はたして私たちは外国語をひたすら反復練習して獲得するのでしょうか？もしかしたら「価値」の優先度が強く働いた学習項目は瞬間的に獲得されているのかもしれません。とくにがんばって記憶練習した覚えもないのに獲得してしまった語があるとすれば、それは学習者にとって「価値」のある単語だった可能性が高いと思われます。そうだとすると、外国語（やその他の科目）を、がんばって学習するよりも、価値を見出して獲得してしまうほうが効果的ではないでしょうか。

価値を見出すためには、「形式→意味」という流れの教育・学習よりも、「意味→形式」という流れのほうが効果があると思います。私たちは（外国語の）形式を見て価値を見出すことはあまりありませんが[8]、意味については、その価値の有無・程度を瞬間的に判断することができるからです。

6.2.2. 連関行列

各種の連関係数を使って連関係数行列を作るために、1 または 0 からな

[8] この例外もあります。あるとき社会人向けのスペイン語コースを担当したとき、受講者から「パハロ」（pájaro：「小鳥」）という言葉の響きが好きで、すぐに覚えてしまった」という感想をいただいたことがあります。そのとき聞き忘れたのですが、この人は「パハロ」の響きだけでなく「小鳥」も好きな人だったのかもしれません。

るデータ行列(Q_{np})の各変数（列）について、2 つずつの変数(X_i, X_j)の組み合わせで、X_i =1, X_j=1 のケース数を示す A(i, j)、X_i =1, X_j=0 のケース数を示す B(i, j)、X_i =0, X_j=1 のケース数を示す C(i, j)、X_i =0, X_j=0 のケース数を示す A(i, j)の行列を作ります。そのためにはじめに次の W_{np} を用意します。

$$W_{np} = 1 - Q_{np}$$

この W_{np} は、データ行列 Q_{np} のすべての成分について、0 と 1 が交換された行列です。

Q_{np}	v1	v2	v3	v4
d1	1	1	0	0
d2	0	0	1	0
d3	0	1	0	0
d4	0	0	1	1
d5	1	1	1	0

W_{np}	v1	v2	v3	v4
d1	0	0	1	1
d2	1	1	0	1
d3	1	0	1	1
d4	1	1	0	0
d5	0	0	0	1

この 2 つの行列を使って(A_{pp}, B_{pp}, C_{pp}, D_{pp})を算出します。

$$A_{pp} = Q_{np}^{T} Q_{np}$$
$$B_{pp} = Q_{np}^{T} W_{np}$$
$$C_{pp} = W_{np}^{T} Q_{np}$$
$$D_{pp} = W_{np}^{T} W_{np}$$

A_{pp} は共起回数を示します。「行列」の転置と積の機能を使ってその成分を確認しましょう。

$$A_{pp} = Q_{np}^{T} Q_{np}$$

$\mathbf{Q}^T$	d1	d2	d3	d4	d5
v1	1	0	0	0	1
v2	1	0	1	0	1
v3	0	1	0	1	1
v4	0	0	0	1	0

X

Q	v1	v2	v3	v4
d1	1	1	0	0
d2	0	0	1	0
d3	0	1	0	0
d4	0	0	1	1
d5	1	1	1	0

=

$\mathbf{Q^TQ}$	v1	v2	v3	v4
v1	2	2	1	0
v2	2	3	1	0
v3	1	1	3	1
v4	0	0	1	1

他の対称行列の成分も確かめましょう。

$B_{pp} = Q_{np}^{T} W_{np}$

Q^T	d1	d2	d3	d4	d5
v1	1	0	0	0	1
v2	1	0	1	0	1
v3	0	1	0	1	1
v4	0	0	0	1	0

X

W	v1	v2	v3	v4
d1	0	0	1	1
d2	1	1	0	1
d3	1	0	1	1
d4	1	1	0	0
d5	0	0	0	1

=

Q^TW	v1	v2	v3	v4
v1	0	0	1	2
v2	1	0	2	3
v3	2	2	0	2
v4	1	1	0	0

$C_{pp} = W_{np}^{T} Q_{np}$

W^T	d1	d2	d3	d4	d5
v1	0	1	1	1	0
v2	0	1	0	1	0
v3	1	0	1	0	0
v4	1	1	1	0	1

X

Q	v1	v2	v3	v4
d1	1	1	0	0
d2	0	0	1	0
d3	0	1	0	0
d4	0	0	1	1
d5	1	1	1	0

=

W^TQ	v1	v2	v3	v4
v1	0	1	2	1
v2	0	0	2	1
v3	1	2	0	0
v4	2	3	2	0

$D_{np} = W_{np}^{T} W_{np}$

W^T	d1	d2	d3	d4	d5
v1	0	1	1	1	0
v2	0	1	0	1	0
v3	1	0	1	0	0
v4	1	1	1	0	1

X

W	v1	v2	v3	v4
d1	0	0	1	1
d2	1	1	0	1
d3	1	0	1	1
d4	1	1	0	0
d5	0	0	0	1

=

W^TW	v1	v2	v3	v4
v1	3	2	1	2
v2	2	2	0	1
v3	1	0	2	2
v4	2	1	2	4

この 4 つの行列から次の式で各種の係数行列を求めます。以下では $_{np}$ を省いて、たとえば A_{np} を A とします。

単純一致 = (A + D) / (A + B + C + D)

J = A / (A + B + C)

H= [(A + D) - (B + C)] / [(A + D) + (B + C)]

Y = (A*D – B*C) / (A*D + B*C)

$P = (A*D – B*C) / [(A + B)(C + D)(A + C)(B + D)]^{1/2}$

$O = A / [(A + B)(A + C)]^{1/2}$

U. = (2A – B – C) / (2A + B + C)

＊連関係数については Anderberg (1973:93-126), Romesburg (1989: 177-209)を参照しました。連関係数行列の A, B, C, D 行列の算出法は河口 (1978: II, 30-31)を参照しました。

●占有度

次のようなサンプル（下左表）を使って「占有度」(Degree of Possession)と名づけるオプションを説明します。積和共起回数を計算すると下右表の対称行列 A_{pp} になります。

Q_{np}	v1	v2	v3	v4
d1	1	1	0	0
d2	0	0	1	0
d3	0	1	0	0
d4	0	0	1	1
d5	1	1	1	0

A_{pp}	v1	v2	v3	v4
v1	2	2	1	0
v2	2	3	1	0
v3	1	1	3	1
v4	0	0	1	1

v1, v2 の間の共起回数は 2 と計算されますが、ここで注目したいのは d1 における v1, v2 の間の共起の様子と、d5 におけるその様子との違いです。上左表を見ると d1 は唯一 v1, v2 だけを共有していますが、d5 では他に v3 でも共有されています。ここで d1 のようなケースのほうが d5 のようなケースよりも重い価値があると解釈し、それを数量的に表現したいと思います。

次は、先の A, B, C, D のそれぞれの対称行列を作成するために用意した Q_{np} と W_{np} です($W_{np} = 1 - Q_{np}$)。

Q_{np}	v1	v2	v3	v4
d1	1	1	0	0
d2	0	0	1	0
d3	0	1	0	0
d4	0	0	1	1
d5	1	1	1	0

W_{np}	v1	v2	v3	v4
d1	0	0	1	1
d2	1	1	0	1
d3	1	0	1	1
d4	1	1	0	0
d5	0	0	0	1

これをそれぞれ次のように変換します。

Q_{np}*	v1	v2	v3	v4
d1	0.500	0.500	0.000	0.000
d2	0.000	0.000	1.000	0.000
d3	0.000	1.000	0.000	0.000
d4	0.000	0.000	0.500	0.500
d5	0.333	0.333	0.333	0.000

W_{np}*	v1	v2	v3	v4
d1	0.000	0.000	0.500	0.500
d2	0.333	0.333	0.000	0.333
d3	0.333	0.000	0.333	0.333
d4	0.500	0.500	0.000	0.000
d5	0.000	0.000	0.000	1.000

たとえば、d1 行には 1 が 2 個出現しているので、それぞれ 0.5 の価値がある、という考え方です。d5 では 1 が 3 個なので、すべて 0.333 という数値を与えます。W_{np} についても同様です。このようにして用意した Q_{np}*と W_{np}*を使って、それぞれの占有度指数を加味した連関係数を算出します。

最後にこの占有度指数を使わない四分点相関係数（下左表 SM）と、使った場合（下右表 SMp）を比較します（単純一致係数 SM）。以下に見るように大小関係の傾向は似ていますが、かなり大きな数値の差が観察されます。

SM.	v1	v2	v3	v4
v1	1.000	0.800	0.400	0.400
v2	0.800	1.000	0.200	0.200
v3	0.400	0.200	1.000	0.600
v4	0.400	0.200	0.600	1.000

SMp	v1	v2	v3	v4
v1	1.000	0.684	0.211	0.211
v2	0.684	1.000	0.087	0.087
v3	0.211	0.087	1.000	0.478
v4	0.211	0.087	0.478	1.000

■スペイン語の普通語 tú と丁敬語 usted

下左表はスペイン語で¿Adónde vas?「どこに行くの」という表現を、Niño(子供), Joven (若者), Mayuor (大人), Anciano (老人)に対してさまざまな条件（親疎・上下関係）で使うときの、それぞれ No (使わない), A veces (ときどき), Siempre (いつも)の使用頻度を調べて集計したものです。

¿Vas?	Niño	Joven	Mayor	Anciano
No	1	5	192	92
A veces	3	22	58	20
Siempre	56	153	110	8

G. & K.	¿Vas?
Positive v.	10600
Negative v.	101941
G. & K.	- .812

このように、横の分類も縦の分類も一定の順序を持った変数であるとき、全体の分布が横と縦の順序にどの程度従っているかを示す係数 GK を算出するために、クロス表で、1 つのマス目にあるデータとそれよりも行と列の位置が大きくなる右下の位置のデータの対の総数である「肯定対総数」(Positive pairs: Ps)を、次のようにして計算します。

Ps(Vas) = 1 * (22+58+20+153+110+8) + 5 * (58+20+110+8) + 192 * (20+8) + 3 * (153+110+8) + 22 * (110+8) + 58 * 8 = 10600

一方、1 つのマス目にあるデータとそれよりも行と列の位置が小さくなる左下の位置データの対の総数である「否定対総数」(Negative pairs: N)を、次のようにして計算します。

Ng(Vas) = 5 * (3+56) + 192 * (3+22+56+153) + 92 * (3+22+58+56+153+110) + 22 * 56 + 58 * (56+153) + 20 * (56+153+110) = 101941

順序連関係数 GK は Ps と Ng の両側相対値です。

GK(Vas) = (P - N) / (P + N) = (10600 – 101941) / (10600 + 101941) = -.812

この数値は大きなマイナス値を示しているので、相手の年齢と普通語の

使用が逆相関の関係になります。

次は同じことを敬語を使った¿Adónde va usted?「どちらに行かれますか？」という表現の使用頻度の集計です。

¿Va usted?	Niño	Joven	Mayor	Anciano
No	55	147	142	18
A veces	5	24	99	33
Siempre	0	9	119	69

G. & K.	¿Va usted?
Positive v.	93267
Negative v.	15854
G. & K.	.709

P(Va usted) = 93267
N(Va usted) = 15854
GK(Va usted) = (93267 - 15854) / (93267 + 15854) = .709

このように、GK を使うことによって、スペイン語の普通体は対話者の年代層と逆連関し、丁寧体はそれと正連関していることがわかります。対話者の年代層だけでなく、各種の変数を比較すると、スペイン語の普通語・丁寧語の選択は上下関係よりも親疎の関係のほうが強く働いていることがわかります。比較した日本語ではその逆の傾向が見つかりました。

＊池田(1976:130-132)を参照しました。

●文字連関行列

行列の成分が数値ではなく文字のデータ行列を扱います。A, B, C...は任意の文字(A, B, ...など)、または文字列(bueno, malo, regular, ...など)とします。このような文字行列の変数の連関行列を「文字連関行列」(Nominal Association Matrix: NAM)とよぶことにします。たとえば、v1-v4 を地方名、d1-d5 はそれぞれの地方で発行された文書、A, B, C, ... を言語特徴、というような資料を想定します。

L_{np}	v1	v2	v3	v4
d1	A	A	B	C
d2	A	A	C	C
d3	A	C	B	C
d4	C	C	C	A
d5	B	B	C	C

Npp	v1	v2	v3	v4
v1	1.000	.600	-.600	-1.000
v2	.600	1.000	-.600	-.600
v3	-.600	-.600	1.000	-.200
v4	-1.000	-.600	-.200	1.000

たとえば、v1 と v2 の相関(0.600)は次のように計算します。両列に同じ文字が使われている回数(a:++)は 4, ある文字が v1 にあって v2 にない場合の数(b: +-)は 1, 逆にそれが v1 になくて v2 にある場合の数も 1 になるので、先の優先係数の式[2a – (b+c)] / [2a – (b+c)]を適用して、[4x2 – (1+1)] / [4x2 + (1+1)] = .600 となります。

この文字連関行列は次のような、1 つの成分の中に、複数の文字がある

場合にも計算できます。

Lt.Oc.	v1	v2	v3	v4
d1	A	A,B	B	C
d2	B,D	B,C,D	B,C	D
d3	A,B	B	B	C
d4	C	C	A	A
d5	B,C	C	B,C	B,C,D

$L_{np.}$	v1	v2	v3	v4
v1	1.000	.500	.067	- .200
v2	.500	1.000	.333	- .467
v3	.067	.333	1.000	- .143
v4	- .200	- .467	- .143	1.000

たとえば、v1 と v2 の文字連関係数(0.520)は次のように計算します。d1 では、v1 の A と v2 の A,B を比べて、両者にある文字数 1 を a(++)とします。v1 にあって v2 にない文字数 0 を b(+-)とします。v2 にあって v1 にない文字数 1 を c(-+)b とします。この a, b, c を他の行 d2, …, d5 でも加算して計算した優先係数の値が文字連関係数(v1, v2) = 0.520 になります。すべての組み合わせ(v1, …, v4)の文字連関係数を計算すると文字連関行列ができます。

6.3. 共起と選択

前のセクションで扱った連関係数では、データの全数(N)を a (+:+), b (+:-), c (-:+), d (-:-)のケースに分けて計算しましたが、N そのものは考慮されませんでした。このセクションでは N を考慮して定義された指標を扱います。ここで a, b, c, d の頻度のほかに、2 つの言語形式の出現回数 X, Y と、共起回数 C: (X:+/ Y:+)と全数 N を使います。

ここで次の関係を確認しておきましょう。

C = a; X = a + b; Y = a + c; N = a + b + c + d

それぞれの数値の関係は次の表で示されます。

X:Y	Y+	Y:−	sum
X:+	C: a	b	X
X:-	c	d	~X
sum	Y	~Y	N

逆に C, X, Y の頻度から各種の連関係数に使われる a, b, c, d の値が導かれます。ここで扱われる数値はすべて非負になります。

a = C; b = X - C; c = Y - C; d = N - a - b - c

6.3.1. 相互情報量

共起係数として用いられる「相互情報量」(Mutual Information: MI)は、共起回数の平均(C/N)と、X と Y の同時確率(X/N)*(Y/N)の比の対数（底=2）と定義されます（石川 2008: 111）。

$$\begin{aligned} MI &= \log_2 \{[(C / N)] / [(X\ Y / N^2)]\} \\ &= \log_2 [(C\ N) / (X\ Y)] \\ &= \log_2 \{[a\ (a + b + c + d)] / [(a+b)(a+c)]\} \end{aligned}$$

上の第 2 式の中の(C N) / (X Y)は C / (XY/N)と書き換えると、共起回数(C)とその期待値との比を示していることがわかります。

上の第 3 式の中の分子[a (a + b + c + d)]と(a+b)(a+c)]が一致したとき、よって

$$\begin{aligned} &a\ (a + b + c + d) - (a+b)(a+c) = 0 \\ &(a^2 + ab + ac + ad) - (a^2 + ac + ab + bc) = 0 \\ &ad - bc = 0 \end{aligned}$$

このとき $MI = \log_2 1 = 0$ になります。

$$\begin{aligned} &ad - bc = 0 \ \rightarrow\ \text{分子=分母} \ \rightarrow\ MI = \log_2 1 = 0 \\ &ad - bc > 0 \ \rightarrow\ \text{分子>分母} \ \rightarrow\ MI > \log_2 1 = 0 \\ &ad - bc < 0 \ \rightarrow\ \text{分子<分母} \ \rightarrow\ MI < \log_2 1 = 0 \end{aligned}$$

上の第 2 式から、bc = 0 のとき分子と分母の差が最大になるので MI が最大になることがわかります。よって

b = 0 or c = 0　または b = c = 0

のときに MI が最大になります。はじめに b = 0 にすると[9]、MI は

$$\begin{aligned} MI &= \log_2 \{[a\ (a + c + d)] / [a(a + c)]\} \qquad \leftarrow\ b = 0 \\ &= \log_2 [(a + c + d) / (a + c)] \end{aligned}$$

このときの対数内の(a + c + d) / (a + c)は明らかに 1 以上です。この式内の c が増加すると次第に分母と分子の値は近くなって分数は 1 に近づき、対数は 0 に近づきます。逆に c が減少すると分子と分母の値の差が大きくなって分数の値は増加し、c = 0 になったときに MI は最大値 $\log_2 [(a + d) / a]$ に到達します。

よって MI は b = c = 0 のときに最大(MI.max.)になります。

[9] 以下の考察ははじめに c = 0 にしても同様です。

MI.max. $= \log_2 [a(a + d) / a^2]$ ← b = c = 0

$= \log_2 [(a + d) / a]$

「規定相互情報量」(Regular Mutual Information: R.MI)は[10]

R.MI = MI / MI.max

$= \{\log_2 [a(a + b + c + d)] / [(a+b)(a+c)]\} / \log_2 [(a + d) / a]$

$= <\ln\{[a(a + b + c + d)] / [(a+b)(a+c)]\} / \ln(2)\}> / [\ln[(a + d) / a] / \ln(2)]$

$= \ln\{[a(a + b + c + d)] / [(a+b)(a+c)]\} / \ln[(a + d) / a]$

●確率から見た相互情報量

言語研究では 2 つの語の結合度を調べるために相互情報量が使われています。これは、共起回数(C)をデータ全体で理論的に期待できる共起得点（期待値=X*Y/N）で割った値の対数（底=2）です。

MI $= \log_2 [C / (X*Y/N)]$

$= \log_2 \{(C/N) / [(X/N)*(Y/N]\}$

$= \log_2 [(C*N) / (X*Y)]$

上の第 2 式中の(C/N) / [(X/N)*(Y/N]は、X, Y の同時確率 P(X,Y) = C/N と X, Y の確率の積 P(X) P(Y)を比で比較しています。

上の第 3 式中の(C*N) / (X*Y)を確率の観点から見直すと

P(Y|X) = C / X ←X の中で Y と共起する条件確率

P(Y) = Y / N ←Y の確率

P(Y|X) / P(Y) = (C / X) / (Y / N) = (C*N) / (X*Y)

ここで C は X と Y の共起回数 ; N は総数を示します。よって、上の式から、対数 $\log_2$ の中の式は、X の中で Y と共起する条件付き確率 P(Y|X)が、本来 Y が起こる確率 P(Y)と比較した比になっていることがわかります。

たとえば、あるスペイン語の資料で調べると、*muy* (='very')という語の頻度が 120, *bien* (='well')の頻度が 167, 全語数が 26578 でした。そうすると、*muy* と *bien* の共起得点が理論的に期待できる値は(120 / 26578) * (167 / 26578)となります。これは、それぞれが出現する確率の積です。そして、実際の資料では *muy* + *bien* が 47 出現しました。これは 47 / 26578 という確率です。そこで相互情報量を計算するために、はじめに共起得点をデータ全体で理論的に期待できる共起得点（期待値）で割った値を求めます。

(47 / 26578) / [(120 / 26578) * (167 / 26578)]

[10] 相互情報量の式から ad < bc のときに負になることがわかりますが、一般に c(-/-)の値は巨大になることから現実的ではありません。

= (47 * 26578) / (120*167) = 62.334

これの対数（底=2）は Log_2 62.334 = 5.962 です。これが相互情報量です。底を 2 とする対数は一般に情報量を示します。たとえば、16 の可能性がある事象の情報量は、16 = 2^4 なので、4 (=$log_2$16)になります。

6.3.2. 単純選択率

これまでに取り上げた各種の連関係数と相互情報量では、どれも 2 つの要素(X, Y)の共起 Co について、X が Y と共起する度合い Co(X, Y)と、Y が X と共起する度合い Co(Y, X)は当然同じ値になります。これは「共起」(cooccurrence)という概念に沿います。

一方、「共起」ではなく「選択」(selection)という視点から見ると、X が Y を選択する度合 Sel(X, Y)と、Y が X を選択する度合 Sel(Y, X)は異なるほうが普通です。このセクションでは次の図式を使って選択の度合いを測る方法を考えます。

X:Y	Y+	Y:−	sum
X:+	C: a (+:+)	b (+:-)	X: a + b
X:-	c (-:+)	d (-:-)	c + d
sum	Y: a + c	b + d	N

はじめに単純に次のように計算する「単純選択率」(Simple Selection Ratio:SSR)を考えます。単純選択率の範囲は明らかに[0, 1]になります。

SSR(X,Y) = P(Y|X) = C / X = a / (a + b)　　← X が Y を選択した率
SSR(Y,X) = P(X|Y) = C / Y = a / (a + c)　　← Y が X を選択した率

6.3.3. 比較選択率

単純な X→Y の選択率 SSR(X,Y)の計算(C/X)では、全体(N)の中での Y の出現率(Y/N)が考慮されていません。X→Y の選択率 SSR(X,Y)が本来の Y の出現率とほとんど同じならば、選択率 SSR(X,Y)はあまり意味がないと考えます。つまり、X が Y の出現に影響しているとは考えられないからです。まして、X→Y の選択率が本来の Y の出現率よりも小さい場合は、逆向きの作用（X が Y の出現を妨げている）を考えなければなりません。

そこで、C/X という条件付き確率 P(Y|X)を、Y/N という Y の確率 P(Y)と比較します。P(Y|X)が P(Y)に比べて大きければ、X が Y を選択する率が、全体の中で Y が選択される率より高い、と考えられるからです。はじめに両者の比(確率比 Probability Ratio: PR)を取ってみましょう。

PR = P(Y|X) / P(Y) = (C/X) / (Y/N) = (C*N) / (X*Y)

このように両者を比で比べると、X と Y を交換しても比は同じ値になってしまいます。

そこで次に確率差(Probability Difference: PD)によって比べます。

$$PD = P(Y|X) - P(Y) = (C/X) - (Y/N) = (C*N - X*Y) / (N*X)$$

この確率差(PD)の式で X と Y を交換すると、分子 C*N - X*Y は同じですが、分母 N*X が異なることがわかります。確率差(PD)を a, b, c, d で表すと

X:Y	Y+	Y:−	sum
X:+	C: a (+:+)	b (+:-)	X: a + b
X:-	c (-:+)	d (-:-)	c + d
sum	Y: a + c	b + d	N

$$\begin{aligned} PD &= (C*N - X*Y) / (N*X) \\ &= [a\,(a + b + c + d) - (a + b)(a + c)] / [N\,(a + b)] \\ &= (a^2 + ab + ac + ad) - (a^2 + ac + ab + bc) / [N\,(a + b)] \\ &= (ad - bc) / [(a + b)(a + b + c + d)] \end{aligned}$$

確率差(PD)は分子 C*N - X*Y = 0 のとき、よって C = X*Y/N, つまり共起回数 C がその期待値と一致したときに最小の 0 になります。

確率差(PD)の b, c をゼロ(0)に近づけていくと、分子 ad - bc は増加し分母 (a + b)(a + b + c + d)は減少するので、PD は次第に増加します。そして、非負の b と c が最小値(0)に達したとき確率差(PD)は最大になります[11]。

そこで確率差の最大値(PD.max)は

$$\begin{aligned} PD.max. &= a\,d / [a\,(a + d)] \qquad \leftarrow\ b = c = 0 \\ &= d / (a + d) \end{aligned}$$

範囲を[0, 1]にした規定した確率差を「比較選択率」(Comparative Selection Ratio: CSR)とします[12]。

$$\begin{aligned} CSR &= PD / PD.max. \\ &= \{(ad - bc) / [(a + b)(a + b + c + d)]\} / [d / (a + d)] \\ &= (ad - bc)(a + d) / [d\,(a + b)\,(a + b + c + d)] \end{aligned}$$

[11] 概念的にも b = c = 0 ということが、X が Y を選択しない回数(b)と、Y が X によって選択されなかった回数(c)が、どちらもゼロ(0)であることを意味するので、そのとき Y 本来の確率 P(Y)と比較した、X が Y を選択するときの確率 P(Y|X)が最大になる、ということが納得できます。

[12] 当然ですが b = c = 0 を CSR にあてはめると

$$CSR(b=c=0) = [a\,d\,(a+d)] / [d\,a\,(a+d)] = 1$$

●muy bien の相互情報量・単純選択率・比較選択率

あるスペイン語の資料で調べると、*muy* (= 'very')という語の頻度(X)が 120, *bien* (= 'well')の頻度(Y)が 167, *muy* + *bien* の頻度(C)が 47, 全語数(N)が 26578 でした。次の表によって、この 4 つの数値(X, Y, C, N)から a, b, c, d の数を求めます。

muy: bien	bien+	Y(−)	sum
muy:+	a = 47	b = 73	a+b = 120
muy:-	c = 120	d = 26338	c+d = 26458
sum	a+c = 167	b+d = 26411	N = 26578

よって *muy*→*bien* の相互情報量(MI)・規定相互情報量(R.MI)・単純選択率(SS)・比較選択率(CS)は

MI(muy, bien) = $\log_2$ (47 * 26578) / (120*167) = 5.962
R.MI(muy, bien)
= ln (47 * 26578) / (120*167) / ln[(47+26338) / 47] = .653
SS(muy, bien) = 47 / 120 = .392
CS(muy, bien)
= (47*26338 - 73*120)*(47+26338) / (26338*120*26578) = .386

一方 *muy*←*bien* のそれぞれの値は

MI(bien, muy) = $\log_2$ (47 * 26578) / (167*120) = 5.962
R.MI(bien, muy)
= ln (47 * 26578) / (167*120) / ln[(47+26338) / 47] = .653
SS(bien, muy) = 47 / 167 = .281
CS(bien, muy)
= (47*26338 - 120*73)*(47+26338) / (26338*167*26578) = .277

相互情報量(MI)は規定化されていないので最大値が定まりません。規定相互情報量(R.MI)は[0, 1]の範囲に規定化されますが、*muy*→*bien* と *muy*←*bien* の方向性は関知しません。単純選択率(SS)と比較選択率(CS)は *muy*→*bien* と *muy*←*bien* の方向性を関知します。単純選択率(SS)は条件付き確率だけで計算しますが、比較選択率(CS)は条件付き確率と選択された語の本来の確率を考慮に入れて比較します。

6.4. 距離

6.4.1. 単純距離

2 つの数値(x, y)の間にある「距離」(D)はその差の絶対値を使って測るこ

とができます。たとえば x = 3, y = 5 であれば D(3, 5) = 2 となります。

$$D(x, y) = |x - y| = |3 - 5| = 2$$

次に x = (x1, x2) = (3, 4), y = (y1, y2) = (5, 2)という 2 次元の平面上の 2 つの座標であれば[13]

$$D(x, y) = [(x_1 - y_1)^2 + (x_2 - y_2)^2]^{1/2} = [(3 - 5)^2 + (4 - 2)^2]^{1/2} = 2.828$$

さらに 3 次元、4 次元、…、として次元数を増やすと次の「ユークリッド距離」(Euclidean distance: ED)になります[14]。

$$ED(x, y) = [(x_1 - y_1)^2 + (x_2 - y_2)^2 + \ldots + (x_k - y_k)^2]^{1/2}$$

$$[\Sigma_{(k=1,n)} (x_k - y_k)^2]^{1/2}$$

次に、2 つの座標に限らず、p 個の座標をもつデータ行列の中の 2 つの列 (x_i, x_j)の間のユークリッド距離は

$$ED(x_i, x_j) = [\Sigma_{(k=1,n)} (x_{ki} - x_{kj})^2]^{1/2} \ (i, j = 1, 2, .., p)$$

ユークリッド距離は、それぞれの対の差を 2 乗して次々に全部足し、その平方根を求めた値です。このままではデータの次元(n)が増えると、距離がどんどん大きくなるので、それぞれの対の差を 2 乗して次々に全部足し、個数(n)で割って平均をとり、その平均の平方根を求めます。これを「単純距離」(Simple Distance: SD)とよびます。よって

$$SD(x_i, x_j) = \{[\Sigma_{(k=1,n)} (x_{ki} - x_{kj})^2] / n\}^{1/2} \ (i, j = 1, 2, .., p)$$

たとえば下左表(X)の A と B の間の単純距離は

$$SD(A, B) = \{(10 - 19)^2 + (11 - 7)^2 + (0 - 0)^2 + (0 - 1)^2] / 4\}^{1/2} = 4.950$$

X	A	B	C	D	E
d1	10	19	14	7	12
d2	11	7	10	0	1
d3	0	0	1	12	1
d4	0	1	2	3	3

X	A	B	C	D	E
A	.000	4.950	2.345	8.411	5.339
B	4.950	.000	3.000	9.233	4.743
C	2.345	3.000	.000	8.231	4.637
D	8.411	9.233	8.231	.000	6.062

[13] ピタゴラスの「三平方の定理」(Pythagorean theorem)を使います。

[14] n = 1 のユークリッド距離は最初に見た絶対値による距離を同じです。

$$D(x, y) = [(x_1 - y_1)^2]^{1/2} = |x - y|$$

ここで絶対値を使うのは、距離は必ず非負になる、という性質があるからです。

E	5.339	4.743	4.637	6.062	.000

距離は互いに近い関係にあるとき小さな値になり、自己との距離はゼロになります。よって相関係数や連関係数とは大小関係が逆になります。また最小値はゼロですが、最大値はデータによって定まりません。

●単純近接

2 つのデータセットの対応する成分間の近接度(Proximity: Prox)の平均を近接(Simple Proximity: SP)とよびます。はじめに近接度(Prox)を次のように定義します[15]。

$$Prox(x, y) = 1 - |x - y| / Max(x, y)$$

上式の x, y は比較する 2 つの値、|x - y|は両者の差の絶対値、Max(x, y)は x と y の最大値(大きな方の値)です。たとえば、(2, 5)の近接度は 1 - |2 - 5| / max(2, 5) = 1 - 3/5 = .4 です。近接度の範囲は[0, 1]です[16]。

単純近接(SP)は 2 つのベクトルの成分間の近接度の平均とします(n: データ数)。

$$SP = \{\Sigma (i)\ Prox[x(i), y(i)]\} / n$$

単純近接(SP)のベースとなる近接度(Prox)は、個別の成分間の近接の度合いをその相対的な数値にして計算するので、たとえば先に見た Prox(2, 5) = .4 と、Prox(20, 50) = 1 - 30/50 = .4 は同じになります。近接にはこの性質があるために、先述の相関や距離で、外れ値が大きく作用する問題を回避することができます。たとえば、次の表で、相関・距離（全限定距離）・近接を比較しましょう。次の表には、d7:v2 と d7:v3 に外れ値があります。

D3	v1	v2	v3
d1	1	3	8
d2	3	5	7
d3	5	7	5
d4	7	8	4
d5	4	9	3
d6	8	9	2
d7	9	41	62

[15] 近接度(Prox)は分離度(Sep)の 1 の補数です（→分散）。Prox = 1 - Sep.

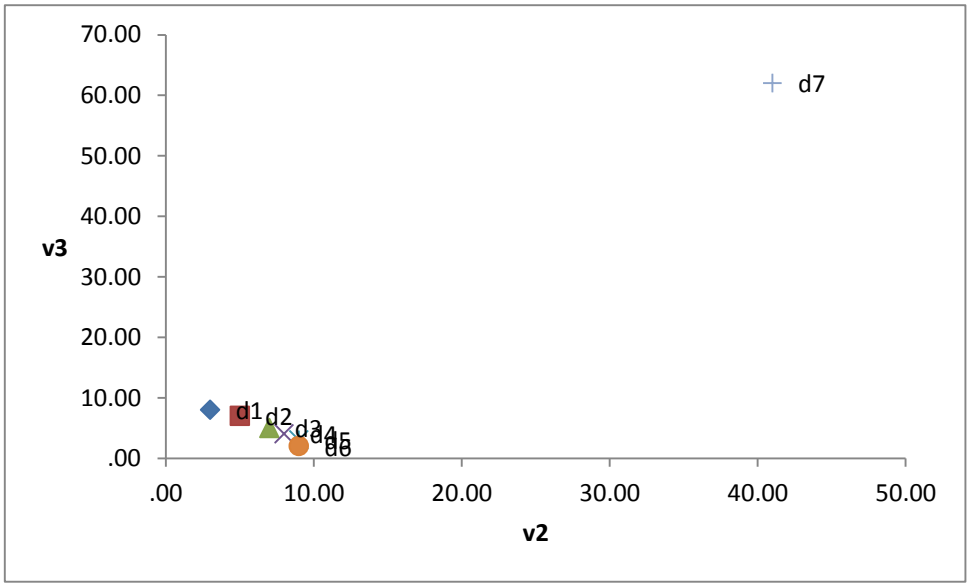
[16] X, Y を非負値(0 または正値)とします。近接度の最大値 1 は X=Y のときで、最小値 0 は X または Y が 0 のときです。X=Y=0 のときは、両者が完全に近接しているので、その近接度を 1 とします。

相関	v1	v2	v3
v1	1.00	.68	.50
v2	.68	1.00	.97
v3	.50	.97	1.00

距離	v1	v2	v3
v1	1.00	.80	.67
v2	.80	1.00	.85
v3	.67	.85	1.00

近接	v1	v2	v3
v1	1.00	.58	.47
v2	.58	1.00	.50
v3	.47	.50	1.00

このように相関(v2, v3)が大きな数値(.97)を示していますが、これは次の図が示すように、d7 の外れ値が強く作用したためです。



距離(v2, v3)も高くなっていますが(.85)、これも d7 の次の x 座標と図を y 座標の差が極端に大きいことが原因です。

6.4.2. 限定距離

先に見た単純距離の範囲を[0, 1]に限定した距離を「限定距離」(Limited Distance: LD)とよびます。距離を[0, 1]の範囲にするために、はじめにデータ全体を行の最大値と最小値を使って限定化します（→「得点」「限定得点」）。このようにデータの行の範囲を[0, 1]にすると成分間の差が 1 を超えることがなくなります。

LSr = D(S(Dnp, MnR(Dnp)), S(MxR(Dnp), MnR(Dnp)))
LSr' = D(Dnp), Rg(Dnp))
LD = SD(LSr) = SD(LSr') ←注[17]

ここで D は行列商、S は行列差、MnR は行の最小値、MxR が行の最大値、SD は単純距離を示します。

[17] 距離を計算するときに差をとるので、Xnp = D(Dnp), Rg(Dnp))のように簡単にしても距離係数の結果は変わりません。

X	A	B	C	D	E
d1	10	19	14	7	12
d2	11	7	10	0	1
d3	0	0	1	12	1
d4	0	1	2	3	3

X	A	B	C	D	E
A	.000	.449	.378	.875	.682
B	.449	.000	.303	.844	.522
C	.378	.303	.000	.728	.450
D	.875	.844	.728	.000	.506
E	.682	.522	.450	.506	.000

6.4.3. 標準距離

たとえば、h1(10, 19, 14, 7, 12)と h4(0, 1, 2, 3, 3)のように、データの規模が大きく異なるとき、そのまま A と B の距離を計算すると不都合なことになります。さらに、たとえば身長(cm)と体重(kg)のように単位が異なるときには明らかに不都合です[18]。そこで、このようなデータの横行の標準偏差が列間の距離に影響することを考慮して、はじめにデータをその行の標準偏差で割って行を標準化した行列（行標準得点行列: Standard Score in row: SSr)を用意し（→「得点」「標準得点」）、その単純距離(SD)を「標準距離」(Standard Distance: StD)として計算します[19]。

SSr = D((Dnp), SdR(Dnp))
StD = SD(SSr)

X	A	B	C	D	E
d1	10	19	14	7	12
d2	11	7	10	0	1
d3	0	0	1	12	1
d4	0	1	2	3	3

X	A	B	C	D	E
A	.000	1.275	1.003	2.225	1.716
B	1.275	.000	.831	2.288	1.392
C	1.003	.831	.000	1.890	1.109
D	2.225	2.288	1.890	.000	1.347
E	1.716	1.392	1.109	1.347	.000

6.4.4. 標準３距離

上左表(SSr)のように標準得点行列の成分の絶対値はしばしば 1 を超えます。そのため距離も 1 を超えることがあります。一般に、データはその標準偏差を 3 倍した値以上または以下になることは極めて稀なので（→「確率」）、先の標準化の分母を標準偏差ではなく、その 3 倍にすることを提案します。次はデータ行列 Dnp の行を、その標準偏差 Sd * 3 で割って変換

[18] 相関係数を求めるときには標準測度が使われているので、単位が異なっていても不都合はありません。
[19] 行から行の平均を引いて、行の標準偏差で割ると行の標準化得点になります。この標準化得点を使っても距離行列の結果は、距離の計算式の分子の引き算のそれぞれの項から平均を引いているので、結果は同じになります。

した結果を標準化 3 行列(Standard Score 3 in row: SS3r)とし、その標準距離(Standard Distance 3: SD3)を計算した結果です。

SS3r = D(S(Dnp, AvR(Dnp)), SdR(Dnp) * 3)
SD3r = 1 - SD(SS4r)

X	A	B	C	D	E
d1	10	19	14	7	12
d2	11	7	10	0	1
d3	0	0	1	12	1
d4	0	1	2	3	3

X	A	B	C	D	E
A	.000	.425	.334	.742	.572
B	.425	.000	.277	.763	.464
C	.334	.277	.000	.630	.370
D	.742	.763	.630	.000	.449
E	.572	.464	.370	.449	.000

限定距離・標準距離・標準３距離は変数の規模が大きく異なるときに使われますが、言語データの中の同じ性質をもつ語の頻度のように同じ条件で計測された得点であれば、むしろ標準化せずに、その頻度差そのものを考慮して単純距離を使うほうがよいことがあります。標準化するとすべての語の頻度の差が均されて大きな情報が失われるからです。たとえば各種のスペイン語テキストを各種の前置詞(a, de, en, con, por, ..., ante, tras)を使って比較するとき、前置詞によって頻度が大きく異なるので限定距離・標準距離･標準３距離を使うとよいでしょう。一方､定冠詞の 4 形態(el, la, los, las)を使ってテキスト間の距離を調べるときには、どれも同じ性質をもつと考えれば、その情報を生かすために単純距離を選択すべきです。

■相関と距離：語末 e の異常な脱落形

相関係数と距離係数の違いを数値とグラフで確認します。次は、中世スペイン語で語末の e が異常に脱落したケースの頻度表と、その相関行列・距離行列です(*a: adelant, en: end, es: est, pa: part, pr: present, v:veint*)。

Año	a	en	es	pa	pr	v
1200			4	13		
1220	8	3	23	18	2	
1240	16	7	8	11	2	
1260	30	3	9	46	40	9
1280	29	17	15	50	35	26
1300	22	1	6	29	59	1
1320	12		6	83	44	11
1340	17		4	22	23	2
1360	10	3	1	13	32	20
1380	51	13	3	29	66	10
1400	20	1	2	64	121	17

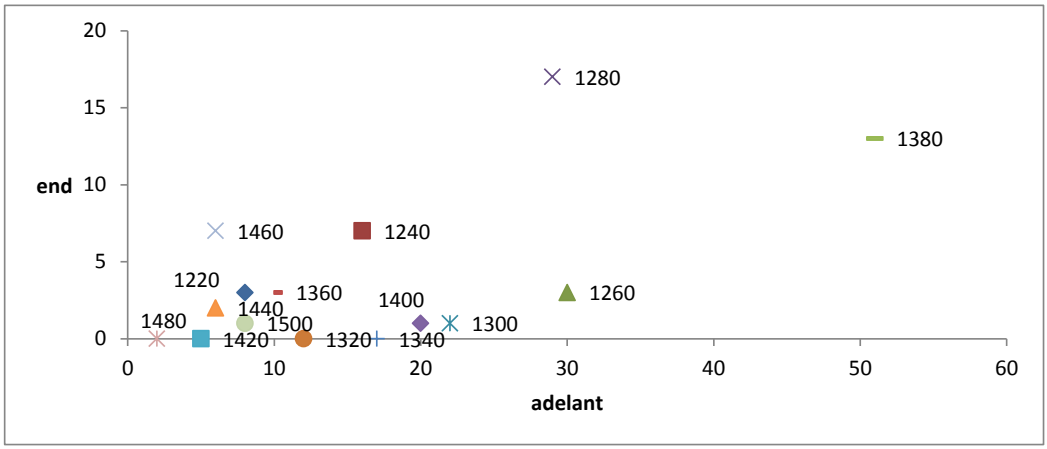
CC	a	en	es	pa	pr	v
adelant	1.000	.645	.172	.405	.508	.450
end	.645	1.000	.318	.079	.062	.512
est	.172	.318	1.000	.237	-.246	.114
part	.405	.079	.237	1.000	.614	.584
present	.508	.062	-.246	.614	1.000	.504
veint	.450	.512	.114	.584	.504	1.000

StD	a	en	es	pa	pr	v
adelant		.377	.405	.475	.603	.410
end	.377		.256	.660	.802	.209
est	.405	.256		.579	.841	.344

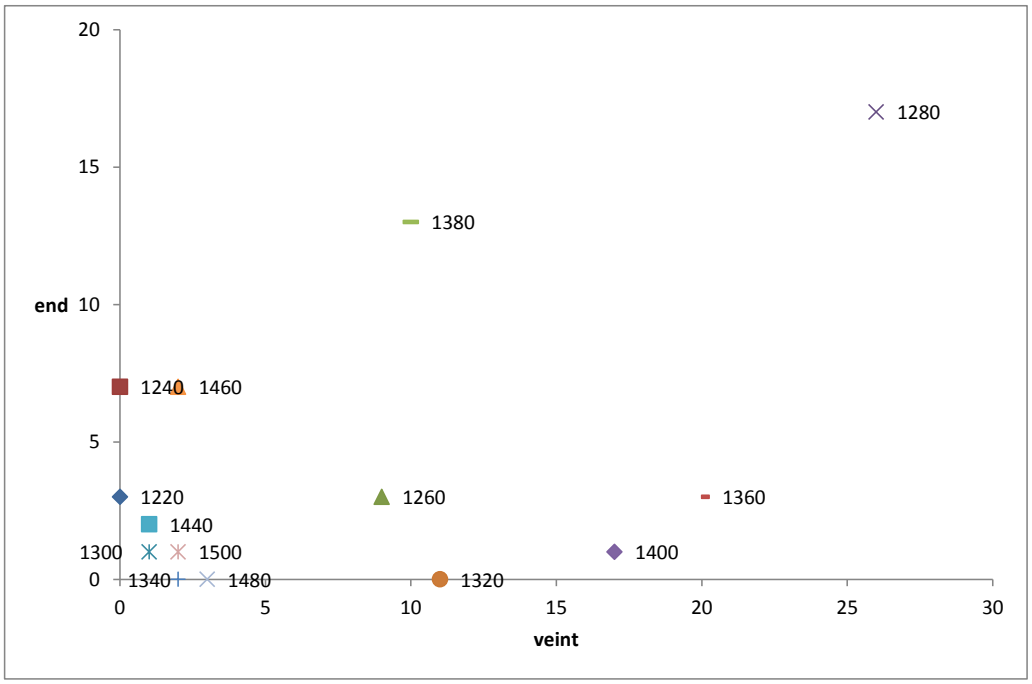
1420	5		1	10	32	
1440	6	2	1	15	24	1
1460	6	7		2	26	2
1480	2		3	17	23	3
1500	8	1		2	16	2

part	.475	.660	.579		.600	.621
present	.603	.802	.841	.600		.736
veint	.410	.209	.344	.621	.736	

相関係数(CC)が一番大きなペアは *end-a(delan)t* です(.645)。一方、距離係数(StD)が一番近いペアは *end-veint* です(.209)。このように両者は一致しません。その理由を探るためにそれぞれのペアの散布図を見ましょう。



上図のように、*end: adelant* は 1280 と 1380 のデータが強く働いて、一定の相関を示しています



一方、上図で *end:veint* の関係を見ると、データが左下に集中していることがわかります。相関はそれほど強くありません(.512)。

このように、相関係数は変数の直線的な方向の「動き」の関係性を示し、距離係数は、変数が占める座標の「位置」の近さを示すので解釈が異なります。データの流れが X 軸上の動きに合わせて Y 軸上で動くと相関が高くなります。一方、X 座標と Y 座標が近いデータが多数を占めると距離が近くなります。

6.4.5. 平均距離

次の表はデータ h1, h2, h3, h4 の属性 X と Y の頻度とその差（X-Y）、X と Y の差の 2 乗$(X-Y)^2$、X の 2 乗(X^2)、Y の 2 乗(Y^2)、X の 2 乗と Y の 2 乗の和(X^2+Y^2)、そして最後の列は X と Y の差の 2 乗を、X の 2 乗と Y の 2 乗の和で割った値$(X-Y)^2/(X^2+Y^2)$を示します[20]。最後の「平均」は最後の列の平均です。

M	X	Y	X-Y	$(X-Y)^2$	X^2	Y^2	X^2+Y^2	$(X-Y)^2/(X^2+Y^2)$
h1	10	19	-9	81	100	361	461	0.176
h2	11	7	4	16	121	49	170	0.094
h3	0	0	0	0	0	0	0	0.000
h4	0	1	-1	1	0	1	1	1.000
							平均	0.317

この最後の平均を「平均距離」(Mean Distance: MD)として定義します。よって平均距離(MD)は

$$
\begin{aligned}
MD &= 1/n \ \Sigma_i [(x_i - y_i)^2 / (x_i^2 + y_i^2)] \\
&= 1/n \ \Sigma_i [(x_i^2 + y_i^2 - 2 x_i y_i)) / (x_i^2 + y_i^2)] \\
&= 1/n \ \Sigma_i [1 - 2 x_i y_i / (x_i^2 + y_i^2)] \\
&= 1/ n \ (n - \ \Sigma_i [2 x_i y_i / (x_i^2 + y_i^2)] \\
&= 1 - 1/n \ \Sigma_i [2 x_i y_i / (x_i^2 + y_i^2)]
\end{aligned}
$$

ここで上式の第 2 項 $1/n \ \Sigma_i [2 x_i y_i / (x_i^2 + y_i^2)]$を「平均近接」(Mean Proximity: MP)として定義します。

$$MP = 1/n \ \Sigma_i [2 x_i y_i / (x_i^2 + y_i^2)]$$

よって MD と MP はそれぞれ 1 の補数になります。

$$MD + MP = 1, \ MD = 1 - MP, \ MP = 1 - MD$$

次の平均距離(MD)

[20] h3 のように$(X-Y)^2 = 0$, $X^2+Y^2 = 0$ のとき、$(X-Y)^2/(X^2+Y^2)$は 0/0 となり数学的には定義できませんが、「距離」という概念を考慮してゼロ(0)として計算します。

$$MD = 1/n \ \Sigma_i [(x_i - y_i)^2 / (x_i^2 + y_i^2)]$$

の分子を見ると $\Sigma_i (x_i - y_i)^2 = 0$ の場合に MD が最小値(=0)になることがわかります。これは $x_i = y_i$ (i=1,2,…,n)の場合です。つまり x_i と y_i が i = 1, 2, …, n ですべて一致する場合です[21]。

次に MD が最大となる場合は、$x_i = 0$ (i=1,2,…,n)、または $y_i = 0$ (i=1,2,…,n)の場合であるはずです[22]。その最大値は $y_i = 0$ (i=1,2,…,n)のとき[23]、

$$MD(x, 0) = 1/n \ \Sigma_i [(x_i - 0)^2 / (x_i^2 + 0^2)]$$
$$= 1/n \ \Sigma_i x_i^2 / x_i^2 = 1/n \ \Sigma_i 1 = 1/n \ n = 1$$

同様にして、$x_i = 0$ (i=1,2,…,n)のときは

$$MD(0, y) = 1/n \ \Sigma_i [(0 - y_i)^2 / (0^2 + \Sigma_i y_i^2]$$
$$= 1/n \ \Sigma_i (y_i^2/y_i^2) = 1/n \ \Sigma_i 1 = 1/n \ n = 1$$

平均近接(MP)の範囲は[0, 1]になりますが条件は平均距離の場合と逆転します。

$MP(x, x) = 1 - MD(x, x) = 1 - 1/n \ \Sigma_i [(x_i - x_i)^2 / (x_i^2 + x_i^2)] = 1 - 0 = 1$
$MP(x, 0) = 1 - MD(x, 0) = 1 - 1/n \ \Sigma_i [(x_i - 0)^2 / (x_i^2 + 0^2)] = 1 - 1 = 0$
$MP(0, y) = 1 - MD(0, y) = 1 - 1/n \ \Sigma_i [(0 - y_i)^2 / (0^2 + y_i^2)] = 1 - 1 = 0$

次は入力行列(M)と、その A, B 列で計算した平均距離(MD)・平均近接(RP)、および D, E 列で計算した平均距離(MD)・平均近接(RP)を示します。MD, RP の式と、導出の途中の計算 X-Y, $(X-Y)^2$, X^2, Y^2, X^2+Y^2 も参照してください。

$$MD = (X-Y)^2 / (X^2+Y^2), \ RP = 1 - MD$$

M	A	B	C	D	E
h1	10	19	14	7	12
h2	11	7	10	0	1
h3	0	0	1	12	1
h4	0	1	2	3	3

はじめに X と Y の平均距離(MD)と平均近接(MP)を見ます。

---

[21] x_i, y_i (i=1,2,…,n)の対の 1 つでも一致しないときは $\Sigma_i (x_i - y_i)^2 = 0$ になりません。

[22] ある到着点までの距離は開始点(0)からの距離が最大になるからです。

[23] この式を見ると、y がすべてゼロ(0)であれば、x がどのような値であろうと、x と y の平均距離は最大の 1 になることがわかります。y のすべてがゼロに近い値でも、x と y の平均距離は最大の 1 に近似します。よって原点（または原点に近接する点）からの距離の比較をするときには、すべての距離が 1（または 1 に近い数値）になるので、使えません。。

M	X	Y	X-Y	$(X-Y)^2$	X^2	Y^2	X^2+Y^2	$(X-Y)^2/(X^2+Y^2)$
h1	10	19	-9	81	100	361	461	0.176
h2	11	7	4	16	121	49	170	0.094
h3	0	0	0	0	0	0	0	0.000
h4	0	1	-1	1	0	1	1	1.000
						MD		0.317
						MP		0.683

次はさらに大きな距離を示す例です。

M	X	Y	X-Y	$(X-Y)^2$	X^2	Y^2	X^2+Y^2	$(X-Y)^2/(X^2+Y^2)$
h1	7	12	-5	25	49	144	193	0.130
h2	0	1	-1	1	0	1	1	1.000
h3	12	1	11	121	144	1	145	0.834
h4	3	3	0	0	9	9	18	0.000
						MD		0.491
						MP		0.509

次は X = Y のときの最小 MD (=0)、最大 MP (=1)の場合です。

M	X	Y	X-Y	$(X-Y)^2$	X^2	Y^2	X^2+Y^2	$(X-Y)^2/(X^2+Y^2)$
h1	14	14	0	0	196	196	392	0.000
h2	10	10	0	0	100	100	200	0.000
h3	1	1	0	0	1	1	2	0.000
h4	2	2	0	0	4	4	8	0.000
						MD		0.000
						MP		1.000

次は Y = 0 のときの最大 MD (=1)、最小 MP (=0)の場合です。

M	X	Y	X-Y	$(X-Y)^2$	X^2	Y^2	X^2+Y^2	$(X-Y)^2/(X^2+Y^2)$
h1	14	0	14	196	196	0	196	1.000
h2	10	0	10	100	100	0	100	1.000
h3	1	0	1	1	1	0	1	1.000
h4	2	0	2	4	4	0	4	1.000
						MD		1.000
						MP		0.000

下左表はデータ行列(M)、下右表はその平均近接対称行列 MP です。

M	A	B	C	D	E
h1	10	19	14	7	12
h2	11	7	10	0	1
h3	0	0	1	12	1
h4	0	1	2	3	3

MP	A	B	C	D	E
A	1.000	.683	.485	.235	.291
B	.683	1.000	.674	.312	.446
C	.485	.674	1.000	.472	.777
D	.235	.312	.472	1.000	.509
E	.291	.446	.777	.509	1.000

プログラム

```
function DisMnM(Xnp) { //平均距離対称行列 (Mean distance)
        var n = NR(Xnp), p = NC(Xnp), Dpp = NewMt(p,p); Dpp[0][0]="[M.Dist.]";
        for(var i = 1; i <= p; i++) {
                Dpp[0][i] = Dpp[i][0] = Xnp[0][i]; Dpp[i][i] = 0; //表頭；表側；対角
成分
        }
        for(var i = 1; i <= p-1; i++) { for(var j = i+1; j <= p; j++) { //距離行列
                var xy = x2 = y2 = 0;
                for(var k = 1; k <= n; k++) {
                        xy = Pow(Xnp[k][i] - Xnp[k][j], 2);
                        x2 = Pow(Xnp[k][i], 2);
                        y2 = Pow(Xnp[k][j], 2);
                        Dpp[i][j] += (x2+y2==0)? 0: xy / (x2+y2);
                } Dpp[j][i] = Dpp[i][j] /= n;
        }} return Dpp;
}
```

●実数データの平均距離・平均近接

以上では説明を簡単にするために非負データの平均距離(MD)・平均近接(MP)を扱いました。実はこの平均距離・平均近接は以下に示すように負のデータも同様に扱うことができます。

はじめに平均近接(MP)は

$$MP = 1/n \quad \Sigma_i [2\ x_i\ y_i / (x_i^2 + y_i^2)]$$

ここで y_i がすべて相手(x_i)の負($-x_i$)であれば後述するように平均近接(MP)が最小(= -1)になります（→直後の項を参照「平均近接の最小値」）。

$$MP(x, -x) = 1/n \quad \Sigma_i \{2\ x_i*(-x_i) / [x_i^2 + (-x_i)^2]\}$$
$$= 1/n \quad \Sigma_i (-2\ x_i^2 / 2\ x_i^2)$$

$$= 1/n \ \Sigma_i -1 = 1/n \ -n = -1$$

$$MP(0, -x) = 1/n \ \Sigma_i \{2*0*(-x_i) / [0^2 + (-x_i)^2]*$$

$$= 1/n \ \Sigma_i [0 / (-x_i)^2] = 1/n \ 0 = 0$$

$$MP(-x, 0) = 1/n \ \Sigma_i \{2*(-x_i)*0 / [(-x_i)^2 + 0^2]\}$$

$$= 1/n \ \Sigma_i [0 / (-x_i)^2] = 1/n \ 0 = 0$$

よって

$$MD(x, -x) = 1 - MP(x, -x) = 1 - (-1) = 2$$

$$MD(0, -x) = 1 - MP(-x, 0) = 1 - 0 = 1$$

このように、非負データの平均距離(MD)の範囲は[0, 1]でしたが、実数データの平均距離(MD)の範囲は[0, 2]になります。また、非負データの平均近接(MP)の範囲は[0, 1]でしたが、実数データの平均近接(MP)の範囲は[-1, 1]になります。

係数	非負データ	実数データ
平均距離係数(MD)	[0, 1]	[0, 2]
平均近接係数(MP)	[0, 1]	[-1, 1]

次は h1, h2 の Y を負(-20, -5)にしたときの平均距離係数(MD)と平均近接係数(MP)を計算した結果です。

M	X:A	Y:B	X-Y	$(X-Y)^2$	X^2	Y^2	X^2+Y^2	$(X-Y)^2/(X^2+Y^2)$
h1	10	-20	30	900	100	400	500	1.800
h2	11	-5	16	256	121	25	146	1.753
h3	0	0	0	0	0	0	0	0.000
h4	0	1	-1	1	0	1	1	1.000
							MD	1.138
							MP	-0.138

●実数データの平均距離の最大値と平均近接の最小値

次の平均距離係数(MD)の式で

$$MD = 1/n \ \Sigma_i (x_i - y_i)^2 / (x_i^2 + y_i^2)$$

一見したところ y が x の単なる負数(-x)ではなく、さらに小さな数、たとえば-x-a になる方が MD が大きくなるのではないかと思われるかもしれません。上の式の右辺の分子の y_i を大きくすると、分子の $(x_i - y_i)^2$ が大きくなるからです。しかし、このとき分母も大きくなるので、MD は 2 より大きくなりません。このことを次のような実験をして具体的に確かめましょう。

M	X:A	Y:B	X-Y	$(X-Y)^2$	X^2	Y^2	X^2+Y^2	$(X-Y)^2/(X^2+Y^2)$
h1	10	-10	20	400	100	100	200	2.000
h2	11	-11	22	484	121	121	242	2.000
h3	2	-2	4	16	4	4	8	2.000
h4	3	-3	6	36	9	9	18	2.000
							MD	2.000
							MP	-1.000

M	X:A	Y:B	X-Y	$(X-Y)^2$	X^2	Y^2	X^2+Y^2	$(X-Y)^2/(X^2+Y^2)$
h1	10	-10	20	400	100	100	200	2.000
h2	11	-11	22	484	121	121	242	2.000
h3	2	-2	4	16	4	4	8	2.000
h4	3	-100	103	10609	9	10000	10009	1.060
							MD	1.765
							MP	-0.765

上の実験では h4:Y の-3 を-100 にしましたが、-1000 にしても-10000 にしても MD は 2 以上にはなりません。そして、h1, h2, h3, h4 の X, Y にどのような値を設定して実験しても MP は[-1, 1]の範囲になります。このことを一般化して数式の中で確かめましょう。

$$MP = 1/n \ \Sigma_i \ [2\ x_i\ y_i / (x_i^2 + y_i^2)]$$

$y_i = -x_i$ の中の 1 つ$(-x_k)$でも$-x_k-a$ になれば(a は任意の正数)、$2\ x_i\ y_i / (x_i^2 + y_i^2)$の分子は

$$2\ x_i\ y_i = 2\ x_i*(-x_k-a) = -2\ x_i*(x_k+a)$$

また $2\ x_i\ y_i / (x_i^2 + y_i^2)$の分母は

$$\begin{aligned} x_i^2 + y_i^2 &= x_i^2 + (-x_k-a)^2 \\ &= x_i^2 + x_i^2 + 2\ x_k\ a + a^2 \\ &= 2\ x_i^2 + 2\ x_k\ a + a^2 \\ &= 2x_i(x_i + a) + a^2 \end{aligned}$$

よって

$$MP(x, -x-a) = 1/n \ \Sigma_i \ \{-2\ x_i*(x_k+a) / [2\ x_i*(x_k+a) + a^2]\}$$

ここで

$$c /(c + d) < 1,\ -c /(c + d) > -1 \quad \ldots\ d > 0$$

$c = 2\ x_i*(x_k+a)$, $d = a^2$ とすると

$$MP(x, -x-a) = 1/n\ \ \Sigma_i \{-2\ x_i*(x_k+a) / [2\ x_i*(x_k+a) + a^2]\}$$
$$= 1/n\ \ \Sigma_i\ -c/(c+d) > 1/n\ \ \Sigma_i\ -1 = 1/n * (-n) = -1$$

よって

$$MP(x, -x-a) > -1 \quad \ldots a > 0$$

このように MP(x, -x-a)は-1 より大きな数になります。そして a=0 のときに MP=-1 になるので、最小値は $y_i = -x_i$ のときに-1 になります。

このことは $y_i = -x_i$ の中の 1 つ$(-x_k)$でも $-x_k+a$ であっても(a は任意の正数)同様であることを以下で確かめましょう。

$2\ x_i\ y_i / (x_i^2 + y_i^2)$の分子は

$$2\ x_i\ y_i = 2\ x_i*(-x_k+a) = -2\ x_i*(x_k-a)$$

また $2\ x_i\ y_i / (x_i^2 + y_i^2)$の分母は

$$x_i^2 + y_i^2 = x_i^2 + (-x_k+a)^2$$
$$= x_i^2 + x_i^2 - 2\ x_k\ a + a^2$$
$$= 2\ x_i^2 - 2\ x_k\ a + a^2$$
$$= 2x_i(x_i - a) + a^2$$

よって

$$MP(x, -x-a) = 1/n\ \ \Sigma_i \{-2\ x_i*(x_k-a) / [2\ x_i*(x_k-a) + a^2]\}$$

ここで

$$c /(c + d) < 1,\ -c /(c + d) > -1 \quad \ldots d > 0$$

$c = 2\ x_i*(x_k-a)$, $d = a^2$ とすると

$$MP(x, -x-a) = 1/n\ \ \Sigma_i \{-2\ x_i*(x_k-a) / [2\ x_i*(x_k-a) + a^2]\}$$
$$= 1/n\ \ \Sigma_i\ -c/(c+d) > 1/n\ \ \Sigma_i\ -1 = 1/n * (-n) = -1$$

よって

$$MP(x, -x-a) > -1 \quad \ldots a > 0$$

6.4.6. 規定距離[24]

「距離」（Distance: D）の基本は次の式で示すように 2 つの数値（x, y）の差です。

$$D(x, y) = |x - y|$$

上式で絶対値を使った理由はたとえば D(2, 5)と D(5, 2)が同じになるからです。具体例では現在地から 2km 離れた地点(X)と、現在地から 5km 離れた地点(Y)の距離は、現在地から 5km 離れた地点(Y)と、現在地から 2km 離れた地点(X)の距離と等しくどちらも 3km になります。

```
-----|-----|-----|-----|-----|-----|-----|-----|
          X   ← 3 →   Y
```

ここで距離は必ず非負数になることを確認します。このことが上の式で絶対値記号を用いた理由です。

さて、絶対値は次のように指数を使った式で計算できます。

$$D(x, y) = \sqrt{(x - y)^2} = [(x - y)^2]^{1/2}$$

よって

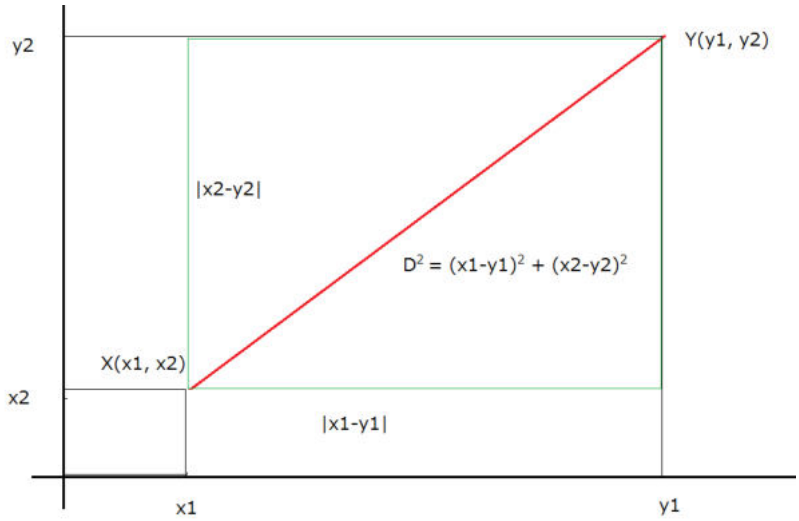
$$D^2(x, y) = (x - y)^2$$

上の例では x, y という 2 地点が直線上のあることを想定していますが、両者が $X(x_1, x_2)$, $Y(y_1, y_2)$という平面座標にあるときの両者の距離は

$$D^2(x, y) = (x_1 - y_1)^2 + (x_2 - y_2)^2$$

上の式はピタゴラスの定理によって下図の赤線の長さ（の 2 乗）を示します。

[24] この節の説明は前節の「平均距離」と類似した展開になります。平均距離と規定距離の特徴の比較については後述します。



同様にして 3 次元空間内にある $X(x_1, x_2, x_3)$と $Y(y_1, y_2, y_3)$の距離は

$$D^2(x, y) = (x_1 - y_1)^2 + (x_2 - y_2)^2 + (x_3 - y_3)^2$$

よって n 次元空間内にある $X(x_1, x_2, \ldots, x_n)$と $Y(y_1, y_2, \ldots, y_n)$の距離はすべての i 個の座標の差の二乗和(Sum of Squared Differences: SSD)になります。

$$D^2(x, y) = (x_1 - y_1)^2 + (x_2 - y_2)^2 + \ldots + (x_n - y_n)^2 = \Sigma\,(x_i - y_i)^2$$
$$SSD(x, y) = [\,\Sigma\,(x_i - y_i)^2]^{1/2} \qquad \ldots(1)$$

たとえば次の X 列と Y 列の差の二乗和(SSD)は

M	X	Y	X-Y	(X-Y)^2
h1	10	19	-9	81
h2	11	7	4	16
h3	0	0	0	0
h4	0	1	-1	1
			SSD	98

以下では、この差の二乗和(SSD)から出発して[0, 1]の範囲に規定した「規定距離」(Regular Distance: RD)を計算する方法を考えます[25]。はじめに先の差二乗和(SSD)の式(1)の両辺を 2 乗して，次のように展開します[26]。

$$SSD^2(x, y) = \Sigma\,(x_i - y_i)^2 = \Sigma\,(x_i^2 + y_i^2 - 2x_iy_i) \quad \ldots(2)$$

[25] 「距離」の項で行った限定化という操作とは異なります。限定化は距離を計算してから範囲が[0, 1]になるように適用しましたが、規定化係数は、以下で見るように、その範囲が[0, 1]になることがはじめから数学的に証明されます。

[26] 式(2)は中学数学で習った式 $(x_i - y_i)^2 = x_i^2 + y_i^2 - 2\,x_i\,y_i$ を $i = 1, 2, \ldots, n$ まで足し上げたことになります。

$$\rightarrow \quad \Sigma (x_i - y_i)^2 = \Sigma (x_i^2 + y_i^2) - 2\Sigma x_i y_i$$
$$\rightarrow \quad \Sigma (x_i - y_i)^2 + 2\Sigma x_i y_i = \Sigma (x_i^2 + y_i^2)$$
$$\rightarrow \quad [\Sigma (x_i - y_i)^2 + 2\Sigma_i x_i y_i] / \Sigma (x_i^2 + y_i^2) = 1$$
$$\rightarrow \quad \Sigma (x_i - y_i)^2 / \Sigma (x_i^2 + y_i^2) + 2\Sigma x_i y_i / \Sigma (x_i^2 + y_i^2) = 1 \ \ldots (3)$$

この(3)の式の左辺第 1 項を「規定距離」(Regular Distance: RD)として定義し、左辺第 2 項を「規定近接」(Regular Proximity: RP)として定義します[27]。

$$RD(x, y) = \Sigma (x_i - y_i)^2 / \Sigma (x_i^2 + y_i^2) \quad \ldots(4)$$
$$RP(x, y) = 2\Sigma x_i y_i / \Sigma (x_i^2 + y_i^2) \quad \ldots(5)^{28}$$

規定距離(RD)と規定近接(RP)は次の関係になります。

$$RD + RP = 1,\ RD = 1 - RP,\ RP = 1 - RD \qquad \leftarrow (3), (4), (5)$$

(4)の規定距離(RD)の式の分子を見ると $\Sigma (x_i - y_i)^2 = 0$ の場合に RD が最小値(=0)になることがわかります。これは $x_i = y_i$ (i=1,2,...,n)の場合です。つまり x_i と y_i の対が i = 1, 2, ..., n ですべて一致する場合です[29]。このとき比較する 2 点は一致するのでその距離 RD は当然ゼロ(0)になります[30]。

次に、X, Y が頻度のような非負数であるとき、RD が最大となる場合と、そのときの最大値を求めます。RD が最大となる場合は、$x_i = 0$ (i=1,2,...,n)、

---

[27] 近接係数については次節で扱います。

[28] 式(5)は次の「コサイン類似度」(Cos)と似ていますが、係数(=2)と分母が異なります(金明哲『テキストデータの統計科学入門』(岩波書店 2009: 161)。

$$Cos(x, y) = \Sigma_i x_i y_i / (\Sigma_i x_i^2 * \Sigma_i y_i^2)^{1/2}$$

Manly (1986: 53)は距離(D_2)として次の式を示しています。この右辺の第 2 項は上のコサイン類似度(Cos)です。

$$D_2(x, y) = 1 - \Sigma_i x_i y_i / (\Sigma_i x_i^2 * \Sigma_i y_i^2)^{1/2}$$

どちらも分母にある $\Sigma_i x_i^2$ と $\Sigma_i y_i^2$ のどちらかゼロの場合にコサイン類似度と距離(D_2)が計算できません。しかし $\Sigma_i x_i^2$ と $\Sigma_i y_i^2$ のどちらかゼロの場合は先に見たように、非負データの規定距離係数(RD)が最大になるときなので重要なポイントになります。実数データの場合は規定距離係数が中間値を示し、規定近接係数(RP)がゼロになりますから、これも重要なポイントです。 (Manly, Bryan F. J. (1986) *Multivariate Statistic Methods. A Primer.* London: Chapman & Hill.)

[29] 1 つの対でも一致しないときは $\Sigma_i (x_i - y_i)^2 = 0$ にはなりません。

[30] 非常に例外的な場合として $\Sigma_i x_i^2 = \Sigma_i y_i^2 = 0$、よって $x_i = y_i = 0$ (i = 1, 2, ..., n)の場合が考えられます。このとき RD の分母が 0 になるので、コンピュータは計算ができません。そこで、このとき分子も 0 になるので、はじめに RD(0, 0) = 0 を定義しておきます。つまり、すべての要素が 0 であるような 2 つの列は完全に一致するので、両者間の距離はゼロである、という意味です。

または $y_i = 0$ (i=1,2,…,n)の場合であるはずです[31]。最初に $y_i = 0$ (i=1,2,…,n)の場合の場合の最大値は[32]

$$RD(x, 0) = \Sigma (x_i - 0)^2 / \Sigma (x_i^2 + 0^2) = \Sigma x_i^2 / \Sigma x_i^2 = 1$$

同様にして $x_i = 0$ (i=1,2,…,n)の場合の最大値は

$$RD(0, y) = \Sigma (0 - y_i)^2 / \Sigma (0^2 + y_i^2) = \Sigma y_i^2 / \Sigma y_i^2 = 1 \quad \ldots(4')$$

次は X と Y の規定距離(RD)の計算例です。

H	X	Y	X-Y	$(X-Y)^2$	X^2	Y^2	X^2+Y^2	
h1	10	19	-9	81	100	361	461	
h2	11	7	4	16	121	49	170	
h3	0	0	0	0	0	0	0	
h4	0	1	-1	1	0	1	1	RD↓
			和→	98	和→		632	.155

RD(x, y) = 98 / 632 = .155

次はさらに大きな距離を示す例です。

H	X	Y	X-Y	$(X-Y)^2$	X^2	Y^2	X^2+Y^2	
h1	7	12	-5	25	49	144	193	
h2	0	1	-1	1	0	1	1	
h3	12	1	11	121	144	1	145	
h4	3	3	0	0	9	9	18	RD↓
			和→	147	和→		357	.412

RD(x, y) = 147 / 357 = .412

[31] ある到着点までの距離は開始点(0)からの距離が最大になるからです。

[32] この式を見ると、y がすべてゼロ(0)であれば、x がどのような値であろうと、x と y の平均距離は最大の 1 になることがわかります。y のすべてがゼロに近い値でも、x と y の規定距離は最大の 1 に近似します。よって原点（または原点に近接する点）からの距離の比較をするときには、すべての距離が 1（または 1 に近い数値）になるので、使えません。

次は X = Y のときの最小 RD (=0)を示します。

H	X	Y	X-Y	$(X-Y)^2$	X^2	Y^2	X^2+Y^2	
h1	14	14	0	0	196	196	392	
h2	10	10	0	0	100	100	200	
h3	1	1	0	0	1	1	2	
h4	2	2	0	0	4	4	8	RD↓
			和→	0		和→	602	.0000

RD(x, y) = 0 / 602 = .000

次は Y = 0 のときの最大 RD (=1)の場合です。

H	X	Y	X-Y	$(X-Y)^2$	X^2	Y^2	X^2+Y^2	
h1	14	0	14	196	196	0	196	
h2	10	0	10	100	100	0	100	
h3	1	0	1	1	1	0	1	
h4	2	0	2	4	4	0	4	RD↓
			和→	301		和→	301	1.000

RD(x, y) = 301 / 301 = 1.000

●実数データの規定距離・規定近接

以上では説明を簡単にするために非負データの規定距離(RD)を扱いました。実はこの規定距離は負のデータも同様に扱うことができることを以下に説明します。

規定近接(RP)は規定距離(RD)の 1 の補数になります。

$$RP(x, y) = 1 - RD(x, y) \qquad \ldots(1)$$

この規定近接(RP)の範囲は明らかに[0, 1]になりますが、その最小値と最大値の条件は、次のように規定距離(RD)の場合の逆です。

RP(x, x) = 1 - RD(x, x) = 1 - 0 = 1 (最大値)
RP(x, 0) = 1 - RD(x, 0) = 1 - 1 = 0 (最小値)
RP(0, y) = 1 - RD(0, y) = 1 - 1 = 0 (最小値)

さて、ここで y_i がすべて相手(x_i)の負($-x_i$)であれば、RP は最小の-1 になります（→後述「実数データの規定距離の最大値」）。このとき[33]:

$$RP(x, -x) = 2\Sigma_i [x_i*(-x_i)] / \Sigma_i [x_i^2 + (-x_i)^2] = -2\Sigma_i x_i / 2\Sigma_i x_i = -1 \quad \ldots(2)$$

[33] 逆に x_i がすべて相手(y_i)の負($-y_i$)であっても同様です。

$RP(0, -x) = 2*\Sigma_i [(-x_i)^2*0] / \Sigma_i [0 + (-x_i)^2] = 0 / \Sigma_i x_i^2 = 0$ …(3)
$RP(-x, 0) = 2*\Sigma_i [0*(-x_i)^2] / \Sigma_i [(-x_i)^2 + 0] = 0 / \Sigma_i x_i^2 = 0$ …(4)

よって

$RD(x, -x) = 1 - RP(x, -x) = 1 - (-1) = 2$ ←(1), (2)
$RD(0, -x) = 1 - RP(0, -x) = 1 - 0 = 1$ ←(1), (3)
$RD(-x, 0) = 1 - RP(-x, 0) = 1 - 0 = 1$ ←(1), (4)

このように規定距離(RD)の範囲は非負データでは[0, 1]でしたが、負数を含めた実数データでは範囲が[0, 2]になります。また、規定近接(RP)の範囲は非負データでは[0, 1]でしたが、実数データでの範囲は[-1, 1]です[34]。

係数	非負データ	実数データ
規定距離(RD)	[0, 1]	[0, 2]
規定近接(RP)	[0, 1]	[-1, 1]

次は h1 の Y を負(-19)にしたときの規定距離(RD)と規定近接(RP)を計算した結果です。

A.B	X:A	Y:B	X-Y	$(X-Y)^2$	X^2	Y^2	X^2+Y^2		
sh1	10	-19	29	841	100	361	461		
h2	11	7	4	16	121	49	170		
h3	0	0	0	0	0	0	0		
h4	0	1	-1	1	0	1	1	RD↓	RP↓
			和→	858		和→	632	1.358	-.358

$RD(x, y) = 858 / 632 = 1.358$
$RP(x, y) = 1 - RD(x, y) = 1 - 1.358 = -358$

[34] 実数データの規定距離係数の範囲は[0, 2]になり、「距離」が非負であることと矛盾しません。一方、実数データの規定近接係数の範囲は[-1, 1]になるので、規定近接係数が負数になることがあります。これは負数を含む実数データでは規定近接係数の分子 $\Sigma_i x_i y_i$ が負になる可能性があるためです。このように「近接」には正負の方向性があり、規定近接係数が 1 に近くなると近接度は増加しますが、規定近接係数が-1 に近くなると「逆方向の近接度」が増加します。これは 2 つの数値の絶対値は近くなるのですが、正負の符号が逆になる状態です。このことはちょうど相関係数の「逆相関」（逆向きの相関）と類似しています。

下左表はデータ行列(M)、下右表はその規定近接対称行列(RP)です。

M	A	B	C	D	E
h1	10	19	14	7	12
h2	11	7	10	0	1
h3	0	0	1	12	1
h4	0	1	2	3	3

RP	A	B	C	D	E
A	1.000	.845	.958	.331	.697
B	.845	1.000	.949	.444	.841
C	.958	.949	1.000	.461	.811
D	.331	.444	.461	1.000	.588
E	.697	.841	.811	.588	1.000

プログラム

```
function DisRgM(Xnp) { //規定距離対称行列 (Regular distance)
        var n = NR(Xnp), p = NC(Xnp), Dpp = NewMt(p,p); Dpp[0][0]="[N.Dist.]";
        for(var i = 1; i <= p; i++) {
                Dpp[0][i] = Dpp[i][0] = Xnp[0][i]; Dpp[i][i] = 0; //表頭；表側；対角
成分
        }
        for(var i = 1; i <= p-1; i++) { for(var j = i+1; j <= p; j++) { //距離行列
                var xy = x2 = y2 = 0;
                for(var k = 1; k <= n; k++) {
                        xy += Pow(Xnp[k][i] - Xnp[k][j], 2);
                        x2 += Pow(Xnp[k][i], 2);
                        y2 += Pow(Xnp[k][j], 2);
                } Dpp[i][j] = Dpp[j][i] = xy / (x2+y2);
        }} return Dpp;
}
```

●実数データの規定距離の最大値と規定近接の最小値

次の規定距離(RD)の式で

$$RD = \Sigma_i (x_i - y_i)^2 / \Sigma_i (x_i^2 + y_i^2)$$

一見したところ y が x の単なる負数(-x)ではなく、さらに小さな数、たとえば-x-a になる方が RD が大きくなるのではないかと思われるかもしれません。上の式の右辺の第 2 項の分子の y_i の絶対値を-10, -20, -100 のように，どんどん大きくすると、分子がどんどん大きくなるからです。しかし、このとき分母も大きくなるので、RD は 2 より大きくなりません。このことを次のような実験をして具体的に確かめましょう。

M	X	Y	X-Y	$(X-Y)^2$	X^2	Y^2	X^2+Y^2		
h1	10	-10	20	400	100	100	200		
h2	11	-11	22	484	121	121	242		
h3	2	-2	4	16	4	4	8		
h4	3	-3	6	36	9	9	18	RD↓	RP↓
			和 →	936	和 →	234	468	2.000	-1.000

ここで h4 の Y をさらに小さくして-10 にすると次のように、規定距離(RD)は小さくなります。

M	X	Y	X-Y	$(X-Y)^2$	X^2	Y^2	X^2+Y^2		
h1	10	-10	20	400	100	100	200		
h2	11	-11	22	484	121	121	242		
h3	2	-2	4	16	4	4	8		
h4	3	-10	13	169	9	100	109	RD↓	RP↓
			和 →	1069		和 →	559	1.912	-0.912

このように X と Y は正負（±）の符号がすべて逆になったとき、規定距離が最大の 2 になります。このことを一般化して数式の中で確かめます。ここでは簡単にするため規定距離(RD)の最大値ではなく、規定近接(RP)の最小値を考えます。両者は互いに 1 の補数になるので、先に見たようにどちらの条件も同じです。

$$RP = 2\,\Sigma_i\, x_i\, y_i \,/\, \Sigma_i\,(x_i^2 + y_i^2)$$

ここで $y_i = -x_i$ として、この$-x_i$の中の 1 つ$(-x_k)$でも$-x_k-a$になれば(a は任意の正数)、それと x_k の積は

$$x_k\,(-x_k - a) = -x_k^2 - x_k\, a$$

上の右辺の第 1 項はΣ_iに含まれるので、RP の分子は

$$\begin{aligned} \text{RP の分子} &= 2\,(\Sigma_i\,[x_i*(-x_i)] - x_k\, a) \\ &= -2\,(\Sigma_i\, x_i + x_k\, a) \end{aligned}$$

そして RP の分母は

$$\begin{aligned} \text{RP の分母} &= \Sigma_i\,(x_i^2 + y_i^2) \\ &= \Sigma_i\,[x_i^2 + (-x_k-a)^2] \\ &= \Sigma_i\,(x_i^2 + x_i^2 + 2\, x_k\, a + a^2) \\ &= 2\,\Sigma_i\, x_i + 2\, x_k\, a + a^2 \\ &= 2(\Sigma_i\, x_i + x_k\, a) + a^2 \end{aligned}$$

ここで

$c/(c+d) < 1,\ -c/(c+d) > -1$　…c と d は正数

よって $c = 2\Sigma_i x_i + x_k a,\ d = a^2$ とすると

$$RP(x, -x-a) = -2(\Sigma_i x_i + x_k a) / [2(\Sigma_i x_i + x_k a) + a^2] > -1$$

このように RP(x, -x-a)は-1 より大きな数になります。そして a=0 のときに RP=-1 になるので、最小値は $y_i = -x_i$ のときに-1 になります。

そして a が-a(任意の負数)であっても同様であることを確かめてみましょう。$-x_i$ の中の 1 つ($-x_k$)でも $-x_k-(-a)$になれば(a は任意の正数)、それと x_k の積は

$$x_k(-x_k + a) = -x_k^2 + x_k a$$

上の右辺の第 1 項は Σ_i に含まれるので、RP の分子は

$$RP\text{ の分子} = 2(\Sigma_i [x_i*(-x_i)] + x_k a) = -2(\Sigma_i x_i - x_k a)$$

また RP の分母は

$$\begin{aligned} RP\text{ の分母} &= \Sigma_i (x_i^2 + y_i^2) \\ &= \Sigma_i [x_i^2 + (-x_k+a)^2] \\ &= \Sigma_i (x_i^2 + x_i^2 - 2 x_k a + a^2) \\ &= 2\Sigma_i x_i - 2 x_k a + a^2 \\ &= 2(\Sigma_i x_i - x_k a) + a^2 \end{aligned}$$

ここで

$c/(c+d) < 1,\ -c/(c+d) > -1$　… $d > 0$

よって $c = 2\Sigma_i x_i - x_k a,\ d = a^2$ とすると

$$RP(x, -x+a) = -2(\Sigma_i x_i - x_k a) / [2(\Sigma_i x_i - x_k a) + a^2] > -1$$

このように RP(x, -x+a)は-1 より大きな数になります。

以上のことは、簡単にするために $x_1, x_2, \ldots, x_n$ のうち 1 個(x_k)だけに a を加えたり引いたりしましたが、すべての項に $a_1, a_2, \ldots, a_n$ を加えたり引いたりしても次のように同様に成り立ちます。

$$RP\text{ の分子} = 2(\Sigma_i [x_i*(-x_i)] + \Sigma_i x_i a_i)$$
$$RP\text{ の分母} = 2(\Sigma_i [x_i*(-x_i)] + \Sigma_i x_i a_i) + \Sigma_i a_i^2$$

ここで

$c/(c+d) < 1,\ -c/(c+d) > -1 \quad \ldots d > 0$

よって $c = 2(\Sigma_i [x_i*(-x_i)] + \Sigma_i x_i a_i),\ d = \Sigma_i a_i^2$ とすると

$$RP(x, -x+a) = -2(\Sigma_i [x_i*(-x_i)] + \Sigma_i x_i a_i) / [2(\Sigma_i [x_i*(-x_i)] + \Sigma_i x_i a_i) + \Sigma_i a_i^2] > -1$$

●規定距離の最大値の厳密な求め方

以上では、説明を簡単にするために規定距離(RD)の最大値を中学数学の範囲内の知識と、極端なケースで「…になるはず」というような私たちの直感を使って求めました。そこで以下では、その最大値とその条件を高校数学の範囲（ベクトルの内積と関数の商の微分）を使ったさらに厳密な方法で求めます。

先の式(5)、規定近接(RP)

$$RP(x, y) = 2\Sigma_i x_i y_i / \Sigma_i (x_i^2 + y_i^2) \qquad \ldots(5)$$

を、ベクトルの積とノルムを使って表現すると[35]

$$RP(x, y) = 2\Sigma_i x_i y_i / (\Sigma_i x_i^2 + \Sigma_i y_i^2) = 2X^T Y / (\|X\|^2 + \|Y\|^2)$$

$$= 2\|X\|\|Y\| / (\|X\|^2 + \|Y\|^2)\cos\theta \quad \leftarrow X^T Y = \|X\|\|Y\|\cos\theta$$ 注[36]

上の式の一部（網掛けの部分）

$$\|X\|\|Y\| / (\|X\|^2 + \|Y\|^2) \quad \ldots(5)$$

の最大値を求めます。上の式の変数が $\|X\|, \|Y\|$ という 2 個になるので 2 つの変数の比を 1 つの変数 c で表し、c だけを使って(5)を表現します。

$$c = \|X\|^2 / \|Y\|^2$$

式(5)の分子と分母をそれぞれ $\|Y\|^2$ で割ると

(5)の分子：$\|X\|\|Y\| / \|Y\|^2 = \|X\| / \|Y\| = c^{1/2}$
(5)の分母：$(\|X\|^2 + \|Y\|^2) / \|Y\|^2 = \|X\|^2 / \|Y\|^2 + 1 = c + 1$

よって式(5)を c の関数で示すと

式(5)：$f(c) = c^{1/2} / (c + 1)$

---

[35] X の「ノルム」$\|X\|$は次のようにして定義されます。

$$\|X\| = (x_1^2 + x_2^2 + \ldots, + x_n^2)^{1/2}$$

[36] 参照：ベクトルの内積(高校数学 B)←余弦定理(高校数学 I)。

次のような分子と分母が関数である式 f(c) = g(c) / h(c)の微分 f'(c)は[37]

$$f'(c) = [g'(c)h((c) - g(c)h'(c)] / h^2(c)$$

これより

$$\begin{aligned} f'(c) &= 1 / (c+1)^2 \, [1/2 \, c^{-(1/2)} \, (c+1) - c^{1/2} \, 1] \\ &= 1 / (c+1)^2 \, 1/2 \, c^{-(1/2)} \, [(c+1) - 2c] \\ &= 1 / (c+1)^2 \, 1/2 \, c^{-(1/2)} \, (1 - c) \\ &= 1/2 \, c^{-(1/2)} \, (1 - c) / (c+1)^2 \\ &= c^{-(1/2)} \, (1 - c) / [2(c+1)^2] \end{aligned}$$

よって f'(c) は c = 1 のときに 0 になります。増減表は

c	0	…	1	…
f'(c)	.	+	0	−
f(c)	0	↑	1/2	↓

この増減表から c=1 のときの f(c)の接線の傾きが 0 になり、そのときの f(c)=1/2 が最大値であることがわかります。

よって

---

[37] 参照：商の導関数(高校数学 III)。次の微分の定義を前提として確認します。

$$f'(x) = \lim_{[h \to 0]} [f(x+h) - f(x)] / h$$

はじめに関数の積を微分します。

(1) $[f(x)g(x)]' = \lim[h \to 0] [f(x+h)g(x+h) - f(x)g(x)] / h$ ←微分の定義

$= \lim_{[h \to 0]} \{[f(x+h)-f(x)]g(x+h) + f(x)[g(x+h)-g(x)]\} / h$
←f(x)g(x+h)を媒介

$= \lim_{[h \to 0]} [f(x+h)-f(x)]g(x+h)/h + \lim[h \to 0] f(x)[g(x+h)-g(x)]/h$
←lim[h→0]を分配

$= f'(x)g(x) + f(x)g'(x)$ …(1) ←微分の定義

次に関数の商を微分します。

(2) $[1/g(x)]' = \lim[h \to 0] [1/g(x+h) - 1/g(x)] / h$ ←微分の定義

$= \lim_{[h \to 0]} \{[g(x) - g(x+h)] / [g(x+h)g(x)]\} / h$ ←通分

$= \lim_{[h \to 0]} \{-[g(x+h) - g(x)] / h * 1 / [g(x+h)g(x)]\}$
第 1 項の分子と第 2 項の分母を整理

$= -g'(x) / [g(x)g(x)] = -g'(x) / g^2(x)$ …(2)

(1), (2)より

$f'(x) = [g(x) / h(x)]' = [g(x) * 1 / h(x)]'$ ←分母を分離

$= g'(x) * 1/h(x) + g(x)*[1/h(x)]'$ ←(1)

$= g'(x) / h(x) + g(x) * -h'(x) / h^2(x)$ ←(2)

$= [g'(x)h(x) - g(x)h'(x)] / h^2(x)$ ←第 2 項の負号を前に

式(5)： $\|X\|\|Y\| / (\|X\|^2+\|Y\|^2)$

は $c = \|X\|^2 / \|Y\|^2 = 1$、つまり $\|X\|^2 = \|Y\|^2$ 、よって $\|X\| = \|Y\|$ のとき最大値 1/2 に達します。よって式(3')の最大値は

式(3')： $2\Sigma_i x_i y_i / (\Sigma_i x_i^2 + \Sigma_i y_i^2)$

$= 2 \|X\|\|Y\| / (\|X\|^2+\|Y\|^2) \cos\theta = 2 * 1/2 * \cos\theta = \cos\theta$

規定近接(RP)が最大であるのは $\|X\| = \|Y\|$ のときなので、次の

$X^T Y = \|X\|\|Y\| \cos\theta$

より

$X^T X = \|X\|\|X\| \cos\theta$
$X^T X = X^T X \cos\theta$

から $\cos\theta=1$ となることは明らかです。以上から規定近接(RP)の範囲が[0, 1]であることがわかりました。このことから、規定近接(RP)の 1 の補数である規定距離(RD)の範囲も[0, 1]になります。このようにして実験的にも、直感的にも、そして数理的にも範囲が必ず[0, 1]になるので、両者とも「規定」(Regular)をつけてよぶことにしました。

なお、上の $X^T X = \|X\|\|X\| \cos\theta$ において X の一方が負であれば $\cos\theta= -1$ になるので、規定距離(RD)の範囲は[-1, 1]になります[38]。

＊この数学的証明では東京大学大学院総合文化研究科・情報学環の倉田博史先生の全面的な支援をいただきました。また先生からご教示いただいた次の式が規定距離・規定近接を考える貴重な出発点になりました。

$\|X-Y\|^2 = \|X\|^2 + \|Y\|^2 - 2 X^T Y$

先生の許可を得て、謝意をこめてここに掲載しました。(2017/2/18)

●平均距離と規定距離の比較

次のように平均距離(MD)と規定距離(RD)の式は似ています。

平均距離： $MD(x, y) = 1/n \ \Sigma_i [(x_i - y_i)^2 / (x_i^2 + y_i^2)]$
規定距離： $RD(x, y) = \Sigma_i (x_i - y_i)^2 / (\Sigma_i x_i^2 + \Sigma_i y_i^2)$

平均距離(MD)の計算では毎回 x と y の差の 2 乗を $(x_i^2 + y_i^2)$ で割った値を足し上げ、その平均を求めた値であり、一方、規定距離(RD)は x と y の差の 2 乗を全部足したものを、x の 2 乗和＋y の 2 乗和で割った値です。この

[38] $X^T -X = \|X\|\|X\| \cos\theta \ \rightarrow \ \cos\theta= -1.$

違いを念頭に置いて具体例を見ながら両者の特徴を比較しましょう。

M	X	Y	X-Y	$(X-Y)^2$	X^2	Y^2	X^2+Y^2	MD
h1	2	3	-1	1	4	9	13	0.077
h2	20	30	-10	100	400	900	1300	0.077
h3	200	300	-100	10000	40000	90000	130000	0.077
h4	225	301	-76	5776	50625	90601	141226	RD 0.041
			和 →	15877		和 →	272539	0.942 0.058

上の平均距離 MD(=.058)の計算では、h1, h2, h3, h4 はそれぞれの距離を相対化して計算し、その平均を求めています。一方、規定距離 RD(=.942)の計算では、小さな値の h1(2, 3)はほとんど最終的な RD の計算に影響を与えていません。RD の値に大きな影響を与えているのは h3, h4 です。

よって、平均距離(MD)はデータの規模に影響されません。一方、規定距離(RD)はデータの規模を感知します(sensitive)。そこで、たとえば X, Y が同じ分量の文書であり、h1, h2 が頻度が低い名詞と動詞、h3 が高頻度の前置詞、h4 が高頻度の接続詞のそれぞれの出現数を示しているとすれば、X と Y の距離を相対化して計算するためには平均距離(MD)を使わなければなりません。規定距離(RD)を使うと、その距離の計算には名詞と動詞の頻度はあまり関与せず、そのほとんどが前置詞・接続詞の頻度によって決定されてしまうからです。一方、h1, h2, h3, h4 がすべて同質の名詞であれば、むしろその頻度そのものを感知する規定距離(RD)を使う可能性も考えられるでしょう。このとき h1, h2, h3, h4 のそれぞれの固有の頻度が重要な情報をもつと考えられるためです[39]。

ここでは平均距離(MD)と規定距離(RD)を比較しましたが、平均近接(MP)と規定近接(RP)を比較しても同じです。

6.5. 差

データ行列の列ごとの平均値や分散などの要約値を差を求めます。そのような差が偶然で起こる程度も確率によって考慮します。

6.5.1. 平均値差

下右表は下左表の列限定得点です。

[39] しかし、たとえ同じ名詞であっても、その中にとくに高頻度の語があれば、これが規定距離係数をほとんど決定してしまいます。そのような場合には平均距離係数も参照すべきです。

D	v1	v2	v3
d1	45	48	66
d2	56	59	54
d3	58	51	78
d4	77	72	20
d5	43	44	32
d6	58	34	90
d7	50	53	100

L	v1	v2	v3
d1	.059	.368	.575
d2	.382	.658	.425
d3	.441	.447	.725
d4	1.000	1.000	.000
d5	.000	.263	.150
d6	.441	.000	.875
d7	.206	.500	1.000

この限定得点の列の平均値は

Ap	v1	v2	v3
平均値	.361	.462	.536

次の表は、それぞれの列の平均値の差を示したものです。距離行列と同じように、このような「平均差行列」(Average Difference Matrix: ADM)では近い関係のものは数値が小さくなります。

M.Df. 0	v1	v2	v3
v1	.000	-.101	-.174
v2	.101	.000	-.073
v3	.174	.073	.000

平均差行列(Rpp)を次のような行列演算で導出します。

$$Ap = AvV(LmV(Xnp))$$
$$ADM = S(Ap^T\ Ap)$$

ここで、Ap は入力行列(Xnp)の列の限定値の平均ベクトル（横ベクトル）です。これを「行列」で定義したように、行列要素の差(S)をベクトルから一様行列の変換を含む演算で出力します。次のように差の行列の大きさは p 行 p 列になります。

$$ADM_{pp} = Ap_{p1} - Ap_{1p}$$

6.5.2. 中央値差

次は、列ごとの中央値の差と、その確率を求めた結果です。

D	v1	v2	v3
d1	45	48	66
d2	56	59	54
d3	58	51	78
d4	77	72	20
d5	43	44	32
d6	58	34	90
d7	50	53	100

MdD.N	v1	v2	v3
v1	.0000	-.0650: -.609^	-.1926: -.794^
v2	-.0650: -.609^	.0000	-.1276: -.707^
v3	-.1926: -.794^	-.1276: -.707^	.0000

MdD.R	v1	v2	v3
v1	.0000	-.0650: -.600^	-.1926: -.788^
v2	-.0650: -.600^	.0000	-.1276: -.702^
v3	-.1926: -.788^	-.1276: -.702^	.0000

6.5.3. 分散値差

同様に「分散差行列」(Variance Difference Matrix: VDM)それぞれの列の分散の差を示します。

V.Df. 0	v1	v2	v3
v1	.000	.011	-.021
v2	-.011	.000	-.032
v3	.021	.032	.000

$$Vp = VrV[LmV(Xnp)]$$
$$VDM = S(Vp^{T}\ Vp)$$

6.5.4. 標準偏差値差

次は、列ごとの標準偏差値の差と、その確率を求めた結果です。

D	v1	v2	v3
d1	45	48	66
d2	56	59	54
d3	58	51	78
d4	77	72	20
d5	43	44	32
d6	58	34	90
d7	50	53	100

SdD.N	v1	v2	v3
v1	.0000	.0181: .594^	-.0327: -.667^
v2	.0181: .594^	.0000	-.0508: -.748^
v3	-.0327: -.667^	-.0508: -.748^	.0000

SdD.R	v1	v2	v3
v1	.0000	.0181: .598^	-.0327: -.654^
v2	.0181: .598^	.0000	-.0508: -.744^

v3	-.0327: -.654^	-.0508: -.744^	.0000

6.5.5. ジニ係数値差

次は、列ごとのジニ係数値の差と、その確率を求めた結果です。

D	v1	v2	v3
d1	45	48	66
d2	56	59	54
d3	58	51	78
d4	77	72	20
d5	43	44	32
d6	58	34	90
d7	50	53	100

GiniD.N	v1	v2	v3
v1	.0000	.1057: .796^	.0899: .759^
v2	.1057: .796^	.0000	-.0158: -.549^
v3	.0899: .759^	-.0158: -.549^	.0000

GiniD.R	v1	v2	v3
v1	.0000	.1057: .800^	.0899: .765^
v2	.1057: .800^	.0000	-.0158: -.552^
v3	.0899: .765^	-.0158: -.552^	.0000

6.5.6. エントロピー差

次は、列ごとのエントロピーの差と、その確率を求めた結果です。

D	v1	v2	v3
d1	45	48	66
d2	56	59	54
d3	58	51	78
d4	77	72	20
d5	43	44	32
d6	58	34	90
d7	50	53	100

N.EntropyD.N	v1	v2	v3
v1	.0000	.0755: .866^	.0614: .817^
v2	.0755: .866^	.0000	-.0140: -.582^
v3	.0614: .817^	-.0140: -.582^	.0000

N.EntropyD.R	v1	v2	v3
v1	.0000	.0755: .855^	.0614: .814^
v2	.0755: .855^	.0000	-.0140: -.580^
v3	.0614: .814^	-.0140: -.580^	.0000