5 データの観察

【目標】「データ」を数値化し、客観的な観察をする。統計的な数値の意味を理解する。

現象を見て単なる印象を述べる場合と、その具体的な頻度を提示する場合とでは、その説得力に大きな違いがあります。また、頻度だけではわかりにくいデータも、頻度数をさまざまな変換をすることによって特徴が明示化されます。この章では、はじめに実測値の和、平均、中央値、中間値、標準偏差を確認し、次にそれらを使って得点分布を変換し分布の特徴を明示化する方法を説明します。

5. 1 データの代表値

次の表は 4 つのスペイン語の単語（場所を示す副詞: *acá, allá, allí, aquí*)と地域ごとのテキストの頻度を示しています。

鍵語	1 Madrid	2 Sevilla	3 México	4 Lima	5 B. A.
acá	0	0	0	11	1
allá	0	1	2	3	3
allí	11	7	10	0	1
aquí	10	19	14	7	12

「鍵語」はテキストを検索したときのキーワードを示します。それぞれの語の分布の特徴をつかむには、和、平均、中央値、中間値、標準偏差などを知っておくと参考になります。これをデータの「代表値」(representative value)を呼ぶことにします。これらの値を示したものが次の図です。それぞれを具体的に見ていきましょう。以下では、それぞれの鍵語を w1, w2, w3, w4 とし、それぞれの地点を L1, L2, …, L5 とします。

◇1 実測値

ここで用いるデータは、Madrid, Sevilla, México, Lima, Buenos Aires (B. A.) における w1 ~ w4（場所を示す副詞: *acá, allá, allí, aquí*)の使用について集計したデータです。何も加工していない生のデータなので「実測値」(実測得点：observed score: o.s.)と呼ぶことができるでしょう。これが出発点です。

これは頻度数(frequency)を示すのでゼロを含む自然数(0, 1, 2, ...)です。数値だけでは分布の様子がわかりにくいので Excel シートの「データバー」を使って視覚化しましょう。◆該当するセルを選択→「ホーム(H)」→「スタイル」グループ→「条件付き書式」→「データバー」
標準のデータバーは色が濃くて、数値が見えにくいことがあります。色を薄くするには、◆該当するセルを選択し、「ホーム(H)」→「スタイル」グループ→「条件付き書式」→「データバー」→「その他のルール」で調整してください。

◇2 和

「和」(Sum: Sm)には横行の値を全部足した「横和」(Sum in row.Sm.r)と、縦列の値を全部足した「縦和」(Sum in column: Sm.c)と、全部の値を全部足した「総和」(Sum in all: Sm.a)があります。w1 の横和 Sm.r は 0 + 0 + 0 + 11 + 1 = 12 です。Madrid の縦和 Sm.c は 0 + 0 + 11 + 10 = 21 です。そして、全部の総和 Sm.a は 0 + 0 + ... + 7 + 12 = 112 になります。Excel 関数 SUM を使ってこれらを算出します。

◇3 平均

「平均」は、データの数値を合計し（和）、データの個数(Count: Cn)で割った値です。これは一般に「算術平均」(Arithmetric Average)と呼ばれています。以下では、特別の場合を除いて「平均」で「算術平均」を示すことにします。

行の個数を Cn.r とし、縦列の個数を Cn.c、全体の個数を Cn.a とすると、横行の平均(Average in row: Av.r)は Sm.r / Cn.r, 縦列の平均(Av.c)は Sm.c / Cn.c、総平均 Av.a は Sm.a / Cn.a になります。Excel 関数 AVERAGE を使います。

速度、濃度、平均、比率など、割り算を使って算出された値の平均は、そのまま合計して個数で割るとうまくいきません。このようなとき、一般に統計学では「調和平均」(Harmonic Average)が使われています[1]。たとえば、ハイキングで一定の行程を往復し、往路は時速 6 km/h、復路は時速 4 km/h だったとします。このとき往復の平均時速を算術平均で出すと(6 + 4) / 2 = 5 になるからといって、平均時速を 5(km/h)としたのでは、不都合なこ

[1] たとえば池田(1976: 40-41)

とが起こります。往復の距離を平均時速で割っても、時間が正しく出てこないのです。たとえば片道 12km だとすると、24 (km) / 5 (km/h) = 4.8(h)になってしまいますが、実際の往路は 12 (km) / 6 (km/h) = 2(h)であり、復路は 12 (km) / 4 (km/h) = 3(h)で、往路と復路を併せて 5 (h)になります。

そこで、次のような調和平均(H.Av.)が使われます。片道の距離を a(km)とすると、a(km) / 6(km/h)が往路の時間になります。同様に、復路の時間は a(km) / 4(km/h)です。往路と復路の平均時間(Av.h)は

Av.h. = (a / 6 + a / 4) / 2
= [(1 / 6 + 1 / 4) / 2] a
= [(2 / 12 + 3 / 12) / 2] a
= [5 / (12 * 2)] a
= (5 / 24) a
= (1 / 4.8) a

この第 2 式と最後の式を取り出すと、

[(1 / 6 + 1 / 4) / 2] a = (1 / 4.8) a
(1 / 6 + 1 / 4) / 2 = 1 / 4.8
1 / [(1 / 6 + 1 / 4) / 2] = 4.8

調和平均 H.av.を一般式で書くと次のようになります[2]。

H.Av.(x, y) = 1 / [(1 / x + 1 / y) / 2]

この調和平均は次の「分数平均」(F.Av.: Fractional Average)の特殊なケースです（分母が同数：→コラム）。分母が異なるときは、次の分数平均(F.Av)を使うことを提案します。

F. Av. (a, b, c, d) = (a + c) / (b + d)

ここで、a, b, c, d は 2 つの分数 x, y の分子(a, c)と分母(b, d)を示します。

x = a / b, y = c /d

ここでは、

[2] ここでは 2 つの値の調和平均を説明しましたが、2 個以上でも同様です。H.av. = 1 / {[Σ (1 / x_i)] / n}, ここで x_i はそれぞれの値を示し、n は x_i の個数を示します。

$x = 12 / 2,\ y = 12 / 3$

となり、それぞれ往路と復路の時速(km/h)を示します。両方の時速の平均を分数平均(F.Av.)を使って示すと、

$$F.Av.\ (12, 2, 12, 3) = (12 + 12) / (2 + 3) = 24 / 5 = 4.8$$

調和平均の算出は複雑で、一見では解釈が難しいのですが、分数平均ならばとても簡単です。直感的に理解できるので説明もしやすいと思います。

比の平均としての「分数平均」

たとえば 1/4 と 2/5 というような 2 つの比率 r_1 と r_2 の平均をとるときは、「算術平均」(A.Av.: Arithmetric Average)、幾何平均(G.Av.: Geometric Average)、調和平均(H.av: Harmonic Average)を使うことが考えられます。

$$A.Av. = (r_1 + r_2) / 2$$
$$G.Av. = \sqrt{(r_1\ r_2)}$$
$$H.Av. = 1 / [(1 / r_1 + 1 / r_2) / 2]$$

一方、比率 r_1 と r_2 のそれぞれの分子(a_1 , b_1)と分母(a_2 , b_2)がわかっているときは($r_1 = a_1 / b_1,\ r_2 = a_2 / b_2$)、$r_1$ と r_2 の分子の和($a_1 + b_2$)を分子とし、r_1 と r_2 の分母($b_1 + b_2$)の和を分母とした分数を使うことも考えられます。これを「分数平均」(F.Av: Fractional Average)と呼ぶことにします。

$$F.Av. = (a_1 + b_2) / (b_1 + b_2)$$

それぞれの平均の結果は類似することがありますが、分数（比率）を扱うとき、分数平均は 2 つの分数の元の数に遡って計算するので、他の平均より正確です。また、結果の解釈もわかりやすいと思います。ちょうど濃度と量の異なる食塩水を混ぜ合わせて出来上がった食塩水の濃度のようなものだからです。たとえば 1/4 と 2/5 という比率の平均は簡単な算術平均(A.Av.)ならば、

$$A.Av. = (1/4 + 2/5) / 2 = 0.325,$$

幾何平均(G.Av.)ならば

$$G.Av. = \sqrt{(\ 1 \times 2 / 4 \times 5)} \fallingdotseq 0.316$$

調和平均(H.av.)ならば、

H.Av. = 1 / [(4 / 1 + 5 / 2) / 2] ≒ 0.308

になります。どちらも分子と分母の大きさに関わりなく一義的に計算されます。ここで提案した分数平均(F.M.)を使うと、次のように計算されます。

F.Av. = (1 + 2) / (4 + 5) ≒ 0.333

10/40 と 4/10 のそれぞれの平均を比べてみましょう。

平均	**1/4, 2/5**	**10/40, 4/10**
A.Av.	0.325	0.325
G.Av.	0.316	0.316
H.Av.	0.308	0.308
F.Av.	0.333	0.280

このように、他の平均と比べて、分数平均では第一項の分子と分母を大きくすると、全体的に薄まって数値が下降していることがわかります。

次の図は、調和平均の説明によく使われる往復（ハイキングなど）の平均速度の計算を示すものです。この図が示すように、距離と時間のそれぞれの和から速度を計算すると、調和平均と分数平均は正しい平均値を出します。

同距離	昨日	今日	和	算術平均	調和平均	分数平均
距離(km)	12	12	24			
時間(h)	2	3	5			
速度(km/h)	6	4	4.80	5.00	4.80	4.80

しかし、往復ではなく、今日は昨日の道を引き返すのではなく、さらに先に進むような場合、次のように両日の距離が異なるのがふつうです。

異距離	昨日	今日	和	算術平均	調和平均	分数平均
距離(km)	12	15	27			
時間(h)	2	3	5			
速度(km/h)	6	5	5.40	5.50	5.45	5.40

このとき、調和平均は距離と時間の和から算定される速度を正しく示してはいません。分数平均は、そのまま距離と時間の和から算定されるので、直感的に理解できると思います。

「分数平均」は、分子の値の和を分母の値の和で割る、という簡単な操作です。2 つの値だけでなく、次のように n 個のデータでも、同じ計算方

法を使うことができます。

$$F.Av. = \Sigma x_i / \Sigma y_i$$

ここで x_i はそれぞれの分子の値、y_i はそれぞれの分母の値を示します。そうすると、y_i がすべて 1 であるときが算術平均になることがわかります。

$$F.Av. = \Sigma x_i / \Sigma 1 = (\Sigma x_i) / n$$

ここで、n はデータの個数を示します。このように算術平均は分数平均の特殊なケースだと考えてもよいでしょう。

たとえば、保険会社が自動車事故保険の掛け金を設定するときに、さまざまな運転手の事故率のデータを勘案することでしょう。そのとき、運転手の集団の事故率を単純に算術平均や調和平均で計算するよりも、運転手の月平均運転時間も考慮に入れないと、正しい全体の事故率の平均が出ないはずです。

（溶液の）濃度、（歩行）速度、（自動車事故）率など、一般に割り算を使って得られた値については、元のデータがあるならば、それに戻って計算すべきだと思います。元のデータがないときは、算出された平均値を慎重に扱わなければなりません。

(1st vers. 2013/3/21; last vers. 2013/5/15)

◇4 中央値

データを大→小になるように並び替えて、その中央にある値が「中央値」(Median: Md)です。中央が偶数になるときは両者の平均をとります。横軸、縦軸、全体の中央値をそれぞれ Md.r, Md.c, Md.a とします。Excel 関数 MEDIAN を使います。

◇5 中間値

ここではデータの最大値と最小値の平均を「中間値」(Central: Ct)と呼ぶことにします[3]。Excel 関数 MAX（最大値）と MIN（最小値）を組み合わせて、

[3] 「中間値」という用語はあまり使われていません。東京大学教養学部統計学教室『統計学入門』（東京大学出版会）(1991:p.34)では「ミッド・レンジ」(mid-range)という用語で説明されています。平均値が mean、中央値が median、中間値が mid-range に対応します。ここでは、イニシャルで区別できるように、それぞれを Average, Median, Central という用語を使うこ

= (MAX(*) + MIN(*)) / 2 という式にします。

◇6 標準偏差

データの平均が同じでも、ばらつきが大きい場合と小さい場合があります。たとえば、{1, 4, 7}というデータ群と、{3, 4, 5}というデータ群では平均はどちらも 4 ですが、それぞれの中身を見るとデータのばらつきが異なります。とくにばらつきが大きいときには、データの扱い方や平均の解釈が代わりますから注意が必要です。

標準偏差の求め方を見ましょう。はじめにそれぞれの値(x)から全体の平均(m)を引き、それを 2 乗したものを全部足して、全体の個数(n)で割ります。これを「分散」(variance: V)と呼びます。分散を求める式は次のようになります[4]。

$$\text{分散(V)} = \frac{(x_1-m)^2 + (x_2-m)^2 \ldots + (x_n-m)^2}{n}$$

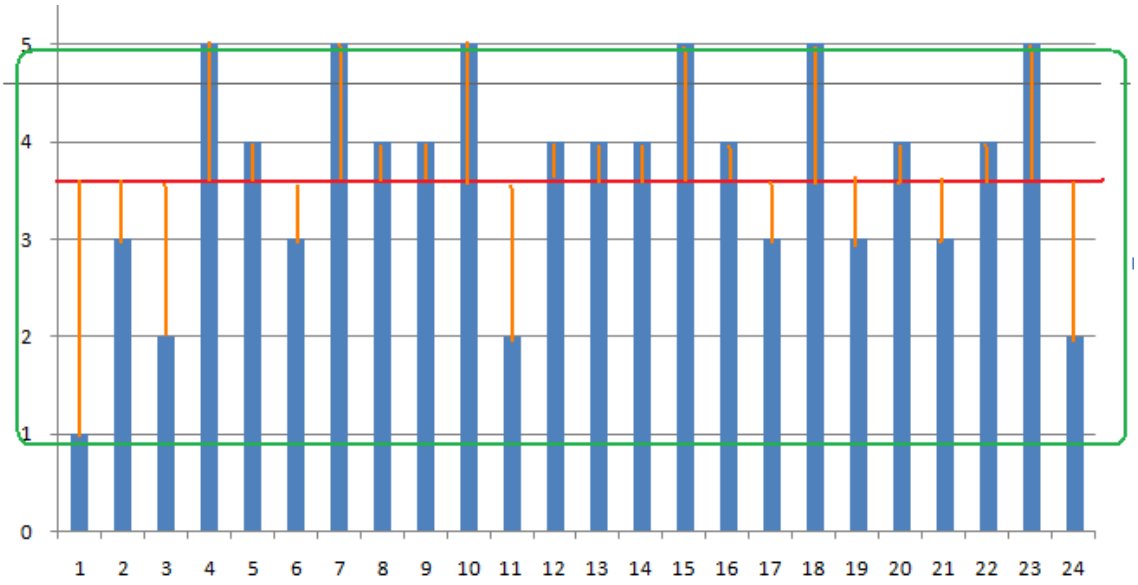
ここで、$x_1, x_2, \ldots x_n$ は個々のデータの値、m はデータの平均、n はデータの個数を示します。分散は、総体としてデータがどれだけ平均から離れて分散しているかを示します[5]。

たとえば次のようなグラフで考えてみましょう。青い縦棒がデータの値（頻度）、赤い横線が平均、オレンジの線が平均からの距離を示します。この距離を全部集めていくわけですが、そのまま足してしまうと、その和はどのようなデータでもゼロになってしまうので、2 乗して足していきます。分散を求めるときは Excel 関数 VARP を使います。

とにします。

[4] 分散の式を展開すると、次のようにまとめられます。分散= $(x1^2 - 2mx1 + m^2 + x2^2 - 2mx2 + m^2 + \ldots + xn^2 - 2mxn + m^2) / n = (x1^2 + x2^2 + \ldots + xn^2) - 2m(x1 + x2 + \ldots + xn) + nm^2) / n = (x1^2 + x2^2 + \ldots + xn^2) - 2nm^2 + nm^2) / n = (x1^2 + x2^2 + \ldots + xn^2) - nm^2) / n = (x1^2 + x2^2 + \ldots + xn^2) / n - m^2$. よって、分散＝2 乗の平均―平均の 2 乗、ということになります。

[5] データの「ばらつき」を見るために平均からの差だけを足していくと、どのようなデータでも和はゼロになってしまいます。それでは、平均からの差の絶対値を足していけばよいのかも知れません。しかし、絶対値は数学的に扱いがやっかいです。計算過程において絶対値は元の数の正負によって場合分けをしなくてはならないからです。それにくらべて、平方和は扱いやすく応用範囲が広くなります。これから見ていく「標準偏差」「標準得点」「相関係数」などの計算で「分散」を使います



さて、分散には「2 乗する」という操作が入っているため平均からの距離が誇張されています。つまり、距離という線分ではなく、むしろ正方形の面積になっているのです。そこで、2 乗和をもとのデータのスケールに戻すために分散の根を求めます。これが「標準偏差」(standard deviation: Sd)です。

$$\text{標準偏差 (Sd)} = \sqrt{\text{分散}}$$

横軸、縦軸、全体の標準偏差をそれぞれ Sd.r, Sd.c, Sd.a とします。標準偏差を求めるときは Excel 関数 STDEVP を使います。

◇7 変動係数

標準偏差は個々のデータの規模（平均）が大きくなると、それに応じて大きくなる性質があります。そこで、こうした規模の違いを超えて比較できるように標準偏差を平均で割った値が「変動係数」(coefficient of variation: Cv)です[6]。標準偏差も平均もデータの規模を反映していますから、標準偏差を平均で割った変動係数はデータの規模に左右されることなく、だいたいのばらつき具合がわかります。変動係数を示すには Excel 関数を組み合わせて=STDEVP(*)/AVERAGE(*)という式を使います。

◇8 正規標準偏差

標準偏差を[0.0 ～ 1.0]の範囲をもつ値にしたものを「正規標準偏差」

[6] 芝祐順他『統計用語辞典』（新曜社）

(Normalized Standard Deviation: N.S.D.)と呼ぶことにします。正規標準偏差は標準偏差(S.D.)をその最大値 S.D. (max)で割ることで求められます[7]。

$$\text{N.S.D.} = \text{S.D.} / \text{S.D. (max)}$$

先に見たように標準偏差 S.D.は次のように定義されます。

$$\text{S.D.} = \sqrt{\{[(x_1 - m)^2 + (x_2 - m)^2 + \ldots + (x_n - m)^2] / n\}}$$

ここで、たとえば{10, 0, 0, 0, 0}というような 1 つだけに数値があるデータを考えましょう。このようなときが変動係数が最大値になるときです。ここで一般化して{a, 0, 0, …, 0}という n 個のデータを考えます。そうすると、上の式の分子の第 1 項だけが$(a - m)^2$ になり、残り n - 1 個はどれも$(0 - m)^2 = m^2$になります。よって変動係数の最大値は、

$$\text{S.D. (max)} = \sqrt{\{[(a - m)^2 + m^2(n - 1)] / n\}}$$

このとき a 以外にデータがないので a が総和になります。よって、a = 和 = nm という関係がわかります。

$$= \sqrt{\{[(nm - m)^2 + m^2(n - 1)] / n\}}$$
$$= \sqrt{\{[(m(n - 1))^2 + m^2(n - 1)] / n\}}$$
$$= \sqrt{\{[m^2(n - 1)^2 + m^2(n - 1)] / n\}}$$
$$= \sqrt{\{m^2[(n - 1)^2 + (n - 1)] / n\}}$$
$$= \sqrt{\{m^2[(n^2 - 2n + 1) + (n - 1)] / n\}}$$
$$= \sqrt{\{m^2[(n^2 - n)] / n\}} = \sqrt{[m^2 n(n - 1) / n]} = \sqrt{[m^2(n - 1)]}$$
$$= m\sqrt{(n - 1)}$$

よって、正規標準偏差(N.S.D.)は

$$\text{N.S.D.} = \text{S.D.} / \text{S.D.(max)} = \text{S.D.} / [m\sqrt{(n - 1)}]$$

正規標準偏差(N.S.D.)と変動係数(C.V.)の違いは、正規標準偏差の分母に$\sqrt{(n - 1)}$が加えられていることです。データ行列は一般に n の数値が大きいので、正規標準偏差は小さくなります。そのような場合は正規標準偏差はむしろ変数（比較的少数）の変動を見るときに使うべきです。

これまで扱ってきたさまざまな値（代表値）を先のデータで計算しまし

[7] この正規化の方法は以下でもしばしば使います。

よう。

実測値	1 Madrid	2 Sevilla	3 México	4 Lima	5 B. A.	和 Sm.r.	個数Cn.r.	平均 Av.r.	標準偏差 Sd.r.	変動係数 C.V.r.	N.S.D.r.
w1	0	0	0	11	1	12	5	2.40	4.32	1.80	0.90
w2	0	1	2	3	3	9	5	1.80	1.17	0.65	0.32
w3	11	7	10	0	1	29	5	5.80	4.53	0.78	0.39
w4	10	19	14	7	12	62	5	12.40	4.03	0.32	0.16
和 Sm.c.	21	27	26	21	17	112					
個数Cn.c.	4	4	4	4	4	Sm.a	20				
平均 Av.c	5.25	6.75	6.50	5.25	4.25		Cn.a	5.60			
標準偏差 Sd.c.	5.26	7.56	5.72	4.15	4.55			Av.a	5.65		
変動係数 C.V.c.	1.00	1.12	0.88	0.79	1.07				Sd.a	1.01	
正規標準偏差 N.S.D.c.	0.58	0.65	0.51	0.46	0.62					C.V.a	0.23
											N.S.D.a

語の使用度

5 つの分野（演劇、小説、随筆、科学技術文、報道文）の言語資料で使われるスペイン語単語の頻度辞典を作成した A. Juilland and E. Chang - Rodríguez (*Frequency dictionary of Spanish words,* The Hague: Mouton, 1964) は単語の使用度(Usage)を示す数値として、

$$U = F \times D$$

という式を提案しました。ここで F は単語の頻度(frequency)を示し、D は分野間の拡散度(dispersion)を示します。つまり、単語の使用度を見るためには、頻度(F)だけでなく、各分野に均等に使用されている度合(D)も勘案すべきだという主張です。

提示されている拡散度の式、

$$D = 1 - \sigma / 2m$$

この σ は標準偏差を示します。分母にある 2 は、√(分野数 5 – 1)のことだと思います。よって D = 1 – 正規変動係数という関係になります。

5. 2 データの得点

前節ではデータ全体の特徴を要約する統計量を見ました。ここでは、データを構成する個々のデータの「得点」(score)に着目し、データそれぞれの特徴を様々な数値を使って全体と比べながら観察します。以下で扱う「得点」の中には「度数」という用語を使って「相対度数」「期待度数」のように一般によく使われるものもありますが、「加重得点」「限定得点」「代表得点」「卓立得点」は一般に使われていません。「標準得点」は「標準スコア」「標準測度」などと呼ばれていますが、ここではデータの個々の数値をすべて「得点」という用語で統一しました。

◇1 相対得点

先に見た実測値の問題点は、横軸と縦軸ごとにスケールが異なるため、比較が難しいということです。たとえば、w1 の 11 と w4 の 10 をそのまま比較することができません。それぞれの和と平均が異なるからです。そこで有効になるのが「相対得点」(Relative Score: R.S.)（割合）です。それぞれの得点 x を和 Sm で割ることで算出できます x = 0 のとき R.S.の最小値は 0 で、x = Sm のとき最大値 1 になります[8]。

R.S. = x / Sm

R.S. [0.0 (x = 0) ≦ 0.5 (x = Sm/2)≦ 1.0 (x = Sm)]

このようにデータの範囲を[0 ~ 1.0]、または[–1.0 ~ +1.0]に変換することを「正規化」(normalization)と呼ぶことにします。データを正規化することによって、さまざまな性格をもつデータを一定の範囲で比較することが可能になります。

[1] 横軸と縦軸の相対得点

相対得点は横軸（横行）についても(Relative Score in row: R.S.r.)、縦軸（縦列）についても(Relative Score in column: R.S.c.)、それぞれ計算することができます。

相対得点（横軸: R.S.r.） = x / Sm.r

相対得点（縦軸: R.S.c.） = x / Sm.c

ここで、x はそれぞれのセルの値です。Sm.r が横軸 row の和（横和）を示し、Sm.c は列 column の和（縦和）を示します。たとえば、相対得点（横軸）の w3 では x = 11 なので、それを Sm.r (= 29)で割ると 11 / 29 = 0.38 になります。◆Excel ではすべて参照を使います。相対得点（横軸）では、= B18 / $G18 のように分母の列文字（ここでは G）を絶対参照します。分子は列も行も相対参照します。それを全範囲にコピーします。次がその結果です。

[8] この数値に 100 を掛けた値が「百分率」(percent)です。

実測値	L1	L2	L3	L4	L5	和
w1	0	0	0	11	1	12
w2	0	1	2	3	3	9
w3	11	7	10	0	1	29
w4	10	19	14	7	12	62
和	21	27	26	21	17	112

相対得点:横軸	L1	L2	L3	L4	L5
w1	.00	.00	.00	.92	.08
w2	.00	.11	.22	.33	.33
w3	.38	.24	.34	.00	.03
w4	.16	.31	.23	.11	.19

◆同様にして相対得点（縦列）を作成します。このとき、分母 Sm.c.は最下行の和のセルを参照します。相対得点（縦軸）では、= B18 / B$22 のように、分母の行番号を絶対参照します。

相対得点:縦軸	L1	L2	L3	L4	L5
w1	.00	.00	.00	.52	.06
w2	.00	.04	.08	.14	.18
w3	.52	.26	.38	.00	.06
w4	.48	.70	.54	.33	.71

[2] 両軸の相対得点

横軸と縦軸を総合した「相対得点（両軸）」(Relative Score in matrix: R.S.m.)を次のように定義します。

R.S.m. = 2x / [(Sm.r.) + (Sm.c.)]

これは横軸の相対得点と縦軸の相対得点の「分数平均」（→コラム）を使います。つまり、横軸の相対得点 x / Sm.r.と縦軸の相対得点 x / Sm.c.のそれぞれの分子を足したものを分子とし（ここでは分子は同じなので、それぞれもセルの値を 2 倍します）、それぞれの分母を足したものを分母としたものです。たとえば w3, Madrid の分数平均は,横軸の平均は 11/29、縦軸の平均は 11/21 なので、(11 + 11) / (29 + 21) = 0.44 になります。◆Excel では=2*B4/($G4+B$8)のように、それぞれの行和、列和を複合参照し、分子を相対参照します。

相対得点:両軸	L1	L2	L3	L4	L5
w1	.00	.00	.00	.67	.07
w2	.00	.06	.11	.20	.23
w3	.44	.25	.36	.00	.04
w4	.24	.43	.32	.17	.30

[3] 全体の相対得点

全体の相対得点（全体: Relative Score in all: R.S.a.）は、それぞれのセルの値を全範囲（pn 個）の和 Sm.a で割ったものです。次のように数値が非常に小さくなる傾向があります。◆Excel では = B18 / $G$22 のように、分母を絶対参照して動かしません。

R.S.a. = x / Sm.a.

相対得点（全体）R.S.a.	1 Madrid	2 Sevilla	3 México	4 Lima	5 B. A.
w1	0.00	0.00	0.00	0.10	0.01
w2	0.00	0.01	0.02	0.03	0.03
w3	0.10	0.06	0.09	0.00	0.01
w4	0.09	0.17	0.13	0.06	0.11
Relative Score in all: R.S.a. = x / Sm.a.					

割合や百分率などの相対得点(R.S.)の問題点は、データの規模が大きくなると分母が大きくなるので、R.S.が小さな数値になりやすいことです。とくに全体の相対得点(R.S.a.)が小さな数値になる傾向があります。

「相対」と「対照」

数値 X と数値 Y を比較するには、「差」(X – Y)と「比」(X / Y)が使えます。さらに、X / (X + Y)という式も考えられます。これは、分子の X や Y を全体(X + Y)の中で相対化しています。これを「相対型」(Relative Type: R.T.)と呼ぶことにします。

相対型(R.T.) = X / (X + Y)

相対型は[0.0 ~ 1.0]の範囲を持ちます。最小値(0.0)は X = 0 のとき、最大値(1.0)は Y = 0 のときに発生します。中間値は X = Y のときに発生します。

また、(X – Y) / (X + Y)という計算もよく使われます。これを「対照型」(Contrastive Type: C.T.)と名づけたいと思います。

対照型(C.T.) = (X – Y) / (X + Y)

次が先に扱ったデータの横軸の相対得点(r.c.R.)を対照型に変換した結果

です。

対照相対得点(行)	1 Madrid	2 Sevilla	3 México	4 Lima	5 B. A.
w1	-1.00	-1.00	-1.00	0.83	-0.83
w2	-1.00	-0.78	-0.56	-0.33	-0.33
w3	-0.24	-0.52	-0.31	-1.00	-0.93
w4	-0.68	-0.39	-0.55	-0.77	-0.61
Relative Score in row (contrast): R.S.r.(c) = (R.S.r.) * 2 - 1					

対照型の範囲は[-1.0 ~ 1.0]になります。0.0 を中心にして、正負が対照的になります。最小値(-1.0)は X = 0 のとき、そして最大値(1.0)は Y = 0 のときに発生します。中間値は 0.0 ですが、やはり X = Y のときに発生します。このように、対照型の最大値と最小値はそれぞれ「割合」と同じ条件で発生しますが、その範囲が異なります。

なお、相対型と対照型は次の関係があります。

相対型×2 – 1 ＝ 対照型

2 [X / (X + Y)] – 1
= 2X / (X + Y) – 1
= [2X – (X + Y)] / (X + Y)
= (X – Y) / (X + Y)

この 2 つの型は便利なモデルなので、あえて「相対型」と「対照型」という名前をつけておくことを提案しました。相対型は一般に「割合」(ratio)とも呼ばれていますが、これは「X / 全体」という式で示されます。ここで「相対型」と呼ぶ概念は本質的には割合と同じですが、分母の中を X と Y, つまり、比較するものと比較されるものを分けて考えます。そのように見ると、以下で扱うように、いろいろなことがわかるからです。「割合」では隠れて見えなかったことが、相対型にすると、自己を含めた全体と比べる、ということからわかることがあるからです。

一方、対照型は「自己と他者の差」と「自己と他者の和」を比べるわけですから、それにどのような意味があるのか、一見しただけではよくわかりません。そこで、相対型が数値をポジティブに評価するためのもの、対照型が数値をポジティブにもネガティブにも評価するためのもの、と考えます。これは相対型のスケール[0.0 ~ 1.0]を 2 倍して[0.0 ~ 2.0]とし、それから 1 を引いて[-1.0 ~ 1.0]にした操作を見るとわかります。対照型を直感的に納得するには、次のように式を変形するとよいでしょう。

(X – Y) / (X + Y) = X / (X + Y) - Y / (X + Y)

つまり、対照型は X の相対得点と Y の相対得点の差を求めたことになります。

中世・近代スペイン語の前置詞 pora / para

次は、中世・近代スペイン語で起きた前置詞の形態変化 pora > para「～のために」を示す相対頻度と対照頻度の比較です。相対頻度を使うと、それぞれの形に注目して変化を観察することができ、対照頻度を使うと、両者を同時に対照させて変化を観察することができます。

F. R. (pora, para)	1260	1280	1300	1320	1340	1360	1380	1400	1420	1440	1460	1480	1500	1520	1540	1560	1580	1600
Ávila			0.90	0.75				0.00										
Burgos	0.86	1.00	1.00	1.00									0.00		0.11			
Cantabria						0.00		1.00				0.50	0.00	0.00	1.00			
Guadalajara								1.00		1.00			0.00	0.00	0.00	0.00	0.03	0.00
Huesca			1.00	0.00			1.00	1.00		1.00	1.00		1.00					
La Rioja	1.00	1.00		1.00					0.27	0.25	0.00		0.00	0.00				
León	1.00	0.57		0.00				0.00	0.00	1.00	0.00	0.00	1.00	0.00				
Madrid								0.00						0.00	0.08	0.20	0.00	0.03
Navarra		0.83	0.50	1.00	1.00	0.93			0.80		1.00				0.00			
Palencia	1.00	1.00	0.00						0.00		0.67	0.00		0.00				
Salamanca	1.00			0.00	0.50	0.42	0.25	0.00	0.69	0.60		0.19	0.75	0.11				0.00
Segovia			0.60						1.00						0.25			
Teruel		1.00		1.00		1.00	0.63	0.95	0.82	0.90	0.67	0.33			1.00			
Toledo						1.00		0.50		0.14	0.50	0.14	1.00		0.00	0.00		
Valladolid		1.00					1.00	0.80				0.00	0.22	0.00	0.00	0.00	0.12	
Zamora	1.00						0.00	0.00	0.25	0.50	0.08					0.00		
Zaragoza	1.00		0.00	1.00	1.00	0.38		0.89	0.92	1.00	0.00		0.86	0.10	0.33			
Total	0.92	0.88	0.76	0.68	0.80	0.64	0.61	0.76	0.69	0.54	0.35	0.12	0.40	0.03	0.12	0.04	0.05	0.02

相対頻度: Pora

F. R. (para, pora)	1260	1280	1300	1320	1340	1360	1380	1400	1420	1440	1460	1480	1500	1520	1540	1560	1580	1600
Ávila			0.10	0.25				1.00										
Burgos	0.14	0.00	0.00	0.00									1.00		0.89			
Cantabria						1.00		0.00				0.50	1.00	1.00	0.00			
Guadalajara								0.00		0.00			1.00	1.00	1.00	1.00	0.97	1.00
Huesca			0.00	1.00			0.00	0.00		0.00	0.00		0.00					
La Rioja	0.00	0.00		0.00					0.73	0.75	1.00		1.00	1.00				
León	0.00	0.43		1.00				1.00	1.00	0.00	1.00	1.00	0.00	1.00				
Madrid								1.00						1.00	0.92	0.80	1.00	0.97
Navarra		0.17	0.50	0.00	0.00	0.07			0.20		0.00				1.00			
Palencia	0.00	0.00	1.00						1.00		0.33	1.00		1.00				
Salamanca	0.00			1.00	0.50	0.58	0.75	1.00	0.31	0.40		0.81	0.25	0.89				1.00
Segovia			0.40						0.00						0.75			
Teruel		0.00		0.00		0.00	0.38	0.05	0.18	0.10	0.33	0.67			0.00			
Toledo						0.00		0.50		0.86	0.50	0.86	0.00		1.00	1.00		
Valladolid		0.00					0.00	0.20				1.00	0.78	1.00	1.00	1.00	0.88	
Zamora	0.00						1.00	1.00	0.75	0.50	0.92					1.00		
Zaragoza	0.00		1.00	0.00	0.00	0.63		0.11	0.08	0.00	1.00		0.14	0.90	0.67			
Total	0.08	0.13	0.24	0.32	0.20	0.36	0.39	0.24	0.31	0.46	0.65	0.88	0.60	0.97	0.88	0.96	0.95	0.98

相対頻度: Para

F. C. (para, pora)	1260	1280	1300	1320	1340	1360	1380	1400	1420	1440	1460	1480	1500	1520	1540	1560	1580	1600
Ávila			-0.80	-0.50				1.00										
Burgos	-0.71	-1.00	-1.00	-1.00									1.00		0.79			
Cantabria						1.00		-1.00				0.00	1.00	1.00	-1.00			
Guadalajara								-1.00		-1.00			1.00	1.00	1.00	1.00	0.94	1.00
Huesca			-1.00	1.00			-1.00	-1.00		-1.00	-1.00		-1.00					
La Rioja	-1.00	-1.00		-1.00					0.45	0.50	1.00		1.00	1.00				
León	-1.00	-0.14		1.00				1.00	1.00	-1.00	1.00	1.00	-1.00	1.00				
Madrid								1.00						1.00	0.85	0.60	1.00	0.94
Navarra		-0.67	0.00	-1.00	-1.00	-0.87			-0.60		-1.00				1.00			
Palencia	-1.00	-1.00	1.00						1.00		-0.33	1.00		1.00				
Salamanca	-1.00			1.00	0.00	0.17	0.50	1.00	-0.38	-0.20		0.62	-0.50	0.78				1.00
Segovia			-0.20						-1.00						0.50			
Teruel		-1.00		-1.00		-1.00	-0.25	-0.89	-0.64	-0.80	-0.33	0.33			-1.00			
Toledo						-1.00		0.00		0.71	0.00	0.71	-1.00		1.00	1.00		
Valladolid		-1.00					-1.00	-0.60				1.00	0.56	1.00	1.00	1.00	0.76	
Zamora	-1.00						1.00	1.00	0.50	0.00	0.85					1.00		
Zaragoza	-1.00		1.00	-1.00	-1.00	0.25		-0.78	-0.84	-1.00	1.00		-0.71	0.81	0.33			
Total	-0.85	-0.75	-0.52	-0.37	-0.60	-0.28	-0.22	-0.52	-0.37	-0.08	0.30	0.76	0.20	0.94	0.76	0.92	0.90	0.95

対照頻度: Pora - Para

◇2 加重得点

[1] 横軸と縦軸の加重得点

たとえば、w1L2 の 19 は横和が 62 ですから、この相対得点は 19/62 = .31 になります。一方、w4L4 の 3 の当体得点はは 3/9 = .33 になり、w1L2 よりも大きな値になります。しかし、私たちの直感では、前者の 19 のほうが後者の 3 よりも「重い」値だと感じられます（→コラム）。

実測値	L1	L2	L3	L4	L5	和
w1	10	19	14	7	12	62
w2	11	7	10	0	1	29
w3	0	0	1	12	1	14
w4	0	1	2	3	3	9
和	21	27	27	22	17	114

このように実測値の得点を比較するとき、その実測値と相対得点の積にすると、実態を表す数値として直感的に納得がいくことがあります。そこで「加重得点」(W.S.: Weighted Score)として次の式を提案します。x=0 のときに W.S.の最小値ゼロになり、x = 和(Sm)のとき、つまりデータの中に x 以外の数値がないときに最大値が x になります。

$$W.S. = O.S. \times R.S. = x \; x \; / \; Sm = x^2 \; / \; Sm$$

$$W.S.: \; 0.0 \; (x=0) \; \leqq \; 0.5 \; (x^2 = Sm \; / \; 2) \; \leqq \; x \; (x = Sm)$$

次が行、列、行列、全体の加重得点を示します。

W.S.row	L1	L2	L3	L4	L5
w1	1.61	5.82	3.16	.79	2.32
w2	4.17	1.69	3.45		.03
w3			.07	10.29	.07
w4		.11	.44	1.00	1.00

W.S.col.	L1	L2	L3	L4	L5
w1	4.76	13.37	7.26	2.23	8.47
w2	5.76	1.81	3.70		.06
w3			.04	6.55	.06
w4		.04	.15	.41	.53

W.S.both	L1	L2	L3	L4	L5
w1	2.41	8.11	4.40	1.17	3.65
w2	4.84	1.75	3.57		.04
w3			.05	8.00	.06
w4		.06	.22	.58	.69

W.S.all	L1	L2	L3	L4	L5
w1	.88	3.17	1.72	.43	1.26
w2	1.06	.43	.88		.01
w3			.01	1.26	.01
w4		.01	.04	.08	.08

この加重得点をさらにそれぞれの範囲の最大値で割って、正規化します(正規加重得点: W.S.+n.:Weighted Score, +normalized）。

W.S.+n. = W. S. / Max(*range*)

加重得点:横軸	L1	L2	L3	L4	L5
w1	0.00	0.00	0.00	10.08	0.08
w2	0.00	0.11	0.44	1.00	1.00
w3	4.17	1.69	3.45	0.00	0.03
w4	1.61	5.82	3.16	0.79	2.32

同様に、縦軸についても加重得点(x^2 / Sm.c.)を求めます。

加重得点:縦軸	L1	L2	L3	L4	L5
w1	0.00	0.00	0.00	5.76	0.06
w2	0.00	0.04	0.15	0.43	0.53
w3	5.76	1.81	3.85	0.00	0.06
w4	4.76	13.37	7.54	2.33	8.47

打率と打数

たとえば、シーズンを通して 10 打数 3 安打という成績の野球選手と 100 打数 25 安打の選手の成績を比べるとき、打率だけを見ると 0.3 と 0.25 になり、前者のほうが優秀ということになります。しかし、安打数で比べるならば後者のほうが優秀です。これを加重得点で比べるならば、0.9 と 6.25 という数値になり、後者のほうが前者の 7 倍(6.944)近い成績になります。このように加重得点のほうが直感に合う数値のように思われます。

[2] 両軸の加重得点

加重得点（両軸：Weighted Score in matrix: W.S.m.）の式は横軸の加重得点と縦軸の加重得点の分数平均です。

$$W.S.m. = (x^2 + x^2) / (Sm.r + Sm.c.) = 2x^2 / (Sm.r + Sm.c.)$$

加重得点:両軸	L1	L2	L3	L4	L5
w1	0.00	0.00	0.00	7.33	0.07
w2	0.00	0.06	0.23	0.60	0.69
w3	4.84	1.75	3.64	0.00	0.04
w4	2.41	8.11	4.45	1.18	3.65

[3] 全体の加重得点

表全体の加重得点（Weighted Score in all: W.S.a）を求めるには、分母に全体の得点(Sm.a.)を使います。表全体の総和(N)で相対化されるために全体的に数値低くなる傾向があります。

$$W.S.m. = x^2 / Sm.a.$$

加重得点:全体	L1	L2	L3	L4	L5
w1	0.00	0.00	0.00	1.08	0.01
w2	0.00	0.01	0.04	0.08	0.08
w3	1.08	0.44	0.89	0.00	0.01
w4	0.89	3.22	1.75	0.44	1.29

◇3 限定得点

実測値の最小値を 0 とし、最大値を 1 として、範囲を[0.0 ~ 1.0]に限定して計算しなおした値を「限定得点」(Limited Score: L.S.)と呼ぶことにします。次のように行、列、全体の、最小値と最大値を使います。

実測値	L1	L2	L3	L4	L5	最小値	最大値
w1	0	0	0	11	1	0	11
w2	0	1	2	3	3	0	3
w3	11	7	10	0	1	0	11
w4	10	19	14	7	12	7	19
最小値	0	0	0	0	1	0	
最大値	11	19	14	11	12		19

$$L.S. = (x - Mn) / (Mx - Mn)$$

$$L.S.: 0.0\ (x = Mn) \leqq 0.5\ (x = (Mx - Mn) / 2) \leqq 1.0\ (x = Mx)$$

ここで Mn が x を含むデータの最小値、Mx がその最大値を示します。x = Mn のとき、L.S.は最小値 0.0 になり、x = Mx のとき、L.S.は最大値 1.0 になります。中点(0.5)は x が Mx と Mn の中間にあるときです。

[1] 横軸と縦軸の限定得点

横軸の限定得点(L.S.r)は次のようになります。

$$L.S.r. = (x - Mn.r.) / (Mx.r. - Mn.r.)$$

ここで　Mn.r.は横軸の最小値を示し、Mx.r.は横軸の最大値を示します。

限定得点:横軸	L1	L2	L3	L4	L5
w1	.00	.00	.00	1.00	.09
w2	.00	.33	.67	1.00	1.00
w3	1.00	.64	.91	.00	.09
w4	.25	1.00	.58	.00	.42

同様にして、次は縦軸の限定得点(L.S.c.)です。

L.S.c. = (x – Mn.c.) / (Mx.c. – Mn.c.)

限定得点:縦軸	L1	L2	L3	L4	L5
w1	.00	.00	.00	1.00	.00
w2	.00	.05	.14	.27	.18
w3	1.00	.37	.71	.00	.00
w4	.91	1.00	1.00	.64	1.00

[2] 両軸の限定得点

横軸の限定得点と縦軸の限定得点の分数平均が両軸の限定得点(Limited Score in matrix: L.S.m.)です。

L.S.m. = [(x – Mn.r.) + (x – Mn.c.)] / [(Mx.r. – Mn.r.) + (Mx.c. – Mn.c.)]
= (2x – Mn.r. – Mn.c.) / (Mx.r.+ Mx.c – Mn.r. – Mn.c.)

限定得点:両軸	L1	L2	L3	L4	L5
w1	.00	.00	.00	1.00	.05
w2	.00	.09	.24	.43	.36
w3	1.00	.47	.80	.00	.05
w4	.57	1.00	.81	.30	.70

[3] 全体の限定得点

全体の限定得点(Limited Score in all: L.S.a.)は行列全体の最小値 Mn.a.と最大値 Mx.a.を使います。

L.S.a. = (x – Mn.a.) / (Mx.a. – Mn.a.)

限定得点:全体	L1	L2	L3	L4	L5
w1	.00	.00	.00	.58	.05
w2	.00	.05	.11	.16	.16
w3	.58	.37	.53	.00	.05
w4	.53	1.00	.74	.37	.63

限定得点と最大値比得点

限定得点はデータの最小値を 0, 最大値を 1 としていますが、最大値を 1 にしただけのスケールも考えられます。これは、この後で扱う「比較得点」の 1 つである「最大値比得点」です。

最大値比得点:横軸	L1	L2	L3	L4	L5
w1	.00	.00	.00	1.00	.09
w2	.00	.33	.67	1.00	1.00
w3	1.00	.64	.91	.00	.09
w4	.53	1.00	.74	.37	.63

◇4 比較得点

個々のセルの値を平均値、中央値、中間値、最小値、最大値というデータの「代表値」(Representative Value)と比較したものを「比較得点」(Comparative Score: C.S.)と呼ぶことにします。比較の仕方として「差」(difference)、「比」(ratio)、「差比」(difference ratio)を考えます。

[1] 比較平均値差得点

「平均値差得点」(Difference to Average Score: D.A.S.)は、それぞれのセルの値(x)の、平均値(Average: Av)からの差を示します[9]。これは x がゼロのとき最小値の-Av となり、x が和(Sm)と同じとき、つまり、データの中で x 以外はすべてゼロのとき、最大値が Sm - Av = AvCn – Av = (Cn – 1) Av になります（Cn はデータ数）。0.0 は中点ではありませんが、中点と同様に重要な「参照値」（= 平均 Av）です。参照値というのは、これを境に数値の意味（方向）が異なる、ということです。

実測値	L1	L2	L3	L4	L5	平均値
w1	0	0	0	11	1	2.40
w2	0	1	2	3	3	1.80
w3	11	7	10	0	1	5.80
w4	10	19	14	7	12	12.40
平均値	5.25	6.75	6.50	5.25	4.25	5.60

D.A.S. = x – Av

D.A.S.: – Av (x = 0) ≦ 0.0 (x = Av) ≦ Sm – Av (x = Sm)

平均値差得点:横軸	L1	L2	L3	L4	L5
w1	-2.40	-2.40	-2.40	8.60	-1.40
w2	-1.80	-.80	.20	1.20	1.20
w3	5.20	1.20	4.20	-5.80	-4.80
w4	-2.40	6.60	1.60	-5.40	-.40

[9] これは「偏差」(deviation)と呼ばれています。

平均値差得点：縦軸	L1	L2	L3	L4	L5
w1	-5.25	-6.75	-6.50	5.75	-3.25
w2	-5.25	-5.75	-4.50	-2.25	-1.25
w3	5.75	.25	3.50	-5.25	-3.25
w4	4.75	12.25	7.50	1.75	7.75

平均値差(両軸：D.A.S. in matrix: D.A.S.m.)は横軸と縦軸の 2 つの平均値差得点の算術平均とします。

D.A.S.m. = [(D.A.S.r.) + (D.A.S.c.)] / 2

平均値差得点：両軸	L1	L2	L3	L4	L5
w1	-3.83	-4.58	-4.45	7.18	-2.33
w2	-3.53	-3.28	-2.15	-.53	-.03
w3	5.48	.73	3.85	-5.53	-4.03
w4	1.18	9.43	4.55	-1.83	3.68

平均値差(全体：D.A.S. in all: D.A.S.a.)では行列全体の平均(Av.a.)を使います。

平均値差得点：全体	L1	L2	L3	L4	L5
w1	-5.60	-5.60	-5.60	5.40	-4.60
w2	-5.60	-4.60	-3.60	-2.60	-2.60
w3	5.40	1.40	4.40	-5.60	-4.60
w4	4.40	13.40	8.40	1.40	6.40

差の平均と算術平均

2 つの値(x, y)がそれぞれ 2 つの数値の差(x = a – b; y = c - d)を示しているとき、x と y の平均は、次のような単純な「算術平均」(arithmetic average: A.A.)で求めることができます。

A.A.(x, y) = [(a – b) + (c – d)] / 2 = (x + y) / 2

[2] 比較平均値比得点

「比較平均値比得点」(Ratio to Average Score: R.A.S.)は実測値を平均値で割った値（比）です。それぞれ横軸、縦軸、全体の平均値比を見ます。x = 0 のときに最小値 0.0 になり、x = 和 (Sm) のとき、和(Sm) / 平均(Av) = 個数(Cn)になります[10]。中点の 1.0 は x = Av のときです。

[10] そこで、(R.A.S.) / Cn という数値で標準化させれば[0.0 ~ 1.0]のスケール

R.A.S. = x / Av

R.A.S.: 0.0 (x = 0) ≦ 1.0 (x = Av) ≦ Cn (x = Sm)

実測値	L1	L2	L3	L4	L5	平均値
w1	0	0	0	11	1	2.40
w2	0	1	2	3	3	1.80
w3	11	7	10	0	1	5.80
w4	10	19	14	7	12	12.40
平均値	5.25	6.75	6.50	5.25	4.25	5.60

平均値比得点:横軸	L1	L2	L3	L4	L5
w1	.00	.00	.00	4.58	.42
w2	.00	.56	1.11	1.67	1.67
w3	1.90	1.21	1.72	.00	.17
w4	.81	1.53	1.13	.56	.97

平均値比得点:縦軸	L1	L2	L3	L4	L5
w1	.00	.00	.00	2.10	.24
w2	.00	.15	.31	.57	.71
w3	2.10	1.04	1.54	.00	.24
w4	1.90	2.81	2.15	1.33	2.82

両軸の「比較平均値比得点」(Ratio to Average Score in matrix: R.A.S.m)は、「比較平均値比得点（横軸）」と「比較平均値比得点（縦軸）」の分数平均とします。

R.A.S.m. = 2 x / (Av.r. + Av.c.)

平均値比得点:両軸	L1	L2	L3	L4	L5
w1	.00	.00	.00	2.88	.30
w2	.00	.23	.48	.85	.99
w3	1.99	1.12	1.63	.00	.20
w4	1.13	1.98	1.48	.79	1.44

全体の平均値比得点(R.A.S.A.)は全体の平均値(Av.a.)を使います。

R.A.S.a. = x / Av.a.

になりますが、これは(R.A.S.) / Cn = x / (Av Cn) = x / Sm になるので、相対得点(r.s)、つまり「割合」[0.0 ~ 1.0]になります。

平均値比得点:全体	L1	L2	L3	L4	L5
w1	.00	.00	.00	1.96	.18
w2	.00	.18	.36	.54	.54
w3	1.96	1.25	1.79	.00	.18
w4	1.79	3.39	2.50	1.25	2.14

[3] 比較平均値差比得点.

平均値差得点はデータのスケールによって左右されるで、平均差得点を平均値で割ってデータのスケールに合わせます（完全な正規化ではありません）。これを「平均値差比得点」(Difference Ratio to Average Score: D.R.A.S.)と名づけます。0.0 は参照値(x = Av)です。

D.R.A.S. = (d.a.s) / Av = (x.－Av) / Av

D.R.A.S.: -1 (x=0) ≦ 0.0 (x = Av) ≦ (Sm – Av) / Av (x=Sm)

平均値差比得点:横軸	L1	L2	L3	L4	L5
w1	-1.00	-1.00	-1.00	3.58	-.58
w2	-1.00	-.44	.11	.67	.67
w3	.90	.21	.72	-1.00	-.83
w4	-.19	.53	.13	-.44	-.03

平均値差比得点:縦軸	L1	L2	L3	L4	L5
w1	-1.00	-1.00	-1.00	1.10	-.76
w2	-1.00	-.85	-.69	-.43	-.29
w3	1.10	.04	.54	-1.00	-.76
w4	.90	1.81	1.15	.33	1.82

平均値差比得点:両軸	L1	L2	L3	L4	L5
w1	-1.00	-1.00	-1.00	1.88	-.70
w2	-1.00	-.77	-.52	-.15	-.01
w3	.99	.12	.63	-1.00	-.80
w4	.13	.98	.48	-.21	.44

平均値差比得点:全体	L1	L2	L3	L4	L5
w1	-1.00	-1.00	-1.00	.96	-.82
w2	-1.00	-.82	-.64	-.46	-.46
w3	.96	.25	.79	-1.00	-.82
w4	.79	2.39	1.50	.25	1.14

差比の平均と「複合平均」

分子に比較項との差をとり、この差と比較項の比を求める「差比」の両軸の計算をするために、はじめに、先に見た「差の平均」（算術平均: A.A.）を求めます。

$$A.A.= [(x - Av.r.) + (x - Av.c.)] / 2$$

次にこれを分子として、Av.r.との比の平均（分数平均: F.A.）を求めます。

$$(A.A + A.A.) / (Av.r. + Av.c.)$$
$$= 2\ A.A / (Av.r. + Av.c.)$$
$$= [(x - Av.r.) + (x - Av.c.)]$$
$$= (2\ x - Av.r. - Av.c.) / (Av.r. + Av.c.)$$

この式は横軸と縦軸のそれぞれの比較項を導入しているので、次の「複合平均」(Complex Average: C.A.)と呼ぶことにします。

$$C.A. = (2\ x - Av.r. - Av.c.) / (Av.r. + Av.c.)$$

[4] 比較中央値得点

比較する相手を、平均値ではなく中央値にして、差、比、差比を計算したものが「中央値得点」(Median Score: M.S.)です。

実測値	L1	L2	L3	L4	L5	中央値
w1	0	0	0	11	1	.00
w2	0	1	2	3	3	2.00
w3	11	7	10	0	1	7.00
w4	10	19	14	7	12	12.00
中央値	5.00	4.00	6.00	5.00	2.00	3.00

(a) 差得点

中央値差得点:横軸	L1	L2	L3	L4	L5
w1	.00	.00	.00	11.00	1.00
w2	-2.00	-1.00	.00	1.00	1.00
w3	4.00	.00	3.00	-7.00	-6.00
w4	-2.00	7.00	2.00	-5.00	.00
中央値差得点:縦軸	L1	L2	L3	L4	L5
w1	-5.00	-4.00	-6.00	6.00	-1.00
w2	-5.00	-3.00	-4.00	-2.00	1.00
w3	6.00	3.00	4.00	-5.00	-1.00
w4	5.00	15.00	8.00	2.00	10.00

中央値差得点：両軸	L1	L2	L3	L4	L5
w1	-2.50	-2.00	-3.00	8.50	.00
w2	-3.50	-2.00	-2.00	-.50	1.00
w3	5.00	1.50	3.50	-6.00	-3.50
w4	1.50	11.00	5.00	-1.50	5.00

中央値差得点：全体	L1	L2	L3	L4	L5
w1	-3.00	-3.00	-3.00	8.00	-2.00
w2	-3.00	-2.00	-1.00	.00	.00
w3	8.00	4.00	7.00	-3.00	-2.00
w4	7.00	16.00	11.00	4.00	9.00

(b) 比得点

比得点では、w1 の横軸の中央値(MdR)が 0.0 なので分母に 0.0 を使うことになり、エラー(#DIV/0!)になります。比得点であるため、両軸は横軸と縦軸の分数平均とします。

中央値比得点：横軸	L1	L2	L3	L4	L5
w1	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!
w2	.00	.50	1.00	1.50	1.50
w3	1.57	1.00	1.43	.00	.14
w4	.83	1.58	1.17	.58	1.00

中央値比得点：縦軸	L1	L2	L3	L4	L5
w1	.00	.00	.00	2.20	.50
w2	.00	.25	.33	.60	1.50
w3	2.20	1.75	1.67	.00	.50
w4	2.00	4.75	2.33	1.40	6.00

中央値比得点：両軸	L1	L2	L3	L4	L5
w1	.00	.00	.00	4.40	1.00
w2	.00	.50	.67	1.20	3.00
w3	4.40	3.50	3.33	.00	1.00
w4	4.00	9.50	4.67	2.80	12.00

中央値比得点：全体	L1	L2	L3	L4	L5
w1	.00	.00	.00	3.67	.33
w2	.00	.33	.67	1.00	1.00
w3	3.67	2.33	3.33	.00	.33
w4	3.33	6.33	4.67	2.33	4.00

(c) 差比得点

中央値比得点（横軸）と同様に、w1 の横軸の中央値(MdR)が 0.0 なので分

母に 0.0 を使うことになり、エラー(#DIV/0!)になります。

中央値差得点:横軸	L1	L2	L3	L4	L5
w1	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!
w2	-1.00	-.50	.00	.50	.50
w3	.57	.00	.43	-1.00	-.86
w4	-.17	.58	.17	-.42	.00

中央値差比得点:縦軸	L1	L2	L3	L4	L5
w1	-1.00	-1.00	-1.00	1.20	-.50
w2	-1.00	-.75	-.67	-.40	.50
w3	1.20	.75	.67	-1.00	-.50
w4	1.00	3.75	1.33	.40	5.00

中央値差比得点:両軸	L1	L2	L3	L4	L5
w1	-1.00	-1.00	-1.00	3.40	.00
w2	-1.00	-.67	-.50	-.14	.50
w3	.83	.27	.54	-1.00	-.78
w4	.18	1.38	.56	-.18	.71

中央値差比得点:全体	L1	L2	L3	L4	L5
w1	-1.00	-1.00	-1.00	2.67	-.67
w2	-1.00	-.67	-.33	.00	.00
w3	2.67	1.33	2.33	-1.00	-.67
w4	2.33	5.33	3.67	1.33	3.00

[5] 比較中間値得点

比較する相手を中間値（Center: C: 最大値と最小値の中間値）にして、差、比、差比を計算したものが「中間値得点」(Center Score: C.S.)です。

実測値	L1	L2	L3	L4	L5	中間値
w1	0	0	0	11	1	5.50
w2	0	1	2	3	3	1.50
w3	11	7	10	0	1	5.50
w4	10	19	14	7	12	13.00
中間値	5.50	9.50	7.00	5.50	6.50	9.50

(a) 差得点

中間値差得点:横軸	L1	L2	L3	L4	L5
w1	-5.50	-5.50	-5.50	5.50	-4.50
w2	-1.50	-.50	.50	1.50	1.50
w3	5.50	1.50	4.50	-5.50	-4.50
w4	-3.00	6.00	1.00	-6.00	-1.00

中間値差得点:縦軸	L1	L2	L3	L4	L5
w1	-5.50	-9.50	-7.00	5.50	-5.50
w2	-5.50	-8.50	-5.00	-2.50	-3.50
w3	5.50	-2.50	3.00	-5.50	-5.50
w4	4.50	9.50	7.00	1.50	5.50

中間値差得点:両軸	L1	L2	L3	L4	L5
w1	-5.50	-7.50	-6.25	5.50	-5.00
w2	-3.50	-4.50	-2.25	-.50	-1.00
w3	5.50	-.50	3.75	-5.50	-5.00
w4	.75	7.75	4.00	-2.25	2.25

中間値差得点:全体	L1	L2	L3	L4	L5
w1	-9.50	-9.50	-9.50	1.50	-8.50
w2	-9.50	-8.50	-7.50	-6.50	-6.50
w3	1.50	-2.50	.50	-9.50	-8.50
w4	.50	9.50	4.50	-2.50	2.50

(b) 比得点

中間値比得点:横軸	L1	L2	L3	L4	L5
w1	.00	.00	.00	2.00	.18
w2	.00	.67	1.33	2.00	2.00
w3	2.00	1.27	1.82	.00	.18
w4	.77	1.46	1.08	.54	.92

中間値比得点:縦軸	L1	L2	L3	L4	L5
w1	.00	.00	.00	2.00	.15
w2	.00	.11	.29	.55	.46
w3	2.00	.74	1.43	.00	.15
w4	1.82	2.00	2.00	1.27	1.85

中間値比得点:両軸	L1	L2	L3	L4	L5
w1	.00	.00	.00	2.00	.17
w2	.00	.18	.47	.86	.75
w3	2.00	.93	1.60	.00	.17
w4	1.08	1.69	1.40	.76	1.23

中間値比得点:全体	L1	L2	L3	L4	L5
w1	.00	.00	.00	1.16	.11
w2	.00	.11	.21	.32	.32
w3	1.16	.74	1.05	.00	.11
w4	1.05	2.00	1.47	.74	1.26

(c) 差比得点

中間値差比得点:横軸	L1	L2	L3	L4	L5
w1	-1.00	-1.00	-1.00	1.00	-.82
w2	-1.00	-.33	.33	1.00	1.00
w3	1.00	.27	.82	-1.00	-.82
w4	-.23	.46	.08	-.46	-.08

中間値差比得点:縦軸	L1	L2	L3	L4	L5
w1	-1.00	-1.00	-1.00	1.00	-.85
w2	-1.00	-.89	-.71	-.45	-.54
w3	1.00	-.26	.43	-1.00	-.85
w4	.82	1.00	1.00	.27	.85

中間値差比得点:両軸	L1	L2	L3	L4	L5
w1	-1.00	-1.00	-1.00	1.00	-.83
w2	-1.00	-.82	-.53	-.14	-.25
w3	1.00	-.07	.60	-1.00	-.83
w4	.08	.69	.40	-.24	.23

中間値差比得点:全体	L1	L2	L3	L4	L5
w1	-1.00	-1.00	-1.00	.16	-.89
w2	-1.00	-.89	-.79	-.68	-.68
w3	.16	-.26	.05	-1.00	-.89
w4	.05	1.00	.47	-.26	.26

[6] 比較最小値得点

比較する相手を最小値（Minimumr: Mn.）にして、差、比、差比を計算したものが「最小値」(Minimumr Score: Mn.S.)です。

実測値	L1	L2	L3	L4	L5	最小値
w1	0	0	0	11	1	.00
w2	0	1	2	3	3	.00
w3	11	7	10	0	1	.00
w4	10	19	14	7	12	7.00
最小値	.00	.00	.00	.00	1.00	.00

(a) 差得点

最小値差得点:横軸	L1	L2	L3	L4	L5
w1	.00	.00	.00	11.00	1.00
w2	.00	1.00	2.00	3.00	3.00
w3	11.00	7.00	10.00	.00	1.00
w4	3.00	12.00	7.00	.00	5.00

最小値差得点:縦軸	L1	L2	L3	L4	L5
w1	.00	.00	.00	11.00	.00
w2	.00	1.00	2.00	3.00	2.00
w3	11.00	7.00	10.00	.00	.00
w4	10.00	19.00	14.00	7.00	11.00

最小値差得点:両軸	L1	L2	L3	L4	L5
w1	.00	.00	.00	11.00	.50
w2	.00	1.00	2.00	3.00	2.50
w3	11.00	7.00	10.00	.00	.50
w4	6.50	15.50	10.50	3.50	8.00

最小値差得点:全体	L1	L2	L3	L4	L5
w1	.00	.00	.00	11.00	1.00
w2	.00	1.00	2.00	3.00	3.00
w3	11.00	7.00	10.00	.00	1.00
w4	10.00	19.00	14.00	7.00	12.00

(b) 比得点

最小値比得点:横軸	L1	L2	L3	L4	L5
w1	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!
w2	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!
w3	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!
w4	1.43	2.71	2.00	1.00	1.71

最小値比得点:縦軸	L1	L2	L3	L4	L5
w1	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	1.00
w2	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	3.00
w3	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	1.00
w4	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	12.00

最小値比得点:両軸	L1	L2	L3	L4	L5
w1	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	2.00
w2	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	6.00
w3	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	2.00
w4	2.86	5.43	4.00	2.00	3.00

最小値比得点:全体	L1	L2	L3	L4	L5
w1	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!
w2	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!
w3	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!
w4	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!

(c) 差比得点

最小値差比得点:横軸	L1	L2	L3	L4	L5
w1	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!
w2	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!
w3	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!
w4	.43	1.71	1.00	.00	.71

最小値差比得点:縦軸	L1	L2	L3	L4	L5
w1	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	.00
w2	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	2.00
w3	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	.00
w4	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	11.00

最小値差比得点:両軸	L1	L2	L3	L4	L5
w1	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	1.00
w2	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	5.00
w3	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	1.00
w4	1.86	4.43	3.00	1.00	2.00

最小値差比得点:全体	L1	L2	L3	L4	L5
w1	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!
w2	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!
w3	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!
w4	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!

[7] 比較最大値得点

比較する相手を最大値（Maximum: Mx.）にして、差、比、差比を計算したものが「最大値得点」(Maximum Score: Mx.S.)です。

実測値	L1	L2	L3	L4	L5	最大値
w1	0	0	0	11	1	11
w2	0	1	2	3	3	3
w3	11	7	10	0	1	11
w4	10	19	14	7	12	19
最大値	11	19	14	11	12	19

(a) 差得点

最大値差得点:横軸	L1	L2	L3	L4	L5
w1	-11.00	-11.00	-11.00	.00	-10.00
w2	-3.00	-2.00	-1.00	.00	.00
w3	.00	-4.00	-1.00	-11.00	-10.00
w4	-9.00	.00	-5.00	-12.00	-7.00

最大値差得点:縦軸	L1	L2	L3	L4	L5
w1	-11.00	-19.00	-14.00	.00	-11.00
w2	-11.00	-18.00	-12.00	-8.00	-9.00
w3	.00	-12.00	-4.00	-11.00	-11.00
w4	-1.00	.00	.00	-4.00	.00

最大値差得点:両軸	L1	L2	L3	L4	L5
w1	-11.00	-15.00	-12.50	.00	-10.50
w2	-7.00	-10.00	-6.50	-4.00	-4.50
w3	.00	-8.00	-2.50	-11.00	-10.50
w4	-5.00	.00	-2.50	-8.00	-3.50

最大値差得点:全体	L1	L2	L3	L4	L5
w1	-19.00	-19.00	-19.00	-8.00	-18.00
w2	-19.00	-18.00	-17.00	-16.00	-16.00
w3	-8.00	-12.00	-9.00	-19.00	-18.00
w4	-9.00	.00	-5.00	-12.00	-7.00

(b) 比得点

最大値比得点:横軸	L1	L2	L3	L4	L5
w1	.00	.00	.00	1.00	.09
w2	.00	.33	.67	1.00	1.00
w3	1.00	.64	.91	.00	.09
w4	.53	1.00	.74	.37	.63

最大値比得点:縦軸	L1	L2	L3	L4	L5
w1	.00	.00	.00	1.00	.08
w2	.00	.05	.14	.27	.25
w3	1.00	.37	.71	.00	.08
w4	.91	1.00	1.00	.64	1.00

最大値比得点:両軸	L1	L2	L3	L4	L5
w1	.00	.00	.00	1.00	.09
w2	.00	.09	.24	.43	.40
w3	1.00	.47	.80	.00	.09
w4	.67	1.00	.85	.47	.77

最大値比得点:全体	L1	L2	L3	L4	L5
w1	.00	.00	.00	.58	.05
w2	.00	.05	.11	.16	.16
w3	.58	.37	.53	.00	.05
w4	.53	1.00	.74	.37	.63

(c) 差比得点

最大値差比得点:横軸	L1	L2	L3	L4	L5
w1	-1.00	-1.00	-1.00	.00	-.91
w2	-1.00	-.67	-.33	.00	.00
w3	.00	-.36	-.09	-1.00	-.91
w4	-.47	.00	-.26	-.63	-.37

最大値差比得点:縦軸	L1	L2	L3	L4	L5
w1	-1.00	-1.00	-1.00	.00	-.92
w2	-1.00	-.95	-.86	-.73	-.75
w3	.00	-.63	-.29	-1.00	-.92
w4	-.09	.00	.00	-.36	.00

最大値差比得点:両軸	L1	L2	L3	L4	L5
w1	-1.00	-1.00	-1.00	.00	-.91
w2	-1.00	-.91	-.76	-.57	-.60
w3	.00	-.53	-.20	-1.00	-.91
w4	-.33	.00	-.15	-.53	-.23

最大値差比得点:全体	L1	L2	L3	L4	L5
w1	-1.00	-1.00	-1.00	-.42	-.95
w2	-1.00	-.95	-.89	-.84	-.84
w3	-.42	-.63	-.47	-1.00	-.95
w4	-.47	.00	-.26	-.63	-.37

◇5 卓立得点

[1] 横軸と縦軸の卓立得点

「自分（セル）が他のメンバー（セル）たちと違う」ことを示す「卓立得点」(Prominent Score: P.S.)という数値を提案します。ここでは 1 つのセルの値(x)、たとえば w3:L3=10 を取り出して説明しましょう[11]。

[11] ここで扱う式は少し複雑なので、これまでのように Sm.r., Sm.c., Sm.a., Cn.r., Cn.c., Cn.a.ではなく、それぞれ s, t, N, p, n, pn を使います。

実測値	L1	L2	L3	L4	L5	和	**個数**
w1	0	0	0	11	1	12	5
w2	0	1	2	3	3	9	5
w3	11	7	10	0	1	29	5
w4	10	19	14	7	12	62	5
和	21	27	26	21	17	112	
個数	4	4	4	4	4		20

ここで、*x* の実測値(=10)を、横行の他の値全体の和(Sm.r. – x = 29 - 10 = 19)と比較します。このとき、そのまま比較するのではなく、*x* に p - 1 = 5 - 1 = 4 を掛けた値(Cn.r. – 1) x と Sm.r. – x を比較します。これは x (1 個)の大きさを、他のセル全部(p - 1 個)と比べると不利になるからです。そこで、セルの数を同じと見なしたときの x の値(Cn.r. - 1) x を考えます。(Cn.r. - 1) x を(Sm.r. – x)と相対化した値は(Cn.r – 1) x / [(Cn.r. – 1) x + (Sm.r. – x)]です。これを横軸の卓立得点(P.S.r.)とします。卓立係数は相対型 X / (X + Y)なので、[0.0 ~ 1.0]のスケール（範囲）になります。

$$\begin{aligned} \text{P.S.} &= (Cn - 1)\,x \,/\, [(Cn - 1)\,x + (Sm - x)] \\ &= (Cn - 1)\,x \,/\, [(Cn - 2)\,x + Sm] \end{aligned}$$

ここで、P.S.は x と x 以外のメンバーの平均(s – x) / (p – 1)を要素とする相対型 X / (X + Y)になっていることがわかります。そこで、最小値(0.0)は X = 0 のときなので x = 0 のときになります。最大値(1.0)は Y = 0 のときなので s－x = 0 のときです。そして、中間値(0.5)は X = Y のときですから、(p – 1)x = (p – 1) (s – x) / (p – 1)、よって x = (s – x) / (p – 1)のときです。これは、x がほかのメンバー(s – x)の平均 (s – x) / (p – 1)のときです。つまり、P.S.は自身とほかのメンバーの平均を比べた値です。それよりも小さければ 0.5 以下になり、大きければ 0.5 以上になります。

$$\text{P.S.: } 0.0\ (x = 0) \leqq 0.5\ [(Cn - 1)\,x = (Sm - x)] \leqq 1.0\ (x = Sm)$$

卓立得点:横軸	L1	L2	L3	L4	L5
w1	.00	.00	.00	.98	.27
w2	.00	.33	.53	.67	.67
w3	.71	.56	.68	.00	.13
w4	.43	.64	.54	.34	.49

セルの数が多くなると、相対得点(R.S.)は小さくなりがちですが、卓立得点(P.S.)ではセルの数(Cn)の大小にあまり左右されない数値が得られます。これは P.S.の式の分子にも分母にも Cn x があるためです。

同様に、横軸と縦軸のそれぞれの卓立得点（P.S.r.; P.S.c.）は

P.S.r. = (Cn.r. – 1) x / [(Cn.r. – 2) x + Sm.r.]

P.S.c. = (Cn.c. – 1) x / [(Cn.c. – 2) x + Sm.c.]

卓立得点:縦軸	L1	L2	L3	L4	L5
w1	.00	.00	.00	.77	.16
w2	.00	.10	.20	.33	.39
w3	.77	.51	.65	.00	.16
w4	.73	.88	.78	.60	.88

[2] 両軸の卓立得点

横軸と縦軸の卓立得点の分数平均を求め、これを「卓立得点」（両軸: prominent score in Matrix: P.S.M.）と定義します。

P.S.m. = [(Cn.r. – 1) x + (Cn.c. – 1) x]
/ {[(Cn.r. – 2) x + Sm.r.] + [(Cn.c. – 2) x + Sm.c.]}

= (Cn.r. + Cn.c. – 2) x / [(Cn.r. + Cn.c. – 4) x + Sm.r + Sm.c]

卓立得点:両軸	L1	L2	L3	L4	L5
w1	.00	.00	.00	.88	.21
w2	.00	.17	.31	.47	.51
w3	.73	.54	.67	.00	.14
w4	.53	.72	.62	.42	.60

[3] 全体の卓立得点

全体の卓立得点(Prominent Score in all: P.S.a)は x を行列全体のそのほかのメンバーと比較します。そのとき、x には行列全体の個数 Cn.a. – 1 を加重して不利にならないようにします。

P.S.a. = [Cn.a – 1) x / [(Cn.a. – 2) x + Sm.a.]

卓立得点:全体	L1	L2	L3	L4	L5
w1	.00	.00	.00	.67	.15
w2	.00	.15	.26	.34	.34
w3	.67	.56	.65	.00	.15
w4	.65	.80	.73	.56	.70

正規得点のスケールの拡大

相対得点、限定得点、卓立得点は[0.0 ~ 1.0]のスケール（範囲）で正規化

されています。その中間点は 0.5 です。このように[0.0 ~ 1.0]のスケール（範囲）で正規化された得点を[-1.0 ~ 1.0]のスケールにするには、その得点を 2 倍して 1 を引きます。[0.0 ~ 1.0]を 2 倍すると[0.0 ~ 2.0]になり、これから 1 を引くと[-1.0 ~ 1.0]になるからです。

一般に、相対型 X / (X + Y)を 2 倍して 1 を引いて、[-1.0 ~ 1.0]のスケールにすると、次のように(X – Y) / (X + Y)という「対照型」になります。これをモデルに使いましょう。

$$\frac{2X}{X+Y} - 1 = \frac{2X-X-Y}{X+Y} = \frac{X-Y}{X+Y}$$

そこで、相対得点 R.S.を次のように対照型にするために、はじめに相対型 X / (X + Y)にします。

$$R.S. = \frac{x}{Sm} = \frac{x}{x+(Sm-x)} \qquad 0.0\ (x=0) \leqq R.S. \leqq 1.0\ (x=Sm)$$

これを対照型にした相対得点を「対照相対得点」(Relative Score in Contrast: R.S.(c))と呼ぶことにします。R.S.(c)は R.S.の x と Sm – x を、それぞれ X, Y として、先の対照型(X – Y) / (X + Y)にしたものです。

$$R.S.C. = \frac{x-(Sm-x)}{x+(Sm-x)} = \frac{2x-Sm}{Sm}$$

$$-1\ (x=0) \leqq R.S.(c) \leqq 1\ (x=Sm)$$

R.S.(c)の左式のほうがわかりやすいですが、Excel で計算するときは右式のほうが簡単です。または R.S.を計算してあれば、それを参照し 2 を掛けて 1 を引きます。次がその結果です。

対照相対得点(行)	1 Madrid	2 Sevilla	3 México	4 Lima	5 B. A.
w1	-1.00	-1.00	-1.00	0.83	-0.83
w2	-1.00	-0.78	-0.56	-0.33	-0.33
w3	-0.24	-0.52	-0.31	-1.00	-0.93
w4	-0.68	-0.39	-0.55	-0.77	-0.61
Relative Score in row (contrast): R.S.r.(c) = (R.S.r.) * 2 - 1					

相対頻度はデータの規模が大きくなると一般に全体の数値が下がり、0.5 を超えることが少なくなります。その対照相対頻度は、上の図のように、ほとんどが負になります。

次に限定得点 L.S.を対照化します。はじめに、L.S.を次のように相対型

(X – Y) / (X + Y)にします。Mn が x を含むデータの最小値、Mx がその最大値を示します。

$$\text{L.S.} = \frac{x-Mn}{Mx-Mn} = \frac{x-Mn}{(x-Mn)+(Mx-x)}$$

0.0 (x=Mn) ≦ L.S. ≦ 1.0 (x=Mx)

上の右式は相対型 X / (X + Y)なので、それを対照型(X – Y) / (X + Y)にしたものが「対照限定得点」(Limited Score (contrast): L.S.C.)です。

$$\text{L.S.C.} = = \frac{(x-Mn)-(Mx-x)}{(x-Mn)+(Mx-x)} = \frac{2x-Mx-Mn}{Mx-Mn}$$

-1.0 (x=Mn) ≦ L.S.c. ≦ 1.0 (x=Mx)

この L.S.C.を最初から計算するには上の右式を使います。L.S.がすでに計算されているならば、それを参照して L.S.C. = (L.S.) × 2 – 1 の計算をします。次がその結果です。

対照限定得点(行)1	1 Madrid	2 Sevilla	3 México	4 Lima	5 B. A.
w1	-1.00	-1.00	-1.00	1.00	-0.82
w2	-1.00	-0.33	0.33	1.00	1.00
w3	1.00	0.27	0.82	-1.00	-0.82
w4	-0.50	1.00	0.17	-1.00	-0.17
Limited Score in row (contrast): L.S.r.(c) = (L.S.r.)*2 - 1					

卓立得点 P.S.は次のように相対型 X / (X + Y)で示されます。

$$\text{P.S.} = \frac{(p-1)x}{(p-1)x+(s-x)}$$

0.0 (x=0) ≦ P.S. ≦ 1.0 (x = s)

よって、「対照卓立得点」(Prominent Score (contrast): P.S.(c))は次のようになります。

$$\text{P.S.C.} = \frac{(p-1)x-(s-x)}{(p-1)x+(s-x)}$$

次がその結果です。

対照卓立得点(行)	1 Madrid	2 Sevilla	3 México	4 Lima	5 B. A.
w1	-1.00	-1.00	-1.00	0.96	-0.47
w2	-1.00	-0.33	0.07	0.33	0.33
w3	0.42	0.12	0.36	-1.00	-0.75
w4	-0.13	0.28	0.08	-0.33	-0.02
Prominent Score in row (contrast): P.S.r.c. = (P.S.r.)*2 - 1					

逆に、[-1.0 ~ 1.0]のスケールを[0.00 ~ 1.00]のスケールにするには、1 を足して[0.0 ~ 2.0]のスケールにして、次に 2 で割って[0.0 ~ 1.0]のスケールにします。

◇6 標準得点

[1] 標準得点

それぞれの横軸、縦軸または行列全体を同じスケールとばらつきで評価するには、和と平均を 0 にすることに加えて、標準偏差が 1 になるようにする必要があります。この操作は平均値差（偏差）を標準偏差で割ることで可能になります。この値を「標準得点」(Standard Score: S.S.と呼びます[12]。

実測値	L1	L2	L3	L4	L5	平均	標準偏差	個数
w1	0	0	0	11	1	2.40	4.32	5
w2	0	1	2	3	3	1.80	1.17	5
w3	11	7	10	0	1	5.80	4.53	5
w4	10	19	14	7	12	12.40	4.03	5
平均	5.25	6.75	6.50	5.25	4.25	5.60		
標準偏差	5.26	7.56	5.72	4.15	4.55		5.65	
個数	4	4	4	4	4			20

標準得点(S.S.)の式は次のとおりです。

$$S.S. = (x - Av) / Sd$$

ここで、x は実測値、Av は平均値、Sd は標準偏差を示します。このように標準得点はそれぞれ元の値から全体の平均値を引いて、さらにその値を全体の標準偏差で割って得られた数値です。次は、このデータを標準得点に置き換えた結果です。

標準得点:横軸	L1	L2	L3	L4	L5
w1	-.56	-.56	-.56	1.99	-.32
w2	-1.54	-.69	.17	1.03	1.03
w3	1.15	.26	.93	-1.28	-1.06
w4	-.60	1.64	.40	-1.34	-.10

[12]「標準得点」は Standarized Measure, Z-Score とも呼ばれています。池田央(1975)『統計的方法 I 基礎』（新曜社）。

標準得点:縦軸	L1	L2	L3	L4	L5
w1	-1.00	-.89	-1.14	1.39	-.71
w2	-1.00	-.76	-.79	-.54	-.27
w3	1.09	.03	.61	-1.27	-.71
w4	.90	1.62	1.31	.42	1.70

標準得点:両軸	L1	L2	L3	L4	L5
w1	-.80	-.77	-.89	1.70	-.52
w2	-1.10	-.75	-.62	-.20	-.01
w3	1.12	.12	.75	-1.27	-.89
w4	.25	1.63	.93	-.45	.86

標準得点:全体	L1	L2	L3	L4	L5
w1	-.99	-.99	-.99	.96	-.81
w2	-.99	-.81	-.64	-.46	-.46
w3	.96	.25	.78	-.99	-.81
w4	.78	2.37	1.49	.25	1.13

標準得点の平均と標準偏差

標準得点の和と平均は 0 になり、標準偏差が 1 になります。これは標準偏差の重要な性質です。

はじめに、標準得点(S.S)の標準偏差がすべて 1 になる理由を確かめておきましょう。はじめに、標準得点の平均(m_{ss})がゼロになることを確かめます。

$$m_{ss} = (S.S._1 + S.S._2 + ... + S.S._n) / n$$

S.S.の定義にしたがって、

$$= [(x_1 - m)/Sd + (x_2 - m)/Sd + ... + (x_n - m)/Sd] / n$$
$$= [(x_1 - m) + (x_2 - m) + ... + (x_n - m)] / (n\ Sd)$$
$$= [(x_1 + x_2 + ... + x_n) - n\ m] / (n\ Sd)$$

ここで、分子の$(x_1 + x_2 + ... + x_n)$は総和を示します。n m は平均の n 倍だから、これも総和となるので、分子はゼロになります。よって標準得点の平均(m_{ss})もゼロです。

次に標準得点の分散(Sd_{ss}^2)は、次のようになります。

$$Sd_{ss}^2 = [(SM_1 - m_{ss})^2 + (SM_2 - m_{ss})^2 + ... + (SM_n - m_{ss})^2] / n$$

先に標準得点の平均 (m_{ss}) がゼロであることを確かめたので、

$$= \{(SM_1 - 0)^2 + (SM_2 - 0)^2 + \ldots + (SM_n - 0)^2\} / n$$

それぞれの標準得点を定義の式に置き換えると、

$$= \{[(x_1 - m)/Sd]^2 + [(x_2 - m)/Sd]^2 + \ldots + [(x_n - m) / Sd]^2]\} / n$$

全体の Sd^2 をくくって外側の分母に移します。

$$= [(x_1 - m)^2 + (x_2 - m)^2 + \ldots + (x_n - m)^2] / (n\ Sd^2)$$

ここで、

$$[(x_1 - m)^2 + (x_2 - m)^2 + \ldots + (x_n - m)^2] / n$$

は、$x_1, x_2, \ldots x_n$ の分散(Sd^2)ですから、先の式は次のようになります。

$$= Sd^2 / Sd^2 = 1$$

標準偏差 Sd は分散の根（ルート）ですから、標準得点の標準偏差も 1 となります。

このようにして尺度を、平均が 0、標準偏差が 1 になるように標準化させた値が標準得点です。標準化前の数値をそのまま比較すると絶対的な尺度になり、全データの中での相対的な価値が勘案されていないことになります。一方、標準得点は平均がゼロ、標準偏差が 1 になるように標準化されているので、点数とか温度とか価格とか（キロ）メートルのような単位がなくなります。これにより、異なる概念（単位）の数値の間の関係も標準得点によって数値化できるようになります。

偏差値

テストでよく使われる「偏差値」は標準得点を 10 倍し 50 を足して計算します。

$$偏差値 = 標準得点 \times 10 + 50$$

そうすると偏差値の平均は 50 になり、標準偏差は 10 になります。標準得点によって、せっかく平均 0, 標準偏差 1 にして標準化したのに、偏差値ではもう一度それを 10 倍して、さらに 50 を足しているのです。これは、私たちが 100 点満点のテストに慣れているためで、そのほうがわかりやすいからでしょう。

[2] 正規標準得点

標準得点をよく観察すると絶対値が 1.00 を超える数値がしばしば現れることがわかります。これは平均との差が標準偏差を超えたことを示しています。偏差値で言えば 40 点以下のケースや 60 以上のケースなので、よく生じる現象です。そこで、標準得点の範囲を[-1.00 ~ 1.00]というスケールで正規化した数値を求めれば、他の正規得点と同様に数値を正規化した尺度で比較することができます。これを「正規標準得点」(Normalized Standard Score: N.S.S.)と名付けることにしましょう。

正規標準得点(N.S.S.)は標準得点(S.S.)を標準得点の理論的な最大値(S.S.max)で割った値とします。先の「正規標準偏差」で見たように、標準偏差の最大値は

$$S.D.max = m\sqrt{(n-1)}$$

そして、標準得点(S.S.)の最大値(S.S.max)は

$$\begin{aligned} S.S.max &= (x - m) / S.D.max \\ &= (x - m) / (m\sqrt{(n-1)}) \\ &= \underline{(x - x/n)} / x/n\sqrt{(n-1)} \\ &= \underline{(nx - x)/n} / x/n\sqrt{(n-1)} \\ &= (n-1)\underline{x/n} / \underline{x/n}\sqrt{(n-1)} \\ &= (n-1) / \sqrt{(n-1)} \\ &= \sqrt{(n-1)} \end{aligned}$$

よって、正規標準得点(N.S.S.)は

$$N.S.S. = S.S. / S.S.max = S.S. / \sqrt{(n-1)}$$

正規標準得点：横軸	L1	L2	L3	L4	L5
w1	-.28	-.28	-.28	1.00	-.16
w2	-.77	-.34	.09	.51	.51
w3	.57	.13	.46	-.64	-.53
w4	-.30	.82	.20	-.67	-.05

正規標準得点：縦軸	L1	L2	L3	L4	L5
w1	-.58	-.52	-.66	.80	-.41
w2	-.58	-.44	-.45	-.31	-.16
w3	.63	.02	.35	-.73	-.41
w4	.52	.94	.76	.24	.98

正規標準得点：両軸	L1	L2	L3	L4	L5
w1	-.43	-.42	-.48	.91	-.28
w2	-.62	-.42	-.35	-.11	-.00
w3	.60	.07	.41	-.68	-.48
w4	.14	.89	.51	-.24	.46

正規標準得点：全体	L1	L2	L3	L4	L5
w1	-.23	-.23	-.23	.22	-.19
w2	-.23	-.19	-.15	-.11	-.11
w3	.22	.06	.18	-.23	-.19
w4	.18	.54	.34	.06	.26

◇7 期待得点

ここで提案する「期待得点」(Expectation Score: E.S.)は、次に示す「期待値」(Expected Frequency: E.F.)を使います[13]。期待値はそれぞれのセルの値が横の和と縦の和から見て、平均に分布しているとすればどのような値として期待されるかを示すものです。「期待される」というよりも「予想される」(expected)と考えたほうがわかりやすいかも知れません。

実測値	L1	L2	L3	L4	L5	和
w1	0	0	0	11	1	12
w2	0	1	2	3	3	9
w3	11	7	10	0	1	29
w4	10	19	14	7	12	62
和	21	27	26	21	17	112

期待値は縦と横の和の割合から計算されます。w1 の和（横和: 0 + 0 + 0 + 11 + 1）が 12 となっています。一方、一番下の和 SmC の横軸に注目すると、1 Madrid の和（縦和: 0+0+11+10）は 21 です。総和は 112 ですから、Madrid の w1 は、横和の 12 回のうち、21 / 112 の割合で出てくると予想されます。つまり、12×(21 / 112) ≒2.25 となります。Excel シートでは横和 Sm.r.を列固定で参照し、縦和 Sm.c.を行固定で参照します。分母の総和 Sm.a.は列も行も固定します。それぞれのセルについての計算結果が次の表です。

[13] 「期待値」(E.F.)は一般に「期待度数」と呼ばれることが多いのですが、ここでは「実測値」と「期待値」を対等に比較する、という意図から両者に「値」という訳語を使います。この訳語「期待値」も使われています。「期待値得点」と、以下で扱う得点(score)は使われていません。

E.F. = (Sm.r. Sm.c.) / Sm.a.

期待値	L1	L2	L3	L4	L5
w1	2.25	2.89	2.79	2.25	1.82
w2	1.69	2.17	2.09	1.69	1.37
w3	5.44	6.99	6.73	5.44	4.40
w4	11.63	14.95	14.39	11.63	9.41

[1] 期待値差

「期待値差」(Difference to Expected Frequency Score: D.E.F.S.)では期待値と実測値の差を計算します。

D.E.F.S. = x – E.F.

期待値差得点	L1	L2	L3	L4	L5
w1	-2.25	-2.89	-2.79	8.75	-.82
w2	-1.69	-1.17	-.09	1.31	1.63
w3	5.56	.01	3.27	-5.44	-3.40
w4	-1.63	4.05	-.39	-4.63	2.59

この表で実測値と期待値の乖離がどの程度がわかります。しかし、スケールが正規化されていないためデータ間で期待値差を比較することはできません。

[2] 期待値比

「期待値比」(Ratio to Expected Frequency Score: R.E.F.S.)を計算するには、実測値／期待値を計算します。

R.E.F.S. = x / E.F.

期待値比得点	L1	L2	L3	L4	L5
w1	.00	.00	.00	4.89	.55
w2	.00	.46	.96	1.78	2.20
w3	2.02	1.00	1.49	.00	.23
w4	.86	1.27	.97	.60	1.28

差ではプラスとマイナスの値で実測値と期待値が比較されますが、比では、実測値も期待値もプラスなので、すべてプラスの数値になり、実測値からの（プラスとマイナスの）差がわかりません。

[3] 期待値差比

差の欠点は、単に実測値と期待値を比較しただけなので、それが絶対化されていることです。それぞれのケースの数値のスケール（相対的な大きさ）に合わせれば、全体を見回した比較ができるようになります。そこで、求めた偏差（のスケール）を期待値（のスケール）で割れば、絶対的な数値ではなく、その数値のスケールに合った相対的な数値が得られます。それが「期待値差比」(Difference Ratio to Expected Frequency Score: D.R.E.F.S.)です[14]。これは差と比を総合した値です。

D.R.E.F.S. = (x – E.F.) / E.F.

期待値差比得点	L1	L2	L3	L4	L5
w1	-1.00	-1.00	-1.00	3.89	-.45
w2	-1.00	-.54	-.04	.78	1.20
w3	1.02	.00	.49	-1.00	-.77
w4	-.14	.27	-.03	-.40	.28

プラスとマイナスの符号は期待値差の場合と同じです。相対誤差は実測値と期待値が同じになったときはゼロになります。これは実測値が予想された値そのものであったことを意味します。たとえば、w3: 2. Sevilla がゼロになっています。これは実測値が予想通りなので情報をもちません。一方、w1: 4. Lima は 3.89 という値になり、期待値よりも実測値がかなり大きいことがわかります。

[4] 期待得点

以上の期待値差、期待値比、期待値差比はどれも[0.0 ~ 1.0]に正規化された数値ではないので扱いがすこし困難です。ここでは、実測値と、期待値の理論的な最大値を比較して正規化した値を求めたいと思います。はじめに期待値の範囲を確認します。期待値が最大になるのは、次のように当該セルの行と列以外のセルがすべてゼロの場合です。

実測値	1 Madrid	2 Sevilla	3 México	和 SmR
w1	10	20	20	50
w2	30	0	0	30
和 SmC	40	20	20	80

[14] 東京大学教養学部統計学教室『統計学入門』（東京大学出版会）(1991:p.247)はこれを「相対誤差」と呼んでいます。

そこで、2 個以上の列と行がある行列を次のように 2 行 x 2 列の行列まとめて表示します。

実測値	当該列	ほかの列	行和
当該行	x	b	x + b
ほかの行	c	d	c + d
列和	x + c	b + d	x + b + c + d

ここでは当該のセルの値(x)は 10 で、そのほかのセルは b = 20 + 20 = 40, c = 30, d = 0 になります。期待値 x のセルの期待値は

E.F.(x) = (x + b)(x + c) / (x + b + c + d)

そして、d = 0 のときの x の期待値は

E.F.(x: d = 0) = (x + b)(x + c) / (x + b + c)

これが「最大期待値」(E.F.max)になります。

E.F.max = E.F.(x: d=0) = (x + b)(x + c) / (x + b + c)

さらに c=0 という状況を考えると、

E.F.(x: c=d=0) = (x + b)x / (x + b) = x

となるので実測値(x)と期待値(x)が一致します。これは、たとえば次のように、当該セルのある行以外の行が全部ゼロの場合です(c=d=0: 50 x 10 / 50 = 10)。

実測値	1 Madrid	2 Sevilla	3 México	和 SmR
w1	10	20	20	50
w2	0	0	0	00
和 SmC	10	20	20	50

これは b=0 のときでも同じです。これは、たとえば次のように、当該セルのある列以外の列が全部ゼロの場合です(b=d=0: 10 x 40 / 40 = 10)。

実測値	1 Madrid	2 Sevilla	3 México	和 SmR
w1	10	0	0	10
w2	30	0	0	30

和 SmC	40	0	0	40

そして、期待値の最小値 E.F.min. (=0)が出現するのは x = 0 のときです。

E.F.min. = 0 (x=0)

以上をまとめると、期待値 E.F.の範囲は次のようになります。

E.F.: 0.0 [x=b=d=0]≦ m. [b=d=0] ≦ (x + b)(x + c) / (x + b + c) [d=0]
E.F.: 0.0 [x=c=d=0]≦ m. [c=d=0] ≦ (x + b)(x + c) / (x + b + c) [d=0]

このときの中点 m.は必ずしも最大値と最小値の中間値ではないのですが、実測値と同じになるので重要な参照値になります。そして、次のように、この中点(m.)は最大値(x + b)(c + c) / (x + b + c)を超えることはありません。

$x \leqq (x + b)(x + c) / (x + b + c)$
$x (x + b + c) \leqq (x + b)(x + c)$
$x^2 + bx + cx \leqq x^2 + cx + bx + bc$
$bc \geqq 0$

ここで、b も c もゼロを含む自然数なので、$bc \geqq 0$ になることは明らかです。そこで、この導出過程を逆に遡れば $x \leqq (x + b)(x + c) / (x + b + c)$にたどり着きます。つまり中点(x)は最大値(x + b)(x + c) / (x + b + c)を超えない、ということです。

「最大期待値」(E.F.max)を計算するために次の式を使います。

E.F.max = (x + b)(x + c) / (x + b + c)
= (x + b)(x + c) / [(x + b) + (x + c) - x)]
= (SmR SmC) / (SmR + SmC – x)

次の図がそれぞれのセルの最大期待値を示しています。

最大期待値	L1	L2	L3	L4	L5
w1	7.64	8.31	8.21	11.45	7.29
w2	6.30	6.94	7.09	7.00	6.65
w3	15.62	15.98	16.76	12.18	10.96
w4	17.84	23.91	21.78	17.13	15.73

「正規期待値得点」(Normalized Expected Frequency Score: N.E.F.S.)は、実測値(x)をこの最大期待値(E.F.max.)で割ったものです。

N.E.F.S.= x / E.F.max

先に見たように、x ≦ E.F.max の関係がありますから、N.E.F.S.R.は[0.0 ~ 1.0]に正規化された数値です。

N.E.F.S.: 0.0 (x = 0) … 0.5 (x = E.F.max / 2) … 1.0 (x = E.F.max)

期待得点	L1	L2	L3	L4	L5
w1	.00	.00	.00	.96	.14
w2	.00	.14	.28	.43	.45
w3	.70	.44	.60	.00	.09
w4	.56	.79	.64	.41	.76

期待得点(E.S.)は、期待値の理論的な最大値（際立った分布において期待される頻度）と比較して実測値を正規化した尺度で評価するものです。これが 1 に近ければ、際立った分布において期待される頻度に近いことを示します。

期待得点(E.S.)を次の式によって対照化すれば、「対照期待得点」(Expectation Score (contrastive): E.S.(c))が得られます。

E.S.(c) = E.S. * 2 - 1

対照期待得点	L1	L2	L3	L4	L5
w1	-1.00	-1.00	-1.00	.92	-.73
w2	-1.00	-.71	-.44	-.14	-.10
w3	.41	-.12	.19	-1.00	-.82
w4	.12	.59	.29	-.18	.53

◇8 逸脱得点

確率的に見て異常な度数を検知する「逸脱得点」(Divergent Score: D.S.)を提案します。

ある事象が起こる確率にはさまざまなものがあります。たとえば、サイコロには{1, 2, 3, 4, 5, 6}という目があるので、1 回サイコロを投げるとき(「試行」と言います)、それぞれの目が出る確率はそれぞれ 1/6 ずつです。これらの目の中の 1 つ、たとえば「1」が出る確率は 1/6 なので、逆に「1」が出ない確率は 1 - 1/6 = 5/6 です。次の表の F (False)は「 1 」が出ないことを示し、T (True)は「 1 」が出ることを示しています。確率の総和が 1 になることを確認してください(5/6 + 1/6 = 1)。

「１」	T の数	確率
F	0	5/6≒0.833
T	1	1/6≒0.167

次にサイコロを 2 回投げる場合(試行回数=2)を考えましょう。たとえば 1 回目が F で 2 回目が T とすると、これを F, T と書きます。この場合も確率の総和は 1 になります(25/36 + 5/36 + 5/36 + 1/36 = 1)。

「１」	T の数	確率
F, F	0	(5/6) (5/6) = 25/36≒0.694
F, T	1	(5/6) (1/6) = 5/36≒0.139
T, F	1	(1/6) (5/6) = 5/36≒0.139
T, T	2	(1/6) (1/6) = 1/36≒0.028

さらに、サイコロを 3 回投げる場合(試行回数=3)を考えます。この場合も確率の総和は 1 になることを確かめてください。

「１」	T の数	確率
F, F, F	0	(5/6) (5/6) (5/6) = 125/216≒0.579
F, F, T	1	(5/6) (5/6) (1/6) = 25/216≒0.116
F, T, F	1	(5/6) (1/6) (5/6) = 25/216≒0.116
T, F, F	1	(1/6) (5/6) (5/6) = 25/216≒0.116
T, T, F	**2**	**(1/6) (1/6) (5/6) = 5/216≒0.023**
T, F, T	**2**	**(1/6) (5/6) (1/6) = 5/216≒0.023**
F, T, T	**2**	**(5/6) (1/6) (1/6) = 5/216≒0.023**
T, T, T	3	(1/6) (1/6) (1/6) = 1/216≒0.005

ここで、たとえばサイコロを 3 回投げて順番を問題にせずに、全部で 2 回「１」が出る場合(T の数=2)の確率を求めると、上の表から、

「１」	T の数	確率
T, T, F	**2**	**(1/6) (1/6) (5/6) = 5/216≒0.023**
T, F, T	**2**	**(1/6) (5/6) (1/6) = 5/216≒0.023**
F, T, T	**2**	**(5/6) (1/6) (1/6) = 5/216≒0.023**

を合わせた確率、つまり、5/216 + 5/216 + 5/216 = 15/216≒0.069 になるこ

とがわかります。これは「1」が 2 回出る場合の確率(5/216)を 3 倍した数です。それぞれの場合の確率 5/216 は$(1/6)^2$ (5/6)、つまり T の確率 1/6 の 2 回分と F の確率 5/6 の 1 回分の積になります。

次に、T, T, F だけでなく、他にも T,F,T と F,T,T があるので、この積 5/216 を 3 倍します。この倍数の 3 を求めるのは、このように少ない試行回数(3 回)ならばすぐ計算できますが、それが多くなると一般式を使わなければなりません。*n* 回の試行で T が *r* 回選ばれる場合の数は *n*C*r* という「組み合わせ」(Combination: *n***C***r*)の値になります[15]。ここでは、T が 2 個で F が 1 個の組み合わせになるので $_3\mathbf{C}_2$ で計算します。そこで、3 回の試行で T が順番を問わずに 2 回出る確率は

$$_3\mathbf{C}_2\,(1/6)^2\,(5/6) = (3 \times 2) / (2 \times 1)\,(1/6)^2\,(5/6) = 15/216 \fallingdotseq 0.069$$

この確率を一般化した式で示すと、

$$_n\mathbf{C}_r\,(p)^r\,(1 - p)^{n-r}$$

になります。ここで *n* はサイコロを投げた総回数（試行数）、*r* は選ばれる回数（成功数）、*p* は T の確率(成功確率:1/6)、1 - *p* は F の確率(失敗確率：5/6)を示します。この確率の分布は「二項分布」(Binomial Distribution)と呼ばれています。

◆二項分布の確率の計算は階乗を多く使うので、*n* や *r* が大きくなると計算が複雑になります。そこで、Excel 関数の BINOMDIST(r, n, p, 0)を使用します。

次は、試行回数= 4 を固定し、成功率を 1/2, 1/3, ..., 1/6 と変化させ、成功回数= 0, 1, 2, 3, 4 のそれぞれの確率を計算した結果です。

---

[15] これは互いに区別のつく 3 個の物{a, b, c}の中から任意の 2 個(= T)を取り出す場合の数と同じです。もし、取り出す順番を考えるならば、ab, ac, ba, bc, ca, cb という 6 個の場合があります。これが「順列」(Permutation: *n*P*r*)で、$_n\mathbf{P}r = n\,(n - 1)(n - 2)\,\ldots\,(n - r + 1)$. ここで、順番を考慮しなければ（「組み合わせ」$_3\mathbf{C}_2$）、ab と ba, ac と ca、bc と cb はそれぞれ同じなので場合の数を 2 で割らなければなりません。この 2 は 2P2 の順列(2! = 2 x 1)です。よって $_3\mathbf{C}_2 = (3 \times 2) / (2 \times 1)$. 一般式は

$$_n\mathbf{C}_r = {}_n\mathbf{P}r \,/\, r! = n\,(n - 1)(n - 2)\,\ldots\,(n - r + 1) \,/\, r! = n! \,/\, [r!(n - r)!]$$

	成功率	x		試行回数	4
成功回数:y	1/2	1/3	1/4	1/5	1/6
0	0.0625	0.1975	0.3164	0.4096	0.4823
1	0.2500	0.3951	0.4219	0.4096	0.3858
2	0.3750	0.2963	0.2109	0.1536	0.1157
3	0.2500	0.0988	0.0469	0.0256	0.0154
4	0.0625	0.0123	0.0039	0.0016	0.0008

たとえば、BINOMDIST(0, 4, 1/2, 0)は 0.0625 を示しています。これはコインを投げて表を出す確率などで 4 回投げて一度も表にならない確率$(1/2)^4$ = 1/16 = 0.0625 を示しています。このように確率が 1/2 のときは、確率の分布が 2 を最大値として、上下対称になります。サイコロの目（たとえば「1」）が出る確率は 1/6 ですが、そのときの成功回数=0 の確率は、$(5/6)^4$ = 0.4823, 成功回数=4 の確率は、$(1/6)^4$ = 0.0008 となって、上下対称ではありません。

ところが次のように試行回数を 4, 5, 6, ..., 20 のように増加させると、次第に分布が上下対称に近づきます。その確率の最大値は、成功率=1/2 のときのように試行回数の中央値ではなく、試行回数と確率の積に近似した成功回数のときの確率になります。たとえば確率が 1/6 で 20 回の試行すれば、成功回数が(1/6) x 20 ≒3 となりますから、成功数=3 の確率が一番高い、ということは直感的にも納得できます。

	成功率	1/6		試行回数	x	
成功回数:y	4	5	6	10	15	20
0	0.4823	0.4019	0.3349	0.1615	0.0649	0.0261
1	0.3858	0.4019	0.4019	0.3230	0.1947	0.1043
2	0.1157	0.1608	0.2009	0.2907	0.2726	0.1982
3	0.0154	0.0322	0.0536	0.1550	0.2363	0.2379
4	0.0008	0.0032	0.0080	0.0543	0.1418	0.2022
5		0.0001	0.0006	0.0130	0.0624	0.1294
6			0.0000	0.0022	0.0208	0.0647
7				0.0002	0.0053	0.0259
8				0.0000	0.0011	0.0084
9				0.0000	0.0002	0.0022
10				0.0000	0.0000	0.0005
11					0.0000	0.0001
12					0.0000	0.0000
13					0.0000	0.0000
14					0.0000	0.0000
15					0.0000	0.0000
16						0.0000
17						0.0000
18						0.0000
19						0.0000
20						0.0000

ここで提案する「逸脱確率得点」(D.P.S.: Divergent Probability Score)は二項分布の確率を利用して求めます。このとき、*r* = 実測値、*n* = 母数、*p* = 全体の中での割合、を使います。

実測値	L1	L2	L3	L4	L5	和
w1	0	0	0	11	1	12
w2	0	1	2	3	3	9
w3	11	7	10	0	1	29
w4	10	19	14	7	12	62
和	21	27	26	21	17	112

上の実測値を使って、たとえば「行」の二項分布得点は、該当するセルの行和(Sm.r.)を *n* とし、列和／総和を *p* とします。w1-1.Madrid(=0)を例にすると、12 回の試行で 0 回起こる確率（成功回数）を、21/112 という全体の確率の二項分布の中での確率を求め(Binomial Score: B.S)、12 x 21 / 112 という成功回数（期待値）での確率（二項分布のの最大値：B.S.max)で割ります。これで得られた商は、最大確率と比較したときの当該確率を正規化した大きさを示すので、「ふつうに起こりうる確率」(0.00～1.00)を示します。ここでは逆に「ふつうには起こりえない逸脱した確率」(0.00 ～ 1.00)を求めたいので、1 からこの数値を引いた数値にします。さらに、実測値が期待値より小さいときは、それをマイナス値にして、評価しやすい形にします(sgn = -1)。

B.D.S. = sgn * [1- B.S. / B.S. (max)]

二項分布逸脱得点（行）B.D.S.r.

二項分布得点:横軸	L1	L2	L3	L4	L5
w1	.08	.04	.04	.00	.30
w2	.15	.24	.31	.16	.11
w3	.01	.17	.06	.00	.04
w4	.12	.06	.12	.04	.09

[B46]=BINOMDIST(B38,$G38,B$42/$G$42,0)

同最大値	L1	L2	L3	L4	L5
w1	.29	.24	.25	.29	.30
w2	.32	.30	.31	.32	.37
w3	.19	.16	.17	.19	.21
w4	.13	.12	.12	.13	.14

[I46]=BINOMDIST($G38*B$42/$G$42,$G38,B$42/$G$42,0)

逸脱得点:横軸	L1	L2	L3	L4	L5
w1	-.72	-.85	-.83	1.00	.00
w2	-.52	-.21	.00	.50	.70
w3	.96	-.04	.65	-.99	-.79
w4	-.08	.52	.00	-.66	.39

[P46]=SIGN(B38-$G38*B$42/$G$42)*(1-B46/I46)

次は D.P.S.を行で求めた場合(D.P.S.c.)です。考え方は同じです。

二項分布得点:縦軸	L1	L2	L3	L4	L5
w1	.09	.05	.05	.00	.30
w2	.17	.25	.28	.15	.11
w3	.01	.17	.06	.00	.04
w4	.13	.05	.15	.02	.09

[B53]=BINOMDIST(B38,B$42,$G38/$G$42,0)

同最大値	L1	L2	L3	L4	L5
w1	.28	.24	.25	.28	.30
w2	.32	.28	.28	.32	.36
w3	.20	.16	.17	.20	.22
w4	.17	.14	.15	.17	.19

[I53]=BINOMDIST($G38*B$42/$G$42,B$42,$G38/$G$42,0)

逸脱確率得点：縦軸	L1	L2	L3	L4	L5
w1	-.67	-.80	-.79	1.00	.00
w2	-.46	-.12	.00	.52	.69
w3	.97	-.05	.66	-.99	-.83
w4	-.19	.68	.00	-.86	.51

[P56]=SIGN(B38-$G38*B$42/$G$42)*(1-B53/I53)

B.D.S.を両軸で求めるときは、B.D.S.R.と B.D.S.C.の分数平均とします。

逸脱確率得点：両軸	L1	L2	L3	L4	L5
w1	-.69	-.83	-.81	1.00	.00
w2	-.49	-.17	.00	.51	.70
w3	.96	-.05	.65	-.99	-.81
w4	-.15	.61	.00	-.77	.46

[P67]=SIGN(B38-$G38*B$42/$G$42)*(1-(B46+B53)/(I46+I53))D.P.S. を全体で求めるときは、分子の B.S.は全体で求め、分母 B.S.(max)は完全に平均化した分布でもとめます。

二項分布得点:全体	L1	L2	L3	L4	L5
w1	.00	.00	.00	.01	.02
w2	.00	.02	.06	.11	.11
w3	.01	.13	.03	.00	.02
w4	.03	.00	.00	.13	.01

[B60]=BINOMDIST(B38,$G$42,1/20,0)

同最大値	L1	L2	L3	L4	L5
w1	.17	.17	.17	.17	.17
w2	.17	.17	.17	.17	.17
w3	.17	.17	.17	.17	.17
w4	.17	.17	.17	.17	.17

[I60]=BINOMDIST($G$42/20,$G$42,1/20,0)

逸脱確率得点：全体	L1	L2	L3	L4	L5
w1	-.98	-.98	-.98	.92	-.89
w2	-.98	-.89	-.68	.39	.39
w3	.92	.25	.83	-.98	-.89
w4	-.83	1.00	-.99	-.25	.96

[P60]=SIGN(B38-$G38*B$42/$G$42)*(1-B60/I60)

◇9 順位得点

[1] 正順位得点

「正順位得点」によって横、縦、全体の範囲で降順の順位をつけます。◆Excel 関数の RANK(*c*,*R*)を使います。*c* は対象のセルを示し、*R* はその範囲を示します。

実測値	L1	L2	L3	L4	L5
w1	0	0	0	11	1
w2	0	1	2	3	3
w3	11	7	10	0	1
w4	10	19	14	7	12

=RANK(B4,$B4:$F4)

正順位得点:横軸	L1	L2	L3	L4	L5
w1	3	3	3	1	2
w2	5	4	3	1	1
w3	1	3	2	5	4
w4	4	1	2	5	3

正順位得点:縦軸	L1	L2	L3	L4	L5
w1	3	4	4	1	3
w2	3	3	3	3	2
w3	1	2	2	4	3
w4	2	1	1	2	1

両軸の順位得点は横軸の順位得点と縦軸の順位得点の単純な平均（算術平均）とします。

正順位得点:両軸	L1	L2	L3	L4	L5
w1	3.0	3.5	3.5	1.0	2.5
w2	4.0	3.5	3.0	2.0	1.5
w3	1.0	2.5	2.0	4.5	3.5
w4	3.0	1.0	1.5	3.5	2.0

正順位得点:全体	L1	L2	L3	L4	L5
w1	16	16	16	4	13
w2	16	13	12	10	10
w3	4	8	6	16	13
w4	6	1	2	8	3

[2] 逆順位得点

逆順位得点は最小値を 1 とした昇順の順位を示します。◆Excel 関数の RANK(*c,R,* 1)を使います。*c* は対象のセルを示し、*R* はその範囲を示します。3 番目の引数として 1 を使います。

逆順位得点:横軸	L1	L2	L3	L4	L5
w1	1	1	1	5	4
w2	1	2	3	4	4
w3	5	3	4	1	2
w4	2	5	4	1	3

逆順位得点:縦軸	L1	L2	L3	L4	L5
w1	1	1	1	4	1
w2	1	2	2	2	3
w3	4	3	3	1	1
w4	3	4	4	3	4

逆順位得点:両軸	L1	L2	L3	L4	L5
w1	1.0	1.0	1.0	4.5	2.5
w2	1.0	2.0	2.5	3.0	3.5
w3	4.5	3.0	3.5	1.0	1.5
w4	2.5	4.5	4.0	2.0	3.5

逆順位得点:全体	L1	L2	L3	L4	L5
w1	1	1	1	16	6
w2	1	6	9	10	10
w3	16	12	14	1	6
w4	14	20	19	12	18

データの抽象化

言語データは言語の現実そのものではなく、分析者が一定の目的を持って複雑な言語の１側面を取り出したものです。したがって、言語データがそのまま言語の現実である、というような過剰な一般化はできません。言語データが示す範囲の中で一定のことがわかった、ということができるぐらいでしょう。

データには多くの数字や記号が並んでいます。観察によって得られた実測値のままではその評価ができないとき、平均と標準偏差という統計量を使って標準得点を求めました。ここで注意したいのは、変換されたデータは実測値そのものではない、ということです。そして、実測値からは標準得点などを求めることができますが、逆に、変換されたデータから実測値を求めることができません。実測値を保存しておかなければ元に戻れないのです。

実測値であれ、変換されたデータであれ、そのデータの傾向がよくわからないとき、私たちは集計表を作成します。しかし、確かに集計表は傾向を見るのには便利なのですが、これは具体的なデータそのものではなく１つの抽象化が施されている、ということです。データから集計表を作ることはできますが、逆に集計表からデータの実態に戻ることはできません。

集計表を見ていても傾向がよくわからないとき、私たちはグラフを作成します。これも同じことですが、グラフから逆に集計表やデータを作成するこ

とはできません。グラフは視覚的なので具体化されたように見えますが、データの現実から見ると実は１つの抽象化をしていることになります。

そして、私たちは数学的な手続きを経て各種の係数を扱います。縦の列と横の行にデータのいろいろな数値が展開されていても係数は 1 つだけの値を返してきます。これは非常に抽象的です。もちろん係数からデータの現実は復元できません。とても個性があり、輝いているような１つのデータがあっても、それは抽象化された係数の中に埋没しています。

それぞれの手法の特徴をよく理解して、可能な限り適切な方法を選択する方法を学びましょう。そして、方法を限定せず、さまざまな方法を組み合わせて、多角的な見方をすることも必要です。柔軟な考え方を身に着けたいと思います。

6 データの相関

【目標】相関係数や Phi 係数などの概念を理解し、2 つのデータ間の関係を数値的に示すことができるようにする。また、カイ二乗検定を使ってクロス集計表の独立性の検定ができるようにする。

前章までの内容は、1 つのデータを代表する値や個々のデータを置き換える数値を扱いました。本章では 2 つの種類のデータを扱い、それらの関係性について統計的に扱います。2 つのデータが関わってきますので、数字の裏にある数式はすこし複雑になりますが、一度理解してしまえば、数値の本質がわかり、その使い方や応用の方法がわかるようになります。はじめには簡単な例や図を使って、統計的な数値を直感的に納得できるようにします。納得できた後でその数学的な根拠を探ります。数学的な根拠がわかったら、自分で手を動かしてそれを実験して確かめましょう。ここまですれば確実にその統計手法が身につきます。

数値の意味がわかったならば、それをたくさん使ってみましょう。そうすれば感覚がだんだんと養われていき、理論的な知識が経験的なスキルによって裏づけられるようになります。知識は使うことで生かされてきます。Excel のようなアプリケーションは、こうした実験をするのにとても便利です。

6.1 量的なデータの相関

◇1 データ

まず頻度やアンケートの結果など、数えたりスケールを測ったりできる量的な数値について扱います。次のデータを見てください。これは西語（スペイン語）の文 1 (Madrid)と文 2(Sevilla)に関して主要な前置詞の頻度を集計したものです。

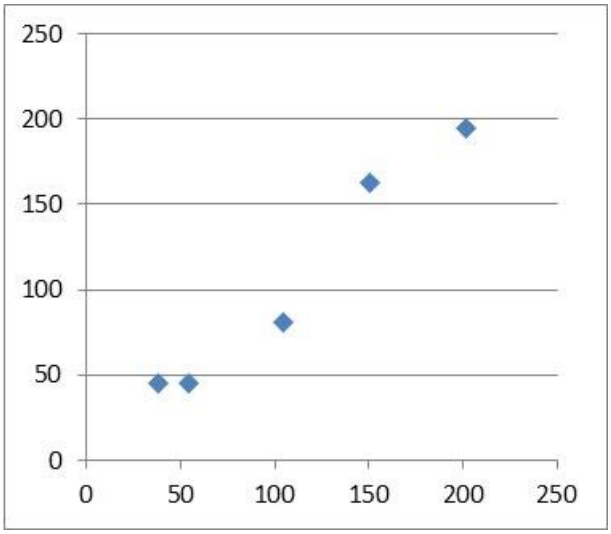
鍵語	1 Madrid	2 Sevilla
a	151	163
con	38	45
de	202	195
en	105	81

por	54	45

ここには「文 1」と「文 2」という 2 つのデータがあります。この 2 つの文は前置詞の観点からみると、どの程度類似しているのでしょうか。本節ではこのような 2 つのデータの関連の強度を計算する方法を見ていきます。

◇2 データ間の関係

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。◆Excel では、英文 1 英文 2 の 2 列を選択し、「挿入」→「グラフ」→「散布図」とします。軸ラベルがあるレイアウトに変更し、それぞれ軸ラベルを編集しておきます。



一見したところ、文 1 と文 2 は比例関係があるようです。この事実を確かめるために横軸の原点（ゼロの位置）を英文 1 の平均までずらし、縦軸の原点を英文 2 の平均までずらして散布図を描き直してみましょう。そのためには、前章で見た標準得点が使われます。これは次のように定義されます。

$$\text{標準得点(SM)} = \frac{x - \text{X の平均値}}{\text{X の標準偏差}}$$

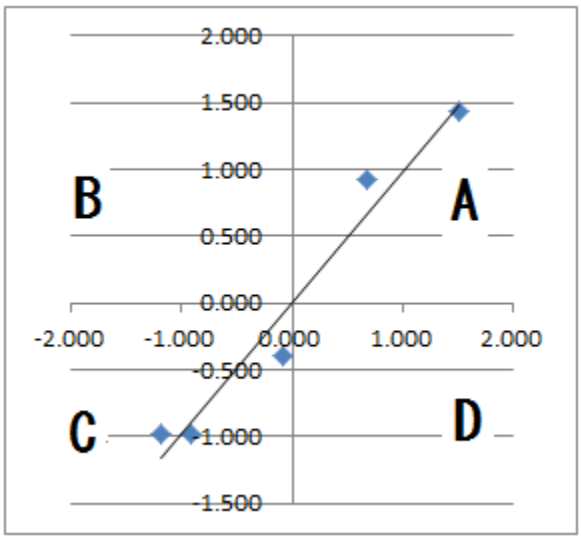
このように標準得点（SM_i）はそれぞれ元の値（x_i）から全体の平均値（m）を引いて、さらにその値を全体の標準偏差（σ）で割って得られた数値です。

次は、このデータを標準得点に置き換えた結果です。つまり全体の平均が

0、標準偏差が 1 になるようにしたものです。

v-1(sm)	v-2(sm)	v-1(sm)
0.674	0.922	0.674
-1.184	-0.980	-1.184
1.513	1.438	1.513
-0.082	-0.400	-0.082
-0.921	-0.980	-0.921

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。



この図を見れば、文 1 と文 2 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、x 軸の値と y 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。

◇3 相関係数

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 の関連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の斜めの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[16]。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、積和を全体の個数で割って積和の平均を出したものが「相関係数」(coefficient of correlation)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。

Xと Y の相関係数（r）

$$
\begin{aligned}
= \{ & [(x_1 - m_x) / \sigma_x] [(y_1 - m_y) / \sigma_y] \\
& + [(x_2 - m_x) / \sigma_x] [(y_2 - m_y) / \sigma_y] \\
& (\ldots) \\
& + [(x_n - m_x) / \sigma_x] [(y_n - m_y) / \sigma_y] \} / n
\end{aligned}
$$

という計算をします。σ_x と σ_y を分母に移すと、

$$
\begin{aligned}
r = [& (x_1 - m_x)(y_1 - m_y) \\
& + (x_2 - m_x) (y_2 - m_y) \\
& + (\ldots) \\
& + (x_n - m_x) (y_n - m_y)] / (n\sigma_x\sigma_y)
\end{aligned}
$$

ここで、

$$
\begin{aligned}
[& (x_1 - m_x)(y_1 - m_y) \\
& + (x_2 - m_x) (y_2 - m_y) \\
& + (\ldots) \\
& + (x_n - m_x) (y_n - m_y)] / n
\end{aligned}
$$

を「共分散」(covariance)と呼び、S_{xy} と書きます。すると先の式は、

$$\text{X と Y の相関係数（}r\text{）} = \frac{S_{xy}}{\sigma_x * \sigma_y}$$

---

[16] 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。

となります。ＸとＹの相関係数 (r) は最終的に

$$\text{相関係数 } (r) = \frac{\text{XとYの共分散}}{\text{Xの標準偏差 * Yの標準偏差}}$$

となります。

相関係数(r)が$-1 \leqq r \leqq 1$になる理由

相関係数は$-1 \leqq r \leqq 1$という範囲に入る標準的な値です。このことを高校数学までに習った判別式を使って確かめてみましょう。

原理的に、相関係数はすべてのデータが一直線に並ぶときに最大になりますから、そのような直線の式を

$$(y - m_y) = a\,(x - m_x)$$

で表します。ここで、m_xと m_yはそれぞれ x と y の平均値を示します。この直線は X と Y の平均値の座標$(m_x,\ m_y)$を通り、傾きは a となります。直線ならば、上の式から、

$$a\,(x - m_x) - (y - m_y) = 0$$

となりますが、実際のデータでは、(x, y)のそれぞれの値、つまり、(x_1, y_1), (x_2, y_2), ... (x_n, y_n)が直線上に並ぶことはふつうありません。その値を(x_i, y_i)として、上の式に当てはめると、$a\,(x_i - m_x) - (y_i - m_y)$ はゼロ(0)ではなくて、プラスになったり、マイナスになったり、さまざまな値をとります。その全体の変動を見るために、その自乗和を計算しましょう。

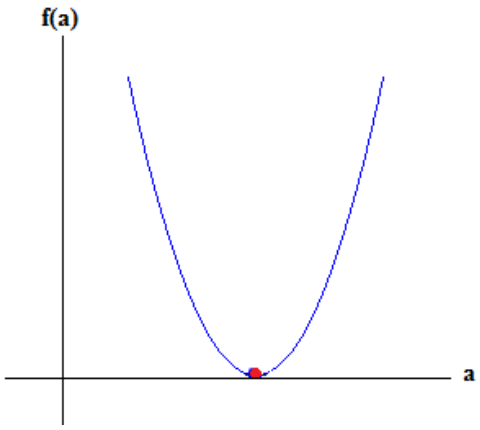
$$f(a) = \Sigma\,[a\,(x_i - m_x) - (y_i - m_y)]^2$$

これは平方和なので負（マイナス）になることはありません。つまり、$f(a) \geqq 0$ です。f(a)を展開しましょう。

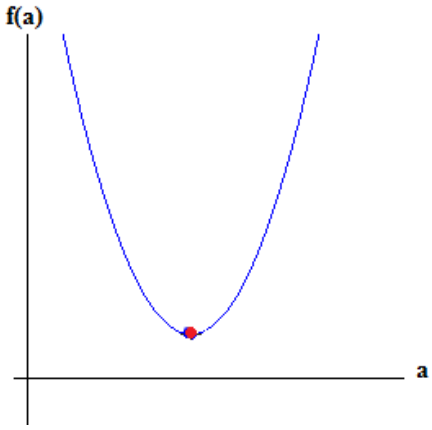
$$\begin{aligned} f(a) &= \Sigma\,[a^2(x_i - m_x)^2 - 2a(x_i - m_x)(y_i - m_y) + (y_i - m_y)^2] \\ &= \Sigma\,a^2(x_i - m_x)^2 - \Sigma\,2a(x_i - m_x)(y_i - m_y) + \Sigma\,(y_i - m_y)^2 \\ &= a^2\,\Sigma\,(x_i - m_x)^2 - 2a\,\Sigma\,(x_i - m_x)(y_i - m_y) + \Sigma\,(y_i - m_y)^2 \end{aligned}$$

このようにf(a)は a の 2 次式になりますが、先に見たようにf(a)≧0 なので、2 次式の放物線の頂点が横軸に接するか、またはその上方にあること

になります[17]。a を横軸に、f(a)を縦軸にしたグラフを描いてみましょう。



【図 6.1e】



【図 6.1f】

f(a)の放物線の頂点がちょうどを横軸上にあるときは（【図 6.1e】）、次の判別式がゼロとなって、解が 1 つになります。放物線の頂点が横軸よりも上にあるときは（【図 6.1f】）解がないので（横軸とぶつからないので）判別式はマイナスになります[18]。

判別式(D) ≦ 0

これを f(a)の式に当てはめます。

$$[2\Sigma(x_i - m_x)(y_i - m_y)]^2 - 4\Sigma(x_i - m_x)^2\Sigma(y_i - m_y)^2 \leqq 0$$
$$[\Sigma(x_i - m_x)(y_i - m_y)]^2 - \Sigma(x_i - m_x)^2\Sigma(y_i - m_y)^2 \leqq 0$$

上の式のそれぞれの要素は、相関係数で使われた要素と同じであることに気づきます。そこで第 2 項を右辺に移動します。

$$[\Sigma(x_i - m_x)(y_i - m_y)]^2 \leqq \Sigma(x_i - m_x)^2\Sigma(y_i - m_y)^2$$

さらに両辺を右辺で割ります。

$$\frac{[\Sigma(x_i - m_x)(y_i - m_y)]^2}{\Sigma(x_i - m_x)^2\Sigma(y_i - m_y)^2} \leqq 1$$

この左辺は、相関係数(r)を自乗したものですから、$r^2 \leqq 1$ となり、よっ

---

[17] ここで「横軸」と言い x 軸と言わないのは、上の 2 次式は x についての 2 次式というよりも、a についての 2 次式を考えているからです。よって「横軸」は「a 軸」のことです。

[18] 2 次方程式 $ax^2 + bx + c = 0$ の判別式(D)は $b^2 - 4ac$ です。

て

$$-1 \leqq r \leqq 1$$

となります。

◇4 相関係数の意味

出力された数値について経験的に次のような解釈できます[19]。

$\|r\| = 0.0$	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

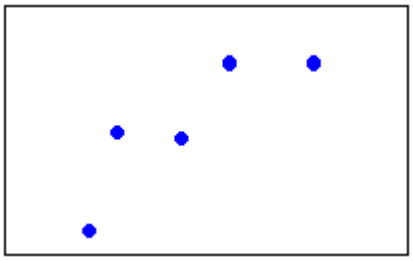
◇5 相関係数についての注意

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。

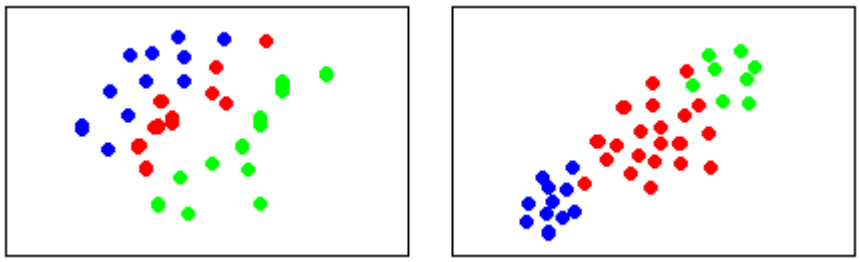
(0) そもそも 2 つが同じデータの場合。たとえば、値とその百分率（%）は　まったく同じデータです。

(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。

[19] 相関係数の範囲は-1≦r≦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。

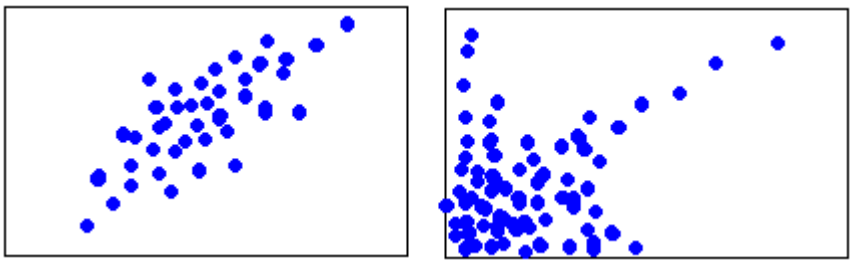


(2) 異質なデータが混在している場合。全く異なるデータを寄せ集めて相関係数を求めると、現象の正しい解釈ができないことがあります。



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[20]。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。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。

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

[20] 先のスペイン語教材のアンケート調査結果がこれと似ています。

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[21]。

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結果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、自然と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、一致していないときは何の要因がありうるか考えてみる必要があるでしょう。

◇6 Excelで相関係数を求める

(1) 次のデータを使用します（前節と同じものです）。

鍵語	1 Madrid	2 Sevilla
a	151	163
con	38	45
de	202	195
en	105	81
por	54	45

(2) 次の計算をします。

[21] 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。

- B7 =SUM(B2:B6)
- B8 =AVERAGE(B2:B6)
- B9 =STDEVP(B2:B6)

(3) B7:B9 をコピーし、C7 に貼付けます。

	A	B	C
1	**鍵語**	1 Madrid	2 Sevilla
2	**a**	151	163
3	**con**	38	45
4	**de**	202	195
5	**en**	105	81
6	**por**	54	45
7	SUM	550	529
8	MEAN	110.000	105.800
9	SD	60.811	62.027

(4) D2 に標準得点の式を入れます。

D2 =(B2-B$8)/B$9

(5) D2 を D2:E6 にコピー。桁数が不統一だと比較しにくいので D, E 列の書式を小数点以下 3 とします。

	A	B	C	D	E
1	**鍵語**	1 Madrid	2 Sevilla	v-1(sm)	v-2(sm)
2	**a**	151	163	0.674	0.922
3	**con**	38	45	-1.184	-0.980
4	**de**	202	195	1.513	1.438
5	**en**	105	81	-0.082	-0.400
6	**por**	54	45	-0.921	-0.980

(6) B7:C9 をコピーして D7 に貼付けます。

	A	B	C	D	E
1	**鍵語**	1 Madrid	2 Sevilla	v-1(sm)	v-2(sm)
2	**a**	151	163	0.674	0.922
3	**con**	38	45	-1.184	-0.980
4	**de**	202	195	1.513	1.438
5	**en**	105	81	-0.082	-0.400
6	**por**	54	45	-0.921	-0.980
7	SUM	550	529	0	0
8	MEAN	110.000	105.800	0.000	0.000
9	SD	60.811	62.027	1.000	1.000

これで正しく標準化されたことがわかります。次に、これらの数値をもとに相関係数を求めてみましょう。まず、それぞれの項目の標準得点の積と全体の積平均を求めます。

F2 =D2*E2

	A	B	C	D	E	F
1	鑵語	1 Madrid	2 Sevilla	v-1(sm)	v-2(sm)	積和
2	a	151	163	0.674	0.922	0.622
3	con	38	45	-1.184	-0.980	1.161
4	de	202	195	1.513	1.438	2.176
5	en	105	81	-0.082	-0.400	0.033
6	por	54	45	-0.921	-0.980	0.903
7	SUM	550	529	0	0	4.894
8	MEAN	110.000	105.800	0.000	0.000	0.979

F2 = D2*E2

F2 を(F3:F6)にコピー

(B7:B8)を(F7:F8)にコピー

これで標準得点をもとに相関係数を求めることができました。

結果を確認するために、Excel 関数を使って相関係数を算出し比較してみましょう。Excel には COREEL という関数が用意されており、対象となる 2 つのデータをコンマ区切りで選択します。

B10 =CORREL(B2:B6,C2:C6)

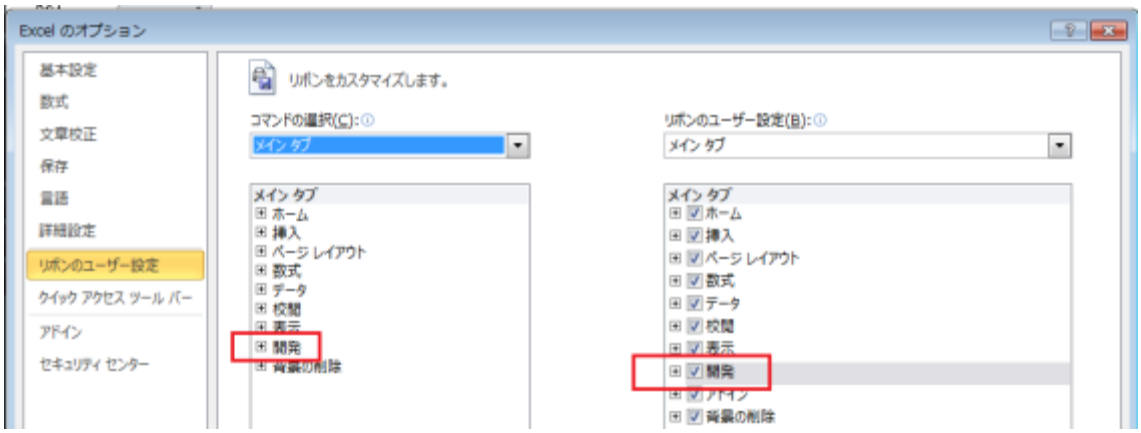
B10 =CORREL(B2:B6,C2:C6)

	A	B	C	D	E	F
1	鑵語	1 Madrid	2 Sevilla	v-1(sm)	v-2(sm)	積和
2	a	151	163	0.674	0.922	0.622
3	con	38	45	-1.184	-0.980	1.161
4	de	202	195	1.513	1.438	2.176
5	en	105	81	-0.082	-0.400	0.033
6	por	54	45	-0.921	-0.980	0.903
7	SUM	550	529	0	0	4.894
8	MEAN	110.000	105.800	0.000	0.000	0.979
9	SD	60.811	62.027	1.000	1.000	
10	Correl	0.979				

F8 と B10 の値が同じになることを確認しましょう。

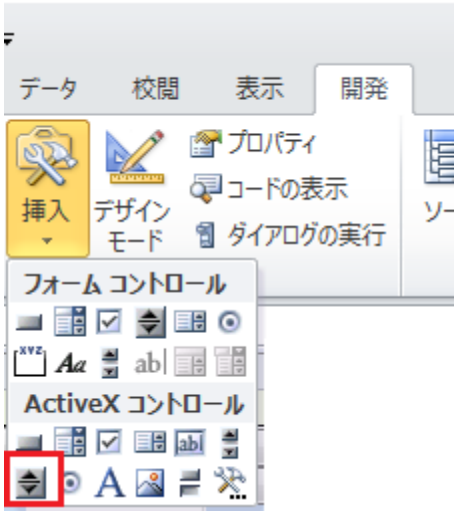
<Tips> それぞれの特徴を見るために値を操作するのに「スピンボタン」を使うと便利です。

(1) はじめにリボンに「開発」タブを設定します。◆「ファイル」→「オプション」→「リボンのユーザー設定」→]を選択し、「リボンのユーザー設定」で「メインタブ」の「開発」のチェックボックスをオンにします。

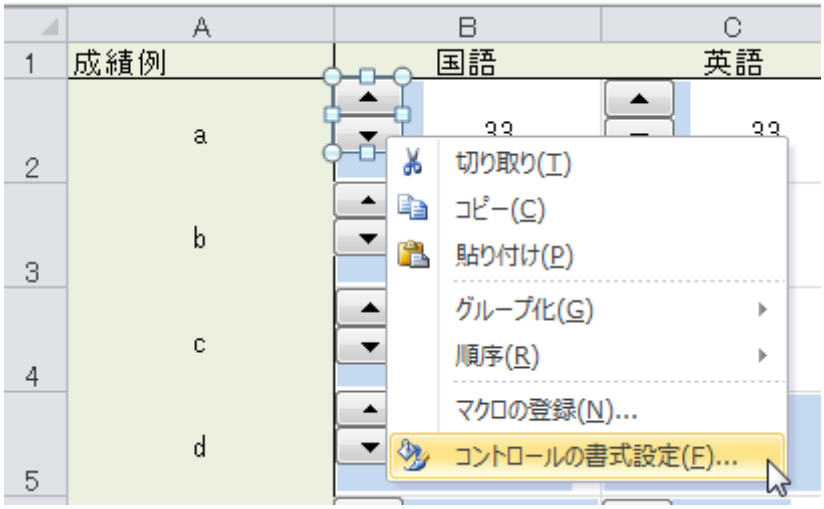


Excel 2007：「Office ボタン」→「Excel のオプション」→「基本設定」→「[開発]タブをリボンに表示する」をチェック

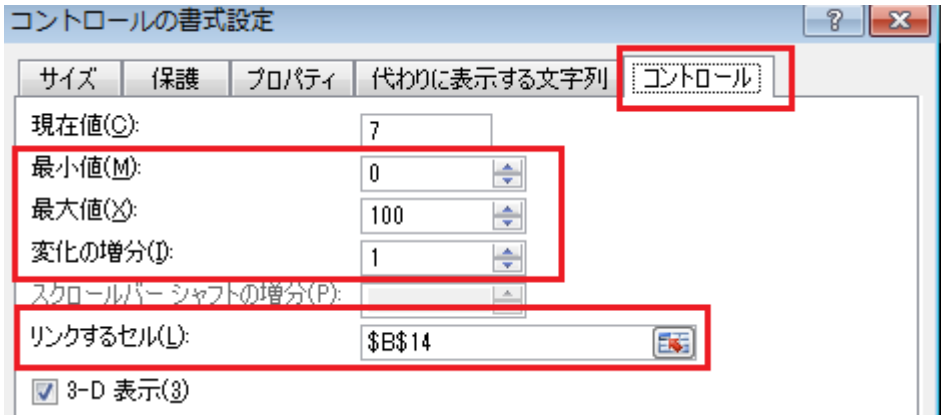
(2)「開発」→「コントロール」→「挿入」→「フォームコントロール」の中のスピンボタンをクリック→シート内の適当な位置にドラッグして配置します。



(3) シートに配置したスピボタンを右クリック→「コントロールの書式設定」



(4) 「コントロール」タブ→「最小値」「最大値」「変化の増分」「リンクするセル」を設定します。「リンクするセル」にスピンボタンによる入力の結果が表示されます。



(5) スピンボタンなどのコントロールは右クリックすることにより、大きさの変更、ドラッグ、コピー、などが可能になります。

スピンボタンは便利なのですが、たとえば 1 から 100 まで移動するときは大変です。スピンボタンをつけたらそれでしか値が操作できなくなるというわけではなく、直接セルに 100 と記入することもできます。

一人称的な研究

私たちは、言語を単なる言語分析用のデータと見ているのではなく、言語作品を鑑賞したり、ことばの伝え合いや共有を経験したり、未知の外国語を学んだりして、言語を生活の中で経験しています。そのとき、感じたり気づいたりすることがあるはずです。言語の現実に触れたときに私たちの内面に

生じる直感や気づきがとても大切です。

言語データ分析は、そのような直感や気づきの「理由」や「姿」を具体的なデータで調べてみるときに役立ちます。このとき言語の経験が最初で、分析はその後になります。自分が経験していることを対象にして分析するときは、何か直感的にぴんと来ることが多いと思います。そこで、なるべく自分で経験した（読んだ、集めた、調べた、実験した、使った、感動した、興味を持った…）言語データから出発して、自分が理解し納得できた方法を適用して、自分の個人的な直感を検証してみることを勧めます。

実際に自分の研究を自分で計画し、試行錯誤をしながら自分の道具を開発し、自分で納得し、自分が個人的に感じたことの理由に接近できれば発展性があるし、何よりもやりがいがあって楽しいことだと思います。このようば研究は「一人称的」であるといえるでしょう。私たちは他者の（本当の）一人称的世界に関心がありますから、そのような他者の関心と研究にも共感します。

6. 2 質的なデータの相関

言語資料を分析するとき頻度などの連続的な数量を扱うこともありますが、プラス・マイナス（＋/－）で示されるような特定の特徴の有無だけを問題にすることもあります。たとえば、「ぬくい」（温かい）、「おとろしい」（面倒くさい）という言葉がある地域で使われるかどうかといったことを扱う場合は、「使う」「使わない」の 2 値のデータになります。これらの語彙リストを作り、使用の有無から地域の関連性を求めるという研究はよく見られます。このような種類のデータ分析には相関係数ではなく今回扱う各種の「類似係数」が適しています[22]。

---

[22] ＊参考：

Ellegard, Alvar. 1959. "Statistical measurement of linguistic relationship." *Language,* 35, p. 131-156.

Kroeber, Alfred L. 1960. "Three quantitative classifications of Romance", *Romance Philology,* 14, pp.189-195.

Kroeber, Alfred L. and Chretien, C. D. 1937. "Quantitative classification of Indo-european languages", *Language,* 13, p.83.

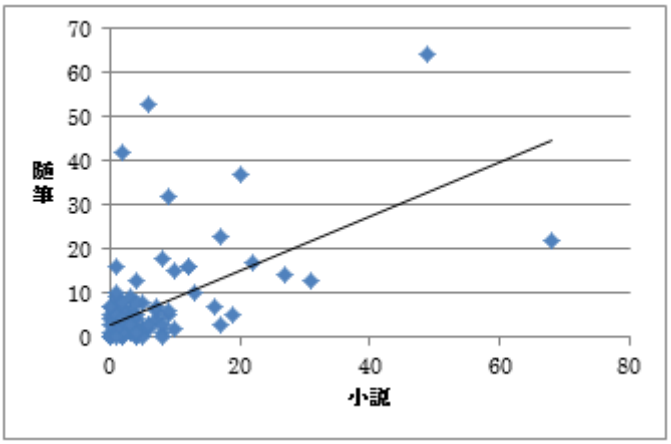
Kroeber, Alfred L. and Chretien, C. D. 1960. "Statistics, Indo-European and taxonomy." *Language,* 36, p. 1-21.

安本美典. 1995.『言語の科学―日本語の起源をたずねる』朝倉書店.

特徴があることだけでなく、それがないことも考慮に入れなければならない場合もあります（つまり、「使われない」ということもその地域を表すデータとなります）。言語の現象に限らず、私たちの日常生活では特徴がある事象（祝祭、病気、事故、降雨など）に注目することが多いのですが、その特徴がないときのことも考えないと、その特徴の本質がわからなくなる場合があります。

◇1 量的データと質的データ

先に見たように、単語の頻得点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。



ここではすべてを単語使用の「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。次の図の「語」の列に続く 2 列が頻度を示しますが、その後の 2 列では 1 が「有」を示し、0 が「無」を示します。頻得点などのような連続的なデータを「量的なデータ」と呼び、このように単に有・無を示すようなデータを「質的なデータ」と呼びます。

（pp.67-85）

語	男性	女性	男性	女性
abajo	5	10	1	1
abandonar	9	6	1	1
abandono	0	0	0	0
abarcar	1	0	1	0
abastecimiento	2	0	1	0
abatir	0	1	0	1
abeja	2	3	1	1
abertura	0	0	0	0
abismo	0	0	0	0
abnegación	0	0	0	0
abogado	3	6	1	1
abonar	5	0	1	0
abono	0	0	0	0
abordar	0	0	0	0
aborrecer	0	6	0	1

言語研究では、たった一度だけ出現するデータを特別に扱うことが一般的です。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 以上を「有」(1)のデータとして基準化する場合が多いです。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があります。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。

◇2 尺度水準

これまでの説明で、質的データ、量的データという２つのタイプに大別しました。２つのデータの大きな特徴は、量的データは質的データに変換可能であるのに対し、質的データは量的データに戻すことは出来ないという点です。こうした質的データと量的データの特徴は、スタンレー[23]によって考案された、「尺度水準」[24]という考え方におおよそ準拠したものです。

尺度水準という考え方に基づけば、すべての数量データは「名義尺度」「順序尺度」「間隔尺度」「比率尺度」という 4 つのタイプのいずれかに分類できます。名義尺度に使用される値は、名前をそのまま数字に置き換え

[23] Stevens, S. S. 1946. “On the Theory of Scales of Measurement”. *Science*. Vol. 103, No. 2684, pp. 677–680.

[24] 尺度水準という考え方は、言語分析に限らず、その他の分野でも広く使われる考え方です。

たものであり、そのデータが、別のデータと同じか、違うかを区別するために割り当てられた数値です。例えば、電話番号は名義尺度であるため、ある番号が、他の番号と同じ番号か、違う番号かを区別するために使用します。

順序尺度の値は、データが大きいか、小さいかを区別するための数値です。例えば、アンケート調査の「好き」「まあまあ好き」「どちらとも言えない」「あまり好きではない」「好きではない」という項目に対し、5, 4, 3, 2, 1 という数値を割り振る場合が順序尺度です。つまり、このとき、数値の中で、４の方が１よりも好きの度合いが優位だとわかります。

間隔尺度の値は、比較できる数値で、一般的には単位を持った値です。例えば、摂氏の温度において、20℃と 18℃を比較したとき、2℃高かった、２℃低かったという間隔を持った値であるため、間隔尺度です。間隔尺度の特徴としては、ゼロという値が本来的な全く存在しないものという意味ではないという点です。たとえば、0℃という値でも、摂氏という温度自体が消えてなくなるわけではなく、0℃が 5℃よりも 5℃低いという便宜上の値です。

比率尺度の値は、比較可能な数値であり、単位を持つという点は間隔尺度の値と同じですが、ゼロになってしまうとそのデータ自体が全く意味を持たなくなるものです。例えば、質量は何グラム増えた、減ったということを判断できますが、これが、0 グラムになると質量というもの自体がなくなります。

このような 4 つの尺度に分けるメリットのひとつは、数値分析できる幅がそれぞれ異なるという点です。名義尺度、順序尺度、間隔尺度、比率尺度の順に、データとして求められた値の数値分析可能な幅が広がっていきます。数値分析が限られたものにしか適応出来ないものを「低水準」、幅広く適応できるものを「高水準」と呼ぶこともあります。そうすると名義尺度は低水準なのに対して、比率尺度は高水準であるということになります。例えば、得点（頻度）は非常に幅広い尺度に適応でき、名義尺度、順序尺度、間隔尺度、比率尺度のいずれにも適応可能です。中央値、最大値、最小値は、順序尺度、間隔尺度、比率尺度に適応できます。和、平均、標準偏差、相関係数は、間隔尺度、比率尺度に対して適応されます。それ以外の複雑な数値分析であっても、比率尺度であれば適応可能である、ということになります。

また、このような尺度を設けるメリットとしては、それぞれの変換可能な方向性があるということです。つまり、高水準なものは低水準なものと

して扱うことができますが、低水準なものは高水準なものとして扱うことはできません。

ここで、４つの尺度と、言語分析における質的・量的データの関係性を整理しておきましょう。一般には、名義尺度と順序尺度は「質的データ」であり、間隔尺度と比率尺度は「量的データ」であるとされます。それは、質的データと量的データの変換方向性によるものからも明らかです。ただし、数値分析可能な範囲が、質的データと量的データのどこまでできるかについては、きれいに対応関係は成立していない場合もあるので注意が必要です。実際に分析するときに、質的データと量的データで数値データを扱い、その関係性が明らかでないときには、上記の 4 つの尺度水準に立ち返ることでそれが何の分析まで行っていいかの方針を決めることができるでしょう。

TIPS 尺度水準と代表値の関係をまとめると次のようになります。×のところは、該当の代表値がその尺度では使えないことを示します。

尺度と代表値		得点	中央値	平均	標準偏差
質的データ	名義尺度	○	×	×	×
	順序尺度	○	○	×	×
量的データ	間隔尺度	○	○	○	○
	比率尺度	○	○	○	○

◇3 四象限と類似係数

2 つのデータの間の関係を見るときに目安になるのが共通して「有」(=1)が起きる回数です。たとえば、先の図では「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。これを「共起回数」と呼びます。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」）が類似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。

単純に共起回数だけでは相対化できないので、次のような 2 × 2 の表を作り、それぞれ a, b, c, d の 4 つを考慮します。a, b, c, d のそれぞれは、高校数学までに習った四象限（quadrants）で示せば、順に第Ⅰ象限(+/+)、第Ⅱ象限(+/-)、第Ⅲ象限(-/+)、第Ⅳ象限(-/-)に相当する値です。a は x も y も「有」(=1)の個数です。b は x が「有」(=1)かつ 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 *a*=4 {*abajo*, *abandonar*, *abeja*, *abogado*}, *b*=3 {*abarcar*, *abastecimiento*, *abonar*}, *c*=2 {*abatir*, *aborrecer*}, *d*=6 {*abandono*, *abertura*, *abismo*, *abnegación*, *abono*, *abordar*}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

類似係数はこれらの数値(*a*, *b*, *c*, *d*)を利用します。*d* を使わない係数もあります。類似度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、類似係数は大きくなる、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、類似係数は小さくなります。以下の７つは、その類似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) はじめに**単純一致係数**(simple matching coefficient)をみましょう。

$$\text{単純一致係数 (s.)} = \frac{a + d}{a + b + c + d} \qquad 0.0 \leqq \text{s.} \leqq 1.0$$

これは、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。

(2) **Russel and Rao 係数**は分子の d を考慮しません[25]。対象 X, Y でともに「＋」である回数だけをカウントします。分母は(1)と同じです。a = 0 のとき最小値 0 になり、b = c = d = 0 のとき最大値 1 になります。

$$\text{Russel and Rao 係数(r.r.)} = \frac{a}{a + b + c + d} \qquad 0.0 \leqq \text{r.r.} \leqq 1.0$$

(3) **Jaccard 係数**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。

$$\text{Jaccard 係数(j.)} = \frac{a}{a + b + c} \qquad 0.0 \leqq \text{j.} \leqq 1.0$$

[25] d の数値の扱い方については、この後説明します。

(4) **Dice 係数**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 1.0 になります。(→後述)

$$\text{Dice 係数(d.)} = \frac{2a}{2a + b + c} \qquad 0.0 \leqq \text{j.} \leqq 1.0$$

(5) **Yule 係数**は ad と bc の差を問題にします。(1)の単純一致係数では a と d を足していますが、Yule 係数では掛けることになります。それから分子は ad と bc の差なので、それがマイナスになることもあります。ad = 0 のとき最小値 -1 になり、bc = 0 のとき最大値 1 になります。ad = bc のときは最小値と最大値の中間 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。

$$\text{Yule 係数(y.)} = \frac{ad - bc}{ad + bc} \qquad -1.0 \leqq \text{y.} \leqq 1.0$$

(6) **Hamann 係数**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 のときは最小値と最大値の中間 0 になります。

$$\text{Hamann 係数 (h.)} = \frac{(a + d) - (b + c)}{(a + d) + (b + c)} \qquad -1.0 \leqq h. \leqq 1.0$$

(7) **Phi 係数**は少し複雑な式です。これは積率相関係数と関係します。(→後述)

$$\text{Phi 係数 (ph.)} = \frac{ad - bc}{\sqrt{(a + b)(a + c)(b + d)(c + d)}} \quad -1.0 \leqq \text{Phi} \leqq 1.0$$

(8) **Ochiai 係数**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。

$$\text{Ochiai 係数(o.)} = \frac{a}{\sqrt{(a + b)(a + c)}} \qquad 0.0 \leqq \text{o.} \leqq 1.0$$

●積率相関係数と 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(ピアソンの積率相関係数)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	a + b + c + d

はじめに総データ数を n とします。

$$n = a + b + c + d$$

先に見たように相関係数（標準得点の積和の平均）の式は次の通りです。

$$\begin{aligned} r = \quad & \{ [(x_1 - m_x) / \sigma_x][(y_1 - m_y) / \sigma_y] \\ & + [(x_2 - m_x) / \sigma_x][(y_2 - m_y) / \sigma_y] \\ & (\ldots) \\ & + [(x_n - m_x) / \sigma_x][(y_n - m_y) / \sigma_y] \} / n \end{aligned}$$

σ_x と σ_y を分母に移すと

$$\begin{aligned} r \quad = [& (x_1 - m_x)(y_1 - m_y) \\ & + (x_2 - m_x)(y_2 - m_y) \\ & + (\ldots) \\ & + (x_n - m_x)(y_n - m_y)] / (\sigma_x \sigma_y n) \quad \cdots \ ① \end{aligned}$$

先に①の分子だけを取り上げましょう。

$$\begin{aligned} r_{分子} \quad = & (x_1 - m_x)(y_1 - m_y) \\ & + (x_2 - m_x)(y_2 - m_y) \\ & + (\ldots) \\ & + (x_n - m_x)(y_n - m_y) \end{aligned}$$

それぞれ展開して

$$\begin{aligned} r_{分子} \quad = & (x_1y_1 - x_1m_y - m_xy_1 + m_xm_y) \\ & + (x_2y_2 - x_2m_y - m_xy_2 + m_xm_y) \\ & + (\ldots) \\ & + (x_ny_n - x_nm_y - m_xy_n + m_xm_y) \\ & \ \vdots \quad \ \vdots \quad \ \vdots \quad \ \vdots \\ & (1) \quad (2) \quad (3) \quad (4) \end{aligned}$$

縦の列をまとめて、

$r_{分子} = (x_1y_1 + x_2y_2 + \ldots + x_ny_n)$ …(1)

$- m_y(x_1 + x_2 + \ldots + x_n)$ …(2)

$- m_x(y_1 + y_2 + \ldots + y_n)$ …(3)

$+ nm_xm_y$ …(4)

ここで、(1) $x_1y_1 + x_2y_2 + \ldots + x_ny_n$ 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。それで結局は

$$x_1y_1 + x_2y_2 + \ldots + x_ny_n = a$$

となります。また

$x_1 + x_2 + \ldots + x_n = a + b$ … Ｘの総和

$y_1 + y_2 + \ldots + y_n = a + c$ … Ｙの総和

$m_x = (a + b) / n$ … Ｘの平均

$m_y = (a + c) / n$ … Ｙの平均

となるので分子は

$r_{分子} = a$ …(1)

$- (a + b)(a + c) / n$ …(2)

$- (a + b)(a + c) / n$ …(3)

$+ (a + b)(a + c) / n$ …(4)

$= a - (a + b)(a + c) / n$

$= [na - (a + b)(a + c)] / n$

$n = a + b + c + d$ なので

$r_{分子} = [(a + b + c + d)a - (aa + ac + ba + bc)] / n$

$= (aa + ab + ac + ad - aa - ac - ab - bc) / n$

$= (ad - bc) / n$ …②

となります。この分子の式はＸとＹに共にある場合の数(a)と、共にない場合の数(d)の積から、片方にしかない２つの場合の数(b と c)の積を引いたものです。a も d もＸとＹのプラス・マイナスが同じ場合です。逆に、b と c はＸとＹのプラス・マイナスが反対になる場合だから、ad - bc がＸとＹの相関を示すのに合理的な数値に関わることが直感的に納得できます。

次に①の分母を $r_{分母}$ とします。

$r_{分母} = \sigma_x \sigma_y n$

$r_{分母}$のうちのXの標準偏差 $\boldsymbol{\sigma_x}$ を取り上げましょう。ルート（根）があるとややこしくなるので、とりあえず２乗したもの（つまり、$\boldsymbol{\sigma_x^2}$ なので分散値）で計算し、後でその根を計算します。

$$\begin{aligned}\sigma_x^2 &= [\,(x_1 - m_x)^2 \\ &\quad + (x_2 - m_x)^2 \\ &\quad + \cdots \\ &\quad + (x_n - m_x)^2\,] / n\end{aligned}$$

それぞれの項を展開して、

$$\begin{array}{llll}\sigma_x^2 & = [(x_1^2 & - 2x_1 m_x & + m_x^2) \\ & + (x_2^2 & - 2x_2 m_x & + m_x^2) \\ & + \cdots & & \\ & + (x_n^2 & - 2x_n m_x & + m_x^2)] / n \\ & : & : & : \\ & (1) & (2) & (3)\end{array}$$

縦の列をまとめて、

$$\begin{aligned}\sigma_x^2 &= [(x_1^2 + x_2^2 + \cdots + x_n^2) && \cdots (1) \\ &\quad - 2m_x(x_1 + x_2 + \cdots + x_n) && \cdots (2) \\ &\quad + (m_x^2 + m_x^2 + \cdots + m_x^2] / n && \cdots (3)\end{aligned}$$

$$\begin{aligned}&= [(x_1^2 + x_2^2 + \cdots + x_n^2) && \cdots (1) \\ &\quad - 2m_x(x_1 + x_2 + \cdots + x_n) && \cdots (2) \\ &\quad + nm_x^2] / n && \cdots (3)\end{aligned}$$

$x_1, x_2, \ldots x_n$ はすべて 1 または 0 です。そこで X の総数は a + b となるので（【図 3.3d】）、次のようになります。

$$x_1 + x_2 + \cdots + x_n = a + b$$
$$x_1^2 + x_2^2 + \cdots + x_n^2 = a + b$$
$$m_x = (a + b) / n$$

これを先の式に代入すると、

$$\sigma_x^2 = [(a + b) \qquad \cdots(1)$$

$$- 2(a + b)^2 / n \quad \ldots(2)$$
$$+ n(a + b)^2 / n^2] / n \quad \ldots(3)$$

$$= \{(a + b) - [2(a + b)^2 + (a + b)^2] / n\} / n$$
$$= [a + b - (a + b)^2 / n] / n$$
$$= [(a + b)n - (a + b)^2] / n^2$$
$$= [(a + b)(a + b + c + d) - (a + b)^2] / n^2$$
$$= (a + b)(c + d) / n^2$$

ここで、σ_x^2 から σ_x に戻します[26]。

$$X\text{の標準偏差 } \sigma_x = \sqrt{(a + b)(c + d)} / n \quad \ldots③$$

同様にして、$r_{分母}$ の σ_y を求めます。

$$\sigma_y^2 = [(y_1 - m_y)^2 + (y_2 - m_y)^2 + \ldots + (y_n - m_y)^2] / n$$
$$= [(y_1^2 - 2y_1m_y + m_y^2) + (y_2^2 - 2y_2m_y + m_y^2) + \ldots + (y_n^2 - 2y_nm_y + m_y^2)] / n$$
$$= [(y_1^2 + y_2^2 + \ldots + y_n^2) - 2m_y(y_1 + y_2 + \ldots + y_n) + nm_y^2] / n$$
$$= [(a + c) - 2(a + c)^2 / n + n(a + c)^2 / n^2] / n$$
$$= (a + c)(b + d) / n^2$$

σ_y^2 も σ_y に戻します。

$$Y\text{の標準偏差 } \sigma_y = \sqrt{(a + c)(b + d)} / n \quad \ldots④$$

上記①に、②と③④を代入すれば、こうして数値が 0 と 1 だけのデータの相関係数（Phi 係数：Phi）は全体で次のようになります。

$$Phi = \frac{(ad - bc) / n}{n\sqrt{(a + b)(c + d)} / n * \sqrt{(a + c)(b + d)} / n}$$

$$= \frac{ad - bc}{\sqrt{(a + b)(a + c)(c + d)(b + d)}}$$

分母は(a, d)と(b, c)をそれぞれ組み合わせて和としたものを全部掛け合わせています。

●Phi 係数と Ochiai 係数

理論的に導き出された Phi 係数を実際に適用してみると不都合なときがあ

[26] つまり、分散値を標準偏差に戻します。

ります。次のデータを比べてみましょう。

	y (+)	y (−)	和
x (+)	100	10	110
x (−)	20	2	22
和	120	12	132

データ (1)

	y (+)	y (−)	和
x (+)	4	10	14
x (−)	20	50	70
和	24	60	84

データ (2)

ここでそれぞれの phi 係数を求めてみます。Phi (1)はデータ(1)、Phi (2)はデータ(2)の Phi 係数です。

$$\text{Phi (1)} = \frac{100\times 2 - 10\times 20}{\sqrt{(100+10)\times(100+20)\times(20+2)\times(10+2)}}$$

$$= \frac{0}{\sqrt{(100+10)\times(100+20)\times(20+2)\times(10+2)}} = 0$$

$$\text{Phi (2)} = \frac{4\times 50 - 10\times 20}{\sqrt{(4+10)\times(4+20)\times(20+50)\times(10+50)}}$$

$$= \frac{0}{\sqrt{(4+10)\times(4+20)\times(20+50)\times(10+50)}} = 0$$

どちらも Phi 係数の分子の ad－bc がゼロとなるので、Phi 係数もゼロになります。しかし、データ(1)とデータ(2)を比べれば(1)のほうがずっと類似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(2)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(0-0)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（本が無限に出版されるとして）だと考えられます。

そこで、先の式で 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。

$$\text{Phi'.} = \lim_{d\to\infty} \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}$$

分母と分子を d で割ります。

$$\text{Phi'} = \lim_{d \to \infty} \frac{a - bc/d}{\sqrt{(a + b)(a + c)(b/d + 1)(c/d + 1)}}$$

それぞれの分母になる d を無限大にすると、分子に何があってもゼロとなります。

$$\text{Phi'} = \frac{a}{\sqrt{(a + b)(a + c)}}$$

これが Phi 係数の修正版（*Ochiai* 係数: *ochi.*）です。とてもシンプルになりました。先のデータ(1), (2)で計算してみましょう。

$$\text{Phi' (1)} = \frac{100}{\sqrt{(100+10)(100+20)}} = 0.870$$

$$\text{Phi'. (2)} = \frac{4}{\sqrt{(4+10)(4+20)}} = 0.218$$

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(1)の方がデータ(2)よりも類似性が高いという直感を裏付けることができます。

●相互情報量と Dice 係数

言語研究ではたとえば 2 つの語の結合度を調べるために、相互情報量という数値を使います。これあｈ、共起得点(a)をデータ全体で理論的に期待できる共起得点（期待値）で割った値の対数（底=2）です。

$$\text{相互情報量} = \log_2 \left(\frac{\text{共起度数・全度数}}{\text{度数X・度数Y}} \right)$$

たとえば､あるスペイン語の資料で *muy* (='very')という語の得点が 120, *bien* (='well')の得点が 167, 全語数が 26578 でした。そうすると、*muy* と *bien* が共起得点が理論的に期待できる値は(120/26578) x (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しょう。

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

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

これの対数（底=2）は 5.962 となります。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 (=$\log_2 16$)となります。

Dice 係数は共起得点を得点(x)と得点(y)の平均で割った値です。ここでは相互情報量のように全語数を計算に含めることはしません。

$$\text{Dice 係数(d.)} = \frac{\text{共起度数}}{(\text{度数(x)} + \text{度数(y)}) / 2} \quad 0.0 \leqq \text{d.} \leqq 1.0$$

分子の共起得点は上の表の a にあたります。得点(x)は a + b にあたります。これは x が y と共起するケース数と y と共起しないケース数の合計になります。同様に得点(y)は a + c です。よって、

$$\text{Dice 係数(d.)} = \frac{a}{(2a + b + c) / 2} = \frac{2a}{(2a + b + c)}$$

b = c = 0 のとき最大値 1 になり、a = 0 のときに最小値 0 になります。Dice 係数は Jaccard の a を 2 倍にしたものです。a と b+c を対照化する、と考えれば、a が 2 数(b, c)と対照化しているので、Dice 係数のほうがつり合いがとれていると思います。

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。この問題は、一般に類似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、多数のデータ間の類似度を見る場合には、問題の両者に存在しない特徴であってもほか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。

◇4 優先係数

以上がよく使われている代表的な類似度係数ですが、そのほかにも次のような類似度係数が考えられます。ここでは、X/(X+Y)という相対型(r: relative)、または(X – Y) / (X + Y)という対照型(c: contrast)によって分類し、さらにd値の有無、積算(mult.)の有無を明記しました[27]。

考えられる類似度係数	X	Y	r:c	d	mult.
1. a / (a + b)	a	b	r	-	-
2. (a - b) / (a + b)	a	b	c	-	-
3. a / (a + c)	a	c	r	-	-
4. (a - c) / (a + c)	a	c	c	-	-
5. a / [a + (b + c)]	a	b + c	r	-	-
6. [a - (b + c)]/[a + (b + c)]	a	b + c	c	-	-
7. 2a / [2a + (b + c)]	2a	b + c	r	-	-
8. [2a - (b + c)] / [2a + (b + c)]	2a	b + c	c	-	-
9. a^2 / (a^2 + bc)	a^2	bc	r	-	+
10. (a^2 - bc)/ (a^2 + bc)	a2	bc	c	-	+
11. a / [a + $(bc)^{1/2}$]	a	√(bc)	r	-	+
12. [a - $(bc)^{1/2}$] / [a + $(bc)^{1/2}$]	a	√(bc)	c	-	+
13. (a + d) / [(a + d) + (b + c)]	a + d	b + c	r	+	-
14. [(a + d) - (b + c)]/[(a + d) + (b + c)]	a + d	b + c	c	+	-
15. ad / (ad + bc)	ad	bc	r	+	+
16. (ad - bc) / (ad + bc)	ad	bc	c	+	+
17. $(ad)^{1/2}$ / [$(ad)^{1/2}$ + $(bc)^{1/2}$]	√(ad)	√(bc)	r	+	+
18. [$(ad)^{1/2}$ - $(bc)^{1/2}$] / [$(ad)^{1/2}$ +$(bc)^{1/2}$]	√(ad)	√(bc)	c	+	+

さらに、9と15は次数が2になっているので、次の式で次数を1に下げることも考えられます[28]。

19. a / $(a^2 + bc)^{1/2}$

20. [ad / (ad + bc)]$^{1/2}$

[27] これらの中で、5, 7, 13, 14, 16はすでに取り上げたものです。ここでは全体を整理するために、これらの公式も含めました。

[28] 10と16は分子が負になることがあるので、根を使うことができません。

上の 8.を「優先係数」(coefficient of preference)と名付けて活用したいと思います。「優先係数」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。2a が b + c と比べてどの程度優先されているのかを示します。優先係数(p.)は Dice 係数の 2a と(b+c)を対照化させた係数です。

$$\text{Preference 係数(p.)} = \frac{2a - b - c}{2a + b + c}$$

p.: -1.0 (a=0) ≦ 0.0 (2a = b+c) ≦1.0 (b=c=0)

◇5 各類似度係数の比較

Phi 係数と Ochiai 係数の選択に限らず、実際の分析でこれらの類似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。発表がそうした係数の数値自体による裏付ける根拠よりも、その先にある類似性を主張することが大きな目標であり、ほかの根拠に十分裏付けられているのであれば、単純一致係数や Russel and Rao 係数や Jacard 係数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。類似度係数が、強い裏付けの根拠として重要な意味を持つならば、Yule 係数や Hamann 係数を選択し、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 Phi 係数を使えばその説明は必要なくなります。Phi 係数にわずかな説明を加えることで Ochiai 係数を使うこともできるでしょう。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。

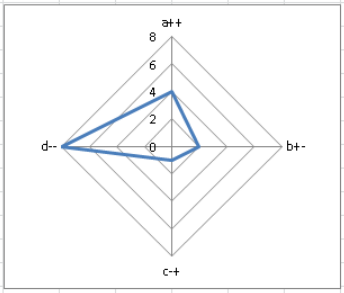
しかし、このような決定は本質的ではなく、実際的な条件に従っています。本質を追究するには、それぞれの係数の性質と分析対象のデータの性質をよく理解して、本質的な条件と実際的な条件のどちらも考慮に入れた上で決定しなくてはいけません。そうすれば、自分でも納得ができますし、自信をもって説明できます。

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。先に見た「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。ここで d 値(0:0)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優先係数(Preference: p.)を選択するとよいでしょう。

性質	s.m.	r.r.	j.	d.	y.	h.	ph	o.	p
d (0:0)を扱う	v	-	-	-	v	v	v	-	-
逆方向(-)を検知	-	-	-	-	v	v	v	-	v
積算がある	-	-	-	-	v	-	v	v	-

逆方向を検知する係数は完全に等質な分布のときゼロになります。これは、次のような実験をするとわかります。



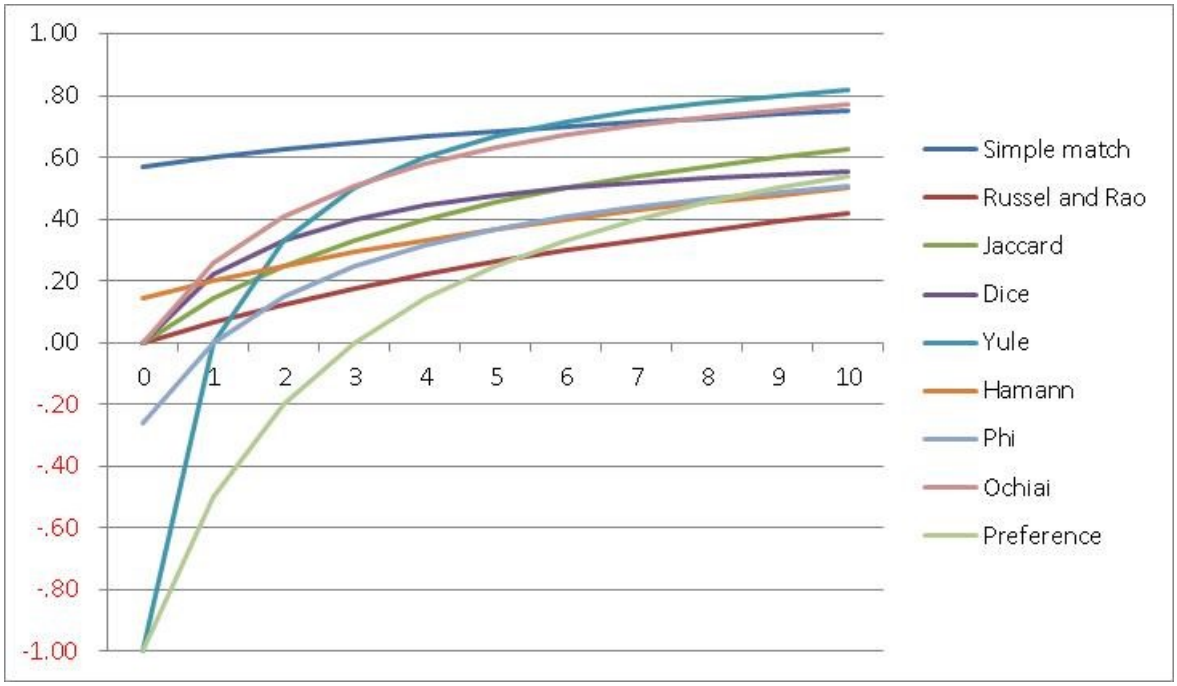
a++	b+-
4	2
c-+	d--
1	8

Co-occurrence	4.000	a
Simple match	0.800	(a+d)/(a+b+c+d)
Russel and Rao	0.267	a/(a+b+c+d)
Jaccard	0.571	a/(a+b+c)
Dice	0.727	2a/(2a+b+c)
Yule	0.882	(ad-bc)/(ad+bc)
Hamann	0.600	((a+d)-(b+c))/((a+d)+(b+c))
Phi	0.577	(ad-bc)/sqrt((a+b)*(a+c)*(d+b)*(d+c))
Ochiai	0.730	a/sqrt((a+b)*(a+c))
Preference	0.455	(2a - (b+c))/(2a+(b+c))

ここで、-1 から 1 の間をとる係数ならばゼロになりますが、他の係数は 0.5 (s.; o.), 0.25 (r.r.), 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 s.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。

次の表と図は b=2, c=4, d=8 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。

Co-occurrence (a+/+)	0	1	2	3	4	5	6	7	8	9	10
b (+/-)	2	2	2	2	2	2	2	2	2	2	2
c (-/+)	4	4	4	4	4	4	4	4	4	4	4
d (-/-)	8	8	8	8	8	8	8	8	8	8	8
Simple match	.57	.60	.63	.65	.67	.68	.70	.71	.73	.74	.75
Russel and Rao	.00	.07	.13	.18	.22	.26	.30	.33	.36	.39	.42
Jaccard	.00	.14	.25	.33	.40	.45	.50	.54	.57	.60	.63
Dice	.00	.22	.33	.40	.44	.48	.50	.52	.53	.55	.56
Yule	-1.00	.00	.33	.50	.60	.67	.71	.75	.78	.80	.82
Hamann	.14	.20	.25	.29	.33	.37	.40	.43	.45	.48	.50
Phi	-.26	.00	.15	.25	.32	.37	.41	.44	.47	.49	.51
Ochiai	.00	.26	.41	.51	.58	.63	.67	.70	.73	.75	.77
Preference	-1.00	-.50	-.20	.00	.14	.25	.33	.40	.45	.50	.54



これを見ると、逆方向を検知しない Simple match, Russel and Rao, Jaccard, Dice の振幅が小さく、とくに Simple match の振幅が小さいことが確認できます。そして、Phi や Ochiai の振幅は小さく、同じ程度の幅であることもわかります。それらに対して Yule と Preference の振幅が大きいことが特徴的です。Yule の上昇は急ですが、Preference は比較的緩やかに上昇します。このことは a[++]の値が高い場合の弁別性を保証します。

ほかにもいろいろな実験をしてそれぞれの係数の性質を調べておく必要があるでしょう。データ分析ではさまざまなデータを扱ったことのある人であれば経験が生かして係数を選択できます。私たちはデータ分析を始めたばかりなのであまり経験はありませんが、何度でも実験で確かめることはできます。実際のデータには数の限りがありますが、実験はいくらでも可能です。また、私たちが経験する実際のデータはかなり偏りがあるのが普通ですが、実験するときは全部自分でコントロールできますから、納得がいくまで確かめることができます。

数値を積算している係数は、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 Ochiai 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。

データの性質として、方向性があるものならば、逆方向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。

分析の手順としては、完全に理解して経験を積む前は、とりあえず全部の係数を比較し、大きく異なる結果を出した係数について、その原因を探り、次にデータと照合して、データの性質を一番よく示している、と思われる係数を選択するとよいでしょう。そのためには、データの性質をよく知っていることと、係数の性質をよく理解していることが必要です。何度でも実験をして確かめてください。

◇6 Excelで質的データを扱う

類似係数を使ってデータを比較するにはまず量的なデータを質的なデータに変換する必要があります。これには IF 関数を使えば便利です。例として次のデータを使用します。

	A	B	C	D	E	F	G	H	I	J	K
1	語	基準手	調査		基準手	調査		a++	b+-	c-+	d--
2	abajo	5	10		1	1		1	0	0	0
3	abandonar	9	6		1	1		1	0	0	0
4	abandono	0	0		0	0		0	0	0	1
5	abarcar	1	0		1	0		0	1	0	0
6	abastecimiento	2	0		1	0		0	1	0	0
7	abatir	0	1		0	1		0	0	1	0
8	abeja	2	3		1	1		1	0	0	0
9	abertura	0	0		0	0		0	0	0	1
10	abismo	0	0		0	0		0	0	0	1
11	abnegación	0	0		0	0		0	0	0	1
12	abogado	3	6		1	1		1	0	0	0
13	abonar	5	0		1	0		0	1	0	0
14	abono	0	0		0	0		0	0	0	1
15	abordar	0	0		0	0		0	0	0	1
16	aborrecer	0	6		0	1		0	0	1	0
17				基準値		0	和	4	3	2	6

(1) はじめに、量的データの質化の基準を設定します。

A17 を質的データに変換するための基準値とします。この値よりも大きい

場合、「1」に変換するというルールにします。0 よりも大きいときに変換する場合は F17=0 と記入しておきます。

(2) IF 関数を使って量的データ(B2)を質的データ(E2)に変換します。

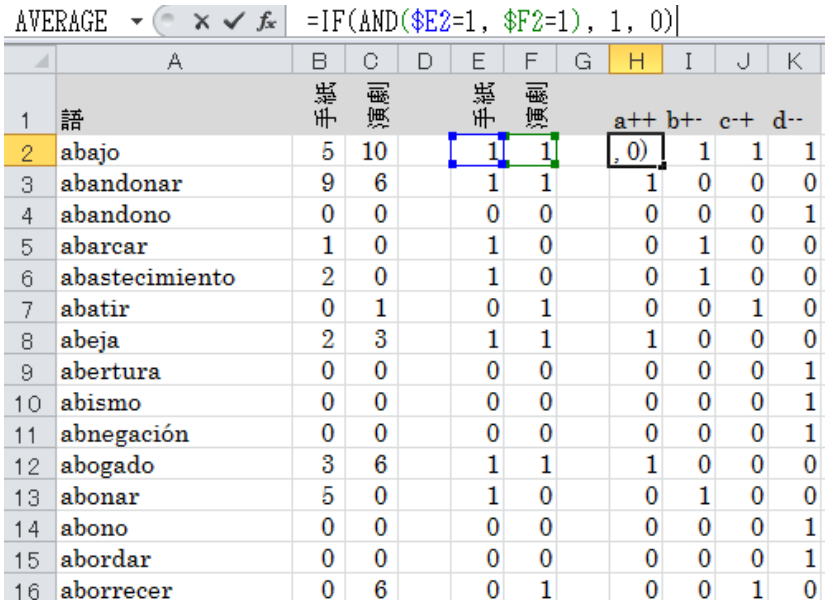
E2=IF(B2>$F$17, 1, 0)[29]

この式の意味は、E2 が基準値の値(0)よりも大きい場合は、1 をそれ以外は 0 を返す、ということです。

(3) E2 を E2:F16 にコピーします。これで 0 より大きい値を 1 と表示することができました。

四象限の計算

次に、さきほどの変換の結果を基に、共通して使われているもの、一方だけ使われているもの、どちらも使われていないものを集計しましょう。



	A	B	C	D	E	F	G	H	I	J	K
1	語	手紙	演劇		手紙	演劇		a++	b+-	c-+	d--
2	abajo	5	10		1	1		, 0)	1	1	1
3	abandonar	9	6		1	1		1	0	0	0
4	abandono	0	0		0	0		0	0	0	1
5	abarcar	1	0		1	0		0	1	0	0
6	abastecimiento	2	0		1	0		0	1	0	0
7	abatir	0	1		0	1		0	0	1	0
8	abeja	2	3		1	1		1	0	0	0
9	abertura	0	0		0	0		0	0	0	1
10	abismo	0	0		0	0		0	0	0	1
11	abnegación	0	0		0	0		0	0	0	1
12	abogado	3	6		1	1		1	0	0	0
13	abonar	5	0		1	0		0	1	0	0
14	abono	0	0		0	0		0	0	0	1
15	abordar	0	0		0	0		0	0	0	1
16	aborrecer	0	6		0	1		0	0	1	0

(1) はじめに E2 と F2 を対象としてデータを入力します。

H2 =IF(AND($E2=1, $F2=1), 1, 0)

この式の意味は、E2（手紙）と F2（演劇）が共に 1 の場合、1 を返し、

---

[29] ここでは基準値を動かすことができるように$F$17 としましたが、下記のように$F$17 を使用しなくても同じ結果を得ることができます。

E2=IF(B2>0, 1, 0)

それ以外は 0 にする、ということです。AND を使って複数の条件を指定していることに注意してください。

H2 を I2:K2 にコピーして、一部を次のように修正します。

I2=IF(AND($E2=1, $F2=0), 1, 0)
J2=IF(AND($E2=0, $F2=1), 1, 0)
K2 IF(AND($E2=0, $F2=0), 1, 0)

(2) H2:K2 を H2:K16 にコピーします。

(3) G17 を書き込み、SUM で H17:K17 を計算します。

H17 =SUM(H2:H16)
H17 を I17:K17 にコピーします。I17 =SUM(I2:I16)
J17 =SUM(J2:J16)
K17 =SUM(K2:K16)

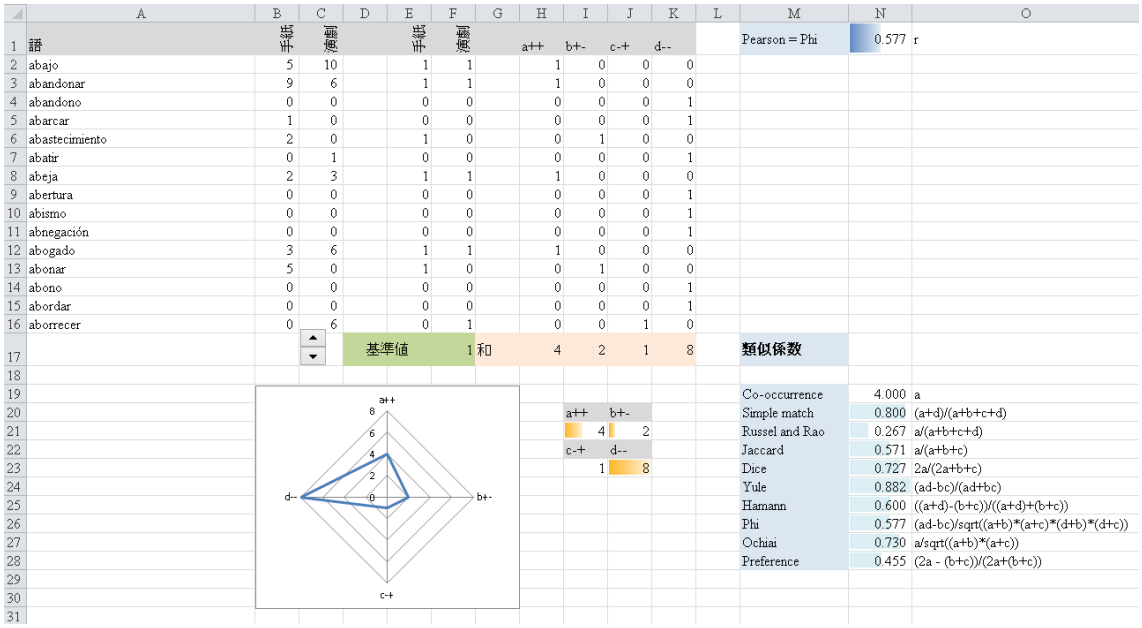
最終的には次のような値になります。

	a++	b+-	c-+	d--
和	4	3	2	6

これで四象限での集計が完了です。

各種の類似係数

それでは各種の類似係数を計算してみましょう。



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	語	手話	動画		手話	動画		a++	b+-	c-+	d--		Pearson = Phi	0.577	r
2	abajo	5	10		1	1		1	0	0	0				
3	abandonar	9	6		1	1		1	0	0	0				
4	abandono	0	0		0	0		0	0	0	1				
5	abarcar	1	0		0	0		0	0	0	1				
6	abastecimiento	2	0		1	0		0	1	0	0				
7	abatir	0	1		0	0		0	0	0	1				
8	abeja	2	3		1	1		1	0	0	0				
9	abertura	0	0		0	0		0	0	0	1				
10	abismo	0	0		0	0		0	0	0	1				
11	abnegación	0	0		0	0		0	0	0	1				
12	abogado	3	6		1	1		1	0	0	0				
13	abonar	5	0		1	0		0	1	0	0				
14	abono	0	0		0	0		0	0	0	1				
15	abordar	0	0		0	0		0	0	0	1				
16	aborrecer	0	6		0	1		0	0	1	0				
17				基準値	1		和	4	2	1	8		類似係数		
18															
19													Co-occurrence	4.000	a
20									a++	b+-			Simple match	0.800	(a+d)/(a+b+c+d)
21									4	2			Russel and Rao	0.267	a/(a+b+c+d)
22									c-+	d--			Jaccard	0.571	a/(a+b+c)
23									1	8			Dice	0.727	2a/(2a+b+c)
24													Yule	0.882	(ad-bc)/(ad+bc)
25													Hamann	0.600	((a+d)-(b+c))/((a+d)+(b+c))
26													Phi	0.577	(ad-bc)/sqrt((a+b)*(a+c)*(d+b)*(d+c))
27													Ochiai	0.730	a/sqrt((a+b)*(a+c))
28													Preference	0.455	(2a - (b+c))/(2a+(b+c))
29															
30															
31															

M19:M27 でそれぞれの係数を求めます。

(1) 共起回数：M19=H17

(2) Simple match 係数：M20=(H17+K17)/(H17+I17+J17+K17)

(3) Russel and Rao 係数：M21=H17/(H17+I17+J17+K17)

(4) Jaccard 係数：M22=H17/(H17+I17+J17)

(5) Yule 係数：M23=((H17*K17)-(I17*J17))/((H17*K17)+(I17*J17))

(6) Hamann 係数：M24=((H17+K17)-(I17+J17))/((H17+K17)+(I17+J17))

(7) Phi 係数：

M25=((H17*K17)-(I17*J17))

/SQRT((H17+I17)*(H17+J17)*(I17+K17)*(J17+K17))

(8) Ochiai 係数：M26=H17/SQRT((H17+I17)*(H17+J17))

(9) Prominence 係数：M27=(H17/(H17+I17)+H17/(H17+J17))/2

(10) Preference 係数：M30=(2*H17-I17-J17)/(2*H17+I17+J17)

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、ということと仮説にしたいと思います。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、それが獲得される、という仮説です。これは、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す。この実験に 12 人が参加しました。

Individuo	(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この仮説に沿うものです。

かなり敷衍して考えてみると、はたして私たちは外国語をくりかえし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要素は瞬間的に獲得し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「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考えられます。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、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。

7 データの検定

6.1 では量的なデータの 2 つの行列に対して、関連性がどの程度あるかという相関係数について見ました。6.2 では質的なデータを対象として、四象限の情報から類似係数を算出する方法を見てきました。ここでは、クロス集計の表から関連度を数値化する方法を扱います[30]。次の表を見て下さい。

	and	but	so	合計
全体	58	43	28	129

単純集計表

[30] ＊参考：池田央．1976.『統計的方法 I 基礎』新曜社．pp.121-132.

	and	but	so	合計
文頭	12	7	11	30
文中	46	36	17	99
全体	58	43	28	129

クロス集計表

上の表は 1 つの指標（英語の等位接続詞）について数値（頻度）を表したものです。一方、下の表は(1)「英語の等位接続詞」と(2)「出現位置」という 2 つの指標を基に集計したものです。このようなものをクロス集計表と呼びます。ここで問題となるのは、この 2 つの指標はお互いに関連しているかどうかということです。具体的に言うと、2 つが関係している場合、「(1)英語の等位接続詞の(2)出現位置は単語によって異なる」という結論になりますし、関係していない場合、「(1)英語の等位接続詞の(2)出現位置は単語に左右されない」（それぞれの現象は「独立」である）ということになります。この判定をする手法が、カイ二乗検定です。ここではカイ二乗検定を理解するために、単純な例として 2-2 の表を用いて説明していきます。

7. 1 検定の方法

なぜカイ二乗検定が必要なのでしょうか。次のようなケースで考えてみましょう。ある現象を数えるにあたって、次のように、それが出現した場合だけを数えるやり方があります。

「方法 A」…効果があったケース：59
「方法 B」…効果があったケース：49

「方法 A」に効果があった場合の数を 59、「方法 B」に効果があった場合の数を 49 として単純に比較すると、確かに「方法 A」のほうが優れている、という結論になるかもしれません。しかし、ここで「方法 A」（そして「方法 B」）に効果があったことを確かめるには、「方法 A」（そして「方法 B」）に効果がなかったケースも調べることが必要です。その結果が次の表です。

実測値1	効果がある	効果がない
方法A	59	35
方法B	49	53

実測値 1 の結果を見ると、やはり「方法 A」のほうが「方法 B」より優れ

ているように見えますが、仮に次の実測値２ようなケースになったときは判断が逆転してしまいます。

実測値2	効果がある	効果がない
方法A	59	65
方法B	49	55

「方法 A」と「方法 B」はどちらも効果がある場合よりも効果がない場合の方が上回り、それぞれの方法の差は 6 ですが、「方法 A」の「効果がない」の数が大きくなっています。

さらに、次の実測値 3 のようなケースがあります。「方法 A」も「方法 B」もどちらも「効果がある」の数が「効果がない」の数よりも上回っています。両者は「効果がある」と「効果がない」の差は 10 となっています。はたして「方法 A」が「方法 B」に比べて効果があると言えるのでしょうか。

実測値3	効果がある	効果がない
方法A	59	49
方法B	49	39

これらは単に「効果がある」という肯定的な反応だけを数えていては見つからなかった問題を示しています。つまり、方法 A と方法 B の差を考えるには、効果があった場合と効果がなかった場合の両方を考える必要があるということです（「方法」と「効果の有無」という 2 つの指標でクロス集計する必要があるということです）。

それでは実測値 1～3 の場合、方法 A と方法 B に差があるといえるのはどれでしょうか。この数値を統計的に算出するのがカイ二乗検定です。この方法を用いることで、差があるかどうかをはっきりと数値で示すことができます。

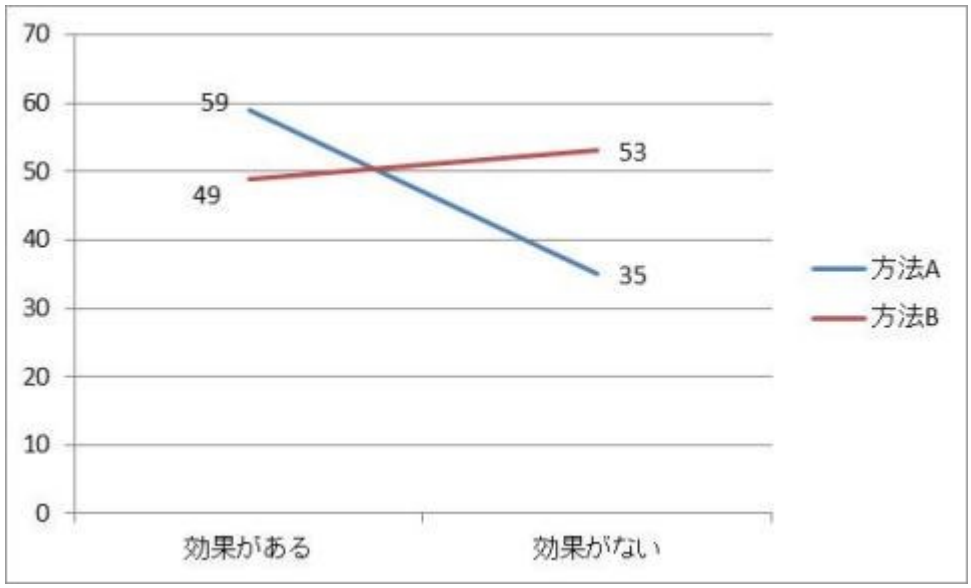
◇1 期待値を計算する

カイ二乗検定の基になるデータは、(1)実測値と(2)期待値です。以下、例として実測値 1 を見ていきましょう[31]。

[31] 比率は「効果がある」の値を「効果がない」の値で割った値です。1 より大きいと「効果がある」ほうが多いことを示します。

実測値1	効果がある	効果がない	計	比率
方法A	59	35	94	1.686
方法B	49	53	102	0.925
計	108	88	196	

このデータをグラフにして視覚化しておきます。



期待値の計算方法に関しては、5 章（→）で見ましたが、ここでは少し違った角度から算出方法を再度考えてみます。説明のために、観測値を次のように呼ぶことにします。

実測値	効果がある	効果がない
方法A	a	b
方法B	c	d

次の表では各セルに「期待される」得点(期待得点：expected score)が示されています。期待得点は次のような式で計算されます。

期待値	効果がある	効果がない
方法 A	$\frac{(a+b)\times(a+c)}{a+b+c+d}$	$\frac{(a+b)\times(b+d)}{a+b+c+d}$
方法 B	$\frac{(c+d)\times(a+c)}{a+b+c+d}$	$\frac{(c+d)\times(b+d)}{a+b+c+d}$

たとえば、方法 A の効果がある期待値は、方法 A の総数($a + b$) 94 に「効

果がある」と期待できる率(a + c) 108、掛けた数値を総数で割った値です。総数 94 が 108:88 に分割されるときに 108 の側に当然期待できる数値、94 x (108 / 196)を示します[32]。

「方法 A」・「効果がある」の期待得点　94 x 108 / 196 = 51.796
「方法 A」・「効果がない」の期待得点　94 x 88 / 196=42.204
「方法 B」・「効果がある」の期待得点　102 x 108 / 196=56.204
「方法 B」・「効果がない」の期待得点　102 x 88 / 196=45.796

期待値	効果がある	効果がない	計
方法A	51.796	42.204	94
方法B	56.204	45.796	102
計	108	88	196

もし実際に観察される値が当然予測される値（期待値）と近いならば、「偶然でも起こるかもしれない分布」ということになります。逆に、もしそれが期待値から大きく外れるならば、観察されたデータは有意味な分布を示していると考えられます。「偶然ではほとんどあり得ない」と考えるのです。つまり、カイ二乗検定のポイントは、「実測値と期待値のズレを見る」というところにあります。

◇2 カイ二乗値を求める

期待値と実測値のずれを総合的に判断するため、すべての升目(a, b, c, d)における実際の観測値と期待値の「相対的な差」の総和で求めます。相対化するには、実測値から期待値を引いたものを期待値で割ります。また、「相対的な差」の合計は、そのままでは 0 になってしまいますので、単純に期待値からの距離を求めるために二乗しておきます（これがカイ「二乗」という名前の由来です）。

$$\text{標準化した値} = \frac{(\text{実測値} - \text{期待値})^2}{\text{期待値}}$$

このような操作を「標準化」と呼びます。データには一定の単位がありますが、標準化すると単位がなくなります。単位がなくなると、どのような

[32] ここで、これらの期待値のすべてが 5 以上であるかどうかを確かめておきます。いずれかが 5 以下だと誤差が大きくてカイ二乗検定には向いていないデータと判断されます。

データでも統計的に同じ処理ができるようになるのです[33]。次がそれぞれの相対的な差です。

標準化	効果がある	効果がない
方法A	1.002	1.230
方法B	0.923	1.133

これらの値を合計した値が「カイ二乗の統計量」（χ^2）と呼ばれるものです。

$$\chi^2 = 1.002 + 1.230 + 0.923 + 1.133 = 4.288$$

式を一般化しましょう。実測値 a, b, c, d の期待値をそれぞれ、a', b', c', d' とし、標準化した値の総和は次のようになります。

$$\chi^2 = \frac{(a - a')^2}{a'} + \frac{(b - b')^2}{b'} + \frac{(c - c')^2}{c'} + \frac{(d - d')^2}{d'}$$

カイ二乗の統計量は、期待値からのズレ（距離）の総和ということになります。この値が大きいほど、期待値とのズレが大きいということが言えます。

7. 2 検定の考え方

値や差を推定する統計は確率に基づいています。確率は全くありえない 0%から、絶対そうである 100%までありますが、たとえば方法 A と方法 B の間に「差が 100%ある」と言い切ることは難しいです。では、どうするかというと「差がないとは言えない」という消極的な言い方をします。この証明には、100%とは反対の 0%から出発します。つまり、「方法 A と方法 B には（全く）差がない」という仮説からスタートするのです。この仮説を帰無仮説(H_0)と呼びます。無に帰したい（棄却したい）仮説ですのでこのように呼ばれます。この逆の「差がある」という仮説を対立仮説(H_1)と呼びます。

---

[33] たとえば、データの絶対的な値を 3 メートルだとして、それが全体の 10 メートルの中での割合を見ると、0.3 という単位（メートル）がなくなった数値になり、この数値は他のケースの割合と同じ尺度で（標準化された尺度で）比較できます。期待値を使った標準化もそれとよく似ています。

H_0: 方法 A と方法 B には差がない

H_1: 方法 A と方法 B には差がある

推測統計が求める確率は H_0 が成立する確率です。たとえば検定の結果、3%と出れば、これは「方法 A と方法 B には差がない可能性が 3%」ということです。逆に言えば、97%の確率で H_1（差がある）が成立します。この場合、H_1 が成立する可能性がかなり高いですので、H_0 は棄却できることになります。

このように棄却する基準のことを「有意水準」と呼びます。一般に 5%と 1%が用いられます。たとえば「5%の有意水準で H_0 が棄却できる」という結論は、H_0 の成り立つ確率が 5%以下（H_1 が成り立つ確率が 95%以上）ということになります。

◇1 検定の評価

カイ二乗統計量は、期待値とのズレであるということを見ました。それではこの値がどの程度大きければ差があるといえるのでしょうか。2-2 の分割表では次のように決まっています。

有意水準	閾値
5%	3.841
1%	6.634

閾値とは、カイ二乗統計量の値がそれ以下であれば成り立たないということです。あらためて先ほどの値を見ると、4.288 ですので、5%の閾値よりも大きいことになります。従って、この結果は「5%水準で有意な差がある」と解釈できます。一方、1%水準の閾値は 6.634 ですので、この水準では H_0 を棄却することはできません。

さきほど「2-2 の分割表では」という但し書きをつけましたが、この点は重要ですので触れておきます。カイ二乗統計量は期待値からのズレの合計であるということを見ましたが、マス目が増えれば増えるほど合計の値が大きくなります。たとえば、2-2 のマスと 4-4 のマスではマスの数は 4 マスと 16 マスですので、平等に扱うのはおかしいでしょう。つまり、有意水準の閾値の値も、マス目の数によって大きくなっていくということになります。

この基準は「自由度」(degree of freedom, df)と呼ばれます。自由度というのは自由に値を決めることができるマスの数のことです。たとえば、2-2

のマスでは、1 つのマスを決めると、縦と横の合計が同じならばm他のすべてのマスの値は自動的に決まってしまうので自由度は 1 ということになります。次の表で方法 A の「効果がある」を 10 とすると、方法 A の「効果がない」は 84、方法 A の「効果がある」は 98、方法 B の「効果がない」は 4 に決まります。

実測値 1	効果がある	効果がない	
方法 A	10	94−10	94
方法 B	108−10	102−(108−10)	102
	108	88	196

なお、n-p のクロス集計表の自由度は、(n-1)-(p-1)で求めることができます。以上のことをまとめて次のように表します[34]。

$$\chi^2 = 4.288 > \chi^2 \text{ (df: 1, p: 0.05)} = 3.841$$

これは「カイ二乗統計量は 4.288 で、自由で 1 の場合の 5%有意水準の 3.841 よりも大きく統計的に有意である」という意味です。

◇2 イェイツの補正（Yates' correction）

2-2 の数値表ではカイ二乗の統計量が一般に大きくなる傾向があります。そのため、先の χ^2 の代わりに次の式を使って少し補正します。

$$\chi^2 \text{ (Yate's cor.)} = \frac{n(|ad - bc| - \frac{n}{2})^2}{(a+b)(a+c)(c+d)(b+d)}$$

そうすると、イェイツの補正をした結果 χ^2 (Yate's cor.)は 3.714 となって、先ほどの値よりも少し小さくなりました。この場合も有意水準 1%で帰無仮説を棄却できないことになります。このようにイェイツの補正を利用することでより慎重な評価ができます。

カイ二乗・イェイツの補正・Phi 係数

イェイツの補正はχ二乗値の分子から n/2 を引いた数値になります。このことを確かめておきましょう。

[34] df は自由度(degree of freedom)、p は確率(probability)を示します。

はじめに次が実測値です。

O	X(+)	X(-)	和
Y(+)	a	b	a + b = s
Y(-)	c	d	c + d = t
和	a + c = u	b + d = v	a + b + c + d =n

次に a, b, c, d それぞれの χ 二乗値を計算します。

$\chi^2(a) = (a - su / n)^2 / (su / n) = [(an - su)^2 / n^2][n / su] = (an - su)^2 / nsu$
$\chi^2(b) = (b - sv / n)^2 / (sv / n) = [(bn - sv)^2 / n^2][n / sv] = (bn - sv)^2 / nsv$
$\chi^2(c) = (c - tu / n)^2 / (tu / n) = [(cn - tu)^2 / n^2][n / tu] = (cn - tu)^2 / ntu$
$\chi^2(d) = (d - tv / n)^2 / (tv / n) = [(dn - tv)^2 / n^2][n / tv] = (dn - tv)^2 / ntv$

この和が χ 二乗(χ^2)です。

$$\chi^2 = [\, tv(an - su)^2 + tu(bn - sv)^2 + sv(cn - tu)^2 + su(dn - tv)^2 \,] / nstuv$$

$$= [\, tv\,(a^2n^2 - 2ansu + s^2u^2)$$
$$+ tu\,(b^2n^2 - 2bnsv + s^2v^2)$$
$$+ sv\,(c^2n^2 - 2cntu + t^2u^2)$$
$$+ su\,(d^2n^2 - 2dnvt + t^2v^2) \,] / nstuv$$

$$= (a^2n^2tv - 2ansutv + s^2u^2tv$$
$$+ b^2n^2tu - 2bnsvtu + s^2vtu^2$$
$$+ c^2n^2sv - 2cntusv + t^2u^2sv$$
$$+ d^2n^2su - 2dnvtsu + t^2v^2su) / nstuv$$

縦列で足します。

$$= [\, n^2\,(a^2tv + b^2tu + c^2sv + d^2su)$$
$$- 2stuvn\,(a + b + c + d)$$
$$+ stuv\,(su + sv + tu + tv) \,] / nstuv$$

$$= [\, n^2\,(a^2tv + b^2tu + c^2sv + d^2su)$$
$$- 2stuvn^2$$
$$+ stuv\,(s + t)(u + v) \,] / nstuv$$

$$= [\, n^2\,(a^2tv + b^2tu + c^2sv + d^2su) - 2n^2stuv + n^2stuv] / nstuv$$

$= n^2 (a^2tv + b^2tu + c^2sv + d^2su - stuv) / nstuv$

$= n (a^2tv + b^2tu + c^2sv + d^2su - stuv) / stuv$

$s = a + b$, $t = c + d$, $u = a + c$, $v = b + d$ なので

$= n [a^2(c + d)(b + d)$
$+ b^2(c + d)(a + c)$
$+ c^2(a + b)(b + d)$
$+ d^2(a + b)(a + c)$
$- (a + b)(c + d)(a + c)(b + d)] / stuv$

$= n [a^2(bc + cd + bd + d^2)$
$+ b^2(ac + c^2 + ad + cd)$
$+ c^2(ab + ad + b^2 + bd)$
$+ d^2(a^2 + ac + ab + bc)$
$– (ac + ad + bc + bd)(ab + ad + bc + cd)] / stuv$

$= n [\underline{a^2bc} + \underline{a^2cd} + \underline{a^2bd} + \underline{a^2d^2}$
$+ \underline{ab^2c} + \underline{b^2c^2} + \underline{ab^2d} + \underline{b^2cd}$
$+ \underline{abc^2} + \underline{ac^2d} + b^2c^2 + \underline{bc^2d}$
$+ a^2d^2 + \underline{acd^2} + \underline{abd^2} + \underline{bcd^2}$
$- \underline{a^2bc} - \underline{a^2cd} – \underline{abc^2} – \underline{ac^2d}$
$- \underline{a^2bd} – \underline{a^2d^2} – abcd – \underline{acd^2}$
$- \underline{ab^2c} – abcd – \underline{b^2c^2} – \underline{bc^2d}$
$- \underline{ab^2d} – \underline{abd^2} – \underline{b^2cd} – \underline{bcd^2}] /stuv$

$= n (a^2d^2 – 2abcd + b^2c^2) / stuv$

$= n (ad – bc)^2 / [(a + b)(a + c)(c + d)(b + d)]$

この式は先に見た χ^2 (Yate's cor.)とわずかに分子の一部が異なるだけです。また、この式は先に見た Phi 係数を二乗して n(= a + b + c + d)を掛けた数値になります。

$\chi^2 = n\ Phi^2$

◇3 Excelでカイ二乗検定をする

それでは Excel でカイ二乗検定を行ってみましょう。カイ二乗検定では実測値と期待値、そして標準化した値を基にして計算しますので、次のようなカイ二乗検定をするためのシートを作成します。

	A	B	C	D
1	実測値1	効果がある	効果がない	和
2	方法A	59	35	94
3	方法B	49	53	102
4	和	108	88	196
5				
6	期待値	効果がある	効果がない	
7	方法A			
8	方法B			
9				
10	標準化	効果がある	効果がない	
11	方法A			
12	方法B			

実測値は横和と縦和の両方を求めておきます。

期待値

「実測値」の和を参照して「期待値」を計算します。B7 に次の式を書き込み、全体にコピーします。なお、表示はセルの書式設定から小数点以下 3 位までの設定にしました。

B7=$D2*B$4/$D$4

I26

	A	B	C	D
1	実測値1	効果がある	効果がない	和
2	方法A	59	35	94
3	方法B	49	53	102
4	和	108	88	196
5				
6	期待値	効果がある	効果がない	
7	方法A	51.796	42.204	
8	方法B	56.204	45.796	
9				
10	標準化	効果がある	効果がない	
11	方法A			
12	方法B			

標準化

(1)「実測値」と「期待値」を参照して期待値との差を標準化した各値を計算します。二乗には^（キャレット）を使います。次の式を入力し、残りのセルにコピーします。

B11 =(B2-B7)^2/B7

CHITEST　=(B2-B7)^2/B7

	A	B	C	D
1	実測値1	効果がある	効果がない	和
2	方法A	59	35	94
3	方法B	49	53	102
4	和	108	88	196
5				
6	期待値	効果がある	効果がない	
7	方法A	51.796	42.204	
8	方法B	56.204	45.796	
9				
10	標準化	効果がある	効果がない	
11	方法A	=(B2-B7)^2/B7		
12	方法B			

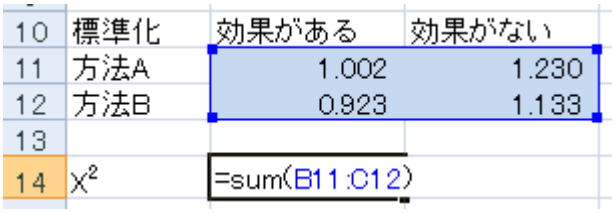
G20

	A	B	C	D
1	実測値1	効果がある	効果がない	和
2	方法A	59	35	94
3	方法B	49	53	102
4	和	108	88	196
5				
6	期待値	効果がある	効果がない	
7	方法A	51.796	42.204	
8	方法B	56.204	45.796	
9				
10	標準化	効果がある	効果がない	
11	方法A	1.002	1.230	
12	方法B	0.923	1.133	

カイ二乗統計量

カイ二乗統計量は標準化した値の合計です。次の式を入力します。

B14＝sum(B11:C12)=4.288



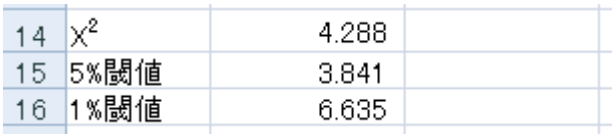
10	標準化	効果がある	効果がない
11	方法A	1.002	1.230
12	方法B	0.923	1.133
13			
14	X^2	=sum(B11:C12)	

有意水準・自由度・限界値

Excel には CHIINV という関数が用意されており、「有意水準」と「自由度」を基に閾値を算出できます。引数は、CHIINV(確率,自由度)です。ここでは自由度 1 の場合の 5%と 1%の閾値を求めてみましょう。

B15=CHIINV(0.05,1)

B16=CHIINV(0.01,1)



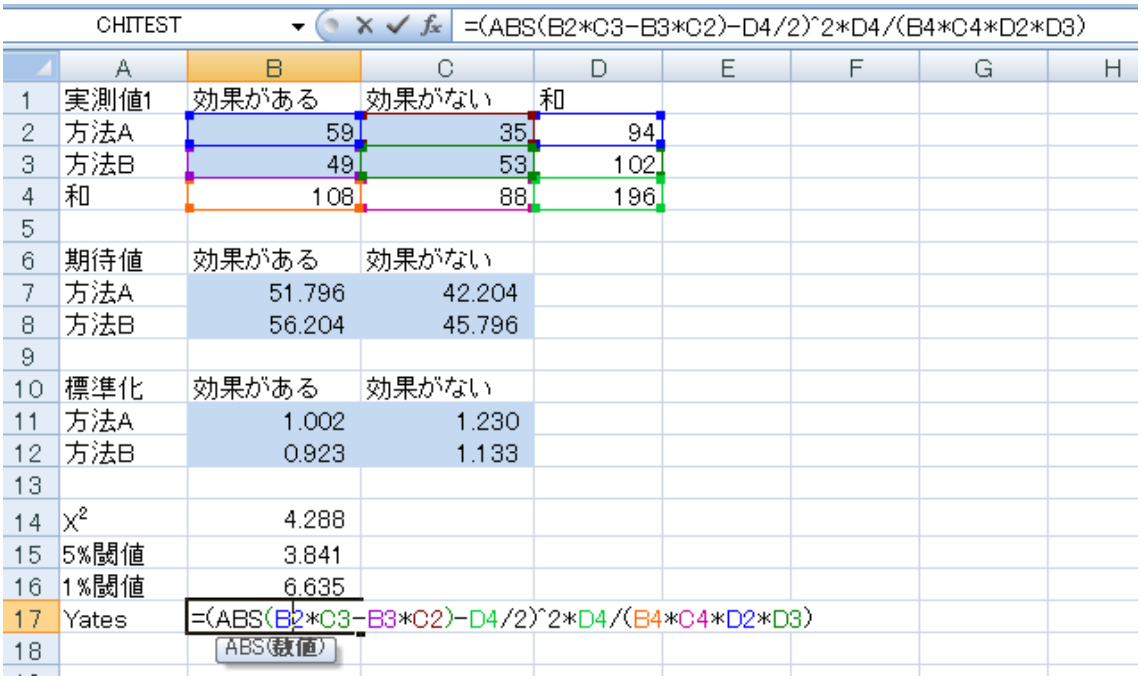
14	X^2	4.288
15	5%閾値	3.841
16	1%閾値	6.635

以上の結果から、カイ二乗統計量は 5%水準の閾値よりも大きく、1%水準の閾値よりも小さいので、5%水準で有意、1%水準ではそうではないということがいえます。

Yates の補正

イェイツの補正を求める関数は残念ながら用意されていませんので、数式を自分で入力します。絶対値に変換するには ABS 関数を利用します。次の式を入力してみましょう。

B17 =(ABS(B2*C3-B3*C2)-D4/2)^2*D4/(B4*C4*D2*D3)=3.714



	A	B	C	D	E	F	G	H
1	実測値1	効果がある	効果がない	和				
2	方法A	59	35	94				
3	方法B	49	53	102				
4	和	108	88	196				
5								
6	期待値	効果がある	効果がない					
7	方法A	51.796	42.204					
8	方法B	56.204	45.796					
9								
10	標準化	効果がある	効果がない					
11	方法A	1.002	1.230					
12	方法B	0.923	1.133					
13								
14	X^2	4.288						
15	5%閾値	3.841						
16	1%閾値	6.635						
17	Yates	=(ABS(B2*C3-B3*C2)-D4/2)^2*D4/(B4*C4*D2*D3)						
18								

CHITEST 関数を使う

Excel には CHITEST 関数が用意されており、これを利用すると実測値と期待値から H_0 が成り立つ確率を直接計算することができます。CHITEST(実測値,期待値)という形で使います。

B18 =CHITEST(B2:C3,B7:C8)

	A	B	C	D
1	実測値1	効果がある	効果がない	和
2	方法A	59	35	94
3	方法B	49	53	102
4	和	108	88	196
5				
6	期待値	効果がある	効果がない	
7	方法A	51.796	42.204	
8	方法B	56.204	45.796	
9				
10	標準化	効果がある	効果がない	
11	方法A	1.002	1.230	
12	方法B	0.923	1.133	
13				
14	X^2	4.288		
15	5%閾値	3.841		
16	1%閾値	6.635		
17	Yates	3.714		
18	CHITEST	0.038		

この計算の結果、0.038 と出ます。これは H_0 が成り立つ可能性が 3.8%であることを示しています。つまり、5%水準では十分に棄却できる値である

ということを示します。

この手法を使うと、標準化の手順が省略できるというメリットと、直接確率を求めることができるというメリットがあります。先ほどまでの結果では 5%水準では有意だが 1%水準では違うということでしたが、3.8%はちょうどこの間に入ります。

実測値 2、実測値 3 について

実測値 2 と 3 について同じように計算するにはシートをコピーして実測値の値を入れ替えればよいでしょう。次のような結果になりました。

	A	B	C	D
1	実測値2	効果がある	効果がない	和
2	方法A	59	65	124
3	方法B	49	55	104
4	和	108	120	228
5				
6	期待値	効果がある	効果がない	
7	方法A	58.737	65.263	
8	方法B	49.263	54.737	
9				
10	標準化	効果がある	効果がない	
11	方法A	0.001	0.001	
12	方法B	0.001	0.001	
13				
14	X^2	0.005		
15	5%閾値	3.841		
16	1%閾値	6.635		
17	Yates	0.004		

	A	B	C	D
1	実測値3	効果がある	効果がない	和
2	方法A	59	49	108
3	方法B	49	39	88
4	和	108	88	196
5				
6	期待値	効果がある	効果がない	
7	方法A	59.510	48.490	
8	方法B	48.490	39.510	
9				
10	標準化	効果がある	効果がない	
11	方法A	0.004	0.005	
12	方法B	0.005	0.007	
13				
14	X^2	0.022		
15	5%閾値	3.841		
16	1%閾値	6.635		
17	Yates	0.000		

カイ二乗統計量もイェイツもかなり小さい値になっています。このことから、実測値 2 と 3 では方法 A と方法 B に差があるとは言えません（H_0 を棄却できません）。

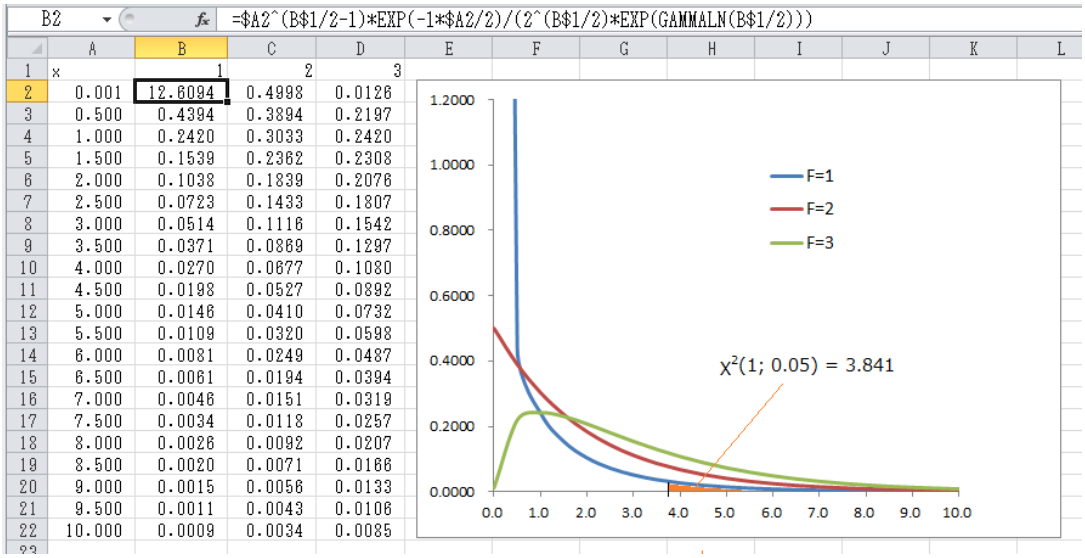
カイ二乗分布

カイ二乗の「限界値」は「有意水準」と「自由度」によって決まります。たとえば、自由度=1、有意水準=0.05 ならば、限界値は 3.841 になります。

この限界値は非常に複雑な数式によって導かれるもので、これを理解することは私たちの「文系のデータ分析」の範囲を超えていると思います。次に示すシートは、カイ二乗分布を示す関数を自由度 1, 2, 3 について求めたものですが、セル[B2]の数式は、次のような関数を使います。このように非常に複雑な式なのです。

B2==$A2^(B$1/2-1)*EXP(-1*$A2/2)/(2^(B$1/2)*EXP(GAMMALN(B$1/2)))

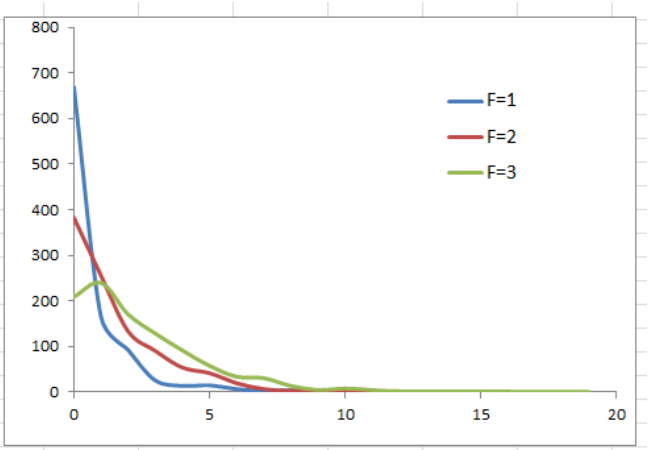
ここで、自由度(F)=1 の線の 3.841 の値の右側の面積が全体の 5%になることを示しています。



x	1	2	3
0.001	12.6094	0.4998	0.0126
0.500	0.4394	0.3894	0.2197
1.000	0.2420	0.3033	0.2420
1.500	0.1539	0.2362	0.2308
2.000	0.1038	0.1839	0.2076
2.500	0.0723	0.1433	0.1807
3.000	0.0514	0.1116	0.1542
3.500	0.0371	0.0869	0.1297
4.000	0.0270	0.0677	0.1080
4.500	0.0198	0.0527	0.0892
5.000	0.0146	0.0410	0.0732
5.500	0.0109	0.0320	0.0598
6.000	0.0081	0.0249	0.0487
6.500	0.0061	0.0194	0.0394
7.000	0.0046	0.0151	0.0319
7.500	0.0034	0.0118	0.0257
8.000	0.0026	0.0092	0.0207
8.500	0.0020	0.0071	0.0166
9.000	0.0015	0.0056	0.0133
9.500	0.0011	0.0043	0.0106
10.000	0.0009	0.0034	0.0085

私たちは実験をすることによってこれを実際に納得することができます。次は、ランダムに 1000 ほどの偏りのないケースを発生させ、自由度=1 のカイ二乗値の頻度を計算した結果です。

x	F1	F2	F3
0	669	383	209
1	166	256	240
2	93	134	171
3	26	91	129
4	14	54	92
5	15	42	58
6	7	20	34
7	3	7	31
8	4	3	14
9	2	3	5
10		4	8
11		1	5
12			2
19			1



それぞれ、先に示した理論的に導かれる連続線の形状に近似していることがわかります。この実験は何度やっても、具体的な数値は変わりますが、グラフの形はそれぞれ類似しています。

ブラックボックス・リープ・ディスコネックション

書店には統計学の参考書が多く並んでいます。「Excel を使ってこのようにすればよい」と説明する手法の本もたくさんあります。実際に手にとって見ると、簡単に統計処理ができるように書かれていて参考になる本もありますが、中には、手法だけを扱って、応用法についての注意などがなく、数学的な背景については大まかに理解していればよい、という姿勢で書かれているものも多いようです。

たしかに、書かれてあるとおりのテクニックを使えばそれなりの結果が出るのですが、どうしてそのような結果が出るのか具体的にわからないことがあります。これでは計算過程がブラックボックスになってしまい、自分が出した結果を説明できません。

参考書の中には説明が「飛躍している」（リープ）と思われるケースもあります。これは説明の段落がどのようにつながるのかわからないような状態です。もしかしたら自分の数学的な知識が不足しているため、リープだと勝手に判断しているのかもしれません。

また、説明の中には「～ということが知られている」「～という公式を使う」というような背景知識に対するリンクになっていることがあります。しかし、私たちが「知られている」という事実や「公式」に疎いとき、背景知識とのリンクは切れてしまっています（ディスコネクション）。

このような理論的な理解がない状態で手法だけを応用してしまうと、結局自分が何をやっているのかわからないのに、自分の名前をつけたレポート・論文・発表を生産してしまうことになりかねません。本人がわかっていないのに、レポート・論文を読む人や発表を聞く人がいるというのは望ましくありません。

そこで、自分にとって、ブラックボックス、リープ、ディスコネクションがあると思われる参考書の説明については、ぜひ自分で実際にいろいろな実験をして納得がいくまで確かめてください。Excel はその実験道具として役立ちます。そして、実験をしながら感覚的に様子がわかったら、今度は統計学や数学の本を読んで数式を理解してください。誰でも難しそうな記号が並んだ数式を目にすると尻込みすることはよくあることですが、そこでじっくり腰を据えて理解してみると案外身近なものであることはよくあることです。理論の理解と実験の順番は逆でも、同時でもよいでしょう。机上の書籍と Excel の往復作業です。いずれにしても自分で納得できた手法を使うことを勧めます。ちょっと面倒かもしれませんが、努力の結果自分が納得できる成果を得たとき、その達成感が次のステップにつながります。

(c) 上田博人（東京大学）Hiroto Ueda (University of Tokyo) 2013.1.17