4. 統計量

行列演算を使ってデータ行列全体の「和」「平均」「分散」「標準偏差」などの統計量(Statistic)を扱います。計算する対象の行列の「行」「列」「全体」について計算します。

4.1. 和

データ行列 D_{np} の横和（横和），縦和（縦和），全体の和を計算します。データ行列 D_{np} の横和 S_{n1} は，次のような行列積で計算します。

$$S_{n1} = D_{np} I_{p1}$$

I_{p1} は P 個の成分をもつ単位縦ベクトルです。

D_{np}	1	2	3
1	6	8	5
2	7	10	6
3	8	4	8
4	9	7	2
5	10	9	4

X

I_{p1}	1
1	1
2	1
3	1

=

S_{n1}	1
1	19
2	23
3	20
4	18
5	23

縦和 S_{1p} は次のような行列積で計算します。

$$S_{1p} = I_{n1}^{T} D_{np} = I_{1n} D_{np}$$

ここで I_{n1}^{T} は単位ベクトル I_{n1} を転置させたものです(=I_{1n})。

I_{1n}	1	2	3	4	5
1	1	1	1	1	1

X

D_{np}	1	2	3
1	6	8	5
2	7	10	6
3	8	4	8
4	9	7	2
5	10	9	4

=

S_{1p}	1	2	3
1	40	38	25

最後に行列全体の総和 S は，横和 S_{n1} または縦和 S_{1p}^{T} の和になります。次は縦和 Sp^{T} の総和 S を求める式です。

$$S = I_{1n} S_{n1} = S_{1p} I_{p1}$$

S_{1p}	1	2	3
1	40	38	25

X

I_{p1}	1
1	1
2	1
3	1

=

S	1
1	103

●正値統計量

下中表は行の和(S)，個数(N)，平均(M)を示します。一方，下右表は空白セルを除いたデータの和(PS)，個数(PN)，平均(PM)を示します。和(S，PS)は同じになりますが，個数(N, PN)と平均(M, PM)は異なります。たとえば，成績処理では 5 回の小テストと出席回数で成績をつけるときに，M + N ではなくて，PM + PN とする方法が考えられます。欠席（空白セル）があると，そのテストが 0 点になって M に大きく影響するためです。言語データでも，同時に多くの個体を比較すると変数に無関係なケースが多くなることがあります。そのときはこの「正値統計量」(positive statistic measure: PSM)を使うことが考えられます。

X	v1	v2	v3	v4	v5
d1	10	19	14	7	12
d2	11	7	10		1
d3			1	12	1
d4		1	2	3	3

行	S	N	M
d1	62	5	12.4
d2	29	5	5.8
d3	14	5	2.8
d4	9	5	1.8

行.P	PS	PN.	PM
d1	62	5	12.4
d2	29	4	7.3
d3	14	3	4.7
d4	9	4	2.3

●群（グループ）の中の統計量

下左表のようなデータ列(v1，v2，v3)＋群列(Group)からなる入力行列を群の分類内で各種の統計量を求めます。下右表は統計量として和を求めた結果です。

D1	x	y	z	Group
d1	5	2	7	a
d2	3	3	2	b
d3	2	0	2	b
d4	4	2	2	c
d5	2	4	3	c
d6	1	8	7	c

和	x	y	z
a	5	2	7
b	5	3	4
c	7	14	12

●変量・個体の同質性・異質性

これから扱う統計量は,すべて縦列でも横行でも計算可能です。しかし，そのように計算された統計量は同質でなければなりません。たとえば，縦列が変数のとき，それらが，単語の長さ，単語内の母音の数，子音の数で

あるようなとき，それぞれの個体（単語）について，これらの数値の和を求めても意味がありません[1]。このことは，年齢と月間読書量のように単位が異なればさらに明らかです。たとえば 12 歳と 5 冊を足した数値 17 は何の意味もありません。一方，変数として比較する文書（文書-1，文書-2，文書-3, ...）を扱っているのであれば，その扱ったすべての文書について，当該の単語が出現した総数を計算することに意味があります[2]。

同じことは個体にもあてはまります。たとえば，個体-1 が定冠詞であり，次の個体-2 が前置詞であり，個体-3, 4, 5, ...，が名詞であって，変数として各文書内の頻度を扱うときは，これら個体のすべての頻度の和が何を意味するのかを見据えておかなければなりません。目的によっては，このような統計量が必要になることもありますが，そのときには異質なデータが混在していることを忘れないようにします。

そこで，そのことを自分にも他者にも明らかにしておくために，「同質個体」(homogeneous individual)，「異質個体」(heterogeneous individual)，「同質変数」(homogeneous variable)，「異質変数」(homogeneous variable) という用語を使ってデータを記述するとよいでしょう。ここで「同質」「異質」というのは同列に扱うことが可能・不可能な数値や名義のことです。

4.2. 平均

4.2.1. 算術平均値

一般に「平均値」(Mean: M) と呼ばれる「算術平均値」(arithmetic mean: AM) はデータの和をその個数 N で割った値です。縦軸の平均 AMv は

$$AMv = I_{n1}{}^{T} D_{np} / N,\ AMv = D[X(Tr(In1, Dnp), N]$$

（I_{n1}：単位ベクトル，D_{np}：データ行列）

D	1	2	3
1	6	8	5
2	7	10	6
3	8	4	8
4	9	7	2
5	10	9	4

縦軸 AMv	1	2	3
算術平均値	8.0	7.6	5.0

次はデータの横軸の平均値 AM_H を示します。

[1] 単語内の母音の数と子音の数の和ならば意味があります。
[2] 各文書の大きさを考慮に入れます（→相対頻度）。

X	v1	v2	v3	v4	v5
d1	10	19	14	7	12
d2	11	7	10	0	1
d3	0	0	1	12	1
d4	0	1	2	3	3

横軸 AM_H	算術平均値
d1	12.400
d2	5.800
d3	2.800
d4	1.800

●算術平均値位置

データの度数が小さな数値の方に偏っていたり，逆に大きな数値に偏っているときは，算術平均値が中央に位置しません。その偏りの様子を見るために，次の表を作成しました。「算術平均値位置」(Arithmetic mean position: AMP)はデータに算術平均値を加えた数値列の中で「算術平均値順位」(Arithmetic mean order: AMO)から[3]，全体の個数の半分 (N+1) / 2 を引いた値です。つまり，中央をゼロ(0)として正負の番号をつけます。

AMP = AMO - (N+1) / 2

「算術平均値相対位置」(Arithmetic mean relative position: AMRP)は，算術平均値の位置をデータと平均値を足した個数の半数(N+1) / 2 で割った値です。

AMRP = AMP / [(N+1) / 2]

よって算術平均値(AM)が最小値に近づくと算術平均値相対位置(AMP)は-1 に近づき，最大値に近づくと算術平均値相対位置(AMP)は+1 に近づきます。算術平均値が中央に位置すれば算術平均値相対位置(AMP)はゼロ(0)です。

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

横軸	AM	AMO	AMP	AMRP
h1	12.4	4	1	.333
h2	5.8	3	0	.000
h3	2.8	5	2	.667
h4	1.8	3	0	.000

たとえば，h1 の算術平均値 12.4 の順位(AMO)は 7, 10, 12 の次になるので 4(番)になります。そして，データの半数は(N+1) / 2 = 3 なので算術平均値位置(AMP)は

AMP = 4 - (5+1) / 2 = 1

[3] 算術平均値順位の関数をプログラムするには，データを昇順に並べ変えなくても，データの中で算術平均より小さな数値の個数を数え，それに 1 を加えた数値を返すようにします。Excel シートでは算術平均値を範囲に含めた Rank 関数が算術平均値順位を返します。

このように，算術平均値位置(AMP)がプラスのときは，算術平均値が中央より上側にあることを示します。このときの算術平均値相対位置(AMRP)は

AMRP = 1 / [(5+1) / 2] = 0.333

こうして h1 の算術平均値が相対的尺度([-1, 1])の上側 1/3 にあることがわかります。

●相対値と対照値

数値 X と数値 Y を比較するとき「差」(X – Y)と「比」(X / Y)が使えます。さらに，X / (X + Y), Y / (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。これは一般に「相対値」(Relative value: Rv)と呼ばれます。

Rv = X / (X + Y)

相対値(Rv)は[0, 1]の範囲を持ちます。その最小値(= 0)は X = 0 のときに発生し，最大値(= 1)は Y = 0 のときに発生します(X / X = 1)。その中間値(0.5)は X = Y のときに発生します: X / (X + Y) = X / 2 X = 1/2。

たとえば「絶対頻度」(Absolute frequency: AF)を「和」(Sum: S)で割った「相対頻度」(Relative frequency: RF)は，その名のとおり相対値です。

RF = AF / S

この式の和(S)は実は AF の考えられる最大値 AF.max を示します。たとえば絶対頻度{2, 3, 5} (S=10)の相対頻度{2/10, 3/10. 5/10} = {0.2, 0.3, 0,5}の中に示される分母(10)は和(S=10)の中で考えられる最大値 AF.max = S = 10 になります。よって先の式は次のように書き換えられます。

RF = AF / AF.max

この式の相対値は[0, 1]の範囲を持ちます。その最小値(= 0)は AF = 0 のときに発生し，最大値(= 1)は AF = AF.max のときに発生します(AF.max / AF.max = 1)。

以下で扱うすべての相対値(R*)は次の一般式の形になります。

R* = * / *.max

また，(X – Y) / (X + Y)という式もよく使われます。これを「対照値」(Contrastive value: Cv)と呼びます。

Cv = (X – Y) / (X + Y)

対照値(Cv)の範囲は[-1, 1]になります。ゼロ(0)を中心にして，正負±1 に伸びます。最小値(-1)は X = 0 のとき，そして最大値(1)は Y = 0 のときに発生します。中間値(0)は X = Y のときに発生します。このように対照値の最大値と最小値はそれぞれ相対値と同じ条件で発生しますが，その範囲が異なります(相対値[0, 1]; 対照値[-1, 1])。

相対値(Rv)と対照値(Cv)の間には次の関係があります。

$$2\,Rv - 1 = Cv$$

なぜならば

$$
\begin{aligned}
&2\,Rv - 1 \\
&= 2X / (X + Y) - 1 && \leftarrow Rv = X / (X + Y) \\
&= 2X / (X + Y) - (X + Y) / (X + Y) && \leftarrow \text{分母を共通にする} \\
&= [2X - (X + Y)] / (X + Y) && \leftarrow \text{共通分母でまとめる} \\
&= (X - Y) / (X + Y) && \leftarrow \text{分子を整理} \\
&= Cv && \leftarrow Cv \text{ の定義}
\end{aligned}
$$

相対値は一般に単に「割合・率」(ratio)とも呼ばれていますが，これらは「X / 全体」という式で示されます。ここで「相対値」と呼ぶ値は本質的に割合・率と同じですが，分母の中を X と Y，つまり比較するもの(X)と比較されるもの(Y)を分けて考えます。割合・率では隠れて見えなかったことが，相対値にすると，自己を含めた全体(X+Y)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，対照値は「自己と他者の差」(X−Y)と「自己と他者の和」(X+Y)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，直ちにはわかりません。対照値を直感的に理解するには，次のように式を変形するとよいでしょう。

$$Cv = (X - Y) / (X + Y) = X / (X + Y) - Y / (X + Y)$$

つまり，対照値(Cv)は X の相対値と Y の相対値の差を求めたことになります。よって相対的な X と Y を比較して評価することになるのです。そこで，相対値が数値をポジティブに評価するためのもの，対照値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。統計量がゼロ(0)から正の+1 に向かって一方向にだけ伸びる概念を示すときは相対値を使い，それがゼロを中心にして正負の±1 に向かって左右の両方向に伸びる概念であれば対照値を使います。

先にも述べたとおり，「相対値」は一般に「割合・率」と呼ばれますが，「対照値」は私たちが調べた限りではその名称が見つかりませんでした。しかし，後述する「Yule 連関係数」「Hamann 連関係数」「Goodman and Kruskal 順序連関係数」などで使われています。私たちもこのテキストの各所で応

用します。

4.2.2. 幾何平均値

次の式で示される平均値は「幾何平均値」(Geometric Mean: GM)と呼ばれます。

$$GM = [\prod X(i)]^{(1/N)}$$

ここで$\prod X(i)$は X(1) * X(2) * … * X(N)という累積する積を示し，指数(^)の 1/N は N 乗根を示します。たとえば，{3, 4}の幾何平均値は(3 * 4)^(1/2) = 3.46 になり，{3, 4, 5}の幾何平均値は(3*4*5)^(1/3) = 3.91 になります[4]。

幾何平均値はデータの成分が倍数や比率であるときの平均値として使われます。たとえば，13, 14, 15 世紀における同一規模の文書内における文字<j>の頻度が，14 世紀全体で前世紀(13 世紀)の 2 倍になり，15 世紀には前世紀(14 世紀)の 10 倍になったとします。この 14, 15 世紀の 2 回の頻度の推移の平均値の倍数として単純に算術平均値を用いると(2+10) / 2 = 6 となり，1 世紀ごとに 6 倍増加したことになります。しかし，たとえば 13 世紀の頻度が 100 であったとすると，世紀間(13, 14, 15 世紀)の推移は(100, 100*2, 200*10) = (100, 200, 2000)になりますから，1900 増加したことになるはずです(2000-100 = 1900)。しかし，算術平均値 (2+10) / 2 = 6 で求めた 6 倍を適用すると 600 になってしまいます(100 * 6 = 600)。そこで，幾何平均値を使うと，(2*10)^(1/2) = 4.472…倍になります。これが 1 世紀あたりの平均値増加率ですから，100 に 4.472 を 2 回掛けると確かに 2000 になります(100 * 4.472… * 4.472… = 1999.878…)。

今度は比率の平均値について考えます。たとえば，スペイン・カスティーリャ地方の中世における語尾母音 e の脱落について，-nd(e)の-d(e)に対する比率が 1/5 = .2 であり，東のアラゴン地方の nd(e)の-d(e)に対する比率が 2/5 = .4 であったとします。ここで，両者の算術平均値を単純に計算すると，(.2 + .4) / 2 = .3 となります。しかし，逆に nd(e)の-d(e)に対する比率は，それぞれ 5/1 = 5, 5/2 = 2.5 になりますから，その算術平均値は(5 + 2.5) / 2 = 3.75 になります。先の.3 の逆数は 1/.3 = 3.33 ですから，これは 3.75 と一致しません。そこで，それぞれの幾何平均値を求めていると，$(.2 * .4)^{1/2}$ = .283,

---

[4] 2 個の数値(a, b)の「幾何平均」を次のように幾何的に理解します。それぞれの数値{a, b}が長方形の横と縦の辺の長さとして，その長方形の面積(a*b)と同じ面積の正方形(x*x)の 1 辺の長さ(x)を示します。

a * b = x * x = x^2→ x = (a * b)^(1/2)

3 個の数値(a, b, c)の幾何平均(x)は，3 辺が{a, b, c}の直方体と同じ体積の立方体の 1 辺(x)になります。

a * b * c = x * x * x = x^3 → x = (a * b * c)^(1/3)

4 個以上の数値(a, b, c, d, …)では図形で示せませんが，同様に積算と n 乗根(n-th root)を延長して考えることができます。「幾何平均」は「相乗平均」とも呼ばれます。

$(5 * 2.5)^{1/2} = 3.53$, そして 3.53 の逆数 1/3.53 = .283 で両者は一致します[5]。

このように,変化率や比率の平均値として算術平均値は適していません。その代わりに幾何平均値を使うべきです。さらに幾何平均値が適したデータとして，後述の昇順で並べ替えて急激な増加を示すデータが挙げられます(→「緩やかな増加と急激な増加」)。

幾何平均値の計算では，$\Pi X(i)$の部分が掛け算の連続になるため，X(i)に大きな数値が多くあるとき,プログラムの実行中にオーバーフロー(扱える数値範囲の最大値を超えてしまうこと)が起こるときがあります。そこで,プログラム(後述)では先の幾何平均値の式の両辺の対数(自然対数 Log)を使います。

$$GM = [\Pi X(i)]^{(1/N)}$$

$$\begin{aligned}
Log(GM) &= Log\{[\Pi X(i)]^{(1/N)}\} && \leftarrow \text{両辺の対数} \\
&= 1/N\ Log\ [\Pi X(i)] && \leftarrow Log\ X^A = A\ Log\ X \\
&= 1/N\ Log\ [X(1) * X(2) * \ldots * X(N)] \\
& && \leftarrow \Pi X(i) = X(1) * X(2) * \ldots * X(N) \\
&= 1/N\ [Log\ X(1) + Log\ X(2) + \ldots + Log\ X(N)] \\
& && \leftarrow Log\ X*Y = Log\ X + LogY \\
&= 1/N\ \ \Sigma Log\ X(i) && \leftarrow X(1) + X(2) + \ldots + X(N) = \ \Sigma X(i) \\
&= \ \Sigma Log\ X(i) / N && \leftarrow \text{分母を整理}[6]
\end{aligned}$$

よって

$$GM = Exp\ (\Sigma Log\ X(i) / N) \qquad \leftarrow Exp(X) = e^X,\ Exp(Log(X)) = X$$

幾何平均値の Excel 関数は GEOMEAN を使います[7]。

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

横軸	幾何平均値
h1	11.7
h2	2.4
h3	.7
h4	1.1

[5] 増加率の幾何平均については清水(1996: 32-33)を参照し,比率の幾何平均については池田(1976: 41)を参照しました。

[6] 対数(Log)の算術平均になります。

[7] Excel の GEOMEAN(範囲)は範囲が非常に大きい数値であるときエラー(#NUM!)を返します。そのときは，次のようにして範囲の対数変換(LN)→平均(AVERAGE)→指数変換(EXP)をすれば幾何平均を計算することができます。

=EXP(AVERAGE(LN(範囲)))

●ゼロを含むデータの幾何平均値

データの成分にゼロ(0)があると，そのデータセットの幾何平均値は成分の掛け算をするので: ΠX(i)]^(1/N)，データの他の数値がどのようであっても，その幾何平均値はすべてゼロ(0)になってしまいます。しかし，言語現象の頻度を様々な場面で計算すると，しばしば頻度がゼロになることがあります。とくに大きな偏りがあるデータでは幾何平均値のほうが算術平均値よりもデータの中心性を示すことが多いので(→後述「緩やかな増加と急激な増加」)，ゼロを含むデータでも，ゼロ(0)にはならない幾何平均値が計算できるとよいでしょう。その方法を次のように考えます。

次の表はゼロ(0)を含むデータの算術平均値(mean)と幾何平均値(gm)を示します。

X	A	B	C	D	E	mean	gm
h1	10	19	14	7	12	12.400	11.744
h2	11	7	10	0	1	5.800	0.000
h3	0	0	1	12	1	2.800	0.000
h4	0	1	2	3	3	1.800	0.000

このように，ゼロ(0)を含むデータ(h2, h3, h4)の幾何平均値(gm)はすべてゼロ(0)になってしまいます。そこで，次にデータ全体に 1 を加えた X'について，同じように算術平均値(mean)と幾何平均値(gm)を計算しました。データ全体に 1 を加えたので，算術平均値(mean)と幾何平均値(gm)からそれぞれ 1 を引いた値(mean-1, gm-1)も載せます。

X'	A	B	C	D	E	mean	mean-1	gm	gm-1
h1'	11	20	15	8	13	13.400	12.400	12.797	11.797
h2'	12	8	11	1	2	6.800	5.800	4.623	3.623
h3'	1	1	2	13	2	3.800	2.800	2.204	1.204
h4'	1	2	3	4	4	2.800	1.800	2.491	1.491

h1→h1'の幾何平均値(gm)の増加は 1 よりもやや大きくなっていますが，差は僅かです。gm-1 では差が僅少になりました。よって，ゼロを含むデータの幾何平均値(geometric mean with zero: gmz)として，次の式が使われます。

$$gmz = gm(X+1) - 1$$

●プログラム (R)

```
gm   = function(A){exp(mean(log(A)))} #幾何平均
gmz   = function(A){exp(mean(log(A+1)))-1} #幾何平均(ゼロを含む)
```

●幾何平均値位置

データの頻度分布が極端に右に長い裾を示すとき(極端に高い数値があるとき)の平均値として幾何平均値(GM)を使います[8]。その選択の正しさを確かめるために，先の算術平均値位置と同様にして，「幾何平均値順位」(Geometric mean order: GMO)から，全体の個数(N+1)の半分 (N+1) / 2 を引いて「幾何平均値順位」(Geometric mean position: GMP)を求めます。

GMP = GMO - (N+1) / 2

「幾何平均値相対位置」(Geometric mean relative position: GMRP)は

GMRP = GMP / [(N+1) / 2]

幾何平均値位置(GMP)・幾何平均値相対位置(GMRP)が幾何平均対照値がマイナス(-)のとき，幾何平均値がデータの下側(左側)にあることを示し，プラス(+)のとき，幾何平均値がデータの上側(右側)にあることを示します。幾何平均値相対位置(GMRP)は幾何平均値のデータ内の位置を相対的な尺度([-1, 1])で示します。

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

横軸	GM	GMO	GMP	GMRP
h1	11.7	3	0	.000
h2	2.4	3	0	.000
h3	.7	3	0	.000
h4	1.1	3	0	.000

上の表を見ると，幾何平均値(GM)は中央にあるので(幾何平均値位置 GMP= 0)，データの中心性を正しく示してます。後述するように，極端に高い数値があるデータでは(たとえば h3)，幾何平均値のほうが算術平均値よりもデータの中心性をよりよく示すことが多いです。上の表と先に見た「算術平均位置」(AMP: 下に再掲)と比べてください。

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

横軸	AM	AMO	AMP	AMRP
h1	12.4	4	1	.333
h2	5.8	3	0	.000
h3	2.8	5	2	.667
h4	1.8	3	0	.000

一般に幾何平均値のほうが算術平均値よりも小さな値になります(→後述「スペイン語語彙頻度データの平均値と幾何平均値」)[9]。

[8] データの頻度分布が正規分布のように左右対称に近いときはデータの中心を示す値として算術平均値(前述)を使います。

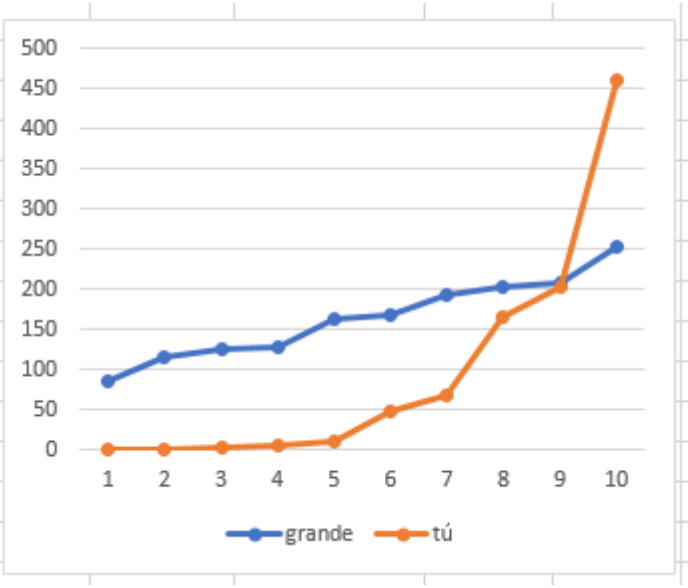
[9] 簡単に a > 0, b > 0 の 2 数について見ておきましょう。

(a + b) / 2 >= (a * b)^(1/2)

●緩やかな増加と急激な増加

次の表と図はスペイン語の形容詞 grande「大きな」と代名詞 tú「君」の 10 種の文書(1 ~ 10)での使用数を示します(それぞれ 10 万語のデータを昇順で並べ替えました)。

*	grande	tú
1	84	0
2	115	1
3	126	2
4	128	4
5	164	11
6	169	47
7	194	67
8	202	165
9	207	204
10	252	461
M	164.1	96.2
GM	156.5	14.6
Md	166.5	29.0



これを見ると，grande は緩やかな増加を示し，tú は急激な増加を示しています。それぞれの算術平均値(M)と幾何平均値(GM)を比べると，両者は grande では近い数値(164.1, 156.5)を示しています。ところが tú では算術平均値(M)と幾何平均値(GM)が大きく異なります(96.3, 18.3)。このように昇順に並べ替えて急激な増加を示す頻度分布では，一般に幾何平均値(GM)のほうが算術平均値(AM)よりもデータの中心性をよく示します。tú の算術平均値(AM) = 96.2 はデータの 7 番と 8 番の間にあって，最大値(10 番)に近くなっていますが，その幾何平均値(GM) = 14.6 は 5 番と 6 番の間にあり，tú の算術平均値(AM)と比べて中央値 (11+47) / 2 = 29.0 に近づいているからです。

中央値(Median)は単純に並べ替えたデータの中央にある値なので，それぞれの数値の大小関係だけが考慮されています[10]。一方，幾何平均値は全体の数値が計算に組み込まれています[11]。

を証明することは次を証明することと同じです。

(a + b)^2 >= [2(ab)^(1/2)]^2

両辺を展開して

a^2 + 2ab + b^2 >= 4*ab

a^2 + 2ab + b^2 - 4*ab >=0

a^2 - 2ab + b^2 = (a - b)^2 >= 0

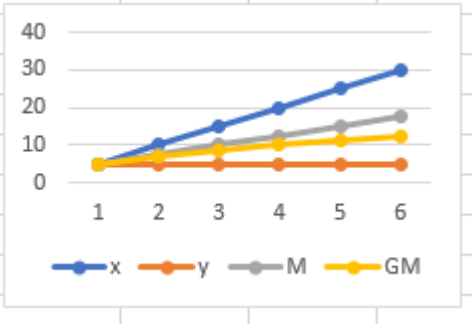
等号(=)は a = b の場合です: (a - b)^2 = 0。

[10] たとえば 9, 10 番のデータが 300, 1000 であっても中央値(29)は変化しません。

[11] 算術平均値(AM)と幾何平均値(GM)の選択の詳細については後述の「幾

次にデータが等差数列(算術数列: arithmetic progression)の場合とそれが等比数列(幾何数列: geometric progression)の場合について，それぞれの算術平均値と幾何平均値を比べます。はじめに，データ(x)が等差数列の場合を見ます。

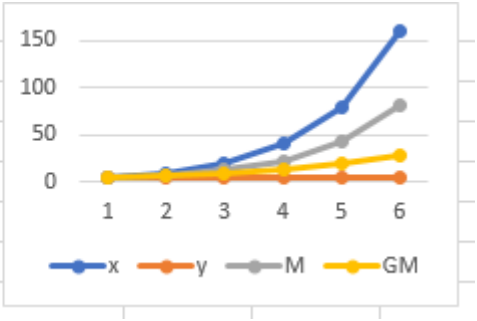
*	x	y	M	GM
1	5	5	5.0	5.000
2	10	5	7.5	7.071
3	15	5	10.0	8.660
4	20	5	12.5	10.000
5	25	5	15.0	11.180
6	30	5	17.5	12.247



このように，算術平均値(M)と幾何平均値(GM)は等差数列(x)の増加にしたがって増加しますが，幾何平均値は算術平均値よりもその増加率が低くなります。しかし，両者は比較的近似しています。

一方，データ(x)が等比数列の場合を見ましょう。

*	x	y	M	GM
1	5	5	5.0	5.000
2	10	5	7.5	7.071
3	20	5	12.5	10.000
4	40	5	22.5	14.142
5	80	5	42.5	20.000
6	160	5	82.5	28.284



上の表と図を見ると，算術平均値(M)は x の数値に忠実にしたがって増加しますが，幾何平均値(GM)は，大きな数値(x)の変化にも同調しながらも，その影響は少なく，むしろ小さな数値(y)に寄り添うようにして変化しています。

算術平均値(M)と幾何平均値(GM)の使い分けは，極端に大きな値の有無がキーになります。極端に大きな値のあるデータの中心として算術平均値を使うと，それがデータの中心ではなく，大きな値に影響された上側の値を示します。極端に大きな値のあるデータの中心として幾何平均値のほうが中心に近くなります。

大量の資料の中には，極端に大きな値があって分布が偏っているデータと，分布に偏りがないデータがあり，同じ基準でデータの中心を示さなければならないときは，やはり算術平均値よりも幾何平均値のほうが，全体的により良く中心を示していることが多いです。偏りのある分布で算術平均値を使ったときの中心の歪みのほうが，偏りのない分布で幾何平均値を使ったときの中心の歪みよりも大きいことが多いからです[12]。次にその例

何標準偏差」で扱います。

[12] しかし，外国語テストの成績のようなデータの和や平均が比較されると

を見ます。

■スペイン語語彙頻度データの平均値と幾何平均値

次はスペイン語語彙の頻度を 10 種のテキスト別に分類したものです[13]。

Word	Car.	Per1.	Ofi.	Lib.	Dra.	Fic.	Ens.	Téc.	Per2.	Man.
de (prep)	4779	10043	11384	6153	3631	6581	7811	8623	8498	4713
que-qué (conj|pn)	5706	2746	2873	2916	5037	2629	3205	3282	3649	4581
y (conj)	3839	2982	3498	4935	2291	3219	3104	3050	3282	2484
a (prep)	3226	2956	2626	2703	2975	2732	2435	2971	2607	3242
en (prep)	1886	3220	2867	2413	1584	2649	2783	2781	3139	1689
ser (v)	1495	1184	1320	3131	2893	1643	2087	1591	1762	2297
haber (v)	1821	1454	735	636	1903	1235	1093	1227	930	1219
por (prep)	1227	1137	1351	1378	857	904	913	1040	1137	912
su-suyo (poses)	534	967	982	1118	572	1401	1285	1250	1153	455
no (av)	1750	571	492	756	3138	1034	1075	822	849	1991

次の表はそれぞれのテキストについての頻度分布を示します。これを見ると，ほとんどが語が[0-500)の頻度範囲の中にあって，それ以上の頻度をもつ語の数は極めて少ないことがわかります。

Freq.	Car.	Per1.	Ofi.	Lib.	Dra.	Fic.	Ens.	Téc.	Per2.	Man.
0	5025	5043	5043	5041	5026	5041	5041	5041	5041	5028
500	18	5	6	7	20	5	6	6	6	17
1,000	6	3	2	2	2	4	3	3	3	4
1,500	3	0	0	0	2	1	0	1	1	2
2,000	0	0	0	1	1	0	2	0	0	2
2,500	0	3	3	2	2	3	1	2	1	0
3,000	1	1	1	1	1	1	2	2	2	1
3,500	1	0	0	0	1	0	0	0	1	0
4,000	0	0	0	0	0	0	0	0	0	0
4,500	1	0	0	1	0	0	0	0	0	2
5,000	0	0	0	0	1	0	0	0	0	0
5,500	1	0	0	0	0	0	0	0	0	0

きは，たとえ上側の外れ値があっても幾何平均ではなく，厳密に算術平均で比較されます。

[13] Car.:手紙，Per1:新聞-1, Ofi:公文書，Lib:書籍，Dra.:演劇，Fic.:小説，Ens.:随筆，Te/c.:科学技術文，Per2:新聞-2, Man.:外国語スペイン語教科書。それぞれのテキストの語数は 10 万であり，全体で 5056 の見出し語(Word)です。この表は降順頻度のトップ 10 語を示します。

一般に語彙の頻度はこのように非常に偏った分布を示します[14]。この分布の特徴を見るために，それぞれのテキストについて，算術平均値(AM)，算術平均値順位(AMO)，算術平均値位置(AMP)，算術平均値相対位置(AMRP)，幾何平均値(GM)，幾何平均値順位(GMO)，幾何平均値位置(GMP)，幾何平均値相対位置(GMRP)，中央値(Md)を次のように計算しました。

縦軸	Car.	Per1.	Ofi.	Lib.	Dra.	Fic.	Ens.	Téc.	Per2.	Man.
AM	16.6	15.8	15.7	15.2	15.5	14.4	14.9	14.6	15.0	15.5
AMO	4564	4199	4277	4391	4565	4419	4378	4338	4427	4504
AMP	2036	1671	1749	1863	2037	1891	1850	1810	1899	1976
AMRP	0.805	0.661	0.692	0.737	0.805	0.748	0.731	0.716	0.751	0.781
GM	0.8	2.3	1.6	2.3	1.1	2.0	2.4	2.3	1.5	0.7
GMO	2380	2238	2301	2344	2644	2430	2221	2260	2202	2724
GMP	-149	-291	-228	-185	116	-99	-308	-269	-327	196
GMRP	-0.059	-0.115	-0.090	-0.073	0.046	-0.039	-0.122	-0.106	-0.129	0.077
Md	1.0	3.0	2.0	3.0	1.0	3.0	3.0	3.0	2.0	0.0

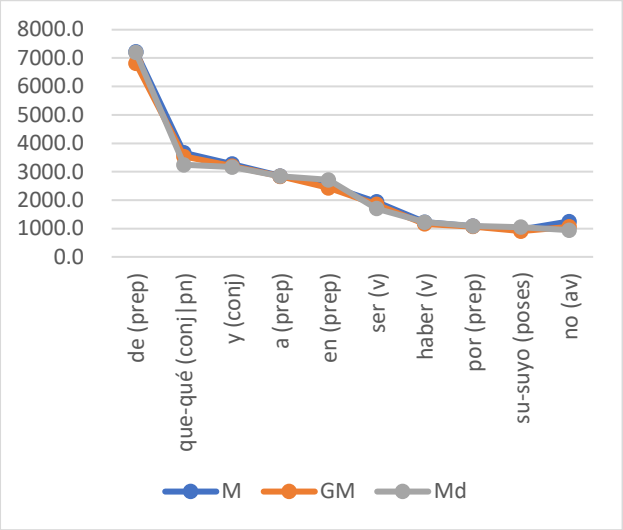
このように偏った分布を示すデータでは，算術平均値よりも幾何平均値のほうが中央値に近く，算術平均値順位は非常に高くなり，一方，幾何平均値順位はほぼ中央(2528)に近く，算術平均値相対位置は最大値の 1.0 に近く，幾何平均値相対位置は中央の 0 に近似します。よって，このように非常に偏った分布を示すデータの中心を見るためには，算術平均値よりも幾何平均値のほうが適しています。

次に，今度はそれぞれの語(Word)について各テキスト内の頻度の分布，算術平均値(AM)，幾何平均値(GM)，中央値(Md)，算術平均値順位(AMO)，幾何平均値順位(GMO)，算術平均値順位－幾何平均値順位(AMO-GMO)を計算しました。

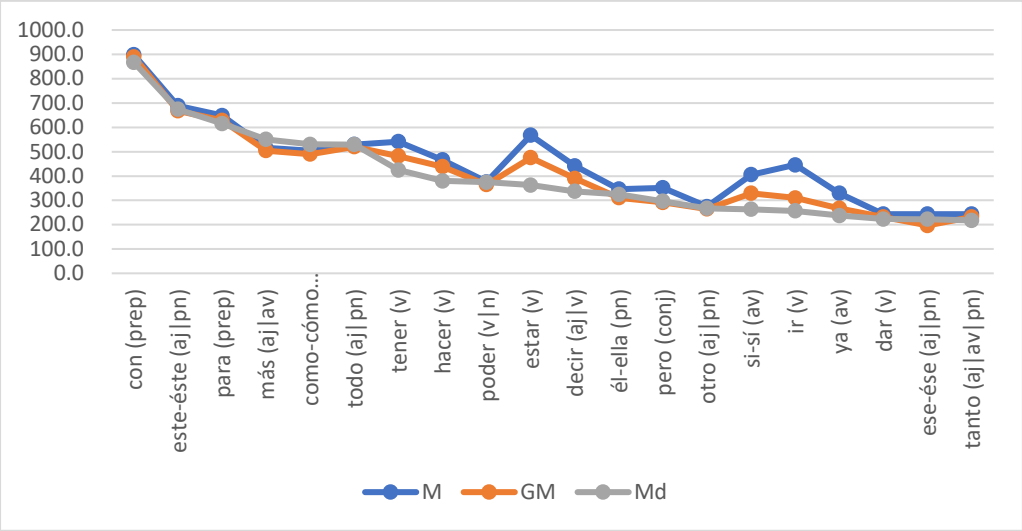
Word	AM	GM	Md	AMO	GMO	MO-GMO
de (prep)	7221.6	6816.9	7196.0	6	6	0
que-qué (conj|pn)	3662.4	3536.1	3243.5	8	7	1
y (conj)	3268.4	3199.7	3161.5	7	6	1
a (prep)	2847.3	2835.9	2844.0	6	6	0
en (prep)	2501.1	2431.8	2715.0	5	5	0
ser (v)	1940.3	1849.1	1702.5	7	7	0
haber (v)	1225.3	1160.3	1223.0	6	5	1
por (prep)	1085.6	1070.8	1088.5	6	6	0
su-suyo (poses)	971.7	908.4	1050.0	5	4	1
no (av)	1247.8	1060.6	941.5	8	7	1

[14] 語彙の頻度だけでなく，文字，音素，形態素などの言語的単位の頻度も非常に偏った分布を示します。

次の図は各語の算術平均値(M)，幾何平均値(GM)，中央値(Md)を示します[15]。数値が大きいので，この図では関係がわかりにくいのですが，全体的に算術平均値(M)よりも幾何平均値(GM)のほうが中央値(Md)に近いようです。

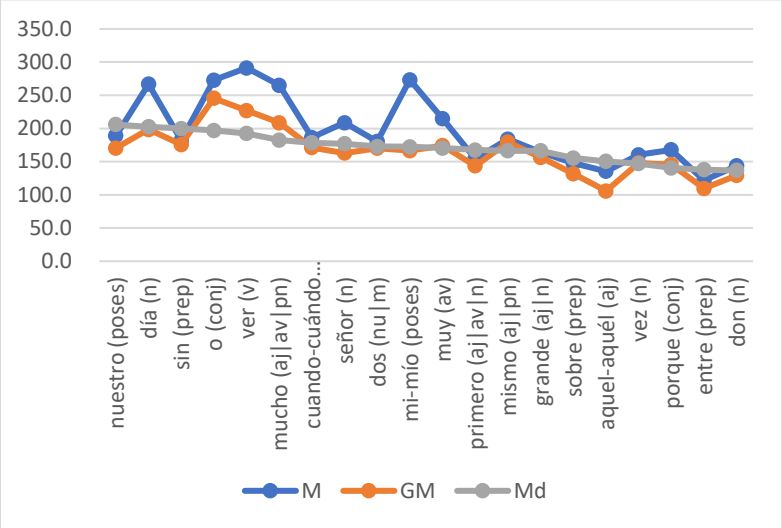


続く 20 語のグラフを見ると，はっきりとその傾向(M より GM が Md に近いこと)がわかります。



それに続く 20 語のグラフの中に，さらにはっきりと同じ傾向が見えます。

[15] 平均値・幾何平均値のそれぞれと中央値との関係を見るために，中央値を降順で並べ替えました。



次の表は，算術平均値順位(AMO)と幾何平均値順位(GMO)をクロスさせた語数を示します：

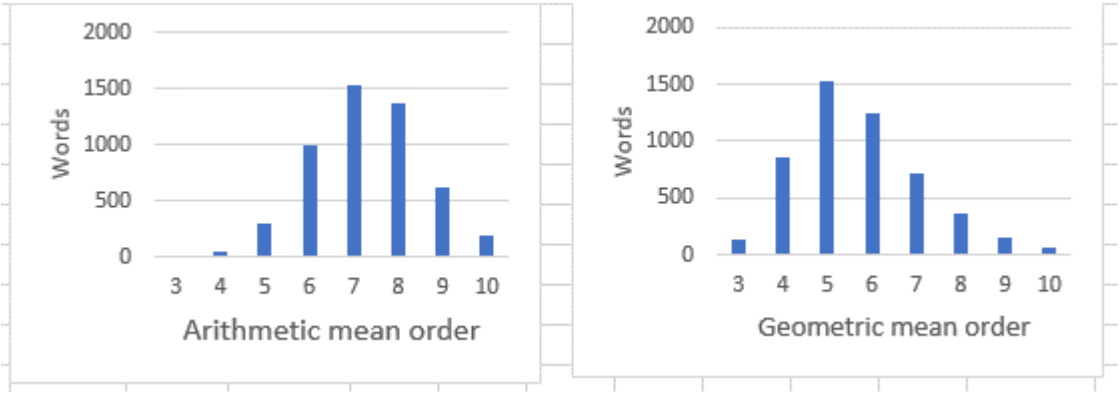
Order	GMO: 3	4	5	6	7	8	9	10	Total
AMO: 3	5								5
4	3	44							47
5	20	70	210						300
6	38	212	371	**380**					1001
7	41	273	469	387	351				**1521**
8	25	189	348	344	252	213			1371
9	8	70	107	114	90	116	118		623
10	1	8	15	12	19	27	43	63	188
Total	141	866	**1520**	1237	712	356	161	63	5056

この表は，全体で算術平均値順位(MO)が幾何平均値順位(GMO)よりも高い(大きい)ことを示しています[16]。そして，算術平均値順位(MO)の頂点(1521)が 7 番であり，幾何平均値順位(GMO)の頂点(1520)が 5 番であることがわかります。よって算術平均値順位(MO)の頂点が幾何平均値順位(GMO)の頂点よりも高いことが確認できます。

上の表を見ると，幾何平均値順位(GMO)が正しく 6 番を示しているとき(380)，算術平均値が 7, 8, 9, 10 のような高い順位を示すことが多いということがわかります(384, 344, 114, 12)。逆に，算術平均値順位(MO)が正しい 6 番を示しているとき(380)，幾何平均値が 7, 8, 9, 10 順位になることは絶対にありません。むしろ幾何平均値順位が 5, 4, 3 番のように下側になることはありますが，それは比較的少数です(371, 212, 38)。そのときのデータではわずかなゼロ(0)や 1 の頻度が全体の中心を下げています。

次に，それぞれの語数の分布をグラフで観察しましょう。

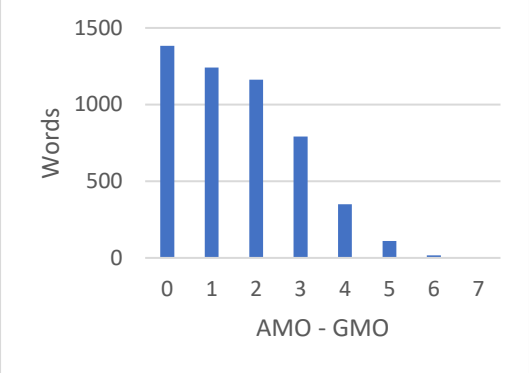
[16] 先に見たように，常に算術平均値≧幾何平均になります。



10 種類のテキストの度数に算術平均値または幾何平均値を加えてあるので，それぞれ 11 個の数値になり，その中央は 6 番になります。算術平均値順位(Arithmetic mean order)の最頻値(頂点)は 7 番ですが，幾何平均値順位(Geometric mean order)の最頻値(頂点)は 5 番です。よって算術平均値の最頻値は中心よりも大きく，逆に幾何平均値の最頻値は中心よりも小さくなっています。

一般に大きな外れ値があるとき，算術平均値は中央値より大きくなりますが，幾何平均値ではその影響は抑えられ，中央値に近づきます。つまり，データの中心性をよりよく示すことが多いことが確認できます。上の 2 つのグラフを比べると，算術平均値順位が 8, 9, 10 で幾何平均値順位よりもずっと高くなっている，つまり中心性を示していないことがわかります[17]。

最後に，算術平均値順位(AMO)と幾何平均値順位(GMO)の差(AMO - GMO)の分布を見ておきましょう。



たしかに両方の順位が同じケース(0)が一番多く，差が 1, 2, …のように大きくなるにつれて該当する語の数が減少しています。これは算術平均値順位(AMO)と幾何平均値順位(GMO)が近い関係にあるように見えます。しかし，差が 0 でなく 1, 2 などのケースが非常に多いこと，そしてケースは少数であっても差が 5, 6, 7 のように大きな数値になることがあることに注意すべきです。

以上で，(1)テキストごとの語彙間の頻度分布では，明らかに幾何平均値

[17] 幾何平均値の位置が 9, 10 になっているデータでは，ほとんどの成分が 0 や 1 で，極めて少数の大きな数値が例外的に特出しています。

のほうが算術平均値よりも中心性を正しく示していることと，(2)語彙ごとのテキスト間の頻度分布では，算術平均値は過大な中心性を示すことがかなり多いこと，一方，幾何平均値はたしかに過小な中心性を示すことがありますが，その数が比較的少ないこと，高頻度語では幾何平均値のほうが算術平均値よりも中央値に近いことを確認しました。

資料：

García Hoz, Víctor. 1953. *Vocabulario usual, vocabulario común y vocabulario fundamental*. C.S.I.C., Madrid.

Juilland, Alphonse and Chang-Rodríguez, Eugene. 1964. *Frequency Dictionary of Spanish Words*. Mouton, The Hague.

4.2.3. 比例平均値

算術平均値は一般に外れ値が少ないデータセットについて適用することがふさわしい平均値です。一方，たとえば{1, 1, 2, 3, 153}のような外れ値(153)があるデータで算術平均値(32)を使うと，多くのデータにとってはその平均値が非現実的な値になります。この 32 という平均値はデータ{1, 1, 2, 3}からも，データ 153 からも，大きく離れているので，データ全体を代表しているように思えません。そのようなデータセットの代表値として，統計学では一般に後述する中央値(=2)が使われます(→「中央値と四分位値」)。しかし，この中央値 2 は 153 にとってはほとんど意味がありません。よってこの中央値もデータ全体を代表していないのです。

そこで，このような外れ値を含むデータについて，むしろデータ全体の規模を反映した代表値として，次のような計算をする「比例平均値」(Proportional Mean: PM)と呼ぶ平均値を使うことを提案します。はじめに算術平均値(AM)を次のようにして求め，次にこれと比較して比例平均値(PM)の方法を説明します（→「比例得点」）。

$$\begin{aligned} AM &= (x_1 + x_2, \ldots, x_n) / n \\ &= (1 + 1 + 2 + 3 + 153) / 5 = 32 \end{aligned}$$

上の式の 1 / n は次のように分配することができます。

$$\begin{aligned} AM &= (x_1 / n + x_2 / n, \ldots, x_n / n) = \Sigma x_i / n \\ &= (1 / 5 + 1 / 5 + 2 / 5 + 3 / 5 + 153 / 5) = 32 \end{aligned}$$

よって算術平均値(AM)はそれぞれの頻度に同じ重さ（ウェイト）1/n (= 1 / 5)を掛けて，その積和を求めたことになります。一方，比例平均値(PM)ではそれぞれの頻度にそれぞれ異なる重さとして，該当する値の大きさに比例した値を総和 N に占める率として設定します。

$$s = (x_1 + x_2, \ldots, x_n) = \Sigma x_i$$

$$PM = (x_1 x_1 / s + x_2 x_2 / s, \ldots, x_n x_n / s)$$
$$= (x_1 x_1 + x_2 x_2, \ldots, x_n x_n) / s$$
$$= \Sigma x_i^2 / s$$
$$= (1^2 + 1^2 + 2^2 + 3^3 + 153^2) / 160$$
$$= 146.4$$

この比例平均値 146.4 はデータ{1, 1, 2, 3, 153}の規模を反映しています。次の表でサンプルデータ X の横和(Sum)，算術平均(AM)，比例平均(PM)を比較してください。

X	v1	v2	v3	v4	v5	Sum	AM	PM
d1	10	19	14	7	12	62	12.4	13.7
d2	11	7	10	0	1	29	5.8	9.3
d3	0	0	1	12	1	14	2.8	10.4
d4	0	1	2	3	3	9	1.8	2.6

このように比例平均はデータセットの規模を代表するので，最大値に近く，最小値からは離れた値になります。

■スペイン語の付加疑問文

次表はスペイン語の付加疑問文の 10 万語あたりの正規頻度(NF)とその和(Sum)，算術平均値(AM)，比例平均値(PM)を比較したものです。

NF.word :100000	1. ¿no?	2. ¿sí?	3. ¿eh?	4. ¿cierto?	5. ¿verdad?	6. ¿cacháis?	7. ¿sabes?	8. ¿viste?	9. ¿hm?	Sum	AM	PM
1.ES.ALC	267.6	111.1	73.6	.0	9.4	.0	39.1	.0	1.6	502.4	55.8	181.1
2.ES.MAD	660.3	217.8	102.8	.0	18.5	.0	56.6	.0	40.0	1096.0	121.8	455.4
3.ES.VAL	223.9	21.1	112.9	.0	2.8	.0	9.2	.0	.0	369.9	41.1	171.4
4.CU.HAB	231.8	3.8	9.0	.0	6.0	.0	.8	.0	.0	251.4	27.9	214.3
5.MX.MON	116.7	65.2	19.2	.0	76.0	.0	.5	.0	1.4	279.0	31.0	86.1
6.CO.MED	16.9	26.6	.0	255.1	.0	.0	.0	.0	.0	298.6	33.2	221.3
7.PE.LIM	1502.6	11.7	2.3	.0	.8	.0	.8	.0	1.6	1519.8	168.9	1485.7
8.CH.STG	34.1	31.8	1.2	13.9	.6	100.6	.6	4.0	14.5	201.3	22.4	63.2
9.UR.MTV	323.9	124.7	22.0	.0	14.1	.0	1.6	86.3	3.1	575.7	64.0	223.4

この表が示すように，頻度の差が非常に大きいので，その算術平均(AM)には代表性がありません。そこでデータの規模を示す比例平均(PM)に注目すると，次に示した 5 地点で昇順順位得点(AR)が大きく異なっています。なお，データの規模を示すはずの総和の順位は算術平均値と同じ順位にな

るので，やはり代表性がありません。

AR	Sum	AM	PM
1-ES-ALC	6	6	4
2-ES-MAD	8	8	8
3-ES-VAL	5	5	3
4-CU-HAB	2	2	5
5-MX-MON	3	3	2
6-CO-MED	4	4	6
7-PE-LIM	9	9	9
8-CH-STG	1	1	1
9-UR-MTV	7	7	7

＊データと分析：PRESEEA en LYNEAL (2017/8/11)
http://lecture.ecc.u-tokyo.ac.jp/~cueda/lyneal/preseea.htm

4.2.4. 調和平均値

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わけにはいきません。たとえば，次のようなケースの平均時速を計算してみましょう。あるグループが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)とすると不都合なことが起こります。なぜなら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 6 km だとすると，往復の所要時間は 12 (km) / 5(km/h) = 2.4 (h)になってしまいますが，実際の往路は 6 (km) / 6 (km/h) = 1 (h)であり，復路は 6 (km) / 4 (km/h) = 1.5 (h)なので，所要時間は往路と復路をあわせて 2.5 (h) のはずです。往復の平均時速は 12 (km) / 2.5 (h) = 4.8 (km/h)でなければなりません。

この平均速度の計算は，単純に総距離を総時間で割った値ですが，いま距離も時間も未知であり，行きと帰りの時速（X km/h と Y km/h）だけから平均時速を出す方法として「調和平均」(Harmonic Mean: HM)が使われます[18]。片道の距離を K / 2 (km)とすると，K / 2 / X が往路の時間になります。同様に復路の時間は K / 2 / Y です。往復の時間(H)は

H = K / 2 / X + K / 2 / Y ←往復の時間
= K / 2 X + K / 2 Y ←各項の分母をまとめる
= (1 / X + 1 / Y) K / 2 ←共通部分をくくる

[18] たとえば池田(1976: 40-41)，清水(1996: 33-34)を参照。

この式から全行程の時速，つまり平均時速（往復の距離／往復の時間: K / H）を引き出します。

```
H = (1 / X + 1 / Y) K / 2          ←往復の時間
1 / H = 1 / [(1 / X + 1 / Y) K / 2]   ←両辺を分母に
K / H = 1 / [(1 / X + 1 / Y) / 2]     ←両辺に K を掛ける
```

調和平均 HM を一般式で書くと次のようになります[19]。

$$HM(X, Y) = 1 / [(1 / X + 1 / Y) / 2] = 2 / (1 / X + 1 / Y)$$

先の例で計算すると次のようになります。

$$HM(6, 4) = 2 / (1 / 6 + 1 / 4) = 4.8$$

なお，この調和平均は次の「分数平均」の特殊なケースです（分子 K が同数）。分子が異なるときは次の分数平均を使います。

調和平均値は複数の速度の平均値，複数の比率の平均値として使われますが，同じ範囲[0, 1]をもつ確率や比率を総合した数値としても利用することができます。→「確率」「二項検定」

4.2.5. 分数平均値

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($R_1 = A_1 / B_1$, $R_2 = A_2 / B_2$)，R_1 と R_2 の分子の和($A_1 + A_2$)を平均の分子とし，R_1 と R_2 の分母($B_1 + B_2$)の和を平均の分母とした分数を使うことを考えます。これを**分数平均**(Fractional Mean: FM)と呼ぶことにします[20]。

$$FM\ (A1/B1, A2/B2) = (A1 + A2) / (B1 + B2)$$

それぞれの平均の結果は連関することがありますが，比率としての分数を扱うとき，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 2 つのコップの食塩水を混ぜ合わせた食塩水の濃度のようなものになるからです。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(AM)ならば

$$AM = (1/4 + 2/5) / 2 = .325$$

---

[19] 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。$HM = N / \Sigma [1 / X(i)]$，ここで X(i)はそれぞれのデータ値を示し，N はデータの個数を示します。いずれかのデータ X(i)が 0 のとき全体の調和平均は 0 になります。

[20] 一般に「加重算術平均」(Weighted arithmetic mean)と呼ばれています。

調和平均(HM)ならば

HM. = 1 / [(4 / 1 + 5 / 2) / 2] = .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(FM)を使うと，次のように計算されます。

FM = (1 + 2) / (4 + 5) = .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
算術平均 AM.	.325	.325
幾何平均 GM.	.316	.316
調和平均 HM.	.308	.308
分数平均 FM.	.333	.280

このように，他の平均と比べて分数平均では第 1 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。

次の表は，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。この表が示すように，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,つまり二日目は一日目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。

このように分数平均は，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求められます。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。

$$FM = (A_1 + A_2 + \ldots, + A_n) / (B_1 + B_2 + \ldots, + B_n) = Sum(A_n) / Sum(B_n)$$

4.2.6. トリム平均値

データの中に極端に大きな値や小さな値（「外れ値」outlier と呼ばれます）があるとき，それが影響して平均値が代表値として役に立たないことがあります。たとえば，{1, 55, 5, 2, 4}のようなデータでは 55 があるために，全体の平均値が 13.4 になり，この平均値が大多数を占める{1, 5, 2, 4}からは大きく外れた値になるので代表値として適していません。

そこで外れ値の影響を除くために中央値（後述）が使われます。そのためにデータを{1, 2, 4, 5, 55}のように大小順に並べ替え，その中央にある値 4 を選びます（データ数が偶数のときは中央にある 2 つの数の平均を使います）。しかし，中央値には中央値以外のデータの大きさは考慮されていません。たとえば，{2, 3, 4, 6, 9}でも，{1, 3, 4, 7, 12}でも，中央値は同じ 4 になります。この場合には中央値よりも平均値の方がデータの代表値として適しています。

外れ値があるデータを代表する値として,「トリム平均値」(trimmed mean: `T.mean`)という数値が使われています[21]。これはソートされた標本値のベクトルの最小のものから t 個(「トリム数」と呼びます)，最大のものから t 個を取り除いて計算した平均値です。次のプロセスでは t を 1, 2 としています。

プログラム(R)

```
T.mean=function(V,t=1,p=F){ #Trimmed mean. V:vt,t:count of trimming,p:process
  n=Len(V); if(t>n/2) stop('t > length(V)/2!')
  V=sort(V); V[c(1:t,(n-t+1):n)]=NA; if(!p) Mean(V) else V}

V=c(0, 13, 18, 20, 42, 157); T.mean(V) # 23.25
V=c(0, 13, 18, 20, 42, 157); T.mean(V,t=2) # 19
```

トリム平均値には中央値と同じ問題が生じます。たとえば{1, 6, 8, 10, 55}のトリム平均値は[6, 7, 8]の平均値(=8)になります。しかし，{0, 6, 8, 10, 550}のトリム平均値も同じ数値になります。そして，トリムされる要素は最大値と最小値なので，たとえば最小値が外れ値ではないときもトリムされています。そして，トリム数は分析者が決定するので，そこに恣意性が生まれる鵜可能性があります。

こうした問題を解決するために，以下では，このトリム平均値を計算するときに取り除く要素の個数(トリム数:t)を徐々に増やす方法を考えます。

[21] 芝・渡部・石塚編『統計用語辞典』(新曜社 1984: 173)では trimmed mean は「調整平均値」と訳されています。

はじめにデータの平均値を計算し，次にデータから最小値と最大値を比較し,中央値からの距離が大きい方を除いたデータの平均値を計算し,次に,最初の中央値からの距離が大きい数値を除いたデータの平均値を計算します。こうして，データがなくなるまで次々に平均値を計算して集めた平均値の合計を平均値の数で割ります(平均値の平均)。中央値からの距離が等しい数値があるときは，それらを一括して除去します。

プログラム(R)

```
T.mean.g=function(V,p=F){
   R=sort(V); m=Median(R); O=Rank.s(-abs(R-m))
   n=Len(V); M=NULL; k=max(O)-1; W=MakeDf(1,n+3)
   M[1]=Mean(R); W[1,]=c(R,median=m,dist.m=NA,mean=Rnd(M[1],1))
   for(i in 1:k){
      S=Where(i,O); R[S]=NA; M[i+1]=Mean(R);
      W=rbind(W,c(R,median=m,dist.m=abs(m-V[S[1]]),mean=Rnd(M[i+1],1)))}
   if(!p) Mean(M)
   else {W=CN(W,c('c'&1:n,VT('median,dist.md,mean'))); RN(W,'r'&1:NR(W))}}
# Gradual trimmed mean. V:vt,n:number of trimming ex.2,p:process

V=c(0,13,18,20,42,157); T.mean.g(V,p=T); CR(); T.mean.g(V)      # 21.80333
V=c(0,13,18,20,42,157); T.mean.g(V,.8,T); CR(); T.mean.g(V,.8)  # 24.33889
```

以下に例 V=c(5,5,6,7,19)を使って具体的にプロセスを示します。

```
V=c(5,5,6,7,19); T.mean.g(V,T); CR(); T.mean.g(V)

   c1 c2 c3 c4 c5 median dist.m mean
r1  5  5  6  7 19      6     NA  8.4
r2  5  5  6  7 NA      6     13  5.8
r3 NA NA  6 NA NA      6      1  6.0

[1] 6.716667
```

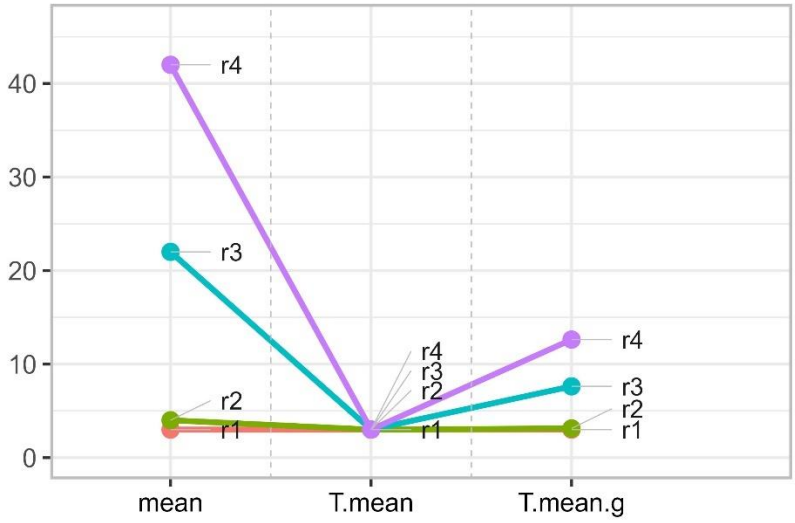
上は要素数が 5 の場合です。中央値は 6 になります。r1 の mean は初期状態の平均値(=8.4)です。このときはトリムされる要素はありません。次の r2 で最初にトリムされる要素は位置 c5 にある最大値 19 です。これが中央値から一番遠い要素です(距離 dist.m ; 19-6=13)。それを NA(not available)にして，NA を除いた平均値を求めると 5.8 になります。同様にして，r3 では列 c1,c2,c4 の要素 5,5,7 がどれも中央値(=6)からの距離が 1 なので，同時にトリムされ，トリムされた要素(6)の平均値は 6.0 になります。これですべての平均値が計算されたので，最後にこれらの 3 個の平均値の平均を計算すると，6.7 (6.716667)になります。

次は，ほかのデータ例を使ってそれぞれの平均値(mean)，中央値(median)，トリム平均値(T.mean)，段階式トリム平均値(T.mean.g)を比べた結果です。

```
U=1:4; V=c(U,5, U,10, U,100, U,200); D=V2D(V,4); D
```

```
Ds=(StatR(D,'T.mean')   -StatR(D,'mean'))/StatR(D,'mean'); Ds1=as.vector(Ds)
Ds=(StatR(D,'T.mean.g')-StatR(D,'mean'))/StatR(D,'mean'); Ds2=as.vector(Ds)
St="mean, median, Equi, T.mean, T.mean.g"; C=DF(D,StatR(D,St),Ds1,Ds2);
L(C,cn=T)
RND(C,"6:7,9,10=1; 8,11,12=3")
G=gLine(C[,c(6,9,10)],m=F,cmx=F,it=F,lb=2,f=12); G; gSave(G,w=400)
```

	c1	c2	c3	c4	c5	mean	median	Equi	T.mean	T.mean.g	Ds1	Ds2
r1	1	2	3	4	5	3	3	0.000	3	3.0	0.000	0.000
r2	1	2	3	4	10	4	3	0.455	3	3.1	-0.250	-0.219
r3	1	2	3	4	100	22	3	0.941	3	7.6	-0.864	-0.653
r4	1	2	3	4	200	42	3	0.970	3	12.6	-0.929	-0.699



データ例(D)の r1=c(1,2,3,4,5)は偏りのないベクトルですが，r2, r3, r4 の最大値は次第に増加して，偏りが大きくなっています(Equi: 0.000→970)。その結果，T.mean は最大値と最小値がトリムされて，すべてのケースのトリム平均値は 3 となります。この結果はトリム平均値は外れ値をまったく考慮しないために起こることです。一方，段階式トリム平均(T.trim.g)は最大値の情報を生かしながら，次々に外れ値の影響を除いています。外れ値が大きく影響する平均値(mean)と比較すると，段階式トリム平均値は平均値と(単純)トリム平均値の間にあることがわかります。偏りのないデータ(r1)では 3 個の平均値は一致します。

オリンピック競技の体操やフィギュアスケートの審判では単純なトリム平均が使用されていますが，単純なトリム平均値(T.mean)では外れ値はまったく考慮されません。一方，段階式トリム平均値(T.mean.g)の計算では，最初は全体の平均値をとり，最後の平均値は中央値だけになります。そして途中の平均値は外れ値である可能性がある最大値と最小値を徐々に取り除いて計算します。そうすると，外れ値はトリム平均値の計算の中で除外されませんが，考慮される回数が少なくなります。逆に中央値に近い数値は考慮される回数が多くなります。このように，単純なトリム平均は外れ値による偏りを回避しますが，最大値･最小値がもつ情報は考慮しません。

一方，段階式トリム平均値はデータがもつ数量的情報を保ちながら外れ値による偏りを回避しています。

4.2.7. ウィンザライズ平均値

切除平均値の計算ではいくつかの最小値と最大値を除外しましたが，ここで扱う「ウィンザライズ平均値」(winsorized mean < Charles P. Winsor 1895-1951)では完全に除外するのではなく，その位置に除外した残りのベクトルの隣接値(最小値と最大値)を代入します(これを「ウィンザライズ」と呼びます)[22]。たとえば，{0,2,7,10,51}の標本では 0 と 51 をウィンザライズし，{2,2,7,10,10}に代えて，その平均値を求めます。ウィンザライズする要素の数は，全体のレンジ(0<r<1)で設定します。たとえば，レンジ 0.8 (80%)のウィンザライズ平均値では，ソートされたデータの左側の 10%と右側の 10%の要素が隣接値代入の対象になります。ここではトリム平均値と同様にして，簡単に両側(最小値と最大値)の要素の個数を指定して処理します。たとえば w=1 ならば，最小値と最大値にそれぞれ 1 個ずつウィンザライズします。

プログラム(R)

```
W.mean=function(V,w=1,p=F){ #V:vt,w:count of winsorization,p:process
   n=Len(V); if(w>n/2) stop('w > length(V)/2!')
   V=sort(V); V[1:w]=V[w+1]; V[(n-w+1):n]=V[n-w]; if(!p) Mean(V) else V}
# Winsorized mean.
```

上の関数では，ウィンザライズする要素の個数(「ウィンザライズ数」と呼びます)の初期設定は w=1 としていますが，この設定はデータ数やデータの分布状態などを考慮して慎重に行わなければなりません。

次は同じデータを用いてトリム平均値とウィンザライズ平均値を比較した結果です。

```
V=c(10,19,14,7,12, 11,7,10,0,1, 0,0,1,12,1, 0,1,2,3,4); D=V2D(V,4)
St="mean, median, min, max, Equi, T.mean, W.mean"
C=DF(D,StatR(D,St)); RND(C,"6,11,12=1; 10=3")

   c1 c2 c3 c4 c5 mean median min max    Equi T.mean W.mean
r1 10 19 14  7 12 12.4     12   7  19   0.125   12.0   12.0
r2 11  7 10  0  1  5.8      7   0  11  -0.300    6.0    5.8
r3  0  0  1 12  1  2.8      1   0  12   0.692    0.7    0.6
r4  0  1  2  3  4  2.0      2   0   4   0.000    2.0    2.0
```

これを見ると，トリム平均値(T.mean)とウィンザライズ平均値(W.mean)の差は大きくありませんが，平衡係数(Balance)の絶対値が大きいとき(r3: 0.692)，ウィンザライズ平均値(0.6)のほうが外れ値の影響を受けた平均値

[22] 芝・渡部・石塚編『統計用語辞典』(新曜社 1984: 13)。

(mean=2.8)から乖離するので，外れ値の影響が少ない，と思われます。しかし，外れ値の影響が少ないということは，外れ値の情報が生かされていないことになります。

ここで，はじめからウィンザライズ数を設定せずにウィンザライズ平均値を計算する方法を提案します。具体的には，データの両側の最小値と最大値の幅を次第に広げていき，それぞれの段階の平均値を計算して，最後に(中央値に達したとき)その平均値の平均を求めます。そのとき，先に見たウィンザライズ平均値の計算では，ソートされたデータの両側にある最小値と最大値の幅を同時に処理しますが，むしろ中央値からの偏差が大きいほうから処理すべきである，という考え方をします。中央値からの偏差が小さければ処理の対象にする理由がないからです。たとえば標本 c(0, 13, 18, 20, 42, 157)の「段階式ウィンザイズ平均値」(gradual winsorized mean)を求めるには，次のプロセスを踏みます。

```
V=c(0,13,18,20,42,157);  W.mean.g(V,p=T); CR(); W.mean.g(V)

   c1 c2 c3 c4 c5  c6 median dist.md mean
r1  0 13 18 20 42 157     19      NA 41.7
r2  0 13 18 20 42  42     19     138 22.5
r3  0 13 18 20 20  20     19      23 15.2
r4 13 13 18 20 20  20     19      19 17.3
r5 18 18 18 20 20  20     19       6 19.0

 23.13333 # Mean (m1:m6)
```

上の r1 の mean が一般の平均値です(41.7)。このデータでは最大値(=157)は中央値(=19)全体から大きく離れています。よって，最小値(=0)よりも最大値を優先させて隣にある値(最大値に続く値=42)と等しくします(r2:c6)。このときの平均は m2=22.5 になります。次に，最小値(=0)と現在の最大値(=42)を比較し，中央値(=20)からの距離(20-13=7, 42-20=22)を比較し，大きい方(42:位置 c5)を選択し，これを続く 20(位置 c3)に置き換えます。このとき，位置 c6 の数値もそれに合わせます。その平均は(m3=15.2)になります。次のステップ(r3)では，最小値(=0)が対象になります。最小値の場合の処理ではその右側にある値(=13)を代入します。実際には，こうした個別の判断をするのではなく，最初にそれぞれの値と中央値との差を計算し，その順番を示すベクトル(O)を用意し，そのベクトルの順序(order)に従って変換の対象を決定します。

プログラム(R)

```
W.mean.g=function(V,p=F){
  R=sort(V); m=Median(R); O=Rank.s(-abs(R-m))
  n=Len(V); M=NULL; k=max(O)-1; W=MakeDf(1,n+3)
  M[1]=Mean(R); W[1,]=c(R,median=m,dist.m=NA,mean=Rnd(M[1],1))
  for(i in 1:k){
```

```
    S=Where(i,O);
    for(w in S){if (w<n/2) R[1:w]=R[w+1] else R[w:n]=R[w-1]}
    M[i+1]=Mean(R)
    W=rbind(W,c(R,median=m,dist.m=abs(m-V[S[1]]),mean=Rnd(M[i+1],1)))}
  if(!p) Mean(M) else {W=CN(W,c('c'&1:n,VT('median,dist.md,mean')))
  RN(W,'r'&1:NR(W))}}
# Gradual winsorized mean. V: vt,p:percent ex.80
# V=c(0,13,18,20,42,157);   W.mean.g(V); CR(); W.mean.g(V,T) # 23.13333
```

次の出力は小さなデータのそれぞれの行の平均値，中央値，ウィンザライズ平均値，段階式ウィンザライズ平均値を比較したものです。

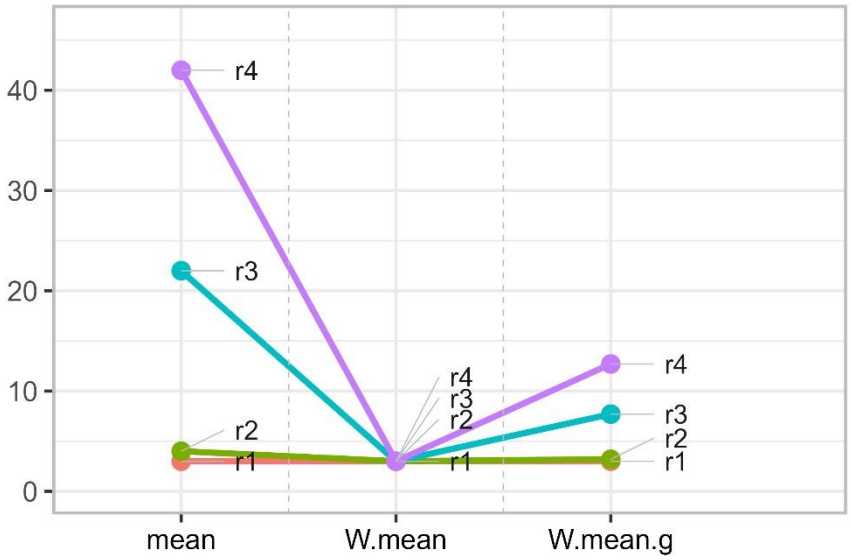
```
V=c(10,19,14,7,12, 11,7,10,0,1, 0,0,1,12,1, 0,1,2,3,4); D=V2D(V,4)
St="mean, median, min, max, Equi,W.mean, W.mean.g"
C=DF(D,StatR(D,St)); RND(C,"6,11,12=1; 10,11=3")

   c1 c2 c3 c4 c5 mean   Equi min max Balance W.mean W.mean.g
r1 10 19 14  7 12 12.4     12   7  19   0.125   12.0     12.0
r2 11  7 10  0  1  5.8      7   0  11  -0.300    5.8      7.1
r3  0  0  1 12  1  2.8      1   0  12   0.692    0.6      1.5
r4  0  1  2  3  4  2.0      2   0   4   0.000    2.0      2.0
```

上のデータ(D)の r3(0,0,1,12,1)を見ると，12 が全体から大きく外れていることがわかります。その平均は 2.8 ですが，ほとんどのデータは平均以下になっています。一方，平均以上は 12 だけです。よって，平均値(=2.8)は全体を代表している，とは言えません。むしろ最大値(=12)によって大きく偏った値と見なされます。このような偏りのあるデータでは平均値を使って全体の概要を示すことができないため，その代わりに中央値(=1)が使用されることがあります。中央値はたしかにデータの中央に位置する値なので，それ以上のデータ数とそれ以下のデータ数が等しく釣合いがとれています。しかし，中央値では最大値(と最小値)の値は完全に無視されています。よって，最大値が 2, 3, 10, 200, 2000 でもまったく同じ中央値(=1)になります。

ここで，{0,0,1,1,12}の最大値として 10, 20, 100, 200, 1000, 2000 を想定して，それぞれの平均値を計算すると次のようになりました。

```
U=1:4; V=c(U,5, U,10, U,100, U,200); D=V2D(V,4); D
Ds=(StatR(D,'T.mean')   -StatR(D,'mean'))/StatR(D,'mean'); Ds1=as.vector(Ds)
Ds=(StatR(D,'T.mean.g')-StatR(D,'mean'))/StatR(D,'mean'); Ds2=as.vector(Ds)
St="mean, median, Equi, W.mean, W.mean.g"; C=DF(D,StatR(D,St),Ds1,Ds2);
L(C,cn=T)
RND(C,"6:7,9,10=1; 8,11,12=3")
G=gLine(C[,c(6,9,10)],m=F,cmx=F,it=F,lb=2,f=12); G; gSave(G,w=400)

   c1 c2 c3 c4  c5 mean median  Equi W.mean W.mean.g    Ds1    Ds2
r1  1  2  3  4   5    3      3 0.000      3      3.0  0.000  0.000
r2  1  2  3  4  10    4      3 0.455      3      3.2 -0.250 -0.219
r3  1  2  3  4 100   22      3 0.941      3      7.7 -0.864 -0.653
```

```
r4  1  2  3  4 200   42      3 0.970      3     12.7 -0.929 -0.699
```



はじめに，平均値(mean)が最大値によって，かなり上方に偏(かたよ)っていることを確認します(r1 → r4)。中央値にはそのような偏りはありませんが，最大値のもつ情報は考慮されず，最大値が増加しても中央値はまったく変化していません。単純ウィンザライズ平均値(W.mean)も同様に，この中央値の問題を抱えています。一方，段階式ウィンザライズ平均値(W.mean.g)では，最大値による偏りの問題を回避しながら最大値の情報を保持している様子がうかがえます。示されたそれぞれの段階式ウィンザイズ平均値の数値も直感的に納得できると思います。

単純トリム平均値と単純ウィンザライズ平均値は切除や代入の対象とする要素の数を設定しなければなりません。その数はデータの分布を観察してから決定します。しかし，そうすると，データごとに処理する要素の数が異なるので，その平均値を比較することができません。また，選択した特定の要素の有無が計算結果に影響するので，分析が恣意的になります。そこで，段階式トリム平均値や段階式ウィンザライズ平均値を用いれば，データの数値を全体的に処理するので分析の個別性・恣意性を回避できます。

4.2.8. 大数平均値

データの外れ値の影響を少なくする方法として前出の「トリム平均値」が考案されました。トリム平均値では最小値と最大値が切除されるので，中央値と同様に，切除された要素の情報は失われます。段階式トリム平均値では，各段階の平均を算出する際，徐々にデータ数が減少しています。初めはデータ全体を含みますが，最後は中央値だけになり，それぞれの平均の母数が異なるので，中央値（付近）のデータの影響度が非常に大きく

なります。ウィンザライズ平均値と段階式ウィンザライズ平均値では最小値と最大値を切除するのではなく，最小値と最大値の位置にそれらに続く数値を代入しています。しかし，やはり最小値と最大値の多くの情報が失われます。

そこで，各段階で最小値と最大値をトリム（切除）するのではなく，データ数の半数以上が含まれるグループ（ソート済み）を左から徐々に切り出しながら，各段階の平均値を求め，その和の平均を求めて「大数平均値」(Majority Mean: M.mean)とする方法を考えてみましょう。以下に昇順にソートしたデータ c(1, 2, 4, 5, 55)を使って具体的にプロセスを示します。データ数は 5 なので，その過半数は 3 個になります。

```
> V=c(1,2,4,5,55); RND(M.mean(V,T),'6=1'); CR(); Rnd(M.mean(V),1)
   c1 c2 c3 c4 c5 mean
r1  1  2  4 NA NA  2.3
r2 NA  2  4  5 NA  3.7
r3 NA NA  4  5 55 21.3

[1] 9.1
```

(r1)　(1 + 2 + 4} / 3　= 2.3

(r2)　(2 + 4 + 5) / 3　= 3.7

(r3)　(4 + 5 +55) / 3　= 21.3

(*)　(2.3 + 3.7 + 21.3) / 3 = 9.1 #M.mean

最初は一番左の要素 c1, c2, c3 の数値 1, 2, 4 の平均値(=2.3)を求めます(r1)。次に，右に続く 3 要素 c2,c3,c4 の数値 2,4,5 の平均値(=3.7)を求めます(r2)。最後に c3,c4,c5 の数値 4,5,55 の平均値(21.3)を計算します(r3)。ここで 3 要素の連続は最後になるので，計算を終了し，r1, r2, r3 で求めた平均値の平均を計算し(=9.1)，これをデータの大数平均値とします。

この大数平均値の方法は全データの中から過半数である 3 連続の標本をすべて抽出しているので，データの抽出法に「偏り」はありません。そして，データの中心にある中央値を含む標本が一番多く，周辺にある数値を含む標本が次第に少なくなり，最小値と最大値を含む標本はそれぞれ 1 個に限られているので，平均値に求められる「中心性」が確保されています[23]。

プログラム(R)

```
M.mean=function(V,p=F){ # Majority mean. V:vt, p:rocess
  n=Len(V); w=floor((n+1)/2); e=floor((n+2)/2); W=M=NULL;
```

[23] 一方，段階式トリム平均値とウィンザライズ平均値では，中央値からの距離の大きい方から切除したり調整したりしているので，それぞれの段階で中心性を確保できません。一方，中央値からの距離を考慮しないで，最小値と最大値の両側から同時に切除したり調整したりすると，中央値からの距離が小さい要素がもつ距離の情報が失われます。

```
   for(i in 1:e){U=V; U[-(i:(i+w-1))]=NA; M[i]=Mean(U); W=rbind(W,U)}
   if(!p) Mean(M)
else {colnames(W)='c'&1:n;rownames(W)='r'&1:e;DF(W,mean=M)}}
# V=c(1,2,4,5,55); M.mean(V) # 9.111111
# M.mean(1:5,F); CR(); M.mean(1:5,T)
```

次は，ほかのデータ例を使ってそれぞれの平均値を比べた結果です。

```
V=c(10,19,14,7,12, 11,7,10,0,1, 0,0,1,12,1, 0,1,2,3,4); D=V2D(V,4)
Dis=(StatR(D,'M.mean')-StatR(D,'mean'))/StatR(D,'mean')
Dis=as.vector(Dis)
St="mean, median, Equi, M.mean"; C=DF(D,StatR(D,St),Dis);
RND(C,"6:7,9=1; 8,10=3")

   c1 c2 c3 c4 c5 mean median   Equi M.mean   Dis
r1 10 19 14  7 12 12.4     12  0.125   12.9 0.039
r2 11  7 10  0  1  5.8      7 -0.300    6.2 0.073
r3  0  0  1 12  1  2.8      1  0.692    3.1 0.111
r4  0  1  2  3  4  2.0      2  0.000    2.0 0.000
```

ここで確認しておきたいことは，大数平均値はデータに偏りが少ないときは平均値に近似し(r1)，完全に偏りがないデータでは(r4: Equi=0.000)では平均値と一致することです。よって，大数平均値はデータの分布の偏りが大きいときは(r3: Equi=0.692)，平均値を大きく修正し(r3: Dis=0.111: 11%)[24]，偏りが小さいときは平均値に近似するので，一般にすべてのデータに同じ平均値を用いて比較するときに有用です。

4.2.9. 外れ値調整平均値の比較

これまでに，外れ値による平均値の偏りを回避する方法として，段階式トリム平均(T.mean.g)，段階式ウィンザライズ平均値(W.mean.g)，大数平均値(M.mean)を提案してきました[25]。最後に，これらの 3 種の平均値を同じ簡単なサンプルを使って比較します。

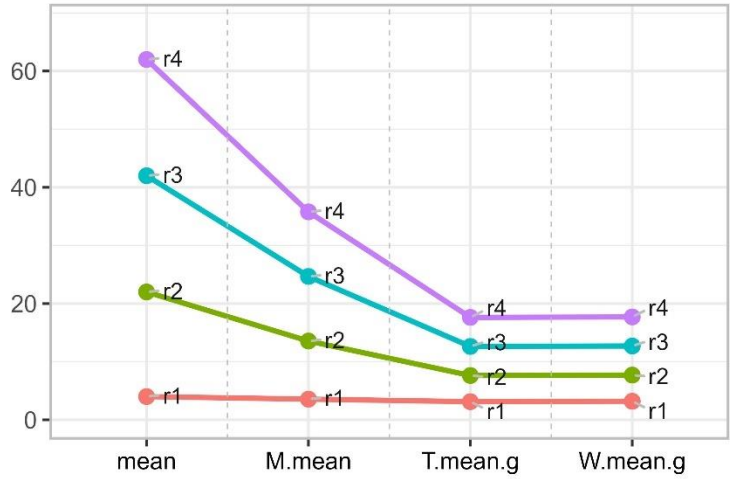
```
U=1:4; V=c(U,10, U,100, U,200, U,300); D=V2D(V,4)
St="mean, median, Equi, M.mean, T.mean.g, W.mean.g"
C=DF(D,StatR(D,St)); RND(C,"6:7,9:11=1; 8=3")
G=gLine(C[,c(6,9:11)],m=F,cmx=F,it=F,lb=3,f=12); G; gSave(G,w=400)

   c1 c2 c3 c4  c5 mean median  Equi M.mean T.mean.g W.mean.g
r1  1  2  3  4  10    4      3 0.455    3.6      3.0      3.1
r2  1  2  3  4 100   22      3 0.941   13.6      6.6      6.7
r3  1  2  3  4 200   42      3 0.970   24.7     10.6     10.7
```

[24] Dis = (M.mean-mean) / mean.

[25] 単純トリム平均値と単純ウィンザライズ平均値は，前述したように，トリム・ウィンザライズする要素数の決定が恣意的・個別的(ad hoc)になるので，ここでは考察しません。

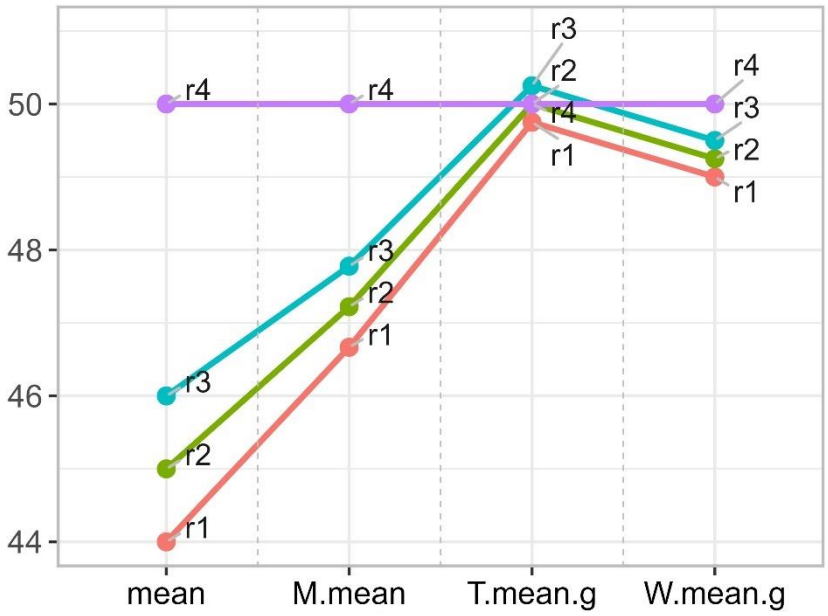
```
r4  1  2  3  4 300   62       3 0.980   35.8      14.6      14.7
```

この結果を見ると，3 種の平均値はそれぞれ最大値による平均値の偏りを回避していますが，その中で大数平均値(M.mean)が一番大きく最大値の情報を生かしていることがわかります。次のグラフを見ると段階式トリム平均(T.mean.g)，段階式ウィンザライズ平均値(W.mean.g)と比べて，大数平均値(M.mean)は，平均値への近似と外れ値の調整をより適切に実現している様子がわかります。r1 と r4 のデータを比較すると，大数平均値は r4 で平均値に近似し，逆に r1 では大きく平均値から乖離します。段階式トリム平均(T.mean.g)と段階式ウィンザライズ平均値でも同じ傾向が見られますが，その変化は小さく感知されにくいと思います。



次に最小値などの下側の外れ値が全体の平均値に及ぼす影響を見ます。

```
V=c(0,40,50,60,70, 5,40,50,60,70, 10,40,50,60,70, 30,40,50,60,70); D=V2D(V,4)
St="mean, median, Equi, M.mean, T.mean.g, W.mean.g"
C=DF(D,StatR(D,St)); L(C,cn=T); RND(C,"6:7,9:11=1; 8=3")

   c1 c2 c3 c4 c5 mean median   Equi M.mean T.mean.g W.mean.g
r1  0 40 50 60 70   44     50 -0.333   46.7     49.8     49.0
r2  5 40 50 60 70   45     50 -0.294   47.2     50.0     49.3
r3 10 40 50 60 70   46     50 -0.250   47.8     50.3     49.5
r4 30 40 50 60 70   50     50  0.000   50.0     50.0     50.0
```



このデータでは，c2:c5 の数値に比べて c1 データは頻度は小さくなっています。大小のバランス(Equi)はどれもマイナスなので平均値は r4 を除いて最小値に向かって傾いています。とくに r1 の偏りは大きく，c1=0 は完全な外れ値と見なされます。r4 では完全に大小関係が平衡しているので(Equi=0)，すべての平均値(mean，T.mean,g，W.mean.g)と中央値は一致しています。

大数平均値(M.mean)は最小値と最大値を同等に扱って除外します。一方，段階式トリム平均(T.mean.g)と段階式ウィンザライズ平均値(W.mean.g)では，中央値からの距離を考慮しながらそれぞれの段階の最小値と最大値を処理します。そして，段階式トリム平均値(T.mean.g)と段階式ウィンザライズ平均値(W.mean.g)を比べると，段階式トリム平均値では，トリムされた要素の情報は失われますが，段階式ウィンザライズ平均値では外れ値は残りの数値によって調整されています。上のグラフを見ると，段階式トリム平均(T.mean.g)と段階式ウィンザライズ平均値(W.mean.g)は類似した結果を示しますが，大数平均値(M.mean)はより平均値(mean)に近いので，外れ値の情報を生かしています。よって，これらの平均値の中で大数平均値(M.mean)がベストだと思われます。

4.3. 中央値と四分位値

中央値(median)はデータを昇順に並べ替えて，その順位のちょうど中央にあるデータの値です。たとえば下の 2 列の成分{8, 10, 4, 7, 9}を昇順に並べ替えて{4, 7, 8, 9, 10}とし，その中央(全体の 50%の位置)にある 3 つめの成分の値 8 が中央値です。データの個数が偶数のときは中央の 2 つのデータの平均をとります。第 1 四分位値(first quartile)は全体の 1/4 の位置(25%

の位置)にあたる値であり，第 3 四分位値(third quartile)は全体の 3/4 の位置(75%の位置)にあたる値です。よって中央値は第 2 四分位値(second quartile)と同じことになります。

S1	v1	v2	v3	v4	v5
h1	10	19	14	7	12
h2	11	7	10	0	1
h3	0	0	1	12	1
h4	0	1	2	3	3

横軸	第 1 四分位値	中央値	第 3 四分位値
h1	10	12	14
h2	1	7	10
h3	0	1	1
h4	1	2	3

四分位値(第 1，第 2，第 3)は Excel 関数の QUARTILE（配列，戻り値:1, 2, 3）を使って求めることができますが，JavaScript にはその関数がないために，次のようにして関数を用意します。

データ D = [d(1), d(2), …, d(n)]を昇順にソートしたデータを S = [s(1), s(2), …, s(n)]とします。ここで，次のように n 個の数値が並ぶ数直線を想定します。

1 ------- q1 ------- q2 ------- q3 ------- n

第 1 四分位値 q1 は全体を 1:3 に内分する点です。そして，第 2 四分位値 q1 は全体を 1:1 に内分し，第 3 四分位値 q1 は全体を 3:1 に内分する点です。

はじめに q1 を求めます。

(q1 - 1) : (n - q1) = 1 : 3
3 * (q1 - 1) = (n - q1)
3 * q1 - 3 = n - q1
4 * q1 = n + 3
q1 = (n + 3) / 4

q2 は

(q2 - 1) : (n - q2) = 1 : 1
q2 - 1 = n - q2
2 q2 = n + 1
q2 = (n + 1) / 2

q3 は

(q3 - 1) : (n - q3) = 3 : 1
q3 - 1 = 3 * (n - q3)
q3 - 1 = 3n - 3 q3
4 q3 = 3n + 1
q3 = (3n + 1) / 4

このようにして求めた q1, q2, q3 が，それぞれ整数であれば，S の配列内の位置によって，それぞれの四分位値が求められます。たとえば，データの個数が 9 個であれば，q1 = 3, q2 = 5, q3 = 7 になり，その位置にある，s(3), s(5), s(7)を第 1 四分位値，第 2 四分位値，第 3 四分位値とします。

q1，q2，q3 が整数でなく小数になる場合は，この小数を切り下げた整数(pl)と切り上げた整数(ph)にそれぞれ配置された数値 s(pl)と s(ph)を比例配分した値を求めます。たとえば，n = 8 のとき q1 = (n + 3) / 4 = 2.75 となるので，配列中の s(2)と s(3)を使って，

第 1 四分位値 = 0.25 * s(2) + 0.75 * s(3)

とします。このとき小数部が 0.75 なので，全体 1 の 0.25 と 0.75 を使って s(2)と s(3)に配分しています。第 2 四分位値と第 3 四分位値についても同様です。

プログラム (JS)

```
function quartile(Xn, sel) {
//四分位数(sel=1:第1四分位数, 2:第2四分位数, 3:第3四分位数) = Excel
  var Sn = sortC(Xn, 1, 1, true), n = Xn.length-1;
        if (sel==1) {var q = (n + 3) / 4;}
        if (sel==2) {var q = (n + 1) / 2;}
        if (sel==3) {var q = (3 * n + 1) / 4;}
        if (Math.round(q) == q) { //qが整数であれば
                return Sn[q][1];
        }
        else { //qが整数でなければ
                var pl = Math.floor(q), xl = Sn[pl][1];
                var ph = Math.ceil(q), xh = Sn[ph][1];
                return (ph - q) * xl + (q - pl) * xh;
        }
}
```

●平均中央対照値

データの分布が正規分布のように左右対称であれば昇順(降順)の頻度分布は平均値(M) と中央値(Md)が同値になります(M = Md)。一方，データが対数正規分布のような分布を示せば右側の裾が長くなり平均値(M)が中央値(Md)よりも大きくなります(M > Md)。そこで平均値(M)と中央値(Md)の「平均中央対照値」(Mean Median Contrast: MMC)を使えば，平均値と中央値の間の大小関係と両者の一致・乖離の程度がわかります。同様にして幾何平均値(GM: 後述)と中央値(Md)の間の「幾何平均中央対照値」(Geometric

Mean Median Contrast: GMMC)を求めます[26]。

MMC = (M - Md) / (M + Md)
GMMC = (GM - Md) / (GM + Md)

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

横軸	平均値	幾何平均値	中央値	MMC	GMMC
h1	12.400	11.744	12.000	.016	-.011
h2	5.800	2.384	7.000	-.094	-.492
h3	2.800	.654	1.000	.474	-.209
h4	1.800	1.125	2.000	-.053	-.280

■スペイン語語彙頻度データの平均中央対照値

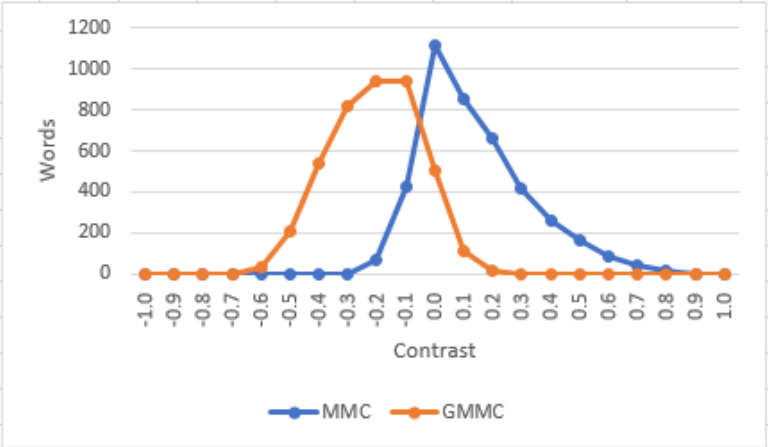
次の表は，先の「スペイン語語彙頻度データ」(→「幾何平均値」)を使って計算した，平均値(M)，幾何平均値(GM)，中央値(Md)，平均中央対照値(MMC)と幾何平均中央対照値(GMMC)です。

Word	M	GM	Md	MMC	GMMC
de (prep)	7221.6	6816.9	7196.0	0.002	-0.027
que-qué (conj|pn)	3662.4	3536.1	3243.5	0.061	0.043
y (conj)	3268.4	3199.7	3161.5	0.017	0.006
a (prep)	2847.3	2835.9	2844.0	0.001	-0.001
en (prep)	2501.1	2431.8	2715.0	-0.041	-0.055
ser (v)	1940.3	1849.1	1702.5	0.065	0.041
haber (v)	1225.3	1160.3	1223.0	0.001	-0.026
por (prep)	1085.6	1070.8	1088.5	-0.001	-0.008
su-suyo (poses)	971.7	908.4	1050.0	-0.039	-0.072

次の表とグラフは，平均値(M)と中央値(Md)，幾何平均値(GM)と中央値(Md)の間の関係を見るために，平均中央対照値(MMC)と幾何平均中央対照値(GMMC)を[-1, 1]の範囲で 0.1 ごとに該当する単語の数を集計した結果です。

[26] データ全体の左右の歪(ゆが)みの程度は後述の「歪度」(わいど)で計測されます。

Contrast	MMC	GMMC
-1.0	0	0
-0.9	0	0
-0.8	0	0
-0.7	0	0
-0.6	0	30
-0.5	0	208
-0.4	0	542
-0.3	2	818
-0.2	67	938
-0.1	428	**942**
0.0	**1115**	507
0.1	849	112
0.2	664	12
0.3	417	2
0.4	262	0
0.5	161	0
0.6	90	0
0.7	39	0
0.8	15	0
0.9	2	0
1.0	0	0



この表とグラフを見ると，平均値の最大値(1115)は中央値とほぼ一致しますが(MMC = 0.0)，平均値が中央値を超えるケースが非常に多くなっています(MMC > 0)。一方，幾何平均値は，その最大値(942)が中央値のやや左側にあり(GMMC < 0)，全体的に中央値の左側に多く分布して，中央値を大きく超えることはありません。これは，平均値が非常に大きな数値に影響されやすく，幾何平均値がその影響が少ないことを示しています。

4.4. 最大値・最小値・中間値・範囲

データ行列の「最大値」(Maximum)，「最小値」(Minimum)，「範囲」(Range)，「中間値」(Mid)を，それぞれの相（行，列，全体）で計算します。範囲は最大値から最小値を引いた値です。中間値は(最大値+最小値) / 2 の値です。中間値は「範囲中央」または「ミッドレンジ」(Mid-Range)と呼ばれますが，ここでは簡単に「中間値」(Mid)と呼ぶことにします。

X	v1	v2	v3	v4	v5
d1	10	19	14	7	12
d2	11	7	10	0	1
d3	0	0	1	12	1
d4	0	1	2	3	3

Horizontal	Minumum	Maximum	Mid	Range
d1	7	19	13.0	12
d2	0	11	5.5	11
d3	0	12	6.0	12
d4	0	3	1.5	3

4.5. 最頻値

データの中で最も多く現れる数値は「最頻値」(Mode)と呼ばれます。

D	v1	v2	v3	v4	v5
d1	10	19	14	7	12
d2	11	7	10	0	1
d3	0	0	1	12	1
d4	0	1	2	3	3

横軸	最頻値	最頻値:頻度
d1	No mode	No mode
d2	No mode	No mode
d3	No mode	No mode
d4	3	3: 2

このデータの d1 と d2 ではどれも異なる数値なので最頻値がありません(No mode)。d3 では 0 と 1 がそれぞれ 2 回ずつ現れているので最頻値が決定できません(No mode)。d4 では 3 が 2 回現れているので，これが最頻値になります。これを「データ最頻値」と呼ぶことにします。

次に,上の表を階級 1:[0, 4]，階級 2:[5, 9]，階級 3:[10, 14]，階級 4:[15, 19]という 4 つの階級に分けると次の結果になります。

D	v1	v2	v3	v4	v5
d1	3	4	3	2	3
d2	3	2	3	1	1
d3	1	1	1	3	1
d4	1	1	1	1	1

横軸	最頻値	範囲	中間値
d1	3	[10, 14]	12
d2	No mode		
d3	1	[0, 4]	2
d4	1	[0, 4]	2

このように一般に最頻値の計算では階級に分けて，一番多くの値をもつ階級の中間値を最頻値とします。これを「階級最頻値」と呼ぶことにします。当然，階級最頻値は階級の設定の仕方で異なります。

●大数最頻値

一般の階級最頻値は階級の設定方法に依存するので，データそのものの最頻値（データ最頻値）とは異なります。そして，データ最頻値はそれぞれのデータが異なる数値を示しているときには役立ちません。また，たとえば{2, 2, 3, 3, 7, 10, 10, 10}のように，最頻値(10)と 2, 3…番目に頻度が高い値(6, 2)が離れているときは，最頻値(10)だけがデータの「最頻性」を代表している，とは言えないでしょう。この場合，{2, 2, 3, 3}というセットのほうが{10, 10, 10}というセットよりも最頻性が高いと考えます。

そこで数値が集中しているデータの探し方として，データの過半数ができるだけ狭い範囲に集中しているデータセットの平均を,「集中した数値」として代表させる方法を考えます。これを**大数最頻値**(Majority mode)と呼びます。ここで「最も多く存在する同一の値」という「最頻値」の概念を「もっとも近い値が半数以上存在するセットの中心」という概念に拡大します。

たとえば d1 = {10, 19, 14, 7, 12}をソートした{7, 10, 12, 14, 19}という行について，次のように個数=5 の過半数 3 個で一番小さい数値範囲のセット（下線）を探します。

セット 1: {7, 10, 12, 14, 19} 範囲: 12 – 7 = 5
セット 2: {7, 10, 12, 14, 19} 範囲: 14 – 10 = 4
セット 3: {7, 10, 12, 14, 19} 範囲: 19 – 12 = 7

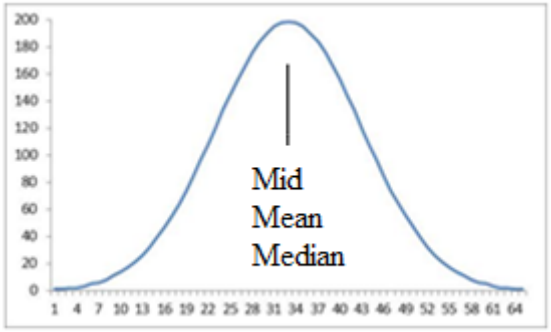
ここでセット 2 の範囲(14-10=4)が一番小さいのでここに近い数値が集中しています。このデータセット{10, 12, 14}の平均 12 を大数最頻値とします。最小の範囲が複数あるときには，セットの幅を 3→4→5 のように 1 つずつ増やして検索を続けます。極端な場合として，{3, 4, 5, 6, 7}のように均等に連続するデータセットでは，どのような幅のデータセットをとっても集約させることができません。そのときは最大の幅としてデータの大きさ(N=5)を使うことになり，この場合は平均値と等しくなります。

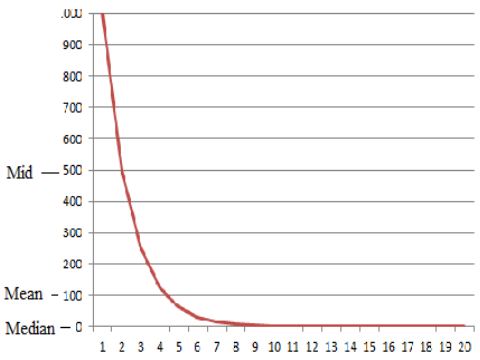
D	v1	v2	v3	v4	v5
d1	10	19	14	7	12
d2	11	7	10	0	1
d3	0	0	1	12	1
d4	0	1	2	3	3

横軸	大数最頻値	大数最頻値:範囲
d1	12.000	10 - 14
d2	9.333	7 - 11
d3	.500	0 - 1
d4	2.667	2 - 3

■言語データの L 字型分布

身長や学力など，「正規分布」と呼ばれる分布を示すデータの頻度とその順位は下図（行：順位，列：頻度）のようになります。つまり，順位の最下位と最上位の数は少なく，多数が平均値の近くに集まります。一方，言語データ（文字，音韻，単語など）は，高順位のデータ（少数）の頻度がきわめて高く，低順位のデータ（多数）の頻度がきわめて低い，という特徴を示します。これは「L 字型分布」と呼ばれています。以下で示すように，正規分布を示すデータと L 字型分布を示すデータは扱い方が異なります。





4.6. 変動

データを説明するときは平均や中央値などのデータの中心を示す値だけでなく，同時にデータの変動も示すとよいでしょう。たとえば同じ平均気温が 20 度の土地でも，1 年を通してほとんど 10-30 度である土地と，それが 0-40 度の土地では気温のあり方は大きく異なります。このセクションでは，そのような変動のあり方を示す数値を扱います。

4.6.1. 分散・標準偏差

2 つのデータ，{D1: 4, 5, 6, 7, 8}と{D2: 2, 4, 6, 8, 10}の平均を比べてみましょう。どちらも平均は 6 で同じですが，データのばらつきが異なります。ばらつき具合を計るには，データの偏差（平均からの差）が必要になります。そこで，{D1: 4-6, 5-6, 6-6, 7-6, 8-6}と{D2: 2-6, 4-6, 6-6, 8-6 10-6}のようにそれぞれ平均を引いて，偏差{D1: -2, -1, 0, 1, 2}と{D2: -4, -2, 0, 2, 4}を作ります。偏差を全部足すと，どちらも 0 になってしまうので，データの散らばり方を比較できません。そこで，それぞれのデータを{D1: $(-2)^2$, $(-1)^2$, 0^2, 1^2, 2^2} y {D2: $(-4)^2$, $(-2)^2$, 0^2, 2^2, 4^2} のように 2 乗して，{D1: 4, 1, 0, 1, 4}と{D2: 16, 4, 0, 4, 16}とし，負の値をなくし，すべて正の値にします。その総和{D1: 4+1+0+1+4}と{D2: 16+4+0+4+16}が「分散」(Variance: V)です。それぞれ，V(D1) = 10, V(D2)=40 になります。

$$Vr = \Sigma (X - M)^2 / N$$

分散の計算ではもとのデータを 2 乗しているため，データの規模よりも大きくなっています。そこで，その平方根をとって，もとのデータの規模に直した数値が「標準偏差」(Standard Deviation: SD)です。この例では SD(D1) = $10^{1/2}$ ≒ 3.16, Sd(D2) = $40^{1/2}$ ≒ 6.32 になります。

$$SD = (Vr)^{1/2}$$

次は行の分散(V)と標準偏差(SD)を導出する行列式です。

$S_{n1} = X_{np}\ I_{p1}$ ←横和列(N:行数 ; P:列数)
$M_{n1} = S_{n1} / P$ ←横平均列(P:列数)
$D_{np} = X_{np} - M_{n1}$ ←偏差行列
$C_{np} = D_{np}^2$ ←偏差 2 乗行列
$W_{n1} = C_{np}\ I_{p1}$ ←偏差 2 乗和列
$V_{n1} = W_{n1} / P$ ←分散列
$SD_{n1} = V_{n1}^{1/2}$ ←標準偏差列

行列関数で示すと

V_{n1} = D(X(E(S(X_{np}, D(X(X_{np}, Ip1), P)), 2), Ip1), P)

SD_{n1} = E(V_{n1}, 1/2)

はじめに横平均(M)を求め，データ行列(X)から横平均（縦ベクトル M）を引いて偏差行列(D)を作ります[27]。このようにして出来上がった行列(D)は，行列のそれぞれの要素が横平均からの偏差を示しています。この大きさの平均を求めますが，偏差の和はゼロになってしまいますから，はじめに行列(D)全体を 2 乗にします(C)。その和を求め(W)，列右数(P)で割った値が横分散列(V)です。そして，横分散列の 2 乗根が横標準偏差列(Sd)です。

X_{np}	v1	v2	v3	v4	v5
d1	10	19	14	7	12
d2	11	7	10	0	1
d3	0	0	1	12	1
d4	0	1	2	3	3

Ip1	1
1	1
2	1
3	1
4	1
5	1

S_{n1}	1
1	62
2	29
3	14
4	9

M_{n1}	1
1	12.40
2	5.80
3	2.80
4	1.80

D_{np}	1	2	3	4	5
1	-2.40	6.60	1.60	-5.40	-.40
2	5.20	1.20	4.20	-5.80	-4.80
3	-2.80	-2.80	-1.80	9.20	-1.80
4	-1.80	-.80	.20	1.20	1.20

C_{np}	1	2	3	4	5
1	5.76	43.56	2.56	29.16	.16
2	27.04	1.44	17.64	33.64	23.04
3	7.84	7.84	3.24	84.64	3.24
4	3.24	.64	.04	1.44	1.44

W_{n1}	1
1	81.20
2	102.80
3	106.80
4	6.80

V_{n1}	1
1	16.24
2	20.56
3	21.36
4	1.36

SD_{n1}	1
1	4.03
2	4.53
3	4.62
4	1.17

●算術平均値±標準偏差

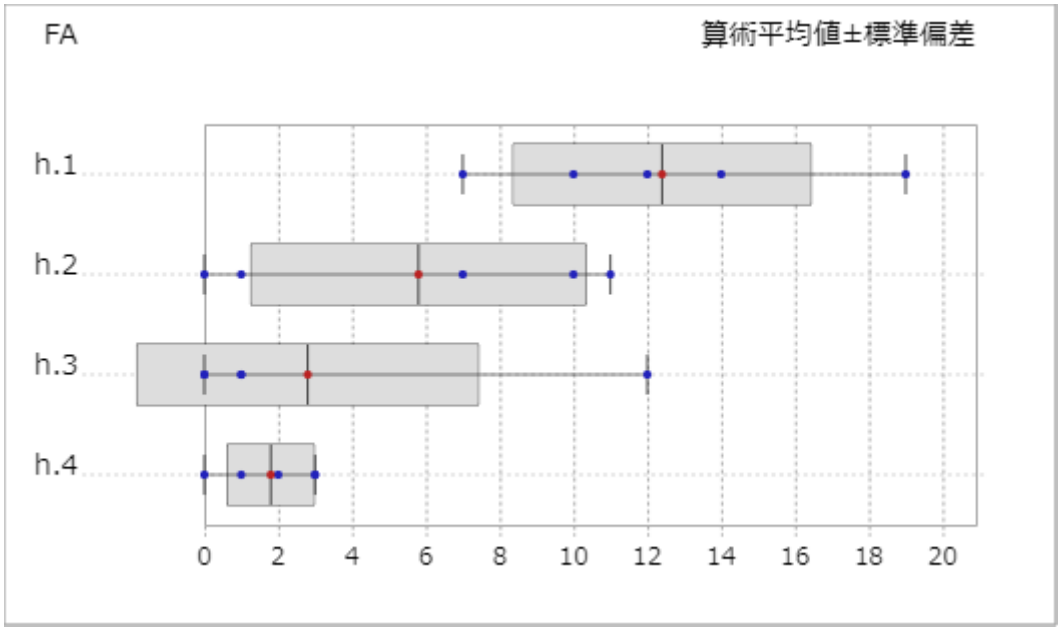
データの頻度分布の中心を記述するとき，一般に平均だけでなく，そのバラツキを示す標準偏差も併記されます。このとき，たとえば次の例{10, 19, 14, 7, 12}では 12.4 ± 4.03 のようにプラスマイナスの記号(±)が使われます。このようにして，多くのデータが「算術平均(AM)＋標準偏差(SD)」と「算術平均(AM)－標準偏差(SD)」の範囲にあることが示されます。

FA	A	B	C	D	E	AM	SD	AM-SD	AM+SD
h.1	10	19	14	7	12	12.400	4.030	8.370	16.430

[27] 行列から縦ベクトルを引くという演算は一般の線形代数の本には定義されていませんが，このテキストでは 1 章で定義してあります。

h.2	11	7	10	0	1	5.800	4.534	1.266	10.334
h.3	0	0	1	12	1	2.800	4.622	-1.822	7.422
h.4	0	1	2	3	3	1.800	1.166	0.634	2.966

次のグラフは，算術平均値を中心として，その左右に平均値―標準偏差と平均値＋標準偏差をボックスで囲んだボックスチャートです。



h.3 を見ると，平均値よりも標準偏差の方が大きいので，平均値―標準偏差の位置がマイナスになっています。多くのデータが「平均値±標準偏差」の範囲にあるのはデータの頻度分布が正規分布に近似しているときです[28]。そうでないときは，多くの場合に先述の幾何平均と後述の幾何標準偏差が有効です。

●標準偏差率

現実のデータの頻度分布が完全な(完全に近い)正規分布を示すことはほとんどありません。そこで，データの平均±標準偏差の範囲の個数(Number of mean ± standard deviation: NMSD)が全体の個数(N)の中で示す率を「標準偏差率」(Standard deviation ratio: SDR)として計算し，NMSD が全体の中で占める割合を見ます。

SDR = NMSD / N

[28] データの頻度が完全に正規分布を示すならば「平均±標準偏差」の範囲に約 68%のデータが入ります。たとえば，平均=50，標準偏差=10 のデータでは

NORMDIST(50+10,50,10,1) = 0.841
NORMDIST(50-10,50,10,1) = 0.159
0.841 - 0.159 = 0.683

次の表は，データ(左表)の算術平均(AM)，標準偏差(SD)，平均値-標準偏差値(AM-SD)，平均値+標準偏差値(AM+SD)，標準偏差率(SDR)を示します。

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

横軸	AM	SD	AM+SD	AM-SD	SDR
h1	12.4	4.0	16.4	8.4	.600
h2	5.8	4.5	10.3	1.3	.400
h3	2.8	4.6	7.4	-1.8	.800
h4	1.8	1.2	3.0	0.6	.400

たとえば，h1 の標準偏差率(SDR)はデータの成分{10, 19, 14, 7, 12}の中で範囲[8.4, 16.4]の中にある成分{10, 14, 12}の個数 3 を総数 5 で割った値.600 になります。

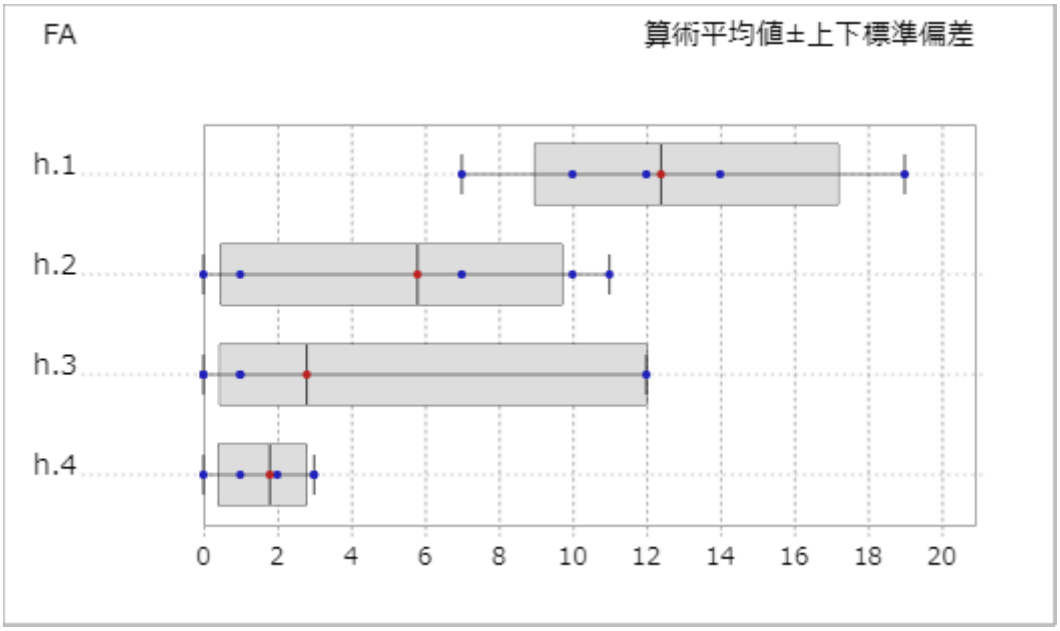
●算術平均値±上側・下側標準偏差

先のデータ(M)の h3 を見ると，「算術平均値－標準偏差」(AM-SD)がマイナス(-1.8)になっています。これはデータの中に平均値を大きく上回る値があって，それが標準偏差を大きくしているからです標準偏差値を平均値の上側だけでなく下側にも適用しているため，このように下側でマイナス値が生じることがあります。

そこで，算術平均値より上にある成分のデータの標準偏差と算術平均値より下にある成分のデータの標準偏差を分けて計算します。それぞれを「上側標準偏差」(Upper standard deviation: USD)，「下側標準偏差」(Lower standard deviation: LSD)と呼びます。次がその計算結果です。

M	A	B	C	D	E	AM	SD	LSD	USD	AM-LSD	AM+USD
h1	10	19	14	7	12	12.4	4.030	3.42	4.802	8.98	17.202
h2	11	7	10	0	1	5.8	4.534	5.324	3.921	0.476	9.721
h3	0	0	1	12	1	2.8	4.622	2.354	9.2	0.446	12
h4	0	1	2	3	3	1.8	1.166	1.393	0.987	0.407	2.787

上の表の「算術平均＋上側標準偏差」(AM+USD)は最大値を上回ることはなく，「算術平均－下側標準偏差」(AM-LSD)は最大値を上回ることはありません。次のグラフで確認してください。



●変動係数

標準偏差はデータの規模（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(SD)を平均(M)で割った値が「変動係数」(Coefficient of Variation: CV)です[29]。

CV = SD / M

標準偏差(SD)も平均(M)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(CV)によって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。そして，変動係数は平均と比べた標準偏差の規模（倍数）を示しているので，その大きさの評価が容易です。

下右表の「確率」は変動係数の乱数累積確率を示します。

X	v1	v2	v3	v4	v5
d1	10	19	14	7	12
d2	11	7	10	0	1
d3	0	0	1	12	1
d4	0	1	2	3	3

横軸	変動係数
d1	.325
d2	.782
d3	1.651
d4	.648

●相対分散・相対標準偏差

分散を[0, 1]の範囲に限定した値を「相対分散」(Relative Variance: R.Vr)と呼びます。相対分散(R.Vr)は分散(Vr)をその理論的な最大値(Vr.max)で割ることで求めます。

[29] 参照：芝他『統計用語辞典』（新曜社）

$$R.Vr = Vr / Vr.max$$

先に見たように分散(Vr)は次のように定義されています(M:平均; N:個数)。

$$Vr = [(X_1 - M)^2 + (X_2 - M)^2 + \cdots + (X_n - M)^2] / N$$

ここで，たとえば{10, 0, 0, 0, 0}というように，要素の 1 つだけに数値があり，残りはすべて 0 であるデータを考えましょう。このようなとき分散が最大値になります。一般化して{K, 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$(K - M)^2$ になり，残り N – 1 個の成分はどれも$(0 - M)^2 = M^2$ になります。よって分散の最大値(Vr.max)は

$$\begin{aligned} Vr.max &= [(K - M)^2 + (N - 1)(0 - M)^2] / N \\ &= [(K - M)^2 + (N - 1) M^2] / N \end{aligned}$$

このとき K 以外にデータがないので K が総和 S になります。
よって

$$K = S = N M \qquad \leftarrow \text{平均 M = 総和 S / 個数 N}$$

分散の最大値 Vr.max は

$$\begin{aligned} Vr.max &= [(K - M)^2 + (N - 1) M^2] / N \\ &= [(N M - M)^2 + (N - 1) M^2] / N && \leftarrow K = N M \\ &= [((N - 1) M)^2 + (N - 1) M^2] / N && \leftarrow M \text{ を外へ} \\ &= (N - 1) M^2 [(N - 1) + 1] / N && \leftarrow M^2 (N - 1) \text{が共通} \\ &= M^2 (N - 1) && \leftarrow N \text{ を整理} \end{aligned}$$

よって，相対分散(R.Vr)は

$$R.Vr = Vr / Vr.max = Vr / [M^2 (N - 1)]$$

同様にして「相対標準偏差」(Relative Standard Deviation: R.SD)は

$$\begin{aligned} R.SD &= (R.Vr)^{1/2} = (Vr / Vr.max)^{1/2} = \{Vr / [M^2 (N - 1)]\}^{1/2} \\ &= SD / [M (N - 1)^{1/2}] \end{aligned}$$

S1	v1	v2	v3	v4	v5		横軸	分散	相対分散	標準偏差	相対標準偏差
h1	10	19	14	7	12		h1	16.240	.026	4.030	.162
h2	11	7	10	0	1		h2	20.560	.153	4.534	.391
h3	0	0	1	12	1		h3	21.360	.681	4.622	.825
h4	0	1	2	3	3		h4	1.360	.105	1.166	.324

たとえば，{1, 0, 0, …, 0}と{10, 0, 0, …, 0}の分散（と標準偏差）を比べ

ると当然後者のほうが大きいのですが，相対分散（と相対標準偏差）は，どちらも最大値 1 になり同じ値になります。これは非常に稀なケースです。

相対標準偏差の指揮の分母に成分数(N)が含まれているので，成分数が大きくなると，小さくなります。とくに成分数が 100 を超えると相対標準偏差は 1/10 になります。相対標準偏差が小さな数値でわかりにくいときは，変動係数(CV)を使うとよいでしょう。

なお，相対分散と相対標準偏差は，それらの最大値を{K, 0, 0, …, 0}という N 個のデータについて想定しているので，頻度などの非負のデータについて計算されるものです。

●安定度

私たちはしばしば個数(N)が少ないデータを扱うことがあり，たとえば，わずか 3 個のデータについて求めた平均値がどの程度信頼してよいかわからないことがあります[30]。それが 30 個であれば，かなり信頼できると思われます。そして，個々のデータがばらついていないときの平均値がほうが，それがばらついている平均値よりも信頼できます。

データ自体のばらつきは分散や標準偏差を計算すればわかりますが，ここではデータ全体に一定の操作によって「ばらつき」を加えたデータセットについて計算された平均や分散などが，どのように安定しているのかを観察します。その操作は，1 つのデータが欠損したと仮定して，それを平均値で補う，という作業をデータ数(N)と同じ回数で繰り返します。安定したデータであれば，この作業によって得られた平均や分散などにばらつきが少なくなるはずです。

たとえば，下表(Ave3)の data 行はデータ{1, 2, 3}とその平均(Ave = 2.000)を示します。

Ave3	x1	x2	x3	Ave	Stability
data	1.000	2.000	3.000	2.000	0.904
t1	2.000	2.000	3.000	2.333	
t2	1.000	2.000	3.000	2.000	
t3	1.000	2.000	2.000	1.667	

ここで，data のそれぞれのデータ値(x1, x2, x3)を data の平均値に変換したものが t1, t2, t3 行です。その変動が[0, 1]の範囲となるように，相対標準偏差(R.SD)を使った次の式で「安定度」(Stability)を定義します(.917)。

$$\text{Stability} = 1 - \text{R.SD} = 1 - \text{SD} / [\text{M} (\text{N} - 1)^{1/2}]$$

[30] たとえば，90 頁の本の中からランダムに 3 頁だけを選んで前置詞 de の平均頻度を求めたとき，その頻度はどの程度信頼できるのでしょうか？3 人のネイティブスピーカーに聞いた文の正誤判断の平均値なども同様です。

ここで SD は平均値{2.333, 2.000, 1,667}の標準偏差，N はその個数(3)，M はその平均(2.333+2.000+1.667) / 3 を示します。先に見たように，相対標準偏差(R.SD)はデータのばらつき具合を[0, 1]の範囲で示すので，その 1 の補数(1 - R.SD)がデータの安定度を示していると考えます。

次はデータ数(N)を 4 にして実験した結果です。このように安定度はさらに増しています(.947)。

Ave4	x1	x2	x3	x4	Ave	Stability
data	1	2	3	3	2.250	0.947
t1	2.250	2.000	3.000	3.000	2.563	
t2	1.000	2.250	3.000	3.000	2.313	
t3	1.000	2.000	2.250	3.000	2.063	
t4	1.000	2.000	3.000	2.250	2.063	

次の表で x4 に外れ値(30)を入れると安定度がかなり低下します(.805)。

Ave4	x1	x2	x3	x4	Ave	Stability
data	1	2	3	30	9.000	0.805
t1	9.000	2.000	3.000	30.000	11.000	
t2	1.000	9.000	3.000	30.000	10.750	
t3	1.000	2.000	9.000	30.000	10.500	
t4	1.000	2.000	3.000	9.000	3.750	

次は個々のデータが平均値に近似する例です{12, 14, 15, 20}。このときの安定度は.972 となっています。よって，この平均値(15.250)はかなり信頼できる，と考えてよいでしょう。

Ave4	x1	x2	x3	x4	Ave	Stability
data	12	14	15	20	15.250	0.972
t1	15.250	14.000	15.000	20.000	16.063	
t2	12.000	15.250	15.000	20.000	15.563	
t3	12.000	14.000	15.250	20.000	15.313	
t4	12.000	14.000	15.000	15.250	14.063	

先の安定度(Stability)の式によれば，安定度は R.Sd が小さくなればなるほど最大の 1.000 に近づきます。R.SD が小さくなるのは，標準偏差(SD)が小さく，個数(N)と平均(M)が大きくなるときですが，標準偏差と平均は連動しているので[31]，安定度を左右するのは標準偏差と個数になります。経験的には安定度が.950 を超えると「かなり安定している」と思われ，それが.990 を超えると「きわめて安定している」と思われます。

[31] たとえばデータ全体を 10 倍にしても（{120, 140, 150, 200}），安定度は変わりません。

この安定性の指標は，平均(Ave)に限らず，和，中央値，四分位値，分散，標準偏差や，後述するさまざまな統計量についても同様に計算することができます。次の表はデータ{12, 14, 15, 20}の標準偏差(SD)の安定性を示します(.837)。表の赤で塗りつぶした数値は data の平均値です。

SD4	x1	x2	x3	x4	SD	Stability
data	12	14	15	20	2.947	0.837
t1	15.250	14.000	15.000	20.000	2.321	
t2	12.000	15.250	15.000	20.000	2.863	
t3	12.000	14.000	15.250	20.000	2.944	
t4	12.000	14.000	15.000	15.250	1.279	

●不等度

データ間の量的な差異が大きい状態を示す数値を「不等度」(Inequality: I)と呼びます。その範囲を[0, 1]とします。データ間の差異の程度を数量化するために，次のような 2 乗和の性質を使います。たとえば，次の x1 と x2 のそれぞれの成分間の差異を求めるために，それぞれを 2 乗し，その和(K=x1^2 + x2^2)を求めます。

X	x1	x2	x1^2	x2^2	K=x1^2+x2^2	K.max	K.min	K.lim	不等度
d1	0	10	0	100	100	100	50	1.00	1.00
d2	1	9	1	81	82	100	50	0.64	0.80
d3	2	8	4	64	68	100	50	0.36	0.60
d4	3	7	9	49	58	100	50	0.16	0.40
d5	4	6	16	36	52	100	50	0.04	0.20
d6	5	5	25	25	50	100	50	0.00	0.00
d7	6	4	36	16	52	100	50	0.04	0.20
d8	7	3	49	9	58	100	50	0.16	0.40
d9	8	2	64	4	68	100	50	0.36	0.60
d10	9	1	81	1	82	100	50	0.64	0.80
d11	10	0	100	0	100	100	50	1.00	1.00

上表が示すように，(0, 10)のように差が最大になるとき K は最大値 100 になり，(5, 5)のように差がないとき K は最小値 25 になります。K の最大値(K.max)は比較する 2 成分の最大値(=和 S)の 2 乗です($S^2 = 10^2 = 100$)。K の最小値(K.min)は比較する 2 成分の平均(M)の 2 乗和($M^2 + M^2 = 2M^2$)になります($5^2 + 5^2 = 50$)。そこで，K が最大値と最小値の幅（範囲）の中で占める位置(限定値：K.lim)を次のように計算します。

K.lim = (K - K.min) / (K.max - K.min)

ここで，データを最初に 2 乗しているので，次数をもとのデータにそろえるために K.lim の根をとった値が不等度(I)です。以下の式では，データ(x)を 2 個から N 個に拡張します(x(1), x(2), … x(N))。よって K は $\Sigma X(i)^2$ に，K の最大値 K.max は S2 に，K の最小値 K.min は NM^2 になります。

$$I = K.lim^{1/2} = [(K - K.min) / (K.max - K.min)]^{1/2}$$
$$= [(\Sigma X(i)^2 - N M^2) / (S^2 - N M^2)]^{1/2}$$

実はこの不等度を，先に見た「相対分散」(RVr)から導くことができます。

$$RVr = Vr / [(N - 1) M^2]$$
$$= \Sigma (X(i) - M)^2 / N / [(N - 1) M^2] \quad \leftarrow 分散(Vr): \Sigma (X(i) - M)^2 / N$$
$$= \Sigma (X(i) - M)^2 / (N (N - 1) M^2) \quad \leftarrow 分母を整理$$
$$= \Sigma (X(i)^2 - 2 M X(i) + M^2) / (N (N - 1) M^2) \quad \leftarrow 分子を展開$$
$$= (\Sigma X(i)^2 - 2 M \Sigma X(i) + N M^2) / (N (N - 1) M^2) \quad \leftarrow 分子\Sigma を分配$$
$$= (\Sigma X(i)^2 - 2 M N M + N M^2) / (N (N - 1) M^2) \quad \leftarrow 和=\Sigma X(i)=NM$$
$$= (\Sigma X(i)^2 - N M^2) / (N (N - 1) M^2) \quad \leftarrow 分子を整理$$
$$= (\Sigma X(i)^2 - N M^2) / (N^2 M^2 - N M^2) \quad \leftarrow 分母を展開$$
$$= (\Sigma X(i)^2 - N M^2) / (S^2 - N M^2) \quad \leftarrow N M = S$$

$$= K.lim^2$$

よって不等度(I)は，次のように相対分散(RVr)の根（「相対標準偏差」Relative Standard Deviation: RSD と呼びます）になります。

$$I = (LVr)^{1/2} = [Vr / [(N - 1) M^2]]^{1/2} = SD / [(N - 1)^{1/2} M] = LSD$$

X	v1	v2	v3	v4	v5
d1	10	19	14	7	12
d2	11	7	10	0	1
d3	0	0	1	12	1
d4	0	1	2	3	3

横軸	相対分散	不等度
d1	.026	.162
d2	.153	.391
d3	.681	.825
d4	.105	.324

不等度(I)と変動係数(CV)の違いは，不等度の分母に $(N - 1)^{1/2}$ を掛けていることです。データ行列は一般に N が大きいので，それに応じて不等度は小さくなります。そのような場合には不等度は個体間の得点の変動ではなく，むしろ比較的少数の変数間の変動を見るときに使うべきです[32]。

■均等度・語の使用度

A. Juilland and E. Chang Rodríguez. *Frequency dictionary of Spanish words,* (The Hague: Mouton, 1964)は，5 つの分野（演劇，小説，随筆，科学技術文，

[32] 代替として後述の平均分離度・平均近接度が有効です。

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，単語の「使用度」(Usage: U)を示す数値として次の式を提案しました。

U = F * D

ここで F は単語の頻度(Frequency)を示し，D は分野間の「拡散度」(Dispersion: D)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(拡散度：D)も勘案すべきだという考え方です。そして，次のような拡散度の式が提示されました。

D = 1 – 標準偏差 / (2 * 平均値)

この分母にある 2 は(分野数 5 – 1)$^{1/2}$ のことだと思います。よって次のような関係になります。このテキストではこの「拡散度」を「均等度」(Equality: E)と呼びます。

均等度(E) = 1 – 不等度(I)

X	v1	v2	v3	v4	v5
d1	10	19	14	7	12
d2	11	7	10	0	1
d3	0	0	1	12	1
d4	0	1	2	3	3

横軸	均等度
d1	.838
d2	.609
d3	.175
d4	.676

■線状拡散度

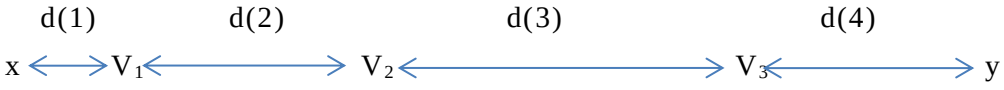
データの性質を見るとき，その頻度と分散を調べることが重要です。集計された頻度データの分散の計算法は先に扱いました。ここでは次のように，連続して続く 1 つの文字データの頻度と拡散度を計算します。

N	Lema
1	l_C
2	¡_B
3	su_T
4	comida_S
5	,_B
6	sin_P
7	aditivo_S
8	!_B
9	el_T
10	aditivo_S
11	desaconsejable_A
…	…

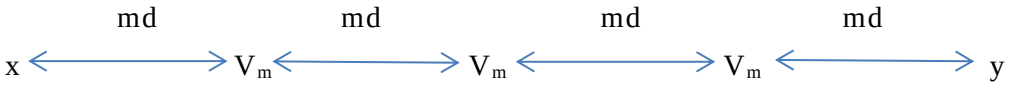
同じデータが全体の中でどのように集中・拡散しているかを示す係数を「線状拡散度」(Linear Dispersion: Lin.D.)と呼び，次のように定義します。

$$Lin.D. = 1 - (\Sigma (d(i) - md)^2 / n)^{1/2} / ((n - 1)^{1/2} * \Sigma (d(i) / n)$$

ここで，d(i)は同じデータが繰り返されるときの，それぞれの間隔(distance)です。md はその平均，n は個数を示します。たとえば上のデータの aditivo の 1 回目の位置と 2 回目の位置は 7 と 10 なので，その間隔は 10 - 7 = 3 になります。md は平均距離(mean distance)を示し，次のようにして計算します。たとえば 5 つの単語が次のように d(1), d(2), ..., d(5)の間隔で出現したとします。



一方，この単語がテキスト内で，完全に等間隔で並ぶと仮定したときの間隔が md です。



d(1)と md の差を計算します。同様に d(2)と md の差を計算します。最初の d(0)と最後の d(n)を加算し，それと md の差を計算します（V の全体が左右に移動しても逸脱度に影響しないためです。その相対標準偏差は平均分布からの全体の逸脱度(0 .0~ 1.0)を示します。「線状拡散度」(L.Disp)はその逆数になるので，1 からこの値を引きます。結果は次のように出力されます。

Lemma	Freq.	F.Rank	F.Permil	Lin.D.	L.D.Rank	Usage	U.Rank
l_C	1	1	.500	1.000	10	1.000	1
¡_B	1	1	.500	1.000	10	1.000	1
su_T	8	4	4.002	.647	7	5.176	4
comida_S	1	1	.500	1.000	10	1.000	1
,_B	250	10	125.063	.882	9	220.534	10
sin_P	2	2	1.001	.190	2	.380	1
aditivo_S	8	4	4.002	.313	4	2.505	2
!_B	1	1	.500	1.000	10	1.000	1
el_T	165	10	82.541	.817	9	134.849	10
desaconsejable_A	2	2	1.001	.389	4	.778	1

それぞれの単語の頻度数(Freq.)，頻度数ランク(F.Rank)，線状拡散度(Lin.D.)，線状拡散度ランク(L.D.Rank)，使用度(Usage)，使用度ランク(U.Rank)が示されています。使用度は頻度と線状拡散度を掛け合わせた値

です。それぞれの値(n)のランク(Rank: 1, 2, …, 10)は最大値(m)が 10 となるように最大値で割って 10 を掛けた結果です。RndUp は小数点以下の繰り上げをする関数です。これによってランクは 1 ~ 10 の範囲の整数になります。なお，頻度の低い語の拡散度はあまり普通でありません。頻度が 1 の語の拡散度は必ず 1 になります。

Rank = RndUp (Freq. / Max * 10)

4.6.2. 幾何標準偏差

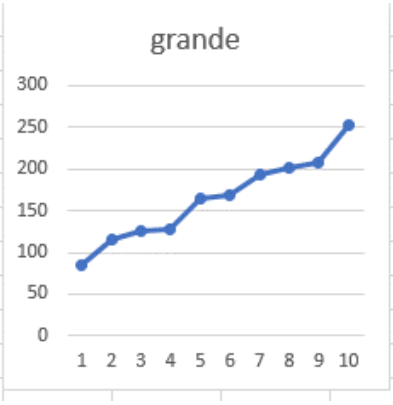
先に見た幾何平均と同様に幾何標準偏差(Geometric Standard Deviation: GSD)を次の Excel 関数式で求めます[33]。

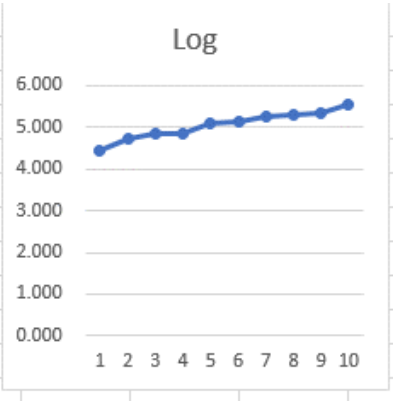
GSD = EXP(STDEVP(LN(範囲)))

この式によって，LN 関数で範囲にあるデータセットを対数変換し(ベクトル)，STDEVP 関数でその標準偏差を求め(スカラー)，EXP 関数でネイピア数(e)を底としたべき乗にします(スカラー)。つまり対数変換したデータセットの標準偏差を指数変換したことになります。

次の表と図はスペイン語の形容詞 grande「大きな」の 10 種類の文書での使用数とその自然対数(Log)を示します(それぞれ 10 万語のデータを昇順で並べ替えてあります)[34]。

*	grande	Log
1	84	4.431
2	115	4.745
3	126	4.836
4	128	4.852
5	164	5.100
6	169	5.130
7	194	5.268
8	202	5.308
9	207	5.333
10	252	5.529





先の幾何標準偏差(GSD)はデータの頻度(grande の列)から直接求めましたが，上の Log のデータ列を使って求めることもできます。

GSD = EXP(STDEVP(Log の範囲)) = 1.371

●プログラム (R)

```
gsd = function(A){exp(sdp(log(A)))} #幾何標準偏差
```

[33] Excel では幾何平均を直接計算する GEOMEAN 関数が用意されていますが，幾何標準偏差を直接計算する関数は用意されていません。
[34] Excel 関数：Log = LN(数値).

```
gsdz = function(A){exp(sdp(log(A+1)))} #幾何標準偏差
```

先の幾何平均のプログラムと同様にデータがゼロ(0)を含む場合にはデータ全体に 1 を足しています(gsdz)。幾何平均と異なり，データが 1 だけ大きくなっても幾何標準偏差は増えることはないので(ほとんど影響しないので)，最終的に-1 を付けません。

●幾何平均値*/幾何標準偏差

先に見たバラツキを含めた中心の範囲を示す「平均±標準偏差」と同様に，次の式で求めた幾何平均(GM)と幾何標準偏差(GSD)を使って「幾何平均(GM) */ 幾何標準偏差(GSD)」を考えます。

GM = EXP(AVERAGE(Log の範囲))
GSD = EXP(STDEVP(Log の範囲))

ここで注意が必要ですが，中心の範囲として GM±GSD ではなくて，GM */ GSD を使います。つまり，上側では GM*GSD のように掛け(*)，下側では GM/GSD のように割ります。

その理由は次のように考えるとわかります。はじめに中心の範囲の上側の境界は[35]

上側境界　= EXP[AVERAGE(Log 範囲) + STDEVP(Log 範囲)]
　　= EXP[AVERAGE(Log 範囲)] * EXP[STDEVP(Log 範囲)]
　　= GM * GSD

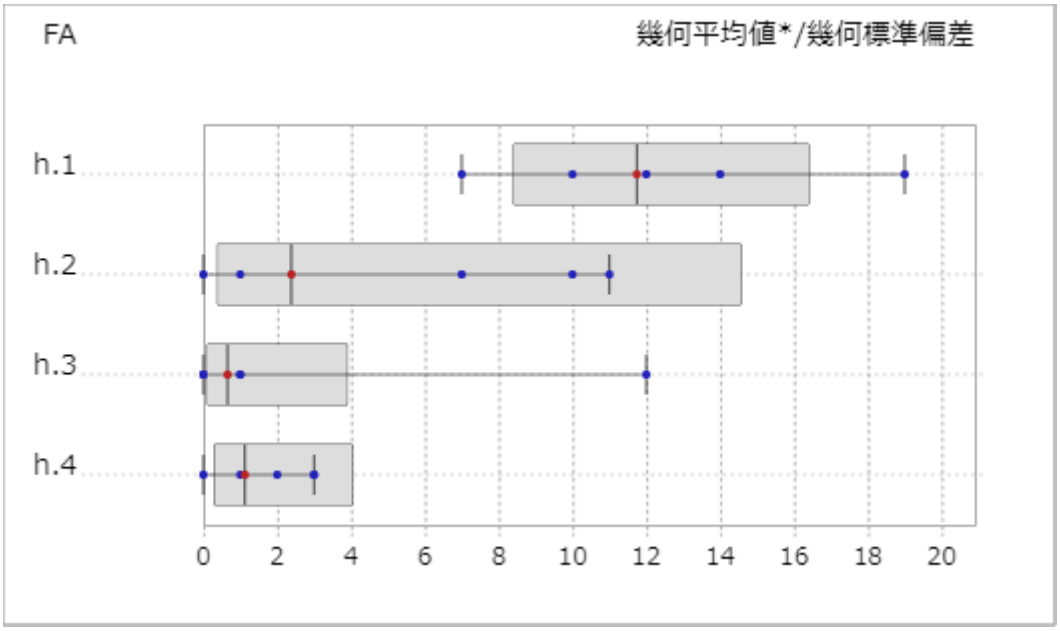
同様に下側の境界は[36],

下側境界　= EXP[AVERAGE(Log 範囲) - STDEVP(Log 範囲)]
　　= EXP[AVERAGE(Log 範囲)] / EXP[STDEVP(Log 範囲)]
　　= GM / GSD

次の表とグラフは幾何平均値(GM)*/幾何標準偏差(GSD)と，その様子を示すボックスチャートです。

FA	A	B	C	D	E	GM	GSD	GM/GSD	GM*GSD
h.1	10	19	14	7	12	11.744	1.396	8.413	16.395
h.2	11	7	10	0	1	2.384	6.106	0.39	14.556
h.3	0	0	1	12	1	0.654	5.942	0.11	3.889
h.4	0	1	2	3	3	1.125	3.578	0.314	4.025

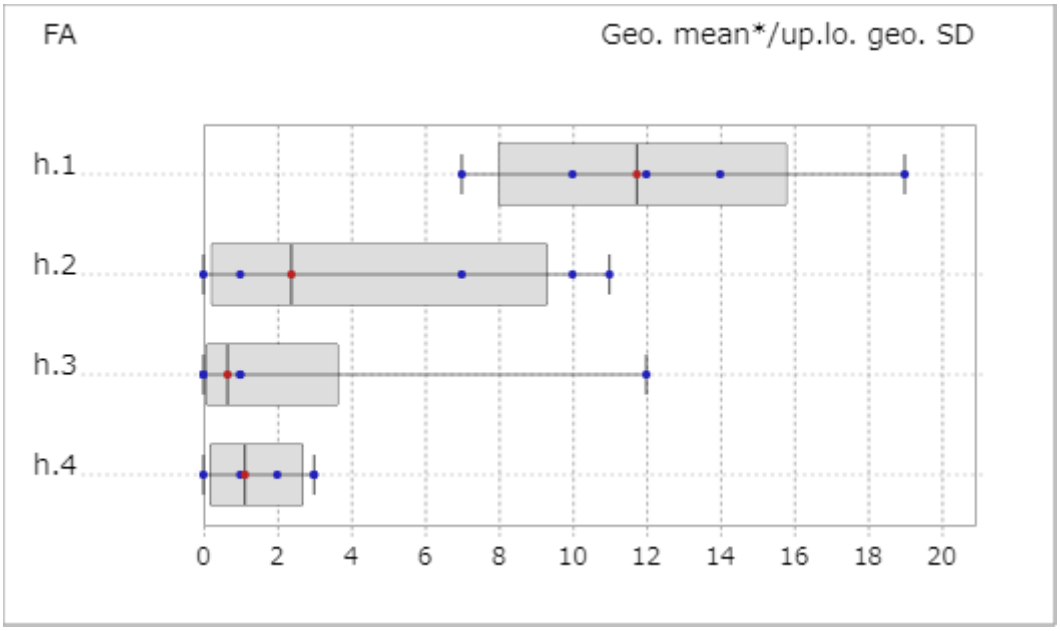
[35] $EXP(x) = e^x$, $EXP(y) = e^y$, $EXP(x + y) = e^{(x+y)} = e^x * e^y = EXP(x) * EXP(y)$.
[36] $EXP(x) = e^x$, $EXP(y) = e^y$, $EXP(x - y) = e^{(x-y)} = e^x / e^y = EXP(x) / EXP(y)$.



●幾何平均値＊／上側下側幾何標準偏差

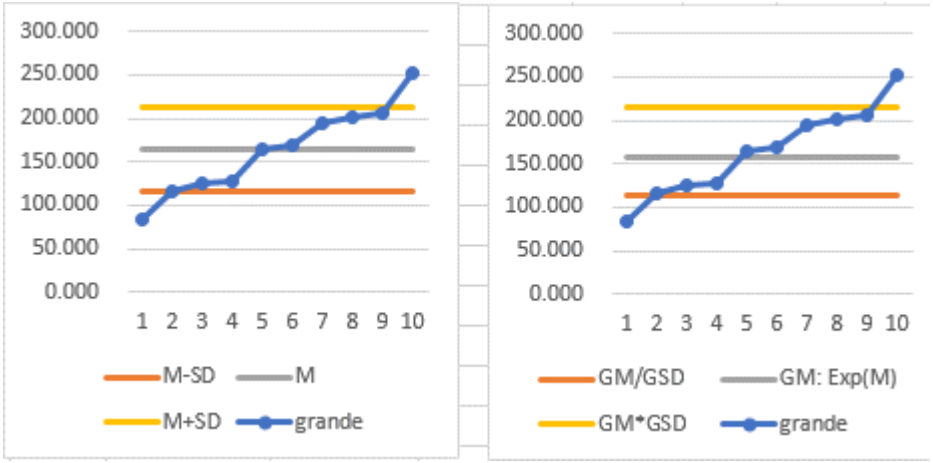
次の表とグラフはデータ例の幾何平均値(GM)＊上側幾何標準偏差(UGSD)／下側幾何標準偏差(LGSD)を示します。（→算術平均値±上側下側標準偏差）

FA	A	B	C	D	E	GM	Geo. SD	LGSD	UGSD	GM/LGSD	GM*UGSD
h.1	10	19	14	7	12	11.7	1.4	1.5	1.3	8.0	15.8
h.2	11	7	10	0	1	2.4	6.1	10.2	3.9	0.2	9.3
h.3	0	0	1	12	1	0.7	5.9	6.5	5.6	0.1	3.6
h.4	0	1	2	3	3	1.1	3.6	5.5	2.4	0.2	2.7



次はデータ{grande: 84, 115, 126, 128, 164, 169, 194, 202, 207, 252}の平均

(M:164.1)と標準偏差(SD:48.5)による M-SD, M, M+SD と，幾何平均(GM:156.5)と幾何標準偏差(SD:1.4)による GM*GSD, GM, GM/GSD を示します。どちらも中心(M, GM)の位置とバラツキ(M+SD~M-SD と GM*GSD~GM/GSD)の範囲が類似しています[37]。



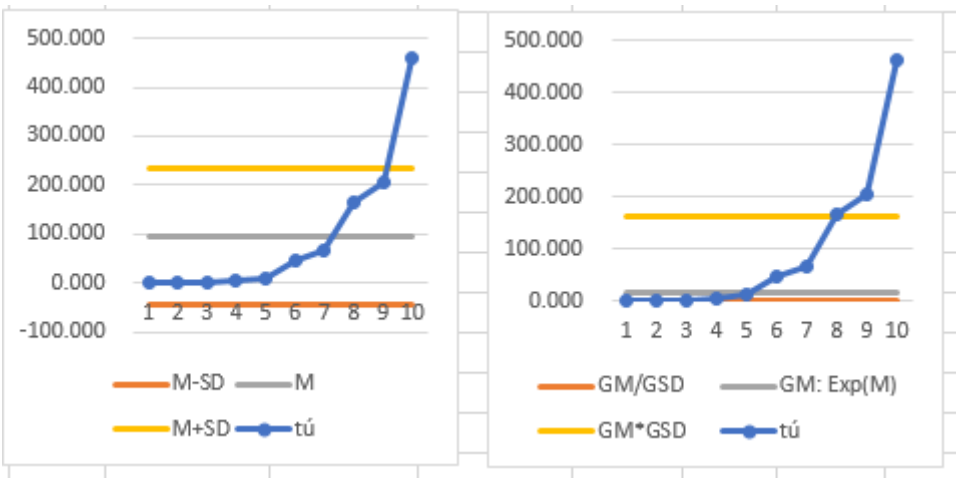
一方，昇順の急激な上昇を示すデータ{tú: 1, 1, 2, 4, 11, 47, 67, 165, 204, 461}の次のそれぞれのグラフは様子が大きく異なります(M: 96.3, SD: 139.8)。はじめに M-SD はゼロ以下になるので(96.3-139.8)，バラツキの範囲としては適していません。そして，M±SD が 10 番を除いてほとんどデータ全体をカバーしています。このようなデータは下右図のように幾何平均(GM:18.4)と幾何標準偏差(GSD:8.9)による GM*GSD, GM, GM/GSD を使ったほうが中心の範囲の見通しがよくなります。幾何平均(GM:18.4)がデータの中心にあり，バラツキ(幾何平均*/幾何標準偏差: 2.0~164.0)が中心の周りの 3 番～7 番をカバーしているからです。

[37] たとえば幾何平均(GM) = 50，幾何標準偏差(GSD) = 2 のとき GM*GSD, GM/GSD のそれぞれの対数正規分布累積確率は

GM*GSD: LOGNORM.DIST(50*2,LN(50),LN(2),1) = 0.841
GM/GSD: LOGNORM.DIST(50/2,LN(50),LN(2),1) = 0.159
0.841 - 0.159 = 0.683

この計算は GM (>0), GSD(>1)にどのような数値を入れても変わりません。そして，その計算結果は正規分布累積確率の計算結果と同じです。

M+SD: NORM.DIST(50+2, 50, 2, 1) = 0.841
M-SD: NORM.DIST(50-2, 50, 2, 1) = 0.159



●Q-Q プロット

これまで見てきたように，昇順の度数が徐々に加算されていくようなデータでは平均(M: 算術平均)と標準偏差(SD)を使って記述することができますが，一方，昇順の度数が徐々に積算されていくようなデータでは幾何平均(GM)と幾何標準偏差(GSD)を使って記述するほうがよいでしょう。より厳密に言えば，データが正規分布(normal distribution)に近似していれば加算型のデータであり，対数変換したデータの分布が正規分布に近似していれば(対数正規分布 log-normal distribution)，積算型のデータである，と考えられます。そこで，データの分布とそれぞれの正規分布との近似を見るために，次に説明するような Q-Q プロット(Q-Q plot)が使われます[38]。

Q-Q プロットを作成するために，はじめに次のようにしてデータ(x)の昇順の度数(たとえば{tú: 1, 1, 2, 4, 11, 47, 67, 165, 204, 461})，データ(x)の対数変換(Log)，等間隔の確率(p)[39]，そして正規分布関数(NORM.DIST)の逆関数(NORM.INV)を使って p に対応する度数(Norm.i)を用意します。

N	x: tú	Log	p	Norm.i	Dif.Norm	LogNorm.i	Dif.Log
1	1	0.000	0.050	-133.7	134.7	0.5	0.5
2	1	0.000	0.150	-48.6	49.6	1.9	0.9
3	2	0.693	0.250	2.0	0.0	4.2	2.2
4	4	1.386	0.350	42.4	38.4	7.9	3.9
5	11	2.398	0.450	78.7	67.7	13.9	2.9
6	47	3.850	0.550	113.9	66.9	24.2	22.8
7	67	4.205	0.650	150.2	83.2	42.7	24.3

[38] Quantile - Quantile plot. (Quantile: 分位数)

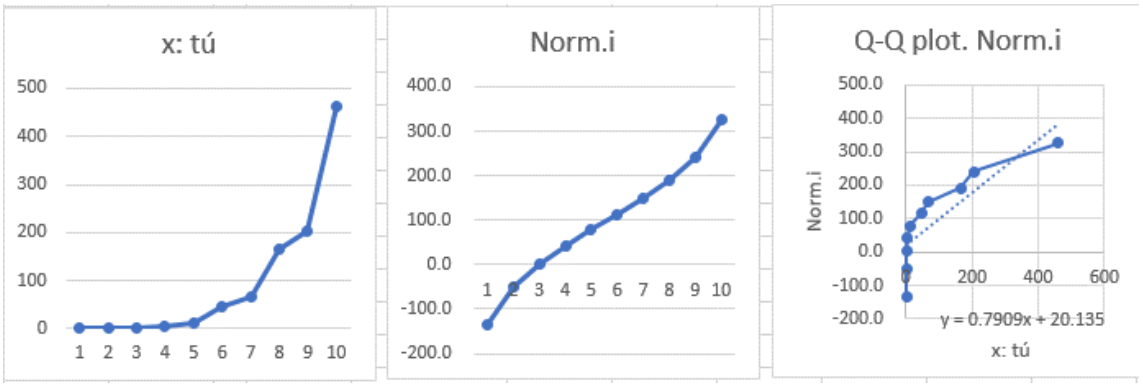
[39] 一般の Q-Q プロットでは順位数(Rank)が多く用いられていますが，ここでは等間隔の確率(p)を求めるために連続数(N)を使っています。両者(N, Rank)の違いは同じ数値があるとき順位は同じになり(同順位)，N は連続数であり続けます。ここで連続数(N)を採用する理由は，データの度数分布を正規分布や対数正規分布と比較するとき，全体を完全に等間隔にした確率を基準(物差し)にするほうがよいからです。

8	165	5.106	0.750	190.6	25.6	80.4	84.6
9	204	5.318	0.850	241.2	37.2	177.7	26.3
10	461	6.133	0.950	326.3	134.7	673.8	212.8

Log = LN(x)

p = (N - 0.5) / 個数(10)

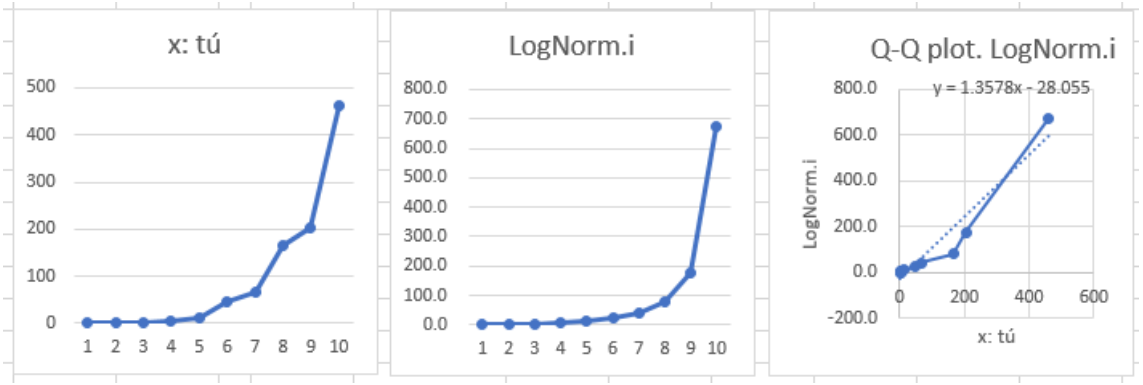
Norm.i = NORMINV(確率 p, x の平均値, x の標準偏差) …注[40]



上左図と上中図はそれぞれ x: tú と Norm.i の折れ線グラフです。両者の動きはかなり違います。そこで，今度はデータ(x)と正規分布逆関数 Norm.i の出力をそれぞれ，横軸と縦軸とした散布図である Q-Q プロットを作成します(上右図)。データ(x)と正規分布逆関数 Norm.i が一致して，傾きが 1(角度が 45 度)の直線になれば完全な正規分布を示します。しかし，この Q-Q プロットによれば両者はかなり乖離しています。上表の Dif.Norm は両者の差の絶対値を示します。

次に，データ(x)と対数正規分布逆関数(LogNorm.i)の Q-Q プロットを作成します。

LogNorm.i = LOGINV(確率 p, Log の平均値, Log の標準偏差) …注[41]



[40] 確率 p =NORMDIST(度数 x, 平均, 標準偏差, 1)
度数 x =NORMINV(確率 p, 平均, 標準偏差)

[41] 確率 p =NORMDIST(LN(度数 x),AVERAGE(LN(範囲)),STDEVP(LN(範囲)),1)
度数 x = LOGINV(確率 p, AVERAGE(LN(範囲)),STDEVP(LN(範囲)))

上中図によればデータ(x)と LogNorm.i がかなり近似しています。その Q-Q プロット(上右図)を見ると両者が傾きが 1 の直線に近い関係にあることがわかります。上表の Dif.Log は両者の差の絶対値を示します。

●正規分布と対数正規分布の判定

Q-Q プロットを使ってデータが正規分布と対数正規分布のどちらに近いかを判定しますが，その違いが目視では判断しにくいことがあります。そこで，上表の Dif.Norm と Dif.Log のそれぞれの総和を SN, SL とし，SN と SL の正規分布残差対照値(Normal Distribution Residual Contrast: NDRC)を次の式で求めます。

$$NDRC = (SN - SL) / (SN + SL)$$

SN と SL が同値であれば(SN = SL)，SN:SL.c = 0 になります。SN = 0 であれば SN:SL.c = -1 であり，正規分布と完全に一致していることを示します。逆に，SL = 0 であれば SN:SL.c = 1 であり，対数正規分布と完全に一致していることを示します。

4.6.3. 不偏分散・不偏標準偏差

無作為で収集されたデータ（標本）は「母集団から抽出されたもの」と考えられます。そこで標本の分散（標本分散:Vr）と母集団の分散（母分散:v）の関係は次の式で示されます。

$$E(Vr) = [(N - 1) / N] v$$

つまり標本分散 Vr の平均 E(Vr)は，母分散 v に(N - 1) / N を掛けた値になります(N:データ数)[42]。この式は次のように導かれます(E:平均関数；V:分散関数；Σ:データ(i)の和関数[i=1...N])[43]。

左辺 E(Vr)

$= E [\Sigma (X_i - M)^2 / N]$ ←分散 Vr の定義: $M = \Sigma X_i / N$

$= (1/N) E [\Sigma (X_i - M)^2]$ ←E(cX) = c E(X)

$= (1/N) E \{\Sigma [(X_i - m) - (M - m)]^2\}$ ←共に母平均 m を引く: m = E(M)

$= (1/N) E \{\Sigma [(X_i - m)^2 - 2 (X_i - m) (M - m) + (M - m)^2]\}$ ←$(X-Y)^2$ を展開

$= (1/N) E [\Sigma (X_i - m)^2 - \Sigma 2(X_i - m) (M - m) + \Sigma (M - m)^2]$ ←Σ を分配

$= (1/N) E [\Sigma (X_i - m)^2 - 2(M - m) \Sigma (X_i - m) + \Sigma (M - m)^2]$ ←2(M - m)を外へ

$= (1/N) E [\Sigma (X_i - m)^2 - 2(M - m)(\Sigma X_i - \Sigma m) + \Sigma (M - m)^2]$ ←Σ を分配

---

[42] 標本のデータ数 N が十分に大きければ(N - 1) / N は 1 に近似するので，標本分散の平均 E(Vr)を近似的に母分散と見なすことができます。

[43] 倉田・星野(2009: 201)を参照しました。

$= (1/N)\ E\ [\ \Sigma\ (X_i - m)^2 - 2(M - m)(\ \Sigma\ X_i - N\ m) +\ \ \Sigma\ (M - m)^2]$ ←$\Sigma\ c = N\ c$

$= (1/N)\ E\ [\ \Sigma\ (X_i - m)^2 - 2N\ (M - m)(\ \Sigma\ X_i/N - m) +\ \ \Sigma\ (M - m)^2]$ ←N を外へ

$= (1/N)\ E\ [\ \Sigma\ (X_i - m)^2 - 2N\ (M - m)(M - m) +\ \ \Sigma\ (M - m)^2]$ ←$\Sigma\ X_i/N = M$

$= (1/N)\ E\ [\ \Sigma\ (X_i - m)^2 - 2N\ (M - m)^2 +\ \ \Sigma\ (M - m)^2]$ ←整理

$= (1/N)\ E\ [\ \Sigma\ (X_i - m)^2 - 2N\ (M - m)^2 + N\ (M - m)^2]$ ←$\Sigma\ c = N\ c$

$= (1/N)\ E\ [\ \Sigma\ (X_i - m)^2 - N\ (M - m)^2]$ ←整理

$= (1/N)\ \{E[\ \Sigma\ (X_i - m)^2 - E[N(M - m)^2]\}$ ←$E(X-Y) = E(X) - E(Y)$

$= (1/N)\ \{\ \Sigma\ E[(X_i - m)^2] - E[N(M - m)^2]\}$ ←$E(X+Y) = E(X) + E(Y)$

$= (1/N)\ \{\ \Sigma\ v - E[N(M - m)^2]\}$ ←$E[(X_i - m)^2] = v$

$= (1/N)\ \{N\ v - E[N(M - m)^2]\}$ ←$\Sigma\ c = N\ c$

$= (1/N)\ \{N\ v - N[E(M - m)^2]\}$ ←$E(cX) = c\ E(X)$

$= (1/N)\ \{N\ v - N\ V(M)\}$ ←$E(M - m)^2 = V(M)$ 注(a)[44]

$= (1/N)\ \{N\ v - N\ v\ /\ N\}$ ←$V(M) = v\ /\ N$ 注(b)[45]

$= [(N - 1)\ /\ N]\ v$ ←整理

よって

$$E(Vr) = [(N - 1)\ /\ N]\ v$$

この式から

$N\ /\ (N - 1)\ E[Vr] = v$ ←両辺に N / (N-1)を掛ける

$E\{[N\ /\ (N - 1)]\ Vr\} = v$ ←$E(c\ X) = c\ E(X)$

$E\{[N\ /\ (N - 1)]\ 1\ /\ N\ \ \Sigma_i\ (X_i - M)\} = v$ ←$Vr = 1\ /\ N\ \ \Sigma_i\ (X_i - M)$

$E[1\ /\ (N - 1)\ \ \Sigma\ (X_i - M)] = v$ ←N の項を整理

ここで「不偏分散」(Unbiased Variance: U.Vr)は次のように定義されます[46]。

$$U.Vr = [1\ /\ (N - 1)]\ \ \Sigma\ (X_i - M)^2$$

よって，先の式を UVr を使って表すと[47]

$$E[1\ /\ (N - 1)\ \ \Sigma\ (X_i - M)] = v$$
$$E(UVr)\ \ = v$$

不偏分散(UV)の根は「不偏標準偏差」(Unbiased Standard Deviation: U.SD)

---

[44] (a): $E(M - m)^2 = E[M - E(M)]^2 = V(M)$

[45] (b): $V(M) = V(\ \Sigma\ Xi\ /\ N) = 1/N^2\ V[\ \Sigma\ (Xi)] = 1/N^2\ \Sigma\ V(Xi) = 1/N^2\ \Sigma\ v$
$= 1/N^2\ N\ v = v\ /\ N$

[46] この不偏分散の式と，次の標本分散(Vr)の式と比較すると，分母の(N - 1)が違うことがわかります。この違いは N が小さいときに顕著になります。

$$Vr = (1\ /\ N)\ \ \Sigma\ (X_i - M)^2$$

[47] よって，不偏分散の期待値が母分散になります。

と呼ばれます。

$$U.SD = UVr^{1/2} = [1 / (N - 1) \ \Sigma_i (X_i - M)]^{1/2}$$

不偏分散(U.Vr)と不偏標準偏差(U.SD)は標本データから母集団の分散と標準偏差をそれぞれ推定するときに使われます。

下右表の「確率」は不偏分散・不偏標準偏差の乱数累積確率を示します。

X	v1	v2	v3	v4	v5
d1	10	19	14	7	12
d2	11	7	10	0	1
d3	0	0	1	12	1
d4	0	1	2	3	3

横軸	不偏分散	不偏標準偏差
d1	20.300	4.506
d2	25.700	5.070
d3	26.700	5.167
d4	1.700	1.304

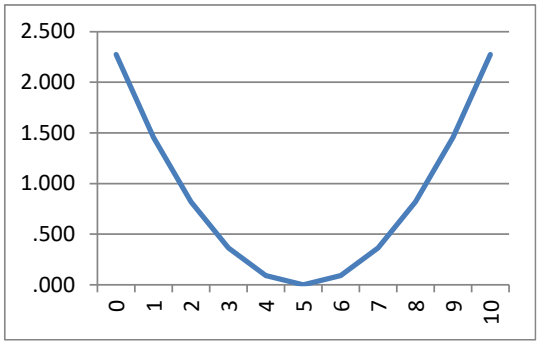
4.6.4. 平均偏差

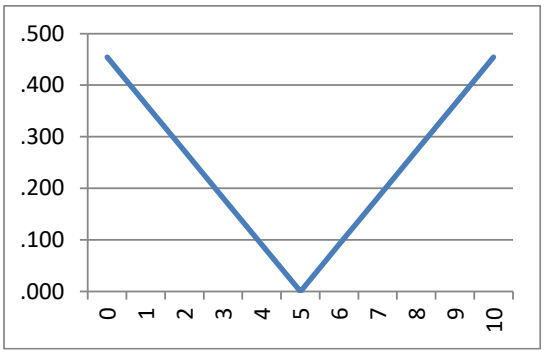
標準偏差の計算では，そのベースとなる偏差（データと平均値の差）が2乗されているために，それが大きくなると極端に標準偏差が増加します。下左図(X)は，データ(0, 1, 2, ..., 10)の標準偏差(StDev=3.16)の計算過程を示します。ここで，X=0 のときの偏差 2 乗/N が平均から離れるにつれて，その増加率が次第に大きくなっていくことがわかります（下左図）。

X	(X-M)^2	(X-M)^2/N
0	25	2.27
1	16	1.45
2	9	0.82
3	4	0.36
4	1	0.09
5	0	0.00
6	1	0.09
7	4	0.36
8	9	0.82
9	16	1.45
10	25	2.27
	StDev	3.16

Y	|X-M|	|X-M|/N
0	5	0.45
1	4	0.36
2	3	0.27
3	2	0.18
4	1	0.09
5	0	0.00
6	1	0.09
7	2	0.18
8	3	0.27
9	4	0.36
10	5	0.45
	MeanDev	2.73

一方，上右表と下右図が示すように偏差の絶対値の平均からの増加率は一定です。





データ全体の偏差を示す指標の 1 つとして，次の**平均偏差**(Mean deviation: MD)と呼ばれる式が使われます(M:平均; N:個数)[48]。

$$MD = [\ \Sigma\ |X(i) - M|\ ] / N$$

次に平均偏差(MD)の最大値(MD.max)を使って，**相対平均偏差**(Relative Mean Devistion: RMD)を設定します。平均偏差の最小値は(3, 3, 3, 3)のようにすべての成分が平均値と同じときに生じる 0 です。平均偏差の最大値(MD.max)はデータセットの和(S)が 1 つのデータに集まった(S, 0, 0, 0)のようなデータセットのときに起こる平均偏差なので

$$\begin{aligned} MD.max &= (|S - M| + |0 - M| + |0 - M| + \ldots) / N \quad \leftarrow M:\text{平均} \\ &= [S - M + (N-1)\ M] / N \quad \leftarrow \text{分子を整理} \\ &= (S - M + NM - M) / N \quad \leftarrow \text{分子を整理} \\ &= (NM - M + NM - M) / N \quad \leftarrow NM = S \\ &= 2M(N - 1) / N \quad \leftarrow \text{分子を整理} \\ &= 2M\ (1 - 1/N) \quad \leftarrow \text{分子を整理} \end{aligned}$$

よって，相対平均偏差(Relative mean deviation: RMD)は

$$\begin{aligned} RMD &= MD / MD.max \\ &= [\ \Sigma\ |X(i) - M| / N] / [2M\ (1 - 1/N)] \\ &= \Sigma\ |X(i) - M| / [2MN\ (1 - 1/N)] \end{aligned}$$

X	v1	v2	v3	v4	v5
d1	10	19	14	7	12
d2	11	7	10	0	1
d3	0	0	1	12	1
d4	0	1	2	3	3

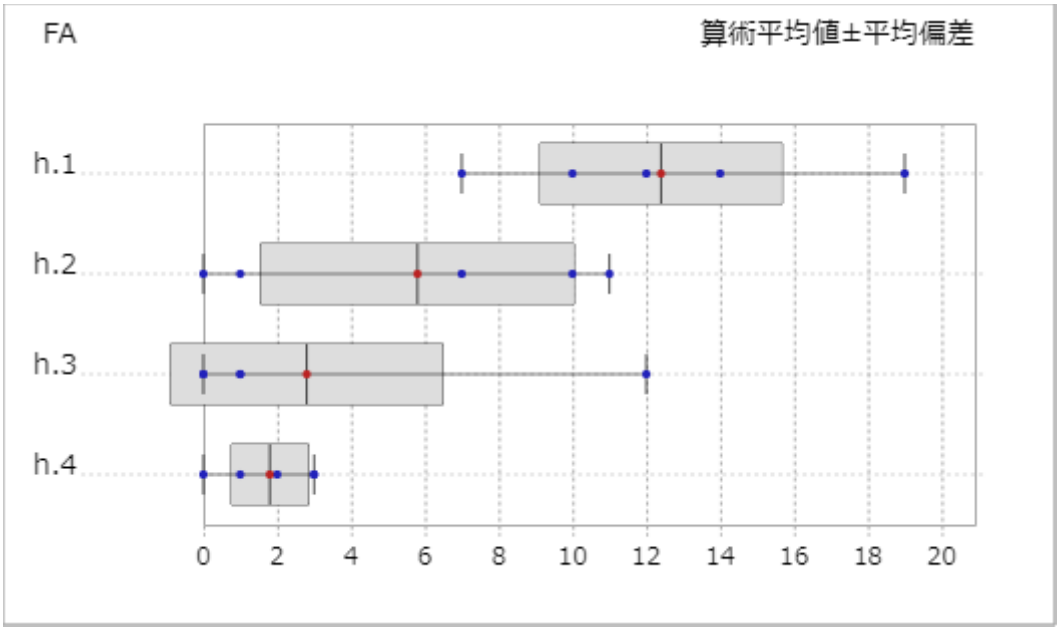
横軸	平均偏差	相対平均偏差
d1	3.280	0.165
d2	4.240	0.457
d3	3.680	0.821
d4	1.040	0.361

[48] 池田(1976: 54-55).

●算術平均値±平均偏差

次の表とグラフはデータ例の算術平均値(AM)±平均偏差(MD)を示します。

FA	A	B	C	D	E	AM	MD	AM - MD	AM + MD
h.1	10	19	14	7	12	12.4	3.28	9.12	15.68
h.2	11	7	10	0	1	5.8	4.24	1.56	10.04
h.3	0	0	1	12	1	2.8	3.68	-0.88	6.48
h.4	0	1	2	3	3	1.8	1.04	0.76	2.84



●上平均偏差と下平均偏差

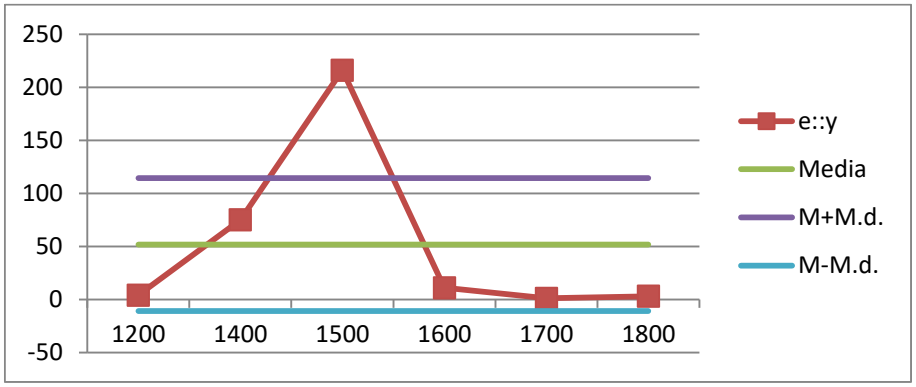
次はある言語データ(D)の年代ごとの頻度(F)と偏差|F-M|を示します。

D	1200	1400	1500	1600	1700	1800				
F	**4**	**75**	**216**	**11**	**1**	**3**	Sum	310.0	M.desv.	62.6
|F-M|	47.7	23.3	164.3	40.7	50.7	48.7	Mean	51.7	SD	85.3

このデータの和(Sum)・平均(Meab)・平均偏差(M.desv,)・標準偏差はそれぞれ 310, 51.7, 62.6, 83.3 です。このとき平均(Mean)を中心にして平均偏差(MD)を上下に伸ばして表を作り，それをグラフ化すると次のようになります。

D	1200	1400	1500	1600	1700	1800
F	4	75	216	11	1	3
Mean	51.7	51.7	51.7	51.7	51.7	51.7
Mean + MD	114.2	114.2	114.2	114.2	114.2	114.2

Mean - MD	-10.9	-10.9	-10.9	-10.9	-10.9	-10.9



このとき平均−平均偏差(Mean - MD)が負（マイナス: -10.9）になりますが，この意味は直感的にわかりにくいと思います。頻度は必ず正値でなければならないはずなのに，平均からの偏差がマイナスになることがあるのはなぜでしょうか？これは平均偏差は平均の上側と下側を区別せずに，平均との差の絶対値を足し上げてから，その平均を求めているために，上側の大きな値が下側にも反映されているからです[49]。

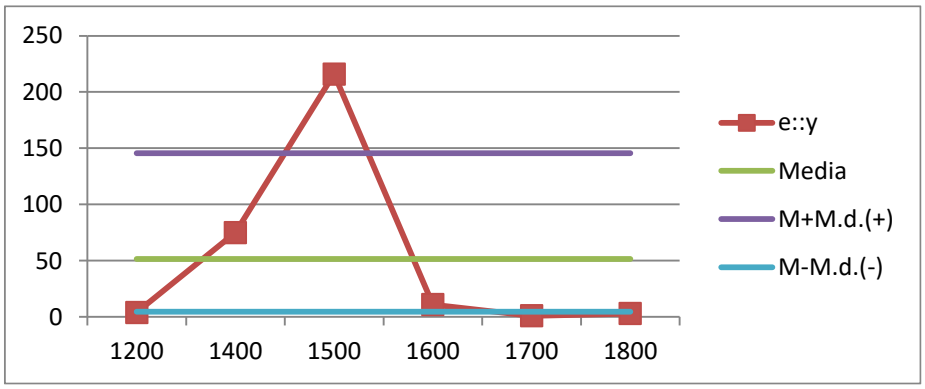
そこで平均の上側と下側を区別して「上平均偏差」(Upper mean deviation: UMD)と「下平均偏差」(Lower mean deviation: LMD)を求めます。上のデータで平均(51.7)の上側の頻度は 75 と 216 であり，下側の頻度は 4, 11, 1, 3 です。その上側平均偏差(UMD)と下側平均偏差 U(LMD)は

$$UMD = (|75 - 51.7| + |216 - 51.7|) / 2 = 93.8$$
$$LMD = (|4 - 51.7| + |11 - 51.7| + |1 - 51.7|+ |3 - 51.7|) / 4 = 46.9$$

次が上側平均偏差(UMD)と下側平均偏差(LMD)を使った頻度分布表とそのグラフです。

D	1200	1400	1500	1600	1700	1800
F	4	75	216	11	1	3
Mean	51.7	51.7	51.7	51.7	51.7	51.7
UMD	145.5	145.5	145.5	145.5	145.5	145.5
LMD	4.8	4.8	4.8	4.8	4.8	4.8

[49] この問題は標準偏差(SD)を使っても同じように生じます。このデータの標準偏差は 85.3 ですから，さらにマイナスの値が大きくなります。一般の統計処理では平均からの±標準偏差を問題にすることが多いのですが，そのときには注意が必要です。



このように，上側平均偏差が頻度の最大値を決して超えないことと同様に，下側平均偏差はマイナスになることはありません。

●比率偏差

平均偏差はデータ行列の数値そのものの平均からの偏差を扱います。しかしデータ行列が大きな全体の一部であるときは，それぞれの列の和を考慮しなければなりません。たとえば，次のデータ行列 Xnp はテキスト t1, t2, ..., t5 の中で語 w1, w2, ..., w4 が使われる頻度を示しているとします。そのとき，それぞれのテキストの総語数が Tp=1000, 1000, 2000, 2000, 2000 だとすると，データ行列の数値をそのまま使って平均偏差を計算することはできません。

Xnp	t1	t2	t3	t4	t5	Sn
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
Tp	1000	1000	2000	2000	2000	8000

Gries (2008)によれば，(1) はじめに，データ行列 X の測定値比率行列 (Observed proportion: Onp)を用意します(Sn: 横和)。

Onp = Xnp / Sn

たとえば O(w1, t1)は X(w1, t1) 10 / 62 = 0.161 になります。

Onp	t1	t2	t3	t4	t5
w1	0.161	0.306	0.226	0.113	0.194
w2	0.379	0.241	0.345	0.000	0.034
w3	0.000	0.000	0.071	0.857	0.071
w4	0.000	0.111	0.222	0.333	0.333

(2) 次にそれぞれのテキストの総語数の期待値比率ベクトル(Expected proportion: Ep)を作成します。

Ep = Tp / Tp.Sum

たとえば Ep(t1) は T(t1) 1000 / Tp.Sum 8000 = 0.125 になります。

Ep	0.125	0.125	0.250	0.250	0.250

(3) 最後に(1)で作成した測定値比率行列 Onp から(2)で作成した期待値比率ベクトル Ep の差(絶対値)の行列を作成します。

|Onp - Ep|	t1	t2	t3	t4	t5	Sum	DPn
w1	0.036	0.181	0.024	0.137	0.056	0.435	0.218
w2	0.254	0.116	0.095	0.250	0.216	0.931	0.466
w3	0.125	0.125	0.179	0.607	0.179	1.214	0.607
w4	0.125	0.014	0.028	0.083	0.083	0.333	0.167

たとえば，|Onp - Ep|(w1, t1) = |0.161 - 0.036| = 0.036 になります。

こうしてできあがった比率偏差行列の横和(Sum)が期待値からの偏差の全体の大きさを示しますが，この最大値は 2.000 になります。このことは次のような偏差が異常に大きな極端な例を見るとわかります。

Xnp	t1	t2	t3	t4	t5	Sn
w1	10	0	0	0	0	10
w2	20	0	0	0	0	20
w3	30	0	0	0	0	30
w4	40	0	0	0	0	40

Tp	200	200000	200000	200000	200000	800200

Onp	t1	t2	t3	t4	t5
w1	1.000	0.000	0.000	0.000	0.000
w2	1.000	0.000	0.000	0.000	0.000
w3	1.000	0.000	0.000	0.000	0.000
w4	1.000	0.000	0.000	0.000	0.000

Ep	0.000	0.250	0.250	0.250	0.250

|Onp - Ep|	t1	t2	t3	t4	t5	Sum	DPn
w1	1.000	0.250	0.250	0.250	0.250	2.000	1.000
w2	1.000	0.250	0.250	0.250	0.250	2.000	1.000

w3	1.000	0.250	0.250	0.250	0.250	2.000	1.000
w4	1.000	0.250	0.250	0.250	0.250	2.000	1.000

この例ではどの語{w1, w2, ..., w5}もテキスト t1 にだけ出現し, その他のテキスト{t2, ..., t5}には出現していません。そして, 縦和のベクトルは t1 が異常に小さく, t1 以外は異常に大きな数値を示しています。このような場合の偏差が無限に最大値に近似します。先と同じ計算をすると, 最後の行列で, テキスト t1 の成分だけが 1.000 になり, その他は 0.250 の成分が 4 個, つまり計 1.000 になります。よって全成分の和は 2.000 になります。そこで, Gries (2008)は「比率偏差」(Deviation of Proportions: DP)として次の式を提示しました。

$$DP(i) = [\ \Sigma |O(i) - E(i)|\ ] / 2$$

よって, DP の範囲は[0, 1)になります[50]。

参考：

Gries, S. Th. (2008). “Dispertions and Adjusted Frequencies in Corpora”, *Internatinal Journal of Corpus Linguistics*. 13, pp. 403-437.

Brezina, V. (2018). *Statistics in Corpus Linguistics. A Practical Guide*. Cambridge University Press, pp. 52-53.

■スペイン語の綴り字の高頻度・高偏差の変化

次の資料はスペイン Madrid 県の各地で発行された公証文書に見られる綴り字の変化のなかで, とくに高頻度・高偏差のものを選んだものです(1000 語正規化頻度)。

Diff	1200	1400	1500	1600	1700	1800	Mean	UMD	LMD	ULMD
& > y	41.7	105.0	224.0	0.0	0.0	0.0	61.8	102.7	51.4	.688
ϕ > ic	0.0	52.1	190.7	101.9	55.9	2.2	67.1	79.1	39.6	.623
ç > c	1.7	17.1	155.3	53.8	5.9	0.0	39.0	65.6	32.8	.633
b > v	0.0	0.7	78.7	91.9	60.0	14.4	41.0	35.9	35.9	.782
z > c	0.0	1.4	59.3	65.6	101.2	3.3	38.5	36.9	36.9	.729

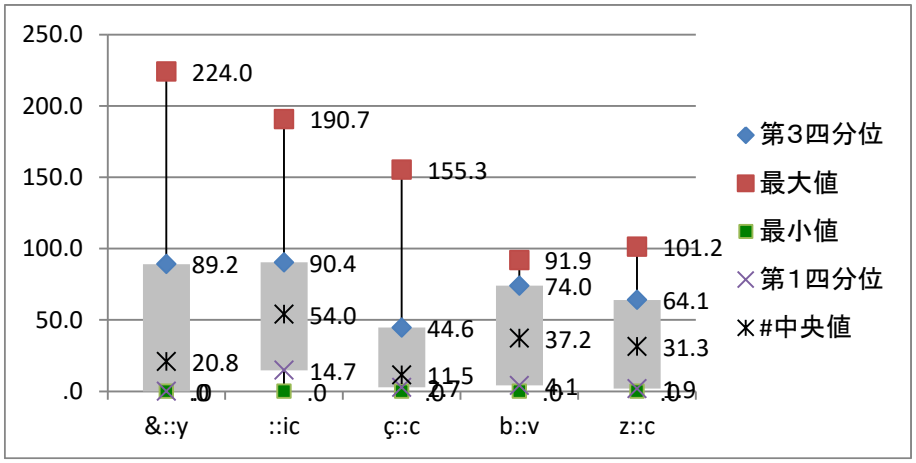
高頻度であることは, 和または平均値で簡単に求めることができますが, 偏差については上下平均偏差(ULMD)を使います。

次の図は中央値と四分位を用いたボックスチャートです[51]。どのデータ

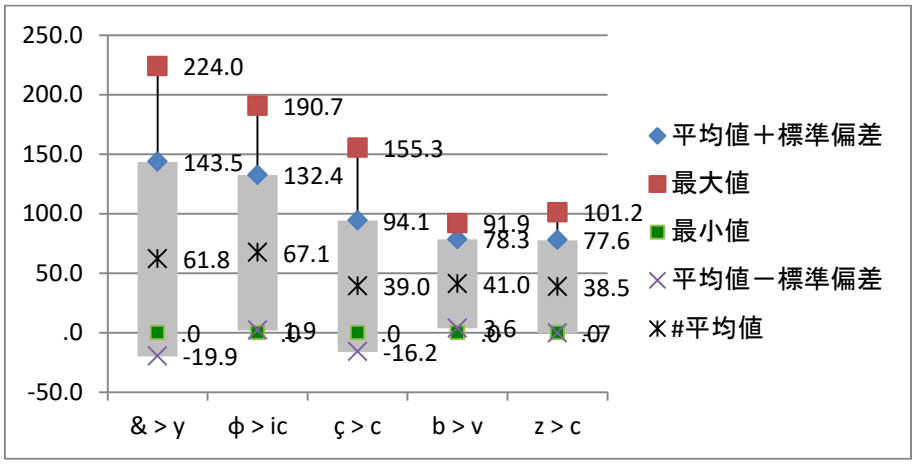
[50] DP の最大値は無限に 1 に近似します。

[51] この図は Excel の株価チャートを利用して作成しましたが, そのデータ形式の仕様によって, 凡例で最大値と最小値が第 3 四分位と第 1 四分位を囲むようにすることができませんでした。Excel の「データの選択」のダ

も中央値が小さいので下に寄っています。第 3 四分位と第 1 四分位で囲まれたボックスの中には全体の 50%が入っています。つまり，半分（だけ）の情報が含まれていることになります。

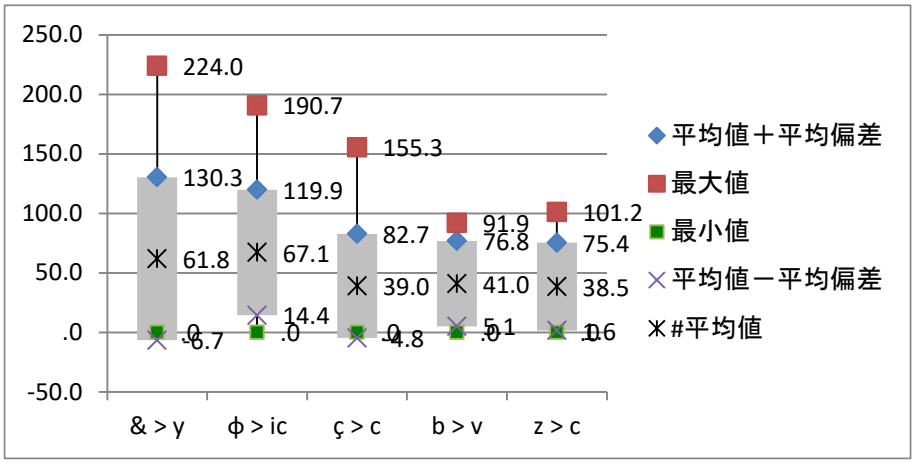


次の図は平均値と標準偏差を用いたボックスチャートです。

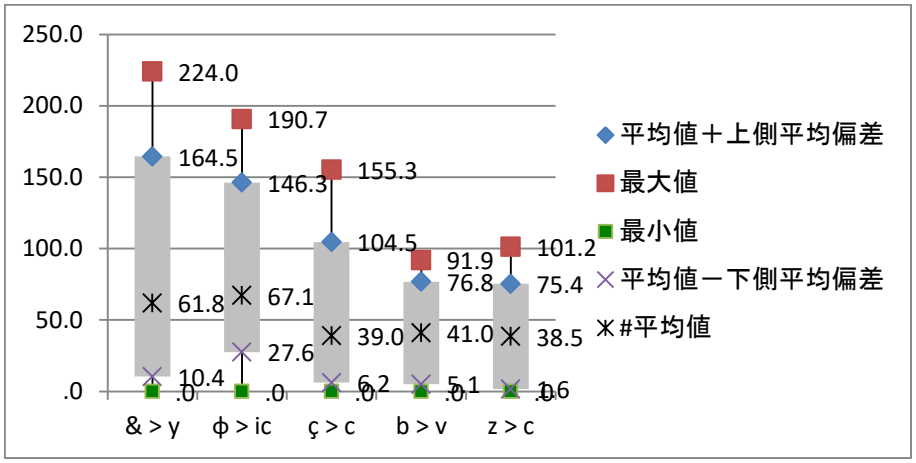


平均値はデータの中心を示し，標準偏差は平均値からの偏差の 2 乗の平均の根を示し，多くのデータが含まれます。先のチャートと比べるとボックスの幅が大きくなっていることがわかります。しかし，下側が最小値の 0 を超えているので，解釈が困難になります。このことは平均値と平均偏差を用いた次のボックスチャートでも同じです。

イアログボックスで凡例項目を移動させるとチャートのボックスの大きさが変化してしまいます。



最後に平均値と上下平均偏差を用いたチャートを示します。



この図は中央値と四分位によるチャートの過小情報(50%)の問題を回避し，また平均値と標準偏差・平均偏差によるチャートの最小値を超える異常性も回避しています。

言語変化の研究では，変化の質的な特徴が取り上げられ，それらを並列して記述されることが多いのですが，高頻度・高偏差の変化は言語の特徴を強く示しているので,これを数量化して区別することに意味があります。

プログラム(VBA)

```
Function MeanDev(Xn, sel) '上側(sel=1)・下側(sel=2)・上下(sel=3)平均偏差
    Dim i&, n&, md!, s1!, s2!, c1&, c2&
    n = nR(Xn): md = amA(Xn) '個数 : 平均
    For i = 1 To n
        If (sel = 1 Or sel = 3) And Xn(i, 1) > md Then s1 = s1 + Abs(Xn(i, 1) - md): c1 = c1 + 1 '上側平均偏差
        If (sel = 2 Or sel = 3) And Xn(i, 1) < md Then s2 = s2 + Abs(Xn(i, 1) - md): c2 = c2 + 1 '下側平均偏差
    Next
```

```
    If sel = 1 Then MeanDev = s1 / c1
    If sel = 2 Then MeanDev = s2 / c2
    If sel = 3 Then MeanDev = ((s1 / c1) + (s2 / c2)) / 2
End Function
```

上側平均偏差･下側平均偏差･上下平均偏差の計算の方法は簡単ですが，実際に計算するのは平均からの偏差の向き（正負）を考慮しないといけないので，多くの資料を扱うことは困難です。そこで上に簡単なプログラムを載せました。

●平均分離度・平均近接度

データの変動を示す係数の 1 つとして，全成分の平均からの分離の程度を計算した**平均分離度**(Separativity from mean: SM)と，その補数**平均近接度**(Proximity to mean: PM)を考えます。はじめに**分離度**(Separation: Sep)を次のように定義します。

$$Sep(X, Y) = |X - Y| / Max(X, Y)$$

上式の X, Y は比較する 2 つの値，|X - Y|は両者の差の絶対値，Max(X, Y)は X と Y の最大値(大きな方の値)です。たとえば，(2, 5)の分離度は|2 - 5| / max(2, 5) = 3/5 = .6 です。分離度の範囲は[0, 1]です[52]。

はじめにデータセットの成分と平均値の分離度の平均(SM.mean)を求めます(M:平均 ; N:個数)。

$$SM.mean = \Sigma Sep[X(i), M] / N$$

たとえば(5, 4, 3, 0)の SM.mean は

$$SM.mean(5, 4, 3, 0) = (|5-3|/5 + |4-3|/4 + |3-3|/3 + |0-3|/3) / 4$$
$$= (.4 + .25 + 0 + 1) / 4 = 1.65 / 4 = .41$$

この SM.mean は，{5, 5, 5, 5}のようにすべての成分が等しいとき，それぞれの成分は平均値(5)と等しいので，最小値は 0 になります。SM.mean の最大値(SM.mean.max)は(5, 0, 0, 0)のように 1 つの成分だけが正値の場合です。このとき，その正値は和(S)と同じになるので，(S, 0, 0, …)という分布の分離性係数(SM)の最大値(SM.mean.max)は

$$SM.mean.max = [|S - M| / max(S, M) + (N-1) |0 - M| / max(0, M)] / N$$
$$= [(S - M) / S + (N-1) M / M] / N \qquad \leftarrow S > M; M > 0$$

[52] X, Y を非負値(0 または正値)とします。分離度の最小値 0 は X=Y のときで，最大値 0 は X または Y が 0 のときです。X=Y=0 のときは，両者が分離していないので，その分離度を 0 とします。

$= [S/S - M/S + (N - 1)] / N$　　←分子を整理
$= [1 - M/(NM) + N - 1] / N$　　←S=NM
$= [N - 1/N] / N = 1 - 1 / N^2$　　←分子を整理：分母を整理

平均分離度(SM)は

$SM = SM.mean / SM.mean.max$
$= \{ \Sigma (i=1,N) Sep[X(i), M]\} / N / (1 - 1 / N^2)$

平均近接度(PM)は

$PM = 1 - SM$

X	v1	v2	v3	v4	v5
d1	10	19	14	7	12
d2	11	7	10	0	1
d3	0	0	1	12	1
d4	0	1	2	3	3

横軸	不等度	均等度	平均分離度	平均近接度
d1	.1625	.8375	.2339	.7661
d2	.3909	.6091	.6024	.3976
d3	.8253	.1747	.8442	.1558
d4	.3239	.6761	.4884	.5116

上のデータの不等度と平均分離度の大小関係はどちらも d1 < d4 < d2 < d4 で同じです。

次の D2 は D1 を 4 回繰り返したものです。

d1: 10, 19, 14, 7, 12
d5: 10, 19, 14, 7, 12, 10, 19, 14, 7, 12, 10, 19, 14, 7, 12, 10, 19, 14, 7, 12

それぞれのデータの不等度・均等度・平均分離度・平均近接度は次のようになります。

横軸	不等度	均等度	平均分離度	平均近接度
d1	.162	.838	.234	.766
d5	.075	.925	.225	.775

この実験が示すように，不等度・均等度はデータ数が多くなるときの下降率・上昇率が平均分離度・平均近接度より高くなりますが，逆に，平均分離度・平均近接度はデータ数にあまり大きく影響されません。その理由は，平均分離度・平均近接度の計算で，平均との偏差の平均がベースになっているためです。これはデータ成分全体の偏差の平均なのでデータ数に影響されません。そして，[0, 1]の範囲にするために最大値 $SM.mean.max = 1 - 1 / N^2$ で割りますが，これはほとんど 1 に近く，しかも N が大きくなるにしたがって非常に速く 1 に近似します。そこでデータ数が多いときには多様性係数でデータの変動を見る方がよいでしょう。

プログラム

```
function Coagulation(Ar) { //凝集性 配列Ar >> X: 0.xxx
        var N = Ar.length, M = Ar[N-1], S = 0; //個数, 最大値, 和
        for(var i=0; i<N; i++) {
                var int = (i==0)? Ar[0]: Ar[i] - Ar[i-1]; //間隔
                S += Math.abs(int - M/N); //間隔と上昇平均値の差の絶対値の和
        }//i
        return (N==1 || M==N)? 1: S / (2*(N-1)*(M-N)/N); // S/理論的最大値
} //function
```

関数 Coagulation に渡す Ar はたとえば[1, 2, 5, 6, 10]のような配列です。Coagulation(1, 2, 3, 6, 10)は 0.75 を返します。ここで配列 Ar の最初の要素を 1 とし，最後の要素をテキストの総語数とします。このように配列の最小値と最大値を同じ値にすることによって，それぞれの語の凝集性と均等性が比較可能になります。

4.6.5. 歪度

平均値を中心にして，データの左右の偏（かたよ）りを計る指標として，「歪度」（わいど）(Skewness: Sk)が使われます（芝・渡部・石塚 1984: 282）。歪度を算出するためには，初めにデータを「標準得点」(Standard score)に変換しなければなりません（後述→「標準得点」）。標準得点はそれぞれの数値から平均(M)を引き標準偏差(Sd.p)で割った値です。歪度は標準得点の 3 乗和をデータ数(N)で割った値です[53]。

$$Sk = \Sigma_i [(X_i - M) / Sd.p]^3 / N$$

データの標準得点は平均よりも大きければプラスになり，小さければマイナスになるので，その 3 乗も，たとえば$(-2)^3 = -8$.のように，プラスとマイナスの符号は変わりません[54]。

次は横軸の平均値と歪度を示す表です。たとえば，d3 の平均値は 2.8 なので，分布は v4:12 という特異な値によって，かなり右に傾いています。そこで，歪度は正の 1.465 になっています。一方，d4 の平均値は 1.8 であり，それを超える 2 と 3 との偏差はそれぞれ 0.2, 1.2，一方，平均値以下の 0, 1 との偏差は，それぞれ 1.8, 0.8 なので，0 に向かう左方向にデータが広

---

[53] 芝・渡部・石塚(1984: 282) (Excel 関数:SKEW.P). 歪度については他の定義もあります(Excel 関数:SKEW，不偏標準偏差 Sd を使用)。

$$Excel.SKEW = \Sigma_i [(X_i - M) / Sd]^3 * N / (N-1) / (N-2)$$

[54] 標準得点のままでもプラス・マイナスの符号は変わりませんが，その分子が偏差（データ値 – 平均）なので，どのようなデータでも和がゼロになってしまうからです。

がっていることがわかります。そこで歪度はマイナス値になっています。

X	v1	v2	v3	v4	v5
d1	10	19	14	7	12
d2	11	7	10	0	1
d3	0	0	1	12	1
d4	0	1	2	3	3

横軸	平均値	歪度
d1	12.400	.367
d2	5.800	-.192
d3	2.800	1.465
d4	1.800	-.363

●相対歪度

歪度(Sk)は[-1, 1]の範囲に相対化されていないので同一の基準による評価が困難です。これを相対化するために，その最大値(Sk.max)を求め，「相対歪度」(Relative Skewness: R.Sk)を次のように定義します。

$$R.Sk = Sk / Sk.max$$

次に，歪度の最大値(Sk.max)を，先の標準偏差(Sd)の最大値(Sd.max)と同様にして求めます。K は{K, 0, 0, 0, 0}のような唯一分布の唯一値です。このとき Sk は最大の Sk.max になります。

$$Sk.max = \{([(K - M)^3 + (N - 1)(0 - M)^3] / SD^3\} / N$$

$$= [(N\,M - M)^3 + (-M)^3 (N - 1)] / (N\,SD^3) \quad \leftarrow K = N\,M$$
$$= [(M (N - 1))^3 + (-M)^3 (N - 1)] / (N\,SD^3) \quad \leftarrow M \text{ を外へ}$$
$$= [(M (N - 1))^3 + (-1)^3 M^3 (N - 1)] / (N\,SD^3) \quad \leftarrow (-M)^3 = (-1)^3 M^3$$
$$= [(M (N - 1))^3 - M^3 (N - 1)] / (N\,SD^3) \quad \leftarrow (-1)^3 = -1$$
$$= [M^3 (N - 1)^3 - M^3 (N - 1)] / (N\,SD^3) \quad \leftarrow \text{共通の } M^3$$
$$= M^3 (N - 1)\,[(N - 1)^2 - 1] / (N\,SD^3) \quad \leftarrow M^3 (N - 1) \text{が共通}$$
$$= M^3 (N - 1)\,(N^2 - 2N + 1 - 1) / (N\,SD^3) \quad \leftarrow (N - 1)^2 \text{を展開}$$
$$= M^3 (N - 1)\,(N^2 - 2N) / (N\,SD^3) \quad \leftarrow 1 - 1 = 0$$
$$= M^3 (N - 1)\,N\,(N - 2) / (N\,SD^3) \quad \leftarrow N \text{ を外へ}$$
$$= M^3 (N - 1)\,(N - 2) / SD^3 \quad \leftarrow N / N = 1$$

先に見たように（→「相対標準偏差」），{K, 0, 0, 0, 0}のような唯一分布のときの SD は

$$Sd.max = M (N - 1)^{1/2}$$

よって

$$Sk.max = M^3 (N - 1)\,(N - 2) / (M (N - 1)^{1/2})^3 \quad \leftarrow Sd.max \text{ を代入}$$
$$= M^3 (N - 1)\,(N - 2) / M^3 (N - 1)^{3/2}$$
$$\leftarrow \text{分母の乗数 3 を配分}$$
$$= (N - 2) / (N - 1)^{1/2} \quad \leftarrow \text{分子と分母の } M^3 (N - 1) \text{が共通}$$

よって，相対歪度(R.Sk)は

$$R.Sk = Sk / Sk.max = Sk * (N - 1)^{1/2} / (N - 2) \quad \leftarrow \text{分母を整理}$$

次は歪度と相対歪度を比較した表です。歪度は範囲が[-1 ~ 1]になりませんが，相対歪度によってデータの偏り方を[-1 ~ 1]の範囲で評価することができるようになりました。

X	v1	v2	v3	v4	v5
d1	10	19	14	7	12
d2	11	7	10	0	1
d3	0	0	1	12	1
d4	0	1	2	3	3

横軸	歪度	相対歪度
d1	.367	.245
d2	-.192	-.128
d3	1.465	.977
d4	-.363	-.242

4.6.6. 尖度

データの分布が平均値に集中してとがった度合いを示す指標として，次のように定義される「尖度」（せんど：Kurtosis: Ku）が使われます[55]。

$$Ku = \Sigma_i [(X_i - M) / Sd]^4 / N$$

ここで，M はデータの平均，Sd はその標準偏差，N は個数を示します。尖度の式を見ると，標準化されたデータ$(X_i - M) / SD$ の 4 乗の平均であることがわかります。よって，データで平均からの標準化された逸脱が 1 以下ならば尖度はさらに小さく，それが 1 以上ならば尖度はさらに大きくなりますから，尖度によって逸脱の程度が強調されます。たとえば N=5 のとき，$(X_i - M) / Sd$ が 1 以上ならば $\Sigma_i [(X_i - M) / Sd]^4$ が 5 以上になるので，$\Sigma_i [(X_i - M) / Sd]^4 / N$ は 1 以上になります。そのとき分子は 4 乗されているので，逸脱の程度が強調されます。

●相対尖度

相対尖度(Relative Kurtosis: R.Ku)を次のように定義します。

$$R.Ku = Ku / Ku.max$$

次に，尖度の最大値(Ku.max)は，{K, 0, 0, 0, 0}のような唯一分布の値です。このとき Ku は最大になります。

$$Ku.max = \{([(K - M)^4 + (N - 1)(0 - M)^4] / Sd^4\} / N$$

$$= [(N M - M)^4 + M^4 (N - 1)] / (N Sd^4) \quad \leftarrow K = N M$$

[55] ほかの定義もありますが，ここでは芝他(1984: 145)に従います。

$= [(M\,(N - 1))^4 + M^4\,(N - 1)] / (N\,Sd^4)$ ←M を外へ
$= [M^4\,(N - 1)^4 + M^4\,(N - 1)] / (N\,Sd^4)$ ←共通の M^4
$= M^4\,(N - 1)\,[(N - 1)^3 + 1] / (N\,Sd^4)$ ←$M^4\,(N - 1)$が共通
$= M^4\,(N - 1)\,(N^3 - 3N^2 + 3N - 1 + 1) / (N\,Sd^4)$ ←$(N - 1)^3$ を展開
$= M^4\,(N - 1)\,(N^3 - 3N^2 + 3N) / (N\,Sd^4)$ ←1 - 1 = 0
$= M^4\,(N - 1)\,N\,(N^2 - 3N + 3) / (N\,Sd^4)$ ←N を外へ
$= M^4\,(N - 1)\,(N^2 - 3N + 3) / Sd^4$ ←N が共通

先に見たように（→「相対標準偏差」），{K, 0, 0, 0, 0}のような唯一分布のときの Sd は

$$Sd.max = M\,(N - 1)^{1/2}$$

よって

Ku.max $= M^4\,(N - 1)\,(N^2 - 3N + 3) / (M\,(N - 1)^{1/2})^4$ ←上式
$= M^4\,(N - 1)\,(N^2 - 3N + 3) / [M^4\,(N - 1)^2]$ ←$M^4\,(N - 1)$が共通
$= (N^2 - 3N + 3) / (N - 1)$ ←分母と分子の共通部分を除去

よって，相対尖度(R.Ku)は

$$R.Ku = Ku / Ku.max = Ku * (N - 1) / (N^2 - 3N + 3)$$ ←分母を整理

下右表の「確率」は尖度と限定尖度の乱数累積確率を示します。

X	v1	v2	v3	v4	v5
d1	10	19	14	7	12
d2	11	7	10	0	1
d3	0	0	1	12	1
d4	0	1	2	3	3

横軸	尖度	相対尖度
d1	2.114	.650
d2	1.281	.394
d3	3.203	.986
d4	1.628	.501

●分散・歪度・尖度

分散(Vr), 歪度(Sk), 尖度(Kr)はそれぞれ分布の「ばらつき」,「ゆがみ」,「とがり」を示します。どの式にも$(X_i - M)]^E/N$ (E=2, 3, 4)が含まれています[56]。

$$Vr = \Sigma_i\,[(X_i - M)]^2 / N$$
$$Sk = \Sigma_i\,[(X_i - M) / Sd]^3 / N$$
$$Ku = \Sigma_i\,[(X_i - M) / Sd]^4 / N$$

上の M はデータの平均，N はデータ数，Sd はデータの標準偏差です。これらデータの分布の様子を示す 3 つの指標について，実際の計算過程を

[56] 芝・南風原(1990:34-35)を参照。

比べてみましょう。次の X はデータ，D は平均からの偏差(X－M)を示します。

X	v1	v2	v3	v4	v5	M	Vr
d1	10	19	14	7	12	12.40	16.240
d2	11	7	10	0	1	5.80	20.560
d3	0	0	1	12	1	2.80	21.360
d4	0	1	2	3	3	1.80	1.360

D	v1	v2	v3	v4	v5	M
d1	-2.40	6.60	1.60	-5.40	-.40	.00
d2	5.20	1.20	4.20	-5.80	-4.80	.00
d3	-2.80	-2.80	-1.80	9.20	-1.80	.00
d4	-1.80	-.80	.20	1.20	1.20	.00

次の S は標準得点((X－M)/Sd)を示し，S^2 は標準得点の 2 乗です。

S	v1	v2	v3	v4	v5	M
d1	-.60	1.64	.40	-1.34	-.10	.00
d2	1.15	.26	.93	-1.28	-1.06	.00
d3	-.61	-.61	-.39	1.99	-.39	.00
d4	-1.54	-.69	.17	1.03	1.03	.00

S^2	v1	v2	v3	v4	v5	M
d1	.35	2.68	.16	1.80	.01	1.00
d2	1.32	.07	.86	1.64	1.12	1.00
d3	.37	.37	.15	3.96	.15	1.00
d4	2.38	.47	.03	1.06	1.06	1.00

上の表を見ると，標準得点(S)の平均(M)はすべてゼロ(0)になっています。この理由は標準得点の分子が偏差なので，先の表(D)で見たように，和・平均が 0 になるためです。また，標準得点の 2 乗(S^2)の平均はすべて 1 になります。この理由は，S^2 の平均の分子も分母も分散(Vr)になるためです。

$$\Sigma i\,[(Xi - M) / Sd]^2 / N = \Sigma i\,(Xi - M)^2 / N\ SD^2 = Vr / Vr = 1$$

これらのことから，標準得点(S)やその 2 乗の平均(S^2)が「ばらつき」「ゆがみ」「とがり」の指標として役に立たないことがわかります。それでは標準得点の 3 乗(S^3)や標準得点の 4 乗(S^4)の平均はどうでしょうか。

S^3	v1	v2	v3	v4	v5	M:Sk
d1	-.21	4.39	.06	-2.41	.00	.37
d2	1.51	.02	.79	-2.09	-1.19	-.19
d3	-.22	-.22	-.06	7.89	-.06	1.47
d4	-3.68	-.32	.01	1.09	1.09	-.36

S^4	v1	v2	v3	v4	v5	M:Ku
d1	.13	7.19	.02	3.22	.00	2.11
d2	1.73	.00	.74	2.68	1.26	1.28
d3	.13	.13	.02	15.70	.02	3.20
d4	5.68	.22	.00	1.12	1.12	1.63

それぞれ異なる数値を示します。S^3 の平均が歪度(Sk)，S^4 の平均が尖度(Ku)です。分布の「ばらつき」を示す分散(Vr)と，「とがり」を示す尖度(Ku)を比べると数値も大小関係も異なることがわかります(Ku: d2 < d4 < d1 < d3; Vr: d4 < d1 < d2 < d3)。

D	v1	v2	v3	v4	v5
d1	-2.40	6.60	1.60	-5.40	-.40
d2	5.20	1.20	4.20	-5.80	-4.80
d3	-2.80	-2.80	-1.80	9.20	-1.80
d4	-1.80	-.80	.20	1.20	1.20

D^2	v1	v2	v3	v4	v5	M:Vr
d1	5.76	43.56	2.56	29.16	0.16	16.24
d2	27.04	1.44	17.64	33.64	23.04	20.56
d3	7.84	7.84	3.24	84.64	3.24	21.36
d4	3.24	0.64	0.04	1.44	1.44	1.36

4.6.7. 中心性係数

平均値はデータの完全な中心(平均値=中央値の状態)にあるのではなく，中央値より小さかったり大きかったりします(中央値は必ずデータの中心にあります)。そこで，平均値よりも大きいデータの個数を Pc (positive count)とし，平均値よりも小さいデータの個数を Nc (negative count)として[57]，次の式で示す中心性係数(Centrality: C)を定義します。

C = (Pc - Nc) / (Pc + Nc)

Pc > Nc ならば C > 0 となり，Pc < Nc ならば C < 0 となり，Pc = Nc ならば C = 0 になります。C の範囲は(-1, 1)です。C は-1, 1 になることはないので，その範囲は[-1, 1]ではありません。

たとえば，d1 行{10, 19, 14, 7, 12}の平均値は 12.4 なので，平均値よりも大きな値{19, 14}の個数は 2 個(Ps)，平均値よりも小さな値{10, 7, 12}の個数は 3 個(Ns)となり，このデータの中心性係数は (2-3) / (2+3) = -0.2 になります。よって，データの個数は平均値より小さい方にやや多い，ということになります。

プログラム (R)

```
central = function(A){A=A-mean(A); p=length(A[A>0]);
n=length(A[A<0]); (p-n)/(p+n)}   #中心性係数(centrality)
```

*	A	B	C	D	E	central
h1	10	19	14	7	12	-0.2
h2	11	7	10	0	1	0.2
h3	0	0	1	12	1	-0.6
h4	0	1	2	3	3	0.2

4.6.8. 平衡係数

中央値からの偏差がプラスになる総量(positive: p)とマイナスになる総

[57] 平均値と同じデータはカウントしません。

量の絶対値(negative: n)を計算し，p と n の対照値[58]を「平衡係数」(Equilibrium: Equi)とします。

Equi = (p - n) / (p + n)

たとえば，d1 行{10, 19, 14, 7, 12}の中央値は 12 なので，p, n, Equi は次のように計算されます。

p = |19 – 12| + |14 — 12| = 7 + 2 = 9
n = |10 – 12| + |7 – 12| = 2 + 5 = 7
Equi = (9 - 7) / (9 + 7) = .125

*	A	B	C	D	E	中央値	中央値正値	中央値負値	平衡指数	歪度
h1	10	19	14	7	12	12	9	7	0.125	0.367
h2	11	7	10	0	1	7	7	13	-0.3	-0.192
h3	0	0	1	12	1	1	11	2	0.692	1.465
h4	0	1	2	3	3	2	2	3	-0.2	-0.363

歪度は平均値からの全体的な乖離の方向を示しますが，平衡係数は中央値からの正と負のデータの平衡性（バランス）を示しています。Equi=0 のときに平衡性は完全になり，中央値より小さい値が多いとき Equi はマイナスになり，中央値より大きい値が多いとき Equi はプラスになります。

プログラム(R)

```
Equi=function(V=D){
  V=unlist(V); V=V-Median(V); p=Sum(V[V>0]); n=-Sum(V[V<0]); (p-n)/(p+n)}
#Equilibrium respecting median
```

*	A	B	C	D	E	balance
h1	10	19	14	7	12	0.125
h2	11	7	10	0	1	-0.300
h3	0	0	1	12	1	0.692
h4	0	1	2	3	3	-0.200

4.6.9. エントロピー

生起確率(p)が大きい事象は「普通のこと」を示すので情報量が少なく，逆に生起確率が小さい事象は「異常なこと」を示すので情報量が多い，と考えられます。この情報量(Information: I)は確率 p を使って次のように数

[58] (X-Y)/(X+Y)の値を X と Y の「対照値」contrast と呼びます。

値化されます。このとき，対数の底(base)は 2 を使います[59]。

$$I = - \log(2)\ p$$

ここで log の前にマイナス(-)がつくのは，確率 p が 1 より小さいため，その対数(log)が必ずマイナスになるためです。そこで log(2) p にマイナス(-)をつけて正の値とします。対数の底を 2 とするのは，情報量を 0, 1 のような 2 値からの選択の数を数える，という約束に従っているためです。

たとえば確率 1/4, 1/8 で起こる事象の情報量は，それぞれ次のように計算されます。

$$I(p:1/4) = - \log(2)\ 1/4 = - \log(2)\ 4^{(-1)} = \log(2)\ 4 = 2$$
$$I(p:1/8) = - \log(2)\ 1/8 = - \log(2)\ 8^{(-1)} = \log(2)\ 8 = 3$$

よって, 確率 1/8 の事象の情報量は確率 1/4 の事象の情報量の 1.5 倍あることがわかります(3 / 2 = 1.5)。これは 8 個の中から 2 分検索(binary search)をして 1 個を選び出すためには 3 回の検索が必要なのに対して(8 > 4 > 2 > 1), 4 個の中から 2 分検索をして 1 個を選び出すためには 3 回の検索で探し出せることを示しています(4 > 2 > 1)。これは 8 (4)枚のトランプを半分ずつ切りながら目的の 1 枚のトランプを探し出すようなことです。

ここで, 情報量(I)を頻度(f)と和(s)から求めた確率(p)で計算すると(p = f / s)，和(s)が同じであれば頻度(f)が大きいほど情報量(I)は小さくなる，ということに注意しましょう。これは，しばしば起こることは「普通のこと」ですから予想しやすいので情報量が少なく，逆に頻度が少ないことは「異常なこと」ですから予想しにくいので情報量が多い，ということを意味します。

複数の成分があるデータ全体の情報量として，それぞれの成分の情報量(I)の期待値が設定され, それは「エントロピー」(Entropy: E)と呼ばれます。たとえば，データ{2, 8}の和が 10 なので，それぞれの確率を 2/10, 8/10 とし，このデータのエントロピー(E)を次のように計算します[60]。

$$E(2, 8) = 2/10 * -\log(2)\ (2/10) + 8/10 * -\log(2)\ (8/10) = .722$$

つまり，それぞれの情報量に全体の中で占める割合を掛け，それを足して情報量の期待値とします。

次の表と図は x + y = 10 とした x, y の値をもつデータ(2 成分)のエントロピーを示します。

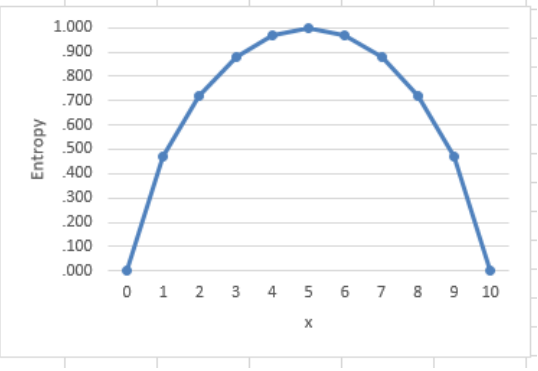
---

[59] 以下では対数の底を下付き文字でなく，見やすいように括弧( )に入れて示します: log(2)。

[60] =2/10*-LOG(2/10,2)+8/10*-LOG(8/10,2)

D	x	y
a	0	10
b	1	9
c	2	8
d	3	7
e	4	6
f	5	5
g	6	4
h	7	3
i	8	2
j	9	1
k	10	0

横軸	エントロピー
a	.000
b	.469
c	.722
d	.881
e	.971
f	1.000
g	.971
h	.881
i	.722
j	.469
k	.000



これを見ると，エントロピーの範囲は[0, 1]であり，成分(確率)の極端な分布({0, 10}, {10, 0})を示すときに最小値(0)になり，完全に均一な分布{5, 5}を示すときに最大値(1)になっています。確率が完全に均一な分布を示すときはまったく予想が立たないので(p = 1/2)，このような状態のときはエントロピーが最大の 1 になります。

データのそれぞれの成分の頻度(x)を和(s)で割った値を，その成分の確率とし(p = x / s)，全部の確率(p)の対数の期待値をエントロピー(E)として，次の式で求めます。

$$E = -\Sigma (i)\ p(i) * \log(2)\ p(i)$$

●相対エントロピー

(10, 12, 9, 11, 15)のような偏りが小さなデータでは，エントロピー(E)は大きくなり(E=2.299)，逆に(1, 3, 5, 16, 99)のように大きな偏りがあるデータのエントロピー(E)は小さくなります(E=1.013)。しかし，エントロピーの数値が[0, 1]の範囲に相対化されていないので，どのように評価すればよいのか判断が困難になることがあります。そこで，エントロピーの最大値 E.max を求め，それを使った「相対エントロピー」(Relative Entropy: RE)が考えられています(RE = E / E.max)。相対エントロピーの範囲は常に[0, 1]になるので，それを使えば情報量を常に同じ尺度で評価・比較できます。

エントロピーの最大値 E.max は，(5, 5, 5, 5, 5)のようにすべての値が同じとき，つまり確率が完全に同じときに現れます(たとえば{p: 1/5, 1/5, 1/5, 1/5, 1/5)。その成分の確率はすべて 1/N ですから(たとえば{p: 1/2, 1/2}, {p: 1/5, 1/5, 1/5, 1/5, 1/5}など), {p: 1/N, 1/N, 1/N, ...}のデータが示す最大のエントロピーE.max は

$$\begin{aligned}
E.max &= -\Sigma\ 1/N \log(2)\ 1/N && \leftarrow p(i) = 1/N \\
&= -\ 1/N\ \ \Sigma \log(2)\ 1/N && \leftarrow \Sigma \text{ 内の } 1/N \text{ は i をもたない} \\
&= -\ 1/N\ \ \Sigma [\log(2)\ 1 - \log(2)\ N] && \leftarrow \log\ a/b = \log\ a - \log\ b \\
&= -\ 1/N\ \ \Sigma -\log(2)\ N && \leftarrow \log(2)\ 1 = 0\ \ \leftarrow 2\text{^}0 = 1 \\
&= -\ 1/N\ N\ [-\log(2)\ N] && \leftarrow \Sigma \text{ 内の } \log(2)\ N \text{ は i をもたない} \\
&= \log(2)\ N
\end{aligned}$$

このようにエントロピーの最大値 E.max はデータの値そのものには関わりがなくデータ数 N だけから計算します[61]: E.max = log(2) N。

よって，相対エントロピー(RE)の式は次の通りです。

$$RE = E / E.max = [-\Sigma p(i) * \log(2)\ p(i)] / \log(2)\ N$$

*	A	B	C	D	E	entropy	entropy.r
h1	10	19	14	7	12	2.2461	0.9673
h2	11	7	10	0	1	NaN	NaN
h3	0	0	1	12	1	NaN	NaN
h4	0	1	2	3	3	NaN	NaN

●ゼロ(0)を含むデータのエントロピー

データの中にゼロ(0)があると，エントロピーの式の中の log(2) p(i)の計算で log2 (0)が-Inf を返すため，結果としてエントロピーは NaN(非数)になります。よって，ゼロ(0)があるデータではエントロピーの計算ができません。しかし，上のデータの h3 と h4 を比べると，明らかに h4 のほうが変動だ少なく，よってエントロピーは大きくなるはずです。そこで，h2, h3, h4 のデータの中のゼロ(0)がなくすために,データ全体に 0.01 を足すことにします。0.01 を足しても，大きくエントロピーに影響しないと思われるからです。このことを実験して確かめましょう。

*	A	B	C	D	E	entropy	entropy.r
h1'	10.01	19.01	14.01	7.01	12.01	2.2462	0.9674
h2'	11.01	7.01	10.01	0.01	1.01	1.7275	0.7440
h3'	0.01	0.01	1.01	12.01	1.01	0.7544	0.3249
h4'	0.01	1.01	2.01	3.01	3.01	1.9025	0.8193

この実験によると，先の h1 のエントロピー・相対エントロピーと今回の h1'　(h1+0.01)のエントロピー・相対エントロピーは僅かな差がありますが，ほぼ近似しています。この差を重視して，全体の分析を放棄するか，または近似値としてプラス 0.01 のデータを使用したエントロピー・相対エントロピーを使用するか，判断が難しいところですが，私は前向きに後者を採用したいと思います。この実験の結果，h4 のほうが h3 よりもエントロピー・相対エントロピーが大きいことがわかりました。

参考：

Hocket, Charles Francis. 1955. *A Manual of Phonology*. Baltimore. Wavery Press. pp. 215-218.

Kucera, Henry and George K. Monroe. 1968. *A Comparative Quantitative*

[61] entropy(5,5,5,5,5) = log2(5) = 2.322; entropy.r(5,5,5,5,5) = 1.

Phonology of Russian, Czech, and German. New York. Elsevier. pp. 65-82.
Shannon, Claude E. and Warren Weaver. 1963. *The Mathematical Theory of Communication*. pp. 8-16.
高岡詠子. 2012『シャノンの情報理論入門』講談社. pp.49-75.

●プログラム

```
entropy =　　function(A){P=A/sum(A); -sum(P*log2(P))} #エントロピー
entropy.r = function(A){P=A/sum(A); -sum(P*log2(P))/log2(length(A))}
#相対エントロピー
```

■スペイン語の母音と子音のエントロピー

次のリストは，スペインの演劇作品(Antnio Buero Vallejo. 1949. *Historia de una escalera. Madrid.* Espasa-Calpe)の音素表記に現れた母音と子音(連続)の頻度数を示します。子音は子音全体，頭子音(音節の初め，母音の前にある子音)，尾子音(音節の後ろ，母音の後にある子音)を数えました。

母音：e: 5087, a: 4546, o: 3666, i: 1826, u: 804.
子音全体：s: 2785, n: 2499, r: 2283, d: 1571, t: 1530, k: 1511, l: 1177, m: 1086, p: 990, b: 904, z: 398, g: 356, x: 254, f: 186, tr: 179, ch: 141, ll: 126, rr: 126, pr: 123, ñ: 103, dr: 102, br: 75, gr: 48, bl: 45, kr: 33, pl: 28, kl: 16, fr: 13, gl: 4, ns: 3.
頭子音：d: 1536, t: 1530, k: 1493, r: 1393, n: 1161, s: 1097, m: 1086, p: 986, b: 904, l: 856, g: 351, z: 342, x: 253, f: 185, tr: 179, ch: 141, ll: 126, rr: 126, pr: 123, ñ: 103, dr: 102, br: 75, gr: 48, bl: 45, kr: 33, pl: 28, kl: 16, fr: 13, gl: 4.
末尾子音：s: 1688, n: 1338, r: 890, l: 321, z: 56, d: 35, k: 18, g: 5, p: 4, ns: 3, f: 1, x: 1.

次の表は，それぞれの標準偏差，相対標準偏差，均等度，エントロピー，相対エントロピーを示します。

縦軸	母音	子音:全体	子音:頭子音	子音:尾子音
標準偏差	1625.874	806.921	541.617	574.118
相対標準偏差	.255	.240	.207	.476
均等度	.745	.760	.793	.524
エントロピー	2.106	3.830	3.999	2.001
相対エントロピー	**.907**	**.781**	**.823**	**.558**

母音は子音全体よりも相対エントロピーがかなり高いので情報量が多く

(.907, .781)，頭子音の相対エントロピーは尾子音の相対エントロピーよりも非常に高いのでそれだけ情報量が非常に多いことがわかります(.823, .558)。

母音と子音全体の均等度とエントロピーの大小関係は同じですが(.745 < .760; 2.106 < 3.830)，相対エントロピーの大小関係は逆転しています(.907 > .781)。この理由は相対エントロピーの分母の計算に N が入っているためです(log(2) N)。母音の数は 5 なので(N=5)，相対エントロピーの式の分母は log(2) N = log(2) 5 = 2.32 ですが，子音全体の数(N)は圧倒的に多く(N=30)，log(2) N = log(2) 30 = 4.91 が相対エントロピーの式の分母になって相対エントロピーの値を大きく下げています。このように，エントロピー(E)はデータの個数(N)が大きいほど大きくなりますが(次の式の Σ は N 個の和を示します)，相対エントロピー(RE)では個数(N)の影響が取り除かれています(次の式の/ log(2) N に注意)。

$$E = -\Sigma\, p(i) \log(2)\, p(i)$$
$$RE = -\Sigma\, p(i) \log(2)\, p(i) / \log(2)\, N$$

4.6.10. 最小値最大値比

データの変動を示す指標として最小値(Min)と最大値(Max)の平等性・格差を考えます。最小値と最大値の差を示す指標は「範囲」（Range. 最大値－最小値）ですが，最小値と最大値の比も重要な指標になります。両者の比として，最小値／最大値と最大値／最小値が考えられますが，後者は最小値がゼロのときに計算ができなくなるので前者をとります。よって**最小値最大値比**(Min-Max Ratio: M.M.R.)は

$$M.M.R. = Min / Max$$

最小値最大値比(M.M.R.)の最小値(0)は Min=0 のときで，最小値最大値比(M.M.R.)の最大値(1)は Min=Max のときに（つまりすべての値が同じときに）生じます。よって，最小値最大値比(M.M.R.)を「平等性」(equality)を示す指標とします。

一方，最小値最大値比(M.M.R.)の 1 に対する補数(1-M.M.R.)は「格差」(gap)を示します。これは次のように**範囲最大値比**(Range Max Ratio: R.M.R.)になります。

$$R.M.R. = 1 - Min / Max = (Max - Min) / Max = Range / Max$$

X	v1	v2	v3	v4	v5
d1	10	19	14	7	12
d2	11	7	10	0	1
d3	0	0	1	12	1

横軸	最小値最大値比	範囲最大値比
d1	.368	.632
d2	.000	1.000
d3	.000	1.000

d4	0	1	2	3	3	d4	.000	1.000

最小値最大値比と範囲最大値比は,最小値と最大値だけを使用するので,異常値の影響を強く受けます。

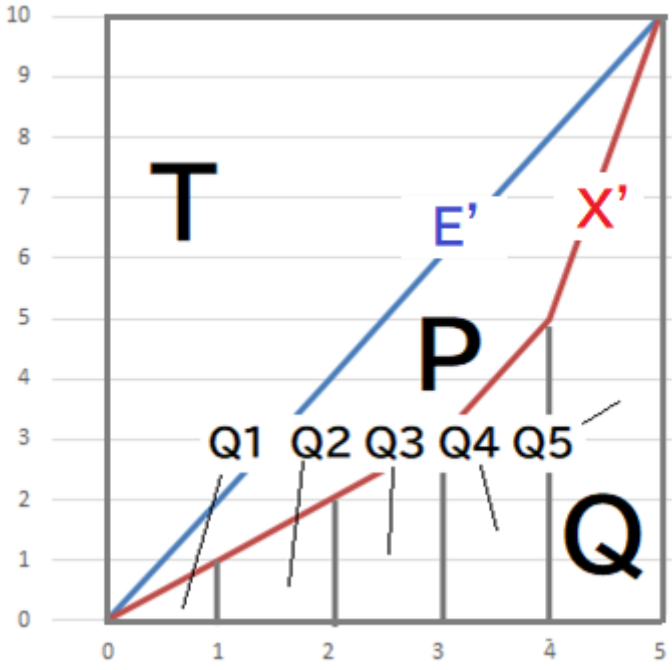
4.6.11. ジニ係数

経済学や社会学で論じられる所得の不平等を示す指数として「ジニ係数」(Gini coefficient: G)が使われます（範囲: [0, 1]）。不平等性を示すジニ係数(G)は偏差の指標になるので，1 - G は平等性，つまり一種の拡散度を示します。その計算の仕方をデータ{2, 5, 1, 1, 1}で説明します(倉田・星野 2009:59)。

はじめにこのデータを昇順にソートして，{X: 1, 1, 1, 2, 5}とします。次に，それぞれの値を累加して，{X': 1, 2, 3, 5, 10}とします。X'のそれぞれの成分は，その成分以下の成分全体の和になります。

X	X'
1	1
1	1 + 1 = 2
1	1 + 1 + 1 = 3
2	1 + 1 + 1 + 2 = 5
5	1 + 1 + 1 + 2 + 5 = 10

このような不平等な分布に対して，完全に平等な分布は{E: 2, 2, 2, 2, 2}で，その累加データは{E': 2, 4, 6, 8, 10}になります。これを X'と E'をグラフにすると次のようになります。



ジニ係数は不平等なX'が完全に平等なE'からどの程度離れているかを示す数値です。そこで上のE'とX'の2つの線に囲まれた範囲の面積(P)を求めます。その最小値はX'の線がE'の線と一致した場合で，そのときの面積はゼロです。その比較値を上のグラフの下三角形の面積(T)とします。上三角形でも同じです。ジニ係数(G)は

[1]　　G = P / T

Tは総和(s=10)を使って求めます。

[2]　　T = n * s / 2　　←n: 個数; s: 和

上の図のTとQの面積は

T　　= 5 * 10 / 2 = 25

Q　　= Q1 + Q2 + Q3 + Q4＋Q5
　　= 1*(0+1)/2 + 1*(1+2)/2 + 1*(2+3)/2 + 1*(3+5)/2 + 1*(5+10)/2
　　= (0+1)/2 + (1+2)/2 + (2+3)/2 + (3+5)/2 + (5+10)/2
　　= (0+1 + 1+2 + 2+3 + 3+5 + 5+10) /2
　　= [(1+2 + 3 +5 + 10) * 2 -10}/2
　　= (1+2 + 3 +5 + 10) - 10/2　= 16

上のQの式はそれぞれの台形の面積の総和(Q)を示します。それぞれの台形の上辺（左側の縦の長さ）と下辺（右側の縦の長さ）の和に高さ(=1)を掛け，それを2で割ります。一般化するために最初の三角形(Q1)も台形と見なしてその上辺をゼロ(0)にします。それぞれの台形の高さは1です。
よって

[3]　　Q　= Σ X'(i) - Σ (i) / 2 = Σ X'(i) - s / 2　←s: 和

よって，ジニ係数(G)は

G　= P / T　　←先の式 [1]
　　= (T - Q) / T　←P + Q = T
　　= 1 - Q / T　　←分母を整理 (1 - 16/25 = 0.36)
　　= 1 - [Σ X'(i) - s / 2] / [(n*s) / 2]　　←[2], [3]
　　= 1 - [2* Σ X'(i) - s] / (n*s)　　←分母を整理

次はx, y, zというデータ例で，これまでの変動を示す値とGini係数を比較した表です。

Gini	x	y	z
1	2	2	3
2	5	2	3
3	1	3	3
4	1	2	1
5	1	1	0

値	x	y	z
標準偏差	1.549	.632	1.265
変動係数	.775	.316	.632
相対標準偏差	.387	.158	.316
拡散度	.613	.842	.684
Gini 係数	.360	.160	.320

先述のように,限定標準偏差や拡散度はNが大きくなると減少しますが,Gini係数はNの増加による影響を受けません。次は同じデータを繰り返した場合ですが,Gini係数は変化していません。これは注目すべき特徴です。

Gini	x	y	z
1	2	2	3
2	5	2	3
3	1	3	3
4	1	2	1
5	1	1	0
6	2	2	3
7	5	2	3
8	1	3	3
9	1	2	1
10	1	1	0
11	2	2	3
12	5	2	3
13	1	3	3
14	1	2	1
15	1	1	0
16	2	2	3
17	5	2	3
18	1	3	3
19	1	2	1
20	1	1	0

値	x	y	z
標準偏差	1.549	.632	1.265
変動係数	.775	.316	.632
限定標準偏差	.178	.073	.145
拡散度	.822	.927	.855
Gini 係数	.360	.160	.320

＊Gini係数については倉田・星野(2009:51-60)を参照しました。

プログラム(R)

```
gini = function(A){ #ジニ係数(A: データ配列)
  s = sum(A); n = length(A); A = sort(A) #s:和; n:個数; A:昇順ソート
  q = 2*sum(cumsum(A)) - s; t = n*s; 1 - q/t #q: 累積和
}
```

●相対ジニ係数

Gini 係数(G)の最大値は{0, 0, 0, 0, 10}のように 1 つだけに数値がある極端なケースになりますが，このとき G = 1 にはなりません。そこで，Gini 係数の最大値 G.max を次のようにして求め，それを使って，Gini 係数を相対化します。

{0, 0, 0, 0, 10}のようなデータ(X)は，その累積和(X': cumsum)も元のデータと同じです(X' = X)。この例では，最初の 4 個目までは累積はすべて 0 だからです。そうすると，先のジニ係数(G)の式の最大値は次のように簡単になります。

$$
\begin{aligned}
G.max &= 1 - [2^* \Sigma X'(i) - s] / (n * s) \\
&= 1 - [2^* \Sigma X(i) - s)] / (n * s) && \leftarrow X' = X \\
&= 1 - (2s - s) / (n * s) && \leftarrow \Sigma X(i) = s \\
&= 1 - s / (n * s) \\
&= 1 - 1 / n \\
&= (n - 1) / n
\end{aligned}
$$

よって，「相対ジニ係数」(Relative Gini: R.Gini)は

$$R.Gini = G / G.max = G / [(n-1) / n] = G * n / (n - 1)$$

このように「相対ジニ係数」(R.Gini)の式は，「ジニ係数」(Gini: G)とほとんど同じで，わずかに* n / (n - 1)の部分が異なります。n が小さいときは Gini と R.Gini は大きく異なりますが，n が大きくなると両者の違いが目立たなくなります。

プログラム(R)

```
gini.r = function(A){n = length(A); gini(A)*n/(n-1)} #相対ジニ係数
```

G	x	y	z
1	2	2	3
2	5	2	3
3	1	3	3
4	1	2	1
5	1	1	0

縦軸	x	y	z
ジニ係数 G	.360	.160	.320
相対ジニ係数 G.r	.450	.200	.400

x の相対ジニ係数 = 0.36 * 5 / 4 = 0.45

●相対ジニ係数の別式

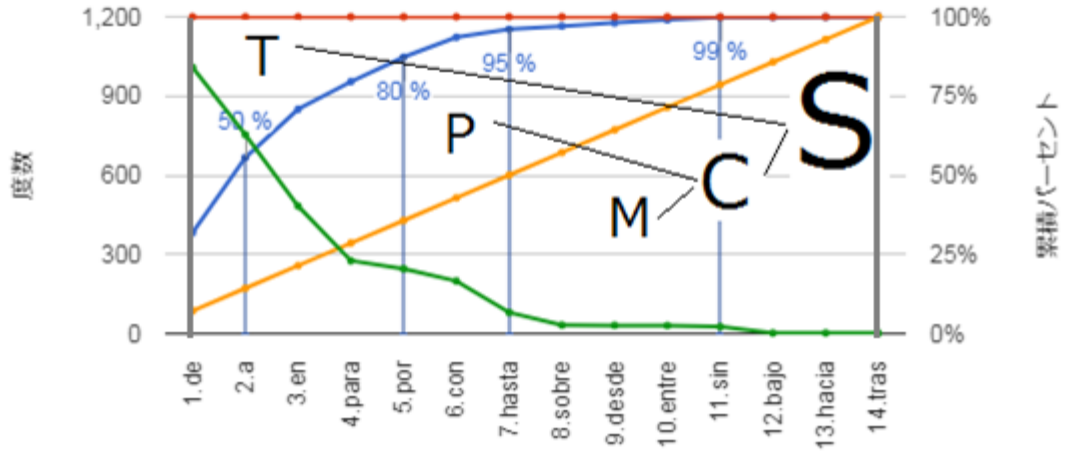
次はスペイン語の資料[62]に出現する前置詞の頻度(和), 累積和, 和の平均, 和の平均の累積（累積和のパーセント(%)を示します。F.度数をキーにして降順で並べ替えました (個数: n = 14)。

Prep	F:度数	C:累積度数	A:平均	M:累積平均	T:総和
de	1008	1008	226.4	226.4	3169
a	753	1761	226.4	452.7	3169
en	483	2244	226.4	679.1	3169
para	276	2520	226.4	905.4	3169
por	245	2765	226.4	1131.8	3169
con	199	2964	226.4	1358.1	3169
hasta	80	3044	226.4	1584.5	3169
sobre	32	3076	226.4	1810.9	3169
desde	31	3107	226.4	2037.2	3169
entre	30	3137	226.4	2263.6	3169
sin	26	3163	226.4	2489.9	3169
bajo	2	3165	226.4	2716.3	3169
hacia	2	3167	226.4	2942.6	3169
tras	2	3169	226.4	3169.0	3169
sum: 和	3169	38290	3169.0	23767.5	44366

上表を見ると最大頻度(*de*)から頻度が急降下していることがわかります。

次の図は頻度(緑線), 累積の%（青線）, 平均の累積（黄線）, 全体%(100%)（赤線）を示したもので「パレート図」(Pareto chart)と呼ばれます。図の左の軸は頻度に対応し, 右の軸は累積頻度・全体・平均のパーセント(%)に対応します。

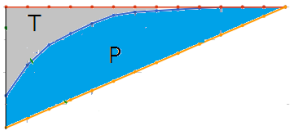
[62] VARITEX: https://lecture.ecc.u-tokyo.ac.jp/~cueda/lyneal/varitex.htm (2018/04/26)



この図を見ると上位 2 語(de, a)が全体の 50%以上の累積度数を占めていることがわかります。全部で 14 語であるので，2/14 = 14.3%の語で全体の半分以上を占めていることになります。同様に，80%の位置で 5/14 (35.7%), 95%の位置で 7/14 (50.08%), 99%の位置で 11/14 (78.6%)に対応しています。

上のように累積度数(P)が急上昇するのは上位に高頻度が偏って集中しているためです。すべての語が同じ頻度であれば，その累積度数を示す線は上の図の左下の 0%に近い点から右上の 100%点に達する斜線（黄）と同じ線になるはずです。そして，逆に，すべての度数が 1 位に集中し，他がゼロであれば，その累積度数を示す線は上の図の 100%の横線（赤）と一致します。そこで，上の図の横線（赤）と曲線（青）で囲まれた T (top)とし，曲線（青）と斜線（黄）で囲まれた領域の総和を P (Pareto: パレート数)とします。そして，「累積度数の集中している度合い」を示す「相対パレート数」(Relative Pareto: RP)を次の式で定義します。

$$RP = P / (T + P)$$



この相対パレート数(RP)の値が大きいほど高頻度語の偏りが高い，と言えるでしょう。曲線（青）が斜線（黄）と一致すると，頻度の分布が完全に平均化していることになり，そのときの RP は 0 になります。また，曲線（青）が横線（赤）が一致すると，頻度の分布が完全に最初の項目に集中していることになり，そのときの RP は 1 になります。よって，RP の範囲は[0, 1]です。

RP = 0 (P = 0)　　RP = 1 (T = 0)

相対パレート数の式の RP = P / (T + P)の T と P は直接計算できません。そこで RP を直接に計算できる C, S, M で表すことにします。

C = P + M　← C: 累積頻度の和（青線と底辺で囲まれた総和），
　　M: 累積平均（斜線と底辺で囲まれた総和）
S = T + C　← S: 累積和（赤線と底辺で囲まれた総和）

とすると

T = S - C
P = C - M
T + P = S - M

よって

RP = P / (T + P) = (C - M) / (S - M)

具体的に先のスペイン語前置詞の例の相対パレート数(RP)を次のように計算します。

C = Σ(i) C(i) = 38290
M = Σ(i) M(i) = 23767.5
S = n * s =14 * 3169 = 44366
RP = (C - M) / (S - M) = (38290 - 23767.5) / (44366 - 23767.5) = 0.705

プログラム (R)

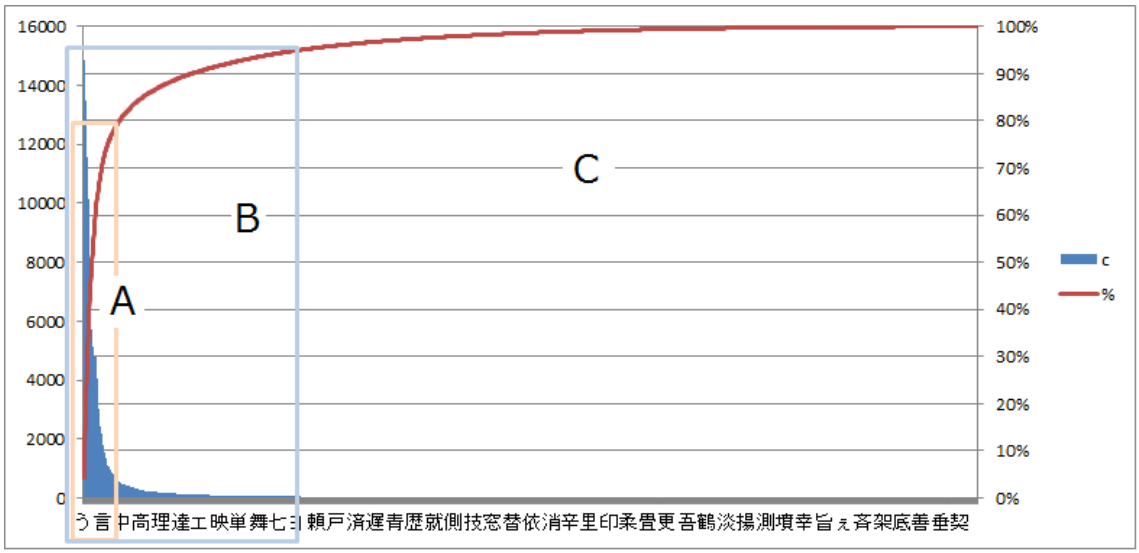
```
RP = function(A){ #相対パレート数 (Relative Pareto)
  A = sort(A, decreasing=T); n = length(A); c = sum(cumsum(A))
  m = sum(cumsum(rep(mean(A), n))); s = n * sum(A); (c - m) / (s - m)
}
```

この相対パレート数(RP)は相対ジニ係数(Gini.r)と同じ値になります。

4.6.12. L 字分布係数

次の図はある資料中の日本語文字（ひらがな，カタカナ，漢字）の降順

頻度分布の ABC 分析（→前節）の結果を示します[63]。



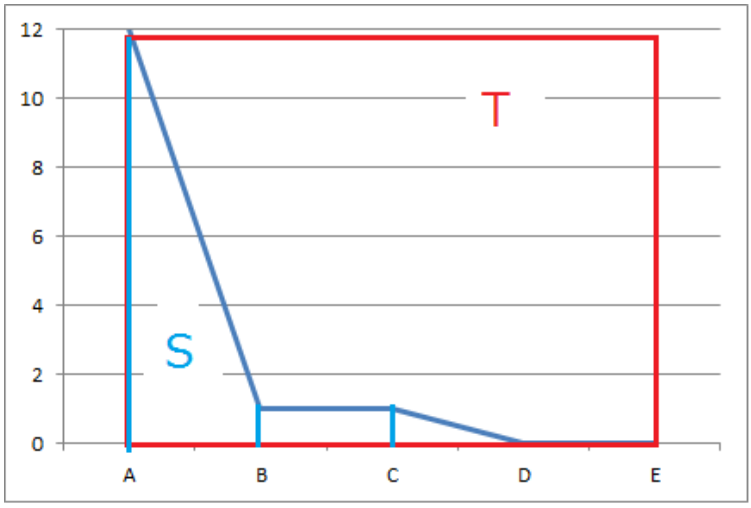
【図】日本語文字の頻度分布[64]

日本語文字の頻度分布では少数の高頻度語と多数の低頻度語の分布の違いが顕著です。このような特徴をもつ頻度分布は「L 字型分布」(L-shaped distribution)と呼ばれます。この L 字型分布の顕著さの度合いを「相対 L 字分布係数」(L-shaped distribution index: LDI)と呼び，次のように定式化します。

たとえば{1, 0, 12, 1, 0}というデータを降順に並べ替えると，{12, 1, 1, 0, 0}のようになります。次の図のような顕著な L 字型分布であれば，赤で示した領域の中で占める青の領域の割合が顕著に小さくなります。

[63] 日本語文字のタイプ数は非常に多いので，グラフの文字はすべての文字を示してありません。たとえばグラフの最初の文字「う」には次の文字が含まれます：う(14881)，の(13674)，い(13463)，ん(12572)，っ(11784)，て(11557)，か(10514)，な(10371)，あ(10168)，で(9823)，と(8568)，た(7952)，そ(7715)，し(5954)，に(5727)，も(5616)，え(5606)，ら(5142)，だ(5045)，は(4895)，ま(4782)，る(4782)，す(4776)，れ(4701)，が(4610)，こ(4057)，ね(3556)，け(3389)，ど(3034)，り(2716)，ー(2474)，を(2441)，く(2417)，よ(2396)，ち(2205)，や(2116)，ゃ(1844)，き(1789)，お(1657).

[64] 資料は次のサイトの C-ORAL-JP を使いました。
http://lecture.ecc.u-tokyo.ac.jp/~cueda/lyneal/lli-uam.htm (2017/08/25)



赤の領域が示す全体(Total: t)は

$t = mx * n$　　mx: 最大値; n: 個数

青の領域が示すデータの和(Sum: s)は

$s = \Sigma(i)\ X(i)$　　X(i): 個々のデータの頻度

全体の中でデータの和 s が占める割合(s / t)は

$s / t = \Sigma(i)\ X(i) / (mx*n)$

$= s / (mx*n)$　　s: データの和

この s / t は顕著な L 字型を示す分布では小さくなりますから，「L 字分布性」(degree of L-shaped distribution: DLD)は

$DLD = 1 - s / t = 1 - s / (mx*n)$

この DLD は，たとえば{10, 0, 0, 0, 0}のように，最初の要素だけに頻度があり他はすべて 0 であるデータで最大になるはずです。このようなデータの最初の要素の頻度 X(1)とデータの和(s)と最大値(mx)が同じです。よって DLD の最大値 DLD.max は

$DLD.max = 1 - mx / (mx*n)$　　← s = mx

$= 1 - 1 / n$

$= (n - 1) / n$

L 字分布係数(LDI: L-shaped distribution index)は

$LDI = DLD / DLD.max$

$= [1 - s / (mx*n)] / [(n - 1) / n]$

= [(mx*n - s) / mx*n] / [(n - 1) / n]
= (mx*n - s) / [mx*(n - 1)]

たとえば{10, 0, 0, 0, 0}のように，最初の要素だけに頻度があり他はすべて 0 であるデータで L 字分布係数(LDI)は最大になります(LDI.max)。このとき，和と最大値は一致します(s = mx)。よって，

LDI.max = (mx*n - mx) / [mx*(n - 1)] ← s = mx
= [mx*(n - 1)] / [mx*(n - 1)] = 1

逆に，たとえば{10, 10, 10, 10, 10}のように，すべて 10 であるデータでは L 字分布係数(LDI)は最小になります(LDI.min)。このとき，mx*n は和(s)と等しくなり，次のように LDI.min はゼロ(0)になります。

LDI.min) = (s - s) / [mx (n - 1)] ← s = mx n
= 0

次はデータ例(M)とその和(S)と相対 L 字分布係数(LDI)の計算例です。比較のために分布の不平等性を示すジニ係数(G)と相対ジニ係数(RG)の計算結果も載せました。h3 分布の偏りを比較すると，LDI の感度が良いと思います。

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

横軸	S	G	RG	LDI
h1	62	.181	.226	.434
h2	29	.428	.534	.591
h3	14	.714	.893	.958
h4	9	.356	.444	.500

プログラム (R)

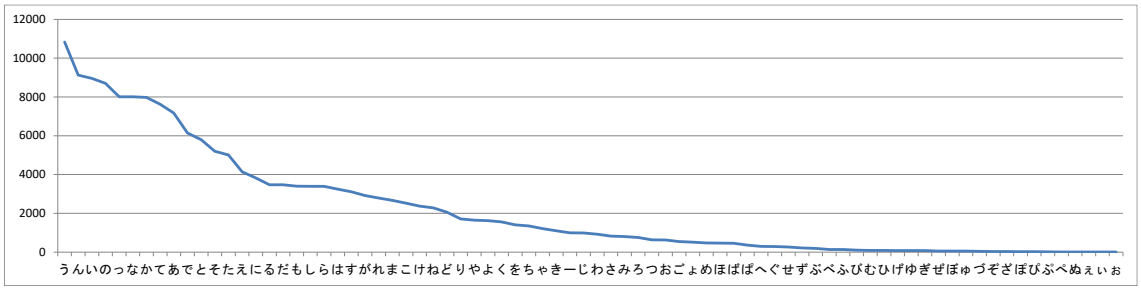
```
ldi = function(A){n=length(A); mx=max(A); (mx*n - sum(A)) / (mx*(n - 1))}
#L 字分布係数 (L-shaped distribution index)
```

上の説明では，L 字型との比較をするためにソートしたデータを用いましたが，ソートしなくても結果は同じになります。

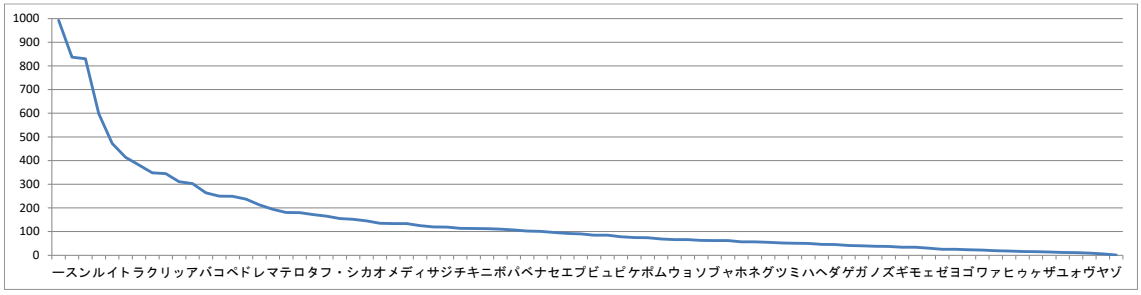
■ひらがな・カタカナ・漢字の L 字頻度分布

次はある日本語資料中のひらがな, カタカナ, 漢字の降順頻度分布図です[65]。

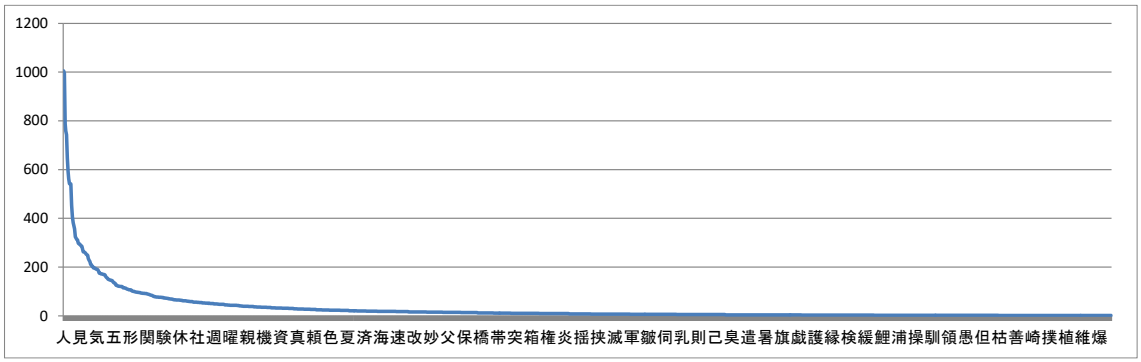
[65] 資料は次のサイトの C-ORAL-JP を使いました。
http://lecture.ecc.u-tokyo.ac.jp/~cueda/lyneal/lli-uam.htm (2017/08/25)



【図-1】ひらがなの降順頻度分布



【図-2】カタカナの降順頻度分布



【図-3】漢字の降順頻度分布

次の表はデータ例(M)とその和(S)，個数(N)，最大値(Mx)，相対ジニ係数(RG)と L 字分布係数(LDI)の計算結果です。

文字	S	N	Mx	RG	LDI
ひらがな	170 868	78	10 828	.638 (r: 2)	.808 (r: 1)
カタカナ	11 860	80	992	.557 (r: 1)	.861 (r: 2)
漢字	44 787	1 443	1 005	.746 (r: 3)	.970 (r: 3)

上の表の(r: *)は順位(rank)を示します。L 字分布係数(LDI)のランク(r)はひらがな→カタカナ→漢字となっていますが，相対ジニ係数(RG)では，ひらがなとカタカナのランク(r)が異なります。図を見ると，L 字型分布の特徴はひらがな→カタカナ→漢字の順で強くなっているので，L 字型分布の観点から見れば L 字分布係数(LDI)のランク(r)が適切です。また，その数値についても L 字分布係数(LDI)のほうが相対ジニ係数(RG)よりも適切だと思います。

4.6.13. 頻度の多様性

(1) TTR

たとえば a, b, c, d という単語が{a, b, c, b, b, a, d}のように並んでいるテキストの「異なり語数(タイプ数)」(number of different words; Type)は 4 語{a, b, c, d}になり，「全語数・延べ語数・トークン数」(total of words, number of running words; Token)は 7 語です。このようなテキストの「語彙の多様性」(lexical diversity)を示す指標として異なり語数(Type)を延べ語数(Token)で割った「タイプ・トークン比」(Type Token Ratio: TTR)が使われます。これは語彙の多様性が語のタイプ数に比例する，と考えられるからです。そしてテキスト内の語のタイプ数はテキストの全語数が多ければ，それに従って多くなるはずなので，それを全語数で割って相対化したものです。

TTR = Type / Token = 4 / 7 = .571

TTR の最大値は Type = Token のときに発生する 1 ですが，TTR の最小値は，その分子の Type が必ず 1 以上になるので，ゼロ(0)にはなりません。そこで TTR の範囲を[0, 1]にするために次のように修正します(修正タイプ・トークン比 TTR.c: corrected)[66]。

TTR.c. = (Type - 1) / (Token - 1) = 3 / 6 = .500

しかしテキストの語数(Token)が多くなっても，タイプ数(Type)は同じように増加するのではなく，ふつう次第に同じ語が繰り返されることが多くなりタイプ数の増加は鈍くなります。したがってテキストの延べ語数が大きくなるとタイプ・トークン比(TTR)は一般に小さくなるので，タイプ・トークン比を使って大きさ（延べ語数）の異なるテキスト間で語彙の多様性を比較することはできません。また，同じ理由で唯一頻度語数を総語数で割った値でテキストの語彙の多様性を比較することもできません。

そこで大きさの異なるテキストの語彙の多様性を比較するときに，テキストの大きさに影響されないように大きさを揃えなければなりません。テキストの大きさをたとえば比較する文書の中で一番小さな文書の語数に揃えた上でタイプ数を数え，それを同じトークン数で割ってタイプ・トークン比(TTR)を求めます。たとえば A (120,000 語), B (250,000 語) y C (502,000 語)という 3 つの文書を比較するとき，B と C の異なり語の頻度を A の規模に合わせて，B (50, 25, 44, ...) を(50 * 120 / 250, 25 * 120 / 25, 44 * 120 /

[66] 分母で Token から 1 を引く理由は，Type = Token のときに最大の 1 にならなければならない，という条件があるためです。なお，TTR と TTR.c. の計算結果は小規模のテキストでは大きな差を示しますが，大規模テキストでは，修正値を使っても結果はほとんど変わりません。

250, ...)とします。同様に C の語の頻度 (88, 52, 60, ...)を(88 * 120 / 502, 52 * 120 / 502, 60 * 120 / 502, ...)とします。このようにして相対化した頻度で計算したタイプ数(Type.r.)とトークン数(Token.r.)の比を「相対化タイプ・トークン比」(Type Token Ratio, relativized: TTR.r.)と呼ぶことにします。

TTR.r. = Type.r. / Token.r.

タイプ・トークン比(TTR)は対象となるテキストのタイプ数を同じテキストの全語数で割った値ですから, 語彙の多様性を示す「内的・個別的な」指標です。「内的・個別的な指標」とはたとえばある地域の（一冊の文学作品に対応します），ある月（作品の章）の気温を示すために，同じ地域で半袖のシャツを着ている人々の割合を基準にして測るようなことになります。そのような指標も有用ですが，他に摂氏(Celcius)の温度のように水の氷点と沸点のような「外的・一般的な」指標（尺度）も必要です。

(2) Zipf

次に語彙の多様性を測定するための外的・一般的な尺度として以下に示す Zipf の公式を使うことを提案します。Zipf (1936: 44-48, 1949: 22-27)はさまざまな言語の大きな文書にある単語の使用頻度に次のような順位と頻度間の関係(rank-frequency relationship)が観察されることを指摘しました。

順位 R(ank) * 頻度 F(requency) = 定数 C(onstant)

この式は順位 R と頻度 F が反比例することを示します。たとえば 1 位の語の頻度（最大値）が 50 であれば，2 位の語の頻度は 25, 3 位の語の頻度は 16, 4 位の語の頻度は 12, ...，となります。この公式を実際のテキストにあてはめて各順位の頻度を見ると，かなりの誤差があることがわかりますが, 頻度分布の全体をおおまかに観察すると一般的傾向として上の式 R * F = C に近い分布を示します。よって Zipf の公式を実際のテキストと理論的な Zipf の計算結果の間に誤差があることから無用のものとせず, むしろこの公式を使って，観測値と理論値の誤差こそが観測されたテキストの語彙の頻度分布の特徴を示している，と考えます。Zipf の公式 R * F = C を使えば，同じ基準でさまざまなタイプ数(Type)を比較できるはずです。その準備として Zipf の公式 R * F = C の重要な性質を次表で確認します。

R	F = C / R	Int(F)	R * F = C
R1: 1	**F1=M: 50.00**	**50**	**50**
2	25.00	25	**50**
3	16.67	16	**50**
4	12.50	12	**50**
5	10.00	10	**50**

…	…	…	…
49	1.02	1	**50**
Rn: 50	**Fn: 1.00**	**1**	**50**
51	0.98	0	**50**

上表の R の列は順位, F = C / R の列は Zipf の公式によって計算された頻度，Int(F)の列は公式によって計算された頻度 F の整数部(integer)，R * F = C の列は R と F の積が C であることを示します。F は頻度を示すので整数になるはずですが，R * F = C の F (= C / F)は必ずしも整数にならないので便宜的に割り算の結果を小数点以下 2 位まで表示しておきました。

上の表で観察できる重要なポイントは，第 1 位の頻度(F1=50)，すなわち最大値(M)が同時に最下位の位置(Rn=50)，すなわち異なり語数(Type=1，2, …, 50)に一致し，式の定数(C = R * F = 50)になる，ということです。

F1 = M = Rn = Type = C

たとえば第 1 位の頻度(F1 = 最大頻度 M)が 50 であれば，最下位の順位(Rn)，すなわち異なり語数(Type)が 50 になり，定数(C)も 50 になります。このようにして，最大頻度(F1=M)さえわかれば，同時に Zipf の公式による最下位順位(Rn)，異なり語数(Type)，定数(C)が決定されるのです。

テキストの最大頻度(M)，たとえば英語のテキストの最大頻度(fm.*the* の頻度)が，実は「異なり語数」(Types: f1.*the*, f2.*of*, f2.*and*, …, … fn.)と同じになる(fm = fn)，ということは直ちには納得しにくいのですが，「Zipf の公式による」という条件をつけて実験すると，どんな最大頻度であっても異なり語数に一致するのです。その理由を次に示します。

最大頻度(F1)が最下位順位(Rn)と一致することは次のことを考えれば理解できます。たとえば最大頻度(F1=M)が 50 であって，F = C / R の公式に従って, 分母の順位(R)を 1 つずつ下げていくことを考えます(R=1, 2, 3, …)。この最大頻度 50 を，順位数(1, 2, 3, …)で次々に割っていくと，割り算の商が 1 以上であるのは分母（順位数）が 50 までです。なぜならば 50 を超えて 51 になると割り算の商が 1 未満になるからです(50/**49**=1.02，50/**50**=1, 50/**51** = 0.98)。

これは F1=M=50 に限らず，F1=M がどんな数であっても同じです。このことは Zipf の公式の重要な性質なので簡単な証明をしておきましょう。

R * F = C　　← Zipf の公式
R1 * F1 = C　　← R1：第 1 位，F1：第 1 位の頻度
1 * F1 = C　　← R1 = 1
1 * M = C　　← F1 = M: 最大頻度
M = C　　← 最大頻度(M) = 定数(C)
F = C / R　　← Zipf の公式: R * F = C

$F = M / R$ ← $M = C$
$F = M / R \geq 1$ ← 頻度(F)は 1 以上でなければならない（上の表）
$M \geq R$ ← 不等式の両辺に R（正数）を掛ける
$R \leq M$ ← 順位(R)の最大値は M(最大頻度)
$Rn = M$ ← Rn：最大順位

したがって「Zipf の公式による異なり語数」(Type.z)は上の表から R1~Rn の個数を示すので，上の式から Rn = M となり，テキストの最大頻度(M)と一致します。

Type.z. = Rn = M

(3) タイプ指数

このように「Zipf の公式によって求められる異なり語数」(Type.z.)は，最大頻度(M)に一致するので頻度順位表の上から順に探したりする必要がなく，最大頻度(M)をそのまま使えばよいので非常に便利です(Type.z. = M)。そこで,テキストの語彙の多様性を示す指標として次の「タイプ指数」(Type index: TI)を考えます。

TI = Type / (Type + Type.z.) = Type / (Type + M)

ここで Type はテキストの異なり語数，Type.z.は Zipf の公式による異なり語数，M はテキストの最大頻度を示します。上の式からタイプ指数(TI)次のように変化することがわかります。

TI = 0 ← Type = 0 のとき
TI = 0.5 ← Type = Type.z.のとき
TI = 1 ← Type = ∞ または Type.z.=0 (M=0)のとき

後述するようにテキストの最大頻度 M は一般にテキストの語数(Total)に直線的に比例します。よってタイプ指数を導出する関数の引数として信頼できます。

(4) 唯一頻度指数

語彙の多様性を示す別の指標として「頻度が 1 である語の数」を考えます。たとえば *accomplice, amiss, amity, ...*のような「稀にしか使われない語」です。テキストの中で「1 度しか使われていない語」であっても，それは「繰り返されていない語」ですから，その数が多ければテキストの語彙の使用が多様であることを示す，と考えられるからです。（逆に「繰り返される語」が多ければテキストの語彙の多様性が低いことになります。）

「頻度が 1 である語」は hapax legomenon (ギリシャ語:「1 度だけ読まれ

る」), 略して hapax と呼ばれます。そこで「頻度が 1 である語の数」を「唯一頻度語数」(Hapax)と呼ぶことにします。テキスト内の唯一頻度語数は 1 度しか使われない語を数え上げなければなりませんが, 語彙の頻度が「Zipf の公式に従う唯一頻度語数」であれば, 先の「異なり語数」(Type)と同じように, 次のようにして最大値(F1 = M)から簡単に求められます。

次表は最大頻度(F1=M)が 50 のときの Zipf の公式に従う順位(R)と頻度(F)を示します。これは先の表と同じですが, ここでは 25 位(R25)の付近を載せてあります。

R	F = C / R	Int(F)	R * F = C
R1: 1	**F1=M: 50.00**	**50**	**50**
2	25.00	25	**50**
3	16.67	16	**50**
…	…	…	**…**
R24: 24	2.08	2	**50**
R25: 25	**F25: 2.00**	**2**	**50**
R26: 26	1.92	**1**	50
R27: 27	1.85	**1**	50
…	…	**…**	**…**
Rn: 50	**Fn: 1.00**	**1**	**50**
51	0.98	0	**50**

この表から頻度（Int(F)）が 1 である語(hapax)は順位 26 番から 50 番までの 25 語であることがわかります。ここでは Hapax = 25 となっていますが, 表を見ると直感的に唯一頻度語数は M / 2 になると思われます。つまり Zipf の公式に従う hapax の数は最大頻度(M)の 1/2 になるようです。このことは F1= 40, 30, ...とする簡単な実験でも確かめられます。そして, 先の「Zipf の公式によって求められる異なり語数」(Type.z.)と同じように, 頻度(F)が 2 の最後の順位を想像すれば 50 / R = 2, よってその順位 R は 25 であることがわかります(50/**24**=2.08, 50/**25**=2, 50/**26**=1.92)[67]。R が 25 を超えると, その商が 2 以下となるからです(50/26=1.92)。このことを先と同じように次に証明します。

F = C / R ← Zipf の公式: R * F = C
F = M / R ← M = C ← 先の証明
F = M / R ≥ 2 ← 頻度(F)は 2 以上でなければならない（上の表）
M ≥ 2 R ← 不等式の両辺に R（正数）を掛ける
R ≤ M / 2 ← 順位(R)の最大値は M(最大頻度)
Hapax.z. = Rn - M / 2 = M - M / 2 = M / 2

[67] 最大値(M)が奇数のときは（たとえば 51）唯一頻度語数(Hapax)は半数(25.5)を切り上げた数値(26)になります。

← Hapax.z.: R ≤ M / 2 の範囲以外（上表の R26~Rn）の個数

よって「Zipf の公式から求められる唯一頻度語数」(Hapax.z)は

Hapax.z = M / 2

そこで「テキストの唯一頻度語数」(Hapax)を「Zipf の公式から求められる唯一頻度語数」(Hapax.z)で割った値を「唯一頻度指数」(Unique frequence index: UFI)とし，これを「語彙の多様性」を示す第二の指標とします。

UFI = Hapax / (Hapax + Hapax.z.) = Hapax / (Hapax + M / 2)

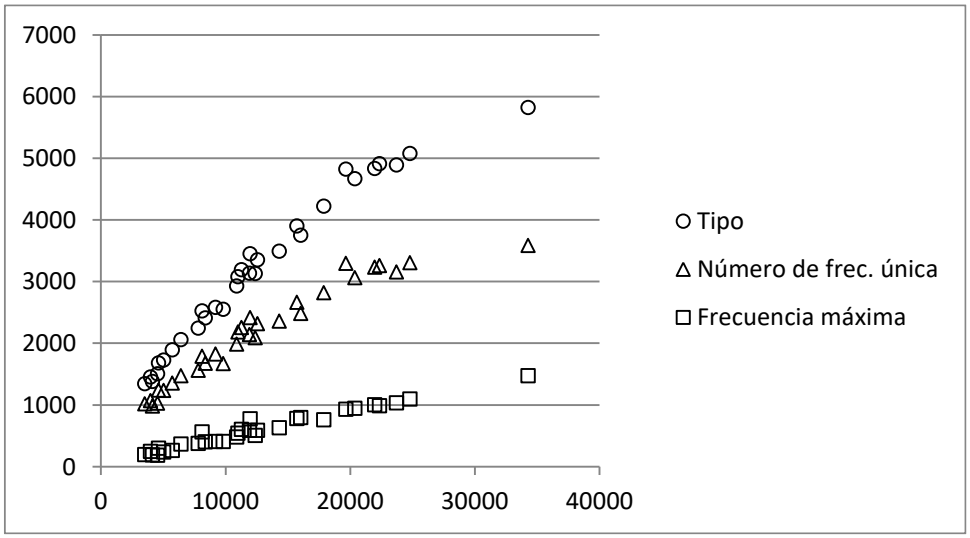
最大値(F1=M)が奇数の場合は Hapax.z = M / 2 に小数点以下の数値(0.5)が出ますが，本来唯一頻度語数は整数のはずです。しかし，ここでの目的は Hapax.z.を求めることではなく，Hapax.z.を使ってテキストの Hapax を相対化することなので,小数点以下の数値を残した精確な数値を使います。

上の式から明らかなように唯一頻度指数(UFI)が次のように変化します。

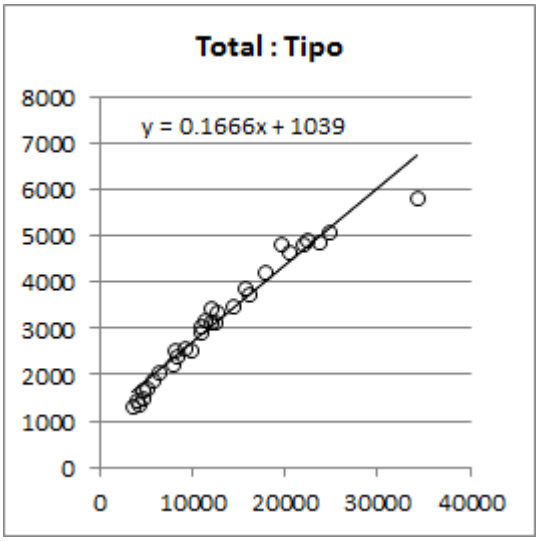
UFI = 0　　← Hapax = 0 のとき
UFI = 0.5　　← Hapax = Hapax.z.のとき
UFI = 1　　← Hapax → ∞ または Hapax.z.=0 (M=0)のとき

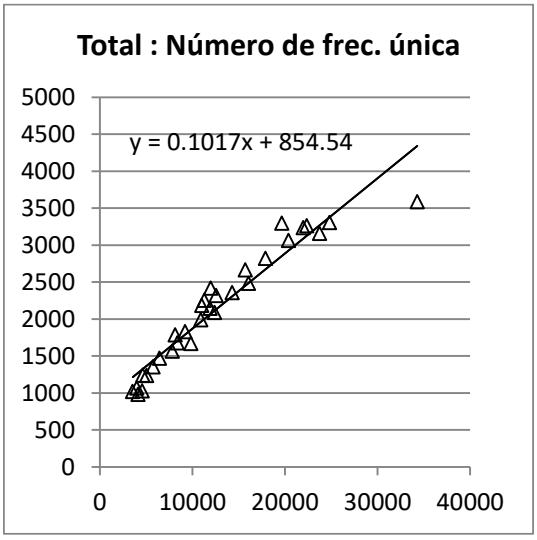
●全語数とタイプ数・唯一頻度語数・最大頻度の関係

次はスペインの小説家 Benito Pérez Galdós(1843-1920)の長編小説 *Fortunata y Jacinta*(1886)の各章(I.1~11; II.1~7; III.1~7; IV.1~6)の語形頻度の総和（Total:横軸）と，各章のタイプ数(Tipo), 唯一頻度語数(Número de frecuencia única: 「1度だけ使われている語の数」Hapax), 最大頻度(Frecuencia máxima: M)の数値（縦軸）を示す散布図です。この図を見ると，どの数値も語数に比例して規則的に上昇していることがわかります。これは直感的にも想像できるので納得しやすい結果です。しかし図をよく観察すると，タイプ数と唯一頻度語数の上昇の傾きが次第に鈍くなっていることもわかります。一方，最大頻度，具体的には定冠詞 el の頻度は非常に規則正しく直線的にテキストの全語数に相関しています。

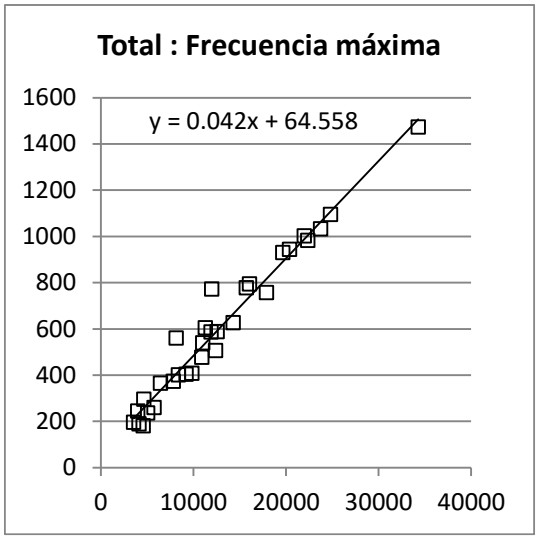


上の図では最大頻度の上昇率が鈍いように見えますが，これはタイプ数や唯一頻度語数と同じ平面で示しているためです。次の 3 図でテキストの全語数と各指数の関係を個別に観察しましょう。





Total y Tipo (R. = .978) / Total y Número de frecuencia única (R.= 959)



Total y Frecuencia máxima (R. = .98 1)

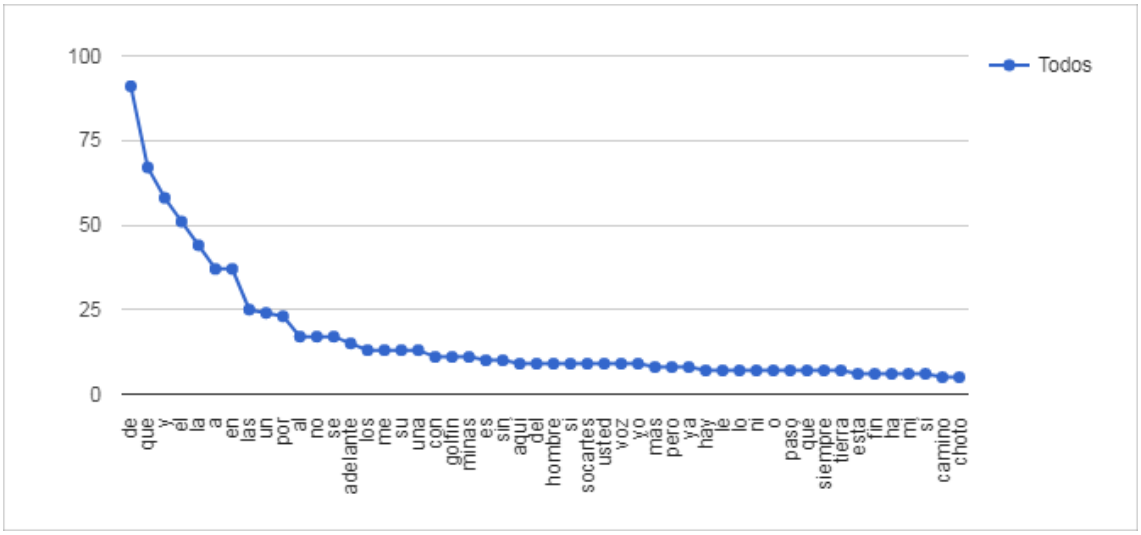
以上の観察から，最大頻度(M)がタイプ数(Type)や唯一頻度語数(Hapax)よりもテキストの規模を精確に示す値であることがわかりました。この結果から，最大頻度を Zipf の公式に適用してタイプ数を求める方法を考えました。

●頻度スペクトル

次は Benito Pérez Galdós(1843-1920)の短編小説 *Marianela* (1878) の第 1 章の語形を頻度を降順に並べた表です。全語数（トークン数）は 788 語です。

FA	Maria.1.
de	91
que	67
y	58
(…)	
abrazaré	1
absolutamente	1
absurdo	1
(…)	

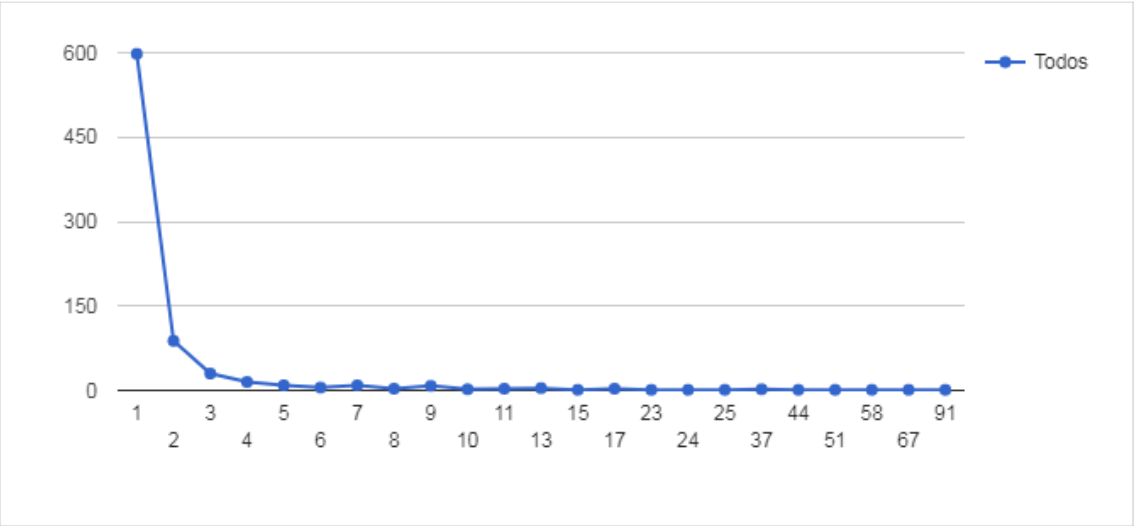
上の表の最大頻度は *de* の 91 語で，最小頻度は 1 語で全部で 598 語になります。上の表を観察すると，最大頻度の de に近い語の頻度は急速に下降し(de, que, y, …)，最小頻度の 1 に近づくと多くの語が 3, 2, 1 の頻度になることがわかります。次の図はその様子を示します。



そこで，各頻度(F: 1, 2, 3, ..., 91)について，その頻度をもつ語数(N)を数えると，次の表とグラフが得られます。

F	N	F^2 * N	F	N	F^2 * N	F	N	F^2 * N	F	N	F^2 * N
1	598	598	7	9	441	15	1	225	44	1	1936
2	88	352	8	3	192	17	3	867	51	1	2601
3	30	270	9	8	648	23	1	529	58	1	3364
4	15	240	10	2	200	24	1	576	67	1	4489
5	9	225	11	3	363	25	1	625	91	1	8281
6	5	180	13	4	676	37	2	2738			

上の F と N の関係をグラフで示すと，ここで N の分布が急激な下降を示していることがわかります。



Zipf(1936: 40-44)は一般に F と N の間に次の関係があることを指摘しました。

$$F^2 * N = C$$

しかし，この Zipf の第二公式も，上の表からも明らかなように，大きな誤差があります。F を 2 乗するために，F の値が多くなるととくに当てはまりが悪くなるようです。Baayen (2001; 8-10)は上の分布表を「頻度スペクトル」(frequency spectrum)と呼び，さまざまな近似式を比較しています。

語彙の頻度分布を観察するとき，すべての語彙が示される頻度の順位による分布だけでなく，頻度が集計された頻度スペクトルの分布も参考になります。

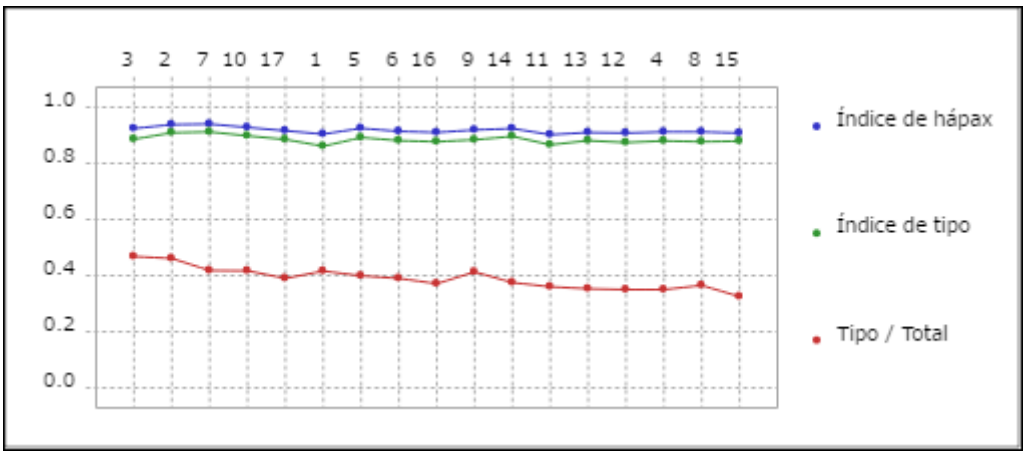
de, que, y などの高頻度語と，次のような頻度が 1 である語(hapax: *abrazaré, absolutamente, absurdo, acabo, acción, acercaba, acobardarás*, ...)の文法的・意味的機能はまったく異なります。その両者の間に位置する語形は頻度がそれぞれ大きな頻度の差異によって数量的な特徴をを示していることが頻度順分布表からも頻度スペクトルの表からもわかります。語の文法的・意味的機能と使用頻度との関係を追究することは言語研究の重要なテーマになります。

■全語数の上昇とタイプ・トークン比の下降

次の表はスペインの作家 Benito Pérez Galdós (1843-1920)の作品 *Trafalgar* (1873)の各章(Capítulo, 1~17)の総語数(トークン数 Total), タイプ数(Tipo), 最大頻度(Frecuencia máxima), 唯一頻度数(Número de frec. única), タイプ・トークン比(Tipo/Total), 相対化タイプ・トークン比(Tipo/Total C.), タイプ指数(Índice de tipo), 唯一頻度指数(Índice de frecuencia única)を示します。各章のデータを語数(Total)で昇順に並べ替えました。

Trafalgar	3	2	7	10	17	1	5	6	16
Total	1300	1402	2000	2183	2298	2503	2591	2723	2777
Tipo	609	646	836	910	896	1040	1033	1063	1030
Máximo	78	65	81	103	116	168	125	144	144
Hápax	472	494	633	663	636	788	768	767	732
Índ. de hápax	0.924	0.938	0.94	0.928	0.916	0.904	0.925	0.914	0.91
Índice de tipo	0.886	0.909	0.912	0.898	0.885	0.861	0.892	0.881	0.877
Tipo / Total	0.468	0.461	0.418	0.417	0.39	0.416	0.399	0.39	0.371

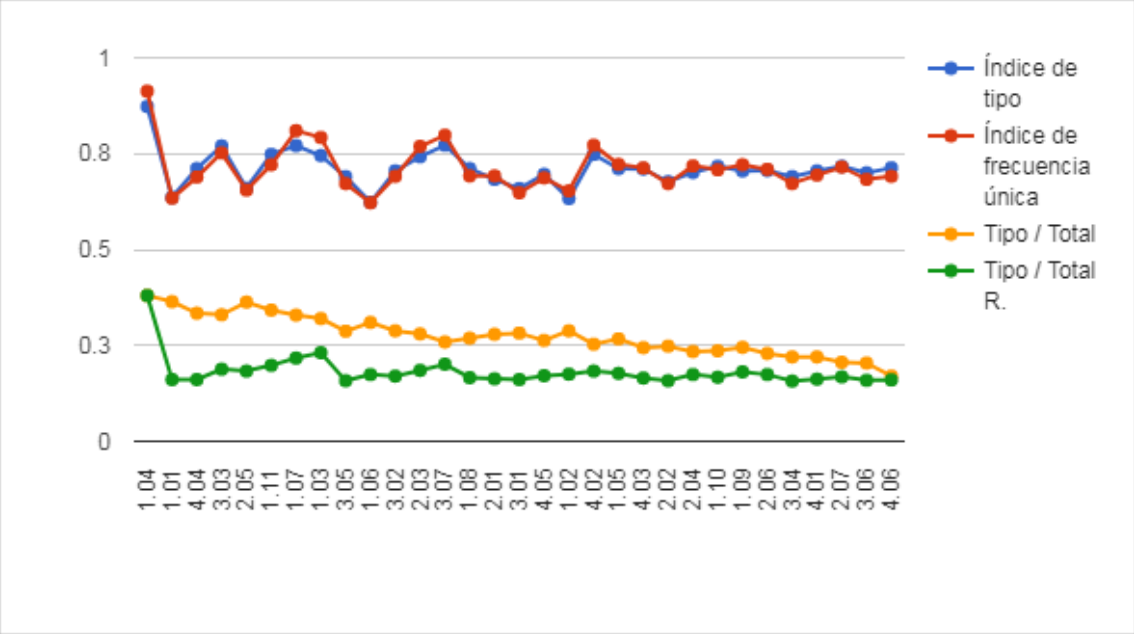
Trafalgar	9	14	11	13	12	4	8	15
Total	3016	3184	3482	3837	4140	4200	4579	4938
Tipo	1245	1195	1255	1355	1450	1470	1671	1603
Máximo	164	139	194	183	209	200	235	220
Hápax	936	844	892	928	1032	1041	1225	1090
Índ. de hápax	0.919	0.924	0.902	0.91	0.908	0.912	0.912	0.908
Índice de tipo	0.884	0.896	0.866	0.881	0.874	0.88	0.877	0.879
Tipo / Total	0.413	0.375	0.36	0.353	0.35	0.35	0.365	0.325



上の表と図を見ると，タイプ・トークン比(Tipo/Total)が全体的に下降していることがわかります。これは各章の総語数が上がるにつれて，異なり語数（トークン数）の相対頻度が下がる，という一般的傾向から説明されます。一方，タイプ指数と唯一頻度指数には総語数による影響がないことがわかります。

■相対化タイプ・トークン比，タイプ指数，唯一頻度指数の収束

次は Benito Pérez Galdós(1843-1920)の長編小説 *Fortunata y Jacinta*(1886)の各章(I.1~11; II.1~7; III.1~7; IV.1~6)の語形頻度の総和を降順に並べ（Total:横軸），各章のタイプ指数(Índice de tipo)，唯一頻度指数(Índice de frecuencia única)，タイプ・トークン比(Tipo/Total)，相対化タイプ・トークン比(Tipo/Total C.)の数値（縦軸）の動きを示す線グラフです。



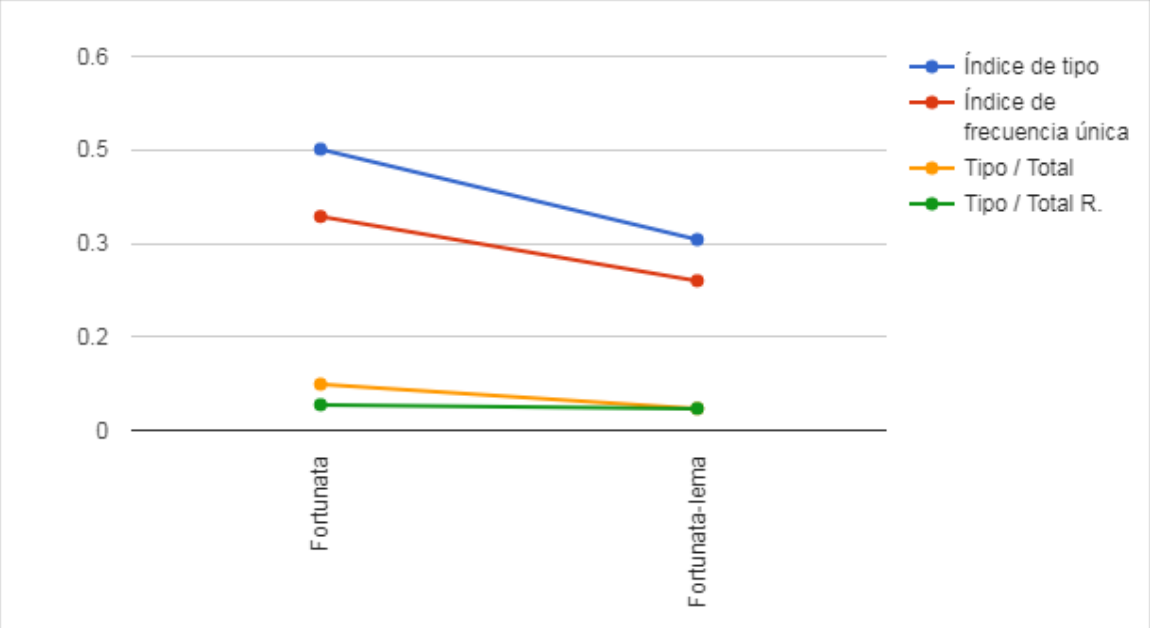
この図を見ると，語形数が比較的少ない左側で(1.04: 語数 3533 ~ 1.05: 12564)，タイプ指数と唯一頻度指数の上下運動が激しく，それが右に進む

につれ次第に安定し，それぞれの値が一定の値に収束していく様子がわかります(4.06: 34284)。具体的には 1.10: 17901 の付近で安定しています。よって，この資料では 18000 語以上の語形を有する章から語彙の多様性を適切に計測できると言えるでしょう。一方，相対化タイプ・トークン比は小さな規模のテキストでも比較的安定しているようですが，これは相対化タイプ・トークン比の感度が小さいからだ，とも言えます。むしろ，感度の強いタイプ指数と唯一頻度指数でテストするほうが精確に収束ポイントを探知できるでしょう。なお，従来のタイプ・トークン比ではテキストの規模の影響を受けているので信頼性が低く，それが減少しても必ずしも語彙の多様性の減少を示していることにはなりません。

■語形と見出し語のタイプ指数と唯一頻度指数

次はスペインの作家 Benito Pérez Galdós (1843-1920)の作品 *Fortunata y Jacinta* (1886)の語形(form)と見出し語形(レンマ：lemma)の各数値を比較したものです。

Diversidad de frecuencia	Fortunata	Fortunata-lema
Total	394606	393797
Tipo	29368	13857
Frecuencia máxima	18247	31362
Número de frec. única	14398	4962
Tipo / Total	0.074	0.035
Tipo / Total C.	0.041	0.035
Índice de tipo	0.451	0.306
Índice de frecuencia única	0.343	0.24



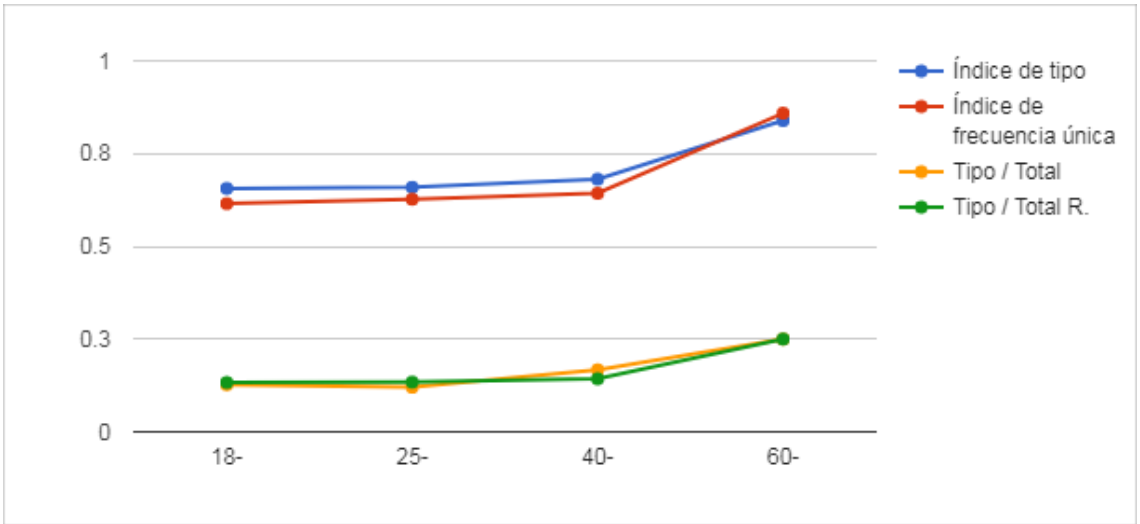
この表と図を見ると，見出し語(Fort.lemma)のタイプ数(Tipo)が小さく，逆にその最大頻度が大きく，そのためにタイプ・トークン比(Tipo/Total)，相対化タイプ・トークン比(Tipo/Total C.)，タイプ指数(Índice de tipo)，唯一頻度指数(Índice de frecuencia única)の数値がすべて小さくなっています。

テキストを分析するとき，分析の目的に応じて語形そのものを対象にすることもありますが，それを見出し語(lemma)に変えてから分析することもあります。見出し語形に変えると，語彙の多様性が低くなります。逆に語形でテキストを分析すると，一般に語彙の多様性が高くなりますが，スペイン語のように屈折形が多い言語では，その多様性がとくに高まります。

■口語スペイン語のタイプ指数と唯一頻度指数

次の表はスペイン語口語資料(C-ORAL-ROM)の資料の話者の年齢(18-, 25-, 40-, 60-)で分類した総語数(トークン数 Total)，タイプ数(Tipo)，最大頻度(Frecuencia máxima)，唯一頻度数(Número de frec. única)，タイプ・トークン比(Tipo/Total)，相対化タイプ・トークン比(Tipo/Total C.)，タイプ指数(Índice de tipo)，唯一頻度指数(Índice de frecuencia única)を示します。

Diversidad de frecuencia	18-	25-	40-	60-
Total	41102	57570	18585	4363
Tipo	5213	6936	3097	1089
Frecuencia máxima	2117	2838	926	209
Número de frec. única	2972	3925	1812	638
Tipo / Total	0.127	0.120	0.167	0.250
Tipo / Total C.	0.133	0.134	0.143	0.250
Índice de tipo	0.656	0.659	0.681	0.839
Índice de frecuencia única	0.615	0.627	0.643	0.859



上のグラフによれば，年齢が上がるにつれて 4 つの指標がすべて上昇しています。タイプ・トークン比が上昇しているのは 40-，60-の年齢層で記録された語数が少ないためだと思われます。語数が多くなると，タイプ・トークン比が下降することは先に述べたとおりです。一方，相対化タイプ・トークン比，タイプ指数，唯一頻度指数は語数に影響されないはずなので，それが上昇しているのは語数によるのではなく，テキスト中のタイプ（異なり語数）と唯一頻度語の割合が上昇している，と判断してよいと思います。

このように相対化タイプ・トークン比，タイプ指数，唯一頻度指数はそれぞれ異なった意味をもつので，私たちは三者を総合して語彙の多様性を判断します。なお，タイプ・トークン比がタイプ指数と一頻度指数よりも一般的に小さな値になりますが，これはタイプ・トークン比の分母が非常に大きな総語数(Total)であるためです。

Cresti, Emanuela / Moneglia Massimo (2005). *C-ORAL-ROM. Integrated Reference Corpora for Spoken Romance Languages*. Paris. John Benjamins.

参考：

Baayen, R. Harald. (2001): *Word frequency distributions*. Kluwer Academic Publishers.

Zipf, George Kingsley. (1936): *The Psycho-bilology of language*. New York. George Routladge & Sons.

____. (1949): *Human Behavior and the Principle of Least Effort*. Cambridge/ Massachusetts: Addison-Wesley.

4.7. 推移

平均，中央値，最頻値などのデータの中心を示す指数と，分散や標準偏差などの変動を示す指数が同じでも，データの成分の並び方（推移）が異なると，データの意味が変わります。ここでは，どのような推移のあり方と，それを数量化する方法を考えます。

4.7.1. 単調性

データの並びに上下の振動のない様子を**単調性指数**(Index of Monotony: IM.)と呼ぶ値によって数量化します。方法は，数値が上昇する値の和(ascendigng: a)と下降する値の和(descending: d）を計算し，両者の対照値を計算します。たとえば，データ

(10, 19, 14, 7, 12)

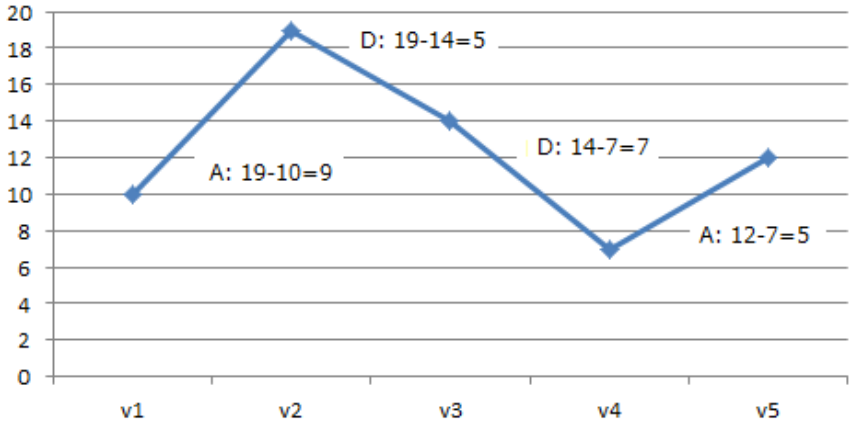
の屈折値(I)は

a: (19-10) + (12-7) = 9 + 5 =14

d: (19-14) + (14-7) = 5 + 7 = 12

IM: a, d の対照値：

IM = (14-12) / (14+12) = 0.077




```
Monotony=function(A){
   a=d=0; for(i in 2:length(A)){
      if(A[i]>A[i-1]) a=a+A[i]-A[i-1] else d=d+A[i-1]-A[i]
   }; (a-d)/(a+d)
} #Monotony [-1, 1] ex: A=c(10, 19, 14, 7, 12); Monotony(A)
```

4.7.2. 単峰性

分布の単峰性(Unimodality)とは，山型の分布が 1 つだけの頂点を持つことを示します。しかし，多少の凸凹があって完全に単峰でなくても，一定の単峰性が存在すると考えます。そこで，データの分布の頂点までの範囲と頂点を超えた範囲に分けて，山型の分布に沿う隣接値の間隔の和(P)と，逆向きの隣接値の間隔の和(N)を計算し，P と N を使って**単峰性**(Unimodality: U)の度合いを次のように定義します。

U = P / (P + N)

たとえば，データ{F.1: 10, 19, 14, 7, 12}では，頂点(=19)に向けて 1 回上昇し(+P: 19-10=9)，頂点を超えてから 2 回下降し(+P: 19-14=5, 14-7=7)，最後に 1 回上昇しています(+N: 12-7=5)。このように，頂点を中心にした山型に沿う行程の距離の和を P 値(9+5+7)とし，それに沿っていない距離の和が N 値(5)とします。よって d1 の**単峰性指数** U(d1)は

U(F.1) = P / (P + N) = (9+5+7)/(9+5+7+5) = 21 / 26 = .808

しかし，**単峰性指数** U のような単純な相対値では，分子と分母の割合が

同じであれば，3/10 も 30/100 も同じ結果になります。また，対照値(N)が小さいと，たとえば，200 / (200+1) = .995 と 300 / (300+1) = .997 のように P の値が大きく異なっても，結果はほとんど変わりません。極端な場合は N=0 のときで，その時は P の値が何であっても，U は 1 になってしまいます。

そこで,エクセル二項分布確率関数 BINOMDIST を使った,次の式(U.Bin)を考えます[68]。

U.Bin = BINOMDIST(P, P+N, 0.5, 1)

この式で事前確率を 0.5 とするのは，単峰性に沿うか，反するか，ということが偶然であるならば，その確率は 0.5 になる，という前提に基づきます。しかし，この式でも N=0 であれば，P がどんな値であっても 1 であることには変わりはありません。すべての場合の確率の和は 1 になるからです。

そこで，累積二項確率を使った期待確率を導入して（→「確率」），次の二項分布**単峰性**(Unimodality by expected binomial probability: U.BinE)を定義します。

U.BinE = BinE(P, P+N, 0.5, 0.99)

この式では，信頼性を示す確率を 99%にして厳しくしました。基準を少し緩めれば 95% (.95)にすることも考えられます。

D	t.1	t.2	t.3	t.4	t.5	P	N	U	U.BinE
F.1	12	21	16	9	14	21	5	.808	0.570
F.2	13	9	12	2	3	14	4	.778	0.480
F.3	2	2	3	14	3	23	0	1.000	0.819
F.4	10	12	10	8	10	6	2	.750	0.293

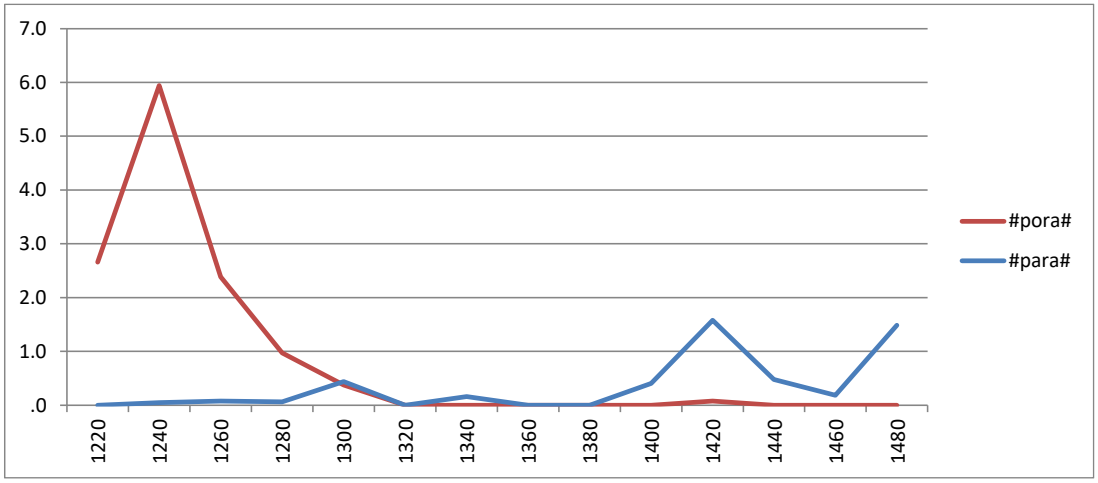
■中世スペイン語の前置詞 pora と para

現代スペイン語の前置詞 *para* (意味は英語 *for*)は中世スペイン語 *pora* (< *por* + *a*) に由来します。年代が確定している公証文書で両者の推移を見ると次のような分布が示されました（千語率）。

Cron.	1220	1240	1260	1280	1300	1320	1340	1360	1380	1400	1420	1440	1460	1480	U.BinE

[68] エクセル関数の 4 番目の引数を 1 にして，累積確率を使います。

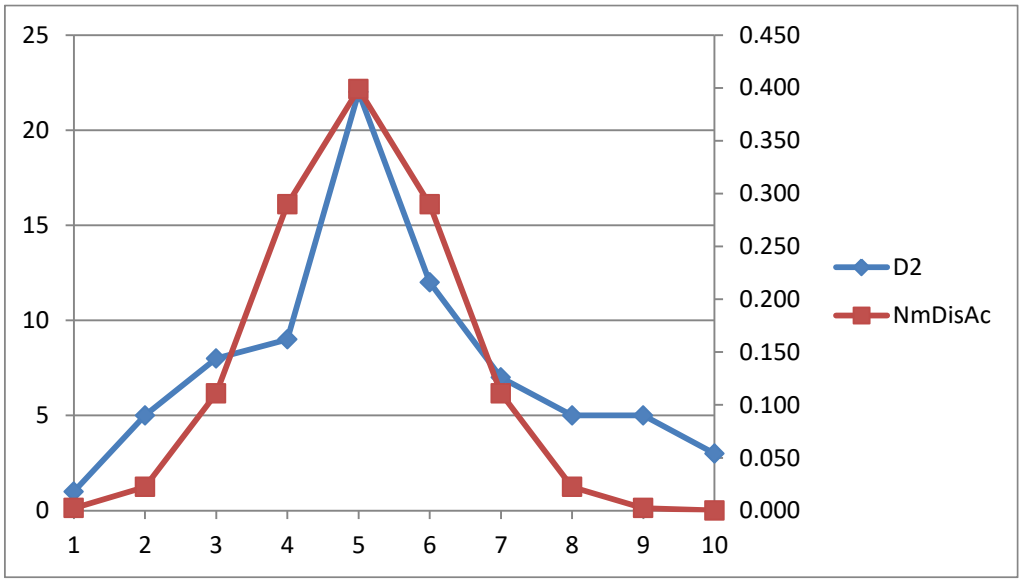
	1220	1240	1260	1280	1300	1320	1340	1360	1380	1400	1420	1440	1460	1480	
pora	2.66	5.94	2.39	.97	.38	.00	.00	.00	.00	.00	.08	.00	.00	.00	.599
para	.00	.05	.08	.06	.44	.00	.16	.00	.00	.41	1.58	.48	.19	1.49	.106



上の表と図を見ると，先行する *pora* が 13 世紀中葉に優勢であり，その後 *para* が少しずつ出現し，15 世紀にピークに達したことがわかります。*pora* の使用が 13 世紀に突出していたため単峰性が.991 に達しています。一方，後続する *para* の比較的低い単峰性(.652)は，その使用に揺れがあったことを示してます。

4.7.3. 正規性

年代順に並べた言語現象の頻度などは，しばしば次の図の青線のような単峰性（頂上が 1 つ）で，最初と最後の頻度が少なく，平均を示す中央で頻度が最大になる傾向があります。これは，言語変化がはじめは少ない頻度で始まり，それが優勢になると一挙に高頻度に達し，衰退すると急激に下降して，最後は小数だけが残り，それもやがて消滅する，という一種の流行のような推移をたどるためです。



この上昇・加工の傾向は，典型的には，確率で見た正規分布に近似しますが，もちろん完全に一致することはありません。そこで，傾向としてどの程度まで正規分布（上図の赤線）に似ているかを示す指標があれば便利です。そのために，次のような実験をして，実測値と正規分布の密度関数との間の相関係数（→後述）を**正規性指数**(Index of Normality: I.Norm)として使うことにします。

I.Norm = Cor(Xn, Sn)

ここで，Cor は 2 つのベクトル間の相関係数を返す関数，Xn は実測値ベクトル，Sn は正規分布密度関数ベクトルです。

プログラムでは，実測値の成分の最大値がある位置を求め，この位置(M)から左右を見て長い方を幅(W)とします。この最大値位置(M)と幅(W)を使って，標準測度(S)の最大値が 4 になるように計算し，それぞれの標準測度から Excel 関数 NormDist を使って正規分布の密度を求め，これをベクトル Sn に代入します。ベクトル Sn は次の表の SS にあたります。

N=10	D	SS	Nm
1	1	-3.200	0.002
2	5	-2.400	0.022
3	8	-1.600	0.111
4	9	-0.800	0.290
5	22	0.000	0.399
6	12	0.800	0.290
7	7	1.600	0.111
8	5	2.400	0.022
9	5	3.200	0.002
10	3	4.000	0.000

ここでデータ(D)の最大値(=22)の位置は 5 になり，これは中央値の位置 5.5 よりやや下にあるため，より長い幅は下側の[6, 10]になります。その最大位置(=10)の標準得点を 4.000 とします。上の表でわかるように，この標準得点に対応する正規分布確率密度はゼロに近似します。正規性指数(I.Norm)の計算は，データ(D)と正規分布確率密度(Nm)の相関係数を使います。ここでは.909 になりました。正規性指数が.8 を超えると実測値の分布が正規分布に近似しているように思われますが，これは目安にすぎません。

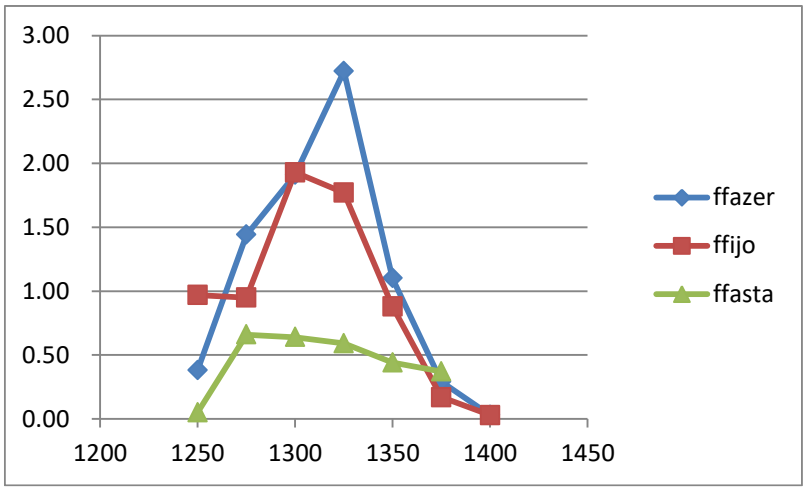
■中世スペイン語の語頭の<ff>-

次は中世スペイン語公証文書に見られる語頭の *ff*-の頻度（千語率）です。

DN	*ffazer*	*ffijo*	*ffasta*
1250	0.38	0.97	0.05
1275	1.44	0.95	0.66
1300	1.91	1.93	0.64
1325	2.72	1.77	0.59
1350	1.10	0.88	0.44
1375	0.29	0.17	0.37
1400	0.03	0.03	

値	*ffazer*	*ffijo*	*ffasta*
正規性指数	.8671	.8737	.3777

このように，*ff*-は 13 世紀中頃から 14 世紀にかけて頻出していますが，その分布は次のグラフを見てわかるように，ラテン語起源の *ffazer*「する」，ffijo「息子」で正規性が高くなっています(FACERE > *ffazer*; FILIU > *ffijo*)。それぞれの正規性指数は.8671, .8737 でした。一方，アラビア語起源の *ffasta*「～まで」の正規性はあまり高くなく(=.3777)，頻度も比較的少ないようでした。

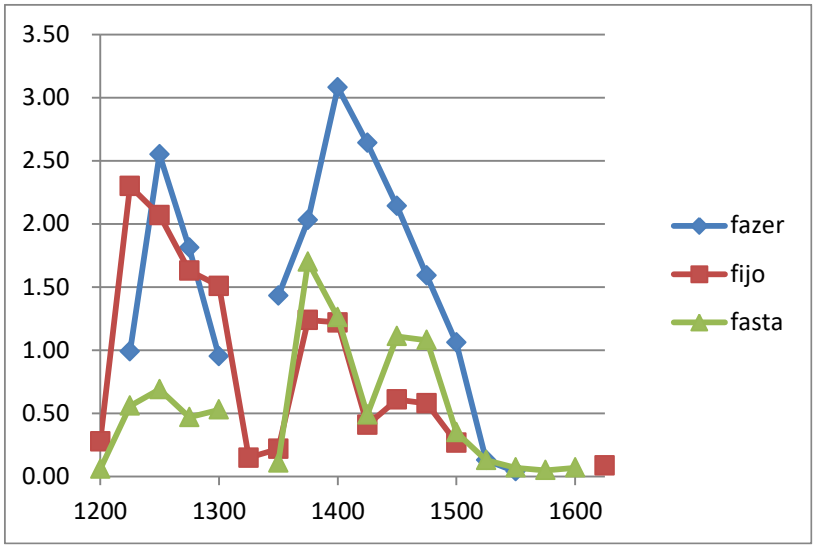


次の表が示すように，これらの語頭 *ff*-の語は語頭 *f*-の語と共存し，*f*-の

語は連綿と長い期間使用されていました[69]。そして，この伝統的な *f*-の語形の正規性はあまり高くありませんでした(.7060, .7083, .5423)。

DN	*fazer*	*fijo*	*fasta*
1200		0.28	0.06
1225	0.99	2.30	0.56
1250	2.55	2.07	0.69
1275	1.81	1.63	0.47
1300	0.95	1.51	0.53
1325		0.15	
1350	1.43	0.22	0.11
1375	2.03	1.24	1.70
1400	3.08	1.22	1.26
1425	2.64	0.41	0.49
1450	2.14	0.61	1.11
1475	1.59	0.58	1.08
1500	1.06	0.27	0.35
1525	0.13		0.13
1550	0.04		0.07
1575			0.05
1600			0.07
1625		0.09	

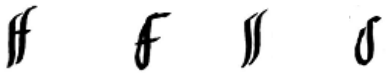
値	fazer	fijo	fasta
正規性指数	.7060	.7083	.5423



このように中世の一時期に生起し消滅した *ff*-の語形の原因として，当時[f-] > [h-] > [ゼロ]というスペイン語特有の音韻変化を意識した過剰訂正

[69] やがて，語頭の f-も消えて，16 世紀に語頭が h-となって現代スペイン語の形 hacer, hijo, hasta が成立しました。

(hypercorrection)であった，という説がありますが，私は，語頭子音の連続が他にもあり(*ss-*. *rr-*. *ll-*)，とくに下図のように当時の「長い *s*, *ss*」(ſ, ſſ)と似ているため,それらによる類推作用が働いたのではないかと思います。*ff-*はスペイン語に限らず，先の音韻変化が起きていない地域にも表れていますが，それは過剰訂正説では説明できません。



(1) ff　(2) f　(3) ſſ　(4) ſ

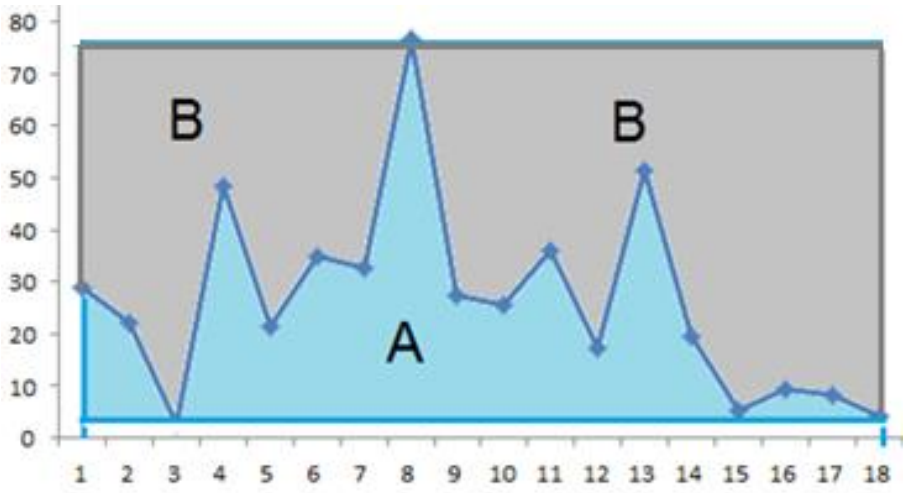
4.7.4. 連続性

一定の数値が続く程度を数量化するために，次のようにして**連続性指数**(Continuity)を設定します。次のデータを例にすると，はじめに，データの数値を結ぶ折れ線と最小値で囲まれる面積(A)を計算します。左端の数値と最大値，右端の数値を求め，データ全体が最大値であったときの面積(T)を計算します(Continuity=.325)。

N	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
D	29	22	3	48	21	35	33	76	27	26	36	17	52	20	5	9	8	4

次のグラフで示すように

T = A + B



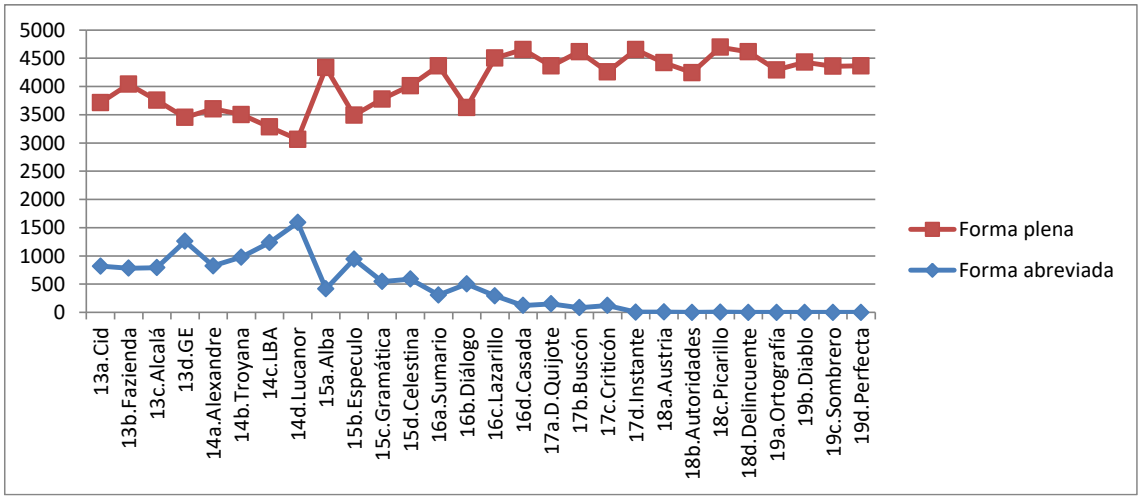
よって，連続性指数(Continuity)は

Continuity = A / (A + B) = A / T

■中世・近代スペイン語の完全形と省略形

次は，中世スペイン語の説話集『ルカノール伯爵』（ドン・フアン・マヌエル作 1330，写本は 15 世紀）の中で頻繁に使われた語 *q<ui>ere*「彼は望む」と *p<ar>a*「～のために」の省略形です。

このような省略形の使用は，次のグラフが示すように，16 世紀以降は少なくなり，現代スペイン語では一部の略語を除けば使用されていません。スペイン語の写本や印刷本の歴史を見るためにサンプリングした資料で，それぞれの連続性指数を計算すると，完全形(Forma plena)が.629，省略形(Forma abreviada)が.278 という数値を示しました。省略形は主に中世と近代初期に限られていたため，定常性指数は低くなります。



4.7.5. 平滑性

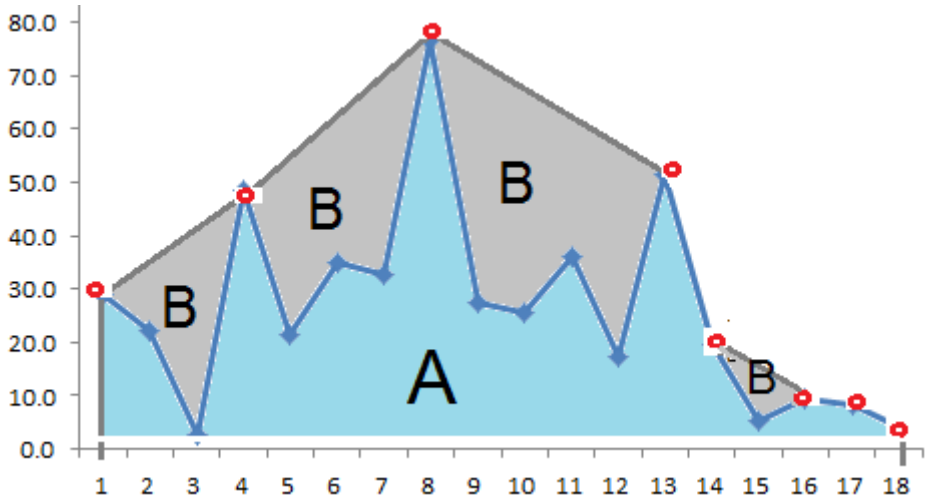
言語現象の歴史的変化や地理的変異の中に，それらが示す数値の増大と減少が平滑に続くことを観察することがあります。ここで扱う**平滑性指数**(Index of Smoothness: IS)は，データの並びの連続性を[0, 1]の範囲で示す指数です。

N	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
D	29	22	3	48	21	35	33	76	27	26	36	17	52	20	5	9	8	4

例として上のデータを使います。次の図の青の線がその推移を示します。このデータから平滑性が最も高い線を作るとすれば，7 つの凸部{29, 48, 76, 52, 9, 8, 4}を結んだ線であると考えます。よって，そのような分布であればデータの推移が完全に平滑している，とみなします。このような線を「平滑線」と呼びます。平滑線を結ぶ点（平滑点：下図の赤丸印）を次のように決めます。両端(=29, 4)と最大値(=76)を平滑点とします。最大値(=76)の

左側の平滑点は，左から右に推移する過程で，左の平滑点の値以上をもつ点を平滑点とします。下図の場合は 4 番目のデータ(=48)が平滑点になります。最大値の左側領域では，これが唯一の平滑点です。最大値の右側領域の平滑点は右端から左に推移しながら平滑点を探します。該当する位置のデータが，その位置の右の平滑点以上であれば，その位置のデータが平滑点とします。下図のデータでは，右から順に{4, 8, 9, 52, 76}が平滑点になります。この平滑線とデータの最小値の線で囲まれた領域の面積を平滑面(T)とします。次の図で

T = A + B



しかし，このデータの実際の線（「観測線」）では，横軸が{2, 3, 5, 6, 7, ...}の位置で数値が下がり，平滑性を損なっています。そこで上図の A の面積が全体の面積(平滑面：T = A+B)の中で占める割合を平滑性指数(IC)とします。よって，平滑性指数(IS)は

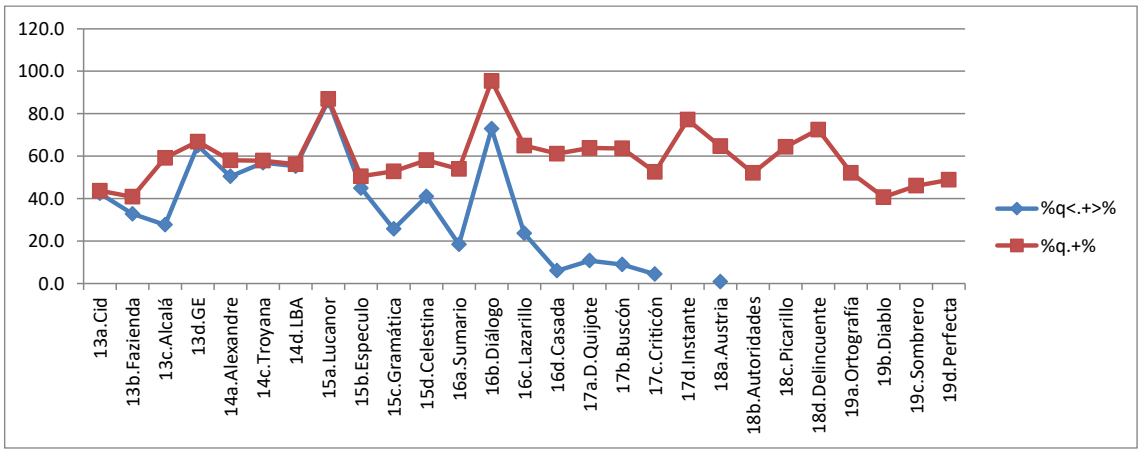
IS = A / T = A / (A + B)

平滑性指数(IS)の最大値(= 1)は B=0 のとき，つまりデータの位置がすべて平滑点となっているときに生じます。一方，最小値(= 0)は，A=0 のとき，つまりデータがないときです。このときは必然的に B=0 となるので，IC = 0 / 0 となり計算不能となります。プログラムでは分子も分母もゼロの場合の分数をゼロで返し，それ以外の場合に分数の計算の結果を返すようにします。

■中世・近代スペイン語の que / qui / qua

資料（13c-19c までの 28 の文書・文学作品）を見ると，中世スペイン語(13-14 世紀)では q の後の母音字は省略されることが多かったのですが，中世から近代の移行期(15-16 世紀)に略形が減少しはじめ，18 世紀になると

完全形にほぼ統一されたことがわかります。



それぞれの語形の平滑性を数量化します。

横軸	平滑性指数(IC):Np
q<.+>	.711: .836^
q.+	.888: .821^

どちらもかなり高い平滑性を示しています。それぞれの語形が歴史的な流れの中で，一部の小さな例外を除けば，自然な推移を辿ってきたことがわかります。文字が省略されるのは，従来，書く労力と紙・羊皮紙を節約するためであったと説明されてきましたが，省略文字の種類は限られ（音節末の<n>, *q* の後の母音，子音の後の<母音+r>または<r+母音>），単語については *nuestro, vuestro, tiempo, tierra* など特定のものに限られていました。また，上で<q>+母音のケースで見たように，その出現には歴史的に平滑性があります。したがって，書き手たちは，労力と資源の節約のために自由に省略形を使っていたのではなく，時代の推移の中で社会的な規範に従っていた，と考えられます。

4.7.6. 定常性

次のような連続するデータ(D)の変化の度合いを知るために，限定化したデータ（LS・範囲：[0, 1]）の隣接する値の差の 2 乗和の平均の根を計算し，これを定常性指数(Constancy)とします（範囲：[0, 1]）。

$$\text{Constancy} = \{ \Sigma [D(i) - (Di + 1)]^2 / (N - 1)\}^{1/2}$$

具体的には

$$\{[(.356- 260)^2 + (.260 - 0)^2 + \ldots + (068 - .014)^2] / (18 - 1)\}^{1/2} = .658$$

N	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
D	29	22	3	48	21	35	33	76	27	26	36	17	52	20	5	9	8	4

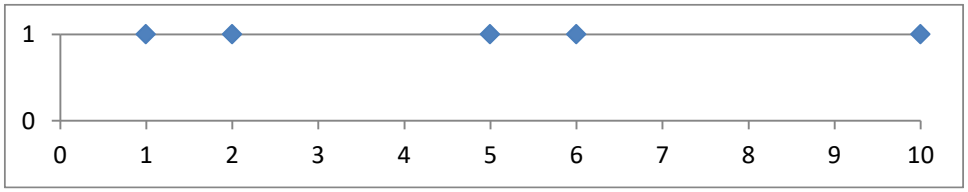
SSc	1	2	3	(…)	17	18
LS	.356	260	0	(…)	.068	.014

X	v1	v2	v3	v4	v5
d1	10	19	14	7	12
d2	11	7	10	0	1
d3	0	0	1	12	1
d4	0	1	2	3	3

横軸	定常性:P/N
d1	.441: .537^
d2	.490: .686^
d3	.350: .230^
d4	.711: .995#

●凝集性と一様性

テキストの中で単語が次図の{1, 2, 5, 6, 10}のように各所に凝集して分布しているか，または一様に遍在しているか，を調べるために「凝集性」(Coagulation)と「一様性」(Uniformity)を使って数量化します。



たとえば上図のように 10 語のテキストの中で, x という単語が 1, 2, 5, 6, 番目の語として使われているとき({1, 2, 3, 6})，それらの間隔(int: interval)はそれぞれ，{1, 1, 3, 1}となります[70]。ここでテキストの最後の位置を f (final)とします。この例では f = 10 です。単語 x が完全に一様に分布していれば，f を含めた間隔の平均は 10 / 5 = 2 となるはずです(int.mean = 2)。そこで実際の間隔(1, 1, 3, 1, 4)と間隔の平均(2)との差を絶対値で示すと，次の表の|int. - int.mean|の列になります(1, 1, 1, 1, 2)。

x	y	int.	int.mean	|int. - int.mean|
x1	1	1	2	|1 - 2| = 1
x2	2	1	2	|1 - 2| = 1
x3	5	3	2	|3 - 2| = 1
x4	6	1	2	|1 - 2| = 1
f	10	4	2	|4 - 2| = 2

上の表の最終列(|int. - int.mean|)，すなわち実際の間隔と間隔の平均の差の絶対値の和 S を求めます。

[70] x1 の間隔は 1, x2 の間隔は 2 - 1 = 1，x3 の間隔は 5 - 2 = 3，x4 の間隔は 5 - 2 = 3.

$$S = \Sigma |int. - int.mean|$$

この S が最大になるのは，次表の y: {1, 2, 3, 4, 10}のように最終の f だけが離れた値をとるときです(S の最大値：S.max.)[71]。

x	y	int.	|int. - int.mean|
x1	1	1	1
x2	2	1	1
x3	3	1	1
x4	4	1	1
f	10	6	4

S.max. = 1 + 1 + 1 + 1 + 4 = 8

S の最大値 S.max を求める式を一般化します。上のそれぞれの表を見ると，f の間隔だけが 6 となり，他のすべて(x1 - x4)の間隔は 1 となっています。f の間隔は f=10 から n - 1 = 4 を引いた数なので，int = m - (n - 1) = 10 - (5 - 1) = 6 になります。よって，この int = m - (n-1)から間隔の平均 m / n を引いた値の絶対値が，間隔と間隔平均の差 S1 になります。

$$\begin{aligned} S1 &= |f - (n - 1) - f / n)| \\ &= |n (f - n + 1) - f / n| \\ &= |(mn - nn + n - f) / n| \end{aligned}$$

[71] これは，x1 だけが離れた値であっても同じです(1 + 4 + 1 + 1 + 1 = 8)。

x	y	asc.	|asc. - asc.mean|
x1	1	1	1
x2	7	6	4
x3	8	1	1
x4	9	1	1
f	10	1	1

また途中の x3 の間隔だけが異なっていても，S は最大になります。

x	y	asc.	|asc. - asc.mean|
x1	1	1	1
x2	2	1	1
x3	8	6	4
x4	9	1	1
f	10	1	1

これらのケースは，どれも 1 か所のデータが極端な値をとっているために凝集度は最大になります。

$$= |(mn - f - nn + n) / n|$$
$$= |[f (n - 1) - n (n - 1)] / n|$$
$$= |(n - 1) (f - n) / n|$$
$$= (n - 1) (f - n) / n \quad \leftarrow \ f \geq n, n \geq 1$$

次に x1, x2, x3, x4 の間隔を見ます。f 以外のすべての語は連続しているために，その間隔はすべて 1 となります。間隔の平均は f / n で求められので，その差は|1 - f / n|です。その個数は x の全語数 n から 1 を引いた数(n - 1 = 5 - 1 = 4)です。よって「x5 以外のすべての語」の間隔の平均と実際の間隔の差の絶対値の和 S2 は[72]

$$S2 = (n - 1) |(1 - f / n)|$$
$$= (n - 1) |(n - f) / n|$$
$$= (n - 1) (f - n) / n \quad \leftarrow \ f \geq n, n \geq 1$$

よって，間隔の平均と実際の間隔の差の絶対値の和 S の最大 S.max.は

$$S.max. = S1 + S2 = 2 (n - 1) (f - n) / n$$

「凝集性」(Coagulation: C)と「一様性」(Uniformity: U)を次のように定義します[73]。

$$C = S / S.max, 0 (S = 0) \leq C \leq 1 (S = S.max)$$
$$U = 1 - C = 1 - S / S.max, 0 (S = S.max) \leq U \leq 1 (S = 0)$$

先の例{1, 2, 5, 6, 10}の S は 1 + 1 + 1 + 1 + 2 = 6, S.max は 2 (5 - 1) (10 - 5) / 5 = 8, よって C = 6 / 8 = 0.750, U = 1 - 0.750 = 0.250 です。

言語データ分析ではテキスト中の語などの言語形式の頻度や平均値だけでなく，その分布状態も考察に含めるべきです。分布状態を示す指標の一つとして凝集性と一様性が役立ちます。

4.8. 区別

4.8.1. 弁別度

たとえばある言語の古文献に<i>と<j>という文字が用いられ，どちらも同じ条件で/i/という音韻を示していたとします。それぞれの頻度(F)が F(<i>)=32 と F(<j>)=2 の間のように大きな差があれば，ほとんどのケース

[72] よって x1 ~ x4 の S1 と x5 の S2 は同じ値になります(1 + 1 + 1 + 1 = 4)。
[73] n = 1 または n = m のとき分母がゼロになるので計算できません。個数 n = 1 のときや，データが{1, 2, 3, 4, 5}のように完全に連続しているときは n = m (= 5)となります。これは完全に凝集している状態なので C = 1, D = 0 とします。

で<i>が使われたことになるので，その弁別する力は強かったと判断できます。一方，それが 32 と 28 のように僅差であれば，<i> ~ <j>はほとんど「自由変異」(free variation)であった，つまり両者は弁別されていなかったと考えられます。そこで，「弁別度」(Distinctive Grade: DG)を次のように定義します。

DG(<i>, <j>) = [F(<i>) – F(<j>)] / F(<i>)

ここで F(<i>)は<i>の頻度を示し，F(<j>)は<j>の頻度を示します。F(<i>)と F(<j>)が等しいと弁別度はゼロになり，F(<j>)がゼロになると<i>の弁別度は 1 になります。

この弁別度はバリアントが 2 つの場合について計算しました。さらに<i>,<j>だけでなく<y>が現れる文献では，次のように計算します。

DG(<i> : <j>, <y>) = {F(<i>) – [F(<j>) + F(<y>)] } / F(<i>)

一般に Fn = F(1, 2, …, n)の中の F(1)の弁別度 DG(1)は

DG(1) = {F(1) – [F(2) + F(3) + … F(n)] } / F(1)
= {F(1) – [Sum(Fn) – F(1)] } / F(1)
= [2*F(1) – Sum(Fn)] / F(1)
= 2 – Sum(Fn) / F(1)

F(1)を F(1, 2, …, n)の最大値(Max(Fn))とすれば

DG(Max(Fn)) = 2 – Sum(Fn) / Max(Fn)

となります。この弁別度は，成分の最大値 F(1) = Max(Fn)が他の成分の和([F(2) + F(3) + … F(n)])よりも小さいとマイナスになり，その理論的最小値が一定になりません。

4.8.2. 対立度

先の弁別度の分母を次のように対照型にして，新たに「対立度」(Oppositive Grade: OG)を設定します。

OG(<i>, <j>) = [F(<i>) – F(<j>)] / [F(<i>) + F(<j>)]

= [F(<i>) – F(<j>)] / Sum(Fn)

一般に F(1, 2, …, n)の中の F(1)の対立度(Og(1))は

OG(1) = {F(1) – [F(2) + F(3) + … F(n)] } / {F(1) + [F(2) + F(3) + … F(n)] }
= {F(1) – [Sum(Fn) – F(1)] } / Sum(Fn)

= [2 *F(1) – Sum(Fn)] / Sum(Fn)
= 2*F(1) / Sum(Fn) – 1

F(1)を F(1, 2, …, n)の最大値 Max(Fn)とすれば

OG(Max(Fn)) = 2 Max(Fn) / Sum(Fn) – 1

となります。

成分の最大値が他の成分の和よりも大きいときには弁別度を使用し，そうでないときは対立度を使用するとよいでしょう[74]。

■15 - 17 世紀のスペイン語の<u>と<v>

従来の研究では 15, 16, 17 世紀のスペインで発刊された書籍では<*u*>と<*v*>が弁別せずに使われていた，と説明されています。次は 15-17 世紀にスペインで発刊された 6 冊の本（冒頭から 2 万字に限る）について，文字 u と v の頻度と弁別度を計算したものです。

全体	1.Nb	2.Rj	3.Lz	4.Cv	5.Qv	6.Gc	Total
<u>	949	820	1.040	1250	1051	849	5959
<v>	165	139	191	194	209	402	1300
弁別度	0.826	0.830	0.816	0.845	0.801	0.527	0.782

このように全体を見るとたしかに比較的弁別度が低いことがわかります。ところが，文字の現れる位置について，それぞれの弁別度を計算してみると，次のようになりました[75]。

位置	1.Nb	2.Rj	3.Lz	4.Cv	5.Qv	6.Gc	Total
#_V	0.974	1.000	0.942	1.000	1.000	1.000	0.996
#_C	1.000	1.000	0.985	1.000	1.000	1.000	0.896
V_V	0.625	1.000	1.000	1.000	1.000	0.939	0.757
V_C	0.971	0.429	0.917	1.000	1.000	0.978	0.929
C_V	0.967	1.000	0.998	1.000	0.998	0.901	0.980
C_C	0.995	1.000	1.000	1.000	0.996	0.997	0.998

たしかに，複数の本を取り上げれば全体的に<*u*>-<*v*>の弁別がないように見えます。しかし，それぞれの本の中では，下線のような弁別度が低い本もありますが，それを除けば文字の位置によって比較的統一されていた

[74] N が大きいと対立度もマイナスになるのが普通です。弁別度も対立度も N が[2, 4]ぐらいの範囲のデータで利用すべきです。
[75] ここでは最大値を示すバリアントの弁別値を計算したので，すべてプラスの値になりました。v が使われるほうがふつうの位置では，u の弁別値がマイナスになります。

ことがわかります。一般にデータの分布にさまざまな要因が隠れているにもかかわらず，それを見ないで全体的な把握をすると，弁別や対立の真の姿を見失うことがあります。

■中世・近代スペイン語の前置詞

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」という意味：英語 *for*)を示す相対頻度と対称頻度の比較です。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，対称頻度を使うと，両者を同時に対称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

対称頻度: Pora - Para

データ行列全体の総和を 1 として，それぞれのセルの値を相対化する方法を「標準化得点」(Normalized Score: NS)と呼びます。

4.9. 有意性

4.9.1. カイ二乗有意性

平均・期待値からの全体的な逸脱の程度を近似的に測るために Pearson のカイ二乗値が使われます（→「確率」）。たとえば正しいサイコロを 60 回振って「1」から「6」の目がどれも平均（期待値 = 60*1/6 = 10）に近くなるような分布が生じる確率は高くなるはずです。一方，その平均から外れた値が多いときには，そのような確率は低くなります。

平均(av)からの外れた値として次に定義される「カイ二乗値」(chi square: chi-n)を使います[76]。

chi-n = Σ(i) [x(i) - av]^2 / av

Excel 関数 CHIDIST(chi, df)はカイ二乗値(chi-n)と自由度(df-n)を引数として[77]，それに対応するカイ二乗確率分布の右側の累積確率を返します[78]。よって，サイコロを投げたときのような偶然でも起こるデータの分布であれば，この累積確率が高くなります。逆に，偶然ではありえないような「有意な」(significant)な分布であれば，当然この確率はきわめて低くなります。このカイ二乗確率累積確率が小さいほど有意性が高いことになるので，有意性を示す値として，カイ二乗確率累積確率（右側確率）の 1 の補数（左側確率）を「カイ二乗有意性」(Chi-square significance: ChiSig-n)と呼ぶことにします。その範囲は[0, 1]です。

ChiSig-n = 1 - CHIDIST(chi-n, df-n)

ここでカイ二乗値(chi-n)は

chi-n = Σ(i) [X(i) - av]^2 / av　←av: X の平均

自由度(df-n)は

---

[76] 後述の 2 次元データのカイ二乗値と区別するために 1 次元データのカイ二乗値 chi-n とします。

[77] 行または列のように，1 次元データのカイ二乗値を計算するときは，データ数から 1 を引いた数が自由度になります。これは和を一定にしたときに自由に起こり得る値の数は個数-1 になるからです。たとえば 5 個のデータのうち自由なデータは 4 個までで,残る 1 個のデータは自由ではなく,[全体の和] - [4 個のデータの和]になります。

[78] たとえば自由度が 1 の場合の 5%点のカイ二乗値は 3.841 なので，CHIDIST(3.841, 1)は.05 を返します。

df-n = N - 1　　←N: データの個数

下左表(H)がデータ行列(S1)のそれぞれの行の，和・平均値・標準偏差・相対標準偏差・カイ二乗有意性を示しています。

S1	v1	v2	v3	v4	v5
h1	10	19	14	7	12
h2	11	7	10	0	1
h3	0	0	1	12	1
h4	0	1	2	3	3

H	和	平均値	標準偏差	相対標準偏差	カイ二乗有意性
h1	62	12.400	4.030	.162	.838
h2	29	5.800	4.534	.391	.999
h3	14	2.800	4.622	.825	AV < 5.0
h4	9	1.800	1.166	.324	AV < 5.0

上の「カイ二乗有意性」を見ると，とくに h4 {0, 1, 2, 3, 3}の有意性が低く(.563 = 56.3%), 続いて h1 (.838 = 83.8%)もあまり有意だとは言えません[79]。一方，h2 や h3 の有意性はきわめて高く，偶然でおこるような分布ではないと言えるでしょう。

次は縦軸の分析結果です

縦軸	v1	v2	v3	v4	v5
和	21.000	27.000	27.000	22.000	17.000
平均値	5.250	6.750	6.750	5.500	4.250
標準偏差	5.262	7.562	5.449	4.500	4.548
相対標準偏差	.579	.647	.466	.472	.618
カイ二乗有意性	1.000	1.000	.999	.998	AV < 5.0

全体のカイ二乗有意性を計算するには, それぞれのセルの期待値得点 e(i, j)（→「得点」）とセルの値 x(i, j)との差（ズレ）の程度を示す値として，次のように定義されるカイ二乗値 chi-np を使います。

chi-np = Σ (i, j) [x(i, j) - e(i, j)]^2 / e(i, j)

このカイ二乗値 chi-np と，次に定義される自由度(df-np)を使って，Excel 関数 CHIDIST(chi-np, df-np)によって右側確率を求めます。

df-np = (n - 1) * (p - 1) ←n:行数 ; p:列数

上の右式の理由は，次のデータで行和と列和を固定すれば，自由になるのはたとえば次の 2 つの例のように，それぞれの数を引いた値の積(3 * 4 = 12)になるからです。

S1	v1	v2	v3	v4	v5
h1	10	19	14	7	12
h2	11	7	10	0	1

S1	v1	v2	v3	v4	v5
h1	10	19	14	7	12
h2	11	7	10	0	1

[79] 有意性を判断するときには 95%や 99%を境界値とするとよいでしょう。

h3	0	0	1	12	1
h4	0	1	2	3	3

h3	0	0	1	12	1
h4	0	1	2	3	3

カイ二乗有意性は CHIDIST(chi-np, df-np)によって求めた数値（右側確率）の 1 の補数（左側確率）です。

ChiSig = 1 - CHIDIST(chi-np, df-np)

全体	値
和	114.000
平均値	5.700
標準偏差	5.658
相対標準偏差	.228
カイ二乗有意性	1.000

この例では全体のカイ二乗有意性が 1 (100%)になりました。このように全体で見れば有意なデータでも，それぞれの行や列がそれほど有意でないこともあります。さらに，平均が 5 を下回ることでカイ二乗で有意性を計算することが不適切なケースも見られます[80]。

4.9.2. 精密有意性

カイ二乗を使った近似的な検定の代わりに Fisher の精密検定が使われることがあります（→「検定」）。有意性を正確に測るときにも同様に Fisher の「精密確率」(exact probability: ep)を使用します。精密確率を求めるためには次の「多項分布」(multinomial distribution)の確率(mdp)を使います（→後述「多項分布個別確率」）。

$$mdp(x_1, x_2, \ldots, x_k) = s! / (x_1! * x_2! *, \ldots, * x_k!) * (1 / k)^s$$
$$\leftarrow \quad s = x_1 + x_2 + \ldots + x_k$$

精密確率(ep)は，和(n)を揃えて可能なすべての整数の組み合わせについて，mdp を計算し，それがデータの mdp よりも小さな値を足し合わせた累積確率です。たとえば，h4 = (0, 1, 2, 3, 3)の mdp(0, 1, 2, 3, 3) = .00258 より小さな mdp を和が同じく 9 (=0+1+2+3+3)になるあらゆる組み合わせで探し，これらを合計します。それらを足し上げていくと.497 (49.7%)という多項分布累積確率になります（→後述「多項分布累積確率」）。この累積確率の 1 の補数を「精密有意性」(Exact Significance: ExSig)と呼びます。

ExSig = 1 - 多項分布累積確率

[80] それぞれのセルの値の有意性については「得点」で扱います。

下左表が入力データ(S1)の 4 つの行についてカイ二乗有意性と精密有意性を計算した結果です。

S1	v1	v2	v3	v4	v5
h1	10	19	14	7	12
h2	11	7	10	0	1
h3	0	0	1	12	1
h4	0	1	2	3	3

横軸	和	カイ二乗有意性	精密有意性
h1	62	.838	.828
h2	29	.999	1.000
h3	14	1.000	1.000
h4	9	.563	.497

●多項分布個別確率

二項分布個別確率(Bi)は次の式で定義されます（→「確率」）。

Bi(x, n, p) = nCx * p^x * (1-p)^(n-x)

← x:生起回数；n:試行回数；p:期待確率

上の式を適用するには Excel 関数=BINOMDIST(x, n, p, 0)を使用すると便利です。累積確率を求めるときは Excel 関数=BINOMDIST(x, n, p, 1)を使用します。

ここで p1 = p, p2 = 1 - p1 とすると，二項分布個別確率は 2 つの期待確率 p1, p2 についてそれぞれの生起回数が x1 回と x2 回であるような確率を求めています。

Bi(x1, x2, p1, p2) = nCx1 * p1^x1 * p2^x2　　←n = x1 + x2

次の組み合わせの公式

nCx = n! / [x! * (n - x)!]

に従って nCx1 を整理すると

nCx1　= n! / [x1! * (n - x1)!]
　　= n! / (x1! * x2!)　　← n = x1 + x2

よって二項分布個別確率(Bi)は

Bi(x1, x2, p1, p2) = n! / (x1! * x2!) * p1^x1 * p2^x2

たとえば, A, B という文字が書かれたカードがそれぞれ 2 枚, 1 枚で全部で 3 枚あるとします。カードの構成は

{A, A, B}

これらの 3 枚のカードを箱に入れて，無作為に 1 枚ずつ取り出しカードの

文字を記録して戻します。このとき A の期待確率は 2/3 = .667 (66.7%)です。この作業（試行）を 10 回して，A が 6 回，B が 4 回生起する個別確率は BINOMDIST(6, 10, 2/3, 0) = .228 (22.8%)であり，その累積確率は BINOMDIST(6, 10, 2/3, 1) =.441 (44.1%)になります。

これを上の式 Bi(x1, x2, p1, p2) = n! / (x1! * x2!) * p1^x1 * p2^x2 に当てはめて計算すると，当然 BINOMDIST(6, 10, 2/3, 0)と同じ結果になります。

Bi(6, 4, 2/3, 1/3) = 10! / (6! * 4!) * (2/3)^6 * (1/3)^4 = .228

次に期待確率が p1, p2, p3 という 3 つの事象について考えます。たとえば，A, B, C という文字が書かれたカードがそれぞれ 3 枚, 2 枚, 1 枚で全部で 6 枚あるとします。カードの構成は

{A, A, A, B, B, C}

このとき，A の期待確率は 3/6, B の期待確率は 2/6, C の期待確率は 1/6 です。ここで先と同じ試行を 10 回行って，そのうち A, B, C がそれぞれ生起する回数を 5 回，3 回，2 回であるような三項分布個別確率(Tri)を求めるには次の式を用います。

Tri(x1, x2, x3, p1, p2, p3)
= nCx1 * n-x1Cx2 * p1^x1 * p2^x2 * p2^x3
←n = x1 + x2 + x3

よって，この場合は次の計算をします。

Tri(5, 3, 2, 3/6, 2/6, 1/6) = 10C5 * 5C3 (3/6)^5 * (2/6)^3 * (1/6)^2

最初の組み合わせ 10C5 は 10 回の試行の中で A を 5 回引くときの場合の数です。これは，{t1, t2, ..., t10}というそれぞれの試行のうち，{t1, t2, t3, t4, t5}や{t1, t2, t3, t4, t6}など 5 回の試行を選択することと同じなので，互いに異なる 10 個のもの{t1, t2, ..., t10}のから 5 個を取り出す場合の数 10C5 になります。そして，次の組み合わせ 5C3 は残る 5 回の試行の中で B を 3 回引くときの場合の数です。最後に残る 2 回の試行の中から C を 2 個引くときの場合の数は 1 に決まっているので無視します。

これに続く積算* p1^x1 * p2^x2 * p2^x3 の意味は二項分布確率と同じです（→「確率」）。

さて，次の組み合わせの公式

nCx = n! / [x! * (n - x)!]

に従って組み合わせの積算 nCx1 * n-x1Cx2 を整理しましょう。

nCx1 * n-x1Cx2 = n! / [x1! * (n - x1)!] * (n - x1)! / [x2! * (n - x1 - x2)!]

= n! / [x1! * x2! * (n - x1 - x2)!]　　←(n - x1)!が共通
= n! / (x1! * x2! * x3!)　　←n = x1 + x2 + x3

よって三項分布個別確率(Tri)は

Tri(x1, x2, x3, p1, p2, p3) = n! / (x1! * x2! * x3!) * p1^x1 * p2^x2 * p2^x3
←n = x1 + x2 + x3

Excel には三項分布個別確率を返す関数はないので，上の式に具体的な生起数(x1, x2, x3)と期待確率(p1, p2, p3)を代入して計算します。

Tri(5, 3, 2, 3/6, 2/6, 1/6) = 10! / (5! * 3!* 2!) * (3/6)^5 * (2/6)^3 * (1/6)^2
= .081

以上で二項分布と三項分布のそれぞれの個別確率を見てきたので，四項分布，五項分布，…も同様にして導きます。よって一般化した「多項分布個別確率」(Multinomial distribution probability: Mu)の式は

Mu(x[i], p[i]) = n! / (x1! * x2! *, …, *xk) * p1^x1 * p2^x2 *, …, * pk^xk

ここで k=2 ならば二項分布個別確率になり，k=3 ならば三項分布個別確率になります[81]。

なお，精密有意性の計算では，それぞれのセル(x1, x2, …, xk)の確率をすべて 1 / セルの数(k)とします。このような等確率の多項分布(Mu')の式は

Mu'(x[i]) = n! / (x1! * x2! *, …, *xk) * (1/k)^x1 * (1/k)^x2 *, …, * (1/k)^xk
= n! / (x1! * x2! *, …, *xk) * (1/k)^(x1 + x2 + …, + xk)
= n! / (x1! * x2! *, …, *xk) * (1/k)^n

理論上の数式ではこのように簡単になりますが，実際のプログラムでは階乗の積算があるので巨大数となりオーバーフローを起こしやすく，さらに確率の n 乗があるのでこれは逆に微小でゼロに近似してしまいます。むしろ整理する前の

Mu'(x[i]) = n! / (x1! * x2! *, …, *xk) * (1/k)^x1 * (1/k)^x2 *, …, * (1/k)^xk

の式を使い，i = 1, 2, …, k までを繰り返すべきです。階乗の積算のオーバーフローを防ぐために，一度対数に変換して，積算を足し算に，乗算を掛け算にして計算した値を最後に指数変換します（→後述「多項分布累積確率」）。

[81] 以上は一石(2004:78-81)を参照しました（→「参考文献」）。

●多項分布累積確率

精密有意性の指標として用いる多項分布累積確率は，すべての可能なケースで，入力されたデータ行について計算した多項分布個別確率よりも小さな確率を探索し，それらを足し上げた累積確率です。膨大な試行回数と計算量になるので，プログラムを使って説明します[82]。

次が，精密有意性の関数です。引数 Zn は n 行 1 列のベクトル（配列）です。

```
Function exactSignificance(Xn) '精密有意性(exact significance)
   Dim K%, SM, FL, i%, MDP#, ES
   K = nR(Xn):   SM = smA(Xn): ReDim FL(SM) '行数：和：階乗対数
   For i = 1 To SM: FL(i) = FL(i - 1) + Log(i): Next
   MDP = multinomial(Xn, FL, K, SM) '多項分布個別確率
   If K = 2 Then ES = 1 - MND2(Xn, FL, K, SM, MDP) '2 項分布累積確率
   If K = 3 Then ES = 1 - MND3(Xn, FL, K, SM, MDP) '3 項分布累積確率
   If K = 4 Then ES = 1 - MND4(Xn, FL, K, SM, MDP) '4 項分布累積確率
   If K = 5 Then ES = 1 - MND5(Xn, FL, K, SM, MDP) '5 項分布累積確率
   If K = 6 Then ES = 1 - MND6(Xn, FL, K, SM, MDP) '6 項分布累積確率
   If K > 6 Then ES = "K > 6"
   exactSignificance = ES
End Function
```

↑はじめに,階乗の対数変換表 FL を用意し,多項分布個別確率を計算し,次にセル数に応じて，2~6 項分布累積確率を計算します。

```
Function multinomial(Xn, FL, K, SM) '多項分布個別確率
   Dim i&, MND: MND = FL(SM) '初期値（多項分布確率の分子）
   For i = 1 To K
      MND = MND - FL(Xn(i, 1)) + Log((1 / K) ^ Xn(i, 1)) '-分母+確率^頻度
   Next
   multinomial = Exp(MND)
End Function
```

↑次式を Log で対数変換して足し上げ，最後に Exp で指数変換します。

Mu'(x[i]) = n! / (x1! * x2! *, …, *xk) * (1/k)^x1 * (1/k)^x2 *, …, * (1/k)^xk

```
Function MND2(Xn, FL, K, SM, MDP) '2 項分布累積確率
      Dim i%, Wn, W: ReDim Wn(K, 1)
```

[82] 群馬大学・青木繁伸氏の次のサイトを参照しました(2016/7/12)。
http://aoki2.si.gunma-u.ac.jp/lecture/Cross/Fisher.html

```
    For i = 0 To SM
        Wn(1, 1) = i: Wn(2, 1) = SM - i
        W = multinomial(Wn, FL, K, SM)
        If W <= MDP Then MND2 = MND2 + W
    Next
End Function
```

↑データの個別確率より小さな確率を足し上げて累積します。

```
Function MND3(Xn, FL, K, SM, MDP) '3 項分布累積確率(和)
    Dim i1%, i2%, Wn, W: ReDim Wn(K, 1)
    For i1 = 0 To SM: For i2 = 0 To SM - i1
        Wn(1, 1) = i1: Wn(2, 1) = i2: Wn(3, 1) = SM - i1 - i2
        W = multinomial(Wn, FL, K, SM)
        If W <= MDP Then MND3 = MND3 + W
    Next
    Call PROC_SHOW(i1, SM) '●経過表示
    Next
End Function
```

↑3 項分布なので，最初の 2 項を変化させます。

●カイ二乗有意性と精密有意性の比較実験

下左表(S1)の h1 行(10, 10, 10, 10, 10)の v1 列を次第に増加させながら，横軸のカイ二乗有意性と精密有意性がどのように変化するのかを実験した結果です。

S1	v1	v2	v3	v4	v5
h1	10	10	10	10	10
h2	12	10	10	10	10
h3	14	10	10	10	10
h4	16	10	10	10	10
h5	18	10	10	10	10
h6	20	10	10	10	10
h7	22	10	10	10	10
h8	24	10	10	10	10
h9	26	10	10	10	10
h10	28	10	10	10	10
h11	30	10	10	10	10
h12	32	10	10	10	10

横軸	和	カイ二乗有意性	精密有意性
h1	50	.0000	.0000
h2	52	.0107	.0065
h3	54	.1195	.1016
h4	56	.3681	.3016
h5	58	.6471	.5781
h6	60	.8454	.7846
h7	62	.9458	.9070
h8	64	.9844	.9643
h9	66	.9963	.9883
h10	68	.9992	.9966
h11	70	.9999	.9991
h12	72	1.0000	.9998

h13	34	10	10	10	10

h13	74	1.0000	1.0000

このように和が大きくなるにつれてカイ二乗有意性と精密有意性は近似していきます。そこで，和が十分に大きい場合は精密有意性の代わりにカイ二乗有意性を使うことが考えられます。そのときのカイ二乗有意性を代用できる「十分な和」の目安はデータのバラツキ（標準偏差）や個数によります。

カイ二乗有意性を計算するときは平均値・期待値が 5 以上でなければ有効でないので，上のようにセルが 5 個ならば和が 5 * 5 = 25 以上でなければなりません。そこで和が小さいときは精密有意性を用いるほうがよいのですが，この精密有意性は非常に多くの計算をするので実際には 7 個以上のセルについての計算は現実的ではありません(和が小さければ可能です)。よって，有意性の計算ではできるかぎり精密有意性を使い，セル数が 8 個以上の場合や，データの最大値が 500 を超えるときや[83]，多項分布累積確率の計算量が 10,000,000 回を超えるようなデータのときにはカイ二乗有意性を使うとよいでしょう。

なお，精密有意性は計算中に二項分布確率を使いますので，入力データは非負の整数でなければなりませんが[84]，カイ二乗有意性の計算では非負の整数だけでなく非負の小数も扱うことができます。

[83] 多項分布個別確率の指数の計算部分でオーバーフローします。
[84] プログラムではデータ値の小数点以下を切り捨てて計算します。