NUMEROS-es.docx

ver. 2017.5.6

Análisis de datos cuantitativos para estudios lingüísticos

Hiroto Ueda (Universidad de Tokio)

en colaboración con

Antonio Moreno Sandoval

(Universidad Autónoma de Madrid)

1998 - 2018

ÍNDICE

0. Para empezar 4
1. Matriz 5
  1.1. Vector identidad y matriz identidad 5
  1.2. Operaciones de elementos matriciales 6
    1.2.1. Entre matriz y matriz 6
    1.2.2. Entre matriz y escalar 7
    1.2.3. Entre matriz y vector 8
  1.3. Producto de matrices 11
    1.3.1. Producto de dos vectores 11
    1.3.2. Producto de matriz y vector 12
    1.3.3. Producto de dos matrices 13
  1.4. Matriz traspuesta 14
  1.5. Funciones matriciales 16
2. Resumen 17
  2.1. Suma 17
  2.2. Media 18
    2.2.1. Media aritmética 18
    2.2.2. Media proporcional 18
    2.2.3. Media armónica 20
    2.2.4. Media truncada 23
    2.2.5. Media mayor 24
  2.3. Máximo, mínimo, rango, mitad 26
  2.4. Mediana 26
  2.5. Moda 27
  2.6. Varianza y desviación típica 28
  2.7. Asimetría y curtosis 33
    2.7.1. Asimetría 33
    2.7.2. Curtosis 35
  2.8. Distinción y oposición 38
    2.8.1. Grado Distintivo 38
    2.8.2. Grado Opositivo 39
3. Puntuación 43
  3.1. Puntuación relativa 43
  3.2. Puntuación opositiva 54
  3.3. Puntuación proporcional 57
  3.4. Puntuación limitada 59
  3.5. Puntuación comparada 62
  3.6. Puntuación estandarizada 68

3.7. Puntuación esperada 73
3.8. Puntuación ordenada 74
3.9. Puntuación divergente 76
3.10. Puntuación asociativa 80
3.11. Puntuación normalizada 83
3.11.1. Puntuación normalizada por suma total 83
3.11.2. Puntuación normalizada por media fraccional 84
3.11.3. Puntuación normalizada de Monsteller 84
3.12. Puntuación probabilística 88
3.12.1. Problema 89
3.12.2. Significatividad 91
3.12.3. Probabilidad Esperada 97
3.12.4. Multiplicador 101
3.13. Puntuación dummy 88
4. Relación 111
4.1. Correlación 111
4.1.1. Coeficiente de correlación 111
4.1.2. Matriz de correlación 117
4.2. Distancia 121
4.2.1. Distancia simple 121
4.2.2. Distancia estandarizada 121
4.2.3. Distancia Minkowski 122
4.2.4. Distancia normalizada en rango 122
4.2.5. Distancia Mahalanobis 125
4.3. Asociación 127
4.3.1. Coeficiente de asociación 127
4.3.2. Matriz de asociación 139
4.4. Asociación ordinal 142
4.5. Asociación nominal 142
4.6. Proximidad 143
4.6.1. Proximidad simple 143
4.6.2. Distancia regular / proximidad regular 145
4.6.3. Distancia media / proximidad media 149
5. Análisis 153
5.1. Análisis de medidas estadísticas 153
5.1.1. Análisis de rango 153
5.1.2. Análisis de centralidad 153
5.1.3. Análisis de variación 154
5.1.4. Análisis de balanza 157
5.1.5. Análisis de oscilación 158

5.2. Análisis de concentración 159
5.2.1. Concentración con criterio exterior 159
5.2.2. Concentración con criterio interior 163
5.2.3. Interpretación de valores de ejes 167
5.2.4. Coeficientes de concentración 169
5.3. Análisis multivariante 186
5.3.1. Regresión múltiple 186
5.3.2. Análisis de componentes principales 199
5.3.3. Análisis discriminante 212
5.3.4. Análisis de correspondencia 220
5.3.5. Análisis factorial 236
5.3.6. Análisis de cluster 243
5.4. Análisis de asociación 243
5.5. Análisis de agrupamiento 248
5.5.1. Agrupamiento por distancia 248
5.5.2. Agrupamiento por probabilidad 251
5.5.3. Agrupamiento por coocurrencia nominal 255
5.6. Dispersión lineal 257
5.7. Análisis de Condición Múltiple 258
5.7.1. Lista de condición múltiple 258
5.7.2. Frecuencia de condición múltiple 260
5.7.3. Coeficiente de condición múltiple 261
5.8. Generalidad y peculiaridad 269
5.9. Análisis por ejes selectivos 273
5.9.1. Análisis por ejes selectivos de matriz de correlación 274
5.9.2. Análisis por ejes selectivos de matriz de proximidad 277
5.9.3. Analisis por ejes selectivos de los casos 280
5.9.4. Análisis por ejes selectivos de los atributos y casos 280
5.9.5. Selección de ejes 282

0. Para empezar

El contenido de este documento se concentra en los métodos de análisis de datos cuantitativos que consideramos útiles para observar los cambios y variaciones de fenómenos lingüísticos. Los métodos tratados nos sirven para aclarar cuestiones numéricas que se nos escaparían si nos dedicáramos simplemente a la observación directa de los datos.

Como somos profesores y estudiantes de letras, es natural que no estemos acostumbrados a tratar las matrices de datos, puesto que no hemos realizado cursos de álgebra lineal con matrices, vectores y escalares. Empezaremos en primer lugar con fundamentos de cálculos matriciales, que nos hará más sencillo y fácil abordar los problemas de la búsqueda de las informaciones numéricas que nos interesan. El nivel de los temas tratados en este manual de texto no es muy alto, de modo que si avanzamos despacio con ejercicios prácticos, podremos llegar a nuestra meta: capacitarnos para realizar tratamientos de datos lingüísticos en sus aspectos cuantitativos.

La programación informática no es obligatoria para personas de letras. Su aprendizaje, sin embargo, es útil para desarrollar sus propios métodos en lugar de seguir y repetir ciegamente las mismas operaciones ofrecidas por los paquetes de software comerciales. De modo que hemos preparado algunos ejemplos de programas para los interesados en el desarrollo propio de los programas. Hemos buscado la manera más sencilla de aprendizaje de codificación en forma de un conjunto de funciones matriciales. La mayoría de los libros de programación ofrecen largos códigos que resultan difíciles de comprender. En su lugar hemos construido una serie de programas divididos en pequeñas partes unifuncionales. El programador debe aprender cómo combinar las relativamente pocas funciones, repetidas en distintos lugares. Hemos elaborado una versión de cada programa en Excel-VBA para uso individual en el escritorio, y otra en PHP para el uso a travás de web. Son NUMEROS-Excel y NUMEROS-web, respectivamente.

Este manual de uso se combina con el de LETRAS-Excel y LETRAS-web, que explica cómo tratar los datos textuales. Es recomendable estudiar los dos al mismo tiempo.

Sitio para descargas: http://lecture.ecc.u-tokyo.ac.jp/~cuedágengóindex.html
2.2. LETRAS para análisis de datos lingüísticos
2.3. NUMEROS para análisis de datos cuantitativos

1. Matriz

Empezamos el curso comprobando algunas operaciones matemáticas con matrices. Incluimos no solamente operaciones tratadas en la mayoría de los libros de álgebra lineal, sino también otras que definimos por primera vez en este documento[1]. Una vez comprendidas estas operaciones, nos sentiremos más seguros del significado de los tratamientos de datos numéricos y de sus modos de utilización. Para los programadores la codificación de estas operciones matriciales resulta más fácil y sencilla, tanto que pueden escribir códigos de pocas líneas.

Supongamos que hemos obtenido un recuento de frecuencia de datos concretos, por ejemplo, la frecuencia de una palabra propia de una región. Hay que comprobar, en primer lugar, si esta cifra es elevada comparada con otras cifras en otras localidades: v1, v2, v3, ... También nos interesa comparar estas cifras con otras palabras, d1, d2, ... Así de esta manera obtenemos una tabla construida de dos ejes, horizontal de localidades y vertical de palabras:

O.S.	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

En lo siguiente trabajaremos con estas tablas bidimensionales. Las tablas de ejemplo son siempre pequeñas pero en realidad puede ser miles de filas y 10 o 20 o más columnas.

1.1. Vector identidad y matriz identidad

Se llama «vector identidad» el vector que contiene el valor 1 en todos sus elementos[2]. Puede ser vertical I_{n1} y horizontal I_{1p}.

---

[1] Utilizaremos formas pasivas, pasiva refleja o pasiva con «ser» o «estar» y participio pasado, cuando se trata de los métodos conocidos; y formas activas cuando tratamos nuestros propios métodos. Hemos buscado informaciones por muchas partes, y al no encontrarnos con informaciones al respecto, bautizamos el método en cuestión con nuestras denominaciones. En algunas partes, proponemos utilizar otros nombres para mantener la organización de nuestro conocimiento, en cuyo caso ofreceremos los nombres comúnmente utilizados en las notas al final de la página.

[2] Existe otra definición de «vector identidad», pero utilizamos esta por ser el

I_{n1}	1
1	1
2	1

I_{1p}	1	2	3
1	1	1	1

En este manual tratamos al vector vertical I_{n1} como una matriz de n filas y 1 columna; y al vector horizontal I_{1p} como una matriz de 1 fila y p columnas. Solemos poner subíndices en matrices y en caso de un escalar (valor real), aparece sin subíndice, por ejemplo M.

Se llama «matrix identidad», la matriz cuadrada diagonal con el valor 1. Transcribiremos la matriz de unidad con I_{pp}, I_{nn}, siempre con subíndices iguales.

U_{pp}	1	2	3
1	1	0	0
2	0	1	0
3	0	0	1

1.2. Operaciones de elementos matriciales

1.2.1. Entre matriz y matriz

Entre los elementos correspondientes de dos matrices iguales en dimensión se pueden dar varias operaciones: suma (adición: A) y resta (sustracción: S). Proponemos realizar también multiplicación (M) y división (D), elevación (E) y logaritmo (L). Obsérvese que el valor del elemento Z(1,1) es 8, que es suma de X(1, 1) y Y(1, 1), es decir, 1 y 7 respectivamente:

X_{np}	1	2
1	1	4
2	2	5
3	3	6

+

Y_{np}	1	2
1	7	10
2	8	11
3	9	12

=

Z_{np}	1	2
1	8	14
2	10	16
3	12	18

$X_{np} + Y_{np} = Z_{np}$, $Z_{np} = A(X_{np}, Y_{np})$

La adición, $X_{np} + Y_{np} = Z_{np}$, y la sustracción, $X_{np} - Y_{np} = Z_{np}$, están definidas en libros de álgebra lineal. En la fórmula derecha, $Z_{np} = A(X_{np}, Y_{np})$, se utiliza la función que definimos como adición: A. La función A devuelve la suma de los dos argumentos, X_{np} y Y_{np}, que vienen entre paréntesis a continuación del nombre de la función (A).

Lo siguiente muestra la operación de multiplicación (M) entre elementos de matrices, que no suelen estar definida en los libros de álgebra. No obstante la

---

más usado en este documento

definimos, puesto que es muy útil en tratamientos subsiguientes. Hay que tener en cuenta que esta operación es distinta de la multiplicación o producto de matrices (X), que explicamos en la sección 1.3.

X_{np}	1	2
1	1	4
2	2	5
3	3	6

*

Y_{np}	1	2
1	7	10
2	8	11
3	9	12

=

Z_{np}	1	2
1	7	40
2	16	55
3	27	72

$X_{np} * Y_{np} = Z_{np}$, $Z_{np} = M(X_{np}, Y_{np})$

Como hemos dicho anteriormente, en este manual tratamos a los vectores como un tipo de matrices, de una sola fila o de una sola columna. De esta manera todas las operaciones matriciales podemos realizarlas sin tratamientos especiales:

X_{n1}	1
1	1
2	2
3	3

*

Y_{n1}	1
1	4
2	5
3	6

=

Z_{n1}	1
1	4
2	10
3	18

$X_{n1} * Y_{n1} = Z_{n1}$, $Z_{n1} = M(X_{n1}, Y_{n1})$

1.2.2. Entre matriz y escalar

Está definida la operación entre matriz y escalar (un valor numérico). Usualmente se hace la multiplicación (M) o la división (D), y nosotros las ampliamos a adición (A), sustracción (S), elevación (E) y logaritmo (L), que utilizaremos en los capítulos siguientes.

X_{np}	1	2
1	1	4
2	2	5
3	3	6

* 5 =

Z_{np}	1	2
1	5	20
2	10	25
3	15	30

$X_{np} * 5 = Z_{np}$, $Z_{np} = M(X_{np}, 5)$

X_{np}	1	2
1	1	4
2	2	5
3	3	6

+ 5 =

Z_{np}	1	2
1	6	9
2	7	10
3	8	11

$X_{np} + 5 = Z_{np}$, $Z_{np} = A(X_{np}, 5)$

X_{n1}	1
1	1
2	2
3	3

^ 2 =

Z_{n1}	1
1	1
2	4
3	9

$X_{n1} \wedge 2 = Z_{n1}$, $Z_{n1} = E(X_{n1}, 2)$

Aquí proponemos una idea que describimos con el nombre de «matriz homogénea». Por ejemplo, en la siguiente operación aditiva transformamos un escalar en una matriz de la misma dimensión que la matriz en la operación, con todos los elementos con el mismo valor que el escalar. Así podemos efectuar la operación admitida de adición normal:

X_{np}	1	2
1	1	4
2	2	5
3	3	6

+ 5 =

X_{np}	1	2
1	1	4
2	2	5
3	3	6

+

Y_{np}	1	2
1	5	5
2	5	5
3	5	5

=

Z_{np}	1	2
1	6	9
2	7	10
3	8	11

De esta manera consideramos que en la operación entre elementos de matrices, un escalar equivale a una «matriz homogénea en todo»:

5 =

Y_{np}	1	2
1	5	5
2	5	5
3	5	5

1.2.3. Entre matriz y vector

Realizamos operaciones entre los elementos de matriz y los de vector. En caso de un vector vertical Y_{n1}, lo transformamos en una «matriz homogénea en columnas», en este caso en dos columnas Y_{np}:

X_{np}	1	2
1	1	4
2	2	5
3	3	6

+

Y_{n1}	1
1	7
2	8
3	9

=

X_{np}	1	2
1	1	4
2	2	5
3	3	6

+

Y_{np}	1	2
1	7	7
2	8	8
3	9	9

=

Z_{np}	1	2
1	8	11
2	10	13
3	12	15

De la misma manera, al tratarse de un vector horizontal Y_{1p}, lo transformamos en una matriz homogénea en filas, en tres filas Y_{np}:

X_{np}	1	2
1	1	4
2	2	5
3	3	6

+

Y_{1p}	1	2
1	7	8

=

X_{np}	1	2
1	1	4
2	2	5
3	3	6

+

Y_{np}	1	2
1	7	8
2	7	8
3	7	8

=

Z_{np}	1	2
1	8	12
2	9	13
3	10	14

El propósito de la conversión de un vector en una matriz homogénea es posibilitar las operaciones matriciales generales. Consideramos equivalentes los vectores y sus correspondientes matrices homogéneas:

Y_{n1}	1
1	7
2	8
3	9

=

Y_{np}	1	2
1	7	7
2	8	8
3	9	9

Y_{1p}	1	2
1	7	8

=

Y_{np}	1	2
1	7	8
2	7	8
3	7	8

Utilizando las matrices homogéneas podemos hacer tal operación como la siguiente:

C	1	2
1	1	2

+

D	1
1	7
2	8
3	9

=

C	1	2
1	8	9
2	9	10
3	10	11

Esta operación es lo mismo que la siguiente:

C	1	2
1	1	2
2	1	2
3	1	2

+

D	1	2
1	7	7
2	8	8
3	9	9

=

C	1	2
1	8	9
2	9	10
3	10	11

(*) Generalización de operaciones de elementos matriciales

Vamos a generalizar las operaciones de elementos matriciales. Se hace fácilmente la derivación de X * Y = Z a (→) Y = Z / X tanto en escalares como en vectores y matrices.

2 * 3 = 6 → 3 = 6 / 2

X_{n1}	1	*	Y_{n1}	1	=	Z_{n1}	1	→	Y_{n1}	1	=	Z_{n1}	1	/	X_{n1}	1
1	1		1	7		1	7		1	7		1	7		1	1
2	2		2	8		2	16		2	8		2	16		2	2
3	3		3	9		3	27		3	9		3	27		3	3

X_{n1}	1	2	*	Y_{n1}	1	2	=	Z_{n1}	1	2	→	Y_{n1}	1	2	=	Z_{n1}	1	2	/	X_{n1}	1	2
1	1	4		1	7	10		1	7	40		1	7	10		1	7	40		1	1	4
2	2	5		2	8	11		2	16	55		2	8	11		2	16	55		2	2	5
3	3	6		3	9	12		3	27	72		3	9	12		3	27	72		3	3	6

Ahora nos interesa saber si se puede hacer lo mismo entre escalar y vector, entre escalar y matriz, entre vector y matriz. Constatamos que introduciendo las matrices homogéneas todas estas derivaciones son posibles:

X	1	2	*	5	=	Z	1	2	→	5	=	Z	1	2	/	X	1	2	=	A	1	2
1	1	4				1	5	20				1	5	20		1	1	4		1	5	5
2	2	5				2	10	25				2	10	25		2	2	5		2	5	5
3	3	6				3	15	30				3	15	30		3	3	6		3	5	5

X	1	2	*	Y	1	=	Z	1	2	→	Y	1	=	Z	1	2	/	X	1	2	=	B	1	2
1	1	4		1	7		1	7	28		1	7		1	7	28		1	1	4		1	7	7
2	2	5		2	8		2	16	40		2	8		2	16	40		2	2	5		2	8	8
3	3	6		3	9		3	27	54		3	9		3	27	54		3	3	6		3	9	9

Las últimas matrices A y B son homogéneas y equivalentes a un escalar (=5) o a un vector (Y). De esta manera, en todos los casos, comprobamos que de X * Y = Z se deriva Y = Z / X. Lo mismo puede decirse de las operaciones de adición y sustracción, de elevación y logaritmo.

El único problema lo encontramos cuando existe cero (0) en algún elemento de X, porque no es realizable la división Z / X. No obstante, este problema se salva si observamos que si un elemento de X es 0, el elemento correspondiente de Z también es 0 en X * Y = Z. Entonces, la división en cuestion Z / X es posible con la definición de la división de 0 por 0 igual a 0, a pesar de que no es admisible en la operaciones generales. En realidad, nos

encontramos con esta necesidad, 0 / 0 = 0, con frecuencia, por ejemplo, en el cálculo de la frecuencia relativa con la columna cero, que veremos en el cap. 3.

1.3. Producto de matrices

El «producto» (multiplicación) de matrices, que es fundamental en los cálculos matriciales, es distinto de la multiplicación de elementos de matrices, tratada en la sección anterior. Para obtener el producto hay que realizar una operación algo complicada, de suma de multiplicaciones de elementos de filas de la primera matriz y los de la columnas de la segunda matriz. Lo veremos con ejemplos más fáciles de vectores, y seguidamente entre matriz y vector; y finalmente entre matrices.

1.3.1. Producto de dos vectores

El producto entre vectores se obtiene por la suma de multiplicaciones de elementos correspondientes de vector fila y vector columna. En el ejemplo siguiente se hace el cálculo de: $X_{13}\ Y_{31} = 1*4 + 2*5 + 3*6 = 32$. En la multiplicación matricial no expresamos un signo de operación, tales como +, -, *, …, sino simplemente yuxtaponemos los dos argumentos participantes en el producto. Para el producto de matrices, utilizamos la función X. En la explicación por tablas utilizaremos el signo de $\times$ para distinguir de otras operadores (+, -, *, …).

$$X_{13}\ Y_{31} = Z,\ Z = X(X_{13},\ Y_{31})$$

X_{13}	1	2	3
1	1	2	3

$\times$

Y_{31}	x
1	4
2	5
3	6

=

Z_{11}	x
1	32

(*) Producto entre vector columna y vector fila

Si, en lugar de entre vector fila y vector columna, efectuamos el producto entre vector columna y vector fila, obtenemos una matriz con elementos individualmente multiplicados de elementos de primera y segunda matriz:

$$X_{31}\ Y_{13} = Z_{33},\ Z_{33} = X(X_{31},\ Y_{13})$$

Y_{31}	x
1	4
2	5
3	6

$\times$

X_{13}	1	2	3
1	1	2	3

=

Y_{31} X_{13}	1	2	3
1	4	8	12
2	5	10	15
3	6	12	18

Esta operación la realizamos en pocas ocasiones, pero también necesarias.

1.3.2. Producto de matriz y vector

Para obtener el producto de matriz y vector columna, se multiplica a filas de matriz por la columna del vector.

$X_{32}\ Y_{21} = Z_{31},\ Z_{32} = X(X_{32},\ Y_{21})$

X_{32}	1	2
1	1	2
2	3	4
3	5	1

$\times$

Y_{21}	1
1	2
2	3

=

Z_{31}	1
1	8
2	18
3	13

Es decir, el primer valor de Z_{31} se obtiene así: (1*2) + (2*3) = 8; el segundo (3*2)+(4*3) =18; y el tercero (5*2)+(1*3) = 13.

A continación demostramos el producto de vector fila y matriz, que se calcula a partir de un elemento que viene de producto de vector fila y la primera columna de la matriz (1*1 + 2*2 + 3*3 = 14) y otro del mismo vector y la segunda columna de la matriz (1*4 + 2*5 + 3*6 = 32):

$X_{13}\ Y_{32} = Z_{12},\ Z_{12} = X(X_{13},\ Y_{32})$

X_{13}	1	2	3
1	1	2	3

$\times$

Y_{32}	1	2
1	1	4
2	2	5
3	3	6

=

Z_{12}	1	2
1	14	32

(*) Producto de matriz y vector identidad

Multiplicando la matriz (X_{np}) por el vector identidad (I_{p1}), se obtiene un vector columna de suma filas de la matriz:

X_{32}	1	2
1	1	2
2	3	4
3	5	1

$\times$

I_{21}	1
1	1
2	1

=

Z_{31}	1
1	3
2	7
3	6

Por la multiplicación de vector fila y una matriz, se obtiene el producto en vector fila, que es la suma vertical de la matriz:

I_{13}	1	2	3
1	1	1	1

$\times$

X_{32}	1	2
1	1	4
2	2	5
3	3	6

=

Z_{12}	1	2
1	6	15

El primer elemento de Z es 6, resultado de 1*1 + 1*1 + 1*4 = 6, que es la suma de la primera columna de X. Ocurre lo mismo en el segundo elemento de Z: 1*4 + 1*5 + 1*6 = 15.

1.3.3. Producto de dos matrices

En la operación de $X_{np}\ Y_{pm} = Z_{nm}$, el elemento Z(i, j) es producto de la multiplicación de fila i de X_{np} y la columna j de Y_{pm}. Por ejemplo Z(1, 1) es producto de fila 1 de X y columna 2 de Y, 1*7 + 2*8 = 23:

$X_{32}\ Y_{23} = Z_{33}$, $Z_{33} = X(X_{32}, Y_{23})$

X_{32}	1	2
1	1	2
2	3	4
3	5	1

$\times$

Y_{23}	1	2	3
1	7	9	2
2	8	1	3

=

Z_{33}	1	2	3
1	1*7+2*8	1*9 + 2*1	1*2 + 2*3
2	3*7+4*8	3*9+4*1	3*2+4*3
3	5*7+1*8	5*9+1*1	5*2+1*3

=

Z_{33}	1	2	3
1	23	11	8
2	53	31	18
3	43	46	13

La multiplicación de dos matrices es solo posible cuando el número de columna de la primera matriz es igual al de filas de la segunda. Hay que tener presente la coincidencia de los subíndices p de la operación siguiente. La dimensión del producto de multiplicación Z_{nm} lleva el primer subíndice, el número de filas de la primera matriz (n) y el de columnas de la segunda (m):

$X_{np}\ Y_{pm} = Z_{nm}$

(*) Conmutación de matrices en la multiplicación

El producto de $X_{nn}\ Y_{nn}$ suele ser distinto del de $Y_{nn}\ X_{nn}$. Por consiguiente, en las operaciones de multiplicación matricial, se describen como «multiplicar X_{nn} por Y_{nn} por derecha» o «multiplicar Y_{nn} por X_{nn} por izquierda». Para simplificar la expresión se utilizan también los verbos «postmultiplicar» y «premultiplicar», respectivamente.

(*) Traslado de escalar

Un escaclar se puede trasladar a cualquier sitio dentro de la multiplicación matricial, puesto que el valor escalar es multiplicado equitativamente a todos los elementos de la matriz.

$$S\ X_{np}\ Y_{pm} = X_{np}\ S\ Y_{pm} = X_{np}\ Y_{pm}\ S$$

(*) Producto por matriz identidad

No se cambia la matriz (Xpp) ni por la postmultiplicación ni por la premultiplicación de una matriz identidad (Ipp).

(a) $X_{pp}\ I_{pp} = X_{pp}$

X_{pp}	x	y	z
1	1	2	3
2	4	5	6
3	7	8	9

×

I_{pp}	x	y	z
1	1	0	0
2	0	1	0
3	0	0	1

=

X_{pp}	x	y	z
1	1	2	3
2	4	5	6
3	7	8	9

(b) $I_{pp}\ X_{pp} = X_{pp}$

I_{pp}	x	y	z
1	1	0	0
2	0	1	0
3	0	0	1

×

X_{pp}	x	y	z
1	1	2	3
2	4	5	6
3	7	8	9

=

X_{pp}	x	y	z
1	1	2	3
2	4	5	6
3	7	8	9

1.4. Matriz traspuesta

La «transposición» de la matriz X_{np} consiste en conmutar los elementos X(i,j) y X(j,i). La «matriz traspuesta» se transcribe con una comilla simple: X_{np}'.

A_{n1}	1
1	1
2	2
3	3

A_{n1}'	1	2	3
x	1	2	3

A_{np}	1	2
1	1	4
2	2	5
3	3	6

A_{np}'	1	2	3
1	1	2	3
2	4	5	6

En este manual la matriz traspuesta X_{np}' la transcribiremos también por el cambio de subíndices: X_{pn}.

$X_{np}' = X_{pn}$

Hay que destacar las propiedades siguientes de la matriz traspuesta, que utilizaremos con frecuencia:

(a) $(X_{np}')' = X_{np}$

X_{np}	1	2
1	1	4
2	2	5
3	3	6

X_{np}'	1	2	3
x	1	2	3
y	4	5	6

$(X_{np}')'$	1	2
1	1	4
2	2	5
3	3	6

(b) $(X_{np} + Y_{np})' = X_{np}' + Y_{np}'$

X_{np}	1	2
1	1	4
2	2	5
3	3	6

+

Y_{np}	1	2
1	7	10
2	8	11
3	9	12

=

Z_{np}	1	2
1	8	14
2	10	16
3	12	18

Z_{np}'	1	2	3
1	8	10	12
2	14	16	18

X_{np}'	1	2	3
1	1	2	3
2	4	5	6

+

Y_{np}'	1	2	3
1	7	8	9
2	10	11	12

=

Z_{np}	1	2	3
1	8	10	12
2	14	16	18

(c) $(X_{np}\ Y_{pm})' = Y_{pm}'\ X_{np}'$

X_{np}	1	2
1	1	4
2	2	5
3	3	6

×

Y_{p1}	x
1	1
2	2

=

Z_{n1}	x
a	9
b	12
c	15

Z_{n1}'	1	2	3
1	9	12	15

Y_{p1}'	1	2
1	1	2

×

X_{np}'	1	2	3
1	1	2	3

=

Z_{1n}	1	2	3
1	9	12	15

2	4	5	6

1.5. Funciones matriciales

Hemos preparado las funciones matriciales siguientes:

Ms(X11): convierte matriz M11 en un escalar M.

Sm(X): convierte un escalar M en matriz M11.

SumR(X_{np}): devuelve un vector columna de las sumas de las filas de una matriz.

SumC(X_{np}): devuelve un vector fila de las sumas de las columnas de una matriz.

SumA(X_{np}): devuelve la suma (escalar) de la matriz

Para obtener SumR y SumC utilizamos las operaciones siguientes que hemos tratado en este capítulo:

$SumR = X_{np}\ I_{p1}$

$SumC = I_{n1}'\ X_{np}$

2. Resumen

Utilizando operaciones matriciales vamos a calcular tales medidas estadísticas que resume los datos como «Suma», «Media», «Varianza», «Desviación típica», etc. Seleccionamos el objeto del cálculo dentro de los «Aspectos» de la matriz: Fila, Columna y Total.

2.1. Suma

Calculamos las sumas de filas, columnas y la suma total. El vector columna de sumas de filas S_{n1} se calcula con la operación siguiente:

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

El vector columna de identidad Ip1 contiene P elementos que son todos 1.

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

El vector fila de sumas verticales S_{1p} se obtiene del producto de $I_{1n}\ D_{np}$:

$$S_{1p} = I_{n1}'\ D_{np} = I_{1n}\ D_{np}$$

Aquí el vector fila de identidad I_{n1}' es un vector traspuesto del vector identidad I_{n1}.

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

Es decir, 1*6+1*7+1*8+1*9+1*10 = 40, etc.

Finalmente, la suma total S (escalar) de la matriz es la suma de las sumas horizontales S_{n1} o de las sumas verticales S_{1p}:

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

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

X

I_{p1}	1
1	1
2	1
3	1

=

S	1
1	103

2.2. Media

2.2.1. Media aritmética

La «Media» se obtiene por la división de la suma de datos por el número de datos (N):

$$\text{Media} = \text{Sum}(D_{np}) / N$$

La tabla inferior derecha muestra la relación entre la Media, Mitad, Rango y Mediana:

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

Columna	1	2	3
Media	8.000	7.600	5.000
Media-Mitad	.000	.600	.000
(Media-Mitad) / Rango	.000	.100	.000
Media-Mediana	.000	-.400	.000

2.2.2. Media proporcional

La media aritmética es apropiada para aplicar al conjunto de datos sin valores extraordinarios. En cambio, cuando se aplica al conjunto que los contiene, por ejemplo {1, 1, 2, 3, 153} donde apreciamos un valor extraordinario, 153, la media aritmética (32) ofrece un valor muy distante de la realidad. Este valor (32) dista tanto de {1, 1, 2, 3} como de 153, de modo que no podemos afirmar su representavidad. En tal caso, se recomienda en la estadística general el uso de mediana (2), Sin embargo, esta mediana, 2 tampoco significa mucho o casi nada para el miembro extraordinario, 153. Por consiguiente la mediana no sirve tampoco para representar el conjunto de datos.

Por lo tanto, para un conjunto de datos dotado de unos valores extraordinarios, proponemos utilizar una clase de la media que denominamos Media Proporcional (MP). Primero repasamos la fórmula de la Media Aritmética (MA) y, seguidamente, explicamos la fórmula de la Media Proporcional (MP) de manera comparada.

$MA = (x_1 + x_2, \ldots, x_n) / n = 1/n \ \Sigma x_i$ n: recuento de casos
$= (1 + 1 + 2 + 3 + 153) / 5 = 32$

La fracción $1 / n = 1/ 5$ se la puede distribuir de la manera siguiente:

$AM = (x_1 / n + x_2 / n, \ldots, x_n / n) = \Sigma x_i / n$
$= (1 / 5 + 1 / 5 + 2 / 5 + 3 / 5 + 153 / 5) = 32$

De esta manera, la Media Aritmética (MA) equivale al producto de suma de la multiplicaciones de la frecuencia por el peso $1/n$ (= 1 / 5). En cambio, en el cálculo de la Media Proporcional (MP), en lugar del valor constante de $1/n$ (= 1 / 5), utilizamos la ratio de la frecuencia con respecto a la totalidad de la frecuencia (N) de manera siguiente:

$N = (x_1 + x_2, \ldots, x_n) = \Sigma x_i$
$WM = (x_1 x_1 / N + x_2 x_2 / N, \ldots, x_n x_n / N)$
$= (x_1 x_1 + x_2 x_2, \ldots, x_n x_n) / N$
$= \Sigma x_i^2 / N$
$= (1^2 + 1^2 + 2^2 + 3^3 + 153^2) / 32$
$= 146.4$

Notamos que el valor de la Media Proporcional (MP), 146.4, refleja la magnitud del conjunto de datos {1, 1, 2, 3, 153}. Compárense la Suma, la Media Aritmética (MA) y la Media Proporcional (MP) en la tabla siguiente (X):

X	v1	v2	v3	v4	v5	Suma	MA	MP
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

Observamos que la Media Proporcional (MP) representa la magnitud del conjunto de datos, de modo que se acerca al valor máximo alejándose de los valores mínimos.

(#) Preguntas confirmativas del español

La tabla siguiente muestra las frecuencias normalizadas por 100000 palabras de las distintas preguntas confirmativas del español con añadidura de Suma, Media Aritmética (MA) y Media Proporcional (MP):

F.N.Pl :100000	1. ¿no?	2. ¿sí?	3. ¿eh?	4. ¿cierto?	5. ¿verdad?	6. ¿cacháis?	7. ¿sabes?	8. ¿viste?	9. ¿hm?	Suma	MA	MP
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

La tabla anterior muestra que la Media Aritmética (MA) falla en representar el conjunto de las frecuencias que ofrece valores sumamente distantes entre sí. Al fijarnos en la Media Proporcional (MP) y en la tabla siguiente de la Puntuación de orden ascendente, notamos que en los 5 lugares indicados no coinciden entre las dos medias. En la Suma, que aparentemente mostraría la magnitud del conjunto de datos, la Puntuación de orden debe ser lógicamente idéntica a la de la Media Aritmética y, por lo tanto, no sirve tampoco como valor representativo.

ARSc	Sum	AM	WM
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

*Datos y análisis: PRESEEA en LYNEAL (2017/8/11)
http://lecture.ecc.u-tokyo.ac.jp/~cueda/lyneal/preseea.htm

2.2.3. Media armónica

De tales valores como la velocidad, densidad, media, ratio, porcentaje, etc. que se obtienen por la división, no podemos derivar la media de ellos haciendo la suma de estos valores por el número de los mismos. Por ejemplo,

intentamos calcular la media de velocidad de las dos caminatas, ida y vuelta, que han hecho un grupo de senderistas. Supongamos que en el camino de ida (6 km), su velocidad ha sido de 6 km/h, y la de la vuelta (6 km) ha sido 4 km/h. La "velocidad media" sería (6 + 4) / 2 = 5, que en realidad no es exactamente la media de las dos velocidades. La prueba de ello es que el cálculo de la suma de horas no se cuadra. El tiempo total de esta caminata resultaría 12 km / 5 km/h = 2.4 h, que es falso. En realidad el tiempo de la ida es 6 km / 6 km/h = 1 h y el de la vuelta: 6 km / 4 km/h = 1.5 h, en total 2.5 h. La diferencia es pequeña pero el primer cálculo sigue siendo distante de la realidad.

Este cálculo se ha hecho con datos de distancias y horas. Pero ¿qué ocurre cuando disponemos solo de los datos de velocidad? En tal caso se usa la «Media armónica» que se calcula de manera siguiente.

Supongamos que la distancia total es K (km), el tiempo total es H, y la velocidad de ida es X km/h y la de vuelta Y km/h. Nuestro propósito es obtener la velocidad media en forma de K / H, expresada por X e Y.

El tiempo total H es la suma del tiempo de ida y el de la vuelta, que son K / 2 / X (mitad de distancia dividida por la velocidad X) y K / 2 / Y, respectivamente.

H = K / 2 / X + K / 2 / Y ← Tiempo total
= K / 2 X + K / 2 Y ← Juntamos los denominadores
= (1 / X + 1 / Y) K / 2 ← Juntamos las partes comunes

A partir de la última fórmula pensamos derivar nuestro objetivo K / H, la velocidad media.

H = (1 / X + 1 / Y) K / 2 ←Tiempo total
1 / H = 1 / [(1 / X + 1 / Y) K / 2]
←Pasar ambas partes al denominador
K / H = 1 / [(1 / X + 1 / Y) / 2] ← Multiplicamos por K

Esta es la fórmula de la Media armónica (M.arm.):

M.arm. (X, Y) = 1 / [(1 / X + 1 / Y) / 2]

En nuestro caso concreto,

M.arm. (6, 4) = 1 / [(1 / 6 + 1 / 4) / 2] = 4.8

La Media armónica es un caso especial de la «Media fraccional» que explicaremos seguidamente, especial en el sentido de que los denominadores son comunes en los dos términos, en nuestro caso, K (kilómetros).(*) Media fraccional

En varias ocasiones en este curso, utilizaremos una versión ampliada de la Media armónica, cuando disponemos de denumeradores y denominadores de dos términos, por ejemplo, «Ratios», resultados de división: $R_1 = A_1 / B_1$, $R_2 = A_2 / B_2$, que denominamos «Meida fraccional» (M. frac.)[3].

M. frac. = $(A_1 + A_2) / (B_1 + B_2)$

Las medias pueden presentar valores parecidos, pero la Media fraccional utiliza las propias cifras que componen la fracción y ofrece un valor más exacto, y el resultado es fácil de entender. Es un caso por ejemplo de la mezcla de dos cantidades distintas de agua con sal, también en cantidades distintas. Comparemos las tres medias «Media aritmética» (M.arit.), que se llama normalmente «Media» a secas, «Media armónica» (M. arm.) y «Media fraccional» (M.frac.) de, por ejemplo, 1/4 y 2/5.:

M.arit.(1/4, 2/5) = (1/4 + 2/5) / 2 = 0.325,

M.arm.(1/4, 2/5) = 1 / [(4 / 1 + 5 / 2) / 2] ≒ 0.308

M.frac.(1/4, 2/5) = (1 + 2) / (4 + 5) ≒ 0.333

Comparemos las tres medias de (1/4, 2/5) y de (10/40, 4/10:

Media	**1/4, 2/5**	**10/40, 4/10**
M.arit.	0.325	0.325
M. arm.	0.308	0.308
M. frac.	0.333	0.280

En esta comparación podemos observar que la media fraccional ofrece un descenso notable cuando aumentamos las cifras de denumerado y de denominador.

La tabla siguiente muestra la comparación de las tres medias en caso de la velocidad en una distancia igual de ayer (ida) y hoy (vuelta):

A. Misma distancia	Ayer	Hoy	Suma	M.arit.	M.arm.	M.frac.
Distancia (km)	12	12	24			
Hora(h)	2	3	5			
Velocidad(km/h)	6	4	4.80	5.00	4.80	4.80

Ahora bien, si hoy no volvemos por el mismo camino, sino seguimos la ruta de distancia diferente, tenemos el cálculo siguiente:

[3] En la literatura de estadística, se llama «Media aritmética ponderada».

A. Distinta dist.	Ayer	Hoy	Suma	M.arit.	M.arm.	M.frac.
Distancia (km)	12	15	27			
Hora(h)	2	3	5			
Velocidad(km/h)	6	5	5.40	5.50	5.45	5.40

En este caso la Media armónica no ofrece una velocidad media correcta. La Media fraccional se calcula directamente con distancia y hora, cuyo significado se entiende fácilmente.

2.2.4. Media truncada

Ante un conjunto de datos numéricos que contiene unos elementos de valor extraordinario, por ejemplo, {1, 55, 5, 2, 4}, la Media aritmética (13.4) no sirve como un resumen del mismo conjunto. La mayoría de los elementos {1, 5, 2, 4} distan de la Media y tampoco la cifra extraordinario 55 es distante de la Media.

Para eliminar la influencia del elemento extraordinario, se utiliza la Mediana, 4, que sí que puede servir como resumen de la mayoría {1, 5, 2, 4}. La Mediana, sin embargo, no considera más que un elemento central y no varía entre, por ejemplo, {2, 3, 4, 6, 9} y {2, 3, 4, 7, 12}. En tal caso, la Media aritmética es mejor.

De esta manera la Media y la Mediana poseen sus méritos y desventajas y a veces no se sabe cuál es mejor para resumir el conjunto. Para salvar esta dificultad, se ha inventado una medida llamada «Media truncada» (ing. *Trimmed mean*), que va sumando primero la Media y después, quitanto el Máximo y Mínimo del conjunto para sacar la segunda Media, y otra vez, quitando el Máximo y Mínimo del conjunto restante, se calcula nuevamente la Media, y así sucesivamente, para juntar todas las Medias hasta llegar al valor de Mediana, y finalmente la Suma de Medias queda dividida por el número de Medias. A continuación mostraremos los procesos concretos utilizando el mencionado conjunto de números, que conviene estar ordenado para que resulten fácil los truncamientos de valor máximo y valor mínimo en cada proceso:

(1) (1 + 2 + 4 + 5 + 55) / 5 = 13.4
(2) (2 + 4 + 5) / 3 = 3.67
(3) (4) / 1 =4
(4) (13.4 + 3.67 + 4) / 3 = 7.02

La tabla siguiente muestra los valores de la Media, la Media truncada y la Mediana,

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

X_{np}	Media	M. trunc.	Mediana
d1	12.400	12.133	12.000
d2	5.800	6.267	7.000
d3	2.800	1.489	1.000
d4	1.800	1.933	2.000

En la Media truncada, en el primer paso se calcula la Media y en el último paso se considera la Mediana y los valores extraordinarios que apartan de la Mediana son considerados menos veces que los valores centrales:

$$L = \text{Int}(\frac{N+1}{2})$$

$$\text{T.ave.} = [\sum_{i=0}^{L-1} \frac{1}{N-2i} \sum_{j=1+i}^{N-i} X(j)] \ / \ L$$

donde L representa el valor entero (Int) de (N+1) / 2, N es el número de datos, y X(j) es cnjunto de datos ordenados.

2.2.5. Media mayor

La Media truncada da un peso grande a los elementos cercanos a la Mediana. Los valores distantes de la Mediana participan en las Medias menos veces. Para salvar esta desigualdad de tratamiento, proponemos un cálculo que llamamos «Media mayor» que es la Media de mayorías de los elementos. Consiste en formar un grupo de la mitad de los primeros elementos ordenados, que abarca la cantidad de elementos igual o mayor de la mitad, según el número de datos, y calculamos la Media de este grupo. Seguidamente formamos de nuevo el segundo grupo, empezando con el segundo de los elementos con la cantidad igual y calculamos de nuevo la Media, y asi sucesivamente vamos calculando (naturalmente por programa) todas las Medias hasta que llegue el corte a la Mediana, donde termina la formación del nuevo grupo. Y finalmente calculamos la Media de las Medias acumuladas, que es la Media mayor. Mostraremos el proceso concreto de los cálculos con el mismo conjunto de datos, 1, 2, 4, 5, 55.

(1) (1 + 2 + 4} / 3 = 2.33
(2) (2 + 4 + 5) / 3 = 3.67
(3) (4 + 5 +55) / 3 = 21.33
(4) (2.33 + 3.67 + 21.33) / 3 = 9.11

Comparemos la Media, la Media truncada y la Media mayor, en la tabla inferior derecha. Creemos que la última, la Media mayor, ofrece unos valores

mas medianos, en el sentido de que considera cada vez la mayoría de los elementos y, efectivamente, ocupa un punto medio entre dos extremos, la Media y la Media truncada.

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

Fila	Media	M. trunc.	M. mayor
d1	12.400	12.133	12.222
d2	5.800	6.267	6.000
d3	2.800	1.489	1.889
d4	1.800	1.933	1.889

Lo siguiente muestra la derivación de Media mayor:

$$L = \text{Int}(\frac{N+1}{2})$$

$$M = \text{Int}(\frac{N+2}{2})$$

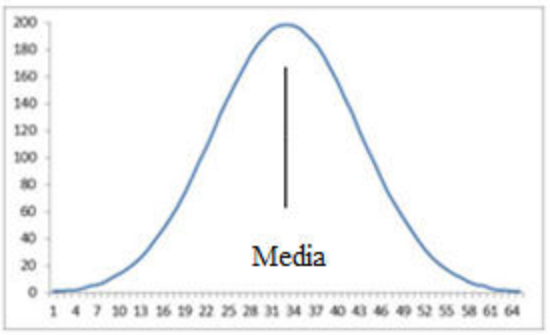
$$\text{M.ave.} = [\sum_{i=1}^{M} \sum_{j=i}^{i+L-1} X(j)] \ / \ (L * M)$$

donde L es la amplitud de grupo de mayoría, N es número de datos, M es el sitio de inicio en el último grupo.

(#) Distribución de forma L de datos lingüísticos

Los datos de estatura del cuerpo o de puntuaciones de asignaturas suelen presentar una curva de distribución (eje horizontal: orden, eje vertical: frecuencia), llamada la «Curva normal» en forma de campana donde se acumulan las altas frecuencias en la Media y descienden los valores tanto por las cifras menores como las mayores de manera pronunciada. También se conoce como «Distribución gaussiana» (Figura 1).

Los datos lingüísticos, fonemas, letras, morfemas, léxicos, etc. suelen presentar una curva peculiar que se llama «Curva de forma L», o «Distribución de ley de potencia», donde se observa la alta frecuencia de los primeros en el orden, y la curva desciende rápidamentte según avanza hacia la derecha. Se contrasta la frecuencia sumamente alta de los primeros pocos elementos más frecuentes y los elemenos de baja frecuencia que son numerosos (Figura 2).



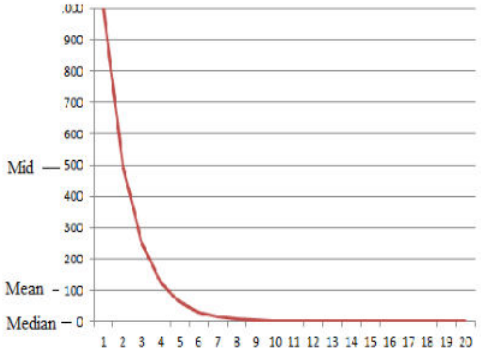


Figura 1. Distribución normal Figura 2. Distribución en L.

2.3. Máximo, mínimo, rango, mitad

Dentro de los tres aspectos de la matriz objeto de análisis se calculan medidas de «Máximo», «Mínimo», «Rango» y «Mitad». El Rango se refiere a la diferencia entre el Máximo y el Mínimo. La Mitad se calcula por la división del Rango por 2, que indica el punto medio del rango.

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

Columna	1	2	3
Mínimo	6	4	2
Máximo	10	10	8
Mitad	8	7	5
Rango	4	6	6

2.4. Mediana

La «Mediana» es el valor que tiene el elemento que se sitúa en el punto medio dentro de la ordenación ascendente (o descendente) de los datos. Por ejemplo, los elementos de la columna 2 son {8, 10, 4, 7, 9}. Ordenamos los datos en forma de {4, 7, 8, 9, 10} y el elemento que ocupa el punto medio de todos es el tercero que lleva el valor de 8, que corresponde a la Mediana. Si el número de elementos es par, la Mediana es la Media de los dos elementos centrales. La tabla inferior derecha muestra la relación entre la Mediana y Mitad:

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

Columna	1	2	3
Mediana	8.000	8.000	5.000
Mediana - Mitad	.000	1.000	.000
(Mediana - Mitad) / Rango	.000	.167	.000

2.5. Moda

Se llama «Moda» la cifra que aparece más veces que otras. En un dato {0, 1, 2, 3, 3} la Moda es 3.

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

Fila	Moda	Moda: Frecuencia
d1	Ninguna moda	Ninguna moda
d2	Ninguna moda	Ninguna moda
d3	Ninguna moda	Ninguna moda
d4	3	3: 2

Las primeras dos filas (d1, d2) no presentan la Moda, por ser todos los miembros distintos de valor. La tercera fila (d3) tampoco presenta la Moda, puesto que hay dos cifras 0, 1, que son de la misma frecuencia máxima 2, es decir cada cifra aparece dos veces, por lo cual no se determina la Moda, la cifra que representa la máxima frecuencia del mismo valor.

(*) Moda mayor

Hemos visto que la Moda no sirve cuando los valores del conjunto son todos distintos. Por otra parte, cuando la Moda se aparta del segundo o tercer valores más frecuentes, no representaría exactamente la "moda" del conjunto. Ampliando el concepto de Moda en el reprentante de máxima concentración. Pensamos por esta razón buscar un rango de mayoría donde se observa la máxima concentración de cifras y la Media de las cifras del mismo rango, y la denominamos «Moda mayor».

Buscamos, por ejemplo, dentro del conjunto, d1 = {10, 19, 14, 7, 12}, ordenado en {7, 10, 12, 14, 19}, un subconjunto de los tres elementos, que ocupa la mayoría, que ofrezca el rango menor, de manera siguiente:

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

donde encontramos el rango menor en la fila 2, de modo que la Moda mayor es la media de 10, 12, y 14, igual a 12. Cuando se encuentran más de un rango menor de la misma distancia, aumentamos la amplitud de rangos, de 3, 4 hasta 5, que es el número de datos. Cuando llega al número de datos (N), la Moda mayor resulta igual a la Media. La tabla siguiente muestra la Moda mayor de cada fila. De esta manera, siempre disponemos de la Moda, cosa que era imposible en la Moda normal.

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

Fila	Moda mayor	Rango
d1	12.000	10.000 - 14.000
d2	9.333	7.000 - 11.000
d3	.500	.000 - 1.000
d4	2.667	2.000 - 3.000

2.6. Varianza y desviación típica

Para describir los datos hay que mostrar no solo los valores de centro, como Media, Mediana, etc., sino también un indicador de variación. Por ejemplo, dos tierras que tienen la temperatura Media de 20 grados son muy diferentes si una tiene la variación entre 10 y 30 grados y la otra, 10 y 40 grados durante un año.

Vamos a comparar los dos datos d1 y d2, por ejemplo: {d1: 4, 5, 6, 7, 8} y {d2: 2, 4, 6, 8, 10}. Ambos datos tienen la misma Media 6, pero sus variaciones son diferentes. Para medir la variación se necesitan datos de «Desviaciones» (diferencias con respecto a la Media), de modo que restamos la Media de los elementos de cada dato, {d1: 4-6, 5-6, 6-6, 7-6, 8-6} と {d2: 2-6, 4-6, 6-6, 8-6 10-6}, y obtenemos las desviaciones {d1: -2, -1, 0, 1, 2} y {d2: -4, -2, 0, 2, 4}. Si sumamos todas las Desviaciones siempre obtenemos un valor 0, que no sirve para comparar los datos. Por esta razón, elevamos al cuadrado los valores de desviaciones: {d1: $(-2)^2$, $(-1)^2$, 0^2, 1^2, 2^2} y {d2: $(-4)^2$, $(-2)^2$, 0^2, 2^2, 4^2} y obtenemos {d1: 4, 1, 0, 1, 4} y{d2: 16, 4, 0, 4, 16}. Ahora podemos sumar estos valores: {d1: 4+1+0+1+4}, {d2: 16+4+0+4+16} que son «Varianzas» (V): V(d1) = 10, V(d2) = 40. El cálculo de la Varianza se ha hecho por elevación al cuadrado, y para volver a la escala original de los datos sacamos la raíz cuadrada, que es la «Desviación típica» (DT): DT(d1) = $10^{1/2}$ = 3.16, DT(d2) = $40^{1/2}$ = 6.32[4].

La derivación matricial de Varianza y Desviación típica de filas es (N: número de filas, P: número de columnas):

$S_{n1} = X_{np}\ I_{p1}$ ←Columna de sumas horizontales (v. 2.1)
$M_{n1} = S_{n1} / P$ ←Columna de medias horizontales
$D_{np} = X_{np} - M_{n1}$ ←Matriz de desviaciones
$C_{np} = D_{np}^2$ ←Matriz de cuadrados de desviaciones
$W_{n1} = C_{np}\ I_{p1}$ ←Columna de sumas de cuadrados de desviaciones
$V_{n1} = W_{n1} / P$ ←Columna de Varianza horizontal

[4] En este manual, en vez del signo de raíz, como $\sqrt{10}$ = 3.16, utilizamos el exponente 1/2, puesto que la operación y la demostración de las fórmulas matemáticas resultan más fáciles y sencillas.

$DT_{n1} = V_{n1}^{1/2}$ ←Columna de Desviación típica horizontal

Utilizando las funciones matriciales:

$V_{n1} = D(X(E(S(X_{np}, D(X(X_{np}, Ip1), P)), 2), Ip1), P)$
$DT_{n1} = E(V_{n1}, 1/2)$

Las matrices y los vectores producidos en las operaciones anteriores son los siguientes:

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

I_{p1}	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

De la misma manera, vamos a calcular el vector fila de Varianza vertical y de la Desviación típica vertical:

$S_{1p} = I_{1n} X_{np}$ ←Fila de sumas verticales (v. 2.1)
$N = CntR(X_{np})$ ←Número de filas de X
$M_{1p} = S_{1p} / N$ ←Fila de medias verticales
$D_{np} = X_{np} - M_{1p}$ ←Matriz de desviaciones
$C_{np} = D_{np}^2$ ←Matriz de cuadrados de desviaciones
$W_{1p} = I_{1n} C_{np}$ ←Fila de sumas de cuadrados de desviaciones
$V_{1p} = W_{1p} / N$ ←Fila de Varianza vertical
$DT_{1p} = V_{1p}^{1/2}$ ←Fila de Desviación típica vertical

Utilizando las funciones matriciales:

$V_{n1} = D(X(I_{1n}, E(S(X_{np}, D(X(I_{1n}, X_{np}), N)), 2)), N)$

$DT_{1p} = E(V_{1p}, 1/2)$

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

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

S_{1p}	1	2	3	4	5
1	21.00	27.00	27.00	22.00	17.00

M_{1p}	1	2	3	4	5
1	5.25	6.75	6.75	5.50	4.25

D_{np}	1	2	3	4	5
1	4.75	12.25	7.25	1.50	7.75
2	5.75	.25	3.25	-5.50	-3.25
3	-5.25	-6.75	-5.75	6.50	-3.25
4	-5.25	-5.75	-4.75	-2.50	-1.25

C_{np}	1	2	3	4	5
1	22.56	150.06	52.56	2.25	60.06
2	33.06	.06	10.56	30.25	10.56
3	27.56	45.56	33.06	42.25	10.56
4	27.56	33.06	22.56	6.25	1.56

W_{1p}	1	2	3	4	5
1	110.75	228.75	118.75	81.00	82.75

V_{1p}	1	2	3	4	5
1	27.69	57.19	29.69	20.25	20.69

DT_{1p}	1	2	3	4	5
1	5.26	7.56	5.45	4.50	4.55

(*) Variación y Dispersión

La Desviación típica, que se utiliza como indicador de variación, tiene la propiedad de aumentar de acuerdo con la escala de los datos. Por esta razón se ha buscado un indicador constante de variación independiente de la escala de datos. Por consiguiente, el Coeficiente de Variación (CV) se calcula de la Desviación Típica (DT) dividida por la Media (M).

$CV = DT / M$

Como el Coeficiente de Variación (CV) no está normalizado, es decir, no vacila entre 0 y 1, buscamos un indicador normalizado de variación, que denominamos «Desviación Típica Normalizada» (DTN), que se calcula por la división de la Desviación Típica (DT) por el valor máximo de la misma

Desviación Típica (DT.max.)[5]:

$$DTN = DT. / DT.max$$

Veamos la manera de buscar la fórmula de DT.max. Partimos de la fórmula de la Desviación típica (DT):

$$DT = \{[(X_1 - M)^2 + (X_2 - M)^2 + \ldots + (X_n - M)^2] / N\}^{1/2}$$

Ahora bien, supongamos que estamos ante un conjunto de datos con un caso extremo de desviación, por ej. {10, 0, 0, 0, 0}, en cuyo caso se presenta el valor máximo de Desviación típica (DT.max.). Para generalizar el problema, utilizamos K en lugar de una cifra concreta: {K, 0, 0, …, 0}. Entonces, solo el primer término de DT es $(K - M)^2$, y todos los restantes son $(0 - M)^2 = M^2$, y por lo tanto, el valor máximo de Desviación típica (DT.max.) es:

$$DT.max. = \{([(K - M)^2 + (N - 1)\ M^2]\}^{1/2}$$

donde, K es igual a la suma de los datos, puesto que los restantes son nulos. Como la suma es igual a la Media (M) multiplicada por el Número (N) de datos (Suma = N M ← M = Suma / N), K es igual a N M:

$$K = Suma = N\ M$$

Por lo tanto:

$$\begin{aligned}
DT.max &= \{[(N\ M - M)^2 + M^2\ (N - 1)] / N\}^{1/2} && \leftarrow\ K = N\ M \\
&= \{[(M\ (N - 1))^2 + M^2\ (N - 1)] / N\}^{1/2} && \leftarrow\ M \text{ al exterior} \\
&= \{[M^2\ (N - 1)^2 + M^2\ (N - 1)] / N\}^{1/2} && \leftarrow\ M^2 \text{ es común} \\
&= \{M^2\ (N - 1)\ [(N - 1) + 1] / N\}^{1/2} && \leftarrow\ M^2\ (N - 1) \text{ es común} \\
&= \{(M^2\ (N - 1)\ N / N\}^{1/2} && \leftarrow\ (N - 1) + 1 = N \quad \leftarrow (N-1) + 1 = N \\
&= [(M^2\ (N - 1)]^{1/2} && \leftarrow\ N / N = 1 \\
&= M\ (N - 1)^{1/2} && \leftarrow\ (M^2)^{1/2} = M
\end{aligned}$$

Por lo tanto, la «Desviación Típica Normalizada» (DTN) es:

$$DTN = DT. / DT.max = DT / [M\ (N - 1)^{1/2}]$$

La diferencia entre Coeficiente de Variación (CV) y Desviación Típica Normalizada (DTN) está en que en el segundo se encuentra el valor de $(N - 1)^{1/2}$ en el denominador. Cuando se trata de los datos cuyo número (N) es grande, el DTN se vuelve pequeño. Recomendamos utilizar el DTN no de modo vertical con N individuos, sino de modo horizontal con P variables, cuyo número suele

[5] Este método de normalización lo utilizaremos en varias ocasiones.

ser relativamente pequeño.

(#) El uso de palabras y dispersión

A. Juilland y E. Chang Rodríguez en su *Frequency dictionary of Spanish words,* (The Hague: Mouton, 1964) propusieron utilizar la fórmula de Uso (U) de palabras por la multiplicación de Frecuencia (Frec) de la palabra en cuestión por su grado de Dispersión (Disp), con los cinco conjuntos de datos: dramas, novelas, ensayos, documentos científicos y noticias:

U = Frec * Disp

De modo que para considerar el grado de uso de palabras, según estos dos autores, hay que ver no solamente sus frecuencia sino también los grados de dispersión y presentaron la fórmula siguiente del grado de dispersión (D):

Disp = 1 – DT / (2 * Media)

Constatamos que el número 2 que se encuentra en el denominador representa (el número de variables 5 – 1)$^{1/2}$. De esta manera, notamos que DT / (2 x Media) es el Desviación Típica Normalizada (DTN). Por lo tanto, para generalizar el valor de Dispersión (Disp), utilizaremos la fórmula siguiente[6];

Disp = 1 – DTN

(*) Índice de oscilación

Cuantificamos el modo de oscilación observada en la secuencia de los datos con una cifra que denominamos «Índice de Oscilación» (IO). Para el cálculo utilizamos el valor de Subida (Sb) y el da Bajada (Bj) de manera siguiente:

Se cuentan 2 veces de subida (Sj) en las secuencias 10→19 y 7→12. Por otra parte 2 veces de bajada (Bj) en 19 → 14 y 14 → 7. Definimos el Índice de Oscilación (IO):

IO = (Sb – Bj) / (Sb + Bj)

Por lo tanto:

IO = (2 – 2) / 2 + 2) = 0

Seguidamente, calculamos las diferencias en Subidas y bajadas entre los

[6] Cuando se trata de un valor normalizado (VN) cuyo rango es de 0 a 1, se puede formular su correspondiente valor inverso por 1 – VN. Esta operación la realizaremos en varias ocasiones

dos valores contiguos. En la Subida contamos con Sb(10→19) = 9, Sb(7→12) = 5, en total Sb = 14. En la Bajada, P(19→14) = 5, N(14→7) = 7, en total, Bj=12:

$$IO(d1) = (14 - 12) / (14 + 12) = .077$$

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

Fila	Oscilación (frecuencia)	Oscilación (distancia)
d1	.000	.077
d2	.000	-.556
d3	.333	.043
d4	1.000	1.000

2.7. Asimetría y curtosis

2.7.1. Asimetría

Como indicador del equilibrio simétrico en la distribución de datos alrededor de la Media, se utiliza la «Asimetría» (As: ing. *Skewness*). Para concentrar los datos alrededor de la Media y normalizar la variación, primero hay que convertir los datos en «Puntuaciones estandarizadas», que consiste en calcular la diferencia con respecto a la Media y dividirla por la Desviación típica. Luego elevamos todos los puntos al cubo y, finalmente, calculamos la Media de estos valores convertidos:

$$As = \Sigma i \, [(Xi - M) / DT]^3 / N$$

donde M es Media, DT es Desviación Típica y N es Número de datos (X)

Los puntos estandarizados, (Xi – M) / DT, presentan valores positivos cuando se trata de los datos más grandes de la Media, y valores negativos cuando los datos más pequeños de la Media, de modo que sus valores elevados a 3 también llevan el mismo signo de positivo o de negativo, por ejemplo $2^3 = 8$; $(-2)^3 = -8$. La Asimetría positiva significa que la distribución de los datos presentan una inclinación a la derecha, y la negativa a la izquierda. La Asimetría presenta valores más o menos normalizados por utilizar los puntos estandarizados, pero su rango no es [-1 ~ 1]. Obsérvese que d3 da el valor de 1.465

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

Fila	Media	Asimetría
d1	12.400	.367
d2	5.800	-.192
d3	2.800	1.465
d4	1.800	-.363

(*) Asimetría normalizada

Para normalizar el valor de Asimetría (As), proponemos utilizar su valor máximo As.max. La fórmula propuesta de de «Asimetría Normalizada» (As.N) es:

$$As.N = As / As.max$$

Seguidamente buscamos el máximo valor de la Asimetría As.max que se presenta con K como el único valor positivo existente dentro del conjunto {K, 0, 0, 0, 0}, y en tal caso la As llega a As.max.

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

$$= [(N\,M - M)^3 + (-M)^3 (N - 1)] / (N\,DT^3) \quad \leftarrow \text{ K = N M}$$
$$= [(M\,(N - 1))^3 + (-M)^3 (N - 1)] / (N\,DT^3) \quad \leftarrow \text{ M, al exterior}$$
$$= [(M\,(N - 1))^3 + (-1)^3\,M^3 (N - 1)] / (N\,DT^3) \quad \leftarrow (-M)^3 = (-1)^3\,M^3$$
$$= [(M\,(N - 1))^3 - M^3 (N - 1)] / (N\,DT^3) \quad \leftarrow (-1)^3 = 1$$
$$= [M^3\,(N - 1)^3 - M^3 (N - 1)] / (N\,DT^3) \leftarrow M^3 \text{ es común}$$
$$= M^3 (N - 1)\,[(N - 1)^2 - 1] / (N\,DT^3) \leftarrow M^3 (N - 1) \text{es común}$$
$$= M^3 (N - 1)\,(N^2 - 2N + 1 - 1] / (N\,DT^3) \quad \leftarrow \text{ desarrolar } (N - 1)^2$$
$$= (M^3 (N - 1)\,N^2 - 2N / (N\,DT^3) \quad \leftarrow \text{ 1 - 1 = 0}$$
$$= (M^3 (N - 1)\,N\,(N - 2) / (N\,DT^3) \quad \leftarrow \text{ N, al exterior}$$
$$= (M^3 (N - 1)\,(N - 2) / DT^3 \quad \leftarrow \text{ N / N = 1}$$

Como hemos visto anteriormente (→Desviación Típica Normalizada), la DT en el conjunto {K, 0, 0, 0, 0} llega al valor siguiente:

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

Entonces,

$$As.max = M^3 (N - 1)\,(N - 2) / (M\,(N - 1)^{1/2})^3$$
$$= M^3 (N - 1)\,(N - 2) / M^3\,(N - 1)^{3/2}$$
$$= (N - 2) / (N - 1)^{1/2}$$

De modo que la «Asimetría Normalizada» (As.N) es:

$$As.N = As / As.max = As * (N - 1)^{1/2} / (N - 2)$$

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

$$AT.max = \{[(N\,M - M)^3 + (-M)^3 (N - 1)] / N\}^{1/3} \quad \leftarrow \text{ K = N M}$$

$= \{[(M\ (N - 1))^3 + (-M)^3\ (N - 1)] / N\}^{1/3}$ ← M al exterior

$= \{[(M\ (N - 1))^3 + (-1)^3\ M^3\ (N - 1)] / N\}^{1/3}$ ← $(-M)^3 = (-1)^3\ M^3$

$= \{[(M\ (N - 1))^3 - M^3\ (N - 1)] / N\}^{1/3}$ ← $(-1)^3 = 1$

Las tablas siguientes muestran los valores de Media, Asimetría y Asimetría Norm[alizada] (As.N). La última (As.N.) tiene el rango de [-1 ~ 1]

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

Fila	Media	Asimetría	Asimetría Norm.
d1	12.400	.367	.245
d2	5.800	-.192	-.128
d3	2.800	1.465	.977
d4	1.800	-.363	-.242

(*) Grado de balanza

Calculamos el diferencias positivas (Ps) con respecto a la Mediana y el de las diferencias negativos (Ng) y con los dos valores calculamos el valor contrastivo de los dos, a lo que damos el nombre de «Índice de Balanza: IB):

IB = (Ps - Ng) / (Ps + Ng)

Por ejemplo, en d1, la Mediana es 12:

P = |19 – 12| + |14 — 12| = 7 + 2 = 9

N = |10 – 2| + |7 – 12| = 2 + 5 = 7

IB = (9 - 7) / (9 + 7) = .125

X	Mediana	Ps	Ng	IB	Asimetría
d1	12	9	7	.125	.367
d2	7	7	13	-.300	-.192
d3	1	11	2	.692	1.465
d4	2	2	3	-.200	-.363

La Asimetría (As) muestra la dirección de la desviación con respecto a la media, mientras que el Índice de Balanza (IB) muestra el equilibrio de los datos positivos y negativos con respecto a la Mediana.

2.7.2. Curtosis

Como indicador de lo puntado que está la distrubición concentrada se utiliza «Curtosis» (Cu), que se define[7]:

[7] Existen otras definiciones. Aqui/ seguimos a Shiba y otros (1984:145).

$$Cu = \Sigma i\, [(Xi - M) / DT]^4 / N$$

donde X es el dato, M es Media, DT es Desviación Típica, N es el Número de los datos. Se utiliza el exponente a los Puntos Estandarizados: (Xi – M) / DT. Explicaremos los detalles más adelante. De los Puntos Estandarizados, véase el capítulo siguiente.

(*) Curtosis normalizada

De la misma manera que la «Asimetría Normalizada», definimos la «Curtosis Normalizada» (Cu.N.) de manera siguiente, con dos valores «Curtosis» (Cu) y el Máximo de Curtosis (Cu.max):

$$Cu.N. = Cu / Cu.max$$

El Máximo de Curtosis (Cu.max) se presenta en una distribución unificada como {K, 0, 0, 0, 0}. En esta distribución dentro de Xi – M, se presenta una vez sin más K – M y los casos restantes (N – 1) dan solo M^4:

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

$= [(N\,M - M)^4 + M^4 (N - 1)] / (N\,SD^4)$ ← K = N M
$= [(M (N - 1))^4 + M^4 (N - 1)] / (N\,SD^4)$ ← M, al exterior
$= [M^4 (N - 1)^4 + M^4 (N - 1)] / (N\,SD^4)$ ← M^4 es común
$= M^4 (N - 1) [(N - 1)^3 + 1] / (N\,SD^4)$ ← $M^4 (N - 1)$ es común
$= M^4 (N - 1) (N^3 – 3N^2 + 3N - 1 + 1] / (N\,SD^4)$ ← desarrollar $(N - 1)^3$
$= M^4 (N - 1) (N^3 – 3N^2 + 3N) / (N\,SD^4)$ ← 1 - 1 = 0
$= M^4 (N - 1)\,N\,(N^2 – 3N + 3) / (N\,SD^4)$ ← N al exterior
$= M^4 (N - 1) (N^2 – 3N + 3) / SD^4$ ← N es común

Como hemos visto anteriormente (→Desviación Típica Normalizada), la DT en el conjungo {K, 0, 0, 0, 0} llega al valor siguiente:

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

Entonces,

Ku.max $= M^4 (N - 1) (N^2 – 3N + 3) / (M (N - 1)^{1/2})^4$ ← ver arriba
$= M^4 (N - 1) (N^2 – 3N + 3) / [M^4 (N - 1)^2]$ ← $M^4 (N - 1)$
$= (N^2 – 3N + 3) / (N - 1)$

De modo que «Curtosis Normalizada» (Cu.N.) es:

$$Cu.N. = Cu / Cu.max = Cu * (N – 1) / (N^2 – 3N + 3)$$

X	v1	v2	v3	v4	v5
d1	12	21	16	9	14
d2	13	9	12	2.00	3
d3	2	2	3	14	3
d4	2	3	4	5	5

Fila	Media	Curtosis	Cu.N.
d1	14.400	2.114	.650
d2	7.800	1.281	.394
d3	4.800	3.203	.986
d4	3.800	1.628	.501

(*) Varianza, asimetría, curtosis

Varianza (Vr), Asimetría (As) y Curtosis (Cu) presenta los grados de Dispersión, Deformación y Agudeza de la distribución de los datos. En cada fórmula se encuentra unos elementos comunes, $(Xi - M)]^{É}N$ (E=2, 3, 4)[8]:

$$Vr = \Sigma i\,[(Xi - M)]^2 / N$$
$$Sk = \Sigma i\,[(Xi - M) / DT]^3 / N$$
$$Ku = \Sigma i\,[(Xi - M) / DT]^4 / N$$

donde M es Media, N es Número, DT es Desviación Típica de los datos. Comparemos las distribuciones de estos valores y procesos de calculación. La tabla inferior izquierda es de los datos (X). La tabla inferior derecha muestra su desviación con respecto a la Media (D):

X	v1	v2	v3	v4	v5	M	DT
d1	10	19	14	7	12	12.40	4.03
d2	11	7	10	0	1	5.80	4.53
d3	0	0	1	12	1	2.80	4.62
d4	0	1	2	3	3	1.80	1.17

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

La tabla siguiente (S) es de los puntos estandarizados: (X－M)/DT. La tabla (S^2) es de los puntos cuadrados de S:

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

En la primera tabla observamos que las Medias de todas las filas resultan 0, por la razón de que el numerador de los Puntos Estandarizados es la Desviación (Xi – M), cuya Suma y, lógicamente, Media son nulas, como se observa en la tabla (D). En la segunda tabla (S^2), todas las Medias son 1, por la razón de que tanto su numerador como su denominador son iguales a la Varianza

[8] Vea/se Shiba y Haebara (1990: 34-35).

(Vr):

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

De esta manera, ni las Medias de Puntos Estandarizados ni las de sus cuadrados sirven para indicar el modo de distribución. Ahora veamos los casos de Puntos Estandarizados de orden 3 (S^3) y de orden 4 (S^4):

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

Efectivamente presentan sus propios valores diferentes. La Media de S^3 corresponde a la Asimetría (As), y la de S4 a la Curtosis (Cu). Tantos sus valores mismos como el orden de los valores se difieren una de otra: Cu: 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

2.8. Distinción y oposición

2.8.1. Grado Distintivo

Por ejemplo, encontramos las letras <i> y <j> en un documento antiguo de español, ambas representantes del mismo fonema /í. Si la frecuencia (F) de cada una son F(i) = 32 y F(j) = 2, significa que en casi todos los casos se utilizan la letra <i> y consideramos que su grado de distinción es alto. Por otra parte, cuando F(i) = 32 y F(j) = 28, significa que el grado de la distinción sería bajo y las dos letras se encontrarían en una situación llamada de variación casi libre. Definimos el «Grado Distintivo» (GD) como:

$$GD(i, j) = [F(i) - F(j)] / F(i)$$

donde, cuando F(i) = F(j), GD es cero, y cuando F(j) es cero, GD es uno.

Cuando hay otros elementos más, por ejemplo <y>, calculamos la Distinción de manera siguiente:

GD(i : j, y) = {F(i) – [F(j) + F(y)] } / F(i)

Generalizando la fórmula, en un conjunto de Fn = F(1, 2, …, n), el grado de Distinción de F(1) es:

GD(F(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)

Si F(1) es Máximo de F(1, 2, …, n), es decir, Max(Fn):

DD(Max(Fn)) = 2 – Sum(Fn) / Max(Fn)

2.8.2. Grado Opositivo

El Grado Distintivo se convierte negativo, cuando el valor máximo F(1) = Max(Fn) es menor que la Suma de los restantes, [F(2) + F(3) + … F(n)]. Por esta razón, proponemos una nueva fórmula de «Grado Opositivo» (GO), donde utilizamos un tipo contrastivo de los dos valores:

GO(i, j) = [F(i) – F(j)] / [F(i) + F(j)]
= [F(i) – F(j)] / Sum(Fn)

Generalizando la fórmula, en F(1, 2, …, n), la Oposiicón de F(1)es:

GO(1) = {F(1) – [F(2) + … F(n)] } / {F(1) + [F(2) … F(n)] }
= {F(1) – [Sum(Fn) – F(1)] } / Sum(Fn)
= [2 F(1) – Sum(Fn)] / Sum(Fn)
= 2 F(1) / Sum(Fn) – 1

Si F(1) es el Máximo de F(1, 2, …, n), Max(Fn):

GO(Max(Fn)) = 2 Max(Fn) / Sum(Fn) – 1

Recomendamos que cuando se trata de los datos cuyo Máximo es mayor que la suma de los restantes, utilicen la Distinción y cuando no, la Oposición.

(#) Las letras <u> y <v> en el español de los siglos XVI, XVI y XVII.

Los estudios previos explican que las dos letras, <u> y <v> se utilizaban sin distinción en los libros publicados en la España de los siglos XV, XI y XVII. Hemos calculado las frecuencias de cada letra en los seis libros: Nebrija (Nb), Rojas (Rj), Lazarillo (Lz), Cervantes (Cv), Quevedo (Qv) y Gracián (Gc).

Hemos hecho recuento de las primeras 20.000 letras de cada obra:

Letra	1.Nb	2.Rj	3.Lz	4.Cv	5.Qv	6.Gc	Total
<u>	949	820	1040	1250	1051	849	5959
<v>	165	139	191	194	209	402	1300
Distinción	0.826	0.830	0.816	0.845	0.801	0.527	0.782

Es cierto que el grado de distinción no es muy alto en estas obras. Sin embargo, si calculamos los grados distintivos con la clasificación de la posición con respecto a los fonemas vocálicos (V) y consonánticos (C) en su contexto inmediato, observamos altos grados en casi todos los casos.。

Posición	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

Ciertamente hay unos libros donde presentan valores reducidos de distinción, marcados en la Tabla. No obstante, no perdamos de vista la tendencia general de la alta distinción entre los usos de las dos letras tratadas.

(*) Búsqueda

Para observar los valores característicos de la matriz, cambiamos el color del número y el del fondo. Por ejemplo:

X	v-1	v-2	v-3	v-4	v-5
d-1	10	40	70	50	20
d-2	20	40	60	50	20
d-3	100	400	700	500	200

Condición: Más de A, A es la Media

Resultado:

Fila	v-1	v-2	v-3	v-4	v-5
d-1	10	40	70	50	20
d-2	20	40	60	50	20
d-3	100	400	700	500	200

(*) Medidas estadísticas positivas

La tabla inferior media muestra la Suma (S), Número de datos (N) y Media (M) de las filas, mientras que la tabla derecha, los mismos valores de los datos de valores positivos a exclusión de cero (0). Podemos comprobar que las dos sumas, S y PS, son iguales, pero los Números (N, PN) y las Medias (M, PM) son diferentes. Por ejemplo, para calcular las puntuaciones de estudiantes de español, consideramos tanto la Media de los puntos de pruebas (M) como las asistencias (N). Si sumamos M y N, un día que un estudiante no ha podido asistir a la clase, el punto de la prueba puede afectar en exceso a la Media. En tal caso podemos considerar la utilidad de «Medidas Estadísticas Positivas» (MEP).

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

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

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

En el análisis de datos linguísticos también tratamos unas tablas grandes, donde se encuentrasn muchos casos de datos no relevantes a las variables, por ejemplo, signos de acento en los documentos medievales. También se pueden calcular todas las medidas tratadas en este capítulo en su tratamiento positivo.

El programa hace el cálculo de Medidas Positivas por medio de una función Pos(X), que devuelve un vector columna con la dimensión de número de datos positivos (NDP), cula formulación matricial es:

$$NDP = Sum(D(X_{n1}, X_{n1}))$$

es decir, dividimos un vector por el mismo vector y obtenemos la secuencia de 1 con casos de cero (0). Como hemos definido 0 / 0 = 0, esta operación es posible. Finalmente con la función Sum(X), obtenemos la suma de los valores 1, que es la dimensión del vector reducido.

(*) Medidas grupales

Calculamos las medidas estadísticas no en filas sino dentro de grupo de filas. Tabla inferior derecha muestra la Suma de los grupos a, b, c:

d1	1	2	3	Group
1	5	2	7	a
2	3	3	2	b
3	2		2	b
4	4	2	2	c
5	2	4	3	c
6	1	8	7	c

Suma	1	2	3
a	5	2	7
b	5	3	4
c	7	14	12

3. Puntuación

Con frecuencia nos vemos en la necesidad de convertir las tablas de frecuencia de datos de acuerdo con determinadas reglas. Vamos a estudiar las distintas reglas que se han propuesto y las que nosotros proponemos.

En este capítulo consideramos los Puntuación de casillas de las tablas que consitituyen la matriz de datos. Se utilizan el término «Frecuencia» en la forma de «Frecuencia absoluta», «Frecuencia relativa», etc., y nosotros utilizamos en su lugar «Puntuación» (ing. *score*), en forma de «Puntuación Absoluts», «Puntuación Relativa», etc, puesto que incluimos en este concepto común varios «Puntuación» que no son necesariamente Frecuencias.

3.1. Puntuación relativa

El problema que presenta los valores de frecuencia absoluta es que la comparación o la evaluación de los valores resulta difícil por la diferencia de escala que hay entre filas y entre columnas. Por ejemplo el valor 11 de d1 y el 10 de d2 no son comparables por la diferencia de sumas de d1 (=62) y d2 (=29).

X	v1	v2	v3	v4	v5	S
d1	10	19	14	7	12	62
d2	11	7	10	0	1	29
d3	0	0	1	12	1	14
d4	0	1	2	3	3	9
T	21	27	27	22	17	114

En esta situación se recurre a los «Puntuación relativa» (PR), que son ratios. Se calcula por la división de los datos por la Suma (S) de conjunto en cuestión[9]. Cuando X = 0, PR resulta que es el Mínimo y cuando X = Suma, PR llega al Máximo 1.

PR = X / S

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

(1) Puntuación relativa de fila y de columna

Los Puntuación relativa se obtienen tanto en filas (PR por fila: PRF) por la suma de filas (SF) como en columnas (PR por columna: PRC) por suma por columna (SC):

[9] Si multiplicamos los P.R por cien, obtenemos el porcentaje (%).

$PRF_{np} = X_{np} / SF_{n1}$
$PRC_{np} = X_{np} / SC_{1p}$

De esta manera, la tabla anterior se queda:

PRF	v1	v2	v3	v4	v5
d1	.16	.31	.23	.11	.19
d2	.38	.24	.34	.00	.03
d3	.00	.00	.07	.86	.07
d4	.00	.11	.22	.33	.33

PRC	v1	v2	v3	v4	v5
d1	.48	.70	.52	.32	.71
d2	.52	.26	.37	.00	.06
d3	.00	.00	.04	.55	.06
d4	.00	.04	.07	.14	.18

(2) Puntuación Relativa de Ambos y de Todo

Definimos los «Puntuación Relativa de Ambos» (PRA) de manera siguiente:

$PRA = 2\ X_{np} / (SF_{n1} + SC_{1p})$

donde utilizamos la Media Fraccional (MF) de los dos Puntuaciones relativas, de fila y de columna (SF_{n1} y SC_{1p}), puesto que se trata de una Media de dos valores resultados de división:

$$\begin{aligned} PRA &= MF(PRF_{np}, PRC_{np}) \\ &= MF(X_{np} / SF_{n1}, X_{np} / SC_{1p}) \\ &= (X_{np} + X_{np}) / (SF_{n1} + SC_{1p}) \\ &= 2\ X_{np} / (SF_{n1} + SC_{1p}) \end{aligned}$$

Por ejemplo, la Media Fraccional (MF) de d1: v1 = MF(10/62, 10/21) = (10 + 10) / (62 + 21) = 0.24. De esta manera consideramos tanto la ratio horizontal como la vertical al mismo tiempo.

PRA	v1	v2	v3	v4	v5
d1	.24	.43	.31	.17	.30
d2	.44	.25	.36	.00	.04
d3	.00	.00	.05	.67	.06
d4	.00	.06	.11	.19	.23

PRT	v1	v2	v3	v4	v5
d1	.09	.17	.12	.06	.11
d2	.10	.06	.09	.00	.01
d3	.00	.00	.01	.11	.01
d4	.00	.01	.02	.03	.03

Obtenemos los «Puntuación Relativa de todo» (PRT) por la división de la matriz dato por la Suma Total (ST) que es un escalar:

$PRT = X_{np} / ST$

Los «Puntuación Relativa», como la Ratio o el Porcentaje, presenta un problema de dar unos valores muy reducidos al tratar unas tablas de dimensión

grande, puesto que la base de división que es la Suma del conjunto suele ser de cantidad considerable. Esta tendencia es especialmente notable en los «Puntuación Relativa de todo» (PRT).

(*) Valores prominentes

[1] Valores Prominentes en fila y columna

Proponemos utilizar «Puntuación Relativa Prominentes» (PRP) que presentan su prominencia dentro del conjunto de números. Los explicamos con un valor de una casilla, por ejemplo, la de d1: v1 (=10):

X_{np}	v1	v2	v3	v4	v5	SF_{n1}	P
d1	10	19	14	7	12	62	5
d2	11	7	10	0	1	29	5
d3	0	0	1	12	1	14	5
d4	0	1	2	3	3	9	5
SC_{1p}	21	27	27	22	17	114	
N	4	4	4	4	4		20

donde SF_{n1} es Suma fila, P es el número de columnas, $SC_{1p,}$ Suma columna, N, número de filas. Comparemos d1: v1 con el resto de la fila (SF – X = 62 – 10 = 52). Ahora no comparamos directamente el valor de d1:v1 (=10) con el resto (=52), sino multiplicamos el valor en cuestión por el número de miembros restantes (=5 - 1 = 4), para equilibrar la escala de los dos términos comparados: 1 contra 4. Ahora comparamos (P – 1) X con (SF – X) y para relativizar el valor de (P – 1)) entre la suma de los dos términos comparados, formulamos la derivación siguiente de «Puntuación Relativa Prominentes en Fila» (PRPF), que tiene el rango de [0.0 ~ 1.0] con el Mínimo (0,0) cuando X = 0, y el Máximo cuando SF = X:

$$PRPF = (P - 1)\, X_{np} \;/\; [(P - 1)\, X_{np} + (SF_{n1} - X_{np})]$$
$$= (P - 1)\, X_{np} \;/\; [(P - 2)\, X_{np} + SF_{n1}]$$

De la misma manera, los «Puntuación Relativa Prominentes en Columna» (PRPC) es:

$$PRPC = (N - 1)\, X_{np} \;/\; [(N - 1)\, X_{np} + (SC_{1p} - X_{np})]$$
$$= (N - 1)\, X_{np} \;/\; [(N - 2)\, X_{np} + SC_{1p}]$$

PRSF	v1	v2	v3	v4	v5
d1	.43	.64	.54	.34	.49
d2	.71	.56	.68	.00	.13
d3	.00	.00	.24	.96	.24
d4	.00	.33	.53	.67	.67

PRSC.	v1	v2	v3	v4	v5
d1	.73	.88	.76	.58	.88
d2	.77	.51	.64	.00	.16
d3	.00	.00	.10	.78	.16
d4	.00	.10	.19	.32	.39

Los «Puntuación Relativa» suelen dar resultados reducidos de acuerdo con el aumento de los miembros, pero «Puntuación Relativa Prominentes» como se observa en las tablas, mantienen los valores destacados a pesar del aumento de números de filas (N) y de columnas (P), debido a que en sus fórmulas aparecen N y P tanto en el numerador como en el denominador.

[2] Valores Prominentes en Ambas y en Todo

Para derivar los «Puntuación Relativa Prominentes en Ambos» (PRPA) utilizamos la operación de la Media Fraccional (MF) de los Puntuación Relativa en Fila y en Columna:

$$\begin{aligned} PRPA &= MF\,(PRPF, PRPC) \\ &= [(P-1)\,X_{np} + (N-1)\,X_{np}] \\ &\quad / \{[(P-2)\,X_{np} + SF_{n1}] + [(N-2)\,X_{np} + SC_{1p}]\} \end{aligned}$$

$$= (P + N - 2)\,X_{np} / [(P + N - 4)\,X_{np} + SF_{n1} + SC_{1p}]$$

Finalmente, en el cáculo de los «Puntuación Relativa Prominentes en Todo» (PRPT) utilizamos la Suma de los restantes de la matriz (S – X) y número de restantes (N*P – 1):

$$\begin{aligned} PRPT &= (N\,P - 1)\,X_{np} / [(N\,P - 1)\,X_{np} + (S - X_{np})] \\ &= (N\,P - 1)\,X_{np} / [(N\,P - 2)\,X_{np} + S] \end{aligned}$$

$$\begin{aligned} PRST &= (N\,P - 1)\,X_{np} / [(N\,P - 1)\,X_{np} + (S - X_{np})] \\ &= (N\,P - 1)\,X_{np} / [(N\,P - 2)\,X_{np} + S] \end{aligned}$$

PRSM	v1	v2	v3	v4	v5
d1	.53	.72	.62	.41	.60
d2	.73	.54	.66	.00	.14
d3	.00	.00	.15	.88	.19
d4	.00	.17	.30	.46	.51

PRSA	v1	v2	v3	v4	v5
d1	.65	.79	.73	.55	.69
d2	.67	.55	.65	.00	.14
d3	.00	.00	.14	.69	.14
d4	.00	.14	.25	.34	.34

(#) Puntuación Relativa y Ratio por Mil: Grafías geminadas medievales

La tabla siguiente (O) muestra las frecuencias de grafías geminadas observadas en documentos notariales de Edad Media española:

O	1260	1280	1300	1320	1340	1360	1380	1400	1420	1440	1460	1480	1500
nn	550	66	143	57	1	2	2	4	4	1	0	2	30
ll	2310	1166	4524	1354	243	367	325	571	902	217	439	589	776
rr	625	327	1563	846	109	309	283	533	290	181	152	249	273

A partir de la tabla anterior, se deriva la tabla de Puntuación Relativa en Fila (R):

PRF	1260	1280	1300	1320	1340	1360	1380	1400	1420	1440	1460	1480	1500
nn	.638	.077	.166	.066	.001	.002	.002	.005	.005	.001		.002	.035
ll	.168	.085	.328	.098	.018	.027	.024	.041	.065	.016	.032	.043	.056
rr	.109	.057	.272	.147	.019	.054	.049	.093	.051	.032	.026	.043	.048

La tabla siguiente (W) muestra un vector fila de número de palabras de los documentos pertenecientes a los años:

W	1260	1280	1300	1320	1340	1360	1380	1400	1420	1440	1460	1480	1500
&	62549	29396	114499	44040	6000	11732	10506	19276	27990	8131	15952	20792	27048

Ahora para obtener la tabla de Ratio por Mil, formulamos la operación matricial siguiente:

$$M_{np} = O_{np} * 1000 / W_{1p}$$

M	1260	1280	1300	1320	1340	1360	1380	1400	1420	1440	1460	1480	1500
nn	9	2	1	1	0	0	0	0	0	0	0	0	1
ll	37	40	40	31	41	31	31	30	32	27	28	28	29
srr	10	11	14	19	18	26	27	28	10	22	10	12	10

En los Puntuación Relativa (R) se comparan dentro de los miembros de la columna, mientras que en la Ratio por Mil (M) se observan las ocurrencias con respecto de la totalidad de palabras recogidas y facilita el estudio cronológico comparado. El descenso notable de <nn> en el signo XIV se debe al uso de la lineta como signo de abreviación, que es el origen de la tilde de la eñe actual.

(*) Valor Relativo y Valor Contrastivo

Cuando se comparan dos valores X e Y, se calcula la diferencia X – Y o se calcula la proporción X / Y. También es útil calcular una fórmula como X / (X + Y) o Y / (X + Y), que son «Valores Relativos» (VR), como hemos visto en esta sección.

$$VR = X / (X + Y)$$

Un Valor Relativo (VR) tiene Rango de [0.0 ~ 1.0] y se vuelve en el Mínimo cero (0) cuando X = 0, y llega al Máximo cuando Y = 0. El punto medio o Mitad (= 0.5) se obtiene cuando X = Y.

Por otra parte también es posible pensar en la fórmula:

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

que denominamos «Valor Contrastivo» (VC). Entre el Valor Relativo (VR) y el Valor Contrastivo (VC) se establece la relación:

$$2\ VR - 1 = VC$$

puesto que:

$$\begin{aligned} & 2\ VR - 1 \\ &= 2X / (X + Y) - 1 \\ &= 2X / (X + Y) - (X + Y) / (X + Y) \\ &= [2X - (X + Y)] / (X + Y) \\ &= (X - Y) / (X + Y) \\ &= VC \end{aligned}$$

La tabla inferior izquierda demuestra los Puntuación Relativa en Fila (R_{np}); y la derecha, su conversión en los valores contrastivos (C_{np}):

$$C_{np} = 2\ R_{np} - 1$$

R_{np}	v1	v2	v3	v4	v5
d1	.16	.31	.23	.11	.19
d2	.38	.24	.34	.00	.03
d3	.00	.00	.07	.86	.07
d4	.00	.11	.22	.33	.33

C_{np}	v1	v2	v3	v4	v5
d1	-0.68	-0.39	-0.55	-0.77	-0.61
d2	-0.24	-0.52	-0.31	-1.00	-0.93
d3	-1.00	-1.00	-0.86	0.71	-0.86
d4	-1.00	-0.78	-0.56	-0.33	-0.33

El Rango de Valor Contrastivo es [-1.0 ~ 1.0] y las cifras negativas y positivas se contrastan alrededor del punto medio (Mitad) que es 0.0. Su Mínimo (-1.0) se presenta cuando X =0, su Máximo (1.0), cuando Y=0 y el punto medio, cuando X = Y. Las condiciones de sus apariciones son las mismas a las del Valor Relativo, pero el Rango es distinto.

Los conceptos de «Valor Relativo» y el de «Valor Contrastivo» son útiles para comprender la esencia de los dos valores en comparación. El Valor Relativo es lo mismo que una Ratio. La Ratio sin embargo se calcula por la división de X por Suma, donde se esconden el valor de X y el Resto. A veces los perdemos de vista, pero conviene saber que en la Ratio se compara entre X y X más el Resto.

Por otra parte, ¿qué significado tiene la comparación entre X – Y y X + Y en el Valor Contrastivo? Aparentemente no encontramos mucho significado, pero acabamos de ver su capacidad de contrastar el resultado en valor negativo y positivo. Conviene transformar la fórmula de Valor Contrastivo (VC) de manera siguiente:

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

Asi, de esta manera llegamos a comprender que con el Valor Contrastivo estamos observando la diferencia entre los dos «Valores Relativos», uno de X y otro de Y.

(#) Preposición española en la edad medieval y moderna

Las dos tablas siguientes muestra la distribución espacio-temporal tanto de «pora» como de «para», en forma de Valor Relativo. La ultima tabla presenta los valores contrastivos de las dos formas y podemos observar sus cambios en dos dimensiones al mismo tiempo.

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

Frecuencia Relativa de «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

Frecuencia Relativa de «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

Frecuencia contrastiva de «pora» y «para»

Los números rojos son negativos, inferior a 0, y el color gris representn la barra de magnitud.

Llamamos «Puntuación normalizada» (PN) a los Puntuación convertidos cuya Suma total resulte 1.

(3) Totalización por suma total

Para el método más sencillo para llegar a unos Puntuación Totalizados se recurre a la división de los datos Xnp por la Suma total:

$$NS.S_{np} = X_{np} / T$$

X_{np}	v1	v2	v3	v4	v5	Sh
d1	10	19	14	7	12	62
d2	11	7	10	0	1	29
d3	0	0	1	12	1	14
d4	0	1	2	3	3	9
Sv	21	27	27	22	17	T:114

NS.S	v1	v2	v3	v4	v5	Sh
d1	.088	.167	.123	.061	.105	.544
d2	.096	.061	.088	.000	.009	.254
d3	.000	.000	.009	.105	.009	.123
d4	.000	.009	.018	.026	.026	.079
Sv	.184	.237	.237	.193	.149	1.000

*Hemos consultado a Ikeda (1976: 121-123).

(4) Totalización por media fraccional

Las fórmulas y la tabla de los «Puntuación Totalizados por Media Fraccional» (PN.MF) son las siguientes:

$$Wnp = 2\ Xnp / (Shn1 + Sv1p)$$
$$PN.MF = Wnp / Sum(Wnp)$$

X_{np}	v1	v2	v3	v4	v5	Sh
d1	10	19	14	7	12	62
d2	11	7	10	0	1	29
d3	0	0	1	12	1	14
d4	0	1	2	3	3	9
Sv	21	27	27	22	17	T:114

PN.MF	v1	v2	v3	v4	v5	Sh
d1	.062	.109	.080	.043	.078	.371
d2	.112	.064	.091		.011	.279
d3			.012	.170	.016	.199
d4		.014	.028	.049	.059	.151
Sv	.174	.187	.212	.262	.164	1.000

(5) Totalización de Mosteller

Dividiendo la tabla por unos determinados valores, llegamos a obtener una tabla cuyas Sumas horizontales sean iguales y Sumas verticales sean iguales también. La Suma total es 1. La tabla así Totalizada (NS.Mos) sirve para comparar los Puntuación con bases comunes de la Suma horizontal (Sh) y de la Suma vertical (Sv). El método se llama «Totalización de Mosteller» y llamos la tabla resultante «Puntuación Totalizados de Monsteler» (PN.Mos):

X_{np}	v1	v2	v3	v4	v5	Sh
d1	10	19	14	7	12	62
d2	11	7	10	0	1	29
d3	0	0	1	12	1	14
d4	0	1	2	3	3	9
Sv	21	27	27	22	17	T:114

PS.Mos	v1	v2	v3	v4	v5	Sh
d1	.068	.091	**.043**	.007	.041	.250
d2	.132	.059	.053	.000	.006	.250
d3	.000	.000	.041	.162	.047	.250
d4	.000	.050	**.063**	.031	.106	.250
Sv	.200	.200	.200	.200	.200	1.000

Para obtener los «Puntuación Totalizados de Monster» (PN.Mos), cuyas Sumas horizontales sean iguales y Suma total sea 1, hay que dividir la matriz Xnp por la columna de Sumas horizontales (Sh) * número de filas (4):

$X^1np = Xnp / (Sh * 4)$

X^1np	v1	v2	v3	v4	v5	Sh
d1	.040	.077	.056	.028	.048	.250
d2	.095	.060	.086	.000	.009	.250
d3	.000	.000	.018	.214	.018	.250
d4	.000	.028	.056	.083	.083	.250
Sv	.135	.165	.216	.326	.158	1.000

Seguidamente, para obtener la fila de Sumas verticales iguales y la Suma total que sea 1, dividimos la matriz anterior X1np por la fila de Sumas verticales (Sv) de la misma * número de columnas (5):

$X^2np = Xnp^1 / (Sv * 5)$

X^2np	v1	v2	v3	v4	v5	Sh
d1	.060	.093	.052	.017	.061	.283
d2	.140	.073	.080	.000	.011	.304
d3	.000	.000	.017	.132	.023	.171
d4	.000	.034	.051	.051	.105	.242
Sv	.200	.200	.200	.200	.200	1.000

En este momento, la columna de Sumas horizontales cambian, de modo que hay que hacer de nuevo la división de la matriz resultante por la columna de Sumas horizontales (Sh) * 4. Seguimos practicando las mismas operaciones, naturalmente con programa, llegamos a los «Puntuación Totalizados de Monsteler» (PN.Mos).

Como hemos visto, los «Puntuación Totalizados de Monsteler» (PN.Mos) mantienen la igualdad tanto de las Sumas horizontales (Sh), como de las Sumas verticales (Sv), pueden cambiar de manera desproporcional los valores orginales, lo que podemos observar en Xnp y PS.Mos. El propósito de PN.Mos es ver las proporsiones de frecuencias con bases iguales de fila y de columna.

(#) Razones de Totalización: dos variantes de la letra «s», números de fallecidos y sobrevivientes

En español de Edad Media y Moderna, tanto la letra «s» poseía dos variantes: la ese corta <s> y la ese larga <ſ>. Se ha observado que la <s> aparece en la posición final de palabra (_#). En realidad, sin embargo, también se encuentra la <s> en la posición inicial (#_) y media (&_&) de palabra. La tabla inferior izquierda muestra la frecuencia de <s> y <ſ> en los primeros 20.000 letras de *Libro de Alexandre* (1300) y la tabla derecha, sus «Puntuación Totalizados por Suma» (PN.S):

/s/	#_	&_&	_#	Sh
<s>	62	2	593	657
<ſ>	314	412	109	835
Sv	376	414	702	1492

PN.S	#_	&_&	_#	Sh
<s>	.042	.001	.397	.440
<ſ>	.210	.276	.073	.560
Sv	.252	.277	.471	1.000

Podemos observar fácilmente la tendencia de aparición de <s> al final de palabra en una tabla de frecuencia absoluta así de dimensión pequeña y valores reducidos. También es útil la tabla de Puntuación Totalizados por Suma (PN.S). Cuando se trata de los datos grandes de dimensión y de valores, normalmente se recurre a las tablas de Puntuación Relativa, horizontales (PRh) y verticales (PRv)

PRh	#_	&_&	_#	Sh

PRv	#_	&_&	_#	Sh

<s>	.094	.003	.903	1.000
<ſ>	.376	.493	.131	1.000
Sv	.470	.496	1.033	2.000

<s>	.165	.005	.845	1.014
<ſ>	.835	.995	.155	1.986
Sv	1.000	1.000	1.000	3.000

Nos damos cuenta de que en la tabla de «Puntuación Relativa horizontales» (PRh), nuestra atención se fija en las filas y, efectivamente, observamos que la ese corta <s> se encuentra mayoritariamente en la posición final (_#: .903). Por otra parte, sin embargo, en la tabla de «Puntuación Relativa verticales» (PRv), también son notables las altas proporciones de la ese alta <ſ> en la posición inicial (#_) y media (&_&) de la palabra, lo cual no hemos notado en la tabla anterior (PRh). Tampoco podemos obtener la misma proporción de las dos variantes en PRh que en PRv, dividiendo .376 por .470 en PRh, que resulta .800, que es distinto de .835 de PRv.

La tabla siguiente muestra los «Puntuación Totalizados por Media Fraccional» (PN.MF):

PN.MF	#_	&_&	_#	Sh
<s>	.052	.002	.377	.430
<ſ>	.224	.285	.061	.570
Sv	.276	.286	.438	1.000

En esta tabla, se calcula la Media Fraccional de la <s> en la posición inicial (#_) tanto de la Ratio horizontal (62/657) como de la vertical (62/314) en forma de: (62 x 2) / (657+314) = .052.

Como ejemplo de mayor envergadura, supongamos que estamos ante una tabla de números de fallecidos y sobrevivientes de la epidemia de cólera en dos ciudades A y B. La tabla inferior derecha muestra los «Puntuación Relativa verticales». Por esta tabla, ¿podemos afirmar que el número de los fallecidos de la ciudad A (.032) es 2.7 veces más grande que el de la ciudad B (.012) por .032 / .12 ≒ 2.66? Si fuera así, al comparar las Ratios de sobrevivientes (.968, .988), la diferencia resulta nada destacable: .968 / .988 ≒ .980.

Xnp	Ciudad A	Ciudad B	Sh
Fallecidos	1300	250	1550
Sobrevivientes	**39000**	20000	59000
Sv	40300	20250	60550

PRv	A	B	Sh
F.	.032	.012	.045
S.	.968	.988	1.955
Sv	1.000	1.000	2.000

En realidad no debemos hacer la comparación de las Ratios entre los conjuntos cuyos números de miembros son diferentes. De ahí viene la necesidad de la Totalización de los datos. La tabla inferior izquierda (PN.S) muestra los «Puntuación Totalizados por Suma» y la derecha (PN.MF), los «Puntuación Totalizados por Media Fraccional»:

PN.S	Ciudad A	Ciudad B	Sh
Fallecidos	.021	**.004**	.026
Sobrevivientes	.644	.330	.974
Sv	.666	.334	1.000

PN.MF.	A	B	Sh
F.	.045	**.017**	.062
S.	.571	.367	.938
Sv	.616	.384	1.000

La tabla anterior izquierda PN.S es simplemente de los «Puntuación Totalizados por Suma». Como todos los Puntuación están divididos por la Suma total (60550), los valores son comparables. De esta manera podemos apreciar los detalles de Ratios, lo cual es difícil de hacer con los datos originales (Xnp). Sin embargo, la cifra de los fallecidos en la Ciudad B viene muy reducida por incluir el caso de sobrevivientes en la Ciudad A, que no tiene mucho que ver con la cifra de los fallecidos en la Ciudad B. En cambio, en el cálculo de los «Puntuación Totalizados por Media Fraccional» (PN.MF), para el caso de los fallecidos en la Ciudad B, se toma en cuenta tanto de la Ratio horizontal con los fallecidos en la Ciudad A, como de la Ratio vertical con los sobrevivientes de la Ciudad B, ambas relevantes para el caso de los fallecidos en la Ciudad B. La cifra resultante .017 es más convincente que el caso anterior (.004).

En los estudios de distintas disciplinas, no solamente en los estudios lingüísticos, se hacen comparaciones de los conjuntos de dimensiones diferentes. Sabemos que no se puede comparar directamente con frecuencias absolutas si la base es diferente. Entonces recurrimos a los Puntuación Relativa en forma de Ratio, Porcentaje (%), Por mil, Por millón, etc. No obstante, en el sentido riguroso de la palabra, la comparación de los Puntuación Relativa no es practicable, cuando las bases son diferentes. En un caso extremo, por ejemplo, sabemos a ciencia cierta que no tiene sentido la comparación entre 250/1000=25% y 3/10=30%. Entonces se cree que se puede hacer la comparación cuando las bases son "parecidas", "cercanas", por ejemplo, entre 25/400, 25/450. ¿Hasta donde se permite hacer la comparación? ¿No existe problema en la comparación de los datos cuyos bases tienen la diferencia de 1.5 o 2 veces más grande que otra, por ejemplo entre 30/100 y 40/200? Una de las razones importantes para recurrir a la Totalización está precisamente en que por ella obtenemos medios que posibilitan las comparaciones igualadas.

(6) Frecuencia por bloques

Cuando las frecuencias se distribuyen de manera sesgada, la suma no representa la tendencia general. Para conocer la tendencia general, pensamos aplicar un método que llamamos «Por bloques»[10].

Supongamos que las palabras que empiezan con *ff-* (*ffazer, ffijo,* etc.) se

[10] Hemos consultado a: Davies, Mark. 2006. *A Frequency Dictionary of Spanish. Core Vocabulary for Learners.* New York. Routledge. p.6-7.

encuentran en las líneas {2, 5, 8, 45, 50}. Dividimos el dato total en 5 bloques de tamaño igual de líneas: {1-10, 11-20, 21-30, 31-40, 41-50}. Entonces, las palabras que están en las líneas {2, 5, 8} pertenecen al grupo de {1-10}; y las de {45, 50}, al grupo de {41-50}. En cambio, los datos que presentan la distribución en líneas {2, 13, 25, 38, 43}, todas las palabras en cuestión aparecen en todos los 5 bloques. La diferencia entre 2 bloques y 5 bloques es grande, a pesar de que la totalidad es la misma (5).

El número de bloques de aparición varía según el número total de bloques. Es posible fijar de antemano el número total de bloques, por ejemplo, en 100. Sin embargo, si determinamos el número total de bloques en el número inferior y más próximo a la totalidad de líneas, se calcula el número de bloques de aparición de manera más precisa.

Por otra parte, al incrementar el número total de bloques, hay un punto a partir del cual no cambia el número de bloques de aparición. Este número puede ser seleccionado.

Las tres tablas siguientes muestran las vicisitudes históricas de las palabras con *f-, ff-, s-, ss-* iniciales. La primera tabla (F.A.) representa las frecuencias absolutas, la segunda, por 100 bloques y la tercera, por 1000 bloques:

F.A.	1200	1250	1300	1350	1400	1450
1.#f[^f]%	2340	3976	1781	2730	4090	3489
2.#ff%	100	1998	**2160**	886	153	127
3.#s[^s]%	3176	8435	4441	5998	8175	7030
4.#ss%	9	1301	2074	939	95	27

100 bloques	1200	1250	1300	1350	1400	1450
1.#f[^f]%	39	68	60	53	55	57
2.#ff%	12	**63**	50	45	35	17
3.#s[^s]%	39	70	62	53	55	58
4.#ss%	5	51	44	39	24	5

1000 bloques	1200	1250	1300	1350	1400	1450
1.#f[^f]%	134	290	180	160	184	215
2.#ff%	22	**243**	149	111	59	31
3.#s[^s]%	135	314	203	171	186	218
4.#ss%	5	140	138	90	36	7

En la primera tabla, el auge de la palabra con *ff-* inicial está en la franja de 1300 (2160), mientras que en las dos tablas siguientes, el auge se adelanta 50 años: 63 y 243 bloques en 1250. De esta manera, la frecuencia absoluta parece presentar la magnitud sesgada, mientras que las dos tablas por bloques muestran

la tendencia general, más fiable. El método por bloques es el cálculo donde se toma en consideración tanto la frecuencia como la dispersión.

3.2. Puntuación opositiva

Utilizamos el concepto de «Grado Opositivo» que hemos visto en el capítulo de las Medidas Estadísticas para formular los «Puntuación opositiva» (PO). La fórmula de «Puntuación opositiva en Fila» (POF) es:

$$\begin{aligned} POF &= [X_{np} - (SF_{n1} - X_{np})] / [X_{np} + (SF_{n1} - X_{np})] \\ &= (2\ X_{np} - SF_{n1}) / S_{n1} \\ &= 2\ X_{np} / SF_{n1} \ - \ 1 \end{aligned}$$

«Puntuación opositiva en Columna» (POC) es:

$$\begin{aligned} POC &= [X_{np} - (SC_{1p} - X_{np})] / [X_{np} + (SC_{1p} - X_{np})] \\ &= (2\ X_{np} - SC_{1p}) / SC_{1p} \\ &= 2\ X_{np} / SC_{1p} \ - \ 1 \end{aligned}$$

Para los «Puntuación opositiva en Ambos» (POA) utilizamos la Media Fraccional (MF) de POF y POC:

$$\begin{aligned} POA &= MF\ (POF, POC) \\ &= [(2\ X_{np} - SF_{n1}) + (2\ X_{np} - SC_{1p})] / (S_{n1} + SC_{1p}) \\ &= (4\ X_{np} - SF_{n1} - SC_{1p}) / (S_{n1} + SC_{1p}) \end{aligned}$$

«Puntuación opositiva en Todo» (POT) es:

$$\begin{aligned} POT &= [X_{np} - (S - X_{np})] / [X_{np} + (S - X_{np})] \\ &= (2\ X_{np} - S) / S \\ &= 2\ X_{np} / S \ - \ 1 \end{aligned}$$

donde SF_{n1} es el vector vertical de las sumas de filas, SC_{1p} es el vector horizontal de sumas de columnas, S es la Suma de toda la matriz (escalar).

POF	v1	v2	v3	v4	v5
d1	-.677	-.387	-.548	-.774	-.613
d2	-.241	-.517	-.310	-1.000	-.931
d3	-1.000	-1.000	-.857	.714	-.857
d4	-1.000	-.778	-.556	-.333	-.333

POC	1	2	3	4	5
d1	-.048	.407	.037	-.364	.412
d2	.048	-.481	-.259	-1.000	-.882
d3	-1.000	-1.000	-.926	.091	-.882
d4	-1.000	-.926	-.852	-.727	-.647

POA	1	2	3	4	5
d1	-.518	-.146	-.371	-.667	-.392
d2	-.120	-.500	-.286	-1.000	-.913
d3	-1.000	-1.000	-.902	.333	-.871
d4	-1.000	-.889	-.778	-.613	-.538

POT	v1	v2	v3	v4	v5
d1	-.825	-.667	-.754	-.877	-.789
d2	-.807	-.877	-.825	-1.000	-.982
d3	-1.000	-1.000	-.982	-.789	-.982
d4	-1.000	-.982	-.965	-.947	-.947

3.3. Puntuación proporcional

Como la frecuencia absoluta (tabla de Puntuación originales) y la frecuencia relativa (tabla de Puntuación Relativa) tienen sus propias características, a la hora de realizar la observación de datos, se complementan una con otra. Para comparar las frecuencias absolutas se utiliza la tabla de Puntuación Relativa para igualar la base de comparación. Sin embargo, a veces nos encontramos con la dificultad de encontrarnos con bases de comparación muy diferentes. Proponemos una solución fácil de este problema en forma de «Puntuación proporcional».

(1) Puntuación proporcional en fila y en columna

Por ejemplo, el dato de d1:v2 19 se convierte en el punto relativo de 0.31, obtenido de la división de 19 por la suma de los valores de la fila, 62, Así, 19 / 62 = .31. Por otra parte, d4:v4 3 da el punto relativo .33 por 3 / 9, que es superior al punto relativo del ejemplo anterior 19 / 62= .32. No obstante, según nuestra intuición el peso de 19 es mucho mayor que el de 3.

X	v1	v2	v3	v4	v5	SF
d1	10	**19**	14	7	12	62
d2	11	7	10	0	1	29
d3	0	0	1	12	1	14
d4	0	1	2	**3**	3	9
SC	21	27	27	22	17	114

Cuando se trata de la comparación de cifras con bases muy diferentes, si multiplicamos la cifra por la ratio, el resultado nos convence intuitivamente. Por ejemplo, damos a la cifra de d1:v2 19 el peso de la ratio de 19 / 62, y también damos a la cifra de d4:v4 3 el peso de 3 / 9. Por lo tanto proponemos una fórmula de «Puntuación proporcional» (PP), tanto de fila (PPF) com de columna (PPC):

$$PPF_{np} = X_{np}\ PR_{np} = X_{np}\ X_{np} / SF_{n1} = X_{np}^2 / SF_{n1}$$

PPF.: 0.0 (X=0) $\leqq$ 0.5 (X^2 = SF / 2) $\leqq$ X (X = SF)

$$PPC_{np} = X_{np}\ PR_{np} = X_{np}\ X_{np} / SC_{1p} = X_{np}^2 / SC_{1p}$$

PPC.: 0.0 (X=0) ≦ 0.5 (X^2 = SC / 2) ≦ X (X = SC)

donde X_{np} es la matriz de frecuencias absolutas, SF_{n1} es el vector vertical de Sumas de fila, Sc_{1p} es el vector horizontal de Sumas de columna. El punto ponderado es el Mínimo de 0 cuando X = 0, y llega al Máximo que es X, cuando X llega a la Suma, es decir cuando no hay más datos que el mismo.

PPF	v1	v2	v3	v4	v5
d1	1.61	**5.82**	3.16	0.79	2.32
d2	4.17	1.69	3.45	0.00	0.03
d3	0.00	0.00	0.07	10.29	0.07
d4	0.00	0.11	0.44	**1.00**	1.00

PRC.	v1	v2	v3	v4	v5
d1	4.76	13.37	7.26	2.23	8.47
d2	5.76	1.81	3.70	0.00	0.06
d3	0.00	0.00	0.04	6.55	0.06
d4	0.00	0.04	0.15	0.41	0.53

(2) Puntuación proporcional en Ambas y en Todo

Los «Puntuación proporcional de Ambas» (PPA) es la Media Fraccional (MF) de PPF y PPC:

PPA. = (X_{np}^2 + X_{np}^2) / (SF_{n1} + SC_{1p}) = 2 X_{np}^2 / (SF_{n1} + SC_{1p})

Para obtener los «Puntuación proporcional de Todo» (PPT) utilizamos la Suma (S) de toda la matriz. Los PPT suelen ser de poco valor por la división por la Suma de la matriz:

W.S.m. = X^2 / Sm.a.

PPA	v1	v2	v3	v4	v5
d1	2.41	8.11	4.40	1.17	3.65
d2	4.84	1.75	3.57	0.00	0.04
d3	0.00	0.00	0.05	8.00	0.06
d4	0.00	0.06	0.22	0.58	0.69

PPT	v1	v2	v3	v4	v5
d1	0.88	3.17	1.72	0.43	1.26
d2	1.06	0.43	0.88	0.00	0.01
d3	0.00	0.00	0.01	1.26	0.01
d4	0.00	0.01	0.04	0.08	0.08

(#) Apócope extrema en el español medieval

La caída la vocal <e> final de palabra, llamada «apócope» es normal en español cuando la vocal viene detrás de una consonantes dental o alveolar: *ciudad(e), papel(e), mes(e),* etc. En español medieval se daba incluso detrás de dos consonantes: *present(e), veint(e), adelant(e), part(e), est(e), end(e),* que se llama «apócope extrema». La tabla siguiente muestra las frecuencias de los casos de la apócope extrema (-CC) y la forma plena (-CCe) junto con su Sumas, Ratios y Puntuación proporcional de Fila (PPF) de las seis palabras indicadas en 1500 documentos notariales divididos en intervalo de 25 años:

Año	-CC	-CCe	Suma	Ratio de -CC	PPF

1075	2	2	4	.500	1.0
1100	7	5	12	.583	4.1
1150	15	5	20	.750	11.3
1175	10	15	25	.400	4.0
1200	25	68	93	.269	6.7
1225	70	173	243	.288	20.2
1250	101	361	462	.219	22.1
1275	228	605	833	.274	62.4
1300	137	418	555	.247	33.8
1325	165	315	480	.344	56.7
1350	102	358	460	.222	22.6
1375	189	312	501	.377	71.3
1400	239	623	862	.277	66.3
1425	52	283	335	.155	8.1
1450	74	535	609	.122	9.0
1475	48	374	422	.114	5.5
1500	45	749	794	.057	2.6
1525	7	386	393	.018	.1
1550		304	304		
1575		365	365		
1600		142	142		
1625		159	159		
1650		158	158		
1675		39	39		
Total	1,516	6754	8270	.183	277.9

Al observar los PPF notamos que la apócope extrema (-CC) es especialmente frecuente en la segunda mitad del siglo XIII y en el siglo XIV. Los estudios anteriores informan que el fenómeno era frecuente en la primera mitad del siglo XIII, pero los casos de los documentos notariales presentan las cifras destacadas en épocas posteriores.

En la tabla en la fila de 1150 tenemos la ratio de –CC alta de .750, puesto que en esta época se encuentran pocos documentos españoles por el empleo usual de latín. Por esta razón los Puntuación proporcional se reducen (11.3) por la poca frecuencia de –CC (15) en dicha fila.

3.4. Puntuación limitada

Denominamos «Puntuación limitada» (PL) a los que calculamos dentro del Rango de [0.0 ~ 1.0], con el Mínimo de 0.0 y el Máximo de 1.0. Para el cálculo de PL preparamos columnas de Mínimos Horizontales (MinH) y de

Máximos Horizontales (MaxH) y filas de Mínimos Verticales (MinV) y de Máximos Verticales (MaxV):

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

Por ejemplo, d1:v1 (= 10) se sitúa dentro del conjunto de {10, 19, 14, 7, 12}, cuyo Rango es 19 – 7 = 12, en un punto donde se avanza 3 Puntuación con respecto al Mínimo: 10 – 7 = 3. Para indicar el sitio que ocupa el valor de 10 dentro del Rango, hacemos el cálculo de (10 – 7) / (19 – 7) = 3 / 12 = .25, lo cual dice que el 10 ocupa un punto de 25% dentro del conjunto.

LS = (X – Min) / (Max — Min)
LS: 0.0 (X = Min) ≦ 0.5 (X = (Max - Min) / 2) ≦ 1.0 (X = MaX)

(1) Puntuación limitada de fila y de columna

Las fórmulas de «Puntuación limitada de Fila» (PLF) y «Puntuación limitada de Columna» (PLC) son:

$PLF_{np} = (X_{np} – MinH_{n1}) / (MaxH_{n1} – MinH_{n1})$
$PLC_{np} = (X_{np} – MinC_{1p}) / (MaxC_{1p} – MinC_{1p})$

donde $MinH_{n1}$ es columna de los Mínimos horizontales, $MaxH_{n1}$ es columna de los Máximos horizontales; $MinC_{1p}$ es fila de los Mínimos verticales, $MaxC_{1p}$ es fila de los Máximos verticales.

LSR.	v1	v2	v3	v4	v5
d1	0.25	1.00	0.58	0.00	0.42
d2	1.00	0.64	0.91	0.00	0.09
d3	0.00	0.00	0.08	1.00	0.08
d4	0.00	0.33	0.67	1.00	1.00

LSC.	v1	v2	v3	v4	v5
d1	0.91	1.00	1.00	0.58	1.00
d2	1.00	0.37	0.69	0.00	0.00
d3	0.00	0.00	0.00	1.00	0.00
d4	0.00	0.05	0.08	0.25	0.18

(2) Puntuación limitada de ambas y de todo

Los «Puntuación limitada de Ambas» (PLA) es la Media Fraccional (MF) de PLF y PLC.

$PLA = [(X_{np} – MinH_{n1}) + (X – MinC_{1p})]$

$$/ [(MaxH_{n1} - MinH_{n1}) + (MaxC_{1p} - MinC_{1p})]$$
$$= (2\ X_{np} - MinH_{n1} - MinC_{1p})$$
$$/ (MaxH_{n1} + MaxC_{1p} - MinH_{n1} - MinC_{1p})$$

Para el cálculo de «Puntuación limitada de Todo» (PLT) utilizamos el Mínimo de todo (MinT) y el Máximo de todo (MaxT):

$$PLT = (X - MinT) / (MaxT - MinT)$$

PLA	v1	v2	v3	v4	v5
d1	0.57	1.00	0.80	0.29	0.70
d2	1.00	0.47	0.79	0.00	0.05
d3	0.00	0.00	0.04	1.00	0.04
d4	0.00	0.09	0.19	0.40	0.36

PLT	v1	v2	v3	v4	v5
d1	0.53	1.00	0.74	0.37	0.63
d2	0.58	0.37	0.53	0.00	0.05
d3	0.00	0.00	0.05	0.63	0.05
d4	0.00	0.05	0.11	0.16	0.16

Los Puntuación limitada son útiles para evaluar la situación de cada punto dentro del Rango del conjunto.

(#) Categoría gramatical y frecuencia del léxico español

La tabla siguiente muestra la distribución de frecuencias de todas las palabras aparecidas en el *Don Quijote* de Cervantes (1605, 1615) con clasificación vertical de categoría gramatical y la horizontal de grado de frecuencia (1 a 10).

Categoría gramatical	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**Total**
Sustantivo	1.656	973	579	349	171	70	10	2			3.810
Verbo	631	399	271	183	93	41	16	9	2	1	1.646
Adjetivo	562	279	191	122	39	25	5	2			1.225
Adverbio	55	36	20	17	18	11	8	4			169
Interjección	10	7	3	1		1					22
Numeral	7	8	8	8	1	3	1				36
Pronombre demostrativo	1	2			1	1	1				6
Pronombre indefinido	2	2	1		8	3					16
Interrogativo			2	1	2	2	1				8
Pronombre personal tónico		1	1	1	3	2	2	2			12
Preposición		3		1	4	4	1	3	2	3	21
Determinante				4	11	10	5	4	3	2	39
Conjunción		1		1	1	1	4	3		2	13
Pronombre personal átono						3	7		3		13
Relativo						1	3			1	5

El léxico suele ser clasificado entre el léxico de función (artículos,

preposiciones, conjunciones, etc.), que son de alta frecuencia y de pocos miembros, por una parte; y el léxico de contenido (nombres, adjetivos, verbos, etc.), que son de poca frecuencia y de numerosos miembros, por otra. Sin embargo, en la tabla anterior observamos que también hay léxico de función de relativamente poca frecuencia; y léxico de contenido de alta frecuencia relativa. Por esta razón hemos preparado una nueva clasificación del léxico español con un doble criterio: categoría gramatical y frecuencia.

Tipo Léxico	Alta Frecuencia ⟷ Baja Frecuencia	
Léxico de Función	Vocablo Gramatical	Vocablo Instrumental
Léxico de Contenido	Vocablo Común	Vocablo Específico

En la literatura de lingüística general se explica que los vocablos de alta frecuencia suelen ser acortados y, por otra parte, suelen conservarse bien las formas irregulares de alta frecuencia. Aparentemente parecen contradictorias estas dos explicaciones puesto que una se refiere al acortamiento y otra a la conservación, tratandose de los mismo vocablos de alta frecuencia.

Al indagar los detalles de los miembros del léxico frecuente, hemos notado que el acortamiento se ha realizado en los vocablos gramaticales de alta frecuencia, por ser palabras átonas, naturalmente pertenecientes al léxico de función, mientras que la conservación de formas irregulares se observa en vocablos comunes de también de alta frecuencia, que son del léxico de contenido. Por lo tanto, deberíamos considerar la causa de cambios lingüísticos no solo en aspectos cuantitativos de frecuencia, sino también en aspectos cualitativos de categoría gramatical.

3.5. Puntuación comparada

Llamamos «Puntuación comparada» (PC) a los Puntuación que se obtienen por la comparación que hacemos con los valores representativos del conjunto: Media, Mediana, Mitad, Mínimo, Máximo, Media Mayor, Moda Mayor.

(1) Puntuación comparada con media.

X_{np}	v1	v2	v3	v4	v5	Media.F
d1	10	19	14	7	12	12.40
d2	11	7	10	0	1	5.80
d3	0	0	1	12	1	2.80
d4	0	1	2	3	3	1.80
Media.C.	5.25	6.75	6.75	5.50	4.25	5.70

Los «Puntuación comparada con Media» demuestran la diferencia con respecto a la Media[11]. Calculamos «Puntuación comparada con Media por Diferencia en fila» (PC.Me.d.f.) y «Puntuación comparada con Media por Diferencia en columna» (PC.Me.d.c.) de manera siguiente:

$$PC.Me.d.f. = X_{np} - MeH_{n1}$$
$$PC.Me.d.c = X_{np} - MeV_{1p}$$

donde MeH_{n1} es columna de Medias horizontales y MeV_{1p} es fila de Medias verticales. Por ejemplo, calculamos el punto comparado con la media en fila de d1:v1=10, con la Media horizontal (10+19+14+7+12) / 5 = 62 / 5 = 12.4 y obtenemos CD.Me.D.r: 10 – 12.4 = -2.4.

PC.Me.d.f.	v1	v2	v3	v4	v5
d1	-2.40	6.60	1.60	-5.40	-0.40
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	-0.80	0.20	1.20	1.20

PC.Me.d.c.	v1	v2	v3	v4	v5
d1	4.75	12.25	7.25	1.50	7.75
d2	5.75	0.25	3.25	-5.50	-3.25
d3	-5.25	-6.75	-5.75	6.50	-3.25
d4	-5.25	-5.75	-4.75	-2.50	-1.25

Los «Puntuación comparada con Media por Diferencia en ambas» (PC.Me.D.a.)[12] y «Puntuación comparada con Media por Diferencia en todo» (PC.Me.D.t.) se calculan de la manera siguiente:

$$PC.Me.D.a. = [(PC.Me.D.f.) + (PC.Me.D.c)] / 2$$
$$PC.Me.D.t. = X_{np} - MeT.$$

PC.Me.d.a.	v1	v2	v3	v4	v5
d1	1.18	9.43	4.43	-1.95	3.68
d2	5.48	0.73	3.73	-5.65	-4.03
d3	-4.03	-4.78	-3.78	7.85	-2.53
d4	-3.53	-3.28	-2.28	-0.65	-0.03

PC.Me.D.t.	v1	v2	v3	v4	v5
d1	4.30	13.30	8.30	1.30	6.30
d2	5.30	1.30	4.30	-5.70	-4.70
d3	-5.70	-5.70	-4.70	6.30	-4.70
d4	-5.70	-4.70	-3.70	-2.70	-2.70

«Puntuación comparada con Media por Ratio en Fila» (PC.Me.r.f.):

$$PC.Me.r.f. = X_{np} / MeH_{n1}$$
$$PC.Me.r.f.: 0.0\ (X = 0) \leqq 1.0\ (X = MeH_{n1}) \leqq P\ (X = SumH)$$

«Puntuación comparada con Media por Ratio en columna» (PC.Me.r.c.):

$$PC.Me.r.c. = X_{np} / MeV_{1p}$$
$$PC.Me.r.c.: 0.0\ (x = 0) \leqq 1.0\ (X = MeV_{1p}) \leqq N\ (x = SumV)$$

[11] Se llama «Desviación».
[12] No utilizamos la fórmula de la Media Fraccional por no tratarse valores de división.

PC.Me.r.f.	v1	v2	v3	v4	v5
d1	0.81	1.53	1.13	0.56	0.97
d2	1.90	1.21	1.72	0.00	0.17
d3	0.00	0.00	0.36	4.29	0.36
d4	0.00	0.56	1.11	1.67	1.67

PC.Me.r.c.	v1	v2	v3	v4	v5
d1	1.90	2.81	2.07	1.27	2.82
d2	2.10	1.04	1.48	0.00	0.24
d3	0.00	0.00	0.15	2.18	0.24
d4	0.00	0.15	0.30	0.55	0.71

«Puntuación comparada con Media por Ratio en ambas» (PC.Me.r.a.):

$$PC.Me.r.b. = 2\ X_{np} / (MeH_{n1} + MeV_{1p})$$

«Puntuación comparada con Media por Ratio en todo» (PC.Me.r.t.):

$$PC.Me.r.a.. = X_{np} / MeT.$$

PC.Me.r.m.	v1	v2	v3	v4	v5
d1	1.13	1.98	1.46	0.78	1.44
d2	1.99	1.12	1.59	0.00	0.20
d3	0.00	0.00	0.21	2.89	0.28
d4	0.00	0.23	0.47	0.82	0.99

PC.Me.r.a.	v1	v2	v3	v4	v5
d1	1.75	3.33	2.46	1.23	2.11
d2	1.93	1.23	1.75	0.00	0.18
d3	0.00	0.00	0.18	2.11	0.18
d4	0.00	0.18	0.35	0.53	0.53

Como los «Puntuación comparada con Media por Diferencia» (PC.Me.D.) se amplían según la escala de los datos, dividimos la Diferencia por la Media para ajustarlos a la medida de los datos. Los llamamos «Puntuación comparada con Media por Diferencia-Ratio» (PC.Me.dr.):

«Puntuación comparada con Media por Diferencia-Ratio en fila» (PC.Me.dr.f.):

$$PC.Me.dr.f = (X_{np} - MeH_{n1}) / MeH1$$

$$PC.Me.dr.f: -1\ (x=0) \leqq 0.0\ (x = MeH1) \leqq SumH–MeH1) / MeH1\ (x=SumH)$$

«Puntuación comparada con Media por Diferencia-Ratio en columna» (PC.Me.dr.c.):

$$PC.Me.dr.c = (X_{np} - MeV_{1p}) / MeV_{1p}$$

$$PC.Me.dr.c: -1\ (x=0) \leqq 0.0\ (x = MeV_{1p}) \leqq SumV - MeV_{1p}) / Me\ (x=SumV)$$

PC.Medr.f	v1	v2	v3	v4	v5
d1	- .19	.53	.13	- .44	- .03
d2	.90	.21	.72	- 1.00	- .83
d3	- 1.00	- 1.00	- .64	3.29	- .64
d4	- 1.00	- .44	.11	.67	.67

PC.Me.dr.c.	v1	v2	v3	v4	v5
d1	.90	1.81	1.07	.27	1.82
d2	1.10	.04	.48	- 1.00	- .76
d3	- 1.00	- 1.00	- .85	1.18	- .76
d4	- 1.00	- .85	- .70	- .45	- .29

«Puntuación comparada con Media por Diferencia-Ratio en ambas» (PC.Me.dr.a.):

$$PC.Me.dr.a. = [(X_{np} - MeR_{n1}) + (X_{np} - MeC_{1p}.)] / (MeR_{n1} + MeC_{1p})$$

$$= (2\ X_{np} - MeR_{n1} - MeC_{1p}) / (MeR_{n1} + MeC_{1p}$$

«Puntuación comparada con Media por Diferencia-Ratio en todo» (PC.Me.dr.t.):

$$PC.Me.dr.a. = (X_{np} - MeA) / MeA$$

PC.Me.dr.b.	v1	v2	v3	v4	v5
d1	.13	.98	.46	- .22	.44
d2	.99	.12	.59	- 1.00	- .80
d3	- 1.00	- 1.00	- .79	1.89	- .72
d4	- 1.00	- .77	- .53	- .18	- .01

PC.Me.dr.a.	v1	v2	v3	v4	v5
d1	.75	2.33	1.46	.23	1.11
d2	.93	.23	.75	- 1.00	- .82
d3	- 1.00	- 1.00	- .82	1.11	- .82
d4	- 1.00	- .82	- .65	- .47	- .47

(2) Puntuación comparada con la mediana

X_{np}	v1	v2	v3	v4	v5	Mediana.F
d1	10	19	14	7	12	12.00
d2	11	7	10	0	1	7.00
d3	0	0	1	12	1	1.00
d4	0	1	2	3	3	2.00
Mediana.C	5.00	4.00	6.00	5.00	2.00	3.00

«Puntuación comparada con la Mediana por Diferencia»

PC.Md.Dfr.	v1	v2	v3	v4	v5
d1	-2.00	7.00	2.00	-5.00	0.00
d2	4.00	0.00	3.00	-7.00	-6.00
d3	-1.00	-1.00	0.00	11.00	0.00
d4	-2.00	-1.00	0.00	1.00	1.00

PC.Md.D.c.	v1	v2	v3	v4	v5
d1	5.00	15.00	8.00	2.00	10.00
d2	6.00	3.00	4.00	-5.00	-1.00
d3	-5.00	-4.00	-5.00	7.00	-1.00
d4	-5.00	-3.00	-4.00	-2.00	1.00

PC.Md.D.a.	v1	v2	v3	v4	v5
d1	1.50	11.00	5.00	-1.50	5.00
d2	5.00	1.50	3.50	-6.00	-3.50
d3	-3.00	-2.50	-2.50	9.00	-0.50
d4	-3.50	-2.00	-2.00	-0.50	1.00

PC.Md.D.t.	v1	v2	v3	v4	v5
d1	7.00	16.00	11.00	4.00	9.00
d2	8.00	4.00	7.00	-3.00	-2.00
d3	-3.00	-3.00	-2.00	9.00	-2.00
d4	-3.00	-2.00	-1.00	0.00	0.00

(3) Puntuación comparada con la mitad

X_{np}	v1	v2	v3	v4	v5	Mitad.F
d1	10	19	14	7	12	13.00
d2	11	7	10	0	1	5.50
d3	0	0	1	12	1	6.00
d4	0	1	2	3	3	1.50
Mitad.C	5.50	9.50	7.50	6.00	6.50	9.50

«Puntuación comparada con la Mitad por Diferencia»

PC.Ct.D.r.	v1	v2	v3	v4	v5
d1	-3.00	6.00	1.00	-6.00	-1.00
d2	5.50	1.50	4.50	-5.50	-4.50
d3	-6.00	-6.00	-5.00	6.00	-5.00
d4	-1.50	-0.50	0.50	1.50	1.50

PC.Ct.D.c.	v1	v2	v3	v4	v5
d1	4.50	9.50	6.50	1.00	5.50
d2	5.50	-2.50	2.50	-6.00	-5.50
d3	-5.50	-9.50	-6.50	6.00	-5.50
d4	-5.50	-8.50	-5.50	-3.00	-3.50

PC.Ct.D.m.	v1	v2	v3	v4	v5
d1	0.75	7.75	3.75	-2.50	2.25
d2	5.50	-0.50	3.50	-5.75	-5.00
d3	-5.75	-7.75	-5.75	6.00	-5.25
d4	-3.50	-4.50	-2.50	-0.75	-1.00

PC.Ct.D.a.	v1	v2	v3	v4	v5
d1	0.50	9.50	4.50	-2.50	2.50
d2	1.50	-2.50	0.50	-9.50	-8.50
d3	-9.50	-9.50	-8.50	2.50	-8.50
d4	-9.50	-8.50	-7.50	-6.50	-6.50

(4) Puntuación comparada con el mínimo

X_{np}	v1	v2	v3	v4	v5	Mínimo.F
d1	10	19	14	7	12	7.00
d2	11	7	10	0	1	.00
d3	0	0	1	12	1	.00
d4	0	1	2	3	3	.00
Mínimo.C.	.00	.00	1.00	.00	1.00	.00

«Puntuación comparada con el Mínimo por Diferencia»

PC.Mn.D.r.	v1	v2	v3	v4	v5
d1	3.00	12.00	7.00	0.00	5.00
d2	11.00	7.00	10.00	0.00	1.00
d3	0.00	0.00	1.00	12.00	1.00
d4	0.00	1.00	2.00	3.00	3.00

PC.Mn.D.c.	v1	v2	v3	v4	v5
d1	10.00	19.00	13.00	7.00	11.00
d2	11.00	7.00	9.00	0.00	0.00
d3	0.00	0.00	0.00	12.00	0.00
d4	0.00	1.00	1.00	3.00	2.00

PC.Mn.D.m.	v1	v2	v3	v4	v5
d1	6.50	15.50	10.00	3.50	8.00
d2	11.00	7.00	9.50	0.00	0.50
d3	0.00	0.00	0.50	12.00	0.50
d4	0.00	1.00	1.50	3.00	2.50

PC.Mn.D.a.	v1	v2	v3	v4	v5
d1	10.00	19.00	14.00	7.00	12.00
d2	11.00	7.00	10.00	0.00	1.00
d3	0.00	0.00	1.00	12.00	1.00
d4	0.00	1.00	2.00	3.00	3.00

(5) Puntuación comparada con el máximo

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

«Puntuación comparada con el Máximo por Diferencia»

PC.Mx.D.r.	v1	v2	v3	v4	v5
d1	-9.00	0.00	-5.00	-12.00	-7.00
d2	0.00	-4.00	-1.00	-11.00	-10.00
d3	-12.00	-12.00	-11.00	0.00	-11.00
d4	-3.00	-2.00	-1.00	0.00	0.00

PC.Mx.D.c.	v1	v2	v3	v4	v5
d1	-1.00	0.00	0.00	-5.00	0.00
d2	0.00	-12.00	-4.00	-12.00	-11.00
d3	-11.00	-19.00	-13.00	0.00	-11.00
d4	-11.00	-18.00	-12.00	-9.00	-9.00

PC.Mx.D.m.	v1	v2	v3	v4	v5
d1	-5.00	0.00	-2.50	-8.50	-3.50
d2	0.00	-8.00	-2.50	-11.50	-10.50
d3	-11.50	-15.50	-12.00	0.00	-11.00
d4	-7.00	-10.00	-6.50	-4.50	-4.50

PC.Mx.D.a.	v1	v2	v3	v4	v5
d1	-9.00	0.00	-5.00	-12.00	-7.00
d2	-8.00	-12.00	-9.00	-19.00	-18.00
d3	-19.00	-19.00	-18.00	-7.00	-18.00
d4	-19.00	-18.00	-17.00	-16.00	-16.00

(6) Puntuación comparada con la media mayor

X_{np}	v1	v2	v3	v4	v5	Media mayor. F.
d1	10	19	14	7	12	12.22
d2	11	7	10	0	1	6.00
d3	0	0	1	12	1	1.89
d4	0	1	2	3	3	1.89
Media mayor. C.	5.17	5.83	6.50	5.33	3.50	4.90

«Puntuación comparada con la Media Mayor por Diferencia»

PC.MjMe.d.f	v1	v2	v3	v4	v5
d1	-2.22	6.78	1.78	-5.22	-.22
d2	5.00	1.00	4.00	-6.00	-5.00
d3	-1.89	-1.89	-.89	10.11	-.89
d4	-1.89	-.89	.11	1.11	1.11

PC.MjMe.d.c	v1	v2	v3	v4	v5
d1	4.83	13.17	7.50	1.67	8.50
d2	5.83	1.17	3.50	-5.33	-2.50
d3	-5.17	-5.83	-5.50	6.67	-2.50
d4	-5.17	-4.83	-4.50	-2.33	-.50

PC.MjMe.d.b	v1	v2	v3	v4	v5
d1	1.31	9.97	4.64	-1.78	4.14
d2	5.42	1.08	3.75	-5.67	-3.75
d3	-3.53	-3.86	-3.19	8.39	-1.69
d4	-3.53	-2.86	-2.19	-.61	.31

PC.MjMe.d.a	v1	v2	v3	v4	v5
d1	5.10	14.10	9.10	2.10	7.10
d2	6.10	2.10	5.10	-4.90	-3.90
d3	-4.90	-4.90	-3.90	7.10	-3.90
d4	-4.90	-3.90	-2.90	-1.90	-1.90

(7) Puntuación comparada con la moda mayor

X_{np}	v1	v2	v3	v4	v5	Moda mayor. F.
d1	10	19	14	7	12	12.00
d2	11	7	10	0	1	9.33
d3	0	0	1	12	1	.50
d4	0	1	2	3	3	2.67
Moda mayor. C.	3.33	2.67	4.33	3.33	1.67	1.09

«Puntuación comparada con la Moda Mayor por Diferencia»

PC.MjMd.D.r	v1	v2	v3	v4	v5
d1	-2.00	7.00	2.00	-5.00	.00
d2	1.67	-2.33	.67	-9.33	-8.33
d3	-.50	-.50	.50	11.50	.50
d4	-2.67	-1.67	-.67	.33	.33

PC.MjMd.D.c	v1	v2	v3	v4	v5
d1	6.67	16.33	9.67	3.67	10.33
d2	7.67	4.33	5.67	-3.33	-.67
d3	-3.33	-2.67	-3.33	8.67	-.67
d4	-3.33	-1.67	-2.33	-.33	1.33

s

PC.MjMd.d.b	v1	v2	v3	v4	v5
d1	2.34	11.67	5.84	-.67	5.17
d2	4.67	1.00	3.17	-6.33	-4.50
d3	-1.92	-1.59	-1.42	10.09	-.09
d4	-3.00	-1.67	-1.50	.00	.83

PC.MjMd.da	v1	v2	v3	v4	v5
d1	8.91	17.91	12.91	5.91	10.91
d2	9.91	5.91	8.91	-1.09	-.09
d3	-1.09	-1.09	-.09	10.91	-.09
d4	-1.09	-.09	.91	1.91	1.91

3.6. Puntuación estandarizada

Para evaluar las filas o las columnas en una misma escala y varianza, conviene convertirlas para que su Media resulte 0, y la Desviación Típica 1, restando de la fila o de la columna en cuestión la Media (M) y dividiendo la diferencia por la Desviación Típica (DT). El resultado se llama «Puntuación estandarizada» (PE).

X_{np}	v1	v2	v3	v4	v5	Mf.	DTf.
d1	10	19	14	7	12	12.40	4.03
d2	11	7	10	0	1	5.80	4.53
d3	0	0	1	12	1	2.80	4.62
d4	0	1	2	3	3	1.80	1.17
Mc.	5.25	6.75	6.75	5.50	4.25	5.70	
DTc.	5.26	7.56	5.45	4.50	4.55		5.66

Las fórmulas de los «Puntuación estandarizada en fila» (PEf) y «Puntuación estandarizada en columna» (PEc) son:

$$PEf_{np} = (X_{np} - Mf_{n1}) / DTf_{n1}$$
$$PEc_{np} = (X_{np} - Mc_{1p}) / DTc_{1p}$$

PEf	v1	v2	v3	v4	v5
d1	-0.60	1.64	0.40	-1.34	-0.10
d2	1.15	0.26	0.93	-1.28	-1.06
d3	-0.61	-0.61	-0.39	1.99	-0.39
d4	-1.54	-0.69	0.17	1.03	1.03

PEc	v1	v2	v3	v4	v5
d1	0.90	1.62	1.33	0.33	1.70
d2	1.09	0.03	0.60	-1.22	-0.71
d3	-1.00	-0.89	-1.06	1.44	-0.71
d4	-1.00	-0.76	-0.87	-0.56	-0.27

Los «Puntuación estandarizada en ambas» (PEa) es la Media Fraccional (MF) de PEf_{np} y PEc_{np}:

$$\begin{aligned} PEb_{np} &= MF(PEf_{np}, PEc_{np}) \\ &= [(X_{np} - Mf_{n1}) + (X_{np} - Mc_{1p})] / (DTf_{n1} + DTc_{1p}) \\ &= (2\ X_{np} - Mf_{n1} - Mc_{1p})] / (DTf_{n1} + DTc_{1p}) \end{aligned}$$

Para los «Puntuación estandarizada en todo» (PEt) utilizamos la Media en todo (Mt) y Desviación Típica en todo (DTt):

$$PEt_{np} = (X_{np} - Mt) / DTt$$

$PEaa_{np}$	v1	v2	v3	v4	v5
d1	0.25	1.63	0.93	-0.46	0.86
d2	1.12	0.12	0.75	-1.25	-0.89
d3	-0.81	-0.78	-0.75	1.72	-0.55
d4	-1.10	-0.75	-0.69	-0.23	-0.01

PEt_{np}	v1	v2	v3	v4	v5
d1	0.76	2.35	1.47	0.23	1.11
d2	0.94	0.23	0.76	-1.01	-0.83
d3	-1.01	-1.01	-0.83	1.11	-0.83
d4	-1.01	-0.83	-0.65	-0.48	-0.48

(*) Media y desviación típica de los Puntuación estandarizada

Los Puntuación estandarizada tienen la propiedad de poseer la Media 0 y la Desviación Típica 1. Vamos al ver la razón de la Media 0.

$$PM_E = (PE_1 + PE_2 + ... + PE_N) / N$$

$$= [(X_1 - M)/DT + (X_2 - M)/DT + \ldots + (X_N - M)/DT] / N$$
$$= [(X_1 - M) + (X_2 - M) + \ldots + (X_N - M)] / (N\ DT)$$
$$= [(X_1 + X_2 + \ldots + X_N) - N\ M] / (N\ DT)$$

donde la parte del numerador $(X_1 + X_2 + \ldots + X_N)$ es la Suma, igual a N M, N veces de Medias. De modo que el numerador total resulta 0 y, por consiguiente, PM_E resulta 0.

Seguidamente vamos a ver la razón de la Desviación Típica (DT) igual a 1.

$$DT_{PE}^2 = [(SM_1 - PM_E)^2 + (SM_2 - PM_E)^2 + \ldots + (SM_N - PM_E)^2] / N$$
$$= \{(SM_1 - 0)^2 + (SM_2 - 0)^2 + \ldots + (SM_N - 0)^2\} / N$$
$$= \{[(X_1 - M)/DT]^2 + [(X_2 - M)/DT]^2 + \ldots + [(X_N - M) / DT]^2]\} / N$$
$$= [(X_1 - M)^2 + (X_2 - M)^2 + \ldots + (X_N - M)^2] / (N\ DT^2)$$

donde

$$[(X_1 - M)^2 + (X_2 - M)^2 + \ldots + (X_N - M)^2] / N = DT^2$$

es decir, se trata de la Varianza. De modo que llegamos a:

$$= DT^2 / DT^2 = 1$$

Como la Desviación Típica (DT) es la raíz cuadrada de la Varianza, también la DT resulta 1.

De esta manera llegamos al concepto de los Puntuación estandarizada con la Media 0 y Desviación 1. Los valores no estandarizados conllevan el nombre de la unidad, por ejemplo, frecuencia de palabra, frecuencia por mil, kilómetros, temperatura, puntuación de pruebas, etc.. Los Puntuación estandarizada están libres de la unidad y podemos evaluar las características de variación independientemente de la escala de los datos.

(*) Propiedad de Puntuación estandarizada

Los Puntuación estandarizada del dato X_{np} y Y_{np} que es A X_{np} + B, como podemos observar en las tablas siguiente.

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

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

$2X_{np}+5$	v1	v2	v3	v4	v5
d1	25	43	33	19	29
d2	27	19	25	5	7
d3	5	5	7	29	7
d4	5	7	9	11	11

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

Para saber la razón de esta propiedad, hay que ver la Media de X y a X + b:

$$M(a\,X+b) = a\,M(X) + b$$

porque

$$\begin{aligned} M(aX+b) &= \Sigma_i\,(a\,X_i + b)] / N && \leftarrow \text{definición de Media} \\ &= (a\ \Sigma_i\,X_i + N\,b) / N && \leftarrow \text{a al esterior; } \Sigma_i b = N\,b \\ &= a\ \underline{\Sigma_i\,X_i / N} + N\,b / N && \leftarrow \text{distribuir / N} \\ [1] \quad &= a\,M(X) + b && \leftarrow \text{def. de Media; N b / N = b} \end{aligned}$$

Por otra parte, entre la Desviación Típica (DT) de X y la de a X + b, existe la relación:

$$DT(a\,X+b) = a\,DT(X)$$

porque

$$\begin{aligned} DT(a\,X+b) &= \{\Sigma_i\,[(aX_1 + b) - M(aX+b)]^2 / N\}^{1/2} && \leftarrow \text{def. de DT} \\ &= \{\Sigma_i\,[aX_1 + b - (a\,M(X) + b)]^2 / N\}^{1/2} && \leftarrow \text{ver [1]} \\ &= \{\Sigma_i\,[aX_1 - a\,M(X)]^2 / N\}^{1/2} && \leftarrow \text{borrar b} \\ &= \{\Sigma_i\,a^2\,[X_1 - M(X)]^2 / N\}^{1/2} && \leftarrow \text{a al exterior} \\ &= a\,\{\Sigma_i\,[X_1 - M(X)]^2 / N\}^{1/2} && \leftarrow \text{a al exterior} \\ [2] \quad &= a\,DT(X) && \leftarrow \text{def. de DT} \end{aligned}$$

Por lo tanto

$$\begin{aligned} PE(aX+b) &= [(aX + b) - M(aX + b)] / DT(aX + b) && \leftarrow \text{de. de PE} \\ &= \{(aX + b) - [a\,M(X) + b]\} / a\,DT(X) && \leftarrow \text{ver [1], [2]} \\ &= [aX - a\,M(X)] / a\,DT(X) && \leftarrow \text{borrar b} \\ &= a\,[X - M(X)] / a\,DT(X) && \leftarrow \text{a al exterior} \\ &= [X - M(X)] / DT(X) && \leftarrow \text{borrar a} \\ &= PE(X) && \leftarrow \text{def. de PE (Puntuación estandarizada)} \end{aligned}$$

Si se multiplica por –a, la fo/rmula de [1] se vuelve:

$$M(-a\,X + b) = --a\,M(X) + b$$

Por lo tanto,

$$SS(-aX + b) = -SS(X)$$

(*) Puntuación estandarizada normalizados

Al observar los Puntuación estandarizada encontramos con frecuencia valores superiores a 1. Esto quiere decir que la diferencia con respecto a la Media supera la Desviación Típica.

Pensamos que también es conveniente poseer unos valores estandarizados cuyo Rango es [-1.00 ~ 1.00], es decir, con cifras normalizadas. Por esta razón proponemos formular los «Puntuación estandarizada normalizados» (PEN). Para normalizar los Puntuación, hay que buscar el Máximo teórico de los Puntuación estandarizada (PEmax). Una vez encontrado el Méximo, ya es cuestión de hacer la división:

$$NPE = PE / PEmax$$

Los Puntuación estandarizada en máximo se presentan cuando la distribución es totalmente parcial de forma {X, A, A, A, A} (X > A}. En tal caso X llega al valor máximo estandarizado. Como hemos visto anteriormente, los Puntuación estandarizada no varían si los datos son multiplicados y/o sumados. Por esta razón, para simplificar la cuestión de buscar el Máximo de PE podemos convertir los datos en {X – A, 0, 0, 0, 0}. Y ahora con X – A = K tenemos los datos {K, 0, 0, 0, 0}. Con este dato, vamos a buscar el valor de punto estandarizado de K.

En primer lugar hay que notar la relación siguiente (M: Media; N: número de datos).

[1] $M = K / N$

porque K = Suma en nuestro caso extremo {K, 0, 0, 0, 0}.

Por otra parte, hemos visto que la Desviación Típica del dato {K, 0, 0, 0, 0} llega al Máximo que es:

[2] $DTmax = M (N - 1)^{1/2}$

Por lo tanto, el Máximo de los Puntuación estandarizada (PEmax) es:

$PEmax = (K - M) / DTmax$ ←def. de PE
$= (K - M) / [M (N -1)^{1/2}]$ ←ver [2]
$= \underline{(K - K / N)} / [K / N (N -1)^{1/2}]$ ←ver [1]
$= \underline{(N K - K)} / N / [K / N (N -1)^{1/2}]$ ←/ N al exterior

$= [(N - 1)\ \underline{K / N}] / [\underline{K / N}\ (N - 1)^{1/2}]$ ←K al exterior

$= (N - 1) / (N - 1)^{1/2}$ ←borrar K / N

$= (N - 1)^{1/2}$ ←$X / \sqrt{X} = \sqrt{X}$

Y finalmente los «Puntuación estandarizada normalizados» (PEN) es:

$PEN = PE / PEmax = PE / (N - 1)^{1/2}$

Las tablas siguientes muestran los «Puntuación estandarizada normalizados» en fila, columna, ambas y todo.

NPEf.	v1	v2	v3	v4	v5
d1	-0.30	0.82	0.20	-0.67	-0.05
d2	0.57	0.13	0.46	-0.64	-0.53
d3	-0.30	-0.30	-0.19	1.00	-0.19
d4	-0.77	-0.34	0.09	0.51	0.51

NPEc.	v1	v2	v3	v4	v5
d1	0.52	0.94	0.77	0.19	0.98
d2	0.63	0.02	0.34	-0.71	-0.41
d3	-0.58	-0.52	-0.61	0.83	-0.41
d4	-0.58	-0.44	-0.50	-0.32	-0.16

NPEa.	v1	v2	v3	v4	v5
d1	0.14	0.89	0.51	-0.25	0.46
d2	0.60	0.07	0.40	-0.67	-0.48
d3	-0.44	-0.43	-0.40	0.92	-0.29
d4	-0.62	-0.42	-0.39	-0.13	0.00

NPEt.	v1	v2	v3	v4	v5
d1	0.17	0.54	0.34	0.05	0.26
d2	0.21	0.05	0.17	-0.23	-0.19
d3	-0.23	-0.23	-0.19	0.26	-0.19
d4	-0.23	-0.19	-0.15	-0.11	-0.11

3.7. Puntuación esperada

Para formular los «Puntuación comparada con Frecuencia Esperada» (PCFE), utilizamos una tabla de «Frecuencias Esperadas» (FE), que muestran frecuencias esperadas desde los Puntuación de vista tanto de la Suma horizontal como de la vertical.

Xnp	v1	v2	v3	v4	v5	Suma: Sh_{n1}
d1	10	19	14	7	12	62
d2	11	7	10	0	1	29
d3	0	0	1	12	1	14
d4	0	1	2	3	3	9
Suma Sv_{1p}	21	27	27	22	17	Total St 114

Las frecuencias esperadas se calculan con la Suma horizontal y la vertical. Por ejemplo, la Suma horizontal de d1 es 62, la Suma vertical de v1 es 21 y la Suma total es 114. Entonces, es lógico que la Frecuencia Esperada de d1:v1 sea el valor que ocupa 62 dentro de la Ratio de 21 / 114, es decir $62 \times (21 / 114) \fallingdotseq 11.42$. Generalizando:

$FE_{np} = (Sr_{n1}\ Sc_{1p}) / St$

FEnp	v1	v2	v3	v4	v5
d1	11.42	14.68	14.68	11.96	9.25
d2	5.34	6.87	6.87	5.60	4.32
d3	2.58	3.32	3.32	2.70	2.09
d4	1.66	2.13	2.13	1.74	1.34

Denominamos «Puntuación comparada con Frecuencia Esperada» (PCFE), calculando la diferencia (d) entre la Frecuencia de Input (X_{np}) y Frecuencia Esperada (FE_{np}), la ratio (r) entre los dos y la diferencia-ratio (dr) entre los dos:

$PCFEd_{np} = X_{np} - FE_{np}$
$PCFEr_{np} = X_{np} / FE_{np}$
$PCFEdr_{np} = (X_{np} - FE_{np}) / FE_{np}$

PCFEd	v1	v2	v3	v4	v5
d1	-1.42	4.32	-0.68	-4.96	2.75
d2	5.66	0.13	3.13	-5.60	-3.32
d3	-2.58	-3.32	-2.32	9.30	-1.09
d4	-1.66	-1.13	-0.13	1.26	1.66

PCFEr	v1	v2	v3	v4	v5
d1	0.88	1.29	0.95	0.59	1.30
d2	2.06	1.02	1.46	0.00	0.23
d3	0.00	0.00	0.30	4.44	0.48
d4	0.00	0.47	0.94	1.73	2.24

PCFEdr	v1	v2	v3	v4	v5
d1	-0.12	0.29	-0.05	-0.41	0.30
d2	1.06	0.02	0.46	-1.00	-0.77
d3	-1.00	-1.00	-0.70	3.44	-0.52
d4	-1.00	-0.53	-0.06	0.73	1.24

En los «Puntuación comparada con Frecuencia Esperada» (PCFE), no hay opción de Eje (Horizontal, Vertical, Ambos y Total), puesto que los calculamos únicamente con la matriz de «Frecuencias Esperadas» (FE).

3.8. Puntuación ordenada

(1) Puntuación ordenada descendentes

Por el orden descendiente de los datos formulamos los «Puntuación ordenada Descendentes» (POD), con el número 1 en el Máximo del conjunto.

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

Los «Puntuación ordenada Descendentes horizontal» (PODh) y «Puntuación ordenada Descendentes vertical» (PODv) son:

PODh	v1	v2	v3	v4	v5
d1	4	1	2	5	3
d2	1	3	2	5	4
d3	4	4	2	1	2
d4	5	4	3	1	1

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

Calculamos los «Puntuación ordenada Descendientes bilateral» (PODb) por la Media Aritmética de PODh y PODv, y los «Puntuación ordenada Descendientes total» (PODt) en el conjunto de datos:

DRSb	v1	v2	v3	v4	v5
d1	3.0	1.0	1.5	3.5	2.0
d2	1.0	2.5	2.0	4.5	3.5
d3	3.5	4.0	3.0	1.0	2.5
d4	4.0	3.5	3.0	2.0	1.5

DRSt	v1	v2	v3	v4	v5
d1	6	1	2	8	3
d2	5	8	6	17	13
d3	17	17	13	3	13
d4	17	13	12	10	10

(2) Puntuación ordenada ascendientes

Por el orden ascendiente de los datos, formulamos los «Puntuación ordenada Ascendentes» (POA), con el número 1 en el Mínimo del conjunto. Los «Puntuación ordenada Ascendentes horizontal» (POAh) y «Puntuación ordenada Ascendentes vertical» (POAv) son:

POAh	v1	v2	v3	v4	v5
d1	2	5	4	1	3
d2	5	3	4	1	2
d3	1	1	3	5	3
d4	1	2	3	4	4

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

Calculamos los «Puntuación ordenada Ascendentes bilateral» (PODb) por la Media Aritmética de POAh y POAv, y los «Puntuación ordenada Ascendentes total» (POAt) en el conjunto de datos:

POAb	v1	v2	v3	v4	v5
d1	2.5	4.5	4.0	2.0	3.5
d2	4.5	3.0	3.5	1.0	1.5
d3	1.0	1.0	2.0	4.5	2.0
d4	1.0	2.0	2.5	3.0	3.5

POAt	v1	v2	v3	v4	v5
d1	14	20	19	12	17
d2	16	12	14	1	5
d3	1	1	5	17	5
d4	1	5	9	10	10

3.9. Puntuación divergente

Proponemos utilizar los «Puntuación divergente» (PD) para detectar las frecuencias anómalas desde el punto de vista probabilístico.

Xnp	v1	v2	v3	v4	v5	Sh
d1	10	19	14	7	12	62
d2	11	7	10	0	1	29
d3	0	0	1	12	1	14
d4	0	1	2	3	3	9
Sv	21	27	27	22	17	S: 114

Derivamos los «Puntuación divergente» de la Probabilidad Binomial (→ **(*)**) con r = Frecuencia (X_{np}), n = Suma horizontal (Sh), Pr = Suma vertical / Suma total. Por ejemplo, consideramos que el dato de d1:v1 (10) corresponde al número de Éxitos dentro del número de Ensayos que es la Suma horizontal y la Probabilidad general corresponde a la división de la Suma vertical por la Suma total: 21 / 114. De esta manera obtenemos la probabilidad del valor 10 dentro de la tabla Bn(X_{np}):

Bn(X_{np})	v1	v2	v3	v4	v5
d1	.122	.050	.118	.037	.081
d2	.007	.171	.065	.002	.047
d3	.058	.023	.099	.000	.256
d4	.160	.245	.304	.167	.106

Por otra parte, calculamos la misma Probabilidad dentro de la tabla de Puntuación esperada (E_{np}), que debe ofrecer el valor máximo de la probabilidad, tratándose de valores esperados (Bn(Enp)):

Enp	v1	v2	v3	v4	v5
d1	11.42	14.68	14.68	11.96	9.25
d2	5.34	6.87	6.87	5.60	4.32
d3	2.58	3.32	3.32	2.70	2.09
d4	1.66	2.13	2.13	1.74	1.34

Bn(Enp)	v1	v2	v3	v4	v5
d1	.13	.12	.12	.13	.14
d2	.19	.17	.17	.19	.21
d3	.27	.25	.25	.26	.29
d4	.33	.30	.30	.31	.37

La división de la probabilidad de X_{np} por la probabilidad de los Puntuación esperada presenta la Ratio de la probabilidad de que aparezca el valor en cuestión. Como nos interesa el grado de divergencia, sacamos el valor reverso de la Ratio de manera siguiente: 1 – Hnp(Xnp) / Hnp(Enp), donde Hnp es la matriz de la probabilidad binomial y Enp, la matriz de Puntuación esperada. También conviene saber si Xnp es positivo o negativo con respecto a los Puntuación esperada y, por esta razón, conservamos el signo de la diferencia entre Xnp y Enp.

Las fórmulas de los «Puntuación divergente horizontales» (PDh) y «Puntuación divergente verticales» (PDv) son:

$$PDh_{np} = Sgn * [1 - H_{np}(X_{np}) / H_{np}(E_{np})]$$
$$PDv_{np} = Sgn * [1 - V_{np}(X_{np}) / V_{np}(E_{np})]$$

donde Sgn es el signo de la Xnp – Enp, Hnp es la matriz de las probabilidades binomiales horizontales y Vnp es la matriz de las probabilidades binomiales verticales:

$$H_{np}(X_{np}) = Binom(X_{np}, P, Sv / T)$$
$$V_{np}(X_{np}) = Binom(X_{np}, N, Sh / T)$$

donde Binom(r, n, p) es la función que devuelve la probabilidad binomial con parámetros de número de Éxitos (r), número de Ensayos (n) y la Probabilidad (p). P es el número de columnas, N es el número de filas, Sv es fila de Sumas verticales, Sh es columna de Sumas horizontales y T es Suma total.

DSr.	v1	v2	v3	v4	v5
d1	-0.06	0.58	0.00	-0.70	0.43
d2	0.96	-0.02	0.61	-0.99	-0.77
d3	-0.78	-0.91	-0.60	1.00	-0.12
d4	-0.51	-0.19	0.00	0.47	0.71

DSc.	v1	v2	v3	v4	v5
d1	-0.16	0.73	0.00	-0.88	0.57
d2	0.97	-0.02	0.61	-0.99	-0.82
d3	-0.76	-0.88	-0.53	1.00	-0.11
d4	-0.44	-0.10	0.00	0.49	0.71

Formulamos los «Puntuación divergente bilaterales» (PDb) como Media Fraccional (MF) de PDh y PDv:

$$PDb = MF(PDh, PDv)$$
$$= Sgn * \{1 - [H_{np}(X_{np}) + V_{np}(X_{np})] / [H_{np}(E_{np}) + V_{np}(E_{np})]\}$$

Para los «Puntuación divergente totales» (PDt), utilizamos la matriz de probabilidades que presenta la matriz homogénea cuyos elementos son 1 / (N*p):

$$E_{np} = 1 / (N*P)\ I_{1n}\ I_{n1}$$
$$A_{np}(X_{np}) = Binom(X_{np}, T, 1 / (N*P))$$

$$DSa = Sgn * [1 - A_{np}(X_{np}) / A_{np}(E_{np})]$$

donde N es número de filas, P es número de columnas. Por $E_{np} = 1 / (N*P) I_{1n} I_{n1}$ formulamos una matriz homogénea de 1 / (N*P). La probabilidad binomial se obtiene de X_{np}, Suma total (T) y la probabilidad de 1 / (N*P).

DSb.	v1	v2	v3	v4	v5
d1	-0.12	0.66	0.00	-0.80	0.51
d2	0.97	-0.02	0.61	-0.99	-0.80
d3	-0.77	-0.89	-0.57	1.00	-0.11
d4	-0.48	-0.15	0.00	0.48	0.71

DSa	v1	v2	v3	v4	v5
d1	-0.81	1.00	-0.99	-0.22	0.96
d2	0.91	0.22	0.81	-0.98	-0.90
d3	-0.98	-0.98	-0.90	0.96	-0.90
d4	-0.98	-0.90	-0.70	0.41	0.41

(*) Probabilidad binomial

Pensamos en la probabilidad utilizando un ejemplo usual del dado. El dado tiene 6 caras con números {1, 2, 3, 4, 5, 6}. Cuando echamos una vez, cuyo acto se llama «Ensayo» (ing. *trial*), la probabilidad de que salga un número determinado, por ejemplo «1», es lógicamente 1/6 y la probabilidad de que no salga este número «1» es 5/6. De esta manera tenemos dos casos de F (Fracaso) y E (Éxito) y con sus correspondientes probabilidades:

«1»	Número de E	Probabilidad
E	1	1/6 ≒ 0.167
F	0	5/6 ≒ 0.833

Procedamos a echar el dado dos veces, es decir ahora el número de Ensayo es 2. Si sale el «1», escribimos E y si no, F. Si en la primera vez no sale «1» y en la segunda vez sale «1», simbolizamos estos dos Ensayos con «F, E». De esta manera manera tenemos ahora la tabla siguiente de probabilidades:

«1»	Número de E	Probabilidad
E, E	2	(1/6) (1/6) = 1/36 ≒ 0.028
E, F	1	(1/6) (5/6) = 5/36 ≒ 0.139
F, E	1	(5/6) (1/6) = 5/36 ≒ 0.139
F, F	0	(5/6) (5/6) = 25/36 ≒ 0.694

Ahora procedamos a echar tres veces el dado, es decir, el número de Ensayo es 3:

«1»	Número de E	Probabilidad
E, E, E	3	(1/6) (1/6) (1/6) = 1/216 ≒ 0.005
E, E, F	2	(1/6) (1/6) (5/6) = 5/216 ≒ 0.023
E, F, E	2	(1/6) (5/6) (1/6) = 5/216 ≒ 0.023

E, F, F	1	(1/6) (5/6) (5/6) = 25/216 ≒ 0.116
F, E, E	2	(5/6) (1/6) (1/6) = 5/216 ≒ 0.023
F, E, F	1	(5/6) (1/6) (5/6) = 25/216 ≒ 0.116
F, F, E	1	(5/6) (5/6) (1/6) = 25/216 ≒ 0.116
F, F, F	0	(5/6) (5/6) (5/6) = 125/216 ≒ 0.579

Ahora bien, sumamos casos en que salga «1» dos veces sin pensar en el orden de la aparición de «1». Pueden ser E, E, F; ó E, F, E; ó F, E, E:

«1»	Número de E	Probabilidad
E, E, F	**2**	**(1/6) (1/6) (5/6) = 5/216 ≒ 0.023**
E, F, E	**2**	**(1/6) (5/6) (1/6) = 5/216 ≒ 0.023**
F, E, E	**2**	**(5/6) (1/6) (1/6) = 5/216 ≒ 0.023**

Ahora tenemos la Suma de las tres probabilidades dentro de los ocho casos: 5/216 + 5/216 + 5/216 = 15/216 ≒ 0.069. Es decir es el múltiplo de 5/216 por 3. Y cada término de esta multiplicación es $(1/6)^2$ (5/6) = 5/216, es decir se trata del producto de dos veces de 1/6 y una vez de 5/6.

Seguidamente tenemos que tener en cuenta que hay tres casos en que aparecen dos Éxitos, de modo que hemos multiplicado el producto de $(1/6)^2$ (5/6) por 3. Cuando se trata de muchos más Ensayos, necesitamos una fórmula generalizada de buscar el número de casos. Este número de casos corresponde a la Combinación de nCr, donde n es el número de Ensayos y r es número de Éxitos. La Combinación nCr se calcula:

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

por ejemplo[13]

$${}_3C_2 = nPr \,/\, r! = (3 * 2) \,/\, (2 * 1) = 3$$

Entonces la probabilidad de los tres casos citados es:

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

Generalizando, la probabilidad de r éxitos en n Ensayos cuya probabilidad teórica es Pr es:

$$\text{Binom}\,(r, n, Pr) = {}_n\mathbf{C}_r\,(Pr)^r\,(1-Pr)^{n-r}$$

[13] Es una situación donde buscamos el número de casos de escoger dos objetos dentro de los tres {a, b, c} en pensar el orden. Si pensamos el orden, hay 6 casos: ab, ac, ba, bc, ca, cb, es decir ${}_3P_2 = 3 * 2 = 6$, que es la Permutación (P): nPr = n (n – 1)(n – 2) … (n – r + 1). Ahora si no hacemos caso al orden, ab es lo mismo que ba, ac es lo mismo que ca y bc es lo mismo que cb. Entonces hay que dividir ${}_3P_2 = 3 * 2 = 6$ por 2: que es 2! (2 * 1). De esta manera ${}_3\mathbf{C}_2 = (3 \times 2) \,/\, (2 \times 1)$.

Esta probabilidad se llama «Probabilidad Binomial». En nuestro caso concreto:

$$\text{Binom } (2, 3, 1/6) = {}_3\mathbf{C_2}\ (1/6)^2\ (1 - 1/6)^{3-2}$$
$$= 3\ (1/6)^2\ (5/6) = 3 * 0.023 = 0.069$$

3.10. Puntuación asociativa

Utilizando los «Coeficientes de Asociación», de los que hablaremos en el Capítulo siguiente, calculamos el grado de Asociación entre la Fila y la Columna en el punto de coocurrencia en forma de «Puntuación asociativa» (PA). Para operación necesitamos preparar de antemano las matrices de A_{np}, que indica la selección tanto de Fila como de Columna, que es la misma que la Matriz de Input (Xnp); Bnp, que indica la frecuencia de la selección positiva de la Fila y la negativa de la Columna; Cnp, que indica la selección negativa de la Fila y la positiva de la Columna; y Dnp, que indica la selección negativa tanto de la Fila como de la Columna.

Xnp	v1	v2	v3	v4	v5	Suma Sh
d1	10	19	14	7	12	62
d2	11	7	10	0	1	29
d3	0	0	1	12	1	14
d4	0	1	2	3	3	9
Suma Sv	21	27	27	22	17	T: 114

Por ejemplo, consideramos el valor de d1:v1 (10) como la frecuencia de la selección positiva de la Fila d1 y de la Columna v1 (A:+/+). La frecuencia de d1(+):v1(–), es decir, (B:+/–) se calcula de Sh (62) – X (10) = 52. La frecuencia de d1(-):v1(+), es decir, (C:-/+) se calcula de Sv (21) – X (10) = 11. Y finalmente, la frecuencia de d1(-):v1(-), es decir, (D:-/-) es: T (114) – A (10) – B (52) – C (11) = 41.

Xnp	v1	v2	v3	v4	v5
d1	A:10	B:52			
d2 d3 d4	C:11	D:41			

Para el elemento d2:v2, también es posible calcular los valores de A, B, C, D, de la manera siguiente:

実測値	v1	v2	v3	v4	v5
d1	D:10	C:19	D:33		
d2	B:11	A:7	B:11		
d3 d4	D:0	C:1	D:22		

Utilizando las operaciones matriciales, hacemos los cálculos siguientes y llegamos a obtener las cuatro matrices siguientes:

$A_{np} = X_{np}$
$B_{np} = Sh_{n1} - X_{np}$
$C_{np} = Sv_{p} - X_{np}$
$D_{np} = S - A_{np} - B_{np} - C_{np}$

Anp	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

Bnp	v1	v2	v3	v4	v5
d1	52	43	48	55	50
d2	18	22	19	29	28
d3	14	14	13	2	13
d4	9	8	7	6	6

Cnp	v1	v2	v3	v4	v5
d1	11	8	13	15	5
d2	10	20	17	22	16
d3	21	27	26	10	16
d4	21	26	25	19	14

Dnp	v1	v2	v3	v4	v5
d1	41	44	39	37	47
d2	75	65	68	63	69
d3	79	73	74	90	84
d4	84	79	80	86	91

Utilizando estas cuatro matrices, podemos obtener los «Coeficientes Asociativos» para formular los «Puntuación asociativa». Por ejemplo, los «Puntuación asociativa por Correspondencia Simple» (PA.CS) de la manera siguiente:

$PA.CS = (A_{np} + D_{np}) / (A_{np} + B_{np} + C_{np} + D_{np})$

PA.CS	v1	v2	v3	v4	v5
d1	.447	.553	.465	.386	.518
d2	.754	.632	.684	.553	.614
d3	.693	.640	.658	.895	.746
d4	.737	.702	.719	.781	.825

«Puntuación asociativa por Jaccard» (PA.J) y «Puntuación asociativa por Jaccard-2» (PA.J2):

$PA.J = A_{np} / (A_{np} + B_{np} + C_{np})$

$PA.J2 = A_{np}*2 / (A_{np}*2 + B_{np} + C_{np})$

PA.J	v1	v2	v3	v4	v5
d1	.137	.271	.187	.091	.179
d2	.282	.143	.217	.000	.022
d3	.000	.000	.025	.500	.033
d4	.000	.029	.059	.107	.130

PA.J2	v1	v2	v3	v4	v5
d1	.241	.427	.315	.167	.304
d2	.440	.250	.357	.000	.043
d3	.000	.000	.049	.667	.065
d4	.000	.056	.111	.194	.231

«Puntuación asociativa por Russel-Rao» (PA.RR) y «Puntuación asociativa por Russel-Rao-3» (PA.RR3):

$PA.RR = A_{np} / (A_{np} + B_{np} + C_{np} + D_{np})$

$PA.RR3 = A_{np}*3 / (A_{np}*3 + B_{np} + C_{np} + D_{np})$

PA.RR	v1	v2	v3	v4	v5
d1	.088	.167	.123	.061	.105
d2	.096	.061	.088	.000	.009
d3	.000	.000	.009	.105	.009
d4	.000	.009	.018	.026	.026

PA.RR3	v1	v2	v3	v4	v5
d1	.224	.375	.296	.164	.261
d2	.243	.164	.224	.000	.026
d3	.000	.000	.026	.261	.026
d4	.000	.026	.051	.075	.075

«Puntuación asociativa por Hamann» (PA.H) y «Puntuación asociativa por Yule» (PA.Y):

$PA.H = [(A_{np} + D_{np}) - (B_{np} + C_{np})] / [(A_{np} + D_{np}) + (B_{np} + C_{np})]$

$PA.Y = [(A_{np} * D_{np}) - (B_{np} * C_{np})] / [(A_{np} * D_{np}) + (B_{np} * C_{np})]$

PA.H	v1	v2	v3	v4	v5
d1	-.105	.105	-.070	-.228	.035
d2	.509	.263	.368	.105	.228
d3	.386	.281	.316	.789	.491
d4	.474	.404	.439	.561	.649

PA.Y	v1	v2	v3	v4	v5
d1	-.165	.417	-.067	-.522	.386
d2	.642	.017	.356	-1.000	-.733
d3	-1.000	-1.000	-.641	.964	-.425
d4	-1.000	-.449	-.045	.387	.529

«Puntuación asociativa por Phi» (PA.Phi) y «Puntuación asociativa por Ochiai» (PA.O):

$$PA.Phi = [(A_{np} * D_{np}) - (B_{np} * C_{np})] / [(A_{np} + B_{np})*(C_{np}+D_{np})*(A_{np}+C_{np})*(B_{np}+D_{np})]^{1/2}$$

$$PA.O = S_{np} / [(A_{np} + B_{np})*(A_{np}+C_{np})]^{1/2}$$

PA.Phi	v1	v2	v3	v4	v5
d1	-.065	.179	-.028	-.222	.136
d2	.294	.006	.148	-.286	-.188
d3	-.178	-.208	-.146	.630	-.082
d4	-.139	-.087	-.010	.104	.151

PA.O	v1	v2	v3	v4	v5
d1	.277	.464	.342	.190	.370
d2	.446	.250	.357	.000	.045
d3	.000	.000	.051	.684	.065
d4	.000	.064	.128	.213	.243

«Puntuación asociativa por Preferencia» (PA.Pr):

$$PA.Pr = [A_{np} * 2 - (B_{np} + C_{np})] / [A_{np} * 2 + (B_{np} + C_{np})]$$

PA.Pr	v1	v2	v3	v4	v5
d1	-.518	-.146	-.371	-.667	-.392
d2	-.120	-.500	-.286	-1.000	-.913
d3	-1.000	-1.000	-.902	.333	-.871
d4	-1.000	-.889	-.778	-.613	-.538

Observamos que en los Puntuación de PA.Pr, casi todos los Puntuación son negativos, menos d3:v4, lo cual demuestra que entre d3 y v4 existe cierto grado de asociación a pesar de los números de contra ejemplos, B y C. Precisaremos las propiedades que poseen los «Coeficientes Asociativos» que aparecen en los «Puntuación asociativa» en el capítulo siguiente (4. Relaciones).

3.11. Puntuación normalizada

Llamamos «Puntuación normalizada» (PN) a los Puntuación convertidos cuya Suma total resulte 1.

3.11.1. Puntuación normalizada por suma total

Para el método más sencillo para llegar a unos Puntuación normalizada se recurre a la división de los datos X_{np} por la Suma total:

$$NS.S_{np} = X_{np} / T$$

X_{np}	v1	v2	v3	v4	v5	Sh
d1	10	19	14	7	12	62
d2	11	7	10	0	1	29
d3	0	0	1	12	1	14
d4	0	1	2	3	3	9
Sv	21	27	27	22	17	T:114

NS.S	v1	v2	v3	v4	v5	Sh
d1	.088	.167	.123	.061	.105	.544
d2	.096	.061	.088	.000	.009	.254
d3	.000	.000	.009	.105	.009	.123
d4	.000	.009	.018	.026	.026	.079
Sv	.184	.237	.237	.193	.149	1.000

＊池田（1976: 121-123）の「総和を基礎にした相対度数」を参照しました。

3.11.2. Puntuación normalizada por media fraccional

Las fórmulas y la tabla de los «Puntuación normalizada por Media Fraccional» (PN.MF) son las siguientes:

$W_{np} = 2\ X_{np} / (Shn1 + Sv1p)$

$PN.MF = Wnp / Sum(Wnp)$

X_{np}	v1	v2	v3	v4	v5	Sh
d1	10	19	14	7	12	62
d2	11	7	10	0	1	29
d3	0	0	1	12	1	14
d4	0	1	2	3	3	9
Sv	21	27	27	22	17	T:114

PN.MF	v1	v2	v3	v4	v5	Sh
d1	.062	.109	.080	.043	.078	.371
d2	.112	.064	.091		.011	.279
d3			.012	.170	.016	.199
d4		.014	.028	.049	.059	.151
Sv	.174	.187	.212	.262	.164	1.000

3.11.3. Puntuación normalizada de Monsteller

Dividiendo la tabla por unos determinados valores, llegamos a obtener una tabla cuyas Sumas horizontales sean iguales y Sumas verticales sean iguales también. La Suma total es 1. La tabla así normalizada (NS.Mos) sirve para comparar los Puntuación con bases comunes de la Suma horizontal (Sh) y de la Suma vertical (Sv). El método se llama «Normalización de Monsteller» y llamamos la tabla resultante «Puntuación normalizada de Monsteller» (PN.Mos):

X_{np}	v1	v2	v3	v4	v5	Sh
d1	10	19	14	7	12	62
d2	11	7	10	0	1	29
d3	0	0	1	12	1	14
d4	0	1	2	3	3	9
Sv	21	27	27	22	17	T:114

PS.Mos	v1	v2	v3	v4	v5	Sh
d1	.068	.091	**.043**	.007	.041	.250
d2	.132	.059	.053	.000	.006	.250
d3	.000	.000	.041	.162	.047	.250
d4	.000	.050	**.063**	.031	.106	.250
Sv	.200	.200	.200	.200	.200	1.000

Para obtener los «Puntuación normalizada de Mosteller» (PN.Mos), cuyas Sumas horizontales sean iguales y Suma total sea 1, hay que dividir la matriz X_{np} por la columna de Sumas horizontales (Sh) * número de filas (4):

$X^1np = Xnp / (Sh * 4)$

X^1np	v1	v2	v3	v4	v5	Sh
d1	.040	.077	.056	.028	.048	.250
d2	.095	.060	.086	.000	.009	.250
d3	.000	.000	.018	.214	.018	.250

d4	.000	.028	.056	.083	.083	.250
Sv	.135	.165	.216	.326	.158	1.000

Seguidamente, para obtener la fila de Sumas verticales iguales y la Suma total que sea 1, dividimos la matriz anterior $X1_{np}$ por la fila de Sumas verticales (Sv) de la misma por el número de columnas (5):

$X^2np = Xnp^1 / (Sv * 5)$

X^2np	v1	v2	v3	v4	v5	Sh
d1	.060	.093	.052	.017	.061	.283
d2	.140	.073	.080	.000	.011	.304
d3	.000	.000	.017	.132	.023	.171
d4	.000	.034	.051	.051	.105	.242
Sv	.200	.200	.200	.200	.200	1.000

En este momento, la columna de Sumas horizontales cambian, de modo que hay que hacer de nuevo la división de la matriz resultante por la columna de Sumas horizontales (Sh) por 4. Seguimos practicando las mismas operaciones, naturalmente con programa, llegamos a los «Puntuación normalizada de Mosteller» (PN.Mos).

Como hemos visto, los «Puntuación normalizada de Monsteller» (PN.Mos) mantienen la igualdad tanto de las Sumas horizontales (Sh), como de las Sumas verticales (Sv), pueden cambiar de manera desproporcional los valores orginales, lo que podemos observar en Xnp y PS.Mos. El propósito de PN.Mos es ver las proporsiones de frecuencias con bases iguales de fila y de columna.

(#) Razones de normalización: dos variantes de la letra «s», números de fallecidos y sobrevivientes

En español de Edad Media y Moderna, la letra «s» poseía dos variantes: la ese corta <s> y la ese larga <ſ>. Se ha observado que la <s> aparece en la posición final de palabra (_#). En realidad, sin embargo, también se encuentra la <s> en la posición inicial (#_) y media (&_&) de palabra. La tabla inferior izquierda muestra la frecuencia de <s> y <ſ> en los primeros 20.000 letras de *Libro de Alexandre* (1300) y la tabla derecha, sus «Puntuación normalizada por Suma» (PN.S):

/s/	#_	&_&	_#	Sh
<s>	62	2	593	657
<ſ>	314	412	109	835

PN.S	#_	&_&	_#	Sh
<s>	.042	.001	.397	.440
<ſ>	.210	.276	.073	.560

Sv	376	414	702	1492

Sv	.252	.277	.471	1.000

Podemos observar fácilmente la tendencia de aparición de <s> al final de palabra en una tabla de frecuencia absoluta así de dimensión pequeña y valores reducidos. También es útil la tabla de Puntuación normalizada por Suma (PN.S). Cuando se trata de los datos grandes de dimensión y de valores, normalmente se recurre a las tablas de Puntuación Relativa, horizontales (PRh) y verticales (PRv)

PRh	#_	&_&	_#	Sh
<s>	.094	.003	.903	1.000
<ſ>	.376	.493	.131	1.000
Sv	.470	.496	1.033	2.000

PRv	#_	&_&	_#	Sh
<s>	.165	.005	.845	1.014
<ſ>	.835	.995	.155	1.986
Sv	1.000	1.000	1.000	3.000

Nos damos cuenta de que en la tabla de «Puntuación Relativa horizontales» (PRh), nuestra atención se fija en las filas y, efectivamente, observamos que la ese corta <s> se encuentra mayoritariamente en la posición final (_#: .903). Por otra parte, sin embargo, en la tabla de «Puntuación Relativa verticales» (PRv), también son notables las altas proporciones de la ese alta <ſ> en la posición inicial (#_) y media (&_&) de la palabra, lo cual no hemos notado en la tabla anterior (PRh). Tampoco podemos obtener la misma proporción de las dos variantes en PRh que en PRv, dividiendo .376 por .470 en PRh, que resulta .800, que es distinto de .835 de PRv.

La tabla siguiente muestra los «Puntuación normalizada por Media Fraccional» (PN.MF):

PN.MF	#_	&_&	_#	Sh
<s>	.052	.002	.377	.430
<ſ>	.224	.285	.061	.570
Sv	.276	.286	.438	1.000

En esta tabla, se calcula la Media Fraccional de la <s> en la posición inicial (#_) tanto de la Ratio horizontal (62/657) como de la vertical (62/314) en forma de: (62 x 2) / (657+314) = .052.

Como ejemplo de mayor envergadura, supongamos que estamos ante una tabla de números de fallecidos y sobrevivientes de la epidemia de cólera en dos ciudades A y B. La tabla inferior derecha muestra los «Puntuación Relativa verticales». Por esta tabla, ¿podemos afirmar que el número de los fallecidos de la ciudad A (.032) es 2.7 veces más grande que el de la ciudad B (.012) por .032 / .12 ≒ 2.66? Si fuera así, al comparar las Ratios de sobrevivientes (.968, .988), la diferencia resulta nada destacable: .968 / .988 ≒ .980.

Xnp	Ciudad A	Ciudad B	Sh
Fallecidos	1300	250	1550
Sobrevivientes	**39000**	20000	59000
Sv	40300	20250	60550

PRv	A	B	Sh
F.	.032	.012	.045
S.	.968	.988	1.955
Sv	1.000	1.000	2.000

En realidad no debemos hacer la comparación de las Ratios entre los conjuntos cuyos números de miembros son diferentes. De ahí viene la necesidad de la normalización de los datos. La tabla inferior izquierda (PN.S) muestra los «Puntuación normalizada por Suma» y la derecha (PN.MF), los «Puntuación normalizada por Media Fraccional»:

PN.S	**Ciudad A**	**Ciudad B**	**Sh**
Fallecidos	.021	**.004**	.026
Sobrevivientes	.644	.330	.974
Sv	.666	.334	1.000

PN.MF.	**A**	**B**	**Sh**
F.	.045	**.017**	.062
S.	.571	.367	.938
Sv	.616	.384	1.000

La tabla anterior izquierda PN.S es simplemente de los «Puntuación normalizada por Suma». Como todos los Puntuación están divididos por la Suma total (60550), los valores son comparables. De esta manera podemos apreciar los detalles de Ratios, lo cual es difícil de hacer con los datos originales (Xnp). Sin embargo, la cifra de los fallecidos en la Ciudad B viene muy reducida por incluir el caso de sobrevivientes en la Ciudad A, que no tiene mucho que ver con la cifra de los fallecidos en la Ciudad B. En cambio, en el cálculo de los «Puntuación normalizada por Media Fraccional» (PN.MF), para el caso de los fallecidos en la Ciudad B, se toma en cuenta tanto de la Ratio horizontal con los fallecidos en la Ciudad A, como de la Ratio vertical con los sobrevivientes de la Ciudad B, ambas relevantes para el caso de los fallecidos en la Ciudad B. La cifra resultante .017 es más convincente que el caso anterior (.004).

En los estudios de distintas disciplinas, no solamente en los estudios lingüísticos, se hacen comparaciones de los conjuntos de dimensiones diferentes. Sabemos que no se puede comparar directamente con frecuencias absolutas si la base es diferente. Entonces recurrimos a los Puntuación Relativa en forma de Ratio, Porcentaje (%), Por mil, Por millón, etc. No obstante, en el sentido riguroso de la palabra, la comparación de los Puntuación Relativa no es practicable, cuando las bases son diferentes. En un caso extremo, por ejemplo, sabemos a ciencia cierta que no tiene sentido la comparación entre 250/1000=25% y 3/10=30%. Entonces se cree que se puede hacer la comparación cuando las bases son "parecidas", "cercanas", por ejemplo, entre 25/400, 25/450. ¿Hasta donde se permite hacer la comparación? ¿No existe problema en la comparación de los datos cuyos bases tienen la diferencia de 1.5 o 2 veces más grande que otra, por ejemplo entre 30/100 y 40/200? Una de las razones

importantes para recurrir a la Normalización está precisamente en que por ella obtenemos medios que posibilitan las comparaciones igualadas.

3.12. Puntuación dummy

La tabla inferior izquierda es una matriz de datos, que se puede convertir en una matriz que denominamos «Puntuación dummy» en forma de la tabla inferior derecha (D(X)):

X	v1	v2	v3
h1	5	2	1
h2	1	3	3
h3	0	1	2
h4	1	2	1

D(X)	h1	h2	h3	h4	v1	v2	v3
c.1	1	0	0	0	1	0	0
c.2	1	0	0	0	1	0	0
c.3	1	0	0	0	1	0	0
c.4	1	0	0	0	1	0	0
c.5	1	0	0	0	1	0	0
c.6	1	0	0	0	0	1	0
c.7	1	0	0	0	0	1	0
c.8	1	0	0	0	0	0	1
c.9	0	1	0	0	1	0	0
c.10	0	1	0	0	0	1	0
c.11	0	1	0	0	0	1	0
c.12	0	1	0	0	0	1	0
c.13	0	1	0	0	0	0	1
c.14	0	1	0	0	0	0	1
c.15	0	1	0	0	0	0	1
c.16	0	0	1	0	0	1	0
c.17	0	0	1	0	0	0	1
c.18	0	0	1	0	0	0	1
c.19	0	0	0	1	1	0	0
c.20	0	0	0	1	0	1	0
c.21	0	0	0	1	0	1	0
c.22	0	0	0	1	0	0	1

Por ejemplo la información de h1:v1 = 5 en la matriz de datos está representada en la columna h1 y v1 con las 5 repeticiones de 1 en la matriz de «Puntuación dummy». La matriz de «Puntuación dummy» es útil para analizar las relaciones entre los casos y los atributos en la misma dimensión al mismo tiempo.

3.13. Frecuencia probabilística

3.13.1. Problema

En la lingüística de corpus, con la búsqueda múltiple (de varias formas al mismo tiempo) con atributos también múltiples (por ejemplo, años), llegamos a obtener una tabla bidimensional, de formas y de variables (años). Ahora estamos en condición de ver el fenómeno no solamente en una franja de años (1200, siglo XIII), sino de observar los cambios lingüísticos a lo largo de la cronología (1200, 1250, etc.) en comparación con otras formas. Veamos un dato real de la Frecuencia Absoluta (FA) de las tres formas con variación ortográfica: <uoz>, <voz>, <boz>[14]:

FA	1200	1250	1300	1350	1400
uoz	3	8	3	11	6
boz	0	3	8	18	35
voz	0	1	1	23	53
Sm	3	12	12	52	94

Estas frecuencias, sin embargo, no son comparables, puesto que las Sumas (Sm) de las tres formas son diferentes {3, 12, 12, 52, 94}. Por ejemplo, la cifra 3 <*uoz*> en 1200 no es comparable con la 8 de la misma forma en 1250. En tal caso, los investigadores recurren a la Frecuencia Relativa (FR), que se calcula por la división de la Frecuencia Absoluta por la Suma, por ejemplo, 3 / 3 = 1.000, 8 / 12 = .667. Si multiplicamos la Frecuencia Relativa por 100, llegamos a la cifra de porcentaje: 1.000 * 100 = 100 (%), 0.667 * 100 = 66.7 (%):

FA	1200	1250	1300	1350	1400
uoz	3	8	3	11	6
boz	0	3	8	18	35
voz	0	1	1	23	53
Sm	3	12	12	52	94

FR (%)	1200	1250	1300	1350	1400
uoz	**100.0**	**66.7**	25.0	21.2	6.4
boz	0.0	25.0	66.7	34.6	37.2
voz	0.0	8.3	8.3	44.2	56.4
Sm	100.0	100.0	100.0	100.0	100.0

Sin embargo, ni la Frecuencia Relativa (FR) ni el porcentaje (%) son adecuados para comparar las cifras con bases distintas. Por ejemplo, 3 entre 3 (FR: 1.000 (100%)) presenta la cifra mayor que 8 entre 12 (FR: 0.667 (66.7%)), a pesar de que pensamos e intuimos que 3 entre 3 es menos importante que 8 entre 12 y mucho menos importante que 80 entre 120. Los aficionados de fútbol saben o intuyen que el futbolista que ha metido 3 goles en 3 partidos es menos

[14] La tabla se ha obtenido en el sitio de «CODEA en LYNEAL»: http://shimoda.lllf.uam.es/ueda/lyneal/codea.htm

importante que el otro ha metido 8 goles entre 12. Todo esto significa que el porcentaje no sirve para la comparación numérica, por la razón de que, por ejemplo, 3 goles en 10 partidos no garantiza 30 goles en 100 partidos, lo que dice precisamente el 30%. Creemos que el porcentaje sirve para describir la proporción que ocupa cada caso dentro del conjunto, pero no sirve para comparar cada caso entre varios conjuntos. Más adelante buscaremos la solución de este problema de la evaluación numérica, propio de la Frecuencia Relativa (FR) y del porcentaje.

Pero antes, veamos el problema de otra frecuencia también utilizada en la lingüística de corpus en general. Se trata de la Frecuencia Normalizada (FN), que se calcula por la división de la Frecuencia Absoluta (FA) por la Totalidad de Palabras (TP) contadas en cada sección, multiplicada por algún Multiplicador (M) apropiado.

FN = FA / TP * M

Por ejemplo, en la franja de 1200 del corpus se han contado 7 736 palabras en total, que es la Totalidad de Palabras (TP). Entonces, la Frecuencia Normalizada de <uoz> en 1200 es 3 / 7 736 * 100 000 = 38.8. Recomendamos utilizar como Multiplicador (M) el número redondeado próximo a la Máxima de la base (TP): 96 059 (en el conjunto de datos de 1400). Llegamos a obtener la tabla inferior derecha (FN):

FA	1200	1250	1300	1350	1400
uoz	3	8	3	11	6
boz	0	3	8	18	35
voz	0	1	1	23	53
TP	7 736	36 052	40 957	64 999	96 059

FN.	1200	1250	1300	1350	1400
uoz	**38.8**	**22.2**	7.3	16.9	6.2
boz	0.0	8.3	19.5	27.7	36.4
voz	0.0	2.8	2.4	35.4	55.2

Sin embargo, aquí en la Frecuencia Normalizada (FN) también se presenta el mismo problema de falta de comparabilidad propia de los datos de bases diferentes, y especialmente, de algunas bases bastante reducidas. No podemos menos de sentir dudas sobre la cifra FN de <uoz> en 1200, 3 entre 7 736 cuya FN es 38.8 en comparación con la FN de la misma forma <uoz> en 1250, 8 entre 36 052 cuya FN es 22.2. Nos preguntamos si 38.8 es realmente comparable con 22.2.

La esencia del problema es la misma tanto en la Frecuencia Relativa (FR) como en la Frecuencia Normalizada (FN) en el sentido de que las dos calculan sobre bases diferentes. Paradójicamente, las dos frecuencias se utilizan precisamente cuando las bases son diferentes, puesto que si las bases son iguales no hace falta recurrir a estas frecuencias y en la Frecuencia Absoluta neta

podemos hacer la comparación numérica sin problema.

El problema de la falta de comparabilidad tratado en este apartado se soluciona por la eliminación del conjunto en cuestión. En el ejemplo de los datos de las tres formas medievales, se trataría de eliminar el conjunto correspondiente a 1200. Es la práctica general en el tratamiento estadístico. Por ejemplo, en el mundo deportivo de beísbol, se calcula la puntuación de los jugadores con participación suficiente en los partidos. Los jugadores que no pasan el umbral establecido están excluidos de la evaluación desde el principio. Pero nos preguntamos qué hacemos con la franja de 1250, donde se registran las frecuencias dentro de la base de casi un tercio de 1400 (37.5%).

Nuestra idea es tratar todos los datos sin distinción, pero con criterios comunes de probabilidad. Nuestro método, que a continuación vamos a explicar, ofrece la evaluación de los datos de manera equitativa, con bases similares o distantes, lo que demuestra la robustez absoluta, en comparación con los métodos tradicionales de la Frecuencia Relativa (FR), inclusive su variante Porcentaje, y la Frecuencia Normalizada (FN), cuya fragilidad hemos visto en los casos de bases distantes en esta sección.

Para ofrecer la solución del problema propio de la Frecuencia Relativa (FR) y de la Frecuencia Normalizada (FN), presentamos una nueva fórmula de frecuencia. Nuestro propósito es buscar un tipo de frecuencia, «Frecuencia Probabilística» (FP), que represente el valor relativo de la Frecuencia Absoluta (FA) dentro del conjunto (base) con unos cálculos sencillos de la probabilidad (Ueda 2017). Se justifica por los resultados de experimentos reales y por nuestra intuición y pensamiento que, por ejemplo, 30 entre 100 es más importante (o significativo) que 3 entre 10, aun siendo ambos iguales de 0.3 (30%) en la probabilidad de ocurrencia. Para demostrarlo, recurrimos a la fórmula probabilística binomial. El camino para llegar a conocer la Frecuencia Probabilística (FP) recorre los tres pasos siguientes: (1) Significatividad (S), (2) Probabilidad Esperada (PE) y, finalmente, (3) Multiplicador (M).

3.13.2. Significatividad

Intuimos y pensamos que, por ejemplo, un futbolista que ha metido 28 goles en 100 partidos es más «importante» y ha contribuido más al equipo que el otro que ha metido 3 goles en 10 partidos, aunque la ratio de goles del primero (28 / 100 = 28%) es menor que la del segundo (3 / 10 = 30%). Al grado de importancia, damos el nombre de «Significatividad» (S) (ing. *significance*). Partimos de unos casos sencillos y especiales para llegar al caso general, aplicable a las frecuencias en general.

Para calcular la Significatividad (S), utilizamos la probabilidad. Para

llegar a entenderla, empezamos con unos ejemplos tan sencillos como de tirada de una moneda, donde cada faz tiene la Probabilidad Esperada (PE) de 0.5 (50%)[15].

La «cara» de moneda, que tiene la probabilidad de 0.5 (50%) de aparecer en una tirada de moneda.

La tabla siguiente muestra las probabilidades de dos eventos (x = 0, 1) de una tirada (número de ensayo n=1) de una moneda: «cara», con valor de 1, o «cruz», con valor 0. La Probabilidad Esperada (PE) de cada una es 0.5, puesto que hay dos posibilidades: cara (1) o cruz (0). En la tabla cada evento viene con su propia Probabilidad de Ocurrencia (PO), que acabamos de ver, la Probabilidad Cumulativa (PC), que se va acumulando con cada Probabilidad de Ocurrencia correspondiente, y la Significatividad (S):

x	Caso	PO(x, 1, 0.5)	PC(x, 1, 0.5)	S(x, 1, 0.5)
cruz: x = 0	(0)	1/2 = 0.5	0.5	0
cara: x = 1	(1)	1/2 = 0.5	0.5 + 0.5 = 1.0	0.5

La columna de Probabilidad de Ocurrencia PO(x) muestra en la primera fila la probabilidad de cruz (x = 0) que es PO(0) = 1/2 = 0.5 y, en la segunda fila, la de cara (x = 1) que es PO(1) = 1/2 = 0.5. La Probabilidad Cumulativa (PC) de x = 0, PC(0), es 0.5, que es igual a PO(0) y la de x = 1, PC(1), es 1.0, que es suma de PO(0) = 0.5 y PO(1) = 0.5. La última Probabilidad Cumulativa (PC) es siempre 1.

Ahora bien, definimos la «Significatividad» (S) de x, S(x, n, e), como correspondiente a la Probabilidad Cumulativa de x - 1, PC (x-1, n, e):

S(x, n, e) = PC(x-1, n, e)

(x: Ocurrencia; n: Ensayos; e: Probabilidad Esperada)

La Significatividad (s) de x = 0, S(0), la definimos como 0, por no existir la Probabilidad Cumulativa de -1:

S(0) = 0

La razón por la que consideramos la Significatividad como la Probabilidad Cumulativa de la Ocurrencia del caso inmediatamente anterior está

[15] Esto significa que si tiramos una moneda, sale siempre una de las dos caras y si tiramos 1000 veces, casi la mitad de las veces (aproximadamente 500 veces), sale la cara y otra mitad de las veces, la cruz. Entonces intuimos y pensamos lógicamente que la Probabilidad Esperada (PE) de cara es 0.5 (50%).

en que pensamos que la suma de las probabilidades anteriores de la Ocurrencia correspondiente es la probabilidad de la significatividad de los números de ocurrencias anteriores. Si echamos una moneda, la Significatividad de la ocurrencia de 1 («cara») es 0.5, lo que es complementaria del Riesgo (no «cara», es decir, «cruz»), que es también 0.5. Por consiguiente,

Significatividad + Riesgo = 1

lo que quiere decir que hay Significatividad de 0.5 (50%) de la aparición de la cara y hay Riesgo de 0.5 (50%). Esto quiere decir que si apostamos por la aparición de la cara, hay un 50% de riesgo, lo que sabemos e intuimos sin recurrir a la teoría de probabilidad.

Hasta aquí hemos visto un caso muy sencillo en que tiramos la moneda solo una vez. ¿Qué ocurre si tiramos la misma moneda dos veces? La tabla siguiente muestra la distribución de Probabilidades de Ocurrencia (PO) que presentan en dos ensayos de tirar una moneda (n = 2). Hay tres casos posibles: x = 0, 1, 2, es decir, (0,0), (1,0) + (0,1) y (1,1):

x	Caso	PO(x, 2, 0.5)	PC(x, 2, 0.5)	S(x, 2, 0.5)
x = 0	(0, 0)	1/4 = 0.25	0.25	0
x = 1	(0, 1); (1, 0)	2/4 = 0.50	0.25 + 0.50 = 0.75	0.25
x = 2	(1, 1)	1/4 = 0.25	0.75 + 0.25 = 1.00	0.75

Esta vez la Probabilidad Esperada (PE) de «cara» es igualmente 0.5. La columna de la Probabilidad de Ocurrencia (PO) muestra que la PO de 0 ocurrencias de cara, PO(0), es 0.25 (cruz, cruz) = (0, 0), es decir 1 de 4 casos. El total de casos son 4, porque se cuentan 4 casos siguientes: (0, 0), (0, 1), (1, 0), (1, 1). La probabilidad de 1 ocurrencia de cara, (cara, cruz) + (cruz, cara); (1, 0) + (0, 1) es 0.5 (2 de 4 casos). Y, finalmente, la probabilidad de 2 ocurrencias de cara, (cara, cara); (1, 1), PO(2) es 0.25, que ocurre 1 de 4 casos. La columna de la Probabilidad Cumulativa (PC) presenta las probabilidades sumadas desde 0 hasta 2 en cada ocurrencia: x = 0, 1, 2.

La última columna de Significatividad (S) corresponde al caso anterior de la Probabilidad Cumulativa (PC). La última Significatividad (S) de n = 2 es S(2) = 0.75, que representa un aumento considerable con respecto al experimento anterior en el que se tiraba solo una vez la moneda: 0.5 (n = 1), lo que quiere decir que 2 entre 2 (S = 0.75) es mucho más «significativo» (importante) que 1 entre 1 (S = 0.5), aun siendo ambos iguales de 100% de Probabilidad Cumulativa (PC). Sin embargo, la Significatividad (S) todavía no llega más que a 0.75 (75%), lo que quiere decir que hay 0.25 (25%) de Riesgo.

Ahora bien, precisamos los 3 parámetros de la Significatividad (S): x:

ocurrencias, *n*: total de las veces de ensayos, *e*: Probabilidad Esperada (PE) en forma de la función S(*x, n, e*):

S(2, 2, 0.5) = PC(1, 2, 0.5) = 0.75

De la misma manera, la Significatividad (S) de *x* = 1 es:

S(1, 2, 0.5) = PC(0, 2, 0.5) = 0.25

Veamos el experimento de 3 tiradas de ensayo (*n* = 3):

x	Caso	PO(*x*, 3, 0.5)	PC(*x*, 3, 0.5)	S(*x*, 3, 0.5)
x = 0	(0,0,0)	1/8 = .125	.125	0
x = 1	(1,0,0), (0,1,0), (0, 0, 1)	3/8 = .375	.500	.125
x = 2	(1,1,0), (1,0,1), (0,1,1)	3/8 = .375	.875	.500
x = 3	(1,1,1)	1/8 = .125	1.000	.875

La Significatividad (S) de la última ocurrencia (*x* = 3) ha aumentado en 0.875 y por consiguiente ahora el Riesgo ha disminuido en 0.125: 1 - 0.875 = 0.125.

S(3, 3, 0.5) = PC(2, 3, 0.5) = 0.875 (87.5%)

Si apostamos que no salga 3 veces la cara, hay probabilidad de 87.5% de ganar la apuesta, que es la Significatividad (S); y el Riesgo de perder la apuesta es de 12.5%. Deberíamos aumentar la Significatividad (S) hasta, por lo menos, 95% (0.95) y, si es posible, hasta 99%, con los Riesgos de 5% o 1%, respectivamente. De esta manera perdemos la apuesta solo 1 de 20 veces, o 1 de 100 veces.

Veamos el experimento de 10 ensayos (*n* = 10):

x	PO(*x*, 10, 0.5)	PC(*x*, 10, 0.5)	S(*x*, 10, 0.5)
x = 0	.001	.001	.000
x = 1	.010	.011	.001
x = 2	.044	.055	.011
x = 3	.117	.172	.055
x = 4	.205	.377	.172
x = 5	.246	.623	.377
x = 6	.205	.828	.623
x = 7	.117	.945	.828
x = 8	.044	.989	.945
x = 9	.010	.999	.989
x = 10	.001	1.000	.999

Finalmente cuando x = 9, obtenemos la Significatividad S(9, 10, 0.5) = 0.989, superior a 95%, y la S(10) = 0.999, superior a 99%, lo que quiere decir que podemos presentar la cifra de 9 entre 10 con Significatividad (S) mayor de 95% y la de 10 entre 10 con Significatividad (S) mayor de 99%. En realidad, al echar la moneda 10 veces si salen 9 veces la cara de la moneda, la probabilidad total de las ocurrencias menores de 9 [0, 1, 2, ..., 8] se suma a 98.9%, que es bastante significativo. Es decir, con la Significatividad de 98.9% podemos afirmar que 9 entre 10 es significativo (importante). Es significativo o importante en el sentido de que 9 o 10 entre 10 ocurren solo con la probabilidad de 0.010 + 0.001 = 0.011 (1.1%). De la misma manera, podemos afirmar que 10 entre 10 posee la Significatividad de 0.999 (99.9%). Compárense con los casos de 1 entre 1 (Significatividad de 50%), 2 entre 2 (75%) y 3 entre 3 (87.5%)[16].

Hasta aquí hemos visto el comportamiento matemático de la Significatividad (S), que depende de los tres parámetros: x: ocurrencias, n: total de las veces de ensayos, e: Probabilidad Esperada (PE). Hemos observado su movimiento de acuerdo con x y n. Ahora veamos qué Significatividad (S) se presenta de acuerdo con el cambio de la Probabilidad Esperada (e). La tabla siguiente muestra la Significatividad (s) de las ocurrencias (x) de eventos dotados de la Probabilidad Esperada (e) de 0.1, por ejemplo, la Probabilidad Esperada (e) de sacar la tarjeta de «1» dentro de las diez tarjetas de {1, 2, ... , 10}:

x	PO(x, 10, 0.1)	PC(x, 10, 0.1)	S(x, 10, 0.1)
x = 0	.349	.349	.000
x = 1	.387	.736	.349
x = 2	.194	.930	.736
x = 3	.057	.987	.930

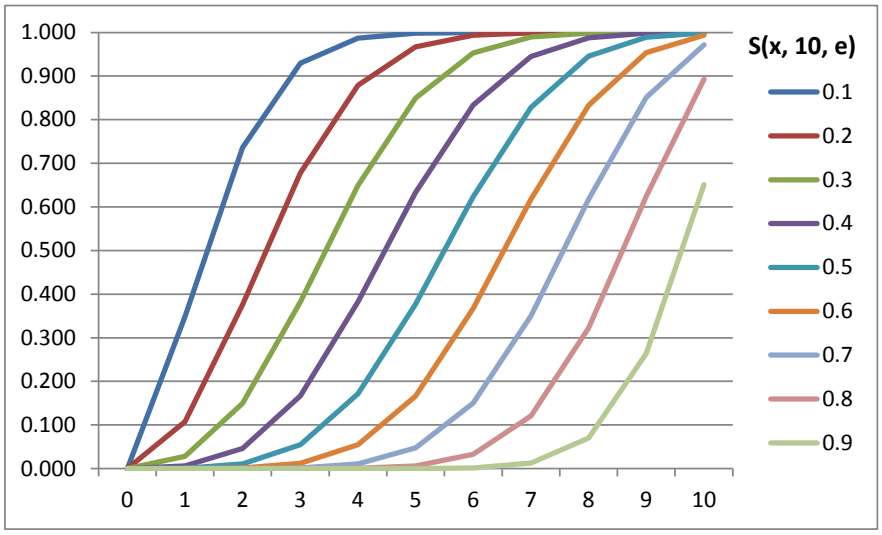
[16] Aquí no se trata de la Probabilidad de Ocurrencia (PO) sino de la Probabilidad Cumulativa (PC) de los casos hasta el caso inmediatamente anterior. Observamos la PO individual, por ejemplo, de un ensayo de 5 caras en 10 monedas, es tan solo .246. Al sumar las probabilidades de 0 a 4 caras, llegamos a la PC de .377. Si sumamos los casos de 0 a 5 caras, llegamos a la PC de .623. Entre .377 y .633 se encuentra la probabilidad esperada de .500. Por otra parte, la razón por la que sumamos los casos del inicio (0) a un caso inmediatamente anterior (4) está en que utilizamos el complemento de la PC (1 - PC) como indicador del grado de significatividad. Por ejemplo, la PC de 9 caras dentro de 10 monedas es .989 y su complemento, .011 (1.1%). La probabilidad de 1.1% es bastante reducida por lo que se rechaza la hipótesis nula de que la moneda no es sesgada.

$x = 4$	.011	.998	.987
$x = 5$	.001	1.000	.998
$x = 6$	.000	1.000	1.000
$x = 7$	.000	1.000	1.000
$x = 8$	.000	1.000	1.000
$x = 9$	.000	1.000	1.000
$x = 10$	.000	1.000	1.000

Por ejemplo, cuando $x = 5$, $n = 10$, $e = 0.1$, S(5, 10, 0.1) resulta ser 0.998, es decir, la suma de las Probabilidades de Ocurrencia (PO) $x = 0, 1, 2, 4$ es 0.998. Por consiguiente, al establecer la norma de la Significatividad (S) en 0.99, el 99% de las ocurrencias corresponden a 0, 1, 2, 3, 4. No aparece casi nunca 5 en adelante (5, 6, 7, ...) y existe una escasa probabilidad de 0.01 (1%).

La tabla siguiente ofrece las Significatividades (S) de acuerdo con las Ocurrencias ($x = 1, 2, \ldots, 10$) y con las Probabilidades Esperada s ($e = 0.1, 0.2, \ldots, 0.9$):

S(x, 10, e)	$e = 0.1$	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
$x = 0$	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
1	0.349	0.107	0.028	0.006	0.001	0.000	0.000	0.000	0.000
2	0.736	0.376	0.149	0.046	0.011	0.002	0.000	0.000	0.000
3	0.930	0.678	0.383	0.167	0.055	0.012	0.002	0.000	0.000
4	0.987	0.879	0.650	0.382	0.172	0.055	0.011	0.001	0.000
5	0.998	0.967	0.850	0.633	0.377	0.166	0.047	0.006	0.000
6	1.000	0.994	0.953	0.834	0.623	0.367	0.150	0.033	0.002
7	1.000	0.999	0.989	0.945	0.828	0.618	0.350	0.121	0.013
8	1.000	1.000	0.998	0.988	0.945	0.833	0.617	0.322	0.070
9	1.000	1.000	1.000	0.998	0.989	0.954	0.851	0.624	0.264
10	1.000	1.000	1.000	1.000	0.999	0.994	0.972	0.893	0.651



Utilizamos la función de Excel BINOMDIST para obtener la Significatividad (S) de 0.349 en la celda de x = 1; e = 0.1:

s = S(*x, n, e*) = BINOMDIST(*x*-1, *n, e*, 1)

(*s*: Significatividad; *x*: Ocurrencia; *n*: Ensayos; *e*: Probabilidad Esperada)

0.349 ← S(1, 10, 0.1) = BINOMDIST(0, 10, 0.1, 1)

Esto quiere decir que al ensayar 10 veces del evento con la Probabilidad Esperada (PE) de 0.1, la ocurrencia 1 (*x* = 1) corresponde a la Significatividad (S) de .349. Las 2 ocurrencias (*x* = 2) del mismo evento corresponde a 0.736:

0.736 ← S(2,10,0.1) = BINOMDIST(1, 10, 0.1, 1)

3.13.3. Probabilidad Esperada

Hemos observado que la Significatividad (*s*) se obtiene por la función de S(*x, n, e*) o la función de Excel BINOMDIST:

s = S(*x, n, e*) = BINOMDIST(*x*-1, *n*, *e*, 1)

(*x*: Ocurrencias, *n*: Ensayos, *e*: Probabilidad Esperada)

Por la función S(*x, n, e*) se obtiene la Significatividad (*s*) por medio de *x*: Ocurrencias, *n*: Ensayos, *e*: Probabilidad Esperada . No obstante, en la práctica del análisis de datos lingüísticos, a diferencia de tales experimentos de la tirada de una moneda o el saque de una tarjeta, generalmente no se conoce la Probabilidad Esperada (PE) de los eventos desde el principio. Los parámetros que se conocen son *x*: Ocurrencias, *n*: Ensayos (Suma) y la Significatividad (*s*) se establece por la parte del usuario. Por esta razón, seguidamente elaboramos la función E que devuelva la Probabilidad Esperada (PE) por medio de *x*:Ocurrencias, *n*:Ensayos y *s*: Significatividad:

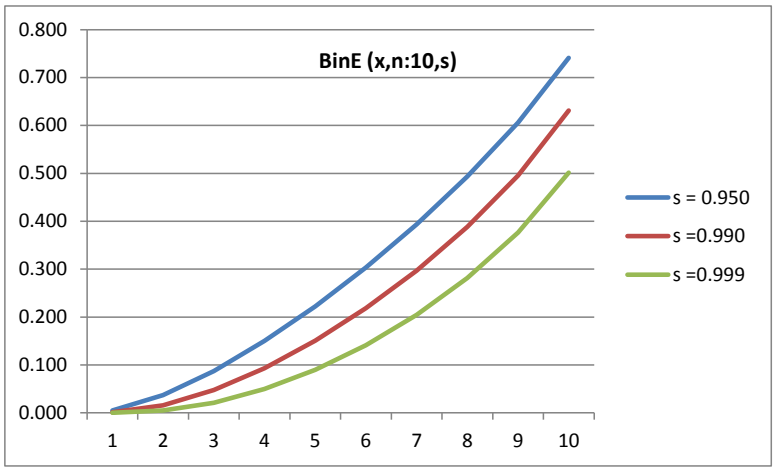
e = E(*x, n, s*) = E(1, 10, 0.95)

La función E(*x, n, s*) devuelve Probabilidad Esperada (PE) que se supone de un evento que ocurre *x* veces en *n* ensayos con la Significatividad *s*. Se trata de presuponer la Probabilidad Esperada (*e*) de, por ejemplo, 5 ocurrencias (*x* = 5) en 10 ensayos (*n* = 10) con la Significatividad (*s*) de, por ejemplo, 0.99 (*s* = 0.99). Con estos tres parámetros, la Probabilidad Esperada (PE) resulta calculable.

La tabla siguiente muestra las Probabilidades Esperada s (PE: *e*) de los eventos de 10 ensayos (*n* = 10), de acuerdo con las Ocurrencias (*x*) de 1 a 10 (*x* =

1, 2, ..., 10), y con distintas Significatividades (s): s = 0.95, 0.99, 0.999. En esta tabla observamos que cuanto mayor es la Ocurrencia (x), tanto mayor es la Probabilidad Esperada (e). Por ejemplo, en la Significatividad (s) de 0.99, la Probabilidad Esperada (PE) de x = 1 es 0.001, mientras que la de x = 10 es 0.631.

E(x, n:10, s)	s = 0.95	s = 0.99	s = 0.999
x = 1	0.005	0.001	0.000
2	0.037	0.016	0.005
3	0.087	0.048	0.021
4	0.150	0.093	0.050
5	0.222	0.150	0.090
6	0.304	0.218	0.141
7	0.393	0.297	0.205
8	0.493	0.388	0.282
9	0.606	0.496	0.376
10	0.741	0.631	0.501



Al mismo tiempo confirmamos que el aumento de la Significatividad (s) causa la disminución de la Probabilidad Esperada (e). Por ejemplo la E(5, 10, 0.95) es 0.222, mientras que la misma con la Significatividad (s) de 0.99 es 0.150 y la misma con la Significatividad de 0.999 es 0.090.

Supongamos que hemos tenido 2 veces éxito (x = 2) en 10 experimentos (n = 10). Con estos datos, sin embargo, no podemos esperar 20 éxitos en 100 experimentos futuros[17]. Veamos cómo se presentan las Probabilidades Esperada (e) al aumentar el número de experimentos n = 10, 100, 1000, ...:

[17] Como veremos inmediatamente, incluso no podemos esperar 2 éxitos en 10 próximos experimentos.

n	E(*n**0.2, *n*, 0.99)
n = 10	0.016
100	0.116
1,000	0.171
10,000	0.191
100,000	0.197
1,000,000	0.199
10,000,000	0.200
100,000,000	0.200
1,000,000,000	0.200

En la tabla anterior con la condición de que la Significatividad sea 0.99 (99%), al obtener 2 éxitos en 10 ensayos, su Probabilidad Esperada (e) es 0.016 (1.6%) y queda muy lejos de la probabilidad de éxito de 0.20 (20%). Cuando n = 10 000 llega a *e* = 0.191 (19.1%). De *n* = 10 000 en adelante, el aumento de la Probabilidad Esperada (PE) es reducido. Finalmente obtenemos *e* = 0.20 (20 %) al llegar a *n* = 10 000 000. Esta característica de la Probabilidad Esperada (PE) es importante, puesto que por ella podemos sabér qué probabilidad teórica hay en cada caso de 2 entre 10, 20 entre 100, 200 entre 1000, así sucesivamente. Nos llama la atención sobre todo los primeros casos donde la magnitud de la base 10, 100, 100 es reducida, lo que causa la poca Probabilidad Esperada (0.016, 0.114, 0.171).

Vamos a efectuar el mismo experimento cambiando la probabiliad esperada (e) en 0.100, 0.200, ..., 1.000.

BinE	n = 10	100	1 000	10 000	100 000	1 000 000
g = 0.10	0.001	0.042	0.079	0.093	0.098	0.099
0.20	0.016	0.116	0.171	0.191	0.197	0.199
0.30	0.048	0.198	0.267	0.289	0.297	0.299
0.40	0.093	0.287	0.364	0.389	0.396	0.399
0.50	0.150	0.381	0.463	0.488	0.496	0.499
0.60	0.218	0.479	0.563	0.589	0.596	0.599
0.70	0.297	0.582	0.665	0.689	0.697	0.699
0.80	0.388	0.691	0.769	0.791	0.797	0.799
0.90	0.496	0.809	0.876	0.893	0.898	0.899
1.00	0.631	0.955	0.995	1.000	1.000	1.000

e = BinE(n*g, n, 0.99)

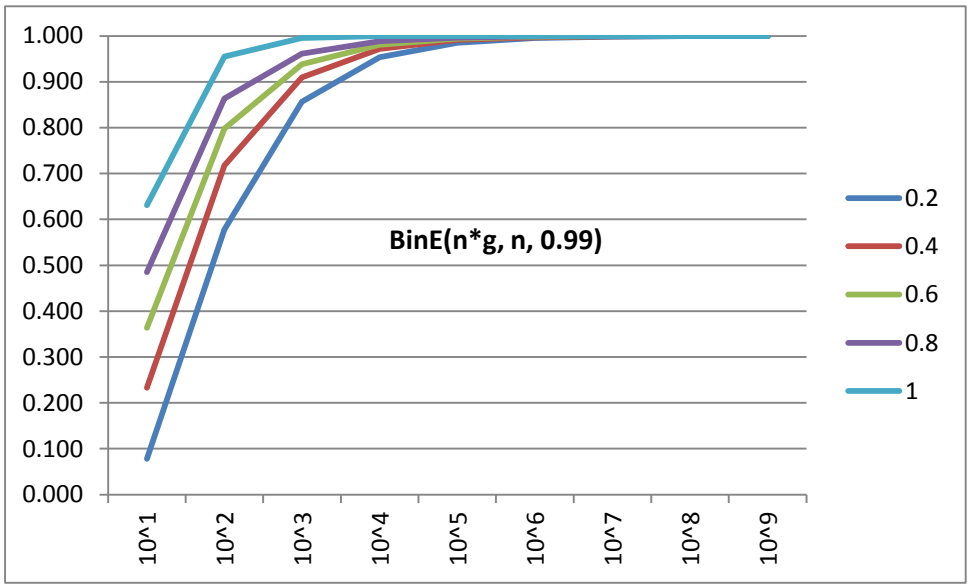
En la tabla anterior, observamos que se reduce el número de ensayos (n) para llegar al objetivo de 20%, 40%, ..., 100%, de acuerdo con la subida de la ratio de éxito (g). Por consiguiente, buscamos la ratio de cumplimiento del

objetivo dividiendo la probabilidad esperada (e) por la ratio de éxito (*g*).

BinE	n = 10	100	1 000	10 000	100 000	1 000 000
g = 0.10	1.0%	42.4%	79.1%	93.1%	**97.8%**	**99.3%**
0.20	7.8%	57.8%	85.7%	**95.4%**	**98.5%**	**99.5%**
0.30	15.8%	66.1%	88.9%	**96.5%**	**98.9%**	**99.6%**
0.40	23.3%	71.8%	91.0%	**97.1%**	**99.1%**	**99.7%**
0.50	30.1%	76.1%	92.6%	**97.7%**	**99.3%**	**99.8%**
0.60	36.4%	79.8%	93.9%	**98.1%**	**99.4%**	**99.8%**
0.70	42.4%	83.1%	**95.0%**	**98.5%**	**99.5%**	**99.8%**
0.80	48.5%	86.3%	**96.1%**	**98.8%**	**99.6%**	**99.9%**
0.90	55.1%	89.9%	**97.3%**	**99.2%**	**99.8%**	**99.9%**
1.00	63.1%	**95.5%**	**99.5%**	**100.0%**	**100.0%**	**100.0%**

BinE(n*g, n, 0.99) / g

Pot ejemplo, la ratio de cumplimiento del objetivo del caso [n=10: g=0.200] es 0.016 / 0.200 = 0.078 (7.8%). Esto quiere decir que dos éxitos dentro de 10 ensayos, si se desea el 99% de significatividad, no se asegura 20%, sino 1.6% (0.016), que llega tan solo a 7.8% del objetivo de 0.2000. Para alcanzar el 95% (0.950) de la ratio de cumplimiento de objetivo, se necesita aproximadamente 10,000 ensayos. La ratio de cumplimiento de objetivo (BinE / g) varía según g y cuanto más la ratio de éxito (g), en menos ensayos se llega a la ratio de cumplimiento, por ejemplo, 95%. Por ejemplo, cuando g = 0.8, en 1000 ensayos la ratio de cumplimiento llega al 96.1% (0.961).



En la tabla anterior y el gráfico, nos damos en cuenta de que cuando g = 1 (100%), se llega a 100% de la ratio de cumplimiento en 1000 ensayos, mientras cuando g = 0.2 (20%), se necesita 1000 000 000 ensayos. Se consigue

la ratio de cumplimiento de 95%, en lugar de 100%, cuando g = 0.2, 0.4, 0.6, y n = 10,000 y cuando g = 1 y n = 100. En cualquier caso, si g es reducida, se necesita gran cantidad de ensayo (n) para dar la seguridad suficiente (99%).

Por lo tanto, al tratar los dato de menos de 1000 (n), especialmente con la ratio de éxito (g) reducida, debemos tener cuidado en el manejo de la frecuencia relativa y normalizada. Los datos lingüísticos suele ser de poca probabilidad (g), por ejemplo la frecuencia de palabras o morfemas en menos de 1% (0.01) de la totalidad. En este caso recomendamos utilizar la probabilidad esperada, la base de la frecuencia probabilística, que suele ser más reducida que la ratio de éxito, pero siempre ofrece la significatividad asegurada de, por ejemplo, 99%.

3.13.4. Multiplicador

Intentamos calcular la «Frecuencia Probabilística» (FP) en forma de Probabilidad Esperada (*e*) * Multiplicador (*m*):

FP = *e* * *m* (*e*: Probabilidad Esperada , *m*: Multiplicador)

La Frecuencia Probabilística (FP) se obtiene por la función de la Probabilidad Esperada E(*x, n, s*) en combinación con el Multiplicador (*m*).

FP = *e* * *m* = E(*x, n, s*) * *m*
21.5 = E(3, 3, 0.99) * 100

Es conveniente que la cantidad del Multiplicador (*m*) sea de la magnitud similar de la Máxima de Suma o Total de Palabras en forma redondeada

La tabla inferior izquierda muestra la Frecuencia Absoluta (FA) y Suma vertical y la tabla derecha es de la Frecuencia Probabilística (FP) con el multiplicador (*m*) = 100. Ahora la Frecuencias Probabilísticas (FP) de [uoz:1200] llega a la cifra de 21.5, a diferencia de la Frecuencia Relativa (FA) de 3 entre 3: 3 / 3 * 100 = 100.0, que es incomparable con, por ejemplo [uoz:1250], FA = 8 / 12 * 100 = 66.7. Resulta que la Frecuencia Relativa (FR) de 3 entre 3 (1.000) es más alta que la de 8 entre 12 (.667), mientras que la Frecuencia Probabilística (FP) de 3 entre 3 es 21.5 y la de 8 entre 12 es 30.2, lo que demuestra la importancia mayor de 8 entre 12.

FA	1200	1250	1300	1350	1400
uoz	3	8	3	11	6
boz	0	3	8	18	35
voz	0	1	1	23	53
Suma	3	12	12	52	94

FP:100	1200	1250	1300	1350	1400
uoz	21.5	30.2	3.9	9.7	1.9
boz	.0	3.9	30.2	20.0	25.9
voz	.0	.1	.1	28.2	43.9

Hemos visto que la Frecuencia Relativa (FR) no es apropiada para realizar la evaluación comparativa de las cifras. En su lugar, hemos introducido la Frecuencia Probabilística (FP) con base de la Suma de las formas en comparación. Ahora se trata de la Frecuencia Probabilística (FP) calculada con el multiplicador 100, con el que se ha utilizado la Significatividad de .99 (99%), lo suficiente para ser bastante confiable. Sin embargo, nos sorprende la magnitud de 21.5 en el caso de 3 entre 3 y la de 30.2 en el caso de 8 entre 12. Son correctos dentro de la Significatividad de .99 (99%), que es condicionada por la cantidad del Multiplicador (100). Pensamos que esto es debido a la adoptación de Suma de las frecuencias de las formas en cuestión como base de comparación. Veamos la posibilidad de la otra base, también muy usual en los estudios lingüísticos de corpus, la cantidad total de palabras o letras.

La tabla inferior izquierda es de la Frecuencia Absoluta (FA) y el número total de palabras (TP):

FA	1200	1250	1300	1350	1400
uoz	3	8	3	11	6
boz	0	3	8	18	35
voz	0	1	1	23	53
TP	7 736	36 052	40 957	64 999	96 059

FP:10^5	1200	1250	1300	1350	1400
uoz	5.6	8.1	1.1	7.3	1.9
boz	.0	1.2	7.1	14.8	23.7
voz	.0	.0	.0	20.5	39.1

La Frecuencia Normalizada (FN), que en los estudios lingüísticos de corpus suele utilizarse con el número total de palabras (TP) en forma de, por ejemplo, 3 / 7 736 * 100 000 = 38.8 en [uoz: 1200] presupone que 3 entre 7 736 corresponda a 38.8 entre 100 000, por la fórmula de proporción:

3 / 7 736 = 38.8 / 100 000

Sin embargo, desde el punto de vista probabilístico esta presuposición no es fiable, lo mismo que la presuposición de que 3 éxitos en 10 ensayos correspondan a 30 éxitos en 100 ensayos, lo que es la base del porcentaje, como hemos visto anteriormente. En la práctica de comparación de cifras, la Frecuencia Probabilística FP es más confiable, con la que se ve un paulatino desplazamiento de <uoz> (1200-1250) por <boz> (1300) hasta <voz> (1350-1400). La misma observación es posible con la Frecuencia Probabilística (FP) de Suma con el multiplicador 100. Sin embargo, la tabla de la Frecuencia

Probabilística (FP) con el número total de palabras (TP) da las cifras más realistas.

Programas

```
Function BinS(x, n, e)
 'Significatividad (x: ocurrencia, n: ensayos, e: probabilidad expectativa)
 If x = 0 Then BinS = 0: Exit Function
 BinS = Application.BinomDist(x - 1, n, e, 1)
End Function
```

```
Function BinE(x, n, s) 'Probabilidad Esperada (H. Ueda 2017)
 '(x: ocurrencia, n: ensayos, s: significatividad)
 Dim i, k, r, mn, mx, sv: BinE = 0: k = 0: If x = 0 Then Exit Function
 r = 10 ^ 6: mn = 0: mx = r
 'r: precisión, mn: mínimo, mx: máximo de búsqueda binaria
 Do
    i = (mx + mn) / 2 'Mitad entre mx y mn
    BinE = i / r 'Candidato de la Probabilidad Esperada
    sc = BinS(x, n, BinE) 'sc: Significatividad del cantidato
    If sc < s - 1 / r Then 'Si sc no llega a s-1/r...
       mx = i 'Bajar el máximo de búsqueda al punto medio (i).
    ElseIf sc > s + 1 / r Then 'Si sc sobrepasa a s-1/r...
       mn = i 'Subir el mínimo de búsqueda al punto medio (i).
    Else 'Si sc se encuentra en s±5/r
       Exit Do 'Salir del bucle.
    End If
 Loop
End Function
```

En el programa (Microsoft Excel VBA) hemos utilizado la técnica de la búsqueda binaria con aplicación del encuentro en el rango. La búsqueda secuencial es impracticable por el tiempo que cuesta en llegar al encuentro.

(*) Frecuencia Probabilística de un ensayo

Cuando m =1 en la Frecuencia Probabilística FP (x, n, m, s), la misma función devuelve la Probabilidad Expectativa calculada en BinE(x, n, s):

PF(x, n, m, s) = BinE(x, n, s) * m
PF(x, n, 1, s) = BinE(x, n, s) * 1 = BinE(x, n, s)

Según la tabla siguiente, Caso A: Probabilidad Expectativa de un ensayo del experimento que ha tenido 9 veces de éxitos en 10 ensayos es más alta que Caso B: Probabilidad Expectativa de un ensayo del experimento que ha tenido 80 veces de éxitos en 100 ensayos.

Evento	s: .95	s: .99
Caso A: Probabilidad Expectativa de un ensayo del experimento que ha tenido 9 veces de éxitos en 10 ensayos (90%)	.606	.496
Caso B: Probabilidad Expectativa de un ensayo del experimento que ha tenido 80 veces de éxitos en 100 ensayos (80%)	.723	.691

De esta manera el número de ensayos (n) es importante. Veamos el movimiento de la Probabilidad Expectativa (e) de acuerdo con el número de pruebas (n): 10, 20, ..., 100 de una hierba medicinal con la ratio de éxito de 90%:

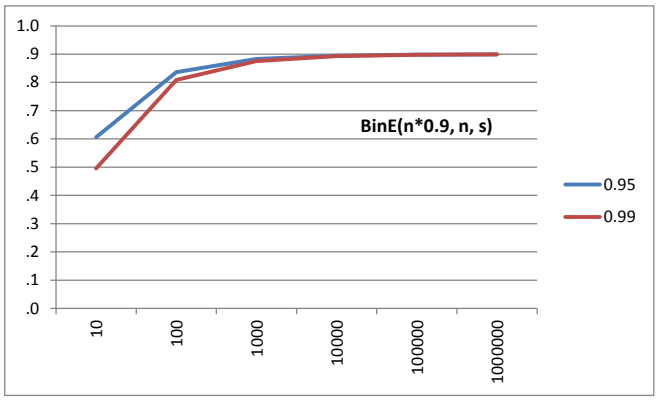
BinE(n*0.9, n, s)	s = 95%	99%
n = 10	0.606	0.496
20	0.717	0.642
30	0.761	0.702
40	0.786	0.736
50	0.801	0.758
60	0.812	0.774
70	0.820	0.786
80	0.827	0.795
90	0.832	0.803
100	0.836	0.809

.606 = PF(10*0.9, 10, 0.95)

De acuerdo con la tabla anterior, aun sabiendo que la hierba medicinal tiene 90% de éxito, si el mismo experimento se ha llevado a cabo solo 10 veces con 9 veces de éxito, la Probabilidad Expectativa de éxito en un medicamento es tan solo 0.606 (60.6%) con la Significatividad de 95%. Si exigimos hasta 99% de Significatividad, la Probabilidad Expectativa desciende a 0.496 (49.6%), es decir menos de la mitad. La Probabilidad Expectativa (e) asciende según el número de ensayos (n). Sin embargo, aun con 100 experimentos, la Probabilidad Expectativa no alcanza a 90%.

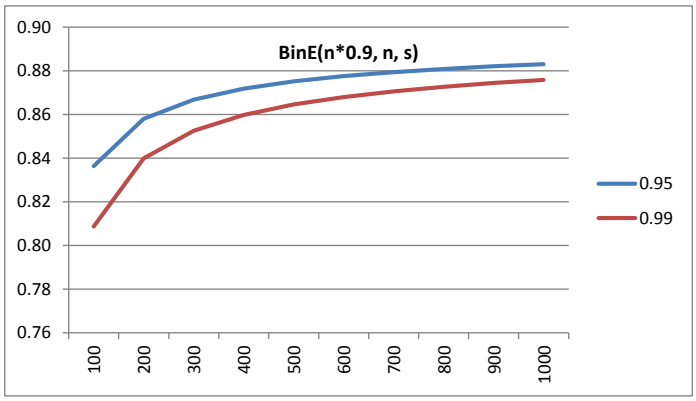
Para conseguir la Probabilidad Expectativa (e) más alta, se puede pensar tanto en el aumento del número de experimentos (n) como en el número de éxitos (x). Primero veamos la primera posibilidad: el aumento de experimentos. Según la tabla siguiente, es necesario contar con 1000 experimentos para obtener la Probabilidad Expectativa deseada de 90% aproximadamente.

BinE(n*0.9, n, s)	s = 95%	99%
n = 10	0.6058	0.4957
100	0.8363	0.8087
1000	0.8830	0.8758
10000	0.8949	0.8928
100000	0.8984	0.8978
1000000	0.8995	0.8993



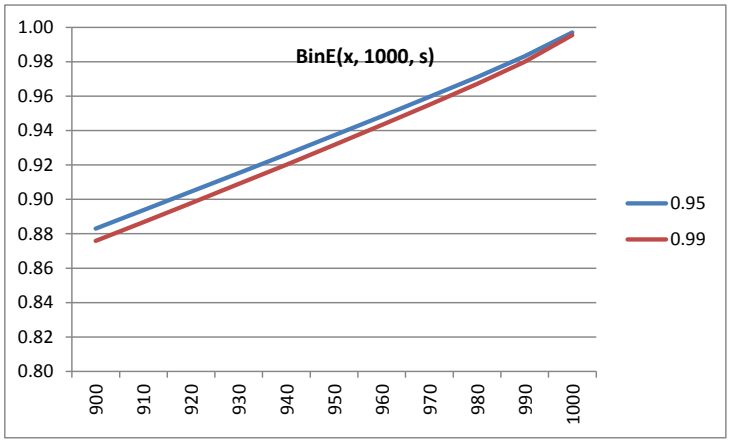
Persiguiendo el movimiento de la Probabilidad Expectativa según n = 100, 200, …, 1000, notamos que la subida inicial es fuerte y la misma ralentiza de manera paulatina posteriormente:

BinE(n*0.9, n, s)	95%	99%
100	0.8363	0.8087
200	0.8580	0.8399
300	0.8668	0.8526
400	0.8718	0.8598
500	0.8751	0.8645
600	0.8775	0.8680
700	0.8794	0.8706
800	0.8808	0.8727
900	0.8820	0.8744
1000	0.8830	0.8758



Al fijar el número de experimentos en 1000, la relación entre el número de éxitos (x) y la Probabilidad Expectativa (e) es corrlativa en forma lineal:

BinE(x, 1000, s)	95%	99%
900	0.8830	0.8758
910	0.8937	0.8868
920	0.9044	0.8978
930	0.9153	0.9090
940	0.9262	0.9202
950	0.9371	0.9316
960	0.9482	0.9432
970	0.9595	0.9549
980	0.9711	0.9671
990	0.9831	0.9800
1000	0.9970	0.9954



Estos experimentos se han llevado a cabo con la premisa de que el éxito y el fracaso de cada experimento de la hierba son independientes uno de otro. Bajo la misma premisa, las 9 veces de éxito en 10 experimentos no garantiza el éxito de 90%, sino que se limita a presentar el 49.5% de éxito con la

Significatividad de 99%. En cambio, un tenista bien preparado, al obtener 9 éxitos en 10 partidos, no poseería tan solo 49.6% del éxito en el próxima partido. A diferencia de la hierba desconocida de su efecto medicinal, el tenista entrenado debería poseer la taza de éxito bastante constante.

(*) Nota sobre el uso de la Frecuencia Probabilística

Suponemos que en una investigación de campo preguntamos a 5 personas sobre el uso de una determinada palabra y 4 de ellas contestan que sí la utilizan. Si la población total del punto encuestado es 1000, y calculando la Frecuencia Probabilística con 99% de Significatividad, llegamos a la cifra siguiente:

PF(4, 5, 1000, 0.99) = 222.1

que representa tan solo 22.2% de la población. No nos convence. Incluso en otra palabra, suponemos que todas las 5 personas encuestadas han dado la respuesta afirmativa. Calculamos su Frecuencia Probabilística:

PF(5, 5, 1000, 0.99) = 398.1

Aun con la respuesta afirmativa total, llegaríamos a la conclusión de que un 39.8% de la población utiliza la misma palabra. Nos parece raro, puesto que preguntando a 5 cubanos si utilizan la palabra «guagua» para autobús, todas las 5 personas contestarán que sí. Y desde el punto de vista probabilística, se supone el uso de la misma palabra en 39.8% de la población a pesar del uso general de la misma palabra en Cuba.

El uso lingüístico de una comunidad humana está basado en una base común de códigos. Por esta razón, si alguien llama «guagua» al autobús, muy probablemente otra persona lo llama así también. En cambio la Frecuencia Probabilística está calculada para tales fenómenos accidentales como la tirada de una moneda o saque de una tarjeta con ojos cerrados. No hay ninguna relación entre el resultado de la primera tirada de la moneda y el de la segunda. La cara salida de la primera tirada no condiciona la de la segunda de ninguna manera. Las dos son independientes.

En cambio cuando las ocurrencias del fenómeno no son independientes uno del otro, como fenómenos lingüísticos, no podemos utilizar la Probabilidad Expectativa ni la Frecuencia Probabilística para obtener la cantidad concreta inferencial. El resultado de las marcas de un futbolista tampoco será completametne independiente unas de las otras, porque el resultado de un partido influiría al del siguiente. En tal caso, tenemos que calcular las cifras probabilísticas suponiendo que sean independientes y considerarlas como datos de referencia no definitiva.

Por otra parte, la Frecuencia Probabilística es útil cuando comparamos las frecuencias con bases muy diferentes. Aun cuando las ocurrencias de varias palabras en comparación no son independientes, se puede comparar las Frecuencias Probabilísticas calculadas con las perfectamente mismas condicioens (ocurrencias, totalidad, multiplicador, Significatividad).

En internet encontramos páginas que ponen ranking de evaluación de gustos personales: «me gusta» o «no me gusta»; «SÍ» o «NO». A veces encontramos un item (A) que ha recibido 8 personas de «SÍ» y 2 de «NO» (80%) y otro item (B) de 55 «SÍ» y 45 «NO» (55%) en este orden de porcentajes (Frecuencia Relativa). Al calcular la Frecuencia Probabilística (FP) de los dos items obtenemos el resultado siguiente:

Item A: FP(8, 10, 1, 0.99) = BinE(8, 10, 0.99) = 0.388
Item B: FP(55, 100, 1, 0.99) = BinE(55, 100, 0.99) = 0.429

De modo que es más razonable poner el item B delante de A en el ranking, en contra del orden por el porcentaje, a pesar de que aquí tampoco se garantiza la independencia total entre las respuestas de la encuesta.

(#) Robustez de la Frecuencia Probabilística: Formas españolas «del» y «al»

En las lenguas románicas, retorromance, italiano, portugués, catalán, francés y español, a excepción de rumano, se encuentra multitud de contracciones de preposición y artículo definido. Dentro de ellas, el español posee solamente dos formas «del» y «al» y en otras combinaciones se separan las dos palabras, *de la, a la, en el,* etc. En los estudios de lingüística general e historia de la lengua española, está explicado que las formas de «de el» y «a el» por su coocurrencia frecuente se han contraído en «del» y «al». Sin embargo, en el Corpus CODEA a partir de 1200 en adelante no se encuentran las formas separadas, el punto de partida de la contracción supuestamente frecuente, más que de manera excepcional. En cambio se encuentran numerosas formas tanto unidas como separadas:

Forma / Fecha	1200	1300	1400	1500	1600	1700
de el	6	2	0	19	51	16
del	1920	1829	2247	2858	1358	426
de la	309	110	145	370	451	171
dela	957	992	1303	1590	427	171

Creemos que desde el punto de vista histórico no ha habido tales procesos de contracción como de *el* > *del, a el* > *al*. Si las formas *la, los, las* con la aféresis de *e-* inicial se debe a la combinación con preposiciones anteriores

(Menéndez Pidal, 1926: 331), y si nos fijamos especialmente en la combinación con la preposición «de», deberían haber existido las formas unidas *dela, delos, delas* antes del nacimiento de las formas actuales del artículo definido: *la, los, las*. Naturalmente la formación de «del» y «al» deben ser coincidentes con las formas unidas: *dela, delos, delas; ala, alos, alas*. Por lo tanto, pensamos que las formas contraídas existieron desde el principio de la historia de la lengua española y no son productos del proceso de la contracción de las formas separadas frecuentes.

Para realizar la observación general de la tendencia de las formas separadas y unidas, es posible utilizar las Frecuencias Absolutas que acabamos de ver. No obstante, los números de las palabras encontradas de cada 100 años tratados son diferentes como muestra la tabla siguiente:

Fecha	1200	1300	1400	1500	1600	1700
Palabras	224,708	230,383	261,564	287,380	125,366	52,938

Por consiguiente, a partir de las Frecuencias Absolutas y las Frecuencias totales de palabras, calculamos las Frecuencias Probabilísticas siguientes (con el multiplicador de 100,000):

FP	1200	1300	1400	1500	1600	1700
de el	.8	.1	.0	3.6	28.6	15.5
del	809.9	751.5	817.6	951.9	1016.4	717.1
de la	120.0	37.8	45.3	113.7	321.6	268.4
dela	394.6	399.5	466.7	521.6	303.5	268.4

Nos fijamos en la Frecuencias Probabilística (FP) de «de el» en 1500, donde se ha disminuido bastante con respecto a la FP del mismo en 1600 y 1700. En la Frecuencia Absoluta, la cifra de 1500 se registraba más alta que en 1700. Sin embargo, la tendencia en la Frecuencia Probabilística es inversa, es decir, se registran más altas en 1600 y 1700, lo que puede ser una evidencia negativa del supuesto proceso de contracción, puesto que naturalmente el proceso de contracción presupondría las altas frecuencias de las formas separadas más en las fechas anteriores que en las posteriores.

Estas observaciones no varían tampoco en las Frecuencias Normalizadas siguientes, puesto que las bases de división no ofrecen diferencias significativas menos en 1600 y, especialmente, 1700.

NF	1200	1300	1400	1500	1600	1700
de el	2.7	.9	.0	6.6	40.7	30.2
del	854.4	793.9	859.1	994.5	1083.2	804.7
de la	137.5	47.7	55.4	128.7	359.7	323.0

dela	425.9	430.6	498.2	553.3	340.6	323.0

Al comparar la Frecuencia Normalizada (FN) y la Frecuencia Probabilística (FP), la NF de «de el» en 1600 es más alta que la FP y en 1700, la NF es casi doble de la FP. A pesar de estas diferencias, la observación general no varía sustancialmente. El uso de la FN no causa gran problema al tratar los datos de las bases parecidas, pero sí en los datos de las bases diferentes. En cambio, la FP modifica las diferencias de las bases utilizando la teoría probabilística, de modo que no produce problemas tanto en los datos de bases parecidas como los de bases diferentes. Podemos afirmar que la FP es robusta en el sentido de que es aplicable a los datos normales como a los datos especiales con bases desiguales.

*CODEA (Corpus de Documentos Españoles Anteriores a 1800):
http://corpuscodea.es/
*Menéndez Pidal, R. 1926-1980. *Orígenes del español.* Madrid. Espasa-Calpe.

4. Relación

Demostraremos las relaciones que existen entre las variables de los datos utilizando los coeficientes de correlación, asociación y orden. También veremos la manera de cuantificar las distancias que hay entre los individuos o elementos. En general trataremos las matrices dotadas de números continuos o de datos cualitativos de forma de 1/0 o verdadero / falso. Por otra parte proponemos también analizar los datos consistentes en letras, propios de análisis de datos lingüísticos.

Todas las operaciones matriciales son posibles tanto en filas como en columnas. Explicamos por defecto los procesos de columnas en la mayoría de las secciones siguientes. Solamente en la Distancia, haremos cálculos de filas. Sin embargo, con la matriz traspuesta, se hace también en otra dirección. La única excepción es la Distancia Mahalanobis, que exige un dato de más filas que columnas.

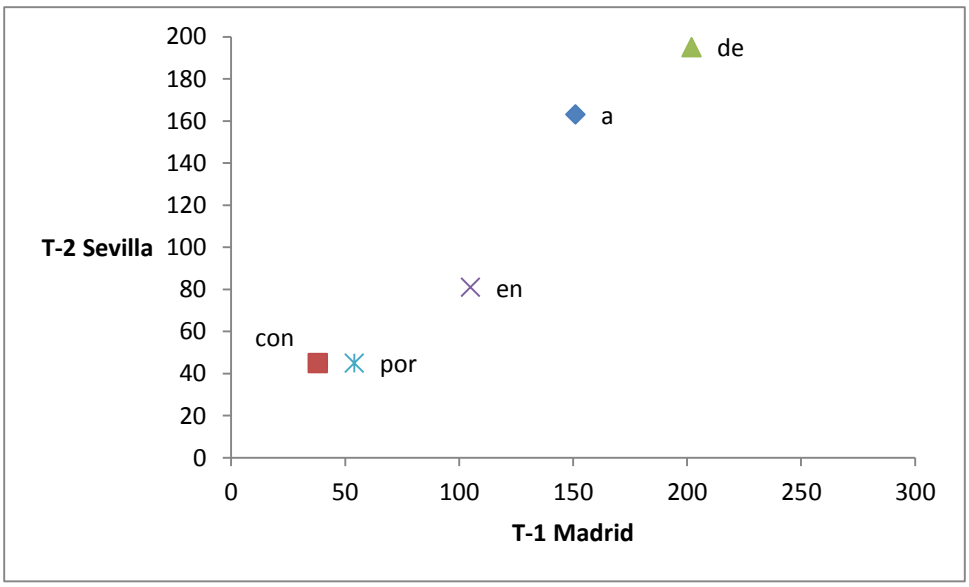
4.1. Correlación

4.1.1. Coeficiente de correlación

El dato siguiente muestra las frecuencias de las preposiciones españolas en los textos de T-1 (Madrid) y T-2 (Sevilla).

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

Nos interesa el grado de correlación que hay entre T-1 y T-2. Veamos el gráfico de distribución constituido del eje horizontal de T-1 y del vertical de T-2:



Así notamos que existe una fuerte relación proporcional entre T-1 y T-2, puesto que cuanto más aumenta el valor de T-1, tanto más aumenta también el valor de T-2. Para obtener el grado cuantificado de la correlación entre los dos textos, calculamos los Puntos Estandarizados (PE), que hemos visto en el Capítulo anterior.

$M1p = (I1p\ Xnp) / N$ ←fila de Medias verticales

$DT1p = [I1p\ (Xnp - M1p)^2 / N]^{1/2}$

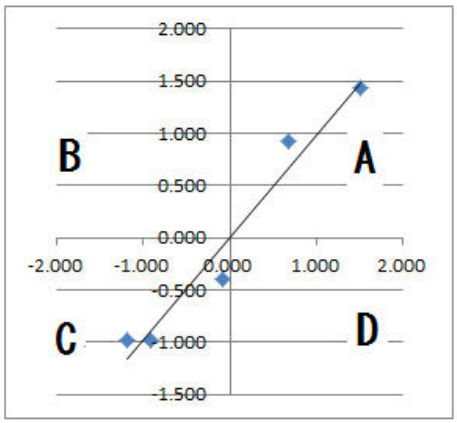
←fila de Desviaciones Típicas verticales

$PEnp = (Xnp - M1p) / SD1p$ ←Matriz de Puntos Estandarizados

La tabla siguientes es de los Puntos Entandarizados, cuya Media es 0 y Desviación Típica 1:

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

Dibujamos de nuevo el gráfico de distribución ahora con los Puntos Estandarizados:



Ahora podemos observar que los puntos están situados alrededor del eje vertical 0 y el horizontal 0, que dividen la area total en cuatro secciones: A, B, C, D. En los puntos que se encuentran en las áreas A y C, los signos de valores «+» y «-» coinciden, de modo que su multiplicación salen positivas «+». Los puntos que se encuentran en las áreas B y D, sus signos son opuestos «+/-» o «-/+», de modo que su multiplicación produce valores negativos «-».

Por esta razón, si sumamos todos estos productos de multiplicación, obtendremos una cifra que indica el grado de correlación entre T-1 y T-2. La distribución lineal que indicamos en el gráfico anterior con una línea ofrecerá un valor máximo de la correlación.

Si hay datos que se sitúan en las áreas de B y D, estos datos disminuyen el grado de correlación y si todos los puntos se encuentran el estas áreas el grado de correlación se vuelve negativo. Con la distribución homogénea de los puntos en las cuatro áreas, el grado de correlación se vuelve nulo, es decir, cero.

Como la Suma de los productos de multiplicación está influida por el número de puntos (N), se divide la Suma por N, es decir buscamos la Media de los productos. Esta es la fórmula de «Coeficiente de Correlación» (CC):

$$CC = \Sigma i\,[(Xi - Mx)/SDx]*[(Yi - My)/SDy] / N \leftarrow \text{def.}$$

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

$$\leftarrow \text{SDx, SDy al exterior}$$

$$CC = SSx_{n1}{}^{T}\ SSy_{n1} / N$$

$$\leftarrow \text{For/mula matricial; PE: Puntos Estandarizados}$$

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

X

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

/ 5

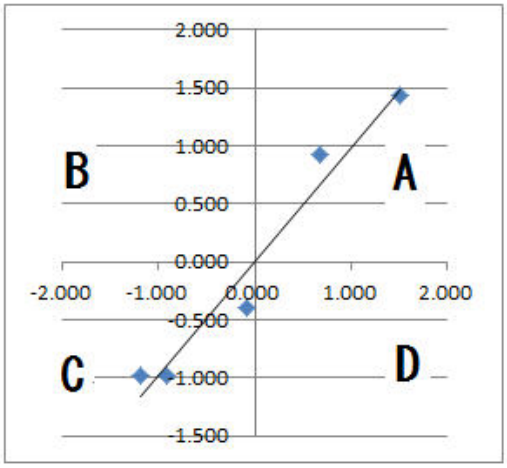
El proceso del cálculo y el resultado son[18]:

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

(*) Rango del coeficiente de correlación

El Rango del «Coeficiente de Correlación» (CC) es [-1 ~ 1]. El Máximo se presenta cuando los Puntos Estandarizados se encuentran en la línea recta de tal manera mostrada en el gráfico. La fórmula de esta línea es:

$$Yn1 = a\ Xn1 + b$$



Como hemos visto en el Cap. anterior, en la sección de la Propiedad de los «Puntos Estandarizados» (PE), los PE de los datos y los PE de los datos multiplicados por a y sumados por b son iguales. De modo que el CC entre Xn1 y (a Xn1 + b) es lo mismo que el CC entre Xn1 y Xn1, que se llama «Autocorrelación». Veamos el CC de Autocorrelacion:

$$CC(X, X) = PEx_{n1}{}^T\ PEx_{n1} / N \qquad \leftarrow \text{def.}$$

[18] Es un dato de ejemplo, si más. Como veremos más adelante, el resultado del cálculo de CC no es fiable, puesto que la distribución proporcional puede ser accidental.

$$= [(Xn1 - M) / SD]^T [(Xn1 - M) / SD] / N \quad \leftarrow \text{def. de PE}$$
$$= \{ \Sigma [(Xi - M) / SD]^2 \} / N \quad \leftarrow \text{elevación a 2}$$
$$= \{ \Sigma [(Xi - M)^2 / SD^2] / N \quad \leftarrow \text{distribuir el exponente 2}$$
$$= \{ \Sigma [(Xi - M)^2 / V] / N \quad \leftarrow \text{Varianza(V)} = SD^2$$
$$= \Sigma [(Xi - M)^2 / N / V \quad \leftarrow \text{V al exterior}$$
$$= V / V = 1 \quad \leftarrow \text{def. de Varianza(V)}$$

Como hemos visto en la Sección de la Propiedad de Puntos Estandarizados, si se multiplica por –a, en lugar de a, los PE se vuelve –PE.

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

lo que demuestra la inclinación a la derecha inferior en el gráfico anterior. Esto se llama la «Correlación Negativa Perfecta».

Por lo tanto el Coeficiente de Correlación (CC) posee el Rango entre el Mínimo de –1 y el Máximo de 1.

(*) Interpretación de coeficiente de correlación

Empíricamente se considera que se puede dar una interpretación del coeficiente de correlación de manera siguiente[19]:

$\|r\| = 0.0$	No existe la correlación entre X e Y.
$0.0 < \|r\| \leqq 0.2$	Casi no existe la correlación entre X e Y.
$0.2 < \|r\| \leqq 0.4$	Existe la correlación débil entre X e Y.
$0.4 < \|r\| \leqq 0.7$	Existe la correlación algo fuerte entre X e Y.
$0.7 < \|r\| \leqq 1.0$	Existe la correlación fuerte entre X e Y.

(*) Cautelas sobre el coeficiente de correlación

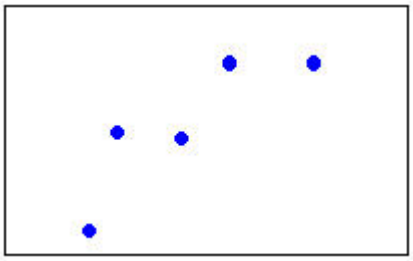
Calculando el «Coeficiente de Correlación» (CC), aparentemente podemos apreciar la relación entre los datos. Sin embargo, tenemos que tener cuidado a la hora de dar interpretaciones al respecto. Sobre todo es peligroso realizar unas interpretaciones a la ligera en los casos siguientes:

(0) Como hemos visto, si se comparan los datos, derivables unos de otros, el CC da lógicamente el valor 1 (ó -1).

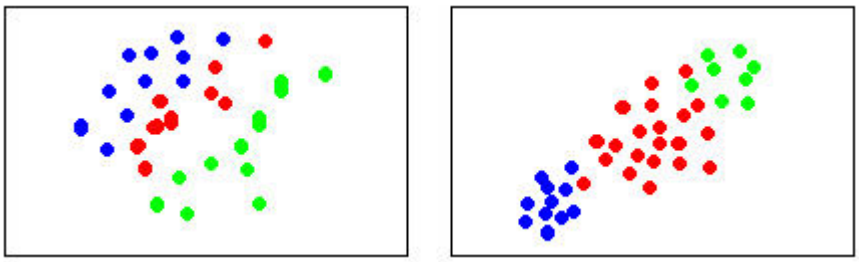
(1) No es fiable si se intenta calcular con número sumamente reducido de datos. Por ejemplo, el siguiente gráfico puede ser resultado accidental[20]:

[19] Utilizamos el signo de valor absoluto para tratar tanto la correlación positiva como la negativa.

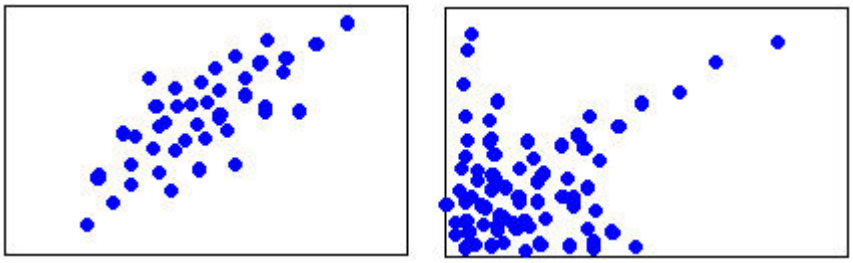
[20] Efectivamente el ejemplo anterior de cinco preposiciones no es apropiado al



(2) A veces ocurre que en el conjunto de datos, existen grupos heterogéneos correlacionales, cuyo conjunto presenta una correlación baja. Véase el gráfico inferior izquierdo. También ocurre que se presenta una correlación aparente en el conjunto de los grupos en que no se observa correlación (gráfico inferior derecho):



(3) Debemos prestar atención a los datos de distribución peculiar. El cálculo del CC es idóneo para los datos que se reúnen mayoritariamente alrededor de la Media con extensiones proporcionalmente desminuidas en ambos lados como muestra el gráfico inferior izquierdo.



Los datos lingüísticos, como hemos visto anteriormente, suele presentar una distribución de forma L y en la correlación entre dos datos, el gráfico presenta una forma peculiar de la imagen superior derecha.

Para analizar los casos concretos, hay que ver el gráfico de distribución para no equivocarse en la interpretación del CC.

Por otra parte, la correlación no garantiza necesariamente la relación causa – efecto. Por ejemplo, entre las horas del estudio y el resultados de

---

ana/lisis de correlacio/n. Lo hemos utilizado simplemente para demostrar las cifras concretas para explicar su correlacio/n aparente.

examen se puede presentar una correlación fuerte[21]. Sin embargo, las horas de estudio no pueden ser necesariamente la causa de la mejora de las puntuaciones de los exámenes. En el fondo de la correlación aparente, podemos ver una base común con las horas de estudio y los resultado de exámenes: interés en los estudios.

Hemos visto solamente operaciones matemáticas del «coeficiente de correlación» (CC). Podemos equivocarnos si no conocemos la esencia de los datos. Hay que tener cuidado cuando un análisis presenta solo cifras de CC, sin presentar los gráficos de distribución. Nosotros los estudiantes de letras tenemos que conocer todos los textos objetos de análisis. Si estamos enterados de qué va el texto, ante unos resultados extravagantes, podemos darnos cuenta de los errores en la formulación matemática e indagar sus causas para avanzar en los estudios posteriores.

4.1.2. Matriz de correlación

Procedamos a presentar la «Matriz de Correlación» que presenta los Coeficientes de Correlación entre múltiples variables (R_{pp}). Para esto necesitamos preparar la Matriz de Puntos Estandarizados (Z_{np}):

$$Z_{np} = (D_{np} - M1p) / DT1p$$

donde M1p es la fila de Medias verticales y DT1p es la fila de Desviaciones Típicas. Si multiplicamos Z_{np} transpuesta y Z_{np}, obtenemos la Matriz de Correlación (R_{pp}):

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

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

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

Rpp	v1	v2	v3
v1	1.000	.643	-.335
v2	.643	1.000	-.545
v3	-.335	-.545	1.000

Como esta fórmula es importante, veamos los elementos de las dos Matrices en multiplicación:

[21] Se puede calcular el coeficiente de correlación tratándose de las unidades distintas: horas y puntuaciones, puesto que para el cálculo utilizamos los Puntos Estandarizados, libres de las unidades.

$$Rpp = Z_{np}{}^{T}\, Z_{np}$$

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

$$= \begin{bmatrix} r_{11} = \text{fila } 1 * \text{col. } 1 & r_{12} = \text{fila } 1 * \text{col. } 2 & r_{13} = \text{fila } 1 * \text{col. } 2 \\ r_{21} = \text{fila } 2 * \text{col. } 1 & r_{22} = \text{fila } 2 * \text{col. } 2 & r_{23} = \text{fila } 2 * \text{col. } 3 \\ r_{31} = \text{fila } 3 * \text{col. } 1 & r_{32} = \text{fila } 3 * \text{col. } 2 & r_{33} = \text{fila } 3 * \text{col. } 3 \end{bmatrix}$$

Por la operación de multiplicaicón matricial:

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

De esta manera constatamos que R_{pp} se obtiene por la Suma de los Productos de multiplicaciones de los elementos de la Matriz Znp, los elementos de la parte diagonal son la Suma de cuadrados de las columnas, los elementos de la parte fuera de la diagonal son las Sumas de productos de multiplicación, que son coeficientes de correlación, la Matriz es simétrica cuyo tamaño es la dimensión de las filas de la primera Matriz de multiplicación o, lo que es lo mismo, la dimensión de las columnas de la segunda Matriz.

(#) «E» átona inicial y «e» átona final

La secuencia inicial latina de «s + consonante» (sC-) se convierte en español en «es + consonante», por ejemplo, *stare* > *estar, scribere* > *escribir.* La frecuencia de este fenómeno, agregación de e átona, es relativamente baja en Navarra y Aragón, dos regiones orientales de España. Por otra parte, se encuentran con frecuencia las formas con caídas de –e final detrás de dos consonantes (-CC), *present, veint,* etc., también en estas regiones. Las tres primeras columnas de la tabla siguiente corresponden a las frecuencia de las formas de *(e)star y (e)scribir* y sus derivados que aparecen en Castilla la Vieja (CV), Navarra (NA) y Aragón (AR) en los 1500 documentos notariales emitidos entre 1200 y 1680, en intervalos de 20 años. Las últimas tres columnas son de las formas con caída vocálica final de palabra de las voces de *present(e), veint(e), adelant(e), part(e), est(e), end(e)*:

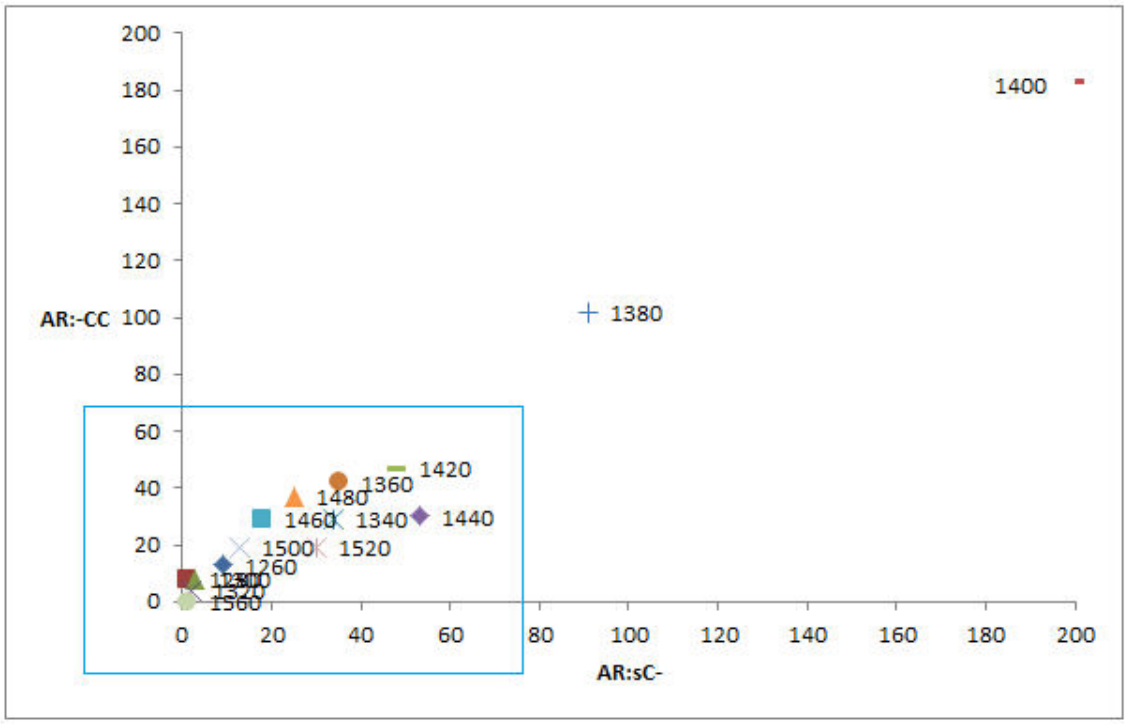
年 : Año	CV:sC-	NA:sC-	AR:sC-	CV:-CC	NA:-CC	AR:-CC
1200						
1220			4	1		1
1240			4	8		7
1260		5	9	5	22	13
1280		8	1	5	27	8
1300		8	3	2	34	8
1320	3	2	2		10	4
1340	1	1	34	1	25	29
1360		1	35		25	42
1380	4	2	91	2	2	102
1400		12	200	3	35	183
1420	4		48			47
1440			53	1	15	30
1460			18			29
1480	5	1	25	3	4	37
1500	3		13	1		19
1520	5		30			19
1540	5					
1560	19		1			
1580	35					
1600	1					
1620	9					
1640	4					
1660						
1680	4					

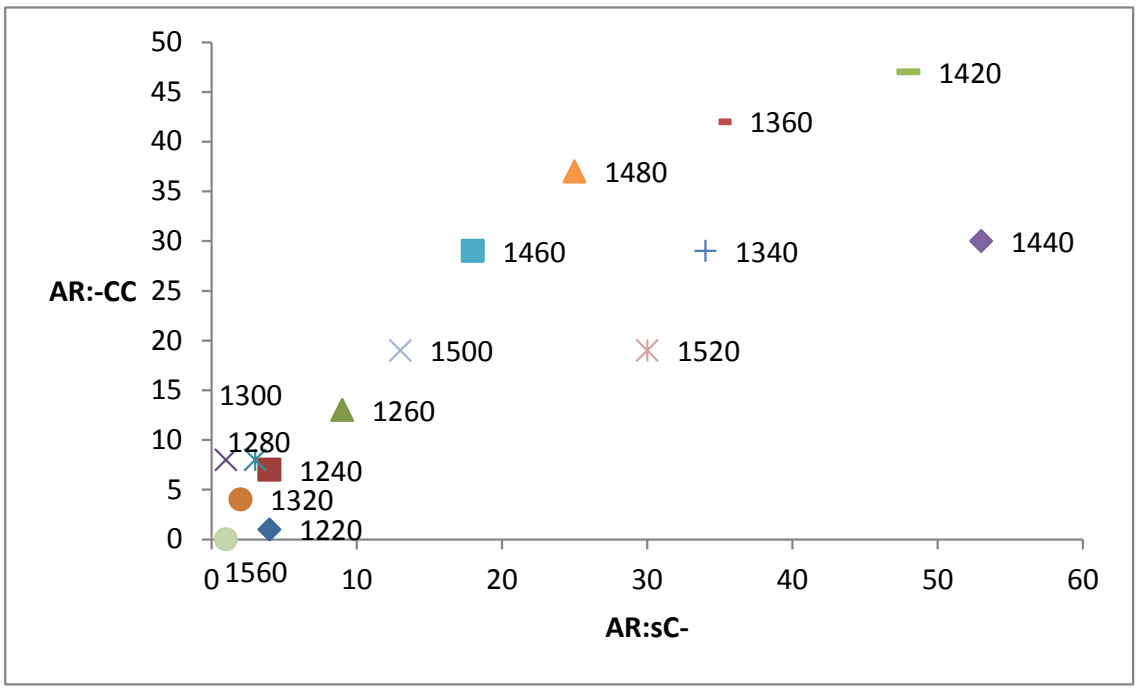
Ambos fenómenos se tratan de la vocal <e> átona y nos preguntamos si existe alguna correlación temporal entre los dos, una en la posición inicial de palabra y la otra en la final. La tabla siguiente es la Matriz de correlación de las seis columnas de la tabla anterior:

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

Efectivamente encontramos valores altos en los dos coeficientes de

correlación entre sC- y –CC en Navarra (.829) y en Aragón (.984). Castilla la Vieja no presenta nada destacable. Los dos gráficos siguientes muestran la distribución de sC- y –CC en Aragón. Al observar el primer gráfico, notamos que los dos puntos extravagantes (1380 y 1400) han influido mucho en la subida del coeficiente de correlación (.984). No obstante, eliminados estos dos puntos, también los restantes siguen presentando la alta correlación cuyo coeficiente resulta ahora .863 en el gráfico inferior:





Los estudios previos han indicado que la apócope extrema es propia de los primeros decenios del siglo XIII, causada por los inmigrantes franceses en Castilla. Creemos que es necesaria la reconsideración al respecto, por varias razones: cronológica (más frecuente en épocas tardías), geográfica (en Navarra y Aragón) y lingüística (correlación entre –CC y sC-).

4.2. Distancia

4.2.1. Distancia simple

Para medir la distancia que existe entre dos filas de matriz, utilizamos la «Distancia Euclídica» que se formula de manera siguiente. Se calcula la diferencia entre los elementos de cada fila en cuestión, se calcula el cuadrado, se suma y finalmente se divide por el número de columnas (P), que llamamos «Distancia Simple» (Dis.S.):

$$Dis.S(i, j) = \{[\Sigma_{k:p} (X_{ik} - X_{jk})^2]^{1/2}\} / P$$

Por ejemplo, calculamos la distancia entre d1 y d2 de la tabla inferior izquierda (D_{np}):

$$\begin{aligned} Dis.S(1, 2) &= \{[\Sigma_{k:p} (X_{1k} - X_{2k})^2]^{1/2}\} / P \\ &= [(45 - 56)^2 + (48 - 59)^2 + (66 - 54)^2]^{1/2} / 3 \fallingdotseq 11.343 \end{aligned}$$

La Distancia aumenta según crece la cifra, lo que significa que la relación de los dos términos en comparación se reduce, lo que es el caso inverdo del «Coeficiente de Correlación» (CC) y de los «Coeficientes de Asociación» (CA). También es diferente de CC y de CA en que siempre ofrece valores positivos, y así sin más, no se determina el valor Máximo:

D_{np}	v1	v2	v3
d1	45	48.000	66
d2	56	59	54
d3	58	51	78
d4	77	72	20
d5	43	44	32
d6	58	34	90
d7	50	53	100

Sp.D.	d1	d2	d3	d4	d5	d6	d7
d1		11.343	10.360	35.195	19.799	17.711	20.050
d2	11.343		14.652	24.262	17.108	25.331	27.006
d3	10.360	14.652		37.265	28.225	12.014	13.565
d4	35.195	24.262	37.265		26.357	47.276	49.967
d5	19.799	17.108	28.225	26.357		35.067	39.808
d6	17.711	25.331	12.014	47.276	35.067		13.229
d7	20.050	27.006	13.565	49.967	39.808	13.229	

4.2.2. Distancia estandarizada

Considerando que las variables poseen Varianza distinta una de otras, que influye en el cálculo de la Distancia, estandarizamos la distancia

dividiéndola por la Desviación Típica de la columna. La distancia así obtenida la llamamos «Distancia Estandarizada» (Dis.E)

$$Dis.E(i, j) = \{[\Sigma_{k:p} (X_{ik} - X_{jk})^2 / DT_k]^{1/2}\} / P$$

donde DT_k es la fila de Desviaciones Típicas de Z_{np} y P es el número de sus columnas.

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

St.D	d1	d2	d3	d4	d5	d6	d7
d1	.000	.872	.775	2.370	.756	1.142	.813
d2	.872	.000	.667	1.521	1.159	1.516	1.074
d3	.775	.667	.000	1.949	1.327	.924	.649
d4	2.370	1.521	1.949	.000	2.388	2.688	2.460
d5	.756	1.159	1.327	2.388	.000	1.567	1.561
d6	1.142	1.516	.924	2.688	1.567	.000	1.107
d7	.813	1.074	.649	2.460	1.561	1.107	.000

4.2.3. Distancia Minkowski

Es posible cambiar los exponentes 2 y 1/2 de Distancia Euclídica en 1 ó más de 2, 3, 4, etc. Cuando el exponente es impar, la diferencia entre los dos elementos, $Z_{ik} - Z_{jk,}$ da un valor negativo inconveniente para sumar las distancias, calculamos los valores absolutos de cada diferencia. Esta se llama «Distancia Minkowski» (Dis.M) y se define:

$$Dis.M(E) (i, j) = \{[\Sigma_{k:p} (X_{ik} - X_{jk})^E]^{1/E}\} / P$$

donde E es exponente. La tabla inferior derecha presenta las Dis.M con el exponente 3:

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

M.D(3)	d1	d2	d3	d4	d5	d6	d7
d1	.000	.905	.874	2.436	.871	1.154	.892
d2	.905	.000	.708	1.568	1.180	1.667	1.195
d3	.874	.708	.000	1.954	1.384	1.076	.687
d4	2.436	1.568	1.954	.000	2.560	2.766	2.506
d5	.871	1.180	1.384	2.560	.000	1.640	1.757
d6	1.154	1.667	1.076	2.766	1.640	.000	1.230
d7	.892	1.195	.687	2.506	1.757	1.230	.000

4.2.4. Distancia normalizada en rango

Proponemos una fórmula de distancia, que llamamos «Distancia

Normalizada en Rango» (DNR). Como hemos visto, las tres distancias mencionadas no presentan valores normalizados y resultan difíciles de evaluar dentro de una escala común. Por esta razón, para que la distancia tenga un rango entre 0 y 1, utilizamos la fórmula siguiente:

$$DNR(i, j) = [\ \Sigma_{k:p}\ (|X_{ik} - X_{jk}|^{E} / Rg_k]^{1/E} / P$$

donde Rg_k es el Rango de la variable en cuestión, es decir la diferencia entre el Máximo y el Mínimo: $Max_k - Min_k$. Cuando Xi_k es Máximo y Xj_k es Mínimo de la columna, da la distancia 1. De modo que el Máximo de la DNR se presenta cuando la fila compone de Máximos verticales y el otro elemento de la fila en comparación es Mínimos verticales. Para el cálculo de DNR utilizamos distancia Minkowski. El resultado es la tabla inferior derecha:

D_{np}	v1	v2	v3
d1	45	48.000	66
d2	56	59	54
d3	58	51	78
d4	77	72	20
d5	43	44	32
d6	58	34	90
d7	50	53	100

DNR	d1	d2	d3	d4	d5	d6	d7
d1		.265	.241	.734	.255	.352	.271
d2	.265		.214	.476	.355	.461	.359
d3	.241	.214		.617	.432	.272	.211
d4	.734	.476	.617		.722	.832	.792
d5	.255	.355	.432	.722		.513	.523
d6	.352	.461	.272	.832	.513		.327
d7	.271	.359	.211	.792	.523	.327	

(#) Correlación y distancia: Formas apocopadas anómalas

Vamos a ver la diferencia entre Coeficiente de Correlación y el de Distancia en matrices y gráficos. Las tablas siguientes muestran los casos de formas apocopadas anómalas en el intervalo de 20 años entre 1200 y 1500 y matrices de correlación (CC) y distancia (NDR): a: *adelant*, en: *end*, es: *est*, pa: *part*, pr: *present*, v: *veint*.

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

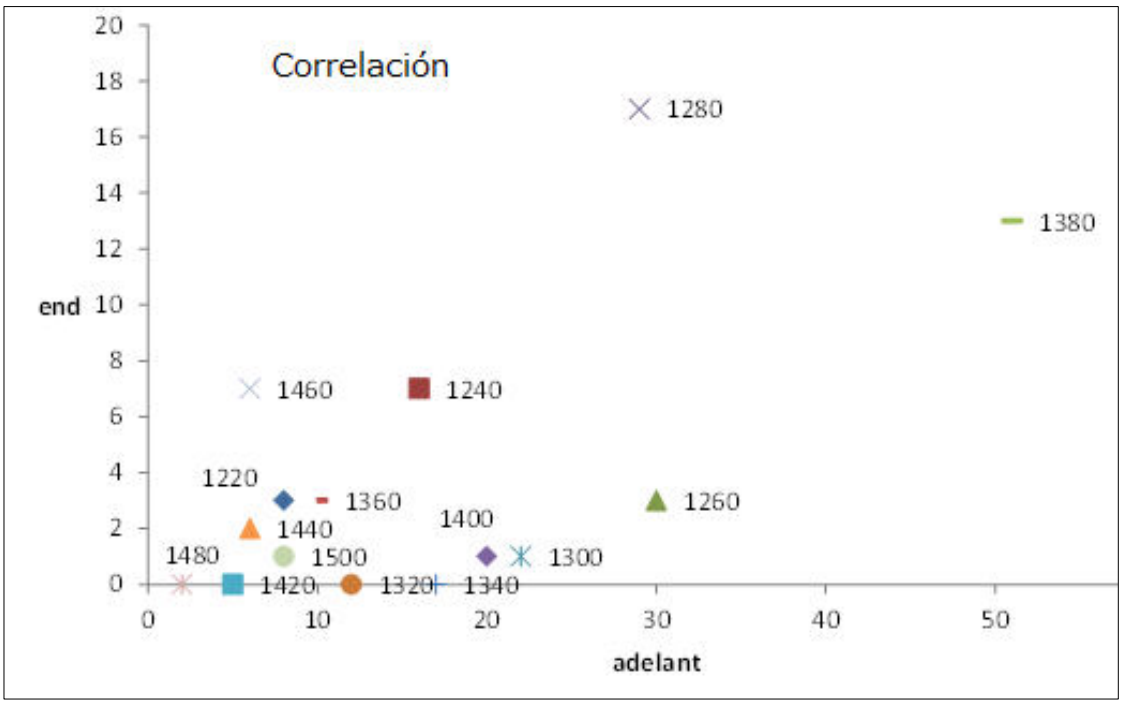
CC	a	en	es	pa	pr	v
adelant	1.000	.645	.172	.405	.508	.450
end	.645	1.000	.318	.079	.062	.512
est	.172	.318	1.000	.237	-.246	.114
part	.405	.079	.237	1.000	.614	.584
present	.508	.062	-.246	.614	1.000	.504
veint	.450	.512	.114	.584	.504	1.000

NDR	a	en	es	pa	pr	v
adelant		.377	.405	.475	.603	.410
end	.377		.256	.660	.802	.209

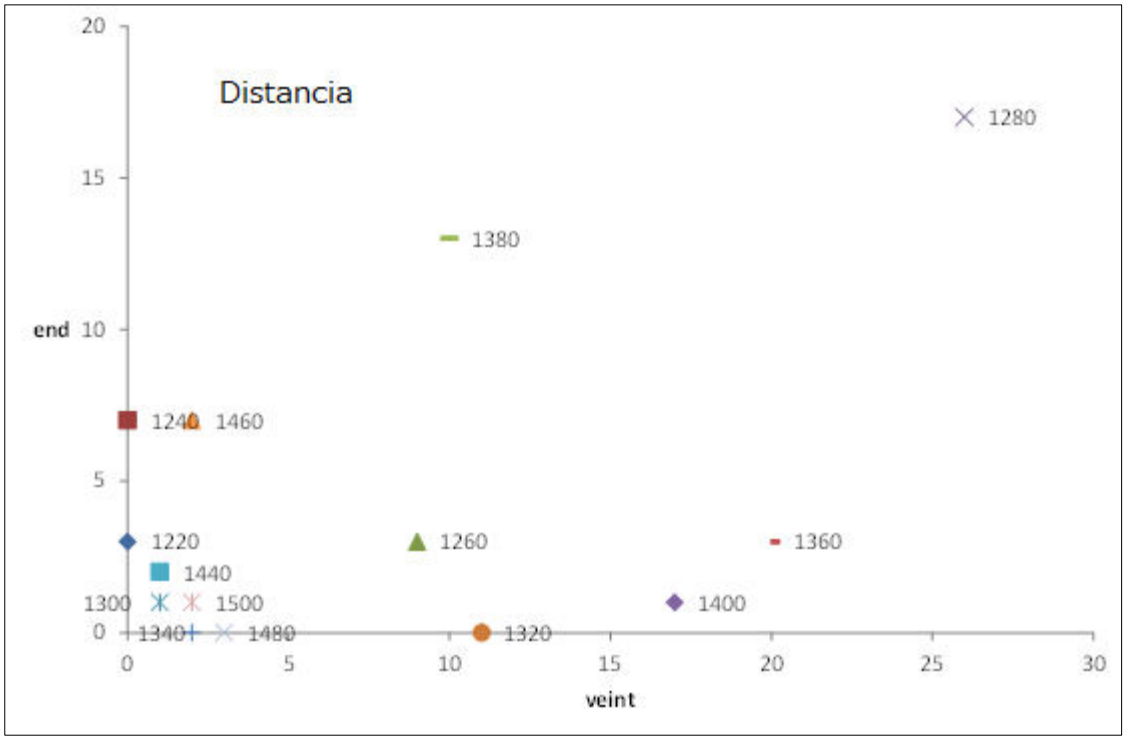
1400	20	1	2	64	121	17
1420	5		1	10	32	
1440	6	2	1	15	24	1
1460	6	7		2	26	2
1480	2		3	17	23	3
1500	8	1		2	16	2

est	.405	.256		.579	.841	.344
part	.475	.660	.579		.600	.621
present	.603	.802	.841	.600		.736
veint	.410	.209	.344	.621	.736	

La pareja en la relación más estrecha en correlación es *end-adelant* (.645) y la pareja más cercana en distancia es *end-veint* (.209), de modo que las parejas no coinciden.



El gráfico superior muestra la correlación causada por los datos 1280 y 1380.



El gráfico de «Distancia» muestra una concentración fuerte en la parte inferior izquierda, donde la distancia entre el valor del eje X y el de Y es mínima. De esta manera, la interpretación de Correlación y la de Distancia son diferentes, puesto que el «Coeficiente de Correlación» (CC) muestra la relación de movimientos de los puntos, mientras que el «Coeficiente de Distancia» (CD) indica la cercanía de las posiciones en el espacio bidimensional. Cuanto más juntos van los dos valores de X e Y, el CC aumenta. La distancia se acorta cuanto más los datos ofrecen valores parecidos de X e Y, preferentemente en los puntos de origen (0).

4.2.5. Distancia Mahalanobis

Si convertimos la matriz objeto de análisis en otra cuya Varianza, lo mismo que en la «Distancia Estandarizada», y ahora «Covarianza» sean 0. La «Covarianza» es numerador de Coeficiente de Correlación:

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

De modo que la nueva matriz presenta su Matriz de Correlación nula. Para obtener la matriz de covarianza nula, hay que realizar la operación de «Análisis de Componentes Principales» (ACP), que trataremos en el Capítulo siguiente:

ACP	E	L	S
d1	- .823	- .544	.325
d2	.635	- .149	.369
d3	- .176	.588	.007
d4	3.171	.218	- .239
d5	- .510	- 1.668	- .270
d6	- 1.383	.789	- 1.025
d7	- .916	.766	.834

Covarianza	v1	v2	v3
v1	2.026	.000	.000
v2	.000	.672	.000
v3	.000	.000	.303

Y ahora estandarizamos la matriz producida de de ACP en Enp:

Enp	v1	v2	v3
d1	- .578	- .663	.591
d2	.446	- .182	.671
d3	- .124	.718	.013
d4	2.228	.266	- .435
d5	- .358	- 2.035	- .491
d6	- .972	.963	- 1.864
d7	- .643	.934	1.515

CC	v1	v2	v3
v1	1.000	.000	.000
v2	.000	1.000	.000
v3	.000	.000	1.000

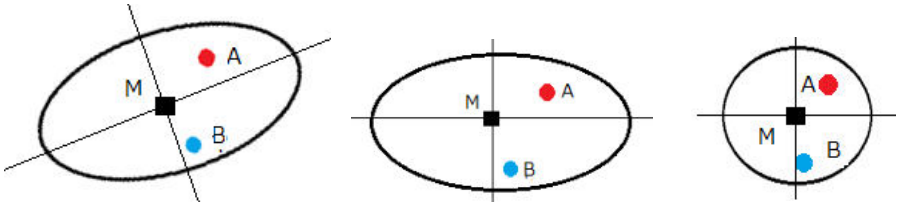
Se llama la «Distancia Mahalanobis» (Dis.M) la calculada a partir de la Matriz Enp:

Dis.M	d1	d2	d3	d4	d5	d6	d7
d1	0.000	0.655	0.903	1.807	1.017	1.715	1.066
d2	0.655	0.000	0.723	1.238	1.346	1.803	1.024
d3	0.903	0.723	0.000	1.406	1.621	1.197	0.927
d4	1.807	1.238	1.406	0.000	1.999	2.063	2.041
d5	1.017	1.346	1.621	1.999	0.000	1.936	2.076
d6	1.715	1.803	1.197	2.063	1.936	0.000	1.960
d7	1.066	1.024	0.927	2.041	2.076	1.960	0.000

Como verems más adelante, en el Análisis de Componentes Principales utilizamos la «Matriz Propia», de modo que el cálculo de Dis.M. no es realizable con los datos cuyo número de filas es menor que el de columnas. De esto hablaremos en el Capítulo siguiente.

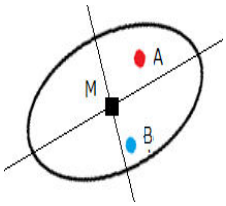
(*) Coordenadas en la Distancia Mahalanobis

Para entender la esencia de la «Distancia Mahalanobis», veamos los tres gráficos siguientes:



El gráfico superior izquierdo muestra la posición del dato A y B con respecto a la de Media tanto de la coordenada X como de la Y. Si el conjunto del dato se distribuye en forma de elipse, observamos la existencia de Covarianza, entre X e Y, es decir, existe cierta correlación, puesto que observamos una inclinación general de los datos. Ahora con el método de «Análisis de Componentes Principales» (ACP) giramos los dos ejes, X e Y, para obtener la distribución de los datos en forma del gráfico superior medio. Esta vez no existe ninguna correlación, y lógicamente ninguna Covarianza. Y finalmente por la estandarización acortamos el largo del eje X porque tiene una Varianza mayor que el eje Y. Entonces obtenemos la distribución de la forma del gráfico superior derecho. La «Distancia mahalanobis» se calcula entre A y B con los datos así convertidos.

Ahora bien, al comparar los dos primeros gráficos, podemos notar la igualdad de distancia tanto sin el giro como con el giro. La distancia cambia en el último gráfico cuando se acorta un eje considerando la Varianza. Entonces, ¿ por qué razón se realiza un giro por ACP? ¿Qué ocurre si acortamos el eje X directamente de manera siguiente?



En realidad es la operación que hemos realizado en el cálculo de la «Distancia Estandarizada». Efectivamente, la Varianza del eje X se ha modificado. Sin embargo, sigue existiendo la Covarianza que influye en el cálculo de la Distancia.

*Para la Distancia Mahalanobis, veáse Okumura (1986:152-155)

4.3. Asociación

4.3.1. Coeficiente de asociación

En el análisis de datos lingüísticos hay ocasiones en que tratamos los

datos cualitativos designados de +/— o «v», en lugar de los datos cuantitativos de valores continuos. Por ejemplo en la tabla siguiente[22], las filas donde tanto la Carta, como el Drama dan el signo «+» son 4: *abajo*, *abandonar*, *abeja*, *abogado*. Este número se llama «Coocurrencias». Como las cifra de «Coocurrencia» varía según el tamaño de los datos, se han propuesto varios «Coeficientes de Asociación», de los cuales presentamos los más usuales en las investigaciones lingüísticas.

Palabra	Carta	Drama		Carta	Drama		a: +/+	b: +/-	c: -/+	d: -/-
abajo	5	10		+	+		1	0	0	0
abandonar	9	6		+	+		1	0	0	0
abandono	0	0		-	-		0	0	0	1
abarcar	1	0		+	-		0	1	0	0
abastecimiento	2	0		+	-		0	1	0	0
abatir	0	1		-	+		0	0	1	0
abeja	2	3		+	+		1	0	0	0
abertura	0	0		-	-		0	0	0	1
abismo	0	0		-	-		0	0	0	1
abnegación	0	0		-	-		0	0	0	1
abogado	3	6		+	+		1	0	0	0
abonar	5	0		+	-		0	1	0	0
abono	0	0		-	-		0	0	0	1
abordar	0	0		-	-		0	0	0	1
aborrecer	0	6		-	+		0	0	1	0

Preparamos la tabla siguiente consruida de dos filas y dos columnas con cuatro celdas: a, b, c, d. La a corresponde al número de casos con reacciones positivas tanto de X como de Y; la b es de la positiva en X y la negativa en Y; la c es de la negativa en X y la positiva en X; la d es de las reacciones negativas en X e Y. En los datos de la tabla anterior: a = 4 {*abajo*, *abandonar*, *abeja*, *abogado*}, b = 3 {*abarcar*, *abastecimiento*, *abonar*}, c = 2 {*abatir*, *aborrecer*}, d = 6 {*abandono*, *abertura*, *abismo*, *abnegación*, *abono*, *abordar*}.

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

Los «Coeficientes de Asociación» utilizan estos valores a, b, c, d. Algunos no utilizan el valor d.

[22] Los datos son de: A. Juilland y E. Chang Rodríguez en su *Frequency dictionary of Spanish words,* (The Hague: Mouton, 1964).

(1) «Coeficiente de Coocurrencia Simple» (CS)

$$CS = (a + d) / (a + b + c + d) \quad 0.0\ (a=d=0) \leqq CS \leqq 1.0\ (b=c=0)$$

(2) «Coeficiente de Jaccard» (J)

$$J = a / (a + b + c) \qquad 0.0\ (a=0) \leqq J \leqq 1.0\ (b=c=0)$$

(3) «Coeficiente de Russel-Rao» (RR)

$$RR = a / (a + b + c + d) \quad 0.0\ (a=0) \leqq RR \leqq 1.0 \quad (b=c=d=0)$$

(4) «Coeficiente de Dice 係数» (D)

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

(5) «Coeficiente de Yule» (Y)

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

(6) «Coeficiente de Hamann» (H)

$$H = [(a+d) - (b+c)] / [(a+d) + (b+c)] \quad -1 \leqq H \leqq 1$$

(7) «Coeficiente de Phi» (Ph)

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

(8) «Coeficiente de Ochiai» (O)

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

(9) Finalmente proponemos nuestra fórmula propia que denominamos «Coeficiente de Preferencia» (Pr):

$$Pr = [2a - (b + c)] / [2a + (b + c)] \quad -1 \leqq Pr \leqq 1$$

Podemos formular algunos coeficientes más pensando en los dos tipos de valores: relativo X/(X+Y) y contrastivo (X-Y) / (X+Y), consideración del valor d (-/-) y existencia de multiplicación. Algunos ya se han tratado anteriormente.

Posibles coeficientes de asociación	X	Y	R:C	d	mult.
1. [a - (b + c)]/[a + (b + c)]	a	b + c	C	-	-
2. 2a / [2a + (b + c)]	2a	b + c	R	-	-
3. [2a - (b + c)] / [2a + (b + c)]	2a	b + c	C	-	-
4. a^2 / (a^2 + bc)	a^2	bc	R	-	+
5. (a^2 - bc)/ (a^2 + bc)	a2	bc	C	-	+

6. $a / [a + (bc)^{1/2}]$ (Ochiai)	a	$(bc)^{1/2}$	R	-	+
7. $[a - (bc)^{1/2}] / [a + (bc)^{1/2}]$	a	$(bc)^{1/2}$	C	-	+
8. $(a + d) / [(a + d) + (b + c)]$	a + d	b + c	R	+	-
9. $[(a + d) - (b + c)]/[(a + d) + (b + c)]$ (Hamann)	a + d	b + c	C	+	-
10. $ad / (ad + bc)$	ad	bc	R	+	+
11. $(ad - bc) / (ad + bc)$ (Yule)	ad	bc	C	+	+
12. $(ad)^{1/2} / [(ad)^{1/2} + (bc)^{1/2}]$	$(ad)^{1/2}$	$(bc)^{1/2}$	R	+	+
13. $[(ad)^{1/2} - (bc)^{1/2}] / [(ad)^{1/2} + (bc)^{1/2}]$	$(ad)^{1/2}$	$(bc)^{1/2}$	C	+	+

Las fórmulas 4 y 10 pueden ser bajadas de elevación:

4'. $a / (a^2 + bc)^{1/2}$
10'. $[ad / (ad + bc)]^{1/2}$

(*) Rasgos no existentes en los datos comparados

Hubo polémica entre Kroeber (1937, 1969) que utilizó Phi y Ellegard (1959) que utilizó Ochiai en cuestiones de lingüística indoeuropea. No se trata de ver cuál es el definitivamente correcto, sino de seleccionar de acuerdo con las características de los datos y objetivo de investigación. Por ejemplo, si en las encuestas se han preguntado Sí o No sobre alguna opinión hay que considerar no solamente las veces que coindicen en las respuestas positivas sino también las negativas.

También cuando tratamos múltiples datos al mismo tiempo, la frecuencia de los rasgos no existentes entre los dos datos en comparación también conlleva información importante.

En cambio si el valor de *d* resulta mayor que *a* con distancia considerable, es recomendable pensar en los coeficientes sin *d*.

(*) Comparación de coeficientes de asociación

En la práctica del análisis hay veces en que vacilamos sobre qué coeficiente utilizar. Se puede pensar en varias maneras de selección y los criterios de selección también pueden variar. Ante el público no muy versado de análisis estadísticos, los coeficientes de Coocurrencia Simple, Russel-Rao, Jacard no necesitan mucha explicación. El uso de Yule o de Hamann exige una explicación suficientemente razonada. Ante la audiencia de analistas expertos es recomendable el uso de Phi o Ochiai. También es posible utilizar varios coeficientes para comparar sus resultados.

Sin embargo, estas decisiones no son teóricas y obedecen a consideraciones prácticas. Para perseguir la eficacia de los coeficientes hay que

tomar decisiones considerando las condiciones tanto teóricas como prácticas.

Comparando las propiedades de los coeficientes notamos que existen rasgos comunes entre ellos. Sobre todo observamos que hay tres rasgos que son importantes: el tratamiento del valor d(-/-), el rango [-1 ~ 1] o [0 ~ 1] y la multiplicación. La tabla siguiente muestra la distribución de los tres rasgos mencionados:

Propiedad	SM	RR	J	D	Y	H	Ph	O	Pr
Componente d (-/-) incluido	v	-	-	-	v	v	v	-	-
Valor negativo	-	-	-	-	v	v	v	-	v
Multiplicación	-	-	-	-	v	-	v	v	-

Si tomamos la decisión de no incluir el valor d (-/-), ver la dirección negativa con el rango [-1 ~ 1], evitar la multiplicación, seleccionamos la última opción Pr.

Si la propiedad de datos posee la dirección, por ejemplo, la encuesta en que se pregunta apoyo u oposición, seleccionamos el coeficiente con valor d (-/-) puesto que la coincidencia de las respuestas negativas es importante. En la investigación léxica de los textos es recomendable la frecuencia de la coincidencia positiva, puesto que la frecuencia de la coincidencia negativa es en teoría ilimitada. No obstante, si el estudio trata un conjunto limitado del léxico, por ejemplo, pronombres, relativos, etc., la frecuencia de coincidencia negativa también conlleva la información de relación entre los dos textos.

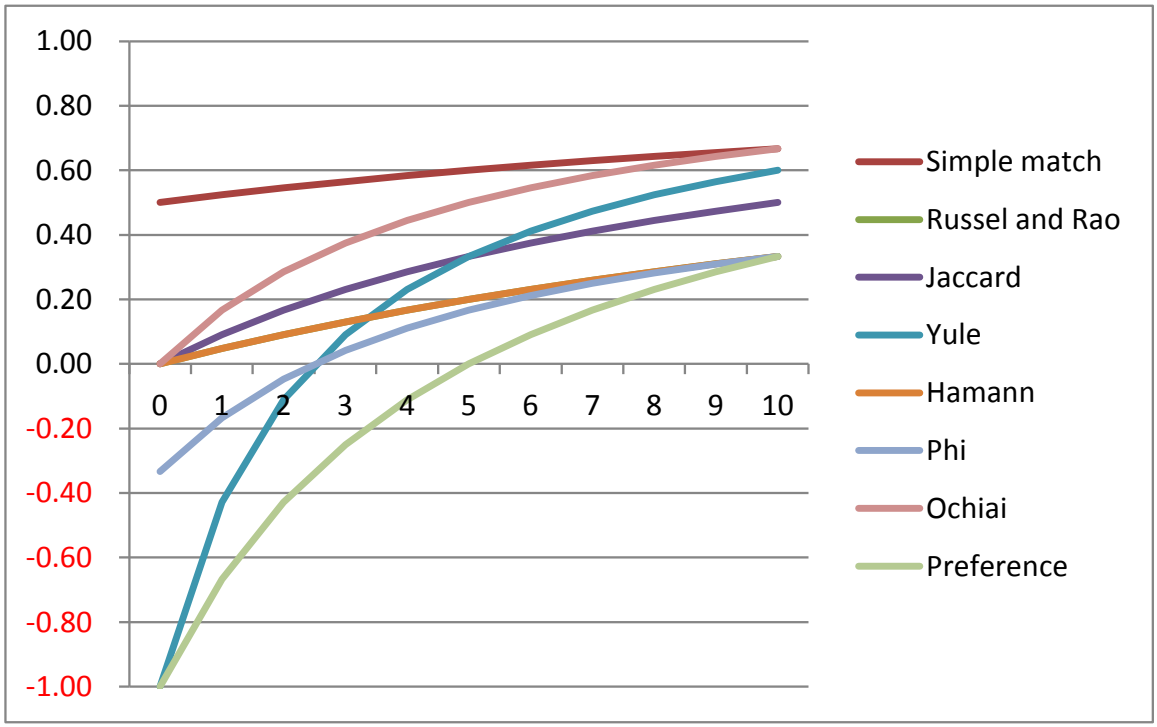
Los coeficientes cuyo rango es [-1.0 ~ 1.0] presenta cero (0) en los datos de distribución homogénea. Hay que tener en cuenta que los otros coeficientes tienen los valores 0.5 (SM, O), 0.25 (RR), 0.33 (J) en la misma distribución. El coeficiente de correlación 0.5 indica un valor de correlación algo fuerte en coeficientes sin valor negativo, mientras que en Y, H y Ph, el mismo valor 0.5 indica una asociación nula.

Los coeficientes de correlación que conllevan el proceso de multiplicación (Y, Ph, O) tienen problema al presentar el valor 0 cuando uno de los términos es 0, a pesar de que el otro es un valor considerable. Además, cuando esto ocurre en la parte del denominador, la operación resulta impracticable.

La tabla y el gráfico siguientes muestran el movimiento de cada coeficiente de acuerdo con la subida de a(+/+), con otros parámetros fijos: b(+/-) = 5, c(-/+) = 5, y d(-/-) = 10:

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

d (-/-)	10	10	10	10	10	10	10	10	10	10	10
SM	0.50	0.52	0.55	0.57	0.58	0.60	0.62	0.63	0.64	0.66	0.67
RR	0.00	0.05	0.09	0.13	0.17	0.20	0.23	0.26	0.29	0.31	0.33
J	0.00	0.09	0.17	0.23	0.29	0.33	0.38	0.41	0.44	0.47	0.50
Y	-1.00	-0.43	-0.11	0.09	0.23	0.33	0.41	0.47	0.52	0.57	0.60
H	0.00	0.05	0.09	0.13	0.17	0.20	0.23	0.26	0.29	0.31	0.33
Ph	-0.33	-0.17	-0.05	0.04	0.11	0.17	0.21	0.25	0.28	0.31	0.33
O	0.00	0.17	0.29	0.38	0.44	0.50	0.55	0.58	0.62	0.64	0.67
Pr	-1.00	-0.67	-0.43	-0.25	-0.11	0.00	0.09	0.17	0.23	0.29	0.33



Se observa que los coeficientes de SM, RR, J y D, que no presenta el valor negativo, su amplitud es lógicamente pequeña. Phi y O son parecidos con amplitud reducida. Y y Pr presentan amplitud grande y la subida de Y es más pronunciada que la de Pr, lo que indica que no posee capacidad distinguidora en los valores altos. El coeficiente Pr, como no incluye el valor de d (-/-), no sufre la consecuencia del aumento de d. La subida rápida de Y se debe al efecto del valor d.

(*) Coeficiente de correlación y coeficiente Phi

El «Coeficiente Phi» es derivable del «Coeficiente de Correlación» (CC) si consideramos que «+» es 1 y «-» es 0:

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

Simbolizamos la Suma total de datos con N:

[1] $N = a + b + c + d$

Como hemos visto anteriormente, la fórmula del CC es:

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

donde Mx es Media de X, My es Media de Y, DTx es Desviación Típica de X, DTy es Desviación Típica de Y.

Numerador de $CC = \Sigma i\ (Xi - Mx)(Yi - My)$

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

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

← distribuir Σ

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

← los elementos que no llevan i, al exterior

Dentro de Xi Yi, los que corresponden a b(1, 0), c(0, 1), d(0, 0), resultan 0. Por lo tanto,

[2] $\Sigma i\ Xi\ Yi = a$

←Suma de la multiplicación cuyo resultado es 1

Por otra parte,

[3] $\Sigma i\ Xi = a + b$ ←Suma de X ←ver la tabla(X:Y)

[4] $\Sigma i\ Yi = a + c$ ←Suma de Y ←ver la tabla(X:Y)

[5] $Mx = \Sigma i\ Xi / N = (a + b) / N$ ←Media de X←[3]

[6] $My = \Sigma i\ Yi / N = (a + c) / N$ ←Media de Y←[4]

De modo que el numerador es:

Numerador de $CC = \Sigma i\ Xi\ Yi - My\ \Sigma i\ Xi - Mx\ \Sigma i\ Yi + N\ Mx\ My$

$= a - (a+c)(a+b)/N - (a+b)(a+c)/N + N\ (a+b)/N\ (a+c)/N$ [2-6]

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

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

$= [Na - (a + b)(a + c)] / N$

$= [(a + b + c + d)a - (aa + ac + ba + bc)] / N$ ←[1]

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

[7] $= (ad - bc) / N$

Ahora veamos uno de los términos del denominador de CC: DTx

$DTx = \{[\Sigma i\ (Xi - Mx)^2]^{1/2} / N\}^{1/2}$ ← Desviación Típica de X

$= \{[\Sigma i\ (Xi^2 - 2\ Xi\ Mx + Mx^2)]^{1/2} / N\}^{1/2}$ ←desarrollar

$= \{[\Sigma i\ Xi^2 - \Sigma i\ 2\ Xi\ Mx + \Sigma i\ Mx^2)] / N\}^{1/2}$ ←distribuir Σ

$= \{[\Sigma i\ Xi^2 - 2\ Mx\ \Sigma i\ Xi + N\ Mx^2)] / N\}^{1/2}$

← los elementos que no llevan i, al exterior

Como todos de Xi es 1 ó 0, la Suma de Xi^2 es:

[8] $\Sigma i\ Xi^2 = a + b$ ←Suma de X^2←ver la tabla(X:Y)

$DTx = \{[(a + b) - 2\ (a + b)^2 / N + (a + b)^2 / N] / N\}^{1/2}$ ←[8], [3], [5]

$= \{[a + b - (a + b)^2 / N] / N\}^{1/2}$ ←$(a + b)^2 / N$ es común

$= \{[(a + b)N - (a + b)^2] / N^2\}^{1/2}$ ←N, al denominador

$= \{[(a + b)(a + b + c + d) - (a + b)^2] / N^2\}^{1/2}$ ←[1]

$= \{(a + b)[(a + b + c + d) - (a + b)] / N^2\}^{1/2}$ ←(a + b) es común

$= [(a + b)(c + d) / N^2]^{1/2}$ ←(a + b) es común

[9] $= [(a + b)(c + d)]^{1/2} / N$ ←N, al exterior

Del mismo modo, el otro término del denominador DTy es:

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

Por lo tanto,

Denominador de CC = N [DTx DTy]

$= N\ \{[(a + b)(c + d)]^{1/2} / N\} * \{[(a + b)(c + d)]^{1/2} / N\}$ ←[9, 10]

$= [(a + b)(c + d)]^{1/2}\} * \{[(a + b)(c + d)]^{1/2} / N$ ←despejar N

[11] $= [(a + b)(c + d)(a + b)(c + d)]^{1/2} / N$

←despejar el exponente 1/2

Por lo tanto el «Coeficiente de Correlación» (CC) es:

$CC = \Sigma i\ (Xi - Mx)(Yi - My) / N$ [DTx DTy]

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

←[7, 11]

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

= Phi ←def.

(*) Coeficiente phi y coeficiente de Ochiai

Al aplicar el «Coeficiente de Phi» (Ph) a veces nos encontramos unos

casos difíciles de tratar. Veamos por ejemplos A y B:

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

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

Ambos datos presentan el Ph con valor cero (0), puesto que el numerador ad－bc de Ph, 0: 100*2 － 10*20 = 0; 4*50 － 10*20 = 0. Sin embargo, el dato A parece presentar un valor de asociación mucho más alto que B, porque hay 100 casos de a(+/+), que ocupa 75.8% de todo, mientras que en B, encontrams tan solo 4 casos de a(+/+) que corresponde a 4.8%

Este problema se debe al tratamiento del valor d(-/-). Se trata del número de elementos que no existe ni en X ni en Y, que es limitado dentro del conjunto de los datos tratados, pero en teoría, fuera del conjunto en cuestión, es ilimitado. Por ejemplo, para ver la similitud de dos personas por los libros que leen. Los tres valores, a, b, c son limitados, pero el de d es ilimitado, puesto que existe sin número de libros que no leen las dos personas.

Vamos a suponer que el valor d de Phi es ilimitado en Phi':

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

$$\begin{aligned}\text{Phi}' &= \lim(d\to\infty)\ (ad - bc) / [(a + b)(c + d)(a + c)(b + d)]^{1/2} \\ &= \lim(d\to\infty)\ [(ad - bc)/d] / \{[(a + b)(c + d)(a + c)(b + d)]^{1/2} /d\} \\ &\qquad \leftarrow \text{ dividir numerador y denominador por d} \\ &= \lim(d\to\infty)\ (a - bc/d) / [(a + b)(c + d)(a + c)(b + d) / d^2]^{1/2} \\ &\qquad \leftarrow \text{ mover d} \\ &= \lim(d\to\infty)\ (a - bc/d) / [(a + b)(c/d + 1)(a + c)(b/d + 1)]^{1/2} \\ &\qquad \leftarrow \text{ distribuir /d} \\ &= a / [(a + b)(a + c)]^{1/2} \qquad \leftarrow \text{ d es ilimitado}\end{aligned}$$

Esta es una versión modificada de Phi, que corresponde al «Coeficiente de Ochiai» cuya forma es sumamente simple. Vamos a aplicarlo a los dos casos citados A y B:

$$\text{Phi}'(A) = 100 / [(199+10)(100+20)]^{1/2} = .870$$
$$\text{Phi}'(B) = 4 / [(4+10)(4+20)]^{1/2} = .218$$

De esta manera nos convence el resultado en que el coeficiente O da una cifra mucho mayor en A que en B.

(*) Información mutua y coeficiente de Dice

En los estudios lingüísticos se utilizan un valor denominado

«Información Mútua» (IM) para ver el grado de combinación de los dos vocablos. Se trata de un logaritmo con base 2 del valor dividido de la frecuencia de la coocurrencia A por un valor teóricamente esperado dentro del grupo:

IM = Log(2) (A*T / X*Y)

donde A es la frecuencia de coocurrencia, T es Suma de todos los vocablos tratados, X es la frecuencia del vocablo *x*, Y es la frecuencia del vocablo *y*. Por ejemplo, supongamos que en un documento español se encuentra 120 veces la palabra «muy», y 167 veces la palabra «bien» y el número total de los vocablos suma a 26578. Entonces, el valor esperado de la coocurrencia de «muy» y «bien» es: (120/26578)*(167/26578) que es la multiplicación de las dos probabilidades de X e Y. En realidad «muy + bien» ha ocurrido 47 veces, que corresponde a la probabilidad de 47/26578. De modo que calculamos la ratio de las dos probabilidades:

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

La «Información Mutua» es el logaritmo con base 2 de esta ratio:

IM = Log(62.334, 2) = 5.962

La base 2 sirve para calcular la información en general.

Utilizando los mismos valores A, X, Y, derivamos al «Coeficiente de Dice» (D), que no incluye el valor T:

$D = A / [(X + Y) / 2]$ $0.0 \leqq D \leqq 1.0$

El valor A corresponde a a(+/+/, X a (a + b), Y a (a + c). Por lo tanto, el «Coeficiente de Dice» (D) es:

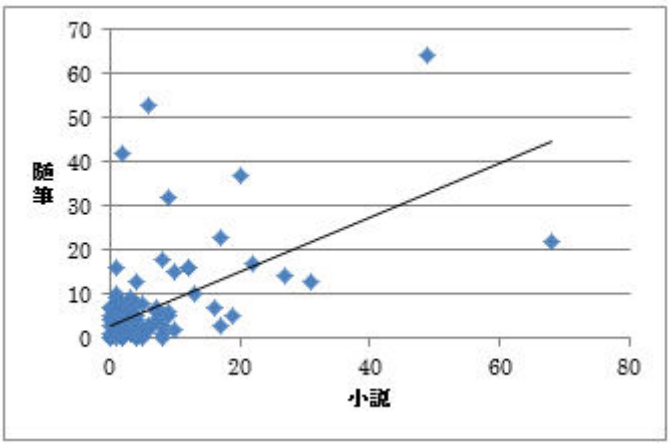
$$D = \frac{a}{(2a + b + c) / 2} = \frac{2a}{(2a + b + c)}$$

En el coeficiente D, comparado con el «Coeficiente Jaccard» (J), el valor de *a* está duplicado. Como se compara el valor a, con dos valores b y c, el coeficiente de D es más equilibrado que el de J.

(*) Dato cuantitativo y dato cualitativo

Como hemos visto anteriormente, la frecuencia de los vocablos presenta una distribución sumamente sesgada y, por esta razón, el «Coeficiente de Correlación» (CC) no es aplicable. Es cierto que podemos trazar la línea de aproximación de la manera siguiente, pero observamos que la gran mayoría se

concentra en la parte inferior izquierda y muy pocos puntos en en el resto:



Una de las soluciones es convertir los «datos cuantitativos» en «datos cualitativos» en forma de «+» y «-» o «V» y «F». De esta manera podemos analizar los datos más o menos equilibrados.

En general, en los análisis lingüísticos, los datos que aparecen solo una vez son tratados de manera especial. Su aparición puede ser accidental. Si ocurre dos veces, la probabilidad de tal fenómeno será sumamente reducida. Cuando tratamos los datos de cantidades enormes, podemos subir el criterio de selección. El resultado varía lógicamente según este criterio, de modo que tenemos que ser conscientes de la selección.

(#) Aprendizaje de la lengua extranjera y prioridad valorativa

Establecemos la hipótesis de que en el aprendizaje del vocabulario y, más ampliamente, en el de la lengua extranjera en general, la prioridad de valor que el aprendiente posee es la causa principal de la adquisición. El «valor» de que hablamos es un concepto diferente del de «importancia», por ejemplo, en forma de vocablos importantes, usuales, fundamentales, etc. cuya selección se establece objetivamente con criterios estadísticos. Estamos hablando del «valor» subjetivo personal que uno tiene sobre vocablos objetos de la adquisición.

Para comprobar la validez de esta hipótesis hemos realizados unos pequeños experimentos con una lista de vocablos. Previamente marcamos la mitad de vocablos que cada sujeto considera que son valiosos para sí mismo. Luego practicamos algunos ejercicios de aprendizaje de memoria y finalmente comprobamos si han aprendido estos vocablos de la lista. Han participando 12 estudiantes matriculados en el curso de Métodos de aprendizaje-enseñanza de español. Hemos realizado el mismo experimento varias veces con distintas listas de vocablos y número variado de participantes. El resultado es la tabla siguiente:

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

(a) +/+: Vocablo valioso (+)」 / Éxito de aprendizaje (+)
(b) +/-: Vocablo valioso (+)」 / Fracaso de aprendizaje (-)
(c) -/+: Vocablo no valioso (-)」 / Éxito de aprendizaje (+)
(d) -/-: Vocablo no valioso (-)」 / Fracaso de aprendizaje (-)

En todos los sujetos participantes del experimentos presentan valores positivos tanto en el «Coeficiente de Yule» (Y) como en el de Hamann (H), lo que demuestra la validez de la hipótesis.

¿En realidad aprendemos la lengua extranjera con largos ejercicios de repetición? Consideramos que adquirimos casi instantáneamente los términos que valoramos positivamente. Si aprendemos unos vocablos sin realizar algunos ejercicios tediosos de memorización, puede ser que son palabras importantes para nosotros. Si no nos equivocamos en la hipótesis, el éxito de aprendizaje de la lengua extranjera (y de otras asignaturas) no dependerá de las horas de estudio, sino de la valoración personal de la materia. Uno de los trabajos importantes del profesor de la lengua extranjera es explicar las razones y despertar intereses personales que pueden tener los estudiantes de la materia, sin obligar a aprenderla ciegamente, con horas y horas de ejercicio. Una vez convencidos de los valores que hay en la materia, el ejercicio mismo se convierte en una práctica significativa.

Para encontrar los valores de la materia, pensamos que la dirección conveniente de enseñanza-aprendizaje es del significado a la forma, más que de la forma al significado. Difícilmente apreciamos el valor observando la forma[23].

[23] Puede haber algunas excepciones. Cuando daba un curso de español dirigido a los trabajadores, una participante me dijo que aprendió la palabra «pájaro» por la impresión agradable que produce la secuencia de los sonidos, pá-ja-ro. Por desgracia, no se me ocurrió preguntarle si le gustan los pájaros.

En cambio podemos sentir inmediatamente la existencia y el grado del valor del significado. En la práctica de la lengua extranjera, y de los procesamiento de análisis matemáticos, la mayoría de la dificultad nace de la forma, sin conocer su significado. La solución será partir del significado para llegar a la forma.

4.3.2. Matriz de asociación

Para elaborar la «Matriz de Asociación» de distintos coeficientes, preparamos de antemano matrices cuadradas de A(i, j) que ofrece la frecuencia Xi = 1, Xj =1, B(i, j) que ofrece la frecuencia Xi = 1, Xj =0, C(i, j) que ofrece la frecuencia Xi = 0, Xj =1, D(i, j) que ofrece la frecuencia Xi = 0, Xj =0. Para preparar las cuatro matrices, A, B, C, D, utilizamos la matriz W cuyos elementos son reversos de Q:

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

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

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

Utilizando estas dos matrices, Q y W, producimos las matrices A, B, C, D:

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

Veamos sus operaciones matriciales y sus resultados:

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

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

X

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

=

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

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

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

X

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

=

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

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

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

X

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

=

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

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

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

X

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

=

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

Con estas 4 matrices, elaboramos las matrices simétricas de distintos «Coeficientes de Asociación»:

Coocurrencia Simple = (A + D) / (A + B + C + D)
Jaccard = A / (A + B + C)
Dice-Sorenson (Jaccard-2) = 2A / (2A + B + C)
Russel-Rao= A / (A + B + C + D)
Russel-Rao-3 = 3A / (3A + B + C + D)
$Hama_{nn}$= [(A + D) - (B + C)] / [(A + D) + (B + C)]
Yule = (A*D – B*C) / (A*D + B*C)
Phi = $(A*D – B*C) / [(A + B)(C + D)(A + C)(B + D)]^{1/2}$
Ochiai = $A / [(A + B)(A + C)]^{1/2}$
Preference = (2A – B – C) / (2A + B + C)

*Para coeficientes de asociación véanse Anderberg (1973:93-126) y Romesburg

(1989: 177-209). Para el método de preparar las matrices A, B, C, D, véase Kawaguchi (1978: II, 30-31).

(*) Grado de posesión

Explicamos la opción que denominamos «Grado de Posesión» (GP) con ejemplos siguientes. A partir de ls datos cualitativos Qnp, obtenemos la matriz de coocurrencia App:

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

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

En App, entre v1 y v2 encontramos la cifra 2, es decir hay 2 veces donde tanto v1 como v2 da 1. Efectivamente en Qnp en d1 y d5 encontramos las correspondencias positivas. Pensamos que el caso de d1 es más importante que d5, puesto que en d1 no hay otros casos donde se da 1, mientras que en d5 hay un caso más, en v3.

Las matrices siguientes han sido utilizadas en la preparación de las matrices, A, B, C, D. W se deriva de Q: $W_{np} = 1 - Q_{np}$.

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

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

Las convertimos de manera siguiente:

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

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

Por ejemplo en la fila d1, aparece 1 en dos ocasiones, de modo que cada uno posee un valor .5. En d5, 1 se divide por el número de la cifra «1»: 1 / 3

= .333. Hacemos lo mismo en otras filas de Q y también en W. Utilizando estas nuevas matrices, preparamos de nuevo las matrices A, B, C, D para elaborar las matrices de coeficientes de asociación. Por ejemplo, comparemos la matriz original de Coocurrencia Simple (CS) y la matriz modificada con el «Grado de Posesión» (CSp). Observamos que entre ambas se mantiene una relación proporcional, pero los valores mismos disminuye de distintas maneras en CSp.

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

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

4.4. Asociación ordinal

Para ver la relación de orden entre las variables o los individuos, utilizamos el «Coeficiente de Asociación Ordinal de Goodman y Kruskal» (GK).

Xnp	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

Gpp	v1	v2	v3	v4	v5
v1	1.000	-.393	.028	.607	-.168
v2	-.393	1.000	.371	.703	.113
v3	.028	.371	1.000	.519	-.175
v4	.607	.703	.519	1.000	-.472
v5	-.168	.113	-.175	-.472	1.000

Por ejemplo para ver la relación entre v1 y v2, primero calculamos el valor P que consiste en la suma de productos entre un elemento de v1 y otro elemento en orden superior de v2. El valor N se obtiene por la suma de productos entre un elemento de v1 y otro elemento en orden inferior de v2:

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

El coeficiente GK es la Ratio de P / N:

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

*Véase Ikeda (1976:130-132).

4.5. Asociación nominal

Trataremos ahora las matrices cuyos elementos no son números sino

letras. Los elementos pueden ser letras, nombres, secuencia de letras, etc. La matriz de asociación de tal dato, la denominamos «Matriz de Asociación Nominal» (MAN). Supongamos, por ejemplo, v1-v4 son regiones, d1-d5 son documentos emitidos en cada región. A, B, C, ... son rasgos lingüísticos observados en cada documento:

Lnp	v1	v2	v3	v4
d1	A	A	B	C
d2	A	A	C	C
d3	A	C	B	C
d4	C	C	C	A
d5	B	B	C	C

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

Por ejemplo, entre v1 y v2 hay asociación de valor .600. Contamos las frecuencias de a(+/+), b(+/-), c(-/+), d(-/-) entre los elementos de v1 y v2, y finalmente obtenemos el valor .600 por el «Coeficiente de Asociación. Preferencia»: [2a – (b+c)] / [2a – (b+c)] = [4x2 – (1+1)] / [4x2 + (1+1)] = .600、

También es posible calcular el mismo valor de MAN en una matriz cuyos elementos son complejos de manera siguiente:

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

Lnp.	v1	v2	v3	v4
v1	1.000	.500	.067	- .200
v2	.500	1.000	.333	- .467
v3	.067	.333	1.000	- .143
v4	- .200	- .467	- .143	1.000

4.6. Proximidad

4.6.1. Proximidad simple

Definimos «Proximidad Simple» (PS) como el promedio de las Proximidades de los miembros correspondientes de los dos columnas de una matriz. La fórmula de la «Proximidad» (Prox) es la siguiente[24]:

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

donde x e y son los dos valores en comparación, |x – y| es el valor absoluto de la diferencia entre x e y, Max(x, y) es el valor máximo entre x e y. Por ejemplo, la proximidad de (2, 5) es 1 - |2 – 5| / Max(2, 5) = 1 – 3/5 = .4. El rango de la

[24] La proximidad (Prox) es valor complementario de «Separatividad» (Sep) respecto a 1.

Prox = 1 - Sep.

Proximidad es [0, 1][25].

La «Proximidad Simple» (SP) es promedio de las Proximidades de los dos vectores en cuestión (n es el número de miembros de cada vector):

$$SP = 1/n \quad \Sigma_i \, Prox(x_i, y_i)$$

La Proximidad Simple (SP) calcula el grado de proximidad convertido en un valor relativo, de modo que por ejemplo Prox(2, 5) = .4 se iguala a Prox(20, 50) = 1 - 30/50 = .4. Por esta característica, podemos evitar el problema del efecto causado por unos valores extraordinarios que hemos visto en el coeficiente de correlación y el de distancia. Comparemos los valores de Correlación (Cor), Distancia (Dis) y Proximidad Simple (PS) en la siguiente matriz, donde observamos los valores extraordinarios en d7:v2 y d7:v3:

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

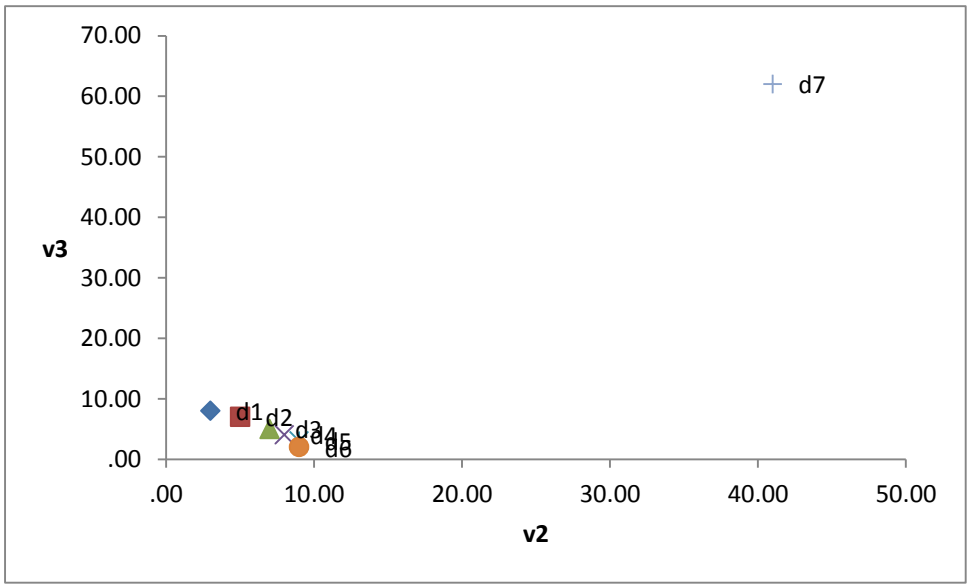
Cor	v1	v2	v3
v1	1.00	.68	.50
v2	.68	1.00	.97
v3	.50	.97	1.00

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

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

De esta manera, la Correlación (v2, v3) presenta un alto valor (.97), que ha producido el valor extraordinario d7, lo que podemos observar en el gráfico siguiente:

[25] Especificamos que x e y son valores no negativos (0 o positivos). El valor máximo de la Proximidad es 1, que se presenta cuando x = y, mientras que el mínimo es 0, cuando x = 0 o y = 0. Definimos que la proximidad de cuando x = y = 0 es 1, por considerar que su coindicencia es perfecta.



Lo mismo puede decirse de la Distancia (v2, v3) que presenta un alto valor (.85), lo que es un efecto de la diferencia grande de las coordenas de d7.

4.6.2. Distancia regular / proximidad regular

Desde el punto de vista matemático, la base de la «Distancia» (D) es la diferencia de los dos valores, x e y:

$D = |x - y|$

La razón por la que utilizamos el valor absoluto (|...|) es que, por ejemplo, D(2, 5) es igual que D(5, 2). En un ejemplo mas concreto, la distantia del punto X, que está a 2 km del punto actual (0), con respecto al punto Y, que está a 5 km es la misma que la distancia de Y con respecto a X:

-----|-----|-----|-----|-----|-----|-----|-----|-----|

X ← 3 → Y

Aquí conviene comprobar que la distancia debe ser un valor no negativo, lo que es razón por la que utilizamos el signo de valor absoluto (|...|), en la ecuacion anterior. Y la misma ecuacion puede expresarse utilizando el exponente:

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

Por lo tanto,

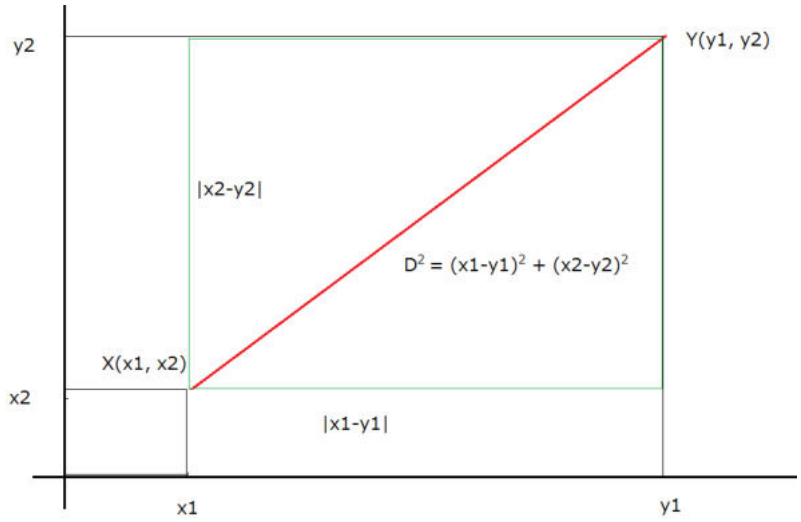
$D^2 = (x - y)^2$

donde suponemos que los valores x e y se sitúan en una misma línea recta.

Si los dos puntos X e Y se encuentran en un plano bidimensional, es decir, si los dos puntos tienen dos coordenadas, $X(x_1, x_2)$, $Y(y_1, y_2)$, respectivamente, la Distancia (D) entre los dos es:

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

lo que representa el largo al cuadrado de la línea diagonal en el grafico siguiente, demostrable por el teorema de Pitágoras:



De la misma manera, la Distancia (D) entre los dos puntos situados en un espacio tridimensional $X(x_1, x_2, x_3)$ e $Y(y_1, y_2, y_3)$ es:

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

Por lo tanto, la Distancia (D) entre los dos puntos situados en un espacio n-dimensional, $X(x_1, x_2, \ldots, x_n)$ e $Y(y_1, y_2, \ldots, y_n)$, es la Suma de Diferencias Cuadradas (SDC):

$$D^2 = (x_1 - y_1)^2 + (x_2 - y_2)^2 + \ldots + (x_n - y_n)^2 = \Sigma_{i=1,n} (x_i - y_i)^2$$

$$SDC(x, y) = \Sigma_{i=1,n} (x_i - y_i)^2 \quad \ldots(1)$$

Por ejemplo, la Suma de Diferencias Cuadradas (SDC) entre las dos coumnas X e Y que están en un espacio de 4 dimensiones es:

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

En lo siguiente partimos de la Suma de Diferencias Cuadradas (SDC) entre los dos vectores X e Y para llegar al concepto de la «Distancia Regular» (DR):

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

$$\rightarrow \Sigma_i (x_i - y_i)^2 = \Sigma_i (x_i^2 + y_i^2) - 2\Sigma_i x_iy_i$$

$$\rightarrow \Sigma_i (x_i - y_i)^2 + 2\Sigma_i x_iy_i = \Sigma_i (x_i^2 + y_i^2)$$

$$\rightarrow [\Sigma_i (x_i - y_i)^2 + 2\Sigma_i x_iy_i] \,/\, \Sigma_i (x_i^2 + y_i^2) = 1$$

$$\rightarrow \Sigma_i (x_i - y_i)^2 \,/\, \Sigma_i (x_i^2 + y_i^2) + 2\Sigma_i x_iy_i \,/\, \Sigma_i (x_i^2 + y_i^2) = 1 \ldots (3)$$

Definimos el primer término de (3) como «Distancia Regular» (DR) y el segundo término de (3) como «Proximidad Regular» (PR).

$$DR(x, y) = \Sigma_i (x_i - y_i)^2 / \Sigma_i (x_i^2 + y_i^2) \quad \ldots(4)$$
$$PR(x, y) = 2\Sigma_i x_i y_i / \Sigma_i (x_i^2 + y_i^2) \quad \ldots(5)^{26}$$

Las relaciones entre la Distancia Regular (DR) y la Proximidad Regular (PR) son:

$$DR + PR = 1,\ DR = 1 - PR,\ PR = 1 - DR \quad \leftarrow(3), (4), (5)$$

Al observar el numerador de (4), entendemos que la Distancia Regular (DR) llega a ser el valor mínimo (0) cuando $\Sigma_i (x_i - y_i)^2 = 0$, es decir, cuando $x_i = y_i$ (i=1,2,...,n), lo que es lógico puesto que los dos puntos en comparación coindicen[27].

Seguidamente buscamos el valor máximo de la Distancia Regular (DR), al tratarse de los datos no negativos, por ejemplo, frecuencias. El máximo debe de presentarse cuando $x_i = 0$ (i=1, 2, ..., n) o $y_i = 0$ (i=1, 2, ..., n)[28]. Cuando $y_i = 0$ (i=1, 2, ..., n), su valor máximo es:

$$DR(x, 0) = \Sigma_i (x_i - 0)^2 / \Sigma_i (x_i^2 + 0^2) = \Sigma_i x_i^2 / \Sigma_i x_i^2 = 1$$

---

[26] La fórmula (5) de la «Proximidad Regular» (PR) es parecida a la del Coeficiente Coseno (Cos), pero difierede él en el multiplicador 2 y el denominador:

$$Cos(x, y) = \Sigma_i x_i y_i / (\Sigma_i x_i^2 * \Sigma_i y_i^2)^{1/2} \quad \ldots(\text{Kin 2009: 161})$$

Manly (1986: 53) presenta la Distancia (D_2), de la cual el segundo término corresponde al coeficiente Coseno:

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

Ambas fórmulas no sos operables cuando uno de $\Sigma_i x_i^2$ y $\Sigma_i y_i^2$ es 0. Sin embargo, como vemos anteriormente, el momento en que uno de $\Sigma_i x_i^2$ y $\Sigma_i y_i^2$ es cero es importante puesto que la Distancia Regular (DR) llega al máximo valor 1 al tratarse de los datos no negativos.

Kin, Meitetsu. (2009) 金明哲『テキストデータの統計科学入門』岩波書店

Manly, Bryan F. J. (1986) *Multivariate Statistic Methods. A Primer.* London: Chapman & Hill.

[27] Se puede considerar un extremo caso excepcional de $\Sigma_i x_i^2 = \Sigma_i y_i^2 = 0$, por lo anto, $x_i = y_i = 0$ (i = 1, 2, ..., n). En tal caso el calculo de la Distancia Regular se hace imposible por ser el denominador 0. Por lo tanto definimos antemano que DR(0, 0) = 0, por la razon de que los dos cero vectores coinciden y su distancia debe de ser 0.

[28] El maximo de la distancia es la distancia con respecto al punto de inicio, que es 0.

De misma manera, cuando $x_i = 0$ (i=1, 2, …, n), DR llega a su valor máximo:.

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

La tabla siguiente demuestra un ejemplo del cálculo de la Distancia Regular (DR):

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

La siguiente es un ejemplo del valor mayor de DR:

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

Cuando X = Y, la Dr ofrece el valor mínimo (0):

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

Cuando Y = 0, la DR llega al ma/ximo (=1):

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

● **Proximidad regular**

Como hemos visto anteriormente, la «Proximidad Regular» (PR) es un valor complementario de la «Distancia Regular» (DR) con respecto a 1:

$$PR(x, y) = 1 - DR(x, y) \qquad ...(1)$$

Naturalmente el rango de la PR es [0, 1]. Sus condiciones, sin embargo, son inversas de las de la DR:

$$PR(x, x) = 1 - DR(x, x) = 1 - 0 = 1 \text{ (Máximo)}$$
$$PR(x, 0) = 1 - DR(x, 0) = 1 - 1 = 0 \text{ (Mínimo)}$$
$$PR(0, y) = 1 - DR(0, y) = 1 - 1 = 0 \text{ (Mínimo)}$$

4.6.3. Distancia media / proximidad media

La tabla siguiente muestra las frecuencias de X e Y, su diferencia (X-Y), la diferencia cuadrada $(X-Y)^2$, X cuadrada, Y cuadrada, la suma de X^2 y Y^2, y finalmente la diferencia cuadrada dividida por la suma de X^2 y Y^2:

M	X	Y	X-Y	$(X-Y)^2$	X^2	Y^2	X^2+Y^2	$(X-Y)^2/(X^2+Y^2)$
h1	10	19	-9	81	100	361	461	.176
h2	11	7	4	16	121	49	170	.094
h3	0	0	0	0	0	0	0	.000
h4	0	1	-1	1	0	1	1	1.000
							Media	.317

Definimos la «Distancia Media» (DM) como la media de diferencia cuadrada dividica por la suma de los valores cuadrados:

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

El segundo término de DM, lo definimos como «Proximidad Media» (PM):

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

Por lo tanto:

$$DM + PM = 1, \ DM = 1 - PM, \ PM = 1 - DM$$

Al observar el numerador de la Distancia Media (DM), sabemos que DM llega a tener el valor mínimo (0) cuando $\Sigma_i (x_i - y_i)^2 = 0$, es decir, cuando $x_i = y_i$ (i=1,2,…,n).

Ahora veamos los casos en que la DM llega a su valor máximo. Cuando $y_i = 0$ (i=1, 2, …, n), la DM llega al máximo. Su valor máximo es:

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

De misma manera, cuando $x_i = 0$ (i=1,2,…,n), la DM llega al máximo. En este caso su valor máximo también es:

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

El rango de la «Proximidad Media» (PM) es igual que el de la «Distancia Media» (DM): [0, 1]. Sus condiciones, sin embargo, son inversas a las de DM:

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

La tabla siguiente demuestra los procesos de los calculos de la «Distancia Media» (DM) y de la «Proximidad Media» (PM):

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

La siguiente muestra una DM mayor:

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

Cuando X = Y, la DM presenta el valor mínimo, y la PM, valor máximo:

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

DM	0.000
PM	1.000

Cuando Y = 0, la DM es máximo (=1) y la PM es mínimo (=0):

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

DM	1.000
PM	0.000

La tabla inferior izquierda es matriz de datos, y la derecha, matriz simétrica de la «Proximidad Media» (PM):.

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

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

● Comparación: Distancia regular / distancia media

Las fórmulas de «Ditancia Regular» (DR) y la de «Distancia Media» (DM) son parecidas:

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

En el cálculo de la DR, la división se hace entre la suma de la diferencia cuadrada y la suma de los valores cuadrados, mientras que en el de la DM, se hace la división de la diferencia por los valores cuadrados y se va sumando todas las veces y finalmente se divide por el núumero de casos (n). Teniendo en cuenta esta diferencia veamos las características de las dos en un ejemplo concreto:

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

En el cálculo de DM (=.058), los valores de h1, h2, h3, h4 están evaluados relativamente dentro de su escala, mientras que en el de RD, los valores pequeños de h1 y h2 no influyen tanto como los de h3 y h4.

Por lo tanto, la DR es sensible a la escala de datos mientras que la DM es estable con respecto al cambio de la escala. Supongamos que X e Y son los dos documentos y h1, h2, h3, h4 representan las frecuencias de un sustantivo, un verbo, una preposición y un artículo definido. En este caso para comparar los dos documentos es más conveniente utilizar la DM. En cambio, si los casos representan los sustantivos de una escala comparable, conviene utilizar la DR, que es sensible a la escala de cada caso.

5. Análisis

Analizamos la matriz de datos desde varios puntos de vista,

5.1. Análisis de medidas estadísticas

5.1.1. Análisis de rango

Examinamos los valores relacionados con el Rango: Mínimo, Mitad, Máximo y Rango:

X	v1	v2	v3
d1	38	18	5
d2	35	10	6
d3	28	44	48
d4	22	30	62
d5	24	29	89

X	v1	v2	v3
Mínimo	22	10	5
Mitad	30	27	47
Máximo	38	44	89
Rango	16	34	84

5.1.2. Análisis de centralidad

Examinamos los valores relacionados con la centralidad de los datos: Media, Mediana, Moda, con respecto a Media, Mediana, Mitad en Diferencia, Valor Contrastivo y Posición en Rango. El cálculo del Valor Contrastivo (Contr) entre, por ejemplo, Media (Me) y Mediana (Md), consiste en:

Contr(Me, Md) = (Me - Md) / (Me + Md)

La fórmula de la Posición en Rango (Pos.Rg) de Media (Me) es:

Pos.Rg(Me) = (Me – Min) / Rg

donde Min es Mínimo y Rg es Rango.

X	v1	v2	v3
Media	29.40	26.20	42.00
Mediana	28.00	29.00	48.00
Diferencia (-)	1.40	-2.80	-6.00
Contraste	.02	-.05	-.07
Mitad	30.00	27.00	47.00
Diferencia (-)	-.60	-.80	-5.00
Contraste	-.01	-.02	-.06

X	v1	v2	v3
Mediana	28.00	29.00	48.00
Media	29.40	26.20	42.00
Diferencia (-)	-1.40	2.80	6.00
Contraste	-.02	.05	.07
Mitad	30.00	27.00	47.00
Diferencia (-)	-2.00	2.00	1.00
Contraste	-.03	.04	.01

Pos. en Rng.	.46	.48	.44

Pos. en Rng.	.38	.56	.51

X	v1	v2	v3
Media mayor	29.11	26.33	41.56
Media	28.00	29.00	48.00
Diferencia (-)	1.11	-2.67	-6.44
Contraste	.02	-.05	-.07
Mitad	30.00	27.00	47.00
Diferencia (-)	-.89	-.67	-5.44
Contraste	-.02	-.01	-.06
Pos. en Rng.	.44	.48	.44

X	v1	v2	v3
Moda mayor	25.00	26.00	66.00
Media	28.00	29.00	48.00
Diferencia (-)	-3.00	-3.00	18.00
Contraste	-.06	-.05	.16
Mitad	30.00	27.00	47.00
Diferencia (-)	-5.00	-1.00	19.00
Contraste	-.09	-.02	.17
Pos. en Rng.	.19	.47	.73

5.1.3. Análisis de variación

Comparamos medidas estadísticas que indican la variación en torno a la Media.

X	v1	v2	v3
d1	38	18	5
d2	35	10	6
d3	28	44	48
d4	22	30	62
d5	24	29	89

X	v1	v2	v3
Varianza	38.24	133.76	1062.00
Dev. est.	6.18	11.57	32.59
Coef. de variación	.21	.44	.78
Desviación est. nrm.	.11	.22	.39
Dispersión	.89	.78	.61

(*) Análisis de variación con clase

La tabla siguiente muestra Suma, Media, Varianza y Desviación Típica (DT) de las filas:

X	v1	v2	v3
d1	38	18	5
d2	35	10	6
d3	28	44	48
d4	22	30	62
d5	24	29	89

Xr	v1	v2	v3
Suma	147	131	210
Media	29.40	26.20	42.00
Varianza	38.24	133.76	1062.00
DT	6.18	11.57	32.59

Por otra parte, a la tabla siguiente añadimos otra columna más: Clase (Cn1). Multiplicamos la Tabla por la columna Clase y, de nuevo, calculamos las medidas anteriores:

Y	v1	v2	v3	Clase
d1	38	18	5	1
d2	35	10	6	2
d3	28	44	48	3
d4	22	30	62	4
d5	24	29	89	5

Yr	v1	v2	v3
Suma	400	435.	854.
Media	2.72	3.32	4.07
Varianza	2.00	1.64	.98
DT	1.41	1.28	.99

M_{1p} = SumV(Xnp * Cn1) ←Suma vertical con clase
M_{1p} = M_{1p} / SumV(Xnp) ←Media vertical con clase
V_{1p} = SumV(Cn1 − M_{1p})^2 * Xnp / SumV(Xnp) ←Varianza vertical
V_{1p} = $V_{1p}^{1/2}$ ←Desviación Típica vertical

donde Vnp es la matriz de datos, Cn1 es la columna de Clase. El objetivo aquí es comparar la variación basada en un criterio expresado cn Clase. En el análisis simple de Media y Varianza se calculan estos valores dentro de datos de manera homogénea (Xr), mientras en Yr, utilizamos los valores de Clase particular en cada fila. Por ejemplo la columna v1 tiene la Media de 29.40, y en Yr presenta 2.72 de acuerdo con la escala de Clase.

Al comparar la Varianza y Desviación Típica (DT) en Xr e Yr, nos damos cuenta de que en Xr, el orden es: v1 < v2 < v3; mientras que en Yr el orden es reverso: v1 > v2 > v3. Observando la matriz de datos, comprobamos que efectivamente en v3 encontramos una concentración numérica en d3, d4, d5, cuyo centro es 4.07, la Media de Yr. La Varianza de v3 parece dispersa en conjunto, mientras que si nos enfocamos en las clases mayores (3, 4, 5), detectamos una concentración grande. La Media de v3 en Xr es 42.00, que se sitúa entre d2 y d3, mientras que la misma en Yr es 4.07, que corresponde al intervalo de d4 y d5.

De esta manera, para observar la Variación, el analista puede tomar dos puntos de vista diferentes: el análisis simple es para hacer la observación general y el «Análisis de variación con clase» con un criterio exterior con un orden establecido, por ejemplo, orden cronológico.

(#) Formas abreviadas en documentos notariales medievales y modernas

En los documentos emitidos en la España de la Edad Media y Moderna, se escribían tales forma abreviadas *d, dl, dla, dlos, dha, dho,* etc. en lugar de las formas plenas: *de, del, dela, delos, dicha, dicho,* etc. La tabla siguiente ofrece las frecuencias por mil normalizada en Medias Fraccionales:

NS.FM	d<e>	d<e>l	d<e>la	d<e>los	d<ic>ha	d<ic>ho	d<ic>hos	dich<o>
1260	348	61	28	174				22
1280	100	66	71					541
1300	629	824	922	686	3			2556
1320	1048	1087	1016	438				5250
1340	215	237	379	103				833
1360	1196	702	805	273				2289
1380	906	1147	1081	451	37	23	65	1372
1400	545	387	396	210	13	24	27	706
1420	981	847	517	331	153	299	195	63
1440	989	1354	938	138	461	548	233	18
1460	914	473	397	158	250	306	303	
1480	2623	1669	902	201	1118	1164	598	
1500	1465	1207	811	412	776	687	541	10
1520	1503	1110	629	231	1021	1052	667	
1540	2707	1854	842	284	1315	1719	865	
1560	660	481	280	121	1533	1192	901	
1580	154	52	88		554	611	457	
1600	558	378			1490	1566	1049	
1620	30				93	63	74	
1640	66				1570	1932	1170	
1660					288	229	78	
1680	43				3566	4953	2579	

En esta tabla observamos que existen cambios cronológicos concentrados. Las formas d<e>, d<e>l, d<e>la, d<e>los llegan a su cumbre en segunda mitad del siglo XV, mientras que las formas de d<ic>ho se encuentran en épocas relativamente tardías. La forma dich<o> es peculiar en su distribución temprana, que parece demostrar la forma apocopada que era frecuente en el siglo XIV.

La tabla siguiente muestra el resultado del Análisis de Variación por Clase.

NS.FM	d<e>	d<e>l	d<e>la	d<e>los	d<ic>ha	d<ic>ho	d<ic>hos	dich<o>
Suma	192590	144323	91631	34261	245069	286587	168847	63807
Media	10.89	10.36	9.07	8.14	17.21	17.51	17.23	4.67
Varianza	18.29	17.43	16.91	19.21	14.90	15.97	15.77	2.59
Desv. típ.	4.28	4.18	4.11	4.38	3.86	4.00	3.97	1.61
C. de variación	.39	.40	.45	.54	.22	.23	.23	.34
Asim. normaliz.	.27	.12	.03	.20	.42	.57	.48	.58
Curt. normaliz.	.20	.11	.07	.16	.33	.49	.39	.45
Pos. de media	1460	1440	1420	1400	1580	1600	1580	1340

De acuerdo con la Media, se sabe el lugar que ocupa el centro de las columnas en la fila de Pos[ición] de media. Por ejemplo la forma d<e> tiene la media de distribución en 1460. Las Medias de las formas *d<e> d<e>, d<e>l, d<e>la, d<e>los* se sitúan en el siglo XV, mientras que las Medias de *d<ic>ha, d<ic>ho, d<ic>hos* se encuentran en el siglo XVI y principios de XVII. La última forma *dich<o>* es peculiar en su situación temprana, en el siglo XIV. Las formas de «de» muestra una distribución prolongada tanto en Varianza como en Curtosis, mientras que las formas de «dicho» presenta una concentración más pronunciada. Estas observaciones corroboran la teoría de que los escribanos mantenían una norma establecida en cada época con los determinados vocablos.

5.1.4. Análisis de balanza

Para ver el equilibrio de la distribución de los datos, consideramos «Grado de Balanza» con respecto a la Mitad, Mediana y Media:

X	v1	v2	v3
d1	38	18	5
d2	35	10	6
d3	28	44	48
d4	22	30	62
d5	24	29	89

Por ejemplo, la Mitad de v1 {38, 35, 28, 22, 24) es (Máximo + Mínimo) / 2 = (38 + 22) / 2 = 30. El «Número positivo» (Ps) con respecto a la Mitad es 2 (38, 35) y el «Número negativo» (Ng) es 3 (28, 22, 24). Y el «Número contrastivo por Mitad» (NC.Mi) es:

NC.Mi = (Ps – Ng) / (Ps + Ng) = (2 – 3) / (2 + 3) = -.20

lo que demuestra que el número de datos va al punto ligeramente inferior de la Mitad.

Ahora bien, en lugar del número de datos, también nos interesa las diferencias que presenta cada dato con respecto a la Mitad. Por ejemplo, dentro del mismo conjunto d1, los datos positivos, 38, 35 presentan las diferencias: 38 – 30 = 8, 35 – 30 = 5, en total 13. Las diferencias negativas son: 30 - 28 = 2, 30 – 22 = 8, 30 – 24 = 6, en total 16. A partir de estos dos números, positivo y negativo, calculamos el «Valor contrastivo por Mitad» (Vc.Mi):

VC.Mi = (Ps – Ng) / (Ps + Ng) = (13 ― 16) / (13+16) = -.10

X	v1	v2	v3
Mitad	30.00	27.00	47.00
Pos.Cnt.	2	3	3
Neg.Cnt.	3	2	2
Balanza	-.20	.20	.20
Asimetría típica normalizada	.05	.09	.15

X	v1	v2	v3
Mitad	30.00	27.00	47.00
Pos.Valor	13	22	58
Neg.Valor	16	26	83
Balanza	-.10	-.08	-.18
Asimetría típica normalizada	.05	.09	.15

Tanto el NC.Mi como el VC.Mi se vueve 0, cuando Ps = Ng; positivo, cuando Ps > Ng; negativo, cuando Ps < Ng. Su rango es [-1 ~ 1], exclusivos los dos extremos.

Lo mismo puede hacerse no solamente con la Mitad, sino también con Media y Mediana. Con la Media, el Valor Contrastivo se vuelve constantemente 0, de modo que utilizamos el Número contrastivo. En cambio, con Mediana, utilizamos el Valor Contrastivo, puesto que el Número contrastivo presenta lógicamente 0 constante.

X	v1	v2	v3
Media	29.40	26.20	42.00
Pos.Cnt.	2	3	3
Neg.Cnt.	3	2	2
Balanza	-.20	.20	.20
Asimetría típica normalizada	.05	.09	.15

X	v1	v2	v3
Mediana	28.00	29.00	48.00
Pos.Valor	17	16	55
Neg.Valor	10	30	85
Balanza	.26	-.30	-.21
Asimetría típica normalizada	.05	.09	.15

5.1.5. Análisis de oscilación

Analizamos índice de Oscilación con números de Subidas y Bajadas y otro con cantidad de Subidas y Bajadass observadas en la secuencia de datos. Par el «Índice de Oscilación en Número» (ION) contamos las veces de subidas y bajadas. Por ejemplo, en d1 {10, 19, 14, 7, 12} hay 2 subidas (Sn) en 10 → 19 y 7 → 12; y 2 bajadas (Bn) en 19 → 14 y 14 → 7. La fórmula de ION es:

$$ION = (Sn - Bn) / (Sn + Bn)$$

Por ejemplo:

$$ION(d1) = (2 - 2) / (2 + 2) = 0$$
$$ION(d3) = (2 - 1) / (2 + 1) = .333$$
$$ION(d4) = (3 - 0) / (3 + 0) = 1.000$$

Seguidamente, calculamos el «Índice de Oscilación en Cantidad» (IOC).

Ahora tomamos en cuenta la cantidad de cada subida y bajada. Por ejemplo, en las subidas de 10 → 19 y 7 → 12, hay cantidad sumada de subidas (Sc): 14 (9 + 5). Y en las bajadas de 19 → 14 y de 14 → 7, hay cantidad sumada de bajadas (Bc): 12 (5 + 7). La fórmula de IOC es:

$$IOC = (Sc - Bc) / (Sc + Bc)$$

Por ejemplo:

$$ION(d1) = (14 - 12) / (14 + 12) = 0.077$$

A	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

A	Sn	Bn	ION	Sc	Bc	IOC
d1	2	2	.000	14	12	.077
d2	2	2	.000	4	14	-.556
d3	2	1	.333	12	11	.043
d4	3	0	1.000	3	0	1.000

Se observa que la fila d2 tiene la oscilación negativa, es decir, la tendencia de bajada, mientras que la fina 4, la oscilación positiva perfecta.

5.2. Análisis de concentración

Denominamos 《Análisis de Concentración》 al método que permite elaborar la distribución concentrada cambiando orden de filas y columnas de acuerdo con los valores que se dan a los dos ejemplos de la matriz de datos. Aparte de los métodos establecidos como el Análisis de Correspondencia, Análisis de Cluster, Análisis de Componentes Principales, y Análisis Factorial, de los que trataremos en la sección siguiente, de nuestra parte proponemos utilizar la Distancia con respecto al punto cero de la Matriz para dotar a sus ejes es decir, el 《Análisis de Concentración por Distancia》 (ACD).

5.2.1. Concentración con criterio exterior

Abordemos primero la «Concentración con criterio exterior». El objetivo es obtener la distribución donde los puntos de reacción se concentren en la parte diagonal de la matriz que corre de la parte superior izquierda a la parte inferior derecha cambiando el orden de filas. Fijése en el cambio del orden de filas de d1-d2-d3-d4 a → d1-d3-d5-d2-d4, y la concentración diagonal de los puntos reactivos «v»:

Lv	v1	v2	v3	v4
d1	v	v		
d2			v	
d3		v		
d4			v	v
d5	v	v	v	

→

Lv	v1	v2	v3	v4
d1	v	v		
d3		v		
d5	v	v	v	
d2			v	
d4			v	v

Una vez reordenada la Matriz, notamos que hay tales agrupamientos de filas como [d1, d3, d5] y [d2, d4] y, al mismo tiempo, de columnas: [v1, v2] y [v3, v4]. La concentración de puntos muestra una tendencia general de distribución, que permite una interpretación unificada tanto de la fila como de la columna.

Para efectuar la ordenación de filas, se necesitan los valores de filas que se obtienen en forma de Media de distancia con respecto al punto cero de la manera siguiente:

$$d1: [(1^2 + 2^2) / 2]^{1/2} = 1.581 \quad (\ldots 1)$$
$$d2: [(3^2) / 1]^{1/2} = 3.000 \quad (\ldots 4)$$
$$d3: [(2^2) / 1]^{1/2} = 2.000 \quad (\ldots 2)$$
$$d4: [(3^2 + 4^2) / 2]^{1/2} = 3.535 \quad (\ldots 5)$$
$$d5: [(1^2 + 2^2 + 3^2) / 3]^{1/2} = 2.160 \quad (\ldots 3)$$

El vector vertical de las Distancias horizontales de filas se formula:

$$Dn1 = [SumH(X_{np} * S_{p1}^2) / SumH(X_{np})]^{1/2}$$

donde SumH es una función que devuelve un vector vertical de la Suma horizontal de las filas:

$$SumH(Xnp) = X_{np}\ I_{p1}$$

Sp1 es un vector vertical cuyos elementos son números secuenciales: {1, 2, ..., P}.

(*) Problema de la misma distancia en distribución diferente

Existen pares de filas diferentes cuyas distancias con iguales, por ejemplo Y.

Y	v1	v2	v3	v4	v5	v6	v7	v8	v9	v10	v11
d1				v		v					v
d2			v					v		v	

Y	Distancia
d1	7.594
d2	7.594

Los cálculos de sus distancia son:

d-1 ... $[(4^2 + 6^2 + 11^2) / 3]^{1/2} = 7.594$
d-2 ... $[(3^2 + 8^2 + 10^2) / 3]^{1/2} = 7.594$

El mismo problema se resuelve utilizando la Distancia de Minkowski con exponente 3:

P2	v1	v2	v3	v4	v5	v6	v7	v8	v9	v10	v11
d2			v					v		v	
d1				v		v					v

P2	valor
d2	8.005
d1	8.128

d1 ... $[(4^3 + 6^3 + 11^3) / 3]^{1/3} = 8.128$
d2 ... $[(3^3 + 8^3 + 10^3) / 3]^{1/3} = 8.005$

(#) Criterio exterior geográfico

Hemos elaborado la matriz siguiente basándonos en el estudio de Cahuzac (1980). Las formas que designan agricultores en Hispanoamérica están ordenadas alfabéticamente en el eje vertical. Los países están colocados de Norte a Sur en el eje horizontal. (México, Cuba, República Dominicana, Puerto Rico, Guatemala, El Salvador, Honduras, Nicaragua, Costa Rica, Panamá, Venezuela, Colombia, Ecuador, Perú, Bolivia, Chile, Paraguay, Uruguay, Argentina).

Agricultor	MX	CU	RD	PR	GU	HO	EL	NI	CR	PN	VE	CO	EC	PE	BO	CH	PA	UR	AR
01 cacahuero											v	v							
02 cafetalista	v	v		v															
03 camilucho																	v	v	v
04 campero																	v	v	v
05 camperuso											v	v							
06 campirano					v	v	v	v	v	v	v	v							
07 campiruso					v	v	v	v	v	v									
08 campista	v			v	v	v	v	v	v	v	v								
09 campusano										v								v	v
10 campuso					v	v	v	v	v										
11 colono			v	v															
12 comparsa																	v	v	v
13 conuquero		v	v	v							v	v							
14 coquero													v	v	v				
15 chagrero												v	v						
16 changador																	v	v	v
17 chilero	v				v	v	v	v	v	v									
18 chuncano																	v	v	v
19 enmaniguado		v	v	v															
20 estanciero																	v	v	v
21 gaucho															v		v	v	v
22 guajiro		v	v																
23 guanaco					v	v	v	v	v	v									
24 guaso		v											v	v	v	v			v
25 huasicama												v	v						
26 huertero	v													v		v			v
27 hulero	v				v	v	v	v	v	v									
28 invernador														v		v	v	v	v
29 jíbaro			v	v															
30 lampero														v	v				v
31 lanudo											v	v	v						
32 llanero											v	v							
33 macanero	v								v										
34 manuto			v							v									
35 montero		v	v																
36 montubio		v	v									v	v	v					
37 paisano			v										v	v					
38 pajuerano		v											v	v					
39 partidario															v			v	v
40 payazo											v	v							
41 piona																	v	v	v
42 ranchero	v	v	v							v									
43 rondín															v				
44 sabanero											v	v							
45 veguero			v							v									
46 viñatero														v		v	v	v	v
47 yanacón														v	v	v			v

Al reordenar las filas por la Distancia con el criterio exterior de columnas (países), se obtiene la matriz concentrada siguiente:

Dst.R	MX	CU	RD	PR	GU	HO	EL	NI	CR	PN	VE	CO	EC	PE	BO	CH	PA	UR	AR
22 guajiro		v	v																
35 montero		v	v																
02 cafetalista	v	v		v															
19 enmaniguado		v	v	v															
11 colono			v	v															
29 jíbaro			v	v															
42 ranchero	v	v	v							v									
33 macanero	v								v										
10 campuso					v	v	v	v	v										
17 chilero	v				v	v	v	v	v	v									
27 hulero	v				v	v	v	v	v	v									
08 campista	v			v	v	v	v	v	v	v	v								
07 campiruso					v	v	v	v	v	v									
23 guanaco					v	v	v	v	v	v									
34 manuto			v							v									
45 veguero			v							v									
13 conuquero		v	v	v							v	v							
06 campirano					v	v	v	v	v	v	v	v							
36 montubio		v	v									v	v	v					
01 cacahuero											v	v							
05 camperuso											v	v							
32 llanero											v	v							
40 payazo											v	v							
44 sabanero											v	v							
38 pajuerano		v											v	v					
37 paisano			v										v	v					
31 lanudo											v	v	v						
15 chagrero												v	v						
25 huasicama												v	v						
14 coquero													v	v	v				
24 guaso		v											v	v	v	v			v
43 rondín															v				
26 huertero	v													v		v			v
47 yanacón														v	v	v			v
30 lampero														v	v				v
09 campusano										v								v	v
28 invernador														v		v	v	v	v
46 viñatero														v		v	v	v	v
21 gaucho															v		v	v	v
39 partidario															v			v	v
03 camilucho																	v	v	v
04 campero																	v	v	v
12 comparsa																	v	v	v
16 changador																	v	v	v
18 chuncano																	v	v	v
20 estanciero																	v	v	v
41 piona																	v	v	v

De esta manera podemos observar las formas concentradas en cada grupo de países.

5.2.2. Concentración con criterio interior

En esta sección intentamos realizar la concentración sin establecer un criterio exterior. De modo que reordenaremos tanto las filas como las columnas.

El dato de ejemplo posee variables v1, v2, … v5, por ejemplo, nombres

de localidades. La tabla inferior derecha muestra el resultado de concentración diagonal, con reordenaciones bilaterales. Esta tabla nos permite analizar tanto los datos desde el punto de vista de localidades, como las localidades desde el punto de vista de datos, con ambas partes agrupadas.

X	v1	v2	v3	v4
d1	v	v		
d2			v	
d3		v		
d4			v	v
d5	v	v	v	

→

Y	v2	v1	v3	v4
d3	v			
d1	v	v		
d5	v	v	v	
d2			v	
d4			v	v

El método de «Concentración con Criterio Interior» consiste en buscar la distribución concentrada (Y) a partir de la matriz de los datos (X). El método es sencillo a pesar de que repetimos varias veces la misma operación. Primero realizamos los mismos cálculos que el caso anterior, con criterio exterior:

$d1: [(1^2 + 2^2) / 2]^{1/2} = 1.581$ (...1)
$d2: [(3^2) / 1]^{1/2} = 3.000$ (...4)
$d3: [(2^2) / 1]^{1/2} = 2.000$ (...2)
$d4: [(3^2 + 4^2) / 2]^{1/2} = 3.535$ (...5)
$d5: [(1^2 + 2^2 + 3^2) / 3]^{1/2} = 2.160$ (...3)

Reordenando las filas de acuerdo con el orden calculado, obtenemos la tabla siguiente:

Lv	v1	v2	v3	v4
d1	v	v		
d3		v		
d5	v	v	v	
d2			v	
d4			v	v

Lv	Dist.
d1	1.581
d3	2.000
d5	2.160
d2	3.000
d4	3.536

Seguidamente, calculamos ahora las distancias de columnas con respecto al punto cero:

$v1: [(1^2 + 3^2) / 2]^{1/2} = 2.236$ (...2)
$v2: [(1^2 + 2^2 + 3^2) / 3]^{1/2} = 2.160$ (...1)
$v3: [(3^2 + 4^2 + 5^2) / 3]^{1/2} = 4.082$ (...3)
$v4: [(5^2) / 1]^{1/2} = 5.000$ (...4)

Según las distancias calculadas, tenemos que cambiar el orden de las columnas v1 y v2:

Lv	v2	v1	v3	v4
d1	v	v		
d3	v			
d5	v	v	v	
d2			v	
d4			v	v

Lv	Dist.
d1	1.581
d3	1.000
d5	2.160
d2	3.000
d4	3.536

Lv	v2	v1	v3	v4
Dist.	2.160	2.236	4.082	5.000

De esta manera, hemos terminado la primera ronda de reordenación vertical y horizontal. Ahora bien, de nuevo calculamos las distancias horizontales de filas:

d1: $[(1^2 + 2^2) / 2]^{1/2}$ = 1.581 (...2)
d3: $[(1^2) / 1]^{1/2}$ = 1.000 (...1)
d5: $[(1^2 + 2^2 + 3^2) / 3]^{1/2}$ = 2.160 (...3)
d2: $[(3^2) / 1]^{1/2}$ = 3.000 (...4)
d4: $[(3^2 + 4^2) / 2]^{1/2}$ = 3.535 (...5)

Ahora tenemos que cambiar el orden de las filas d1 y d3:

Lv	v2	v1	v3	v4
d3	v			
d1	v	v		
d5	v	v	v	
d2			v	
d4			v	v

Lv	Dist.
d3	1.000
d1	1.581
d5	2.160
d2	3.000
d4	3.536

Lv	v2	v1	v3	v4
Dist.	2.160	2.550	4.082	5.000

Y calculamos las distancias verticales de columnas:

v-2: $[(1^2 + 2^2 + 3^2) / 3]^{1/2}$ = 2.160 (...1)
v-1: $[(2^2 + 3^2) / 2]^{1/2}$ = 2.550 (...2)
v-3: $[(3^2 + 4^2 + 5^2) / 3]^{1/2}$ = 4.082 (...3)
v-4: $[(5^2) / 1]^{1/2}$ = 5 (...4)

En este momento, tanto las filas como las columnas están ordenadas correctamente y terminamos las operaciones de reordenación.

(#) Criterio interior geográfico

Vamos a analizar los mismos datos de Cahuzac (1980) con el método de Concentración con criterior interior:

Dst.A	GU	HO	EL	NI	CR	PR	PN	MX	RD	VE	CU	CO	EC	PE	CH	BO	AR	UR	PA
10 campuso	v	v	v	v	v														
07 campiruso	v	v	v	v	v		v												
23 guanaco	v	v	v	v	v		v												
17 chilero	v	v	v	v	v		v	v											
27 hulero	v	v	v	v	v		v	v											
08 campista	v	v	v	v	v	v	v	v		v									
33 macanero					v			v											
06 campirano	v	v	v	v	v		v			v		v							
11 colono						v			v										
29 jíbaro						v			v										
34 manuto							v		v										
45 veguero							v		v										
02 cafetalista						v		v			v								
42 ranchero							v	v	v		v								
19 enmaniguado						v			v		v								
13 conuquero						v			v	v	v	v							
22 guajiro									v		v								
35 montero									v		v								
01 cacahuero										v		v							
05 camperuso										v		v							
32 llanero										v		v							
40 payazo										v		v							
44 sabanero										v		v							
31 lanudo										v		v	v						
36 montubio									v		v	v	v	v					
37 paisano									v				v	v					
15 chagrero												v	v						
25 huasicama												v	v						
38 pajuerano											v		v	v					
26 huertero								v						v	v		v		
14 coquero													v	v		v			
24 guaso											v		v	v	v	v	v		
09 campusano							v										v	v	
47 yanacón														v	v	v	v		
30 lampero														v		v	v		
43 rondín																v			
28 invernador														v	v		v	v	v
46 viñatero														v	v		v	v	v
39 partidario																v	v	v	
21 gaucho																v	v	v	v
03 camilucho																	v	v	v
04 campero																	v	v	v
12 comparsa																	v	v	v
16 changador																	v	v	v
18 chuncano																	v	v	v
20 estanciero																	v	v	v
41 piona																	v	v	v

En general, al tratar los datos lingüísticos, el analista establece de antemano unos criterios de clasificación y realiza el análisis basado en los mismos criterios sin considerar el criterio interior latente en la estructura de los

datos. Llamamos a este método «Precategorización». El analista limita su capacidad de análisis con el criterio que él mismo establece. En lugar de seguir siempre con los mismos criterios preestablecidos, también es posible categorizar los datos después de análisis con el criterio interior, por lo que el método va a ser más flexible y nos da la ocasión de encontrar nuevos descubrimientos. Llamamos a este método «Poscategorización». Ambos métodos son posibles, pero el último no se adopta con frecuencia en los estudios lingüísticos.

5.2.3. Interpretación de valores de ejes

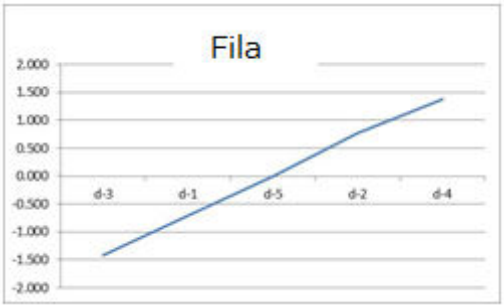
Los valores de distancia que hemos utilizado como medios de reordenación muestran grados de cercanía entre los individuos (filas) por una parte, y variables (columnas) por otra.

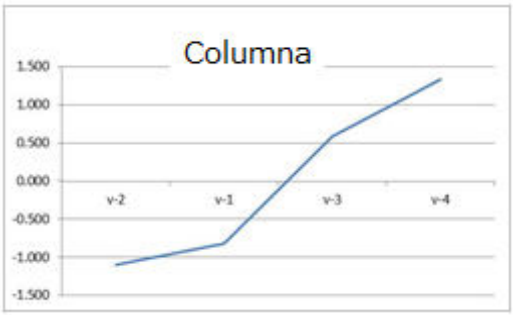
Lv	v2	v1	v3	v4
d3	v			
d1	v	v		
d5	v	v	v	
d2			v	
d4			v	v

Lv	Dist.
d3	1.418
d1	0.709
d5	0.014
d2	0.760
d4	1.381

Lv	v2	v1	v3	v4
Dist.	1.097	0.821	0.582	1.336

Veamos sus gráficos. La subida de los valores de filas es constante, mientras que la de columnas la subida entre v2 y v1 es leve, lo que significa que las dos columnas son cercanas en su distribución de los datos:

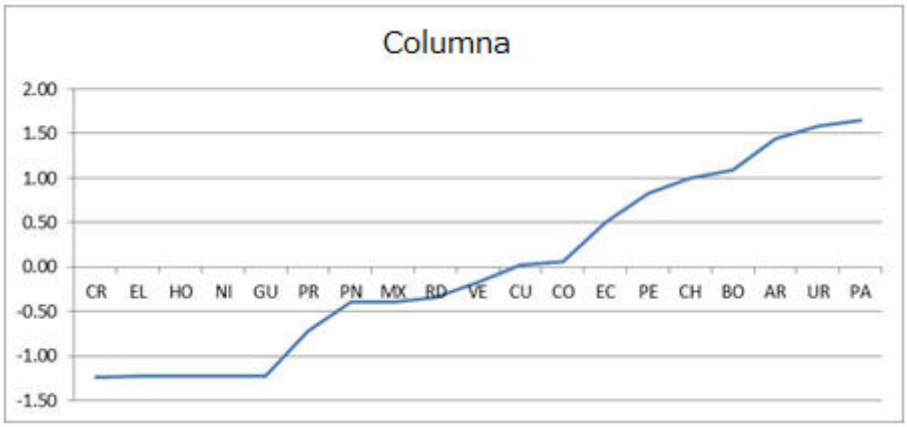




(#) Clasificación geográfica

El gráfico siguiente representa los valores de columnas resultados del Análisis con criterio interior. Observamos grosso modo grupos siguientes: Países centroamericanos (CR, EL, HO, NI, GU); México y países caribeños (MX, DR, VE, CU; Países andinos (CO, EC, PE, CH, BO); Países de La Plata (AR, UR,

PA).:



La línea recta de los países centroamericanos (CR, EL, HO, NI, GU) indica la distribución homogénea de los datos, lo que se observa fácilmente en la tabla anterior (Criterio interior geográfico). Ciertamente las cifras y los gráficos sirven para observar las tendencias globales. Sin embargo, se tratan de los resultados de determinadas transformaciones y abstracciones. Si volvemos a la matriz de los datos reordenados, lo que se observa en cifras y gráficos se presenta más convincente. Las cifras y los gráficos nos ayudan a dar una conclusión final, pero es conveniente comprobar los datos reales en las tablas de la matriz.

(#) Concentración de la matriz de relación

Los gráficos siguientes muestran la matriz de Asociación con el coeficiente de «Preferencia» (gráfico superior) y el resultado de la Concentración (gráfico inferior):

Preference	CU	RD	PR	MX	GU	EL	HO	NI	CR	PN	VE	CO	EC	PE	BO	CH	PA	UR	AR
CU	1,000	0,200	-0,200	-0,500	-1,000	-1,000	-1,000	-1,000	-1,000	-0,789	-0,778	-0,600	-0,294	-0,368	-0,750	-0,714	-1,000	-1,000	-0,840
RD	0,200	1,000	-0,059	-0,778	-1,000	-1,000	-1,000	-1,000	-1,000	-0,429	-0,800	-0,636	-0,579	-0,619	-1,000	-1,000	-1,000	-1,000	-1,000
PR	-0,200	-0,059	1,000	-0,385	-0,692	-0,692	-0,692	-0,692	-0,714	-0,750	-0,467	-0,765	-1,000	-1,000	-1,000	-1,000	-1,000	-1,000	-1,000
MX	-0,500	-0,778	-0,385	1,000	-0,143	-0,143	-0,143	-0,143	0,067	-0,059	-0,750	-1,000	-1,000	-0,765	-1,000	-0,667	-1,000	-1,000	-0,826
GU	-1,000	-1,000	-0,692	-0,143	1,000	1,000	1,000	1,000	0,867	0,412	-0,500	-0,778	-1,000	-1,000	-1,000	-1,000	-1,000	-1,000	-1,000
EL	-1,000	-1,000	-0,692	-0,143	1,000	1,000	1,000	1,000	0,867	0,412	-0,500	-0,778	-1,000	-1,000	-1,000	-1,000	-1,000	-1,000	-1,000
HO	-1,000	-1,000	-0,692	-0,143	1,000	1,000	1,000	1,000	0,867	0,412	-0,500	-0,778	-1,000	-1,000	-1,000	-1,000	-1,000	-1,000	-1,000
NI	-1,000	-1,000	-0,692	-0,143	1,000	1,000	1,000	1,000	0,867	0,412	-0,500	-0,778	-1,000	-1,000	-1,000	-1,000	-1,000	-1,000	-1,000
CR	-1,000	-1,000	-0,714	0,067	0,867	0,867	0,867	0,867	1,000	0,333	-0,529	-0,789	-1,000	-1,000	-1,000	-1,000	-1,000	-1,000	-1,000
PN	-0,789	-0,429	-0,750	-0,059	0,412	0,412	0,412	0,412	0,333	1,000	-0,579	-0,810	-1,000	-1,000	-1,000	-1,000	-1,000	-0,818	-0,846
VE	-0,778	-0,800	-0,467	-0,750	-0,500	-0,500	-0,500	-0,500	-0,529	-0,579	1,000	0,600	-0,765	-1,000	-1,000	-1,000	-1,000	-1,000	-1,000
CO	-0,600	-0,636	-0,765	-1,000	-0,778	-0,778	-0,778	-0,778	-0,789	-0,810	0,600	1,000	-0,158	-0,810	-1,000	-1,000	-1,000	-1,000	-1,000
EC	-0,294	-0,579	-1,000	-1,000	-1,000	-1,000	-1,000	-1,000	-1,000	-1,000	-0,765	-0,158	1,000	0,111	-0,467	-0,692	-1,000	-1,000	-0,833
PE	-0,368	-0,619	-1,000	-0,765	-1,000	-1,000	-1,000	-1,000	-1,000	-1,000	-1,000	-0,810	0,111	1,000	-0,059	0,333	-0,600	-0,636	-0,077
BO	-0,750	-1,000	-1,000	-1,000	-1,000	-1,000	-1,000	-1,000	-1,000	-1,000	-1,000	-1,000	-0,467	-0,059	1,000	-0,333	-0,765	-0,579	-0,130
CH	-0,714	-1,000	-1,000	-0,667	-1,000	-1,000	-1,000	-1,000	-1,000	-1,000	-1,000	-1,000	-0,692	0,333	-0,333	1,000	-0,467	-0,529	-0,048
PA	-1,000	-1,000	-1,000	-1,000	-1,000	-1,000	-1,000	-1,000	-1,000	-1,000	-1,000	-1,000	-1,000	-0,600	-0,765	-0,467	1,000	0,818	0,538
UR	-1,000	-1,000	-1,000	-1,000	-1,000	-1,000	-1,000	-1,000	-1,000	-0,818	-1,000	-1,000	-1,000	-0,636	-0,579	-0,529	0,818	1,000	0,714
AR	-0,840	-1,000	-1,000	-0,826	-1,000	-1,000	-1,000	-1,000	-1,000	-0,846	-1,000	-1,000	-0,833	-0,077	-0,130	-0,048	0,538	0,714	1,000

Dst.cct.	EL	HO	NI	GU	CR	PN	MX	VE	PR	CO	RD	CU	EC	PE	CH	BO	AR	UR	PA
EL	1.000	1.000	1.000	1.000	0.867	0.412	-0.143	-0.500	-0.692	-0.778	-1.000	-1.000	-1.000	-1.000	-1.000	-1.000	-1.000	-1.000	-1.000
HO	1.000	1.000	1.000	1.000	0.867	0.412	-0.143	-0.500	-0.692	-0.778	-1.000	-1.000	-1.000	-1.000	-1.000	-1.000	-1.000	-1.000	-1.000
NI	1.000	1.000	1.000	1.000	0.867	0.412	-0.143	-0.500	-0.692	-0.778	-1.000	-1.000	-1.000	-1.000	-1.000	-1.000	-1.000	-1.000	-1.000
GU	1.000	1.000	1.000	1.000	0.867	0.412	-0.143	-0.500	-0.692	-0.778	-1.000	-1.000	-1.000	-1.000	-1.000	-1.000	-1.000	-1.000	-1.000
CR	0.867	0.867	0.867	0.867	1.000	0.333	0.067	-0.529	-0.714	-0.789	-1.000	-1.000	-1.000	-1.000	-1.000	-1.000	-1.000	-1.000	-1.000
PN	0.412	0.412	0.412	0.412	0.333	1.000	-0.059	-0.579	-0.750	-0.810	-0.429	-0.789	-1.000	-1.000	-1.000	-1.000	-0.846	-0.818	-1.000
MX	-0.143	-0.143	-0.143	-0.143	0.067	-0.059	1.000	-0.750	-0.385	-1.000	-0.778	-0.500	-1.000	-0.765	-0.667	-1.000	-0.826	-1.000	-1.000
VE	-0.500	-0.500	-0.500	-0.500	-0.529	-0.579	-0.750	1.000	-0.467	0.600	-0.800	-0.778	-0.765	-1.000	-1.000	-1.000	-1.000	-1.000	-1.000
PR	-0.692	-0.692	-0.692	-0.692	-0.714	-0.750	-0.385	-0.467	1.000	-0.765	-0.059	-0.200	-1.000	-1.000	-1.000	-1.000	-1.000	-1.000	-1.000
CO	-0.778	-0.778	-0.778	-0.778	-0.789	-0.810	-1.000	0.600	-0.765	1.000	-0.636	-0.600	-0.158	-0.810	-1.000	-1.000	-1.000	-1.000	-1.000
RD	-1.000	-1.000	-1.000	-1.000	-1.000	-0.429	-0.778	-0.800	-0.059	-0.636	1.000	0.200	-0.579	-0.619	-1.000	-1.000	-1.000	-1.000	-1.000
CU	-1.000	-1.000	-1.000	-1.000	-1.000	-0.789	-0.500	-0.778	-0.200	-0.600	0.200	1.000	-0.294	-0.368	-0.714	-0.750	-0.840	-1.000	-1.000
EC	-1.000	-1.000	-1.000	-1.000	-1.000	-1.000	-1.000	-0.765	-1.000	-0.158	-0.579	-0.294	1.000	0.111	-0.692	-0.467	-0.833	-1.000	-1.000
PE	-1.000	-1.000	-1.000	-1.000	-1.000	-1.000	-0.765	-1.000	-1.000	-0.810	-0.619	-0.368	0.111	1.000	0.333	-0.059	-0.077	-0.636	-0.600
CH	-1.000	-1.000	-1.000	-1.000	-1.000	-1.000	-0.667	-1.000	-1.000	-1.000	-1.000	-0.714	-0.692	0.333	1.000	-0.333	-0.048	-0.529	-0.467
BO	-1.000	-1.000	-1.000	-1.000	-1.000	-1.000	-1.000	-1.000	-1.000	-1.000	-1.000	-0.750	-0.467	-0.059	-0.333	1.000	-0.130	-0.579	-0.765
AR	-1.000	-1.000	-1.000	-1.000	-1.000	-0.846	-0.826	-1.000	-1.000	-1.000	-1.000	-0.840	-0.833	-0.077	-0.048	-0.130	1.000	0.714	0.538
UR	-1.000	-1.000	-1.000	-1.000	-1.000	-0.818	-1.000	-1.000	-1.000	-1.000	-1.000	-1.000	-1.000	-0.636	-0.529	-0.579	0.714	1.000	0.818
PA	-1.000	-1.000	-1.000	-1.000	-1.000	-1.000	-1.000	-1.000	-1.000	-1.000	-1.000	-1.000	-1.000	-0.600	-0.467	-0.765	0.538	0.818	1.000

De esta manera se observan las asociaciones que existen entre los países lingüísticamente cercanos.

5.2.4. Coeficientes de concentración

Utilizamos varios coeficientes de correlación que sirven para apreciar el grado de concentración.

(1) Número de división máxima

La tabla inferior izquierda muestra el estado de la división máxima con el punto divisor (4, 2) d2:v1:

Cor.A	v2	v1	v3
d1	v		
d4	v	v	
d5	v	v	
d2	v	v	v
d3		v	v

Cor.A	v2 v1	v3
d1		
d4	A	B
d5		
d2		
d3	C	D

Al buscar el punto de división que dé mayor número de reacción en la sección superior izquierda y la sección inferior derecha, encontramos el punto (4, 2). Para calcular la cifra indicadora de la división máxima, dividimos en cada punto divisor cuatro partes: A, B, C, D, de manera de la tabla superior derecha. Calculamos el valor positivo (Ps) por la suma de Medias de A y D; el valor negativo por la misma de B y C. El valor de división (Z) se calcula con el valor contrastivo entre Ps y Ng:

$$Z = (Ps - Ng) / (Ps + Ng)$$

En las partes positivas A y D se encuentran 7 + 1 = 8 reacciones en la

superficie de 8 + 1 = 9 puntos, mientras que en las partes negativas, 1 + 1 = 2 puntos de reacción en la superficie de 4 + 2 = 6 puntos. Utilizando estas cifras calculamos los valores de Ps, Ng y finalmente Z:

Ps = (7 + 1) / (8 + 1) = .889, Ng = (1 + 1) / (4 + 2) = .333
Z = (.889 - .333) / (.889 + .333) = .456

En realidad, utilizamos programa para hacer la búsqueda exhaustiva calculando en cada punto el valor Z y determinamos el punto divisor máximo donde dé el máximo valor Z.

La división siguiente también da el máximo valor (.456):

Dst.A	v1	v2	v3
d1	v		
d4	v	v	
d5	v	v	
d2	v	v	v
d3		v	v

Sin embargo, en cuanto al número de puntos positivos en A y D, 3 + 4 = 7 es inferior al caso anterior, 7 + 1 = 8. El alto valor del punto divisor (3, 1) se debe a la ocupación máxima de superficie A y D. Nos interesa más el número de reacciones positivas, de modo que damos peso al valor Z con un número de reacciones posirivas (7) dividido por el número total de superficie multiplicado por 10: 7 / (N * P * 10) = 7 / 150 = .0467, que es leve pero sirve para diferenciar los valores idénticos de división.

(2) Valor de división máxima

Consideramos no solamente los números de reacción en las cuatro partes, A, B, C, D, sino también sus posiciones dentro de la matriz, puesto que consideramos que los puntos distantes del punto divisor llevan un valor más alto que los cercanos. Por ejemplo, para calcular la distancia del punto (1, 1) con respecto al punto divisor (4, 2) utilizamos la distancia euclidiana (E):

$$E(1, 1, 4, 2) = [(1 - 4)^2 + (1 - 2)^2]^{1/2} = (9 + 1)^{1/2} = 3.162$$

Calculamos las mismas distancias de todos los puntos con la división de A, B, C, D, y obtenemos el valor contrastivo Z para encontrar su Máximo. El resultado es el siguiente:

Dst.A	v1	v2	v3
d1	v		
d4	v	v	
d5	v	v	
d2	v	v	v
d3		v	v

Contr.coef.	Valor
MaxDiv.C: 4-2	.600
MaxDiv.V: 4-2	.770

(#) Vocales abiertas en dialecto andaluz

Se observa que en el dialecto andaluz, se pierden consonantes finales de palabra, y en algunas regiones, se abren vocales anteriores. Preparamos la tabla siguiente contando los casos en Manuel Alvar y Antonio Llorente: *Atlas lingüístico y etnográfico de Andalucía*, 1973. El signo (+) indica la abertura vocálica y (++) la abertura grande de la vocal (puntos normalizados por el número de localidades de cada provincia):

R	CA	SE	H	MA	CO	GR	J	AL
1533B:miel:el>e+	9	10	17	15	20	46	29	30
1533C:miel:el>e:	4	6	11	16	12	16	11	3
1615A:caracol:-ól>ó+(:)	3	3	2	5	15	19	14	11
1615B:caracol:-ól>ó(:)	15	27	18	16	3	6	1	2
1616A:árbol:-ol>o+		1	2		6	6	8	6
1616B:árbol:-ol>o	17	30	23	26	18	23	11	11
1618A:sol:-ól>ó+(:)	3	9	7	13	13	19	12	11
1618B:sol:-ól>ó(:)	15	21	15	13	1	6	1	1
1623A:beber:-ér>é+l			2	1	10	19	11	20
1623B:beber:-ér>é+	3	7	4	6	13	15	17	8
1623C:beber:-ér>é	15	24	19	19	2	4		
1626C:tos:o++				2	7	18	10	12
1626C:tos:-ós>ó+	5	7	7	13	18	27	17	19
1626D:tos:-ós>ó	11	22	10	9	2	2	1	
1627B:nuez:-éθ>é+	5	13	7	17	20	39	25	26
1627C:nuez:e++				5	14	26	18	18
1627C:nuez:-éθ>é	12	16	12	9	3	1	1	
1629B:voz:-óθ>ó+	3	5	3	12	22	44	30	30
1629C:voz:-óθ>ó	14	23	18	13	2	2	1	1
1689A:niños:-os>-o+		2	1	4	22	44	31	30
1689B:niños:-os>oh[os)		1	4		2	8	3	8
1690A:pared:-éd>é+		8	6	10	17	24	19	11
1693B:redes:redes>re+	4	6	14	12	3	16	6	6
1694B:clavel:-él>é+,	3	6		15	20	40	24	29

1694C:clavel:-él>ér					5	1	1
1695A:claveles:e-es>-e-e+	2		4	2	4	2	3
1695B:claveles:e-es>-e+-e+	1		7	18	33	24	21
1695C:claveles:-e-es>-e-e:	1		3	1	1	2	1

Veamos el Número y Valor de división máxima, calculada en la tabla concentrada:

Contr.coef.	Valor
MaxDiv.C: 7-4	.581
MaxDiv.V: 7-4	.755

De acuerdo con el punto divisor máximo (7, 4), hemos dividido la tabla de la manera siguiente:

Dst.A	CA	SE	H	MA	CO	GR	J	AL
1626D:tos:-ós>ó	11	22	10	9	2	2	1	
1627C:nuez:-éθ>é	12	16	12	9	3	1	1	
1623C:beber:-ér>é	15	24	19	19	2	4		
1629C:voz:-óθ>ó	14	23	18	13	2	2	1	1
1618B:sol:-ól>ó(:)	15	21	15	13	1	6	1	1
1615B:caracol:-ól>ó(:)	15	27	18	16	3	6	1	2
1616B:árbol:-ol>o	17	30	23	26	18	23	11	11
1533C:miel:el>e:	4	6	11	16	12	16	11	3
1693B:redes:redes>re+	4	6	14	12	3	16	6	6
1618A:sol:-ól>ó+(:)	3	9	7	13	13	19	12	11
1695C:claveles:-e-es>-e-e:		1		3	1	1	2	1
1623B:beber:-ér>é+	3	7	4	6	13	15	17	8
1695A:claveles:e-es>-e-e+		2		4	2	4	2	3
1690A:pared:-éd>é+		8	6	10	17	24	19	11
1626C:tos:-ós>ó+	5	7	7	13	18	27	17	19
1533B:miel:el>e+	9	10	17	15	20	46	29	30
1627B:nuez:-éθ>é+	5	13	7	17	20	39	25	26
1615A:caracol:-ól>ó+(:)	3	3	2	5	15	19	14	11
1694B:clavel:-él>é+,	3	6		15	20	40	24	29
1629B:voz:-óθ>ó+	3	5	3	12	22	44	30	30
1616A:árbol:-ol>o+		1	2		6	6	8	6
1694C:clavel:-él>ér						5	1	1
1695B:claveles:e-es>-e+-e+		1		7	18	33	24	21
1689B:niños:-os>oh[os)		1	4		2	8	3	8
1627C:nuez:e++				5	14	26	18	18

1689A:niños:-os>-o+	2	1	4	22	44	31	30
1626C:tos:o++			2	7	18	10	12
1623A:beber:-ér>é+l		2	1	10	19	11	20

En esta tabla observamos que el fenómeno de la abertura vocálica es propio del andaluz oriental, concretamente en Córdoba (CO), Jaén (J), Granada (GR) y Almería (AL). En la parte occidental, en Cádiz (CA), Sevilla (SE), Huelva (H) y Málaga (MA), se encuentran valores altos en la parte superior izquierda (área A) de la tabla, la que corresponde a los casos de la no abertura vocálica.

* Véase: Manuel Alvar. 1973. *Estructuralismo, geografía lingüística y dialectología actual*, p.203.

(3) Distancia media secuencial

Los datos concentrados merecen la atención aun si no presentan altos valores de correlación. Por ejemplo, en la Concentración por Análisis de Cluster, de que hablaremos más adelante, presentan unas distribuciones concentradas peculiares no en las partes diagonales. En las partes concentradas observamos una fuerte asociación entre filas y columnas.

Para calcular la «Distancia Media Secuencial» buscamos todas las parejas posibles dentro de la matriz y calculamos la diferencia de la coordenada X e Y para medir la distancia entre los dos puntos:

$$SMD = \Sigma_i \Sigma_j \Sigma_a \Sigma_{b(i<>a,\ j<>b)} \{[(i-a)^2 + (j-b)^2 / (N^2 + P^2)]^{1/2} |X_{ij}\ X_{ab}|\} / N$$

donde (i, j) es coordenada de Xij, y (a, b) es coordenada de Xab. N es el número de parejas. De esta manera calculamos la distancia euclidiana entre los dos puntos Xij y Xab. La multiplicamos por el valor de los dos puntos. En datos cualitativos, existe solo 1 ó 0.

Por otra parte, en la matriz de datos cuantitativos, como la siguiente, consideramos la cantidad de cada uno como peso de la distancia. De esta manera prestamos atención no solamente a la distancia, sino también el peso de los datos.

P2	v1	v2	v3	v4
d1	1	1	2	3
d2	2	4	3	4
d3	1	3	2	3
d4	3	3	2	4
d5	2	3	2	4

(4) Distancia media referencial

Para precisar la distancia tomamos en cuenta no simplemente el número secuencial, sino el valor referencial tanto de fila como de columna, de la manera siguiente:

$$RMD = \Sigma_i \Sigma_j \Sigma_a \Sigma_{b(i<>a,\ j<>b)} [(V_i - H_a)^2 + (V_j - H_b)^2 / Mx]^{1/2} |X_{ij}\ X_{ab}| / N$$

donde V y H es vector de los valores de filas y columnas respectivamente. Mx es:

$$Mx = (Vmax - Vmin)^2 + (Hmax - Hmin)^2$$

donde Vmax, Vmin, Hmax, Vmin son Máximo y Mínimo de los elementos del vector V y Máximo y Mínimo de los elementos del vector H, respectivamente.

X	v1	v2	v3	v4
d1	v	v		
d2			v	
d3		v		
d4			v	v
d5	v	v	v	

→

Y	v2	v1	v3	v4
d3	v			
d1	v	v		
d5	v	v	v	
d2			v	
d4			v	v

Cct por dist.	X	Y	Diferencia
DMS	.591	.549	-.042
DMR	.630	.588	-.042

(5) Coeficiente de correlación secuencial

X	v1	v2	v3	v4
d1	v	v		
d2			v	
d3		v		
d4			v	v
d5	v	v	v	

→

Y	v2	v1	v3	v4
d3	v			
d1	v	v		
d5	v	v	v	
d2			v	
d4			v	v

Considerando las tablas como gráficos de distribución con ejes construidos de números secuenciales, calculamos el «Coeficiente de Correlación Secuencial» (CCS) entre los dos ejes, horizontal (X) y vertical (Y):

Datos(X, Y) = (1, 1) (2, 1) (2, 2) (3, 1) (3, 2) (3, 3) (4, 3) (5, 3) (5, 4)
CCS = 0.820

(6) Coeficiente de correlación referencial

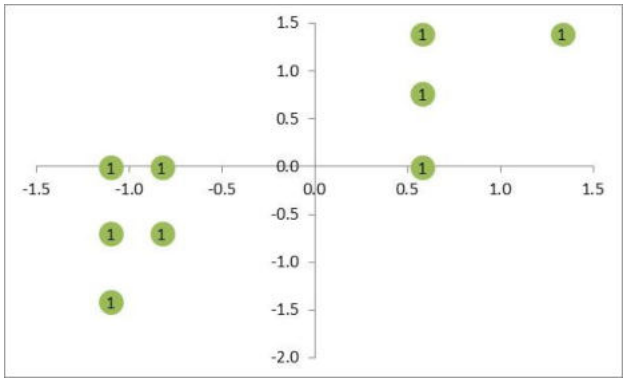
En realidad, cada punto no se sitúa con intervalos secuenciales, sino lleva su propio valor referencial horizontal y vertical:

Y	v2	v1	v3	v4
d3	v			
d1	v	v		
d5	v	v	v	
d2			v	
d4			v	v

Lv	Valor
d3	1.42
d1	0.71
d5	0.01
d2	0.76
d4	1.38

Lv	v2	v1	v3	v4
Valor	1.10	0.82	0.58	1.34

El gráfico de burbuja siguiente muestra las posiciones de los puntos con ejes estandarizados:



El «Coeficiente de correlación referencial» (CCR) se calcula con estos valores de los dos ejes:

Datos(X, Y) = (-1.10, -1.42) (-1.10, -0.71) … (1.34, 1.38)
RCC = 0.35

Coeficiente	Sin cct.	Con cct.	Diferencia
CCS	0.226	0.820	0.594
CCR	0.563	0.835	0.273

(7) Coeficiente de unión

En la tabla inferior izquierda (X), encontramos dos contactos entre d1/v1 y d1/v2 y entre d4/v3 y d5/v3. Encontramos en total 5 contactos, mientras que en la tabla Y, 9 contactos.

X	v1	v2	v3	v4
d1	v	v		
d2			v	
d3		v		
d4			v	v
d5	v	v	v	

→

Y	v2	v1	v3	v4
d3	v			
d1	v	v		
d5	v	v	v	
d2			v	
d4			v	v

La frecuencia de las uniones se calculan no solamente en los datos cualitativos (X, Y), sino también en los cuantitativos, como la tabla inferir izquierda (A), convirtiendo los puntos superiores a la Media (2.6) en «v» (tabla B):

A	v1	v2	v3	v4
d1	1	1	2	3
d2	2	4	3	4
d3	1	3	2	3
d4	3	3	2	4
d5	2	3	2	4

→

B	v1	v2	v3	v4
d1				v
d2		v	v	v
d3		v		v
d4	v	v		v
d5		v		v

Para normalizar las Frecuencias de unión, buscamos el Máximo de uniones (U_{max}). dividiendo la Frecuencia de unión por el Máximo de unión obtenemos el «Coeficiente Normalizado Unión» (CNU). En CNU consideramos solo el número de uniones y no hacemos caso de los valores de puntos.

El Máximo de Unión (U_{max}) se determina por el número de reacciones (R). Por ejemplo, cuando R = 2, Umax es 1. Lo expresamos Umax(2) = 1. Cuando R = 3, en ambos casos siguientes, Umax(3 = 2:

«R=**3;** U_{max} **=2»**

«R=**3;** U_{max} **=2»**

Cuando R = 4, en la figura inferior izquierda, U = 3, y en la inferior derecha, U =4, de modo que Umax(4) = 4.

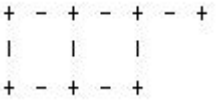
«R=**4;** U_{max} **=3»**

«R=**4;** U_{max}**=4»**

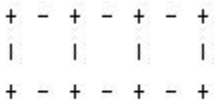
Cuando R = 5, 6, 7, 8, U_{max} se presenta siempre cuando la distribución contiene un cuadrado:

«R=**5;** U_{max} **=5»**

«R=**6;** U_{max} **=7»**

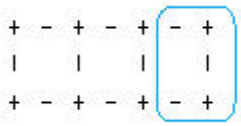


«R=**7;** U_{max} **=8»**

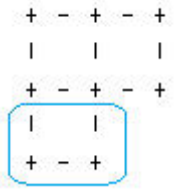


«R=**8;** U_{max}**=10»**

Cuando R = 8, tanto en la figura inferior izquierda como en la derecha, da U_{max} = 10, puesto que ambos casos se tratan de la unión de un rectángulo y dos puntos:



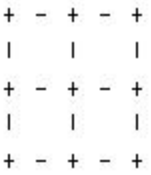
«R=**8;** U_{max}**=10»**



«R=**8;** U_{max}**=10»**

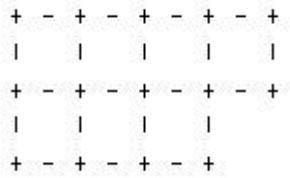
Cuando R = 9, el número de unión varía según el modo de la combinación. U_{max} (= 12) se presenta en la distribución cuadrada:

«R=**9;** U_{max} **=11»**



«R=**9;** U_{max}**=12»**

Veamos el caso de R = 14, donde da U_{max} = 20:

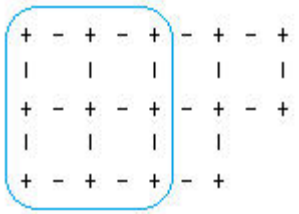


«R=**14;** U_{max}**=20»**

De esta manera, U_{max} se presenta en la combinación de cuadrado, rectángulo y puntos restantes. Para saber el número de uniones en el cuadrado, buscamos la longitud de cuadrado:

$$S = Int(R^{1/2})$$

donde Int(X) es la función que devuelve el número entero de X. En este caso Int(14) = 3. En el cuadrado de 3 x 3, contamos 2 * 3 uniones horizontales y otras 2 * 3 uniones verticales:

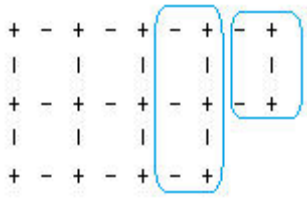


«R=**9;** U_{max}=**12»**

Entonces el número de uniones del cuadrado (w) es:

$W = 2\ S\ (S - 1)$

El resto se divide entre el restángulo y puntos restantes:



R=**5,** U_{max}=**8**

En la parte rectangular, se busca el número de columnas (C) en la fórmula siguiente:

$C = Int((R - S^2) / S)$

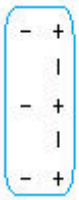
donde R es el número de reacciones y S es longitud del cuadrado que hemos visto anteriormente:

$C = Int((R - S^2) / S) = Int((14 - 3^2) / 3) = 1$

Cada columna contiene S uniones horizontales y S- 1 uniones verticales:

$S + S - 1 = 2\ S - 1$

Por ejemplo, cuando S = 3, se calcula el número de uniones: 2*3-1 = 5.



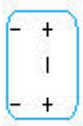
«R=**3,** U_{max}=**5»**

Por lo tanto, el número de uniones del rectángulo (T) es:

$T = C\ (2\ S - 1) = \quad Int[(R - S^2) / S]\ (2S - 1)$

Por último calculamos el número de uniones de los puntos restantes (Q):

14 – 9 – 3 = 2.



«R=**2,** U_{max} **=3»**

Utilizamos el operador Mod, que indica el resto de la división en número entero, por ejemplo 14 Mod 3 = 2 (14 / 3 = 4..2). El número de uniones de puntos restantes es:

$$Q = (R \text{ Mod } S) + (R \text{ Mod } S) - 1 = 2 (R \text{ Mod } S) - 1$$

Por lo tanto, Umax(R) es:

$$Umax(R) = 2 S (S - 1) + Int[(R - S^2) / S] (2S - 1) + Q$$

donde $S = Int(R^2)$, R es número de reacciones, Q es 2 (R Mod S) – 1, cuando R Mod S > 0.

Las tablas siguientes muestran las distribuciones de puntos tanto anterior como posterior a la concentración. La última tabla muestras los coeficientes de unión de las dos tablas anteriores:

X	v1	v2	v3	v4
d1	v	v		
d2			v	
d3		v		
d4			v	v
d5	v	v	v	

Y	v2	v1	v3	v4
d3	v			
d1	v	v		
d5	v	v	v	
d2			v	
d4			v	v

Contr.coef.	X	Y
Union Coef.	.417	.750

(8) Medida de asociación en orden

Cuando los datos y sus variables están ordenados cuantitativa o cualitativamente, se calcula la «Medida de Asociación en Orden de Goodman y Kruskal» (GK) con el valor contrastivo entre el número de los datos positivos y el de los negativos. Utilizamos GK como uno de los coeficientes de concentración, puesto que nuestros datos concentrados están en condición necesaria para calcular GK. Las tablas cruzadas obtenidas en las encuestas también pueden ser sometidas al cálculo de GK sin cambiar el orden de los dos ejes. El ejemplo concreto de la utilización de GK lo veremos seguidamente.

(#) «Tú» y «usted»

La tabla siguiente muestra muestran las frecuencias de uso de «tú» y

«usted» según se dirige al niño, joven, persona mayor y anciano, en distintas situaciones de comunicación interpersonal con variables de «poder» y «solidaridad».

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

G. & K.	¿Vas tú?
V. Positivo	10600
V. Negativo	101941
G. & K.	- .812

Tanto en las variables verticales, como en las horizontales existe el orden conceptual de menor a mayor, de modo que no podemos cambiar el orden ni de filas ni de columnas. Para ver en qué grado los datos respetan estos órdenes, contamos los «casos positivos» (Ps) en forma de la suma de multiplicaciones de casillas con las casillas situadas en posiciones relativas inferiores derechas.

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

Por otra parte, nos interesan también los «casos negativos» (Ng) en forma de la suma de multiplicaciones de casillas con las casillas situadas en la posiciones relativas inferiores, ahora, izquierdas:

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

La «Medida de Asociación en Orden de Goodman y Kruskal» (GK) es valor contrastivo entre Ps y Ng:

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

La GK presenta una alta cifra negativa, lo que indica que entre la Edad del interlocutor y el Uso de «tú» existe un asociación negativa fuerte.

Ahora veamos lo mismo con otra expresión: ¿Adónde va usted?

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

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

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

De esta manera, constatamos cuantitativamente la asociación positiva fuerte entre el uso de «usted» con la Edad del interlocutor. En realidad los mismos cálculos son aplicables con otros parámetros sociolingüísticos y llegamos a la conclusión de que en el español actual la variable indicadora de solidaridad pesa más que la de poder.

*Para «Medida de Asociación en Orden de Goodman y Kruskal», véase Ikeda y Shiba (1976: 130-132).

(9) Índice de asociación predicativa

Como indicador del grado de asociación entre las columnas y las filas, se utiliza el «Índice de Asociación Predicativa» (IAP) de Goodman y Kruskal. Primero explicamos el método de calcular IPA de modo vertical (IAPv).

G.K.Pred.	y1	y2	y3
x1	8	3	1
x2	1	2	7
x3	3	2	9
x4	2	8	1
Sv	14	15	18

El valor máximo del vector horizontal de sumas (Sv) es 18 (y3). Por consiguiente al tratar de predecir la selección entre y1, y2, y3, utilzando solo Sv, el error suma a 14 + 15 = 29 (P):

$$P = 14 + 15 = 29$$

Por otra parte, utilizando la información que proporcionan x1 ~ x4 para predecir la selección de y1, y2, y3, el error suma a Q:

$$Q = (3 + 1) + (1 + 2) + (3 + 2) + (2 + 1) = 15$$

Se define el Índice de asociación predicativa vertical (IPAv) como siguiente:

$$IAPv = (P - Q) / P$$

De esta manera se ha intentado calcular el valor relativo de la disminución de los errores (P – Q) a base de P. Cuando Q = 0, IAPv llega a cobrar el máximo valor 1.

$$IPAv = (29 - 15) / 29 \fallingdotseq .483$$

Es conveniente utilizar los valores máximos tanto de la suma vertical (SvM) como de cada columna RM*i* para simplificar y facilitar el cálculo:

$P = S - SvM$ ←Suma total (S) – Máximo de la suma vertical (SvM)
$Q = S - \Sigma_i RM_i$
←Suma total (S) – Suma de Máximos de filas (RM)

$IAPv = (P - Q) / P$
$= [(S - SvM) - (S - \Sigma i\ RMi)] / (S - SvM)$
$= (\Sigma i\ RMi - SvM) / (S - SvM)$

Lo calculamos con los datos de la tabla:

$IAPv = [(8 + 7 + 9 + 8) - 18] / (47 - 18) = 14 / 29 \fallingdotseq .483$

Seguidamente calculemos el «Índice de Asociación Predicativa horizontal» (IAPh):

G.K.Pred.	y1	y2	y3	Sh
x1	8	3	1	12
x2	1	2	7	10
x3	3	2	9	14
x4	2	8	1	11

De la misma manera que IAPv:

$P = S - ShM$ ←Suma total (S) – Máximo de la suma vertical (ShM)
$Q = S - \Sigma_i CM_i$
←Suma total (S) – Suma de Máximos de columnas (CM)

$IPAv = (P - Q) / P$
$= [(S - ShM) - (S - \Sigma i\ CMi)] / (S - ShM)$
$= (\Sigma_i CM_i - ShM) / (S - ShM)$

Lo calculamos con los datos de la tabla:

$IPAh = [(8 + 8 + 9) - 14] / (47 - 14) = 11 / 33 \fallingdotseq .333$

Para calcular el «Índice de Asociación Predicativa total» (IAP), se utiliza la media fraccional de IAPv e IAPh:

$IPA = [(\Sigma_i RM_i - SvM) + (\Sigma_i CM_i - ShM)]$
$/ [(S - SvM) + (S - ShM)]$
$= (\Sigma i\ RMi + \Sigma i\ CMi - SvM - ShM) / (2S - SvM - ShM)$

Lo calculamos con los datos de la tabla:

IPA = (14 + 11) / (29 + 33) = 25 / 62 ≒ .403

El «Índice de Asociación Predicativa» es constante entre la matriz original y la matriz concentrada correspondiente.

* Para el «Índice de Asociación Predicativa», hemos consultado a Ikeda (1976: 127-130).

(10) Medida de asociación de Cramer

Se utiliza la «Medida de Asociación de Cramer» (MAC) como indicadora del grado de asociación que existe entre las columnas y las filas. Veamos un ejemplo (Xnp) con su correspondientes puntos esperados (Enp):

Enp = (Sh * Sv) / S

donde Sh, Sv, S son Sumas horizontales, Sumas verticales y Suma total de la matriz (Xnp).

Xnp	v1	v2	v3	Sh
d1	45	48	66	159
d2	56	59	54	169
d3	58	51	78	187
d4	77	72	20	169
Sv	236	230	218	S: 684

Enp	v1	v2	v3
d1	54.860	53.465	50.675
d2	58.310	56.827	53.863
d3	64.520	62.880	59.599
d4	58.310	56.827	53.863

Los valores de la matriz siguiente (Cnp) son:

$Cnp = (Xnp - Enp)^2 / Enp$

Cnp	v1	v2	v3
d1	1.772	0.559	4.634
d2	0.092	0.083	0.000
d3	0.659	2.245	5.681
d4	5.991	4.051	21.289

CMA 0.185

Se define chi cuadrado (χ^2) como conjunto de la matriz Cnp:

$$\chi^2 = \Sigma_i \Sigma_j [(X_{ij} - E_{ij})^2 / E_{ij}]$$

donde X es una casilla de Xnp y E es su valor esperado correspondiente.

La «Medida de Asociación de Cramer» (MAC) se define con la fórmula

siguiente:

$$CMA = \{ \chi^2 / \chi^2_{max} \}^{1/2}$$

donde χ^2_{max} es el Máximo de χ^2, que se busca de la manera siguiente. Primero calculamos el valor esperado (E) con la Suma horizotal (Sh), Suma vertical (Sv) y la Suma total (S):

$$E_{ij} = Sh_i \, Sv_j / S$$

Entonces,、

$$\begin{aligned}
\chi^2 &= \Sigma_i \Sigma_j [(X_{ij} - Sh_i \, Sv_j / S)^2 / (Sh_i \, Sv_j / S)] \\
&= \Sigma_i \Sigma_j [(S \, X_{ij} - Sh_i \, Sv_j) / S]^2 / (Sh_i \, Sv_j / S) \\
&= \Sigma_i \Sigma_j \{[(S^2 \, X_{ij}^2 - 2 \, S \, X_{ij} \, Sh_i \, Sv_j + Sh_i^2 \, Sv_j^2) / S^2] \, (S / Sh_i \, Sv_j)\} \\
&= \Sigma_i \Sigma_j [(S^2 \, X_{ij}^2 / Sh_i \, Sv_j - 2 \, S \, X_{ij} + Sh_i \, Sv_j) / S] \\
&= \Sigma_i \Sigma_j (S \, X_{ij}^2 / Sh_i \, Sv_j) - 2 \Sigma_i \Sigma_j \, X_{ij} + \Sigma_i \Sigma_j (Sh_i \, Sv_j / S)
\end{aligned}$$

donde una parte del segundo término $\Sigma_i \Sigma_j \, X_{ij}$ es igual a la Suma total (S) y una parte del tercer término $\Sigma_i \Sigma_j \, Sh_i \, Sv_j$ es S^2.
Por lo tanto,

$$\begin{aligned}
\chi^2 &= S \; \Sigma_i \Sigma_j (X_{ij}^2 / Sh_i \, Sv_j) - 2S + S \\
[1] \qquad &= S \, [\Sigma_i \Sigma_j (X_{ij}^2 / Sh_i \, Sv_j) - 1]
\end{aligned}$$

Ahora bien, el chi cuadrado llega a su Máximo cuando se presenta la distribución única con la Suma horizontal y la vertical coincidentes del valor de la casilla en cuestión. En tal caso la Medida de asociación se supone que llega a su Máximo, puesto que entre la selección de la fila y la de la columna, no existe más que una sola conexión:

Xmax	v1	v2	...		vP	Sh
d1	X_1	0	0	0	0	X_1
d2	0	X_2	0	0	0	X_2
:	0	0	...	0	0	...
dN	0	0	0	X_M	0	X_M
Sv	X_1	X_2	...	X_M	0	S

donde M corresponde al Mínimo entre el número de filas N y el número de columnas P:

[2] $M = \min(N, P)$

De esta manera, en la parte cuadrada se considera solo los casos de la

selección exclusiva. En esta situación el chi cuadrado llega a su Máximo:

$$\chi^2_{max} = S\,[\,\Sigma_i \Sigma_j\,(X_{ij}^2 / Sh_i\,Sv_j) - 1] \quad \leftarrow[1]$$

$$= S\,(X_1^2 / X_1X_1 + X_2^2 / X_2X_2 + \ldots + X_M^2 / X_M\,X_M - 1) \quad \leftarrow\text{tabla}$$

$$= S\,(M - 1) \quad \leftarrow\text{todos los términos son 1, que ocurre M veces}$$

$$= S\,[\min(N, P) - 1] \quad \leftarrow[2]$$

Por lo tanto, la «Medida de Asociación de Cramer» (MAC) es:

$$CMA = \{\chi^2 / \chi^2_{max}\}^{1/2} = \{\chi^2 / [S\,(\min(N, P) - 1)]^{1/2}$$

Se abre la raíz cuadrada, puesto que en su cálculo se ha utilizado cifras cuadradas.

El coeficiente de Cramer no varía si cambiamos el orden de filas y/o el de columnas, de modo que tanto anterior como posterior a la concentración, se mantiene invariable. Lo utilizamos, por lo tanto, para ver el grado de asociación constante entre filas y columnas de la matriz:

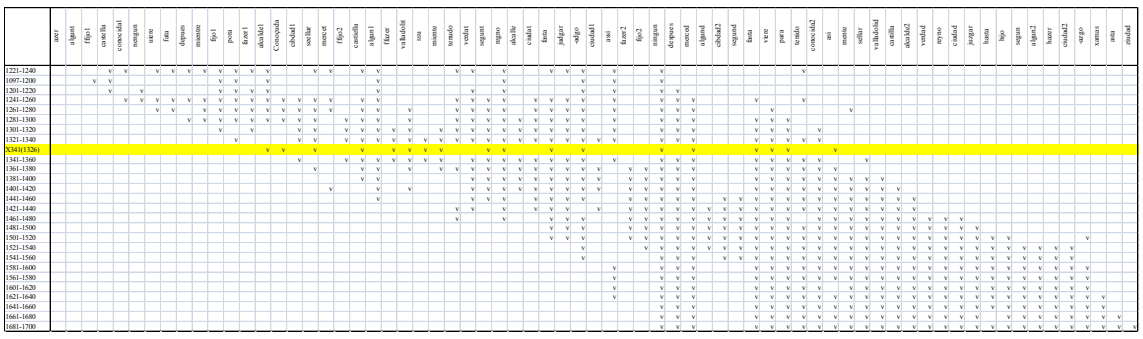
X	v1	v2	v3	v4
d1	v	v		
d2			v	
d3		v		
d4			v	v
d5	v	v	v	

Y	v1	v2	v3	v4
d1	v	v		
d2			v	
d3		v		
d4			v	v
d5	v	v	v	

Contr.coef.	X	Y
MaxDiv.C: 1-2	.333	.778
MaxDiv.V: 1-2	.195	.896
Sq.Dist	.591	.549
Rf..Dist	.630	.588
Sq.Correl.	.226	.820
Rf.Correl.	.563	.835
Union Coef.	.417	.750
G & K Rank	.167	.917
Cramer	.616	.616

(#) Datación de los documentos antiguos

La tabla siguiente procede del estudio de Kawasaki (2012), que intenta datar los documentos no fechados a partir de los rasgos lingüísticos observados en los documentos fechados.



En la figura muestra la situación que ocupa el documento no fechado dentro de la distribución concentrada de puntos con coordenadas de años (vertical) y rasgos (horizontal). Para llevar a cabo una datación fiable hay que repetir multitud de experimentos con rasgos apropiados de alta frecuencia y de baja variabilidad geográfica.

5.3. Análisis multivariante

Desde esta sección en adelante empezamos a utilizar operaciones matriciales relativamente más avanzadas. Ahora que estamos acostumbrados a las operaciones fundamentales, ya no vamos a denominar los vectores como E_{n1}, sino simplemente E_n y se supone que los vectores son siempre verticales, a menos que están traspuestas: E_n^T. Para los elementos de la matriz y del vector, utilizaremos minúsculas.

5.3.1. Regresión múltiple

El método llamado «Regresión Múltiple» se utiliza para buscar una fórmula de regresión con múltiples variables Xnp y una variable exterior (Y_n). Se formula la ecuación con un vector de peso (W_p), con el que se intenta calcular de nuevo la variable exterior En. Se evalúa mejor cuando se presentan cuanto menos errores entre el variable exterior (Y_n) y el vector producto de la ecuación (E_n). Por esta razón se busca el peso W_p que dé el Mínimo de la Suma cuadrada entre Y_n y E_n. El peso W_p sirve para observar el grado de influencia de las variables y también para estimar la variable exterior incógnita.

Veamos por ejemplo una tabla de puntuaciones de 5 personas en 3 materias. La variable exterior Point es el resultado general de una asignatura. Nuestro objetivo es buscar una ecuación con las tres variables que da la variable exterior con un error mínimo posible.

X_{np}	v1	v2	v3	POINT
d1	6	8	5	12
d2	7	10	6	11
d3	8	4	8	13
d4	9	7	2	7
d5	10	9	4	14

Supóngase una fórmula que da a cada dato (Xnp) distintos pesos (Wp) y un dato constante W80) en forma de intercepto [i = 1, 2, …, N]:

[1] $E(i) = W(0) + W(1)\ X(i, 1) + W(2)\ X(i, 2) + \ldots + W(_p)\ X(i, _p)$

El intercepto constante se agrega a todos los individuos. De modo que se supone que cada individuo posee el valor 1 para recibir el peso W(0). Para la operación matricial preparamos un vector de identidad Ip, fuera de la matriz de datos Xnp:

$$E(i) = I_p\ W(0) + X(i,\ 1)\ W(1) + X(i,\ 2)\ W(2) + \ldots + X(i,\ _p)\ W(_p) \qquad [i = 1\ldots n]$$

La fórmula matricial es:

$$E_n = X_{np}\ W_p$$

donde la primera columna de Xnp es un vector de identidad.

Suponemos un vector de errores (o residuales) que es la diferencia entre el vector de la variable exterior preexistente y el vector que produce la fórmula anterior:

[2] $$R_n = Y_n - E_n = Y_n - X_{np}\ W_p$$

Ahora calculamos la Suma cuadrada del vector de errores en forma de S:

$$\begin{aligned} S &= R_n{}^T\ R_n = (Y_n - X_{np}\ W_p)^T\ (Y_n - X_{np}\ W_p) \qquad \leftarrow [2] \\ &= [Y_n{}^T - (X_{np}\ W_p)^T]\ (Y_n - X_{np}\ W_p) \quad \leftarrow \text{転置行列の性質(T)} \\ &= Y_n{}^T\ Y_n - Y_n{}^T\ X_{np}\ W_p - (X_{np}\ W_p)^T\ Y_n + (X_{np}\ W_p)^T\ X_{np}\ W_p \quad \leftarrow \text{展開} \\ &= Y_n{}^T\ Y_n - Y_n{}^T\ X_{np}\ W_p - Y_n{}^T\ (X_{np}\ W_p) + W_p{}^T\ X_{np}{}^T\ X_{np}\ W_p \quad \leftarrow T \\ &= Y_n{}^T\ Y_n - 2\ Y_n{}^T\ X_{np}\ W_p + W_p{}^T\ X_{np}{}^T\ X_{np}\ W_p \end{aligned}$$

En esta fórmula el vector Wp es incógnito. El objetivo del método de «Regresión Múltiple» es buscar Wp cuando la fórmula de S diferenciada por Wp da el valor cero en forma de vector horizontal de cero: $O_p{}^T$. Imagínense que vamos a buscar unos coordenados multidimensionales donde la curva da el Mínimo, sin inclinación).

El primer término $Y_n{}^T\ Y_n$ no lleva Wp, y al diferenciarlo por Wp, resulta 0. La diferenciación del segundo y tercer término la explicaremos más adelante. Utilizamos el signo de la direrencial $\frac{\partial S}{\partial a}$ en forma de Df(S, w), que significa la diferencial de S por w:

$$\begin{aligned} &Df(S,\ W_p) = -\ 2\ Y_n{}^T\ X_{np} + 2\ X_{np}{}^T\ X_{np}\ W_p = O_p{}^T \\ &2\ X_{np}{}^T\ X_{np}\ W_p = O_p{}^T +\ \ 2\ Y_n{}^T\ X_{np} \end{aligned}$$

[3] $$X_{np}{}^T\ X_{np}\ W_p = Y_n{}^T\ X_{np}$$

Para obtener Wp, premultiplicamos ambos lados por $(X_{np}{}^T\ X_{np})^{-1}$:

$$(X_{np}{}^T\ X_{np})^{-1}\ (X_{np}{}^T\ X_{np})\ W_p = (X_{np}{}^T\ X_{np})^{-1}\ Y_n{}^T\ X_{np} \quad \leftarrow [3]$$

$I_{pp}\ W_p = (X_{np}{}^T\ X_{np})^{-1}\ Y_n{}^T\ X_{np}$ ←$A\ A^{-1} = I_{pp}$

$W_p = (X_{np}{}^T\ X_{np})^{-1}\ Y_n{}^T\ X_{np}$ ←$I_{pp}\ A = A$

$W_p = (X_{np}{}^T\ X_{np})^{-1}\ X_{np}{}^T\ Y_n$ ←Propiedad de matriz traspuesta

De esta manera obtenemos el vector horizontal Wp que contiene cuatro elementos: intercepto, v1, v2, v3:

Weight	P: Intercept	v1	v2	v3	Std res.
Value	-3.819	.740	.462	1.157	1.545

El vector de valores estimados (E_n) se obtiene por la fórmula [1]. El vector de errores (residuales: R_n) y la Suma cuadrada de R_n dividida por N es el Residual estandarizado:

$R_n = Y_n - E_n$

$Std.R. = (R_n{}^T\ R_n / N)^{1/2}$

X	POINT	Expected	Residual
d1	12	10.104	1.896
d2	11	12.926	-1.926
d3	13	13.207	-.207
d4	7	8.392	-1.392
d5	14	12.371	1.629

(*) Matriz inversa

(1) Definición de la matriz inversa

Si las dos matrices cuadradas están en la relación siguiente, una de las dos es la Matriz inversa de la otra:

$X_{pp}\ Y_{pp} = I_{pp}$ (Matriz de identidad) → $Y_{pp} = X_{pp}{}^{-1}$

La Matriz inversa de X_{pp} se simboliza por $X_{pp}{}^{-1}$, y posee las propiedades siguientes, que se utilizan en los cálculos estadísticos con frecuencia:

(a) $X_{pp}\ X_{pp}{}^{-1} = I_{pp}$

X_{pp}	1	2
1	7	8
2	9	10

×

$X_{pp}{}^{-1}$	1	2
1	-5	4
2	4.5	-3.5

=

I	1	2
1	1	0
2	0	1

(b) $X_{pp}^{-1} X_{pp} = I_{pp}$

X_{pp}^{-1}	1	2
1	-5.0	4.0
2	4.5	-3.5

×

X_{pp}	1	2
1	7	8
2	9	10

=

I_{pp}	1	2
1	1	0
2	0	1

(2) Propiedades de la matriz inversa

(a) $(X_{pp}^{-1})^{-1} = X_{pp}$

X_{pp}	1	2
1	7	8
2	9	10

→

X_{pp}^{-1}	1	2
1	-5.0	4.0
2	4.5	-3.5

→

$(X_{pp}^{-1})^{-1}$	1	2
1	7	8
2	9	10

(b) $(X_{pp}\ Y_{pp})^{-1} = Y_{pp}^{-1} X_{pp}^{-1}$

X_{pp}	1	2
1	1	3
2	2	4

×

Y_{pp}	1	2
1	7	8
2	9	1

=

$X_{pp}Y_{pp}$	1	2
1	34	11
2	50	20

→

$(X_{pp}Y_{pp})^{-1}$	1	2
1	0.154	-0.085
2	-0.385	0.262

Y_{pp}^{-1}	1	2
1	-2.00	1.500
2	1.00	-0.500

×

X_{pp}^{-1}	1	2
1	-0.015	0.123
2	0.136	-0.108

=

$Y_{pp}^{-1} X_{pp}^{-1}$	1	2
1	0.154	-0.085
2	-0.385	0.262

(c) $(X_{pp}^{T})^{-1} = (X_{pp}^{-1})^{T}$

X_{pp}	1	2
1	7	8
2	9	10

→

X_{pp}^{T}	1	2
1	7	9
2	8	10

→

$(X_{pp}^{T})^{-1}$	1	2
1	-5.0	4.5
2	4.0	-3.5

X_{pp}	1	2
1	7	8
2	9	10

→

X_{pp}^{-1}	1	2
1	-5.0	4.0
2	4.5	-3.5

→

$(X_{pp}^{-1})^{T}$	1	2
1	-5.0	4.5
2	4.0	-3.5

(3) Método de buscar la matriz inversa

El método de buscar la matriz inversa consiste en ir tranformando tanto la matriz objeto X_{pp}, como la matriz de Identidad I_{pp} con las mismas matrices transformativas T_{pp} (de matrices trasformativas hablaremos más adelante). Premultiplicando las dos matrices X_{pp} e I_{pp} por T_{pp} hasta que X_{pp} se convierta en la matriz de identidad, la I_{pp} inicial llega a ser la matriz inversa de X_{pp}. Este procesamiento se llama el «Método de Gauss-Jordan».

En los procesos utilizamos dos tipos de matriz transformativa:

(a) Matriz que cambia el orden de las filas
(b) Matriz que multiplica una fila por un número real y le agrega un múltiplo de otra fila.

Vamos a ver como la matriz de identidal inicial se convierte en la matriz inversa de X_{pp} en los procesos siguientes:

0. $X^{(0)}, Z^{(0)} = I$ ←Estado inicial de las matrices X y Z
1. $X^{(1)} = T^{(1)} X^{(0)}, Z^{(1)} = T^{(1)}I$ ←premultiplicar $T^{(1)}$ a $X^{(0)}$ y $Z^{(0)}$=I
2. $X^{(2)} = T^{(2)} T^{(1)} X^{(0)}, Z^{(2)} = T^{(2)} T^{(1)}I$ ←premultiplicar$T^{(2)}$ de nuevo
   (…) ←premultiplicar $T^{(3)}$, …,$T^{(k)}$ sucesivamente
3. $I = T^{(k)} \ldots T^{(2)} T^{(1)} X^{(0)}$ ←$X^{(0)}$ llega a ser I por multitud de T
4. $Z^{(k)} = T^{(k)} \ldots T^{(2)} T^{(1)} I$ ←$Z^{(0)}$ llega a ser $Z^{(k)}$ por multitud de T
5. $I X^{(0)-1} = T^{(k)} \ldots T^{(2)} T^{(1)} X^{(0)} X^{(0)-1}$ ←postmultiplicar $X^{(0)-1}$ a 3
6. $X^{(0)-1} = T^{(k)} \ldots T^{(2)} T^{(1)} I$ ←5. I A = A; A A^{-1} =I
7. $Z^{(k)} = X^{(0)-1}$ ←fórmula derecha (f.d.) de 4= f.d. de 6

De esta manera la $Z^{(k)}$ llega a ser matriz inversa de $X^{(0)}$.

Vamos a probar estas operaciones con un ejemlo de X(0) y preparamos una matriz de identidad Z(0). Nuestro objetivo es convertir X(0) en la matriz de identidad por varias premultiplicaciones por T_{pp} que conviertan la matriz Z_{pp} en la matriz de identidad.

X(0)	1	2	3
1	**0**	2	1
2	2	1	2
3	2	1	1

Z(0)	1	2	3
1	1	0	0
2	0	1	0
3	0	0	1

Primero para convertir X(1,1) en 1, intentamos realizar la operación siguiente:

F1 ← F1 / X(1, 1)

Esta fórmula significa que se divide la primera fila F1 por X(1, 1) y se renueva la fila F1. Sin embargo, esta vez, como X(1, 1) es 0, la división por X(1, 1) no es realizable. En tal caso, se hace el cambio de orden de filas:

F1 ← F2, F2 ← F1

$X^{(1)}$	1	2	3
1	**2**	**1**	**2**
2	**0**	**2**	**1**
3	2	1	1

$Z^{(1)}$	1	2	3
2	**0**	**1**	**0**
1	**1**	**0**	**0**
3	0	0	1

Ahora sí es posible la división:

F1 ← F1 / X(1, 1) ← F1 / 2

$X^{(2)}$	1	2	3
1	**2/2=1**	**1/2**	**2/2=1**
2	0	2	1
3	2	1	1

$Z^{(2)}$	1	2	3
1	**0/2=0**	**1/2**	**0/2=0**
2	1	0	0
3	0	0	1

Seguidamente utilizando la nueva F1 convertimos F2 y F3 para que las casillas de la columna 1 (C1) se conviertan en 0. Esta vez, convertimos F3 por la operación siguiente:

F3 ← F3 – X(3, 1) F1 ← F3 – 2 F1

$X^{(3)}$	1	2	3
1	1	1/2	1
2	0	2	1
3	**2-2*1=0**	**1-2*(1/2)=0**	**1-2*1=-1**

$Z^{(3)}$	1	2	3
1	0	1/2	0
2	1	0	0
3	**0-2*0=0**	**0-2*1/2=-1**	**1-2*0=1**

Hemos terminado la formación de la columna 1 (C1). Ahora hacemos lo mismo en la columna 2:

$X^{(4)}$	1	2	3
1	1	1/2	1
2	0	**2**	1
3	0	0	-1

$Z^{(4)}$	1	2	3
1	0	1/2	0
2	1	0	0
3	0	-1	1

Esta vez, X(2, 2) = 2 no es 0, de modo que dividimos F2 por X(2, 2) directamente.

F2 ← F2 / X(2, 2) ← F2 / 2

$X^{(5)}$	1	2	3
1	1	1/2	1
2	**0/2=0**	**2/2=1**	**1/2**
3	0	0	-1

$Z^{(5)}$	1	2	3
1	0	1/2	0
2	**1/2**	**0/2**	**0/2**
3	0	-1	1

Convertimos filas F1 y F3 por las operaciones siguientes:

F1 ← F1 – X(1, 2) F2 ← F1 – 1/2 F2

F3 ← F3 – X(3, 2) F2 ← F3 – 0 F2

$X^{(6)}$	1	2	3
1	**1-(1/2)*0 =1**	**1/2-(1/2)*1 =0**	**1-(1/2)*(1/2)=3/4**
2	0	1	1/2
3	**0-0*0=0**	**0-0*1=0**	**-1-0*(1/2)=-1**

$Z^{(6)}$	1	2	3
1	**0-(1/2)*(1/2) =1/4**	**1/2-(1/2)*0 =1/2**	**0-(1/2)*0 =0**
2	1/2	0	0
3	**0-0*(1/2)=0**	**-1-0*0=-1**	**1-0*0=1**

Terminamos la columna 2 y ahora haremos lo mismo en la columna 3.

$X^{(7)}$	1	2	3
1	1	0	3/4
2	0	1	1/2
3	0	0	**-1**

$Z^{(7)}$	1	2	3
1	1/4	1/2	0
2	1/2	0	0
3	0	-1	1

F3 ← F3 / X(3, 3) ← F3 / -1

$X^{(8)}$	1	2	3
1	1	0	3/4
2	0	1	1/2
3	**0/-1=0**	**0/-1=0**	**-1/-1=1**

$Z^{(8)}$	1	2	3
1	1/4	1/2	0
2	1/2	0	0
3	**0/-1=0**	**-1/-1=1**	**1/-1=-1**

F1 ← F1 – X(1, 3) F3 ← F1 – 3/4 F3

F2 ← F1 – X(2, 3) F3 ← F1 – 1/2 F3

$X^{(9)}$	1	2	3
1	1-(3/4)x0 =1	0-(3/4)x0 =0	3/4-(3/4)x1 =0
2	0-(1/2)x0 =0	1-(1/2)x0 =1	1/2-(1/2)x1 =0
3	0	0	1

$Z^{(9)}$	1	2	3
1	1/4-(3/4)x0 =-1/4	1/2-(3/4)x1 =-1/4	0-(3/4)— 1 =3/4
2	1/2-(1/2)x9 =1/2	0-(1/2)x1 =-1/2	0-(1/2)x-1 =1/2
3	0	1	-1

Por estas operaciones la matriz X llega a ser matriz identidad y ahora la matriz Z es la inversa de X inicial.

$X^{(k)}$	1	2	3
1	1	0	0
2	0	1	0
3	0	0	**1**

$Z^{(k)}$	1	2	3
1	-1/4	-1/4	3/4
2	1/2	-1/2	1/2
3	0	1	-1

Por programa se realiza estas operaciones y efectivamente obtenemos la matriz inversa cuya propiedad se comprueba por su multiplicación matricial:

X	1	2	3
1	0	2	1
2	2	1	2
3	2	1	1

X

X^{-1}	1	2	3
1	-.250	-.250	.750
2	.500	-.500	.500
3	.000	1.000	-1.000

=

$X X^{-1}$	1	2	3
1	1	0	0
2	0	1	0
3	0	0	1

Hemos consultado a Hasegawa (2000:129-136). Para elaborar el programa, véase Nawata (1999:58-80).

(4) Demostración de operaciones con matriz inversa

Se utilizan las operaciones siguientes. Para comprender sus significados haremos demostraciones matemáticas.

[1] $I^{-1} = I$

$I\ I^{-1} = I$ ←def. de matriz inversa: $X\ X^{-1} = I$, donde $X = I$
$I^{-1} = I$ ←$I\ X = X$, $X=I$

[2] $(A^{-1})^{-1} = A$

$A^{-1}\ (A^{-1})^{-1} = I$ ←$X\ X^{-1} = I$
$A\ A^{-1}\ (A^{-1})^{-1} = A\ I$ ←premultiplicar A
$I\ (A^{-1})^{-1} = A\ I$ ←$X\ X^{-1} = I$
$(A^{-1})^{-1} = A$ ←$X\ I = X$; $I\ X = X$

[3] $(A\ B)^{-1} = B^{-1}\ A^{-1}$

$(A\ B)\ (A\ B)^{-1} = I$ ←$X\ X^{-1} = I$, $X = A\ B$
$(A\ B)\ (A\ B)^{-1} = A\ A^{-1}$ ←$A\ A^{-1} = I$
$(A\ B)\ (A\ B)^{-1} = A\ I\ A^{-1}$ ←$A = A\ I$
$(A\ B)\ (A\ B)^{-1} = A\ B\ B^{-1}\ A^{-1}$ ←$I = B\ B^{-1}$
$(A\ B)^{-1} = B^{-1}\ A^{-1}$ ←quitar A B de los dos lados

[4] $A\ A^{-1} = A^{-1}\ A$

$A\ A^{-1} = I$ ←$A\ A^{-1} = I$
$(A^{-1}\ A)\ (A\ A^{-1}) = (A^{-1}\ A)\ I$ ←premultiplicar$A^{-1}\ A$
$A^{-1}\ A\ A\ A^{-1} = A^{-1}\ A$ ←$X\ I = X$, $X=A^{-1}\ A$
$I\ A\ A^{-1} = A^{-1}\ A$ ←$X\ I = X$, $X=A^{-1}\ A$
$A\ A^{-1} = A^{-1}\ A$ ←$I\ A = A$

∗ Para [2] y [3] hemos consultado a Adachi (2005:110-111).

(*) Matriz transformativa

Premultiplicando una matriz identidad modificada, que denominamos «Matriz transformativa», podemos realizar transformaciones de filas. Utilizamos estas matrices transformativas en la operaciones que derivan la matriz inversa.

(a) R1 ← 0

T_{pp}	1	2	3
1	0	0	0
2	0	1	0
3	0	0	1

×

A_{pp}	1	2	3
1	1	2	3
2	4	5	6
3	7	8	9

←

R_{pp}	1	2	3
1	0	0	0
2	4	5	6
3	7	8	9

(b) R1 ← R2

T_{pp}	1	2	3
1	0	1	0
2	0	1	0
3	0	0	1

×

A_{pp}	1	2	3
1	1	2	3
2	4	5	6
3	7	8	9

←

R_{pp}	1	2	3
1	4	5	6
2	4	5	6
3	7	8	9

(c) R1 ~ R2 (cambio de filas)

T_{pp}	1	2	3
1	0	1	0
2	1	0	0
3	0	0	1

×

A_{pp}	1	2	3
1	1	2	3
2	4	5	6
3	7	8	9

←

R_{pp}	1	2	3
1	4	5	6
2	1	2	3
3	7	8	9

(d) R1 ← 3 R1 (m/ultiplo de fila)

T_{pp}	1	2	3
1	3	0	0
2	0	1	0
3	0	0	1

×

A_{pp}	1	2	3
1	1	2	3
2	4	5	6
3	7	8	9

←

R_{pp}	1	2	3
1	3	6	9
2	4	5	6
3	7	8	9

(e) R2 ← R2 + R1

T_{pp}	1	2	3
1	1	0	0
2	1	1	0
3	0	0	1

×

A_{pp}	1	2	3
1	1	2	3
2	4	5	6
3	7	8	9

←

R_{pp}	1	2	3
1	1	2	3
2	5	7	9
3	7	8	9

(f) R2 ← R2 + 2 R1

T_{pp}	1	2	3
1	1	0	0
2	2	1	0
3	0	0	1

×

A_{pp}	1	2	3
1	1	2	3
2	4	5	6
3	7	8	9

←

R_{pp}	1	2	3
1	1	2	3
2	6	9	12
3	7	8	9

(g) R2 ← 3 R2 + 2 R1

T_{pp}	1	2	3
1	1	0	0
2	2	3	0
3	0	0	1

×

A_{pp}	1	2	3
1	1	2	3
2	4	5	6
3	7	8	9

R_{pp}	1	2	3
1	1	2	3
2	14	19	24
3	7	8	9

En la última operación (g) observamos que el elemento diagonal multiplica la fila correspondiente y el elemento no diagonal multiplica la fila de su número de columna y agrega la fila multiplicada a la fila en cuestión.

* Véase Shiba (1975: 197-199).

(*) Diferencial de matriz

En varias ocasiones en el análsis multivariante, realizamos la diferencial de matriz por vector.

[1] Veamos un caso de la diferencia de Tpp por Wp:

$$T_{pp} = Yp^{T} X_{np} Wp = [y_1, y_2, \ldots, y_p] \begin{bmatrix} x_{11} & x_{12} & \ldots & x_{1p} \\ x_{21} & x_{22} & \ldots & x_{2p} \\ \ldots & \ldots & \ldots & \ldots \\ x_{n1} & x_{n2} & \ldots & x_{np} \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ \ldots \\ w_p \end{bmatrix}$$

$$Wp = \begin{bmatrix} w_1 \\ w_2 \\ \ldots \\ w_p \end{bmatrix}$$

Para saber su significado, desarrollamos T_{pp}:

$$\begin{aligned} T_{pp} = \; & [y_1 x_{11} + y_1 x_{12} + \ldots + y_1 x_{1p}, \\ & y_1 x_{21} + y_2 x_{22} + \ldots + y_2 x_{2p}, \\ & \ldots, \\ & y_1 x_{n1} + y_2 x_{n2} + \ldots + y_p x_{np}] \begin{bmatrix} w_1 \\ w_2 \\ \ldots \\ w_p \end{bmatrix} \end{aligned}$$

$$\begin{aligned} [1] \quad & = y_1 x_{11} w1 + y_1 x_{12} w2 + \ldots + y_1 x_{1p} w_p \\ & + y_1 x_{21} w1 + y_2 x_{22} w2 + \ldots + y_2 x_{2p} w_p \\ & + \ldots \\ & + y_1 x_{n1} w1 + y_p x_{n2} w2 + \ldots + y_p x_{np} w_p \end{aligned}$$

En lugar de $\frac{\partial S}{\partial a}$ utilizamos Df(S, w), que signifida: diferenciar S por w:

$Df(T_{pp}, w1) = y_1 x_{11} + y_2 x_{21} + \ldots + y_1 x_{n1}$ primera columna de [1]
$Df(T_{pp}, w2) = y_1 x_{12} + y_2 x_{22} + \ldots + y_2 x_{n2}$ segunda columna de [1]
...
$Df(T_{pp}, w_p) = y_1 x_{1p} + y_2 x_{2p} + \ldots + y_p x_{np}$ p columna de [1]

Juntamos estas diferenciales en:

$Df(T_{pp}, Wp) = Df(Yp^T X_{pp} Wp, Wp) = Yp^T X_{pp}$ ←vector vertical

Compárese con la diferencial que hemos aprendido en el colegio:

$Df(yxw, w) = yx$

[2] Veamos el segundo caso donde en el objeto se encuentra el cuadrado de Wp:

$$T_{pp} = Wp^T X_{pp} Wp = [w1, w2, \ldots, w_p] \begin{bmatrix} x_{11} & x_{12} & \ldots & x_{1p} \\ x_{12} & x_{22} & \ldots & x_{2p} \\ \ldots & \ldots & \ldots & \ldots \\ x_{1p} & x_{2p} & \ldots & x_{pp} \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ \ldots \\ w_p \end{bmatrix}$$

$$Wp = \begin{bmatrix} w_1 \\ w_2 \\ \ldots \\ w_p \end{bmatrix}$$

donde X_{pp} es una matriz simétrica:

$$T_{pp} = [w1\ x_{11} + w1\ x_{12} + \ldots + w1\ x_{1p},\ w1\ x_{21} + w2\ x_{22} + \ldots + w2\ x_{2p},\ \ldots,\ w1\ x_{n1} + w2\ x_{n2}\ w2 + \ldots + w_p\ x_{np}\ ] \begin{bmatrix} w_1 \\ w_2 \\ \ldots \\ w_p \end{bmatrix}$$

$$= w1\ x_{11}\ w1 + w1\ x_{12}\ w2 + \ldots + w1\ x_{1p}\ w_p + w2\ x_{12}\ w1 + w2\ x_{22}\ w2 + \ldots + w2\ x_{2p}\ w_p + \ldots + w_p\ x_{1p}\ w1 + w_p\ x_{2p}\ w2 + \ldots + w_p x_{pp}\ w_p$$

$$= x_{11}\ w1^2 + w1\ x_{12}\ w2 + \ldots + w1\ x_{1p}\ w_p + w2\ x_{12}\ w1 + x_{22}\ w2^2 + \ldots + w2\ x_{2p}\ w_p + \ldots + w_p\ x_{1p}\ w1 + w_p\ x_{2p}\ w2 + \ldots + x_{pp}\ w_p^2$$

Los elementos que contienen w1 son de primera fila y de primera columna de la fórmula anterior:

$Df(T_{pp}, w1) = 2w1\ x_{11} + 2(w2\ x_{12 +} \ldots _{+}\ w_p\ x_{1p}) = 2(w1\ x_{11} + w2\ x_{12 +} \ldots _{+}\ w_p\ x_{1p})$

De la misma manera, los elementos que contienen w2 son de segunda fila y segunda columna:

$Df(T_{pp},\ w2) = 2w2\ x_{12} + 2(w2\ x_{22 +} \ldots _{+}\ 2w_p\ x_{2p}) = 2(w2\ x_{12} + w2\ x_{22 +} \ldots _{+}\ w_p\ x_{2p})$

...

Y finalmente:

$Df(T_{pp},\ w_p) = 2w_p\ x_{1p} + 2(d2\ x_{2p +} \ldots _{+}\ w_p\ x_{pp}) = 2(w_p\ x_{1p} +_{\ 2}\ x_{2p +} \ldots _{+}\ w_p\ x_{pp})$

Juntamos todas estas diferencias en:

$Df(T_{pp},\ Wp) = Diff.\ (Wp^T\ X_{pp}\ Wp,\ Wp) = 2\ X_{pp}\ Wp$

Compárese con:

$Df(wxw,\ w) = 2xw$

(#) Cambios cronológicos de letras españolas

La tabla inferior izquierda (X) muestra las frecuencias de determinadas letras españolas de obras del siglo XIII al XIX y el resultado del Análisis de Regresión Múltiple (<*> representa la abreviación de letras:

X: Obra	<*>	ñ	è	á	τ	Y
Cid	836				144	1207
Fazienda	902				157	1220
Alcalá	921				444	1230
GE	1,349				301	1270
Alexandre	877				78	1300
Lucanor	1,877				227	1330
Troyana	1,105				399	1350
LBA	1,366				146	1389
Alba	464	156			543	1433
Especulo	1,024	52			215	1450
Gramática	577	51		4	192	1492
Celestina	573	41			131	1499
Sumario	329	70			322	1514
Diálogo	561					1535
Lazarillo	297	33			142	1554

Z: Obra	Y	Esperado	Residual
Cid	1207	1396	-189
Fazienda	1220	1382	-162
Alcalá	1230	1249	-19
GE	1270	1266	4
Alexandre	1300	1421	-121
Lucanor	1330	1241	89
Troyana	1350	1249	101
LBA	1389	1335	54
Alba	1433	1485	-52
Especulo	1450	1419	31
Gramática	1492	1482	10
Celestina	1499	1491	8
Sumario	1514	1474	40
Diálogo	1535	1492	43
Lazarillo	1554	1505	49

Casada	139	40				1583	Casada	1583	1598	-15
Quijote	165	57	3		2	1605	Quijote	1605	1621	-16
Buscón	93	47	7		1	1626	Buscón	1626	1617	9
Criticón	147	45	20			1651	Criticón	1651	1616	35
Instante	4	21	94	2		1677	Instante	1677	1641	36
Austria	7	60	39			1704	Austria	1704	1665	39
Autoridades		27	3	196		1726	Autoridades	1726	1780	-54
Picarillo	4	123	108			1747	Picarillo	1747	1798	-51
Delincuente		42		229		1787	Delincuente	1787	1831	-44
Ortografía		35		93		1815	Ortografía	1815	1694	121
Diablo		55		223		1841	Diablo	1841	1845	-4
Sombrero		89		222		1874	Sombrero	1874	1894	-20
Perfecta		63		184		1899	Perfecta	1899	1820	79

Su intercepto y valores de cada variable son:

Intercepto	<*>	ñ	è	á	τ	Res. est
1554.853	- .112	1.475	.572	.936	- .457	70.948

Los valores negativos de la abreviación <*> y la conjunción en forma de «τ» apuntan la correlación inversa de su frecuencia y el orden cronológico. Por otra parte, la letra española por excelencia <eñe> y vocales con acento gráfico grave <è>y agudo <á> muestran su correlación recta con los años de publicación. Sin embargo, la Media de residuales (Std. res) indica 71 años, lo que significa que es difícil realizar la predicción precisa con las frecuencias de letras sueltas.

(*) Método de cuantificación de tipo I.

El «Método de cuantificación de tipo I» se utiliza para realizar el análisis de regresión múltiple con datos cualitativos, por ejemplo:

X	v1	v2	v3	Punto
d1		v		12
d2	v	v	v	11
d3	v		v	13
d4	v	v		7
d5	v	v		14

Xr	Punto	Esperado	Residual
d1	12.000	12.000	.000
d2	11.000	11.000	.000
d3	13.000	13.000	.000
d4	7.000	10.500	-3.500
d5	14.000	10.500	3.500

Peso	P: Intercepto	v1	v2	v3	Res. est.
Valor	14.000	-1.500	-2.000	.500	2.214

Debemos tener cuidado con los casos siguientes:

Y	v1	v2	v3	Punto
d1	v	v		12
d2	v	v	v	11
d3	v		v	13
d4	v	v		7
d5	v	v		14

Z	v1	v2	v3	Punto
d1		v		12
d2	v	v		11
d3	v		v	13
d4	v	v		7
d5	v	v		14

En la tabla superior izquierda (X), la columna v1 está seleccionada en todos los individuos, lo que significa que no lleva la información distintiva. En la tabla derecha (Z), las dos columnas v2 y v3 están en relación de distribución complementaria. Esto quiere decir que al seleccionar una de las dos, la otra está determinada de manera unívoca. En tales casos no existe la matriz inversa que posibilite el análisis. Conviene buscar otras variables que no presenten estas características: distribución homogénea y complementaria.

5.3.2. Análisis de componentes principales

Multiplicando los valores de variables por un vector de pesos para que la Varianza de todas las variables de los datos resulte Máxima y, al mismo tiempo, la Correlación entre todas las variables resulte 0, las variables así multiplicadas cobran un nuevo sentido sintético. Los mismos pesos también pueden dar a los datos mismos para ver sus posiciones dentro de las nuevas variable sintéticas. Por ejemplo, la nueva variable que muestra una correlación grande con la variable de Matemáticas y con la de Ciencia dentro de las puntuaciones de exámenes, se considera como indicadora de peso de Ciencias exactas. El método se llama «Análisis de componentes principales» de Pearson.

La nueva variable que sea multiplicada por un peso para que su Varianza cobre su máximo valor facilita la interpretación de la variación de los datos de la mejor manera posible. Y otra variable que le sigue en su Varianza, presenta la segunda mejor explicación de los datos. Las dos variables, que muestran la Correlación cero (0), dan sus propias explicaciones no solapadas. El número de tales variables nuevas es el mismo que el de las variables de los datos objeto. Sin embargo, las variables sucesivas presentan cada vez menos Varianza, por lo que disminuye su capacidad explicativa, de modo que es suficiente analizar las primeras variables nuevas.

Para proceder a su cálculo, preparamos los puntos estandarizados (X_{np}) con el vector de Medias veticales y otro de Desviaciones Típicas:

$$X_{np} = (D_{np} - M_p) / S_p$$

Postmultiplicamos esta matriz X_{np} por el vector incógnito W_p para

formular el vector vertical Z_n:

[1] $Z_n = X_{np}\ W_p$

Buscamos la Varianza (V) de este vector compuesto:

[2] $V = (Z_n{}^T\ Z_n) / N$
$= (X_{np}\ W_p)^T\ (X_{np}\ W_p) / N$ ←[1]
$= W_p{}^T\ X_{np}{}^T\ X_{np}\ W_p / N$ ←$(A\ B)^T = B^T\ A$
$= W_p{}^T\ (X_{np}{}^T\ X_{np} / N)\ W_p$ ←N es escalar, movible
$= W_p{}^T\ R_{pp}\ W_p$ ← $R_{pp} = X_{np}{}^T\ X_{np} / N$ (cap. 4)

Ponemos la condición de Wp, cuya suma de productos sea 1. Sin tal condición, existe sinnúmero de Wp:

[3] $W_p{}^T\ W_p = 1$

Con esta restricción [3], para buscar el Máximo de la Varianza (V) [2], calculamos la diferencial de F, con el «multiplicador de Lagrange», que vamos a ver posteriormente, que tiene que ser 0, en el Máximo donde no existe la inclinación:

$F = W_p{}^T\ R_{pp}\ W_p - L\ (W_p{}^T\ W_p - 1)$

La diferencia (Df) de F por Wp es:

[4a] $Df(F, W_p) = 2\ R_{pp}\ W_p - 2\ L\ W_p = 0$

[4b] $R_{pp}\ W_p = L\ W_p$ ←mover Wp al lado derecho, dividir ambos por 2

Esta forma [4b] se llama «Ecuación característica», de la que se derivan tanto el «Valor propio» L como «Vector propio» Wp. El Valor propio corresponde a La Varianza como se demuestra seguidamente:

$V = W_p{}^T\ R_{pp}\ W_p$ ←[2]
$= W_p{}^T\ L\ W_p$ ←[4b]
$= L\ W_p{}^T\ W_p$ ←L es escalar, movible
$= L$ ←[3]

Existen tantos valores propios y los vectores propios como las variables del dato objeto, y los vectores se llaman «Componentes» que se ordenan por la magnitud de su correspondiente «Valor propio».

La tabla inferior izquierda es de las puntuaciones de los individuos de d1 a d7, de variables M: Matemáticas, S: Ciencia y L, Latín. La tabla PCAd son los puntos de los tres componentes, #1, #2, #3. PCAv es la matriz propia y, por último, PCAe: valores propios, es decir, Varianzas de los componentes:

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

PCAd	#1	#2	#3
d1	-.823	-.544	.325
d2	.635	-.149	.369
d3	-.176	.588	.007
d4	3.171	.218	-.239
d5	-.510	-1.668	-.270
d6	-1.383	.789	-1.025
d7	-.916	.766	.834

PCAv	#1	#2	#3
E	.569	.616	-.545
L	.635	.093	.767
S	-.523	.782	.338

PCAe	#1	#2	#3
E.value	2.026	.672	.303

(*) Vector de valores propios y matriz propia

La tabla inferior izquierda es la matriz de datos, la derecha es su matriz de correlación:

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

R_{pp}	M	S	L
M	1.000	.643	-.335
S	.643	1.000	-.545
L	-.335	-.545	1.000

Con una matriz cuadrada R, si existe tal relación, que se llama «ecuación característica»:

$$R_{pp}\ W_p = L\ W_p$$

el escalar L se llama «valor propio» y el vector Wp, «vector propio». Como existen P valores propios y P vectores propios correspondientes a la Matriz R_{pp}, llamamos al conjunto de valores propio el «Vector de valores propios», y al conjunto de vectores propios, «Matriz propia»:

$$R_{pp}\ E_{pp} = L_p\ E_{pp}$$

La tabla inferior izquierda es la matriz Rpp, la media, la matriz propia, y la derecha, el producto de multiplicación de las dos:

R	M	S	L
M	1.000	.643	-.335
S	.643	1.000	-.545
L	-.335	-.545	1.000

X

E	#1	#2	#3
M	.569	.616	-.545
S	.635	.093	.767
L	-.523	.782	.338

=

R E	#1	#2	#3
M	1.152	.414	-.165
S	1.286	.062	.232
L	-1.060	.526	.102

La tabla inferior izquierda es el vector de valores poprios (Lp), la media es la matriz propia y la derecha es el producto de multiplicación de las dos:

L	#1	#2	#3
Value	2.026	.672	.303

E	#1	#2	#3
M	.569	.616	-.545
S	.635	.093	.767
L	-.523	.782	.338

L E	#1	#2	#3
M	1.152	.414	-.165
S	1.286	.062	.232
L	-1.059	.526	.102

De esta manera se comprueba:

$$R_{pp}\ E_{pp} = Lp\ E_{pp}$$

Comprobamos también que la matriz propia está constituida de vectores de longitud 1 y de producto interior 0.

$$E_{pp}^{T}\ E_{pp} = I_{pp} \text{ (matriz identidad)}$$

E^T	M	S	L
#1	.569	.635	-.523
#2	.616	.093	.782
#3	-.545	.767	.338

X

E	#1	#2	#3
M	.569	.616	-.545
S	.635	.093	.767
L	-.523	.782	.338

=

E^T E	1	2	3
1	1.000	.000	.000
2	.000	1.000	.000
3	.000	.000	1.000

(*) Ortogonalidad de matriz propia

El producto de multiplicación de los dos vectores propios presenta 0:

$$E_{p(i)}^{T}\ E_{p(j)} = 0 \quad [i \neq j]$$

Vamos a derivar esta propiedad de manera siguiente:

1. $R_{pp}\ E_{p(j)} = L_{(j)}\ E_{p(j)}$ ←def. de valore propio y vector propio

Hacemos operacione idénticas a los dos lados de 1.

2. $E_{p(i)}^{T}\ R_{pp}\ E_{p(j)} = E_{p(i)}^{T}\ L_{(j)} * E_{p(j)}$ ←premultiplicar $E_{p(i)}^{T}$
3. $= L_{(j)}\ E_{p(i)}^{T}\ E_{p(j)}$ ← $L_{(j)}$ es escalar, movible

Transformamos el lado izq. de 1.

4. $E_{p(i)}^{T}\ R_{pp}\ E_{p(j)} = [R_{pp}^{T}\ E_{p(i)}]^{T}\ E_{p(j)}$ ← $B^T A = (A B)^T$

5. $= [E_{p(j)}{}^T R_{pp} E_{p(i)}]^T$ ←$B^T A = (A^T B)^T$
6. $= [E_{p(j)}{}^T L_{(i)} E_{p(i)}]^T$ ←Fórmula propia: $R E_p = L E_p$
7. $= L_{(i)} [E_{p(j)}{}^T E_{p(i)}]^T$ ←L es escalar, movible
8. $= L_{(i)} E_{p(i)}{}^T E_{p(j)}$ ←$(A^T B)^T = B^T A$

Los lados izq. de 2. y 4. son iguales:

9. $L_{(j)} E_{p(i)}{}^T E_{p(j)} = L_{(i)} E_{p(i)}{}^T E_{p(j)}$ ←3. = 8.
10. $[L_{(i)} - L_{(j)}] E_{p(i)}{}^T E_{p(j)} = 0$ ←trasladar el lado izq. a dcho
11. $E_{p(i)}{}^T E_{p(j)} = 0$ $[L_{(i)} \neq L_{(j)}]$

El que la suma de multiplicaciones de los elementos de los dos vectores sea 0 significa que los dos presentan una ortogonalidad, es decir se cruzan con ángulo recto. Por otra parte presuponemos que la longitud de los vectores propios es 1:

12. $E_{p(i)}{}^T E_{p(i)} = 1$

Considerando 11. y 12. de todos los vectores constituyentes de la matriz propia, se formula lo siguiente:

13. $E_{pp}{}^T E_{pp} = I_{pp}$ [E_{pp} es matriz identidad]

* Sobre la ortogonaliad de vectores propios, hemos consultado Adachi (2005).

(*) Decomposición espectral

La matriz Rpp dentro de la fórmula de la matriz propia:

$R_{pp} E_{pp} = L_p E_{pp}$

es descomponible de manera siguiente. Se llama «decomposición espectral», que se utiliza en el «método iterativo potencial» de que trataremos en seguida.

a. $R_{pp} = L_{(1)} E_{(1)} E_{(1)}{}^T + L_{(2)} E_{(2)} E_{(2)}{}^T + \ldots + L_{(p)} E_{(p)} E_{(p)}{}^T$

donde, (1), (2), …, (p) representan los valores propios y sus correspondientes vectores propios.

Para derivar esta ecuación, preparamos de antemano las siguientes propiedades de la matriz propia Epp.

b1. $E_{pp}{}^T E_{pp} = I_{pp}$ ←diagonaliad de Epp
b2. $E_{pp}{}^{-1} E_{pp} = I_{pp}$ ←def. de matriz inversa: $X^{-1} X = I$
b3. $E_{pp}{}^T = E_{pp}{}^{-1}$ ←b1, b2
b4. $(E_{pp}{}^T)^{-1} E_{pp}{}^T = I_{pp}$ ←def. de matriz inversa: $X^{-1} X = I$

b5 $(E_{pp}^{-1})^T E_{pp}^T = I_{pp}$ ←propiedad de matriz inversa: $(X^T)^{-1} = (X^{-1})^T$

b6 $(E_{pp}^T)^T E_{pp}^T = I_{pp}$ ←b3

b7 $E_{pp} E_{pp}^T = I_{pp}$ ←propiedad de matriz traspuesta: $(X^T)^T = X$

b8 $E_{pp}^T E_{pp} = E_{pp} E_{pp}^T = I_{pp}$ ←b1, b7

Ahora bien, empecemos con la fórmula de la matriz propia:

c1. $R_{pp} E_{pp} = L_p E_{pp}$ ←matriz propia

c2. $R_{pp} E_{pp} E_{pp}^T = L_p E_{pp} E_{pp}^T$ ←premultiplicar E_{pp}^T

c3. $R_{pp} E_{pp}^T E_{pp} = L_p E_{pp} E_{pp}^T$ ←b8: $E_{pp}^T E_{pp} = E_{pp} E_{pp}^T$

c4. $R_{pp} E_{pp}^{-1} E_{pp} = L_p E_{pp} E_{pp}^T$ ←b3: $E_{pp}^T = E_{pp}^{-1}$

c5. $R_{pp} I_{pp} = L_p E_{pp} E_{pp}^T$ ←c4, b2: $E_{pp}^{-1} E_{pp} = I_{pp}$

c6. $R_{pp} = L_p E_{pp} E_{pp}^T$ ←R I = R

Seguidamente desarrollamos el productio de la multiplicación matricdial $E_{pp} E_{pp}^T$.

E_{pp}	(1)	(2)	(…)	(p)
1	e_{11}	e_{12}	…	e_{1p}
2	e_{21}	e_{22}	…	e_{2p}
…	…	…	…	…
p	e_{p1}	e_{p2}	…	e_{pp}

X

E_{pp}^T	1	2	…	p
(1)	e_{11}	e_{21}	…	e_{p1}
(2)	e_{12}	e_{22}	…	e_{p2}
(…)	…	…	…	…
(p)	e_{1p}	e_{2p}	…	e_{pp}

=

$e_{11}e_{11}+e_{12}e_{12}+…+e_{1p}e_{1p}$	$e_{11}e_{21}+e_{12}e_{22}+…+e_{1p}e_{2p}$	…	$e_{11}e_{p1}+e_{12}e_{p2}+…+e_{1p}e_{pp}$
$e_{21}e_{11}+e_{22}e_{12}+…+e_{2p}e_{1p}$	$e_{21}e_{21}+e_{22}e_{22}+…+e_{2p}e_{2p}$	…	$e_{21}e_{p1}+e_{22}e_{p2}+…+e_{2p}e_{pp}$
…	…	…	…
$e_{p1}e_{11}+e_{p2}e_{12}+…+e_{pp}e_{1p}$	$e_{p1}e_{21}+e_{p2}e_{22}+…+e_{pp}e_{2p}$	…	$e_{p1}e_{p1}+e_{p2}e_{p2}+…+e_{pp}e_{pp}$

Para cada vector vertical que constituye Epp, $E_{(1)}$, $E_{(2)}$, … $E_{(p)}$, que son vectores propios, desarrolamos sus elementos y el producto de multiplicaciones $E_{(1)} E_{(1)}^T$. No presenta escalar por no ser $E_{(1)}^T E_{(1)}$, sino $E_{(1)} E_{(1)}^T$, Primero vamos a ver el caso del primer vector:

$E_{(1)} E_{(1)}^T$ =

E	(1)
1	e_{11}
2	e_{21}
…	…
p	e_{p1}

X

E^T	1	2	…	p
(1)	e_{11}	e_{21}	…	e_{p1}

=

$e_{11}e_{11}$	$e_{11}e_{21}$	…	$e_{11}e_{p1}$
$e_{21}e_{11}$	$e_{21}e_{21}$	…	$e_{21}e_{p1}$
…	…	…	…
$e_{p1}e_{11}$	$e_{p1}e_{21}$	…	$e_{p1}e_{p1}$+

Aquí observamos que los elementos de $E_{(1)}\ E_{(1)}{}^T$ son los primeros términos de los elementos correspondientes de $E_{pp}\ E_{pp}{}^T$. Vamos a ver el caso del segundo vector:

$E_{(2)}\ E_{(2)}{}^T =$

E	(2)
1	e_{12}
2	e_{22}
…	…
p	e_{p2}

X

E^T	1	2	…	p
(2)	e_{12}	e_{22}	…	e_{p2}

=

$e_{12}e_{12}$	$e_{12}e_{22}$	…	$e_{12}e_{p2}$
$e_{22}e_{12}$	$e_{22}e_{22}$	…	$e_{22}e_{p2}$
…	…	…	…
$e_{p2}e_{12}$	$e_{p2}e_{22}$	…	$e_{p2}e_{p2}$

Aquí también observamos que los elementos de $E_{(2)}\ E_{(2)}{}^T$ son los segundos términos de los elementos correspondientes de $E_{pp}\ E_{pp}{}^T$. De la misma manera, el producto de los p-ésimos vectores:

$E_{(p)}\ E_{(p)}{}^T =$

E	(p)
1	e_{1p}
2	e_{2p}
…	…
p	e_{pp}

X

E^T	1	2	…	p
(p)	e_{1p}	e_{2p}	…	e_{pp}

=

$e_{1p}e_{1p}$	$e_{1p}e_{2p}$	…	$e_{1p}e_{pp}$
$e_{2p}e_{1p}$	$e_{2p}e_{2p}$	…	$e_{2p}e_{pp}$
…	…	…	…
$e_{pp}e_{1p}$	$e_{pp}e_{2p}$	…	$e_{pp}e_{pp}$

De esta manera comprobamos que la suma de $E_{(i)}\ E_{(i)}{}^T$ (i = 1, 2, …, p) es igual a $E\ E^T$:

$$E_{pp}\ E_{pp}^{T} = E_{(1)}\ E_{(1)}^{T} + E_{(2)}\ E_{(2)}^{T} + \ldots + E_{(p)}\ E_{(p)}^{T}$$

Por lo tanto,

$$R_{pp} = L_p\ E_{pp}\ E_{pp}^{T} \quad \leftarrow \text{c6.}$$
$$= L_{(1)}\ E_{p(1)}\ E_{p(1)}^{T} + L_{(2)}\ E_{p(2)}\ E_{p(2)}^{T} + \ldots + L_{(p)}\ E_{p(p)}\ E_{p(p)}^{T}$$
$$\leftarrow L_{(1)},\ L_{(2)},\ \ldots,\ L_{(p)} \text{ son escalares}$$

*Para la decomposición espectral hemos consultado Adachi (2005) e Iwasaki y Yoshida (2006).

(*) Método reiterativo potencial

Para obtener el vector de valores propios y la matriz propia, se utiliza el «Método reiterativo potencial». Se busca primero el máximo de valores propios y sus correspondientes vectores propios y seguidamente, se busca los segundos, terceros, así sucesivamente con las matrices restantes dentro de la descomposición espectral.

En la fórmula de la matriz propia, $R_{pp}\ E_{pp} = L_p\ E_pp$, pueden existir un sin número de tanto Lp como Epp, de modo que se pone una restricción de que la longitud del vector propio sea 1:

[1] $$E_{p(i)}^{T}\ E_{p(i)} = 1 \quad (i = 1, 2, \ldots, p)$$

Otra condición es que $E_{p(i)}$ y $E_{p(j)}$ $[i \neq j]$ sean ortogonales:

$$E_{p(i)}^{T}\ E_{p(j)} = 0 \quad (i, j = 1, 2, \ldots, p;\ i \neq j)$$

Por lo tanto,

$$E_{pp}^{T}\ E_{pp} = I_{pp} \text{ (matriz identidad)}$$

Supongamos que el estado inicial de la suma $S_p^{(0)}$ de los vectores propios $E_{p(1)}, E_{p(2)}, \ldots, E_{p(p)}$ es:

$$S_p^{(0)} = E_{p(1)} + E_{p(2)} + \ldots + E_{p(p)}$$

Se supone que se van premultiplicando Rpp a ambos lados:

$$S_{p(1)} = R_{pp}\ S_p^{(0)} = R_{pp}\ E_{p(1)} + R_{pp}\ E_{p(2)} + \ldots + R_{pp}\ E_{p(p)} \quad \leftarrow \text{premult. R}$$
$$= L_{(1)}\ E_{p(1)} + L_{(2)}\ E_{p(2)} + \ldots + L_{(p)}\ E_{p(p)} \quad \leftarrow R_{pp}\ E_p = L\ E_p$$

$$S_{p(2)} = R_{pp}^{2}\ S_p^{(0)} = L_{(1)}^{2}\ E_{p(1)} + L_{(2)}^{2}\ E_{p(2)} + \ldots + L_{(p)}^{2}\ E_{p(p)} \quad \leftarrow \text{premult. R}$$

(…) $\leftarrow$ premult. R sucesivamente

$S_p^{(k)} = R_{pp}^{k} S_p^{(0)} = L_{(1)}^{k} E_{p(1)} + L_{(2)}^{k} E_{p(2)} + \ldots + L_{(p)}^{k} E_{p(p)}$

Vamos a detectar el máximo de $L_{(1)}$, $L_{(2)}$, …, $L_{(p)}$ en forma de $L_{(m)}$:

$L_{(m)} > L_{(1)}, L_{(2)}, \ldots, L_{(p)}$

Por lo tanto,

$S_p^{(k)} = L_{(1)}^{k} E_{p(1)} + \ldots + \mathbf{L_{(m)}^{k}} E_{p(m)} + \ldots + L_{(p)}^{k} E_{p(p)}$ ←**$L_{(m)}$** es L máximo

$= L_{(m)}^{k} [L_{(1)}^{k} / L_{(m)}^{k} E_{(p)}1 + \ldots + E_{p(m)} + \ldots + L_{(p)}^{k} / L_{(m)}^{k} E_{p(p)}]$

←$L_{(m)}^{k}$ al exterior

Se supone que si se va aumentando k, todos los términos menos $E_{p(m)}$ se vuelve casi cero (0), puesto que su denominador se aumenta excesivamente.

[2] $S_p^{(k)} \fallingdotseq L_{(m)}^{k} E_{p(m)}$ (k →極大)

Y, en este momento, la ratio entre $S_p^{(k-1)}$ y $S_p^{(k)}$ va a ser igual a la ratio entre $L_{(m)}^{k-1}$ y $L_{(m)}^{k}$, es decir, $L_{(m)}$. En realidad solo trabajamos con L Ep premultiplicándolo por Rpp hasta que Ep no presente un mínimo de cambio, puesto que los restantes teóricamente son ignorables.

En la seguna ronda en la que se busca el segundo valor propio y su vector propio se trabaja con la matriz de Rpp(2), que es la restante de Rpp en la decomposición espectoral:

$R_{pp} = L_{(1)} E_{(1)} E_{(1)}^{T} + L_{(2)} E_{(2)} E_{(2)}^{T} + \ldots + L_{(p)} E_{(p)} E_{(p)}^{T}$

De la fórmula anterior, restamos el primer término $L_{(1)} E_{(1)} E_{(1)}^{T}$ y ahora tabajos con:

$R_{pp}(2) = R_{pp}(1) - L(1) E_{p(1)} E_{p(1)}^{T}$

。Con la nueva $R_{pp(2)}$, se hace los mismos procesos, y así se repite los mismos procedimientos con $R_{pp(3)}$, $R_{pp(4)}$, …, $R_{pp(p)}$.

El algoritmo del programa para derivar el vector de los valores propios Lp y la matriz propia Epp es lo siguiente. Primero se prepara el Ep, el primer vector constituyente de Epp, con un estado inicial del vector identidad:

$E_p \leftarrow I_p$

Seguidamente:

[4] $E_p \leftarrow R_{pp} E_p$ ←premultiplicar Ep por Rpp

[5] $E_p \leftarrow E_p / (E_p^T E_p)^{1/2}$ ←poner la longitud de Ep en 1 ...[1]
[6] Si el cambio de Ep es grande, se vuelve a [4]; si es suficientemente pequeño, se termina y procede a la segunda ronda.

Se repite multitud de veces los procesos de [4][5] hasta que no se presente el cambio significativo, es decir llega al número k de [2], cuando $Rpp^{(k)}$ $Ep^{(k)}$ (=$Sp^{(k)}$) está todavía sin aplicar el proceso de [5], es decir mantiene el estado anterior $Rpp^{(k-1)}$ $Ep^{(k-1)}$ (=$Sp^{(k-1)}$) multiplicado por $(Ep^{(k-1)} Ep^{(k-1)})^{1/2}$, que corresponde a $L_{(m)}$ de [2], es decir el valor propio:

$$L_{(1)} = (Ep_{(1)} Ep_{(1)})^{1/2}$$

De esta manera se ha obtenido tanto el valor propio $L_{(1)}$ como el vector propio $E_{p(1)}$. Veamos el estado de la descomposición espectral:

$$R_{pp} = L_{(1)} E_{p(1)} E_{p(1)}^T + L_{(2)} E_{p(2)} E_{p(2)}^T + \ldots + L_{(p)} E_{p(p)} E_{p(p)}^T$$

de donde restamos $L_{(1)} E_{p(1)} E_{p(1)}^T$:

$$R_{pp}^{(2)} = R_{pp}^{(1)} - L_{(1)} E_{p(1)} E_{p(1)}^T$$

Y de nuevo calculamos el valor propio máximo, $L_{(2)}$, con su correspondiente vector propio $E_{p(2)}$ con $R_{pp}^{(2)}$. Y así sucesivamente buscamos cuantos valores propios y vectores propios necesitemos y completamos la matriz propia final, que es conjunto de vectores propios.

*Hemos consultado la explicación y el programa de Shirai (2009: 99-101)).

(*) Multiplicador de Lagrange

Se utiliza el «multiplicador de Lagrange» para realizar el cálculo diferencial con condiciones. Para buscar el Máximo de la función Y

[1] $Y = f(x_1, x_2, \ldots, x_n)$

se hace el cálculo diferencia de Y por variables $x_1, x_2, \ldots, x_n$:

$$Df(Y, x_1)=0,\ Df(Y, x_2)=0,\ \ldots,\ Df(Y, x_n)=0$$

de esta manera se obtienen los valores de $x_1, x_2, \ldots, x_n$. Ahora bien, al efectuar el cálculo, tenemos que respetar una condición:

[2] $G = g(x_1, x_2, \ldots, x_n) = 0$

Para satisfacer esta condición, se utiliza el multiplicador de Lagrange L y se fomula la función W de nuevo:

[3] $W = Y - L\,G$

$= f(x_1, x_2, \ldots, x_n) - L\,g(x_1, x_2, \ldots, x_n)$

Y se hace el cálculo diferencial de W por $x_1, x_2, \ldots, x_n, L$.

[4] $Df(W, x_1) = 0, Df(W, x_2) = 0, \ldots, Df(W, x_p) = 0, \underline{Df(W, L) = 0}$

Al aplicar [3] $W = Y - L\,G$ a [4], se obtienen:

$Df(W, x_1) = Df(Y, x_1) - L\,Df(G, x_1) = 0$
$Df(W, x_2) = Df(Y, x_2) - L\,Df(G, x_2) = 0$
(…)
$Df(W, x_n) = Df(Y, x_n) - L\,Df(G, x_n) = 0$

Y la última fórmula de [4] subrayada da:

$Df(W, L) = Df(Y - L\,G, L) = -G = 0$ [Y はゼロ]

Por lo tanto,

$G = g(x_1, x_2, \ldots, x_n) = 0$

De esta manera observamos que al realizar el cálculo efectivamente G se vuelve 0.

* Hemos consultado Kobayashi (1967:89-90).

(*) Concentración por análisis de componentes principales

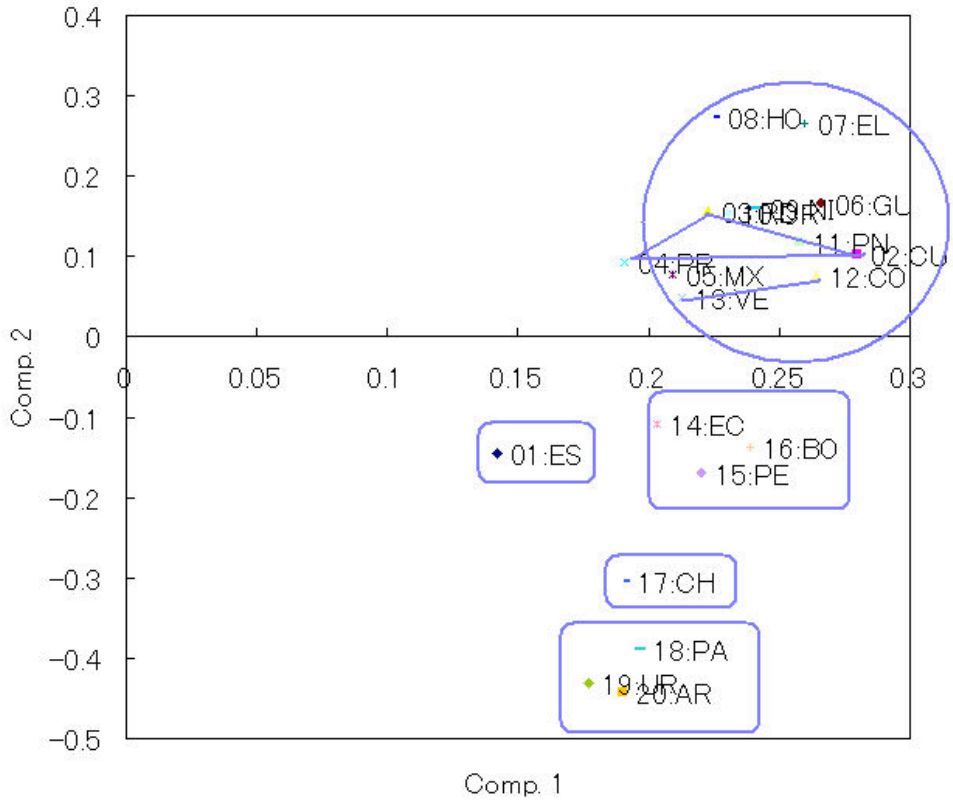
Al realizar el análisis de concentración con los pesos de variables puntos aplicados obtenidos por el análisis de componentes principales, se puede cambiar el orden de los dos ejes según estos valores para conseguir una distribución concentrada:

PCA.Cct	Latin	English	Physics
B	88	28	20
C	64	43	32
A	59	56	54
F	48	45	66
E	51	58	78
G	22	32	90
D	16	50	100

(#) Análisis de componentes principales del léxico variable español

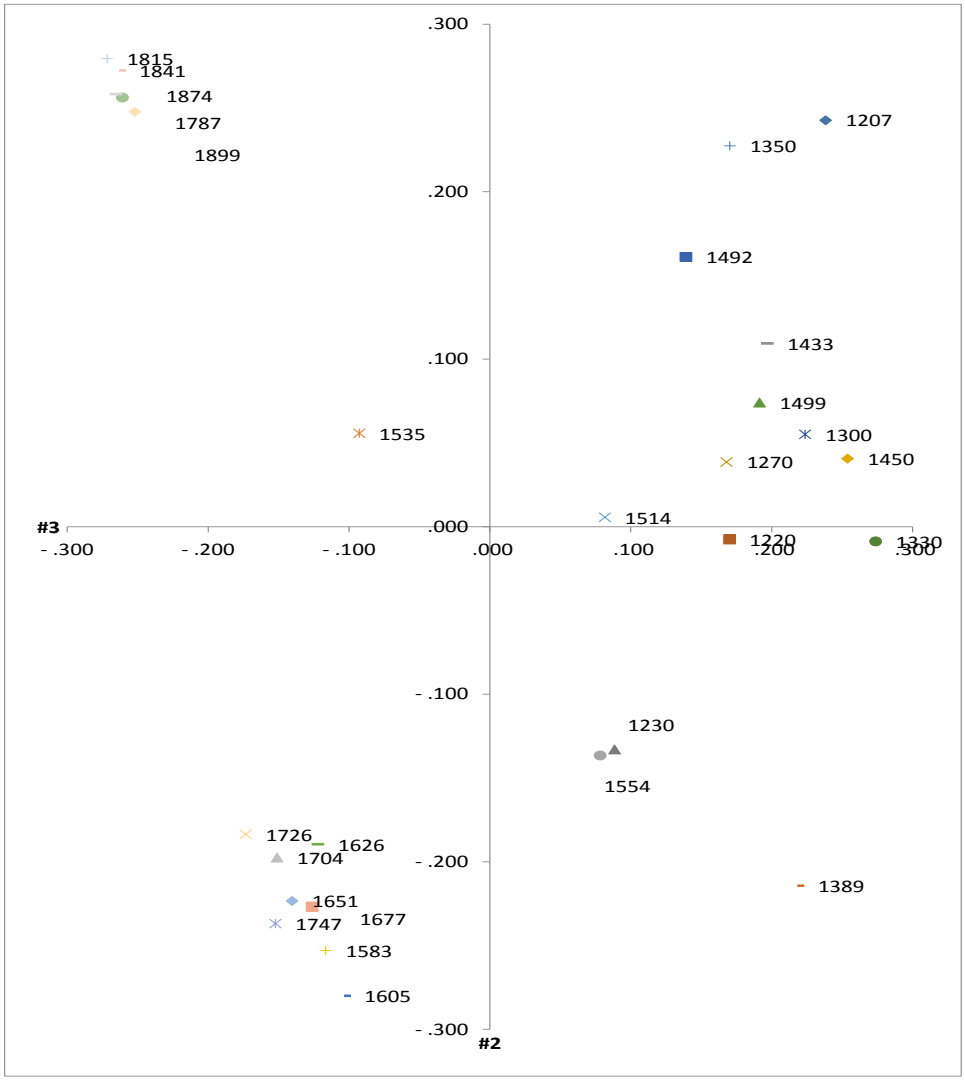
El gráfico siguiente muestra el resultado de análisis de componentes principales con los datos de 842 vocablos variantes en el español moderno en 20 países

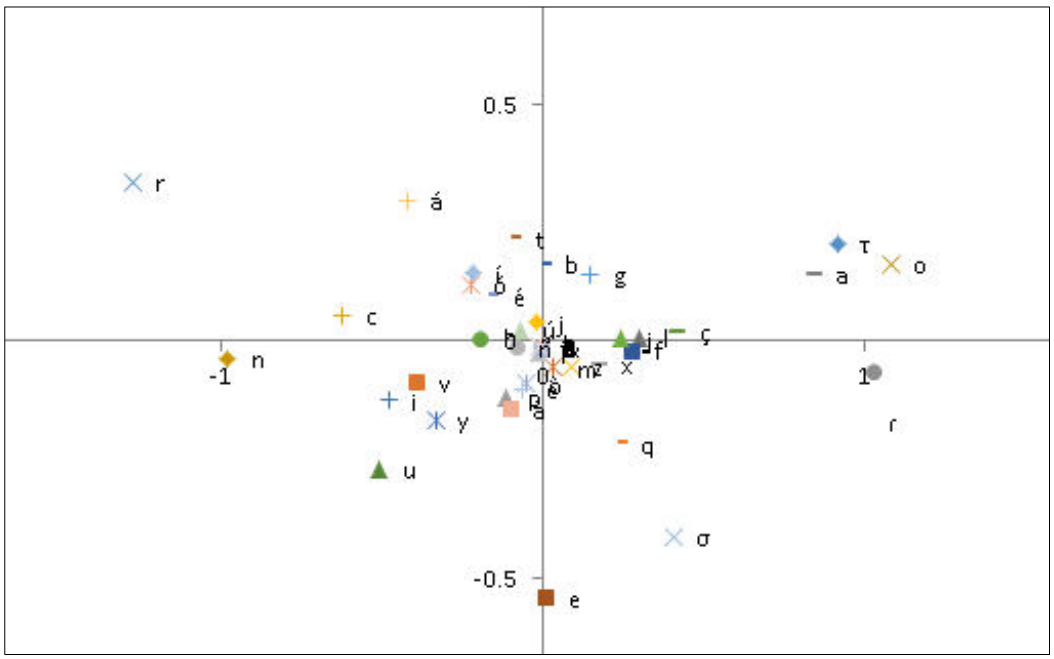
hispanohablantes. El eje horizontal indica el primer componente, y el vertical el segundo. El primer componente divide España (ES) del resto, y el segundo muestra el orden geográfico de norte a sur de Hispanoamérica: México y Centroamérica, países caribeños (MX; HO, EL, GI, PN; PR, CU, DR), Colombia (CO) y Venezuela (VE), países andinos (EC, BO, PE), Chile (CH) y la Plata (PA, IR, AR). La variación léxica, de esta manera, muestra un continuum geográfico.



(#) Análisis de componentes principales de las frecuencias de letras española en edad media y moderna

Preparamos una matriz con las frecuencias muestreadas de 28 obras del siglo XIII al XIX para ver su variación frecuencial. No hemos obtenido el primer componente interpretable. No obstante, el segundo componente parece dividir las épocas clásicas (Edad Media y Moderna) y la Edad Contemporánea. El tercer componente indica la división entre los siglos XVII, XVII por una parte, y el siglo XIX por otra. En cuanto a las letras, las variantes de <s>, <d> y <r> son importantes.


.300
.200
.100
.000
- .100
- .200
- .300
#3
- .300
- .200
- .100
.000
.100
.200
.300
#2
1815
1841
1874
1787
1899
1207
1350
1492
1433
1499
1535
1300
1270
1450
1514
1220
1330
1230
1554
1726
1626
1704
1651
1747
1677
1583
1605
1389


0.5
-0.5
-1
1
r
á
t
b
g
τ
a
o
c
ç
n
v
i
y
u
q
σ
e

5.3.3. Análisis discriminante

El método que se utiliza para realizar una inferencia de la variable cualitativa (Y) a partir de las variables cuantitativas (v1, v2, ...) con unos pesos desconocidos se llama «Análisis Discriminante» de Fisher. Se busca el vector de peso, que sirve para evaluar el significado cuantitativo de las variables. Los mismos pesos sirven para hacer la predicción de la variable cuantitativa objeto.

X	v1	v2	v3.	Y
d1	6	8	5	
d2	7	10	6	
d3	8	4	8	v
d4	9	7	2	
d5	10	9	4	v

La tabla puede ser por ejemplo, puntuaciones de pruebas de v1: Lectura, v2: Composición, v3: Vocabulario e Y, evaluación final de «Sobresaliente».

Primero se estandarizan las variables cuantitativas;

[1] $X_{np} = \mathrm{Std}(X_{np})$ ...estandarizar: (X – Media) / Desviación estándar

X_{np}	v1	v2	v3.
d1	-1.414	.194	.000
d2	-.707	1.166	.500
d3	.000	-1.748	1.500
d4	.707	-.291	-1.500
d5	1.414	.680	-.500

Supongamos que al X_{np} se multiplica por un vector desconocido¿ W_p:

[2] $Z_n = X_{np}\, W_p$

Una vez obtenido el vector W_p, se obtiene también Z_n por esta fórmula. Comprobamos que la Media (M) de Z_n es cero (0):

$$
\begin{aligned}
M &= (\Sigma_{(i:N)} Z_n) / N && \leftarrow \text{def. de Media} \\
&= \Sigma_{(i:N)} (X_{np}\, W_p) / N && \leftarrow [2] \\
&= \Sigma_{(i:N)} (X_{i1}W_1 + X_{i2}W_2 + \ldots + X_{ip}W_p) / N && \leftarrow \text{elementos de M} \\
&= (\Sigma_{(i:N)}X_{i1}W_1 + \Sigma_{(i:N)}X_{i2}W_2 + \ldots + \Sigma_{(i:N)}X_{ip}W_p) / N \\
& && \leftarrow \text{distribuir } \Sigma \\
&= (W_1 \Sigma_{(i:N)}X_{i1} + W_2 \Sigma_{(i:N)}X_{i2} + \ldots + Wp \Sigma_{(i:N)}X_{ip}) / N
\end{aligned}
$$

←anteponer constante

Como la matriz X_{np} está estandarizada, su Suma vertical es 0:

$$\Sigma_{(i:N)}X_{i1} = \Sigma_{(i:N)}X_{i2} = \ldots = \Sigma_{(i:N)}X_{ip} = 0$$

Por lo tanto, los numeradores de los tres términos de M son 0:

[3] $M = 0$

La variación total (S) de Z_n es:

$$\begin{aligned} S &= \Sigma_{(i:N)}(Z_i - M)^2 && \leftarrow \text{def. de variación} \\ &= \Sigma_{(i:N)}Z_i^2 && \leftarrow [3]\ M = 0 \end{aligned}$$

Dividimos el vector Z_n entre el grupo de los estudiantes con la evaluación «sobresaliente», vector Z_v y otros no sobresalientes, vector Z_c, cuyos números son Nv y Vc, y cuyas medias son Mv y Mc.

La Suma de las variaciones de Zv y Zc se llama «Variación entre grupos» (Sw):

$$Sw = \Sigma_{(i:Nv)}(Zv_i - Mv)^2 + \Sigma_{(i:Nc)}(Zc_i - Mc)^2$$

La Media de Zn es 0, pero no lo son necesariamente Mv y Mc, puesto que ni Zv ni Zc están estandarizados.

Suponiendo que los elementos de cada grupo son todos iguales, la suma de los cuadrados de diferencia entre su Media y Media total (M) se llama «Suma de cuadrados entre los dos grupos» (Sb). La Sb muestra el grado de variación de cada grupo dentro del conjunto, cuya Media es cero:

$$\begin{aligned} Sb &= \Sigma_{(i:Nv)}(Mv - M)^2 + \Sigma_{(i:Nc)}(Mc - M)^2 \\ &= \Sigma_{(i:Nv)}Mv^2 + \Sigma_{(i:Nc)}Mc^2 && \leftarrow [3]\ M = 0 \\ [4] \quad &= Nv\ Mv^2 + Nc\ Mc^2 && \leftarrow \text{Multiplicados por contstantes} \end{aligned}$$

La variación total (S) es igual a Sw + Sb, lo que se demuestra de la manera siguiente:

$$\begin{aligned} Sw &= \Sigma_{(i:Nv)}(Zv_i - Mv)^2 + \Sigma_{(i:Nc)}(Zc_i - Mc)^2 \\ &= \Sigma_{(i:Nv)}(Zv_i^2 - 2\,Zv_i\,Mv + Mv^2) && \leftarrow \text{desarrollar} \\ &+ \Sigma_{(i:Nc))}(Zc_i^2 - 2\,Zc_i\,Mc + Mc^2) \\ \\ &= \Sigma_{(i:Nv)}Zv_i^2 - \Sigma_{(i:Nv)}2Zv_i\,Mv + \Sigma_{(i:Nv)}Mv^2 && \leftarrow \text{distribuir } \Sigma \\ &+ \Sigma_{(i:Nc))}Zc_i^2 - \Sigma_{(i:Nv)}2Zc_i\,Mc + \Sigma_{(i:Nv)}Mc^2 \end{aligned}$$

$$= \Sigma_{(i:Nv)} Zv_i^2 - 2\,Mv\,\Sigma_{(i:Nv)} Zv_i + Nv\,Mv^2 \quad \leftarrow \text{adelantar Mv}$$
$$+ \Sigma_{(i:Nv)} Zc_i^2 - 2\,Mc\,\Sigma_{(i:Nv)} Zc_i + Nc\,Mc^2 \quad \leftarrow \text{adelantar Mc}$$

$$= \Sigma_{(i:Nv)} Zv_i^2 - 2\,Mv\,Nv\,Mv + Nv\,Mv^2 \quad \leftarrow \Sigma_{(i:Nv)} Zv_i = Nv\,Mv$$
$$+ \Sigma_{(i:Nv)} Zc_i^2 - 2\,Mc\,Nc\,Mc + Nc\,Mc^2 \quad \leftarrow \Sigma_{(i:Nv)} Zv_i = Nv\,Mc$$
$$\leftarrow \text{Suma = número * media}$$

$$= \Sigma_{(i:Nv)} Zv_i^2 - 2\,Nv\,Mv^2 + Nv\,Mv^2 \quad \leftarrow \text{juntar Mv}$$
$$+ \Sigma_{(i:Nv)} Zc_i^2 - 2\,Nc\,Mc^2 + Nc\,Mc^2 \quad \leftarrow \text{juntar Mc}$$

[5] $$= \Sigma_{(i:Nv)} Zv_i^2 - Nv\,Mv^2 + \Sigma_{(i:Nv)} Zc_i^2 - Nc\,Mc^2$$

por lo tanto,

$$Sw + Sb = \Sigma_{(i:Nv)} Zv_i^2 - Nv\,Mv^2 + \Sigma_{(i:Nc)} Zc_i^2 - Nc\,Mc^2 \quad \leftarrow [5]\ Sw$$
$$+ Nv\,Mv^2 + Nc\,Mc^2 \quad \leftarrow [4]\ Sb$$

$$= \Sigma_{(i:Nv)} Zv_i^2 + \Sigma_{(i:Nc)} Zc_i^2 = St$$

Seguidamente veremos la ratio de «Suma de cuadrados entre los dos grupos» Sb dentro de la variación total S, lo que se llama «Ratio de correlación» (RC).

[6] $$RC = Sb / S = Sb / (Sw + Sb)$$

Por ejemplo, si todos los elementos de cada grupo son de valor medio del grupo, la Sw se vuelve 0, lo que quiere decir que los elementos de cada grupo se concentran en un punto y, por [6], la RC se vuelve 1. Y cuando las Medias de cada grupo son iguales a la Media total (Mv = M, c = M), la Sw se vuelve 0, y no se discriminan los dos grupos y la RC llega a 0.

Expresamos S con el vector desconocido Wp:

$$S = Z_n^T Z_n$$
$$= (X_{np} W_p)^T (X_{np} W_p) \quad \leftarrow [2]\ Z_n = X_{np} W_p$$
$$= W_p^T X_{np}^T X_{np} W_p \quad \leftarrow \text{operación matricial}$$

donde representamos

[7] $$S_{pp} = X_{np}^T X_{np}$$

de modo que

[8] $$S = W_p^T S_{pp} W_p, \quad \leftarrow [1]$$

Y también expresamos Sb con Wp:

[9] $Sb = Nv\ Mv^2 + Nc\ Mc^2$ ← [4]

$= Nv\ (Sv_p{}^T/Nv\ W_p)^2$ ← Sv_p: Suma del grupo v en X_{np}

$+ Nc\ (Sc_p{}^T/Nc\ W_p)^2$ ← Sc_p: Suma del grupo c en X_{np}

$= Nv\ (Sv_p{}^T\ W_p)^2 / Nv^2$ ← Nv es escalar

$+ Nc\ (Scp^T\ W_p)^2 / Nc^2$ ← Nc es escalar

$= (Sv_p{}^T\ W_p)^2 / Nv$ ← Nv es escalar

$+ (Scp^T\ W_p)^2 / Nc$ ← Nc es escalar

$= (Sv_p{}^T\ W_p)^T\ (Sv_p{}^T\ W_p) / Nv$ ← operación matricial

$+ (Scp^T\ W_p)^T\ (Scp^T\ W_p) / Nc$ ← operación matricial

$= W_p{}^T\ Sv_p\ Sv_p{}^T\ W_p / Nv$ ← operación matricial

$+ W_p{}^T\ Scp\ Scp^T\ W_p / Nc$ ← operación matricial

$= W_p{}^T\ (Sv_p\ Sv_p{}^T / Nv + Sc_p\ Sc_p{}^T / Nc)\ W_p$

$= W_p{}^T\ B_{pp}\ W_p$ ← Bpp corresponde a [9b]

[9b] $Bpp = Sv_p\ Sv_p{}^T / Nv + Sc_p\ Sc_p{}^T / Nc$

Por lo tanto RC en [6] se convierte en:

[10] $RC = Sb / S = W_p{}^T\ B_{pp}\ W_p / W_p{}^T\ S_{pp}\ W_p$

$S_{pp} = X_{np}{}^T\ X_{np}$ ←[7]

$B_{pp} = Sv_p\ Sv_p{}^T / Nv + Sc_p\ Sc_p{}^T / Nc$ ←[9]

El objetivo del Análisis Discriminante es buscar el vector W_p con el máximo RC en [10], es decir, encontrar la mejor manera posible de discriminar los dos grupos. De modo que realizamos el cálculo diferencial de RC por W_p, con la restricción de $W_p{}^T\ S_{pp}\ W_p$←[8]) = 1. Pensamos en la función $F(W_p)$ con el multiplicador de Lagrange L y restricción de ST – 1 = 0:

$F\ (W_p) = Sb - L\ (S - 1)$ ← multiplicador de Lagrange L

$= W_p{}^T\ B_{pp}\ W_p\ -\ L\ (W_p{}^T\ S_{pp}\ W_p\ -\ 1)$ ←[8], [9], [9b]

Buscamos W_p de cuando la fórmula diferenciada de F por W_p sea 0:

$Diff.(F, W_p) = 2\ B_{pp}\ W_p\ -\ 2\ L\ S_{pp}\ W_p = 0$ ← diferencial de matriz

Por lo tanto,

[11] $(B_{pp}\ -\ L\ S_{pp})\ W_p = 0$

$S_{pp}{}^{-1}\ (B_{pp}\ -\ L\ S_{pp})\ W_p = S_{pp}{}^{-1}\ 0$ ← premultiplicar $S_{pp}{}^{-1}$

$(S_{pp}^{-1} B_{pp} - S_{pp}^{-1} L S_{pp}) W_p = 0$ ← distribuir S_{pp}^{-1}

$(S_{pp}^{-1} B_{pp} - L S_{pp}^{-1} S_{pp}) W_p = 0$ ← L es escalar

$(S_{pp}^{-1} B_{pp} - L I_{pp}) W_p = 0$ ← $S_{pp}^{-1} S_{pp} = I_{pp}$ (matriz ident.)

$S_{pp}^{-1} B_{pp} W_p - L I_{pp} W_p = 0$ ← distribuir W_p

$S_{pp}^{-1} B_{pp} W_p - L W_p = 0$ ← $I_{pp} W_p = W_p$

De esta manera llegamos precisamente a la pauta del problema de valor propio: $R_{pp} A_p - L A_p = 0$. De aquí buscamos el valor propio L y el vector propio Wp a partir de la matriz cuadrada $S_{pp}^{-1} B_{pp}$.

Por otra parte de [11], se deduce que el valor propio L corresponde a la Ratio de Correlación RC, de la manera siguiente:

$(B_{pp} - L S_{pp}) W_p = 0$ ← [11]

$W_p^T (B_{pp} - L S_{pp}) W_p = W_p^T 0$ ← premultiplicar W_p^T

$W_p^T B_{pp} W_p - W_p^T L S_{pp} W_p = 0$ ← desarrollar

$W_p^T B_{pp} W_p - L W_p^T S_{pp} W_p = 0$ ← mover escalar L

$Sb - L S = 0$ ← $W_p^T B_{pp} W_p = Sb$, $W_p^T S_{pp} W_p = S$

$Sb = L S$ ← mover L S a la derecha

$L = Sb / S$ ← Sb / S = RC (Ratio de Correlación)

Hay libros donde se explica la Ratio de Correlación en forma de raíz cuadrada, puesto que la forma anterior se ha derivado de las variaciones en cifras elevadas al cuadrado. Para distinguir los dos, utilizamos el nuevo término «Ratio de Correlación en Raíz» (RCR):

$$RCR = (Sb / St)^{1/2}$$

Std..	Read	Write	Vocab.	POINT	V. esp.	P. est.	Eval.
d1	-1.414	.194	.000			-1.090	Ok
d2	-.707	1.166	.500			-.297	Ok
d3	.000	-1.748	1.500	v	v	1.088	Ok
d4	.707	-.291	-1.500			-.408	Ok
d5	1.414	.680	-.500	v	v	.707	Ok

En la tabla anterior (Std), la columna de «P. est.» (Puntos estandarizados) corresponde al vector Zn de [2].

En la columna de V. esp (Valor esperado), figura v cuando el valor de P.est es positivo. En la columna de Eval. (Evaluación), figura Ok cuando POINT y V.esp. coinciden tanto positiva como negativamente.

En la tabla siguiente de Var. (Variables), el Peso es el vector propio Wp, seguidos de Suma, M (Media) y DT (Desviación Típica):

Var.	Read	Write	Vocab.
Peso	.761	-.070	.644
Sum	40.000	38.000	25.000
M.	8.000	7.600	5.000
DT	1.414	2.059	2.000

T. eval.	RP	RCR.
Valor	1.000	.927

Finalmente en la tabla anterior derecha, damos la evaluación general con la «Ratio de Precisión» (RP), que se deriva del número de Ok dividido el número total y la «Ratio de Correlación en Raíz» RCR.

* Hemos consultado Mino (157-161) e Ishii (2014: 140-149).

(*) Discriminante desconocido

Al aplicar el vector de los pesos W_p derivado de los datos X_{np} a los datos D_{np} cuyo Discriminante es desconocido preparamos la matriz de puntos estandarizados (Ynp) utilizando la Media y Desviación Típica de X_{np}, y la postmultiplicamos Wp:

$$Y_{np} = [D_{np} - M(X_{np})] / Sd(X_{np})$$
$$E_n = Y_{np} W_p$$

(#) Cuantificación de tipo II

Se llama «Cuantificación de tipo II» al análisis discriminante realizado con el datos de puntos cualitativos como el siguiente:

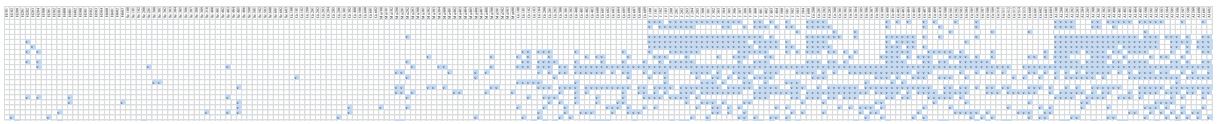
English-5	Read	Write	Vocab.	POINT
d1		v		
d2	v	v	v	
d3	v		v	v
d4	v	v		
d5	v	v		v

(#) Discriminación entre el andaluz oriental y el occidental

La tabla siguiente muestra las frecuencias de rasgos fonéticos observadas en las provincias de Andalucía dividida a priori entre el Occidente (H, SE, CA, MA) y el Oriente (CO, J, GR, AL). Ordenamos los datos de manera descendiente del valor contrastivo entre Oeste y Este:

Ñ1000	H	SE	CA	MA	CO	J	GR	AL	Oeste	Este	Cntr
1602B:disgusto:-sg->x						710	217	600	0	1527	1.000
1660B:unos granos:s=g>x						581	174	233	0	988	1.000
1694C:clavel:-él>ér						32	109	33	0	174	1.000
1663B:las juergas:s=xwe>xwe				38		774	348	833	38	1955	.961
1577A:naranja:-nx->nx	42					839	196	900	42	1934	.958
1647A:las lentejas:-x->x	42					871	217	833	42	1922	.958
1624A:decir:-ír>í+l	83				280	516	413	633	83	1843	.913
1626C:tos:o++				77	280	323	391	400	77	1394	.895
1623A:beber:-ér>é+l	83			38	400	355	413	667	122	1835	.875
1627C:nuez:e++				192	560	581	565	600	192	2306	.846
1632D:los árboles:s=a>a			59	77	80	387	370	767	136	1603	.844
1581C:carne:-rn->ln		65			240	355	65	33	65	693	.830
1695B:claveles:e-es>-e+-e+		32		269	720	774	717	700	301	2912	.812
1631C:los ojos:s=o>o				269	240	742	783	667	269	2431	.801
1620A:mar:-ár>ál	125	32		38	320	452	370	500	196	1641	.787
1616A:árbol:-ol>o+	83	32			240	258	130	200	116	828	.755
1614A:peregil:-íl>í+(:)		32	59	77	320	258	239	267	168	1084	.732
1613A:zagal:-ál>á+(:)	42	97	59		360	194	348	267	197	1168	.711

Utilizando estos rasgos fonéticos, hemos aplicado el Análisis Discriminante a los datos originales de los individuos encuestados.



El resultado es el siguiente:

Var.	Weight	Sum	M.	St.dev.
1613A:zagal:-ál>á+(:)	.147	44.000	.191	.393
1631C:los ojos:s=o>o	.137	92.000	.400	.490
1614A:peregil:-íl>í+(:)	.120	39.000	.170	.375
1635A:las vacas:s=b>ph	.107	41.000	.178	.383
1623A:beber:-ér>é+l	.102	63.000	.274	.446
1581C:carne:-rn->ln	.100	23.000	.100	.300
1620E:mar:-ár>ár	.090	36.000	.157	.363
1627C:nuez:e++	.073	81.000	.352	.478
1620A:mar:-ár>ál	.072	59.000	.257	.437
1632D:los árboles:s=a>a	.072	57.000	.248	.432
1695B:claveles:e-es>-e+-e+	.061	104.000	.452	.498

1602B:disgusto:-sg->x	.040	50.000	.217	.412
1616A:árbol:-ol>o+	.034	29.000	.126	.332
1663B:las juergas:>xwe	.031	66.000	.287	.452
1693A:redes:redes>rede	.028	49.000	.213	.409
1624A:decir:-ír>í+l	.018	63.000	.274	.446
1626C:tos:o++	.015	49.000	.213	.409
1660B:unos granos:s=g>x	.011	33.000	.143	.351
1694C:clavel:-él>ér	.002	7.000	.030	.172
1647A:las lentejas:-x->x	-.003	63.000	.274	.446
1577A:naranja:-nx->nx	-.030	63.000	.274	.446

T. eval.	RP	RCR.
Valor	.943	.910

Observamos que los rasgos principales para discriminar el andaluz oriental del occidental son mayoritariamente las vocales abiertas por efecto de la caída de las consonantes finales de palabra. También merece la atención de un rasgo peculiar: el cambio de [s] + [g] en [x], por ejemplo en *unos granos*.

* Los datos son de Manuel Alvar y Antonio Llorente: *Atlas lingüístico y etnográfico de Andalucía*, 1973.

(*) Análisis de varianza

Se utiliza «Análisis de varianza» para saber el valor significativo de la diferencia de las Varianzas entre las variables (M1, M2, M3):

Xnp	M1	M-2	M-3
A	44	34	33
B	39	29	32
C	42	33	35
D	45	36	32
E	48	30	31

ANOVA	Variation	D.f.	Variance	F.ratio	P. 5%:1%:
Among g,	410.800	2	205.400	28.137	3.885
Within g.	87.600	12	7.300		6.927
All	498.400	14	35.600		0

Para realizar el análisis se buscan la Variación entre variables (Sb), la Variación dentro de cada variable (Sw) y la Variación total (S). El objetivo es calcular la Ratio de Sb en Sw y ver si es significativo estadísticamente.

Primero se calculan el vector de Medias verticales (M_p) y Media total (M):

$$M_p = I_p^T X_{np} / N$$
$$M = \Sigma (X_{np}) / (N * P)$$

Seguidamente calculamos Sb, Sw y S:

$$Sb = N \quad \Sigma \quad (M_p - M)^2$$
$$Sw = \quad \Sigma \quad (X_{np} - M_p)^2$$
$$S = \quad \Sigma \quad (Xnp - M)^2$$

Calculamos los grados de libertad de toda la matriz: N * P - 1. Restamos 1, puesto que si se determina la Suma y Suma menos 1, el restante se determina automáticamente, es decir, no tiene libertad. De misma manera la libertad entre grupos es P – 1; la libertad dentro de grupos es (N – 1) * P. Las Varianzas de cada categoría son:

Varianza total: V = S / (N * P – 1)
Varianza entre grupos: Vb = Sb / (P – 1)
Varianza dentro de grupos: Vw = Sw / [(N – 1) * P]

Y finalmente se obtiene la Ratio Fisher (RF):

RF = Vb / Vw

Si RF excede los criterios anteriormente establecidos, por ejemplo 5% ó 1%, se rechaza la hipótesis nula de que no existe la diferencia de varianzas entre grupos. En la última columna se ofrecen los valores de límite de 5%, 1%, seguidos de la probabilidad de la Ratio Fisher.

5.3.4. Análisis de correspondencia

El «Análisis de correspondencia» es el método desarrollado por Juan-Paul Benzécri. Independientemente Chikio Hayashi descubrió el algoritmo equivalente con los datos cualitativos. El método consiste en buscar los valores de peso tanto en en los individuos (X1, X2, ...) como en las variaables (Y1, Y2, ...) para que la distribución de las frecuencias presenten la mayor correlación posible. Los valores inferidos a los dos ejes sirven para ver sus pesos y también para reordenar la tabla. El resultado nos facilita interpretar de nuevo la distribución de las frecuencias.

D_{np}	Y_1: English	Y_2: Physics	Y_3: Latin	S_n
X_1: Ana	9	14	18	41
X_2: Juan	17	7	11	35
X_3: Mary	15	13	14	42
X_4: Ken	5	18	8	31
T_p	46	52	51	149

El objetivo del método es buscar los dos vectores desconocidos de individuos X_n: (X_1, X_2, X_3, X_4) y de variables Y_p: (Y_1, Y_2, Y_3).

Vamos a imponer una primera condición a los dos vectores: que la Media de $S_i * X_i$ (i=1, 2, …, n) y la de $T_i * Y_i$ (i=1, 2, …, n) sean 0. S_n es vector de sumas horizontales, T_p es vector de sumas verticales, N es escalar de Suma total:

$$S_n = SumR(D_{np});\ T_p = SumV(D_{np});\ N=Sum(D_{np})$$

[1a] Mx

$$= [(9+14+18)*X_1 + (17+7+11)*X_2 + (15+13+14)*X_3 + (5+18+8)*X_4)] / 149$$
$$= (41X_1 + 35X_2 + 42X_3 + 31X_4) / 149 = S_n^T X_n / N = 0$$

[1b] My

$$= [(9+17+15+5)*Y_1 + (14+7+13+18)*Y_2 + (18+11+14+8)*Y_3] / 149$$
$$= (46Y_1 + 52Y_2 + 51Y_3) / 149 = T_p^T Y_p / N = 0$$

La segunda condición es que la Varianzade $S_i * X_i$ (i=1, 2, …, n) y la de $T_i * Y_i$ (i=1, 2, …, n) sean 1. Para formular la restricción preparamos dos matrices diagonales cuyos elementos son sumas horizontales (S_{nn}) y sumas verticales (T_{pp}):

[2a] $Vx = [41(X_1 - Mx)^2 + 35(X_2 - Mx)^2 + 42(X_3 - Mx)^2 + 31(X_4 - Mx)^2 / 149$
$= (41X_1^2 + 35X_2^2 + 42X_3^2 + 31X_4^2) / 149$ ←[1a] $Mx = 0$
$= X_n^T S_{nn} X_n / N = 1$

[2b] $Vy = [46(Y_1 - My)^2 + 52(Y_2 - My)^2 + 51(Y_3 - My)^2] / 149$
$= (46Y_1^2 + 52Y_2^2 + 51Y_3^2) / 149$ ←[1b] $My = 0$
$= Y_p^T T_{pp} Y_p / N = 1$

donde

$$S_{nn} = dg(S_n);\ T_{pp} = dg(T_{1p})\ [dg: matriz\ diagonal]$$

S_{nn}	1	2	3	4
1	41			
2		35		
3			42	
4				31

T_{pp}	1	2	3
1	46		
2		52	
3			51

Considerando a D_{np} como un gráfico de distribución, la correlación (R) entre el eje X (X_n) y el eje Y (Y_p) es:

[3] $R = [9(X_1 - Mx)(Y_1 - My)$

$$
\begin{aligned}
&+ 14(X_1 - Mx)(Y_2 - My) \\
&+ 18(X_1 - Mx)(Y_3 - My) \\
&+ 17(X_2 - Mx)(Y_1 - My) \\
&+ \ldots \\
&+ 8(X_4 - Mx)(Y_3 - My)] / 149 \\
\\
&= (9X_1Y_1 + 14X_1Y_2 + \ldots + 8X_4Y_3) / 149 \qquad \leftarrow Mx = My = 0 \\
&= X_n^T D_{np} Y_p / N
\end{aligned}
$$

El objetivo es buscar los dos vectores X_n e Y_p de cuando R sea el Máximo. Para maximizar R formulamos S con las dos condiciones de Varianzas Vx = 1, Vy = 1, con dos multiplicadores de Lagrange: Lx y Ly[29]. Realizamos el cálculo diferencial de S por X_n e Y_p, cuyos resultados son vector cero (O_n, O_p):

$$S = (X_n^T D_{np} Y_p) / N - Lx [(X_n^T S_{nn} X_n) / N - 1] - Ly [(Y_p^T T_{pp} Y_p) / N - 1]$$

[4a] $Df(S, X_n) = D_{np} Y_p / N - 2 Lx S_{nn} X_n / S = O_n$ (cero)

[4b] $Df(S, Y_p) = D_{np}^T X_n / S - 2 Ly T_{pp} Y_p / S = O_p$ (cero)

[5a] $D_{np} Y_p / S = 2 Lx S_{nn} X_n / S$
← mover el segundo término de [4a] al lado derecho
$X_n^T D_{np} Y_p / S = 2 Lx X_n^T S_{nn} X_n / S$ ← premult. X_n^T a los dos lados
$R = 2 Lx$ ← [3] $R = X_n^T D_{np} Y_p / S$; [2a] $X_n^T S_{nn} X_n / S = 1$

[5b] $D_{np}^T X_n / S = 2 Ly T_{pp} Y_p / S$
← mover el segundo término de [4b] al lado derecho
$X_n^T D_{np} / S = 2 Ly Y_p^T T_{pp} / S$ ← mover matriz ; T_{pp}: matriz diagonal
$X_n^T D_{np} Y_p / S = 2 Ly Y_p^T T_{pp} Y_p / S$ ← postmult. Y_p a los dos lados
$R = 2 Ly$ ← [3] $R = X_n^T D_{np} Y_p / S$; [2b] $Y_p^T T_{pp} Y_p / S = 1$

Por [5a], [5b],

[6] $R = 2 Lx = 2 Ly$

[7a] $D_{np} Y_p = R S_{nn} X_n$ ← [5a] $D_{np} Y_p / S = 2 Lx S_{nn} X_n / S$; [6] $R = 2 Lx$
$R S_{nn} X_n = D_{np} Y_p$ ← cambio de lados
$S_{nn} X_n = D_{np} Y_p / R$ ← mover escalar R
$S_{nn}^{-1} S_{nn} X_n = S_{nn}^{-1} D_{np} Y_p / R$ ← premult. S_{nn}^{-1} a los dos lados
$X_n = S_{nn}^{-1} D_{np} Y_p / R$ ← $S_{nn}^{-1} S_{nn} = I_{nn}$

---

[29] De esta manera, resultan:
$Df(S, Lx) = (X_n^T S_{nn} X_n) / S - 1 = 0$ ← [2a]
$Df(S, Ly) = [(Y_p^T T_{pp} Y_p) / S - 1 = 0$ ← [2b]

[7b] $D_{np}{}^{T} X_n = R\ T_{pp}\ Y_p$ ←[5b] $D_{np}{}^{T} X_n / S = 2\ Ly\ T_{pp}\ Y_p / S$; [6] $R = 2\ Ly$

Introduciendo [7a] a; Xn de [7b],

[8] $D_{np}{}^{T}\ 1/R\ S_{nn}{}^{-1}\ D_{np}\ Y_p = R\ T_{pp}\ Y_p$

$D_{np}{}^{T}\ S_{nn}{}^{-1}\ D_{np}\ Y_p = R^2\ T_{pp}\ Y_p$ ←mover escalar R

$D_{np}{}^{T}\ S_{nn}{}^{-1}\ D_{np}\ (T_{pp}{}^{1/2})^{-1}\ T_{pp}{}^{1/2}\ Y_p = R^2\ (T_{pp})^{1/2}\ (T_{pp})^{1/2}\ Y_p$

←$(T_{pp}{}^{1/2})^{-1}\ T_{pp}{}^{1/2} = I_{pp}$; $(T_{pp})^{1/2}\ (T_{pp})^{1/2} = T_{pp}$ (See infra)

where donde abreviamos por A_p,

[9] $(T_{pp})^{1/2}\ Y_p = A_p$

[8] se hace:

$D_{np}{}^{T}\ S_{nn}{}^{-1}\ D_{np}\ (T_{pp}{}^{1/2})^{-1}\ A_p = T_{pp}{}^{1/2}\ R^2\ A_p$

$(T_{pp}{}^{1/2})^{-1}\ D_{np}{}^{T}\ S_{nn}{}^{-1}\ D_{np}\ (T_{pp}{}^{1/2})^{-1}\ A_p = (T_{pp}{}^{1/2})^{-1}\ T_{pp}{}^{1/2}\ R^2\ A_p$

←両辺に$(T_{pp}{}^{1/2})^{-1}$を左積

$(T_{pp}{}^{1/2})^{-1}\ D_{np}{}^{T}\ S_{nn}{}^{-1}\ D_{np}\ (T_{pp}{}^{1/2})^{-1}\ A_p = R^2\ A_p$ ←$(T_{pp}{}^{1/2})^{-1}\ T_{pp}{}^{1/2} = I_{pp}$

donde abreviando por $(T_{pp}{}^{/2})^{-1}\ D_{np}{}^{T}\ S_{nn}{}^{-1}\ D_{np}\ (T_{pp}{}^{/2})^{-1} = A_{pp}$

$A_{pp}\ A_p = R^2\ A_p$

De esta manera llegamos a la fórmula de valor y vector propios. El programa busca tanto R^2 como A_p al mismo tiempo. Y_p se obtiene por [9],

$Y_p = [T_{pp}{}^{1/2}]^{-1}\ A_p$

El vector Y_p es reducido por ser 0 la suma de productos con S_n es 0 y varianza es 1 por [1a]. Para escalar a la dimensión de los datos, lo multiplicamos por la raíz cuadrada de la Suma total $S^{1/2}$. También conviene multiplicar por el coeficiente de correlación (R, para reflejar la dimensión de importancia[30].

$Y_p' = Y_p * Sum(D_{np})^{1/2} * R_p$

El vector X_n se obtiene por [7a]

$X_n = S_{nn}\ D_{np}\ Y_p / R$

* Hemos consultado a Okumura(1986), Takahashi(2005), Mino(2005)

[30] Takahashi (2005: 127-129).

(*) Elevación de la matriz a 1/2 y a -1/2

Se define $A_{pp}^{1/2}$ de la matriz cuadrada A_{pp} como X_{pp} que tenga relación de

$$X_{pp}\, X_{pp} = X_{pp}^{2} = A_{pp}:$$

$$X_{pp}^{2} = X_{pp}\, X_{pp} = A_{pp},\ X_{pp} = A_{pp}^{1/2}$$

Y si A_{pp} posee una matiz inversa A_{pp}^{-1}, Y_{pp} en la relación $Y_{pp}\, Y_{pp} = A_{pp}^{-1}$ es A_{pp} elevada a -1/2: $A_{pp}^{-1/2}$

$$Y_{pp}^{2} = Y_{pp}\, Y_{pp} = A_{pp}^{-1},\ Y_{pp} = A_{pp}^{-1/2}$$

(*) Matriz inversa de la matriz diagonal

Cuando T_{pp} es una matriz diagonal, los elementos de su matriz inversa T_{pp}^{-1} son valores inversos de T_{pp}:

T_{pp}	1	2	3
1	A		
2		B	
3			C

T_{pp}^{-1}	1	2	3
1	1/A		
2		1/B	
3			1/C

Cuando T_{pp} es una matriz diagonal, los elementos de su matriz elevada a 1/2 $T_{pp}^{1/2}$ son raíces de T_{pp}:

T_{pp}	1	2	3
1	A		
2		B	
3			C

$T_{pp}^{1/2}$	1	2	3
1	$\sqrt{A}$		
2		$\sqrt{B}$	
3			$\sqrt{C}$

Por lo tanto, $(T_{pp}^{1/2})^{-1}$ es:

$(T_{pp}^{1/2})^{-1}$	1	2	3
1	$1/\sqrt{A}$		
2		$1/\sqrt{B}$	
3			$1/\sqrt{C}$

(*) Corrrespondencia entre individuos y variables

La tabla inferior izquierda (D_{np}) es el dato de entrada y la tabla inferior derecha es el vector de peso X_n, correspondientes a los individuos:

D_{np}	Y_1: English	Y_2: Physics	Y_3: Latin
X_1: Ana	9	14	18

CA.d. (X_n)	#1	#2
x1: Ana	-.094	-.211

X_2: Juan	17	7	11
X_3: Mary	15	13	14
X_4: Ken	5	18	8

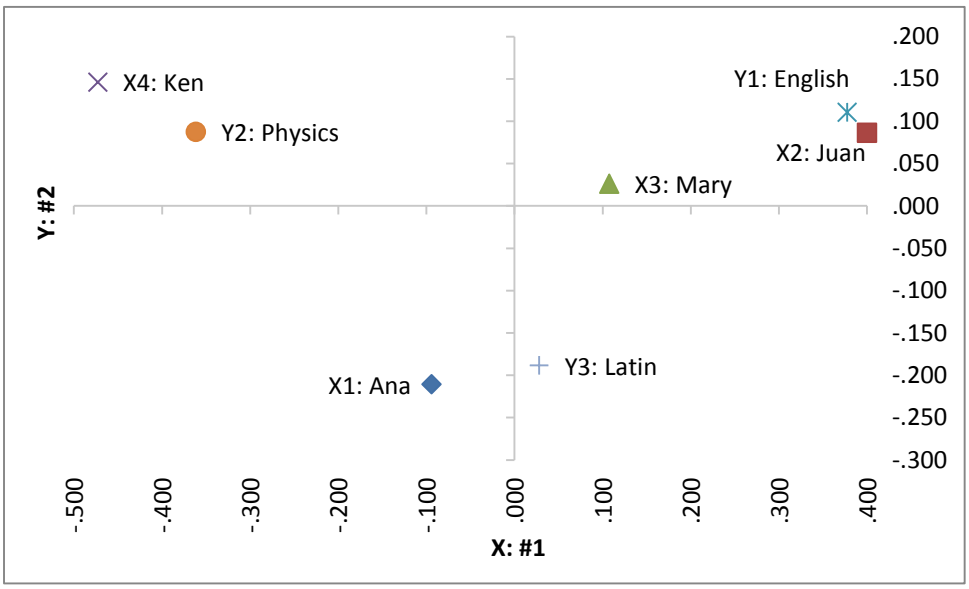
x2: Juan	.400	.086
x3: Mary	.108	.026
x4: Ken	-.473	.146

La tabla inferior izquierda muestra la correlación y la derecha el vector de peso Yp, correspondiente a las variables:

Corresp.	1	2
Correl.	.300	.136

CA.v. (Yp)	#1	#2
y1: English	.377	.110
y2: Physics	-.361	.087
y3: Latin	.028	-.189

El gráfico siguiente muestra la distribución de los individuos y las variables situados en los dos vectores #1 y #2:



Se observa qu hay relación fuerte entre Ken y Physics, Juan y English, Ana y Latin. Mary se encuentra neutral con cierta inclinación a Engish.

(*) Concentración

Se obtiene una distribución concentrada si reordenamos tanto los individuos como variables según los vectores X_n e Y_p. El Análisis de Correspondencia ofrece el mayor grado de correlación, lo que es precisamente su objetivo.

Crp.cct	y2: Latin	y3: Physics	y1: English
x4: Ken	18	8	5
x1: Ana	14	18	9
x3: Mary	13	14	15
x2: Juan	7	11	17

(*) Análisis de correspondencia unilateral

Hemos visto que en el Análisis de Correspondencia se buscan los vectores desconocidos tanto de individuos como de variables. En este lugar proponemos realizar un análisis unilateral donde determinamos uno de los dos vectores y buscamos el otro desconocido. Al vector preestablecido dotamos unos puntos estandarizados de los números secuenciales (1, 2, ..., N o P). Fijando este vector como un criterio externo, buscamos el otro vector desconocido, que también suponemos que su Media es 0 y Varianza 1. Nuestro objetivo es maximizar el coeficiente de correlación, lo mismo que el Análisis de Correspondencia bilateral.

Veamos primero el caso en que fijamos el vector Y_p de las variables y buscamos el vector desconocido de individuos X_n.

Test: D_{np}	Y_1: English	Y_2: Latin	Y_3: Science	S_{n1}
X_1: Ana	9	14	18	41
X_2: Juan	17	7	11	35
X_3: Mary	15	13	14	42
X_4: Ken	5	18	8	31
T_{1p}	46	52	51	149

Pongamos la condición de que las Medias de ambos vectores son 0:

$S_n = SumR(D_{np})$; $T_{1p} = SumV(D_{np})$; $N=Sum(D_{np})$
$S_{nn} = dg(S_n)$; $T_{pp} = dg(T_{1p})$ [dg: matriz simétrica]

[1a] $Mx = (41X_1 + 35X_2 + 42X_3 + 31X_4) / 149 = S_n^T X_n / N = 0$

La varianzas de X_n es 1:

[2] $Vx = X_n^T S_{nn} X_n / N = 1$

La correlación (R) de D_{np} es:

[3] $R = [9(X_1–Mx)(Y_1–My)$
$+ 14(X_1–Mx)(Y_2–My)$
$+ 18(X_1–Mx)(Y_3–My)$
$+ 17(X_2–Mx)(Y_1–My)$

$+\ldots$

$+ 8(X_4 - Mx)(Y_3 - My)] / 149$

$= (9X_1Y_1 + 14X_1Y_2 + \ldots + 8X_4Y_3) / 149$ ←$Mx = My = 0$

$= X_n^T D_{np} Y_p / N$

Para maximizar R, formulamos S con el multiplicador Lagrange L con la condición de [2]

$S = (X_n^T D_{np} Y_p) / N - L [Vx - 1]$

$= (X_n^T D_{np} Y_p) / N - L [(X_n^T S_{nn} X_n) / N - 1]$

La diferencial de S por X_n es O_n (vector cero):

[4] $Df(S, X_n) = D_{np} Y_p / N - 2 L S_{nn} X_n / N = O_n$ (cero)

[5] $D_{np} Y_p / N = 2 L S_{nn} X_n / N$

←mover el segundo término de [4] a la dercha

$X_n^T D_{np} Y_p / N = 2 L X_n^T S_{nn} X_n / N$ ←premult. X_n^T a los dos lados

$R = 2 L$ ←[3] $R = X_n^T D_{np} Y_p / N$; [2] $X_n^T S_{nn} X_n / N = 1$

[6] $D_{np} Y_p = R S_{nn} X_n$

←[5] $D_{np} Y_p / N = 2$ v1 $S_{nn} X_n / N$; [6]. $R = 2$ v1

$R S_{nn} X_n = D_{np} Y_p$ ←cambiar el lado izquierdo y el derecho

$S_{nn} X_n = D_{np} Y_p / R$ ←mover el escalar R

$S_{nn}^{-1} S_{nn} X_n = S_{nn}^{-1} D_{np} Y_p / R$ ←postmult. S_{nn}^{-1} a los dos lados

$X_n = S_{nn}^{-1} D_{np} Y_p / R$ ←$S_{nn}^{-1} S_{nn} = I_{nn}$

De la misma manera que en el análisis de correspondencia bitateral, multiplicamos Xn por R (coeficiente de correlación):

$X_n" = X_n * R = S_{nn}^{-1} D_{np} Y_p / R * R = S_{nn}^{-1} D_{np} Y_p$

Por otra parte, para buscar el vector de variables Y_p, fijando el vector de individuos (X_n), desviamos los procesos a partir de [2] en:

[1b] $My = (46Y_1 + 52Y_2 + 51Y_3) / 149 = T_p^T Y_p / N = 0$

[2b] $Vy = Y_p^T T_{pp} Y_p / N = 1$

[3] $R = X_n^T D_{np} Y_p / N$

Y para maximizar R se formula con el multplicador Lagrange:

$S = (X_n^T D_{np} Y_p) / N - L [Vy - 1]$

$= (X_n^T D_{np} Y_p) / N - L [(Y_p^T T_{pp} Y_p) / N - 1]$

[4b] $Df(S, Y_p) = D_{np}^T X_n / N - 2 L T_{pp} Y_p / N = O_p$ (cero)

[5b] $D_{np}^T X_n / N = 2 L T_{pp} Y_p / N$ ←mover el segundo término de [4b]
$X_n^T D_{np} / N = 2 L Y_p^T T_{pp} / N$ ←$A^T B = B^T A$; T_{pp} matriz diagonal
$X_n^T D_{np} Y_p / N = 2 L Y_p^T T_{pp} Y_p / N$ ←posmult. Y_p a los dos lados
$R = 2 L$ ←[3] $R = X_n^T D_{np} Y_p / N$; [2b] $Y_p^T T_{pp} Y_p / N = 1$

[6b] $D_{np}^T X_n = R T_{pp} Y_p$ ←5b. $D_{np}^T X_n / N = 2 v2 T_{pp} Y_p / N$; 6. $R = 2 v2$
$R T_{pp} Y_p = D_{np}^T X_n$ ←intercambiar los dos lados
$T_{pp} Y_p = D_{np}^T X_n / R$ ←mover escalar R
$T_{pp}^{-1} T_{pp} Y_p = T_{pp}^{-1} D_{np}^T X_n / R$ ←premult. T_{pp}^{-1} a los dos lados
$Y_p = T_{pp}^{-1} D_{np}^T X_n / R$ ←$T_{pp}^{-1} T_{pp} = I_{pp}$

Lo mismo que el caso anterior, por la misma razón, excluimos R:

$Y_p" = Y_p * R = T_{pp}^{-1} D_{np}^T X_n / R * R = T_{pp}^{-1} D_{np}^T X_n$

(#) Diferencia geográfica de rasgos fonéticos andaluces

La tabla siguiente muestra las frecuencias relativas de los casos de la vocal abierta y cerrada en distintas provincias de Andalucía:

R / N * 100	H	SE	CA	MA	CO	J	GR	AL
1533B:miel:el>e+	17	10	9	15	20	29	46	30
1533C:miel:el>e:	11	6	4	16	12	11	16	3
1615A:caracol:-ól>ó+(:)	2	3	3	5	15	14	19	11
1615B:caracol:-ól>ó(:)	18	27	15	16	3	1	6	2
1616A:árbol:-ol>o+	2	1			6	8	6	6
1616B:árbol:-ol>o	23	30	17	26	18	11	23	11
1618A:sol:-ól>ó+(:)	7	9	3	13	13	12	19	11
1618B:sol:-ól>ó(:)	15	21	15	13	1	1	6	1
1623A:beber:-ér>é+l	2			1	10	11	19	20
1623B:beber:-ér>é+	4	7	3	6	13	17	15	8
1623C:beber:-ér>é	19	24	15	19	2		4	
1626A:tos:-ós>ó+h	6	2	2	4	7	13	17	9
1626C:tos:o++				2	7	10	18	12
1626C:tos:-ós>ó+	7	7	5	13	18	17	27	19
1626D:tos:-ós>ó	10	22	11	9	2	1	2	
1627A:nuez:-éθ>é+h	5	2	1		2	3	8	3
1627B:nuez:-éθ>é+	7	13	5	17	20	25	39	26
1627C:nuez:e++				5	14	18	26	18
1627C:nuez:-éθ>é	12	16	12	9	3	1	1	

1629A:voz:-óθ>óh	5	3		1	1	1	5	3
1629B:voz:-óθ>ó+	3	5	3	12	22	30	44	30
1629C:voz:-óθ>ó	18	23	14	13	2	1	2	1
1689A:niños:-os>-o+	1	2		4	22	31	44	30
1689B:niños:-os>oh[os)	4	1			2	3	8	8
1690A:pared:-éd>é+	6	8		10	17	19	24	11
1693A:redes:redes>rede	3	2	1	1	4	12	8	18
1693B:redes:redes>re+	14	6	4	12	3	6	16	6
1693C:redes:redes>reh	1		2		1	4	7	
1694A:clavel:-él>-él	3	2	1	3	6	5	11	15
1694B:clavel:-él>é+,		6	3	15	20	24	40	29
1694C:clavel:-él>ér						1	5	1
1695A:claveles:e-es>-e-e+		2		4	2	2	4	3
1695B:claveles:e-es>-e+-e+		1		7	18	24	33	21
1695C:claveles:-e-es>-e-e:		1		3	1	2	1	1
1695D:claveles:e-es>-e-eh	3	1			5	4	9	5

Realizamos el Análisis de correspondencia unilateral, con un criterio exterior fijo: provincias ordenadas de oeste a este. Nuestro objetivo es buscar el vector de los rasgos fonéticos por el que reordenamos las filas como presentamos en la tabla siguiente:

Cor.R	H	SE	CA	MA	CO	J	GR	AL
1629C:voz:-óθ>ó	18	23	14	13	2	1	2	1
1627C:nuez:-éθ>é	12	16	12	9	3	1	1	
1626D:tos:-ós>ó	10	22	11	9	2	1	2	
1623C:beber:-ér>é	19	24	15	19	2		4	
1618B:sol:-ól>ó(:)	15	21	15	13	1	1	6	1
1615B:caracol:-ól>ó(:)	18	27	15	16	3	1	6	2
1616B:árbol:-ol>o	23	30	17	26	18	11	23	11
1693B:redes:redes>re+	14	6	4	12	3	6	16	6
1629A:voz:-óθ>óh	5	3		1	1	1	5	3
1533C:miel:el>e:	11	6	4	16	12	11	16	3
1627A:nuez:-éθ>é+h	5	2	1		2	3	8	3
1618A:sol:-ól>ó+(:)	7	9	3	13	13	12	19	11
1695C:claveles:-e-es>-e-e:		1		3	1	2	1	1
1623B:beber:-ér>é+	4	7	3	6	13	17	15	8
1690A:pared:-éd>é+	6	8		10	17	19	24	11
1533B:miel:el>e+	17	10	9	15	20	29	46	30
1626C:tos:-ós>ó+	7	7	5	13	18	17	27	19

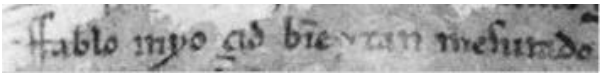
1695A:claveles:e-es>-e-e+		2		4	2	2	4	3
1627B:nuez:-éθ>é+	7	13	5	17	20	25	39	26
1626A:tos:-ós>ó+h	6	2	2	4	7	13	17	9
1693C:redes:redes>reh	1		2		1	4	7	
1615A:caracol:-ól>ó+(:)	2	3	3	5	15	14	19	11
1695D:claveles:e-es>-e-eh	3	1			5	4	9	5
1689B:niños:-os>oh[os)	4	1			2	3	8	8
1616A:árbol:-ol>o+	2	1			6	8	6	6
1694A:clavel:-él>-él	3	2	1	3	6	5	11	15
1629B:voz:-óθ>ó+	3	5	3	12	22	30	44	30
1694B:clavel:-él>é+,		6	3	15	20	24	40	29
1693A:redes:redes>rede	3	2	1	1	4	12	8	18
1695B:claveles:e-es>-e+-e+		1		7	18	24	33	21
1689A:niños:-os>-o+	1	2		4	22	31	44	30
1627C:nuez:e++				5	14	18	26	18
1623A:beber:-ér>é+l	2			1	10	11	19	20
1626C:tos:o++				2	7	10	18	12
1694C:clavel:-él>ér						1	5	1

Las altas frecuencias en la sección superior izquierda muestran que en las provincias de la Andalucía occidental (H, SE, CA, MA), se detectan casos de no abertura vocálica, mientras que la sección inferior derecha se reúnen altas frecuencias correspondientes a los casos de la vocal abierta en las provincias orientales: CO, J, GR. AL. Se trata de una tendencia general y, por supuesto, se encuentran numerosos casos en las secciones restantes. No obstante, son relativamente pocos en comparación con las secciones mencionadas.

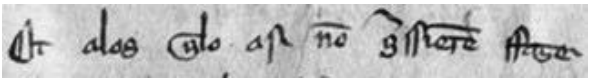
* Origen de datos: Manuel Alvar y Antonio Llorente: *Atlas lingüístico y etnográfico de Andalucía*, 1973.

(#) Cronología relativa de las letras en documentos notariales del español medieval y moderno.

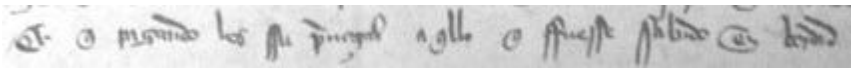
Las fotocopias siguiente muestran distintos tipos de letras utilizadas en las obras literarias y documentos notariales:



(a) Cid, 1207, Letra gótica libraria {7} *ffablo myo çid bien e tan meſurado*



(b) CODEA:0287, Madrid, 1340, Letra de albalaes
{14} *Et a los que lo aſi non quiſſieren ffazer*



(c) CODEA:3931, Madrid 1386, Letra gótica cortesama
{31} *E que pagando los ſſu prinçipal aquello que ffueſſe ſabido en verdad*

Se enumeran otros tipos de letras que suman a 14. La tabla siguiente muestra la cronología de los tipos de letras colocadas en orden alfabético en el eje horizontal:

Letra	Carolina	Cort	De al.	De priv	G. c.	G. c. al	G. c. p.	G. lib	G. r.	Gótica	H. c	H. r.	Prec	Proc.
1075	1													
1100	2													
1125	1													
1150	7							1		1				
1175	2							5	1					
1200	3							14	5	10				
1225	3			1				48	5	12				
1250	1		2	4	14			45	4	8				
1275			14	5	32	22		13	3	1				
1300			4		46	1		8						
1325			1		29			3		1				
1350			1		25			3					5	
1375		6			7		3						6	
1400		4			2			1		1		1	20	
1425		12							1	1			3	
1450		30							2					1
1475		9							3	1		1		
1500		20							4			1	3	
1525		8							1		1	1		2
1550		1									3	11		
1575		1									2	3		2
1600											4	3		
1625											8			1
1650											4			
1675											7			

Fijando la cronología en el eje vertical, realizamos el análisis de correspondencia unilaterial para obtener la distribución diagonal. Hemos obtenido el resultado del nuevo orden de letras de la manera siguiente:

Letra	(1) Carolina	(2) G. lib	(3) De priv	(4) Gótica	(5) G. c. al	(6) De al.	(7) G. c.	(8) G. r.	(9) G. c. p.	(10) Prec	(11) Cort	(12) H. r.	(13) Proc.	(14) H. c
1075	1													
1100	2													
1125	1													
1150	7	1		1										
1175	2	5						1						
1200	3	14		10				5						
1225	3	48	1	12				5						
1250	1	45	4	8		2	14	4						
1275		13	5	1	22	14	32	3						
1300		8			1	4	46							
1325		3		1		1	29							
1350		3				1	25			5				
1375							7		3	6	6			
1400		1		1			2			20	4	1		
1425				1				1		3	12			
1450								2			30		1	
1475				1				3			9	1		
1500								4		3	20	1		
1525								1			8	1	2	1
1550											1	11		3
1575											1	3	2	2
1600												3		4
1625													1	8
1650														4
1675														7

Letra: (1) Carolina, (2) Gótica libraria, (3) De privilegios, (4) Gótica, (5) Gótica cursiva [albalaes], (6) De albalaes, (7) Gótica cursiva, (8) Gótica redonda, (9) Gótica cursiva [precortesana], (10) Precortesana, (11) Cortesana, (12) Humanística redonda, (13) Procesal, (14) Humanística cursiva

*Materiales: CODEA +2015 «Corpus de Documentos Españoles Anteriores a 1700» (Pedro Sánchez Prieto Borja, GITHE: (Grupo de Investigación de Textos para la Historia del Español, Universidad de Alcalá). Contiene 1502 textos provenientes de distintas regiones de España de l siglo XI al XVII.

(#) Área concentrada: Palabras del concepto 'agricultor' en Latinoamérica.

La figura siguiente muestra el resultado del Análisis Correspondiente Bilateral de las las palabras del concepto 'agricultor' en Latinoamérica (origen: Cahuzac 1980)[31]:

[Concent.]	PA	UR	AR	BO	CH	PE	EC	CU	MX	RD	PN	PR	CO	C5	VE
16 changador	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
18 chuncano	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
20 estanciero	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
12 comparsa	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
41 piona	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
03 camilucho	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
04 campero	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
21 gaucho	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0
39 partidario	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0
28 invernador	1	1	1	0	1	1	0	0	0	0	0	0	0	0	0
46 viñatero	1	1	1	0	1	1	0	0	0	0	0	0	0	0	0
43 rondín	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0

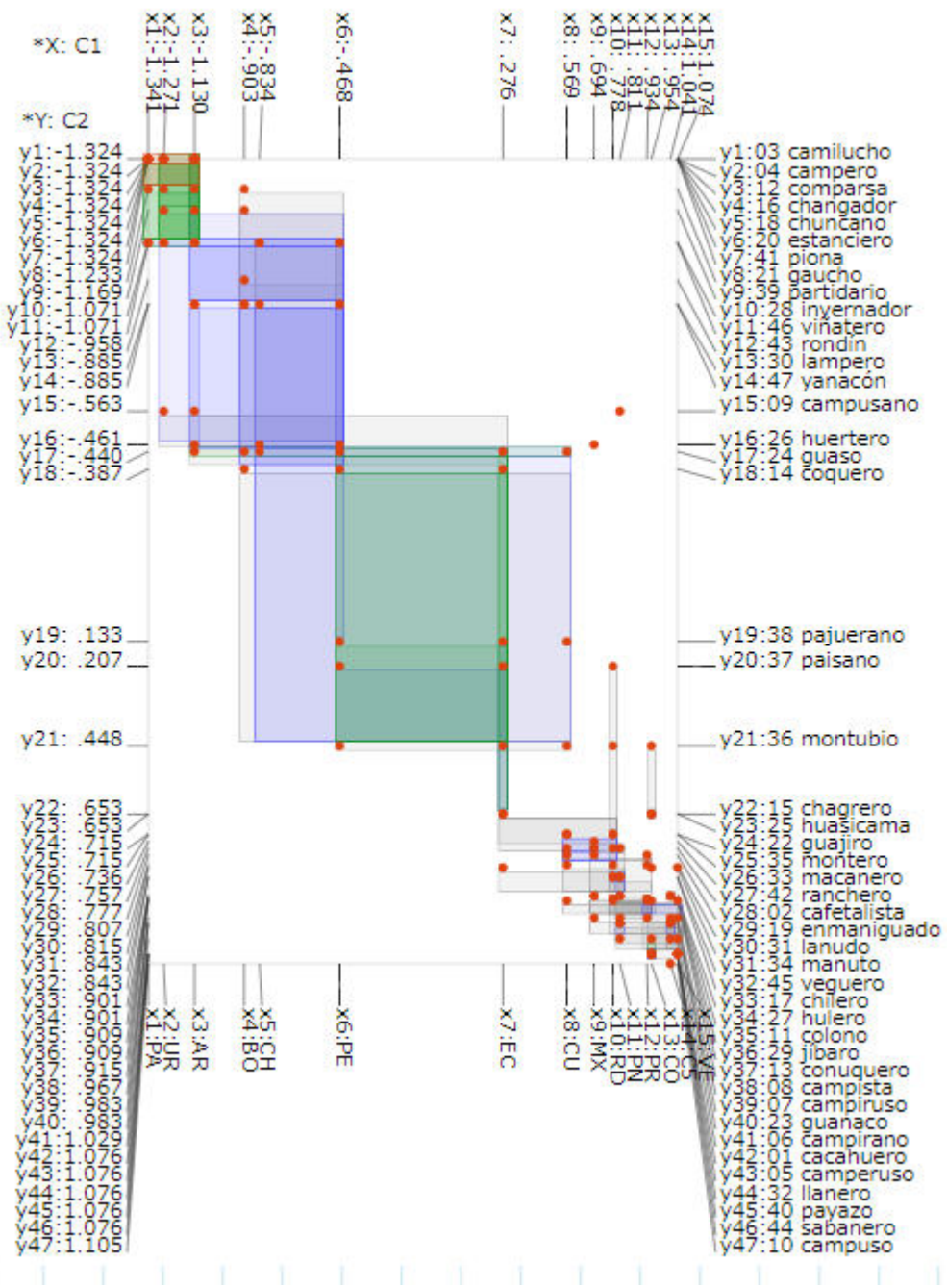
En esta figura observamos que las palabras 16, 18, ..., 21 se concentra en PA (Paraguay), UR (Uruguay), AR (Argentina). Denominamos «Área Concentrada» a la zona donde se reúnen reacciones positivas. En la tabla de distribución concentrada, encontramos varias Áreas Concentradas alrededor de la línea diagonal. Por el programa buscamos, a partir de un punto determinado, los puntos sucesivos inferiores del lado derecho para encontrar el punto que ofrezca la mayor resultado de la multiplicación de media y significatividad y ordenamos la tabla de manera descendente:

[31] Cahuzac, Philippe. (1980) "La División del español de América en zonas dialectales: Solución etnolingüística o semántico-dialectal." *Lingüística Española Actual,* 10, pp. 385-461.

C1	A1	C2	A2	X1	Y1	X2	Y2	Suma	Total	Media	ProEx
03 camilucho	PA	20 estanciero	AR	1	1	3	6	18	18	1.000	.847
04 campero	PA	41 piona	AR	1	2	3	7	18	18	1.000	.847
12 comparsa	PA	21 gaucho	AR	1	3	3	8	18	18	1.000	.847
16 changador	PA	21 gaucho	AR	1	4	3	8	15	15	1.000	.819
03 camilucho	UR	20 estanciero	AR	2	1	3	6	12	12	1.000	.779
04 campero	UR	41 piona	AR	2	2	3	7	12	12	1.000	.779
12 comparsa	UR	21 gaucho	AR	2	3	3	8	12	12	1.000	.779
16 changador	UR	39 partidario	AR	2	4	3	9	12	12	1.000	.779
18 chuncano	PA	21 gaucho	AR	1	5	3	8	12	12	1.000	.779
18 chuncano	UR	28 invernador	AR	2	5	3	10	12	12	1.000	.779
20 estanciero	UR	46 viñatero	AR	2	6	3	11	12	12	1.000	.779
20 estanciero	PA	46 viñatero	AR	1	6	3	11	17	18	.944	.762
41 piona	UR	46 viñatero	AR	2	7	3	11	10	10	1.000	.741
24 guaso	PE	36 montubio	EC	6	17	7	21	10	10	1.000	.741
41 piona	PA	46 viñatero	AR	1	7	3	11	14	15	.933	.721
21 gaucho	UR	46 viñatero	AR	2	8	3	11	8	8	1.000	.688

Por ejemplo, el caso de 03:*camilucho* se sitúa en la coordenada (1, 1) en la tabla de distribución concentrada y, por lo tanto, la coordenada (1, 1) va a ser el punto de partida. El programa llega a la coordenada (3, 6) y la zona cercada entre (1, 1) y (3, 6) ofrece la suma de reacciones 18 y el recuento total de casos: 3 * 6 = 18. Ahora la media es 18 / 18 = 1, que es el valor máximo. La probabilidad expectativa es .847. Al reordenar la tabla, asignamos la primera clave en ProEx, la segunda en Media (Mean) y la tercera en el número de casos (Count).

El gráfico siguiente muestra las primeras 170 áreas concentradas. Cada área representa las relaciones que hay entre casos y atributos.



5.3.5. Análisis factorial

El método del «Análisis factorial» (AF) está basado en un principio opuesto al del «Análisis de componentes principales» (ACP). En ACP se busca unos ejes que expliquen la mayor cantidad de variables posibles en la matriz de datos, mientras que en el AF se busca el vector cuyos miembros se diferencien unos de otros de la manera máxima. Por ejemplo, en la matriz de puntuaciones, el vector que explique cómo las puntuaciones de español e inglés se diferencian del vector de matemáticas y ciencias. El primero es la variable compuesta de ciencias humanas y el segundo, de ciencias exactas. Dentro de multitud de

métodos de AF, explicamos el «Método varimax directo» de Horst, siguiendo a Shiba (1975:90-103).

El objetivo es maximizar la varianza del vector desconocido factorial (Ap) para el cual los factores (A_1, A_2, …, A_p) se distancian en el grado máximo. En la fórmula siguiente, en lugar de varianza utilizamos variación (V*) para simplificar el número de datos (N). M es Media de los elementos de A_p y P es el número de variables:

$$\begin{aligned} V^* &= \Sigma (A_i - M)^2 \\ &= \Sigma (A_i^2 - 2\,M\,A_i + M^2) \\ &= \Sigma A_i^2 - 2\,M\,\Sigma A_i + P\,M^2 \\ &= \Sigma A_i^2 - 2\,(\Sigma A_i)^2 / P + P\,(\Sigma A_i)^2 / P^2 \qquad \leftarrow\ M = (\Sigma A_i) / P \\ &= \Sigma A_i^2 - (\Sigma A_i)^2 / P \end{aligned}$$

Lo expresamos en matrices (ver más adelante):

$$V^* = A_p^{\,T} (I_{pp} - I_p\, I_p^{\,T} / P)\, A_p$$

Para que los valores negativos mantengan su signo, en su lugar elevamos al cuadrado los miembros de Ap que simbolizamos con $A_p^{(2)}$:

[1] $$V^{**} = A_p^{(2)\,T} (I_{pp} - I_p\, I_p^{\,T} / P)\, A_p^{(2)}$$

Introduciendo la matriz diagonal,

$$A_{pp} = \begin{bmatrix} A_1 & & & \\ & A_2 & & \\ & & \ldots & \\ & & & A_p \end{bmatrix}$$

la fórmula [1] se hace:

[1b] $$V^{**} = A_p^{\,T} A_{pp} (I_{pp} - I_p\, I_p^{\,T} / P)\, A_{pp} A_p$$ ← ver más adelante

El vector factorial Ap es el producto de Rpp postmultiplicado por Tp, desconocido. Extipulamos que la longitud de T_p sea 1:

[2] $$A_p = R_{pp}\, T_p$$

[2b] $$T_p^{\,T} T_p = 1$$

Para maximizar V** con la condición de [2b] preparamos una formula W con el multiplicador de Lagrange L:

$$\begin{aligned} W &= V^{**} - L\,(T_p^{\,T} T_p - 1) \\ &= A_p^{\,T} A_{pp} (I_{pp} - I_p\, I_p^{\,T} / P)\, A_{pp} A_p - L\,(T_p^{\,T} T_p - 1) \qquad \leftarrow [1b)] \end{aligned}$$

$$= T_p^T R_{pp}^T A_{pp} (I_{pp} - IpIp^T / P) A_{pp} R_{pp} T_p - L (T_p^T T_p - 1) \quad \leftarrow [2]$$

El resultado del cálculo diferencial de W por el desconocido Tp es 0:

$$Df(W, T_p) = 2 [R_{pp}^T A_{pp} (I_{pp} - I_p I_p^T / P) A_{pp} R_{pp} T_p - L T_p] = 0$$

[3] $R_{pp}^T A_{pp} (I_{pp} - I_p I_p^T / P) A_{pp} R_{pp} T_p = L T_p$ ← mover $L T_p$ al lado derecho

lado izq. $= R_{pp}^T A_{pp} (I_{pp} - I_p I_p^T / P) A_{pp} A_p$ ← [2] $A_p = R_{pp} T_p$

$= R_{pp}^T (A_{pp} I_{pp} A_{pp} A_p - A_{pp} I_p I_p^T A_{pp} A_p / P)$ ←R_{pp}^T al exterior

$= R_{pp} (A_{pp} A_p^{(2)} - A_p A_p^T A_p / P)$

←R_{pp} es simétrica; $A_{pp} I_{pp} = A_{pp}$; $A_{pp} A_p = A_p^{(2)}$; $I_p^T A_{pp} = A_p^T$

$= R_{pp} (A_p^{(3)} - A_p A_p^T A_p / P)$ ←$A_{pp} A_p^{(2)} = A_p^{(3)}$

Por lo tanto [3] se convierte en [3b]:

[3b] $R_{pp} (A_p^{(3)} - A_p A_p^T A_p / P) = L T_p$

donde establecemos:

[4] $B_p = A_p^{(3)} - A_p A_p^T A_p / P$

de modo que [3b] se hace:

[3c] $R_{pp} B_p = L T_p$

El vector derivado de la fórmula siguiente con R_{pp} y su vector de carga W_p se llama «vector estructural» (ver más adelante):

$$R_{pp} W_p / (W_p^T R_{pp} W_p)^{1/2}$$

En lugar de W_p utilizamos B_p de [4] y utilizamos su vector estructural como Ap:

[5] $A_p = R_{pp} B_p / (B_p^T R_{pp} B_p)^{1/2}$

[6] $B_p = A_p^{(3)} - A_p A_p^T A_p / P$ ←[4]

De esta forma no podemos calcular ni A_p ni B_p que satisfagan [5] y [6] al mismo tiempo, puesto que B_p del lado derecho de [5] se estipula en [6] y A_p del lado derecho de [6] se estipula en [5], de modo que ambas fórmulas dependen una de otra. En realidad en el programa, preparamos B_p con datos provisionales y se calcula A_p en [5] y se calcula B_p en [6]; seguidamente se calcula Ap de nuevo en [5] y B_p en [6], así sucesivamente hasta que no presente cambio significativo en A_p. De esta manera A_p y B_p llegan a sus valores

correspondientes que satisfacen [5] y [6]. A_p es el primer vector factorial.

Al encontrarse el primer factor, se calcula la matriz restante de R_{pp} y se repite los mismos procesos para encontrar el segundo. También se buscan otros factores según el interés del analista. En cada proceso se calculan los puntos que correspondientes a los individuos con datos estandarizados Z_{np} y B_p.

(*) Elevación de los elementos de matriz

Entendemos las operaciones desarrollando sus elementos:

[1] $$V^* = \Sigma A_i^2 - (\Sigma A_i)^2 / P = A_p^T (I_{pp} - I_p I_p^T / P) A_p$$

Comprobamos los elementos del lado derecho:

$$A_p^T (I_{pp} - I_p I_p^T / P) A_p$$

$$= A_p^T \left(\begin{bmatrix} 1 & & & \\ & 1 & & \\ & & \ldots & \\ & & & 1 \end{bmatrix} - \begin{bmatrix} 1 \\ 1 \\ \ldots \\ 1 \end{bmatrix} [1, 1, \ldots, 1] / P \right) A_p$$

$$= A_p^T \left(\begin{bmatrix} 1 & & & \\ & 1 & & \\ & & \ldots & \\ & & & 1 \end{bmatrix} - \begin{bmatrix} 1 & 1 & \ldots & 1 \\ 1 & 1 & \ldots & 1 \\ \ldots & \ldots & \ldots & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix} / P \right) A_p$$

Simbolizando M = 1 / P

$$V^* = A_p^T \left(\begin{bmatrix} 1 & & & \\ & 1 & & \\ & & \ldots & \\ & & & 1 \end{bmatrix} - \begin{bmatrix} M & M & \ldots & M \\ M & M & \ldots & M \\ \ldots & \ldots & \ldots & M \\ M & M & M & M \end{bmatrix} \right) A_p$$

$$= [A_1, A_2, \ldots, A_p] \begin{bmatrix} 1-M & -M & \ldots & -M \\ -M & 1-M & \ldots & -M \\ \ldots & \ldots & \ldots & \ldots \\ -M & -M & \ldots & 1-M \end{bmatrix} A_p$$

$$= [A_1(1 - M) + A_2(-M) + \ldots + A_p(-M),$$
$$A_1(-M) + A_2(1 - M) + \ldots + A_p(-M),$$
$$\ldots$$
$$A_1(-M) + A_2(-M) + \ldots + A_p(1 - M)] \begin{bmatrix} A_1 \\ A_2 \\ \ldots \\ A_p \end{bmatrix}$$

$$= [A_1(1 - M) + A_2(-M) + \ldots + A_p(-M)] A_1$$
$$+ [A_1(-M) + A_2(1 - M) + \ldots + A_p(-M)] A_2$$
$$+ \ldots$$

$$+ [A_1(-M) + A_2(-M) + \ldots + A_p(1 - M)]\ A_p$$

$$= A_1^2 + A_2^2 + \ldots + A_p^2 - M\ (A_1 + A_2 + \ldots + A_p)^2$$

$$= A_1^2 + A_2^2 + \ldots + A_p^2 - (A_1 + A_2 + \ldots + A_p)^2 / P \quad \leftarrow M = 1 / P$$

$$= \Sigma A_i^2 - (\Sigma A_i)^2 / P \quad = V^*$$

[2] $V^{**} = A_p^{(2)\,T} (I_{pp} - I_p\ I_p^T / P)\ A_p^{(2)} = A_p^T\ A_{pp} (I_{pp} - I_p\ I_p^T / P)\ A_{pp}\ A_p$

Comprobamos los elementos de $A_p^T\ A_{pp}$; y de $A_{pp}\ A_p$:

$$A_p^T\ A_{pp} = [A_1, A_2, \ldots, A_p] \begin{bmatrix} A_1 & & & \\ & A_2 & & \\ & & \ldots & \\ & & & A_p \end{bmatrix} = [A_1^2, A_2^2, \ldots, A_p^2] = A_p^{(2)\,T}$$

$$A_{pp}\ A_p = \begin{bmatrix} A_1 & & & \\ & A_2 & & \\ & & \ldots & \\ & & & A_p \end{bmatrix} \begin{bmatrix} A_1 \\ A_2 \\ \ldots \\ A_p \end{bmatrix} = \begin{bmatrix} A_1^2 \\ A_2^2 \\ \ldots \\ A_p^2 \end{bmatrix} = A_p^{(2)}$$

(*) Vector estructural

La matriz de correlación los datos estandarizados es (donde N = número de datos):

[1] $R_{pp} = Z_{np}^T\ Z_{np} / N$

Obtenemos vector variable F_n por la multiplicación de $Z_{np}\ W_p$:

[2] $F_n = Z_{np}\ W_p$

La varianza de F_n es (La Media de F_n es 0):

[3] $V(F_n) = F_n^T\ F_n / N$
$= (Z_{np}\ W_p)^T\ Z_{np}\ W_p / N$
$= W_p^T\ Z_{np}^T\ Z_{np}\ W_p / N$
$= W_p^T\ R_{pp}\ W_p$

El vector estandarizado de F_n es G_n:

[4] $G_n = F_n / V(F_n)^{1/2} = Z_{np}\ W_p / (W_p^T\ R_{pp}\ W_p)^{1/2}$ $\leftarrow$[2], [3]

Ap es vector de correlación entre Z_{np} y G_n:

$A_p = Z_{np}^T\ G_n / N$
$= Z_{np}^T\ Z_{np}\ W_p / (W_p^T\ R_{pp}\ W_p)^{1/2} / N$ $\leftarrow$[4]
$= R_{pp}\ W_p / (W_p^T\ R_{pp}\ W_p)^{1/2}$ $\leftarrow$[1]

* Hemos consultado a Shiba (1975). Este autor denomina a Ap «vector estructural» y llama la atención sobre su importancia. En nuestra derivación del vector factorial, Bp coresponde a Wp:

(#) Experimento sobre simbolismo fonético

La tabla inferior izquierda muestra las impresiones individuales sobre la sensación psicológica de una sílaba en la escala de -3 a +3 en cinco categorías: Grande, Agudo, Claro, Duro, Grave. A su derecha damos el resultado del análisis factorial en sus puntos y variables (abajo):

Ss	Grande	Agudo	Claro	Duro	Grave
1.pa	-1	2	2	2	-3
2.ba	2	-3	-3	1	2
3.ta	-1	2	1	2	-2
4.da	2	-1	-3	1	2
5.ka	0	3	1	3	-1
6.ga	3	-2	-3	2	3
7.sa	-2	2	2	1	-2
8.za	2	-1	-2	0	3
9.ma	0	-1	-1	-2	0
10.na	0	-1	0	-2	0
11.ra	1	-2	2	-3	0

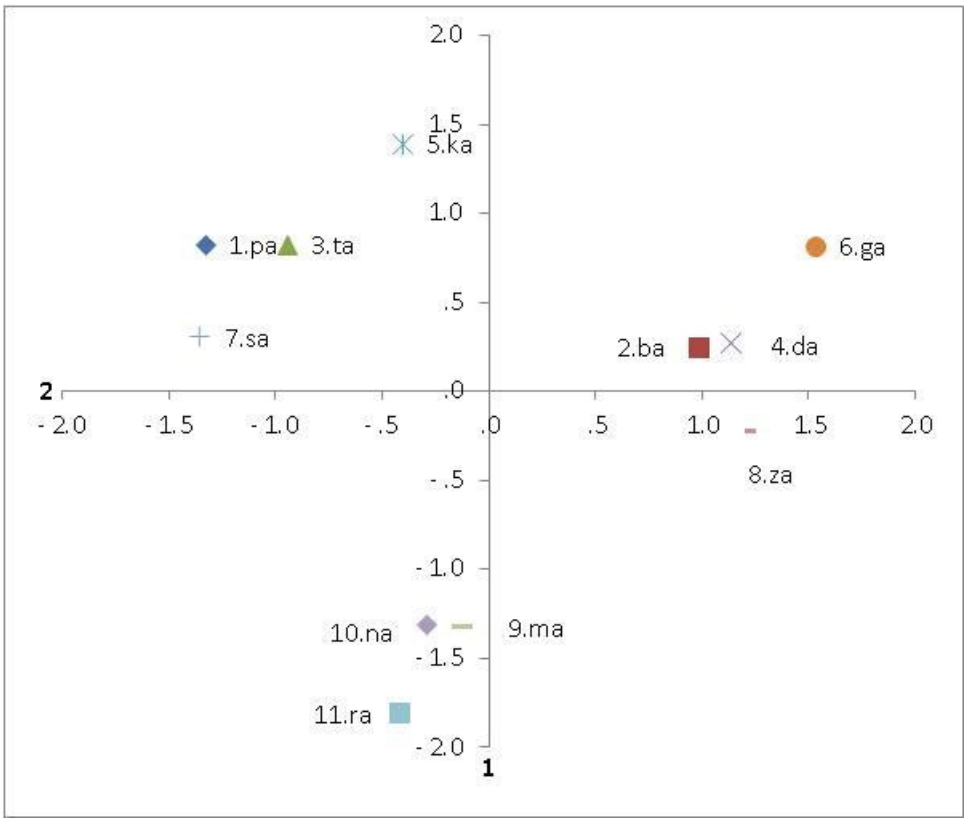
FA.d.	#1	#2	#3	#4
1.pa	-1.329	.823	-.733	.196
2.ba	.989	.246	-2.171	-.667
3.ta	-.946	.823	-.057	-.463
4.da	1.134	.274	.664	-.761
5.ka	-.400	1.390	1.430	.912
6.ga	1.535	.813	-.684	.708
7.sa	-1.355	.305	-.138	-.397
8.za	1.203	-.221	1.416	.522
9.ma	-.127	-1.328	.601	-1.663
10.na	-.286	-1.311	.257	-.589
11.ra	-.419	-1.815	-.585	2.202

FA.v.	#1	#2	#3	#4
Grande	.960	-.063	-.106	.217
Agudo	-.728	.557	.399	.004
Claro	-.940	-.063	.072	.327
Duro	.008	1.000	-.013	-.015
Grave	.979	-.150	.016	.048

El primer factor (#1) reacciona fuertemente con Grande y Grave, de modo que puede representar «Fuerza» mientras que el segundo factor reacciona con Duro y Agudo, que simboliza «Cortante». Las correlaciones entre factores son lógicamente nulas:

Correlation	1	2	3	4
1	1.000	.000	.000	.000
2	.000	1.000	.000	.000
3	.000	.000	1.000	.000
4	.000	.000	.000	1.000

En el gráfico de distribución de las sílabas con el factor #1 en el eje horizontal y #2 vertical. En el factor #1 se contrastan entre sonidos sonoros y sordos. En la dimensión vertical correspondiente al factor #2, el contraste es entre sonidos explosivos y nasales-líquidos.



(*) Concentración

Reordenando tanto los individuos (sílabas) como las variables (impresiones) se obtiene la figura siguiente, donde se muestra la distribución concentrada por el análisis factorial:

nota: los valores positivo se muestran en color azul y los negativos en naranja

Fct.cct	c. Clear	b. Sharp	d. Hard	a. Big	e. Heavy
7.sa	2	2	1	-2	-2
1.pa	2	2	2	-1	-3
3.ta	1	2	2	-1	-2
11.ra	2	-2	-3	1	0
5.ka	1	3	3	0	-1
10.na	0	-1	-2	0	0
9.ma	-1	-1	-2	0	0
2.ba	-3	-3	1	2	2
4.da	-3	-1	1	2	2
8.za	-2	-1	0	2	3
6.ga	-3	-2	2	3	3

5.3.6. Análisis de cluster

(En preparación)

5.4. Análisis de asociación

Es una técnica que se utiliza en el comercio de venta en internet. Tenemos la experiencia de pedir, por ejemplo, un libro A por internet. Al seleccionar el libro A, recibimos algunas recomendaciones de otros libro B y C. ¿En qué datos se basa este sistema de recomendación? Es un sistema sencillo basado en los datos almacenados de las otras ventas anteriores. Vamos a ver como se hace el cálculo de valores que se utiliza en los sitios de venta masiva: internet, supermercado, grandes almacenes, etc.

Estamos ante una matriz pequeña de ejemplo, constituida de 5 filas (d1:5) y 4 columnas (A, B, C, D). En realidad analizamos, por ejemplo, en Varilex más de 9800 filas (formas lingüísticas variantes) con 21 columnas (países)[32]. Sin embargo, siempre es mejor empezar con los datos más sencillos posibles. Una vez entendida la base, ya es cuestión de ampliar las dimensiones.

Dat	A	B	C	D
d1	1	1	0	0
d2	0	0	1	0
d3	0	1	0	0
d4	0	0	1	1
d5	1	1	1	0

En primer lugar calculmos «Support» (Soporte) entre A y B, que es la razón de coincidencias al respecto a la totalidad (número) de datos: 5. Entre A y B encontramos 2 coincidencias (d1 y d5), donde observamos el valor 1 tanto en A como en B. De modo que Soporte entre A y B es:

Soporte(A, B) = 2 / 5 = .400

Consideramos que cuanto mayor es el Soporte, tanto mayor el grado de asociación entre A y B. El máximo valor es 1.000. Es lo mismo que la «Correspondencia simple», que hemos aprendido anteriormente. Es sime/trico, es decir, siempre resulta «Soporte»(A, B) = «Soporte»(B, A).

En segundo lugar, se busca el grado de «Confidence» (Confianza), que es la probabilidad condicionada de A => B. Se trata de la probabilidad de

[32] http://lecture.ecc.u-tokyo.ac.jp/~cueda/varilex/

seleccionar B, a la hora de seleccionar A. Se calcula la coincidencia de A y B (=2), dividida por las veces de A (=2):

Confianza(A, B) = 2 / 2 = 1.000

Esta cifra corresponde a la probabilidad de comprar B, de los que han comprado A. No debemos pensar que la «Confianza» es vice versa entre A y B. A se confía mucho en B, pero no tanto B en A, puesto que:

Confianza(B, A) = 2 / 3 = 0.667

Es decir, el producto B tiene más compradores que el producto A. Casi dos tercios de los compradores de B, compran A, mientras que la totalidad de los de A compra B. El valor ma/ximo de «Confianza» es 1.000, como hemos visto en «Confianza»(A, B).

Y el «Análisis de asociación» no termina con estos dos valores, sino toma en cuenta otro valor (último) que es también significativo: «Lift» (Elevación). Ahora se compara la «Confianza»(A, B) con respecto a la probabilidad de B:

Elevación(A, B) = Confianza(A, B) / Probabilida(B)
= (2 / 2) / (3 / 5) = 1 / 0.6 = 1.667

Se trata del grado (la ratio) de contribución de la compra de A en la compra de B. La cantidad de B se obedece a la probabilidad de B, que se obtiene naturalmente por la división de la frecuencia de B por la totalidad (3 / 5 = .600), donde todavía no interviene el factor A. Y ahora nos interesa la Confianza (A, B) comparada con la probabilidad de B, que corresponde al grado de contribución de la venta de A en la de B, porque B, desde el principio, tiene cierta probabilidad (.600) y este valor se eleva en la «Confianza»(A, B). Es decir, cuanto mayor la «Elevación», tanto mayor el grado de contribución de A en B. La «Elevación» tampoco es bidireccional, sino unidireccional: A => B, es decir la «Elevación»(A, B) es necesariamente igual a la «Elevación»(B, A).

Hemos visto que se han tomado en consideración las tras cifras indicadoras de asociación: «Soporte», «Confianza» y «Elevación». Hay que considerar estas tres cifras al mismo tiempo y, por esta razón, proponemos una cifra más de «Sythesis» (Síntesis) en forma de multiplicación de las tres anteriores:

Síntesis = Soporte * Confianza * Elevación

El siguiente cuadro representa la dirección de influencia con LHS (Left Hand Side: lado de mano izquierda) y RHS (Right Hand Side: lado de mano

derecha): LHS => RHS. L.f. y R.f. son las frecuencias de cada columna. L.p. y R.p. son probabilidades de LHS y RHS, que se obtienen por la división de frecuencia por el número de datos. Cooc. es la coocurrencia de Lhs y Rhs. Seguidamente el programa calcula Suport (Soporte), Confid.(Confianza) y Lift (Elevación). Hemos ordenado la tabla en Synt., Lift, Cnf. y Sup. en este orden manera descendiente (de mayor a menor):

#	L	R	LHS	=>	RHS	L.f.	R.f.	L.p.	R.p.	Cooc.	Sup.	Cnf.	Lift	Synt.
1	1	2	A	=>	B	2	3	.400	.600	2	.400	1.000	1.667	.667
2	2	1	B	=>	A	3	2	.600	.400	2	.400	.667	1.667	.444
3	4	3	D	=>	C	1	3	.200	.600	1	.200	1.000	1.667	.333
4	3	4	C	=>	D	3	1	.600	.200	1	.200	.333	1.667	.111
5	1	3	A	=>	C	2	3	.400	.600	1	.200	.500	.833	.083
6	3	1	C	=>	A	3	2	.600	.400	1	.200	.333	.833	.056
7	2	3	B	=>	C	3	3	.600	.600	1	.200	.333	.556	.037
8	3	2	C	=>	B	3	3	.600	.600	1	.200	.333	.556	.037
9	1	4	A	=>	D	2	1	.400	.200	0	.000	.000	.000	.000
10	2	4	B	=>	D	3	1	.600	.200	0	.000	.000	.000	.000
11	4	1	D	=>	A	1	2	.200	.400	0	.000	.000	.000	.000
12	4	2	D	=>	B	1	3	.200	.600	0	.000	.000	.000	.000

Cuando el dato es grande, el cálculo de todos estos valores va a ser laborioso, de modo que necesitamos elaborar el programa para la computación. Y, lo más importante, estos valores se obtienen en todas las combinaciones de parámetros. A-B, A-C, A-D, B-C, B-D, C-D. Y ahora en lugar de los parámetros simples, también nos interesan los compuestos. AB -C, AB-D, BC-D, etc.

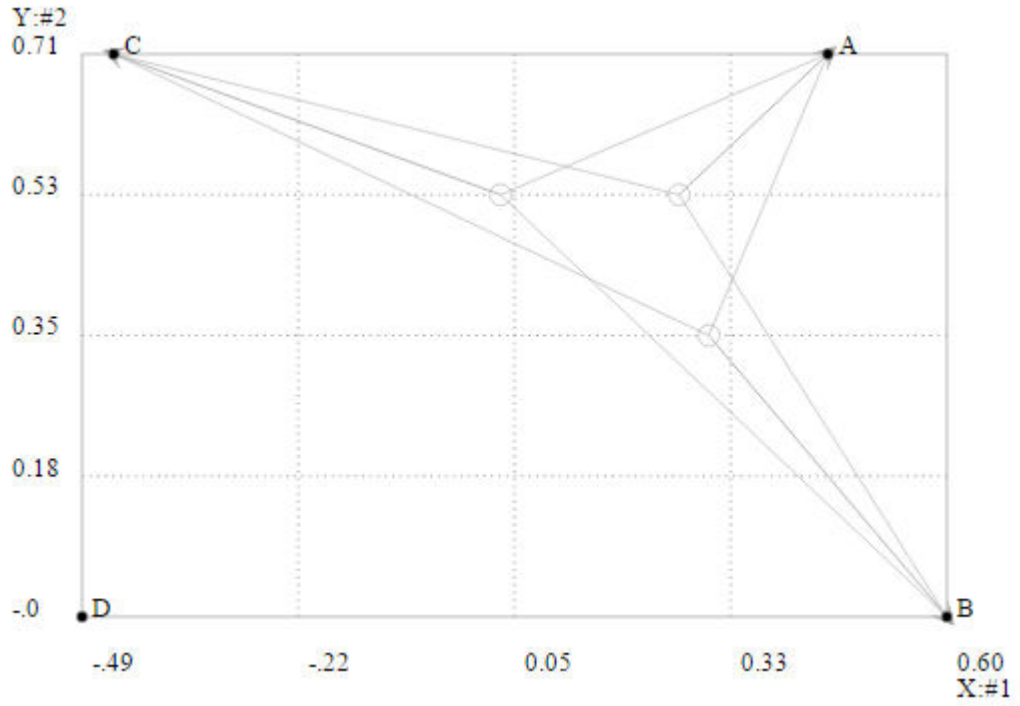
#	L.	R.	LHS	=>	RHS	L.f.	R.f.	L.p.	R.p.	Cooc.	Sup.	Cnf.	Lift	Synt.
1	2:3	1	B:C	=>	A	1	1	.200	.200	1	.200	1.000	5.000	1.000
2	1:3	2	A:C	=>	B	1	2	.200	.400	1	.200	1.000	2.500	.500
3	1:2	3	A:B	=>	C	2	3	.400	.600	1	.200	.500	.833	.083
4	1:4	3	A:D	=>	C	0	1	.000	.200	0	.000	.000	.000	.000
5	1:2	4	A:B	=>	D	2	1	.400	.200	0	.000	.000	.000	.000
6	1:3	4	A:C	=>	D	1	0	.200	.000	0	.000	.000	.000	.000
7	1:4	2	A:D	=>	B	0	1	.000	.200	0	.000	.000	.000	.000
8	2:4	1	B:D	=>	A	0	0	.000	.000	0	.000	.000	.000	.000
9	2:4	3	B:D	=>	C	0	1	.000	.200	0	.000	.000	.000	.000
10	2:3	4	B:C	=>	D	1	0	.200	.000	0	.000	.000	.000	.000
11	3:4	1	C:D	=>	A	1	1	.200	.200	0	.000	.000	.000	.000
12	3:4	2	C:D	=>	B	1	2	.200	.400	0	.000	.000	.000	.000

Estas combinaciones se explotan con el número de parámetros, por ejemplo 21 países hispanohablantes: (21 * 20) = 420. Y nos perdemos en el resultado voluminoso con tras cifras indicadores de asociación. De ahí se hable de la «minería de datos», con el objetivo de buscar unas piedras preciosas dentro de la cantidad de materiales recogidos.

Ante una situación complicada de relaciones, es útil unos métodos de visualización. Se han propuesto a dibujar unos gráficos de redes con líneas que combinan entre los parámetros en relación dotada de más asociación. De nuestra parte, hemos pensado utilizar el método del Análisis de Componentes Principales, que nos ofrece los ejes que garantizan la agrupaciones de parámetros en el espacio bidimensional no correlato:

E.mat.	#1	#2	#3	#4
A	.449	.707	.317	-.445
B	.599	-.000	.230	.767
C	-.449	.707	-.317	.445
D	-.489	.000	.864	.122

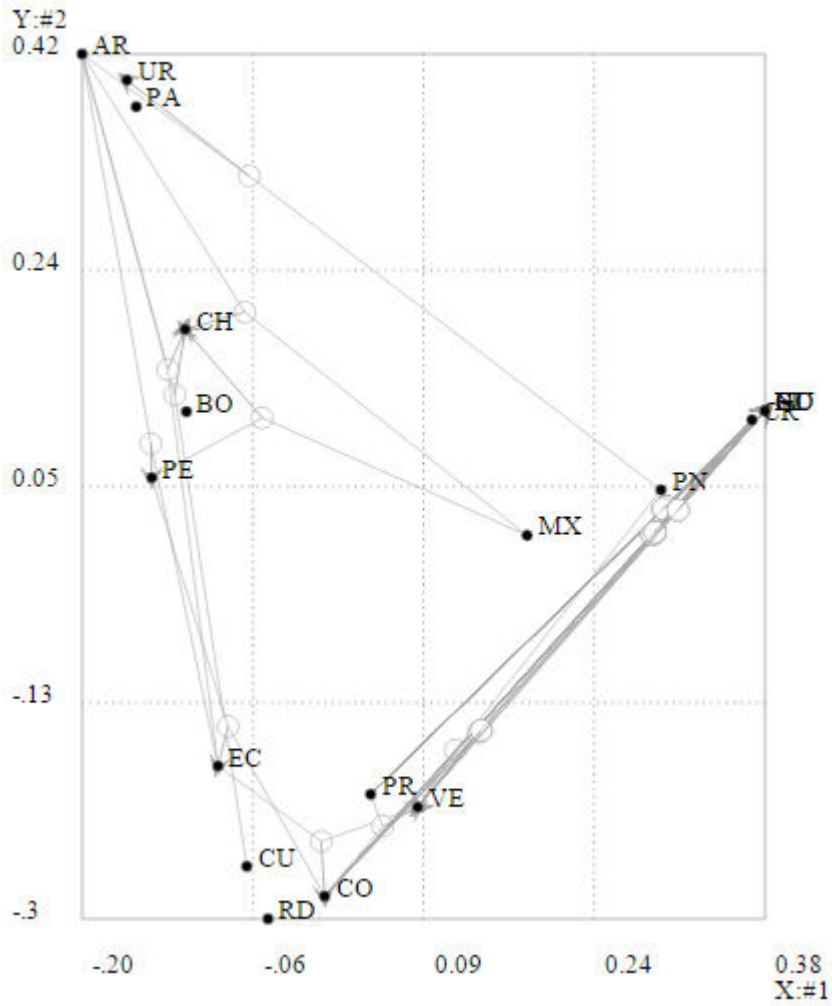
Utilizando estos valores para los ejes X y Y para situar los parámetros A, B, C, D, de acuerdo con los valores de vectores eigen.



En realidad no reproducimos todos los trazos de relaciones, sino nos limitamos a exponer vínculos más importantes (con las menores distancias). En el gráfico de 21 países de Varilex observamos algunos trazos entre los países situados con distancia considerable. Es debido a que los países están situados en

el espacio de tan solo dos dimensiones (X e Y). En realidad tenemos 21 vectores eigen, es decir, podemos imaginarnos en el espacio multidimencional (de 21 dimensiones). De ahí convenga dibujar las conexiones en forma de líneas entre los parámetros, que son países hispanohablantes.

Con los datos de Cahuzac (1980) hemos realizado el «Análisis de asociación» y hemos obtenido el siguiente gráfico de redes:



http://lecture.ecc.u-tokyo.ac.jp/~cueda/varilex/ (2016/2/26)

Bibliografía

Cahuzac, Philippe. (1980) "La División del español de América en zonas dialectales: Solución etnolingüística o semántico-dialectal." *Lingüística Española Actual,* 10, pp. 385-461.

Michael Hahsler, Bettina Grun, Kurt Hornik, Christian Buchta. "Introduction to arules – A computational environment for mining association rules and frequent item sets"
https://lyle.smu.edúIDA/arules/ (24/2/2016)

5.5. Análisis de agrupamiento

En esta sección analizamos la relación entre los datos y la columna que indica el grupo que pertenece el dato y seguidamente intentamos clasificar los datos cuyos grupos son incógnitos. Lamamos este método el «Análisis de agrupamiento» para distinguirlo del método de Regresión múltiple que analiza los datos con una columna numérica exterior y del Análisis discriminante que analiza los datos con una columna binominal exterior. El Análisis de agrupamiento trata los datos con una columna multinominal exterior.

5.5.1. Agrupamiento por distancia

Por la distancia calculada entre los vectores horizontales inferimos el Grupo a los datos cuyo grupo es desconocido basándonos en los datos con la columna de grupo sabido. Dentro de la matriz D, buscamos los datos (d1, d2, ...) o el grupo de datos (a, b, c) más cercanos al dato X:

D	v1	v2	v3	Grupo
d1	5	2	7	a
d2	3	3	2	b
d3	2		2	b
d4	4	2	2	c
d5	2	4	3	c
d6	1	8	7	c

X	v1	v2	v3
x1	4	2	5
x2	3	7	6

Se calcula la distancia del vector d1 y del x1 por la distancia euclídica (Dist):

$$Dist(d1, x1) = \{ \Sigma i \, [D_{np}(1, i) - X_{np}(1, i)]^2 \}^{1/2}$$

Los elementos de d1 son (5, 2, 7) y los de x1 son (4., 2, 5), de modo que se calcula la distancia de la manera siguiente:

$$Dist(d1, x1) = [(5 - 4)^2 + (2 - 2)^2 + (7 - 5)^2]^{1/2}$$
$$= (1^2 + 0^2 + 2^2)^{1/2} = 5^{1/2} \fallingdotseq .236$$

Realizamos el mismo cálculo también con d2, d3, ... y el vector que ofrece el Mínimo de las distancias infiere el nombre de su grupo a x1. Hacemos lo mismo también con x2. El resultado es el siguiente:

Dt.g	v1	v2	v3	Grp.i	Val.
x1	4	2	5	d1:a	2.236
x2	3	7	6	d6:c	2.449

(*) Agrupamiento por media del grupo

En lugar de los valores que tienen los datos individuales, utilizamos también la Media del los miembros de cada grupo:

D	v1	v2	v3
a	5.000	2.000	7.000
b	2.500	1.500	2.000
c	2.333	4.667	4.000

El resultado de:

Dt.g	v1	v2	v3	Grp.i	Val.
x1	4	2	5	a	2.236
x2	3	7	6	c	3.145

También podemos hacer lo mismo con Mediana, Mitad y Media mayor. La tabla siguiente es el resultado de agrupamiento por Media mayor:

Dt.g	v1	v2	v3	Grp.i	Val.
x1	4.0	2.0	5.0	a	2.236
x2	3.0	7.0	6.0	c	3.446

(*) Agrupamiento por distancia estandarizada

La tabla siguiente contiene una columna de valores con Media y Varianza muy diferentes de las otras. Esta columna sesga considerablemente la distancia:

D2	v1	v2	v3	Grupo
d1	5	2	56	a
d2	3	3	33	b
d3	2		21	b
d4	4	2	22	c
d5	2	4	45	c
d6	1	8	72	c

X2	v1	v2	v3
x1	4	2	50
x2	3	7	60

Para evitar el sesgo, estandarizamos la matriz de manera siguiente (Inp es la matria input, M es Media, Dt es desviación típica):

$$X_{np} = [I_{np} - M(I_{np})] / Dt(I_{np})$$

D2	v1	v2	v3
d1: a	1.633	-.588	.649
d2: b	.000	-.196	-.693
d3: b	-.816	-1.373	-1.393
d4: c	.816	-.588	-1.335
d5: c	-.816	.196	.007
d6: c	-1.633	1.765	1.583

Resultado:

Dt.g	v1	v2	v3	Grp.i	Val.
x1	.816	-.588	.299	a	.888
x2	.000	1.373	.883	c	1.330

(*) Agrupamiento por distancia Mahalanobis

Utilizando la distancia Mahalanobis, podemos prescindir no solamente la Media y la Varianza de cada variable (columna), sino también la magnitud Covarianza entre las variables:

D2	#1	#2	#3
d1: a	-.493	1.811	-.907
d2: b	-.389	-.397	.750
d3: b	-1.009	-1.636	-1.030
d4: c	-1.095	-.031	1.681
d5: c	.341	-.732	-.277
d6: c	2.010	-.485	-.454

Dt.g	v1	v2	v3	Grp.i	Val.
x1	-.399	.866	-.981	a	.953
x2	1.036	.603	1.218	c	1.493

(*) Revisión del dato de input

La tabla siguiente es el resultado del análisis de agrupamiento por la Distancia estandarizada y Media del mismo dato D2. Al calcular la distancia con el grupo por Media, a veces nos encontramos con los datos cuyo grupo inferido es discordante de su grupo inicial, concretamente d4. En la tabla inferior derecha están representadas las cifras de Pos[isito] que es número de los casos concordantes que da «Ok», Neg[ativo], el de los casos negativos que da «No» y

RP (Ratio de precisión) que es Pos / N (número de datos):

Dt.g	v1	v2	v3	Disc.	Grp.i	Eval.	Val.
d1	1.633	-.588	.649	a	a	Ok	.000
d2	.000	-.196	-.693	b	b	Ok	.797
d3	-.816	-1.373	-1.393	b	b	Ok	.797
d4	.816	-.588	-1.335	c	b	No	1.274
d5	-.816	.196	.007	c	c	Ok	.385
d6	-1.633	1.765	1.583	c	c	Ok	2.267

Grp.	Pos.	Neg.	RP.
Val.	5.000	1.000	.833

De esta tabla sabemos que «d4» se acerca más al grupo «b», a pesar de que pertenece al grupo «c». Efectivamente el vector horizontal de «d4» es más parecido a los miembros del grupo «b». Si la clasificación se ha hecho sin criterio razonable, se puede realizar una reclasificación. Llamamos «Precategorización» al método rígido que respeta el criterio preestablecido, y «Postcategorización» al método flexible que permite una nueva clasificación basada en el análisis numérico. En los estudios lingüísticos, se adopta más veces el primer método. Sin embargo, con el método de Postcategorización nos acercamos más a la realidad del los datos librándonos del criterio exterior preestablecido.

5.5.2. Agrupamiento por probabilidad

Existe un método en que se utiliza la probabilidad en forma de frecuencia relativa dentro del grupo. Se la aplica con una fórmula bayesiana (Thomas Bayes) para encontrar su grupo que ofrece máxima probabilidad. El método se llama «Clasificador bayesiano ingenuo»

Basándose en los datos como Q, se intenta inferir el grupo de los datos de como Y:

Q	v1	v2	v3	Grp
d1	v	v		a
d2	v		v	a
d3	v			a
d4		v		a
d5	v		v	b
d6		v	v	b

Y	v1	v2	v3
x1	v	v	
x2			v

La tabla siguiente es la de probabilidades de la ocurrencia de «v» dentro de cada grupo. Se calcula con la frecuencia dividida por el número de casillas, por ejemplo la probabilidad de «v» del grupo «a» es (3 + 1 / (4 + 2) ≒ .667. Se

agrega 1 al numerador y 2 (número de grupos) del denominador. Lo explicamos posteriormente:

Likel.	v1	v2	v3
a	.667	.500	.333
b	.500	.500	.750

D.data	v1	v2	v3	Disc.	Grp.i	Eval.	mx:mn
d1	v	v		a	a	Ok	.711
d2	v		v	a	b	No	.006
d3	v			a	a	Ok	.711
d4		v		a	a	Ok	.495
d5	v		v	b	b	Ok	.006
d6		v	v	b	b	Ok	.339

Grp.	Pos.	Neg.	AR
Val.	5.000	1.000	.833

D.pred	v1	v2	v3	Grp.i	mx:mn
x1	v	v		a	.711
x2			v	b	.339

donde la probabilidad del vector «x1» (1, 1, 0) en el grupo «a», y otra de x1 en «b» son:

$$P(X=a|Y=x1) = (4/6) * (.667) * (500) * (1 - .333)$$
$$P(X=b|Y=x1) = (2/6) * (.500) * (.500) * (.750)$$

En el último término de P(X=a|Y=x1), invertimos la probabilidad de .333 en 1 - .333, puesto que así se representa la probabilidad de «no» ocurrencia de «v» en v3. De esta fórmula explicaremos más adelante. Aquí simplemente notamos que en todos los cálculos anteriores se hacen multiplicaciones, donde no conviene utilizar la cero probabilidad. De modo que se hace la operación de probabilidad antes mencionada: Se agrega 1 al numerador y 2 (número de grupos) del denominador.

La última tabla da dos valores: Ct (mx, mn) es un valor contrastivo entre el Máximo y el Mínimo de todas las probabilidades que se han calculado entre grupos. El último «Grupo» es el nombre del grupo inferido.

D.pred	Ct(mx, mn)	Grupo
x1	.711	a
x2	.339	b

(*) Teorema de Bayes

Se calculan la probabilidad de que X e Y ocurren al mismo tiempo:

$P(X, Y) = P(X)\ P(Y|X)$
$P(X, Y) = P(Y)\ P(X|Y)$

La primera fórmula P(X, Y) representa la probabilidad coocurrente (conjunta), que es la probabilidad de X, P(X), multiplicada por la probabilidad de Y con la condición de que ocurre, X P(Y|X). Por ejemplo, supongamos que X es diamante de naipes, e Y número «7». La probabilidad de diamante es: P(X) = 1/4 y la de «7» dentro de las naipes de diamante es P(Y|X) = 1/13. Entonces, la probabilidad de diamante «7» P(X,Y) es P(X) = 1/4 multiplicado por P(Y|X) 1/13 = 1 / (4 * 13). La segunda fórmula se interpreta de la misma manera. Como las dos lleva la misma fórmula en el lado izquierdo, se deriva:

$P(X)\ P(Y|X) = P(Y)\ P(X|Y)$

de donde llegamos al «Teorema de Bayes»:

$P(X|Y) = P(X)\ P(Y|X) / P(Y)$

Como este teorema es importante, explicamos un ejemplo de aplicación. La tabla siguiente muestra los documentos recogidos en dos regiones A y B. Se han encontrado 4 documentos en la región A y 13 en la región B, en total 17. La columna de P(X) muestra la probabilidad de P(X=A) 4/17 y P(X=B) 13/17, respectivamente. En la columna P(Y|X) están las frecuencias relativas de un fenómenos lingüístico, por ejemplo, caída de la vocal final, dentro de los documentos encontrados. En los 4 documentos se han encontrado 3 documentos de la región A con la caída vocálica, mientras que en la región B, 5 entre 13:

X	P(X)	P(Y|X)	P(X) P(Y|X)	P(X) P(Y|X) / P(Y) = P(X|Y)
A	4/17	3/4	4/17 * 3/4 = 3/17	(3/17) / (8/17) = 3/8
B	13/17	5/13	13/17 * 5/13 = 5/17	(5/17) / (8/17) = 5/8
Suma	1		8/17 = P(Y)	1

La P(X) se llama «probabilidad anterior» por no considerar la P(Y), que viene después. P(Y|X) se llama «verosimilitud» (ing. *likelihood*) porque muestra la probabilidad dentro de cada grupo (región). Ahora bien, como hemos visto anteriormente, la multiplicación de la probabilidad anteior y la verosimilitud P(X) P(Y|X) es la «probabilidad conjunta» (ing. *joint probability*). Por ejemplo, la probabilidad conjunta de A 3 / 17 muestra la la Ratio de los documentos en que se encuentran las caídas vocálicas. Pasa lo mismo en B: P(X) P(Y|X) = 5/17.

Preste atención al hecho de que en el cálculo de la probabilidad conjunta, el numerador del primer término se repite en el denominador del segundo término. Lo entendemos si consideramos que el numerador en la probabilidad anterior va a ser la base (denominador) en el cálculo de la verosimilitud. Lo hemos expresado todo en fracciones en lugar de decimales, para poder seguir sus procesamientos.

Ahora, la Suma de las dos verosimilitudes (3/17 + 5/17=8/17) es el denominador del Teorema de Bayes P(Y), que es la probabilidad del fenómeno (8) dentro de todos los documentos (17).

Finalmente formulamos P(X|Y) siguiendo al Teorema de Bayes: P(X|Y) = P(X) P(Y|X) / P(Y). Se trata de la probabilidad de X, con la condición de Y. En nuestro caso concreto, si se presenta la caída vocálica, la probabilidad de que el documento sea de la región A es: (3/17) / (8/17) = 3/8; y la de que sea de la región B es: (5/17) / (8/17) = 5/8. Se trata de la Ratio obtenido de la división de la probabilidad conjunta P(X) P(Y|X) por la suma de las dos, que es P(Y).

Cuando se presentan más eventos, por ejemplo, la vocal es <e>, delante de consonante, etc., se multiplica más probabilidades en forma de Y_2, Y_3, ..., Y_p:

$$P(Y|X) = P(Y_1|X)\ P(Y_2|X)\ \ldots\ P(Y_p|X)$$

* Hemos consultado Takamura (2000: 99-117), y Kato, Hamuro y Yada (2008: 111-115).

(#) Precategorización y postcategorización del dialecto andaluz.

Preparamos una matriz de 230 localidades por 164 rasgos fonéticos del dialecto andaluz. La analizamos con el método de «Agrupamiento por probabilidad». Extraemos una pare, de los últimos rasgos en las localidades de la provincia de Huelva. Efectivamente la mayoría de las localidades están agrupadas en la Hueva (H). Sin embargo, hay localidades clasificadas como de Cádiz (Ca) o de Sevilla (Se):

154	155	156	157	158	159	160	161	162	163	164	Grp.d	Grp.i	Eval.
v											H	H	Ok
v											H	H	Ok
v									v	v	H	H	Ok
v									v	v	H	H	Ok
v											H	H	Ok
	v		v						v	v	H	H	Ok
v			v								H	H	Ok
v											H	H	Ok
v											H	H	Ok

			H	H	Ok
v			H	H	Ok
	v		H	H	Ok
			H	H	Ok
		v	H	H	Ok
			H	H	Ok
			H	H	Ok
v			H	Ca	No
v			H	Ca	No
v			H	Ca	No
v			H	Ca	No
	v		H	Ca	No
v			H	H	Ok
	v		H	Se	No
	v		H	Se	No

En total las localidades correctamente agrupadas son 173, que ocupa 75%:

Grp.	Pos.	Neg.	Pres.
Val.	173	57	.752

Fuera del método que se busca el Agrupamiento con los datos individuales, todos los métodos no presentan grupos inferidos coincidentes con los iniciales. El método de Agrupamiento con los datos individuales, en realidad, se trata de indentificación más que de agrupamiento. Solo se busca un individuo más cercano para encontrar el agrupamiento de dato con grupo desconocido. De esta manera, no es adecuado para el análisis lingüístico, aunque reconocemos su utilidad práctica.

Volviendo a la tabla anterior, la columna de «Grp.d» corresponde a la división administrativa, que es una «Precategorización», mientras que la de «Grp.i» es el resultado de la «Poscategorización», facilitada por el Análisis de agrupamiento. En la práctica, después de estudiar los datos con las categorías preestablecidas, podemos tratar los datos con nuevos puntos de vista para ver la división lingüística de la región, aparte de la división administrativa, que no tiene mucho que ver con la lingüística.

5.5.3. Agrupamiento por coocurrencia nominal

La tabla inferior izqueirda posee una matriz nominal con una columna de grupo (Grp). Comparando con esta matriz, intentamos agrupar unos datos cuyos

grupos son desconocidos.

El método más sencillo es buscar la fila que presenta más coincidencias de letras. Llamamos a este método «Agrupamiento por coocurrencia nominal». El análisis de los datos iniciales siempre ofrece lógicamente el agrupamiento coincidente:

D	v1	v2	v3	Grp
d1	A	D	H	a
d2	A	D	I	b
d3	A	F	H	b
d4	A	E	H	c
d5	B	F	G	c
d6	B	F	I	c

Co.g	v1	v2	v3	Grp	Grp.i	Eval.	値
d1	A	D	H	a	d1: a	Ok	1.000
d2	A	D	I	b	d2: b	Ok	1.000
d3	A	F	H	b	d3: b	Ok	1.000
d4	A	E	H	c	d4: c	Ok	1.000
d5	B	F	G	c	d5: c	Ok	1.000
d6	B	F	I	c	d6: c	Ok	1.000

La tabla siguiente es el resultado de agrupamiento de los datos cuyos grupos son desconocidos, a los que han inferido el grupo en «Grp.i»:

X	v1	v2	v3
x1	B	D	J
x2	B	E	H

Co.g	v1	v2	v3	Grp.i	値
x1	B	D	J	d1: a	.333
x2	B	E	H	d4: c	.667

En lugar de individuos, también es posible comparar con la Media del grupo. Por ejemplo, x1 de X coincide con d1 de la matriz D en utilizar la letra «D», que cobra el valor de 1 entre 3: 1 / 3 ≒ .333. También d2 posee la letra «D», pero su valor es 1 dentro de 6.

D	v1	v2	v3	Grp
d1	A	D	H	a
d2	A	D	I	b
d3	A	F	H	b
d4	A	E	H	c
d5	B	F	G	c
d6	B	F	I	c

Co.g	v1	v2	v3	Grp	Grp.i	Eval.	値
d1	A	D	H	a	a	Ok	1.000
d2	A	D	I	b	a	No	.667
d3	A	F	H	b	a	No	.667
d4	A	E	H	c	a	No	.667
d5	B	F	G	c	c	Ok	.556
d6	B	F	I	c	c	Ok	.556

X	v1	v2	v3
x1	B	D	J
x2	B	E	H

Co.g	v1	v2	v3	Grp.i	値
x1	B	D	J	a	.333
x2	B	E	H	c	.444

x2 con el grupo «c», hay 4 coocurrencias B, B, E, H, que son 4 entre 9, 4 / 9 ≒ .444, que es el Ma/ximo entre los tres grupos «a», «b», «c».

5.6. Dispersión lineal

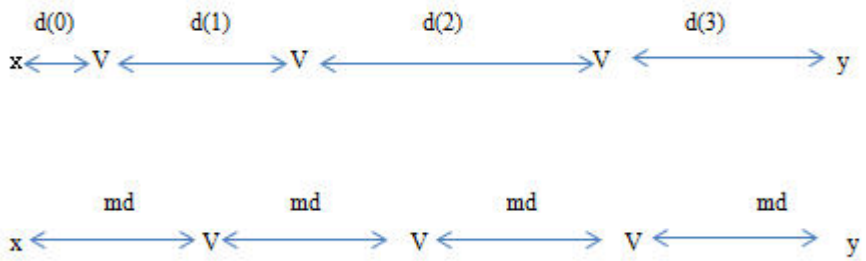
Al indagar las características de los datos, es importante averiguar no solamente su Frecuencia sino también su Dispersión. Cuando se trata de los datos dotados de frecuencia, se practica el cálculo de su Varianza, como hemos visto anteriormente. Aquí tratamos simplemente una larga secuencia de los datos nominales sin más, para ver su dispersión en su aparición dentro de la misma secuencia:

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
…	…

Definimos «Dispersión lineal» (DL), que sirve para ver el grado de dispersión dentro de la secuencia, con la fórmula siguiente:

$$DL = 1 - [\Sigma (d(i) - md)^2 / n)^{1/2} / [(n - 1)^{1/2} * M)$$

donde d(i) representa la distancia que hay entre sus sucesivas apariciones, md es su Media, n es el número (frecuencia) del mismo dato, y M es la Media: $\Sigma (d(i) / n$. Por ejemplo, la palabra «aditivo» aparece en el lugar 7 y 10, de modo que la distancia es 3. La md es la distancia media, que se calcula por la división de la totalidad de datos por el número del dato en cuestión:



Calculamos la diferencia entre d(1) y med, la de d(2) y md, y asi

sucesivamente. Sumamos la primera distancia d(0) con la última d(n), puesto que así podemos evitar el sesgo del cálculo si el conjunto de datos aparece en unos determinados lugares dentro de la secuencia total. Su «Desviación Típica Normalizada» (DTN) oscila entre 0 y 1. La «Dispersión lineal» (DL) es la inversa de la DTN.

Lema	Frec.	Rango.F	F.p.m	DL	Rango.DL	Uso	Rango.U.
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

El programa ofrece la Frecuencia (Frec.), Rango de frecuencia (Rango.F.), Frecuencia por mil (F.p.m), Dispersión lineal (DL), Rango de DL (Rango.DL), Uso (Uso) y Rango de Uso (Rango.U.):

U = Frecuencia * DL
Rango = RndUp (Frec. / Max * 10)

donde RndUp es la función que devuelve el valor subido de cifras decimales, por ejemplo RndUp(1.7) = 2. Max es el Máximo de Frecuencia. De esta manera Rango oscila entre 1 y 10. Los datos de poca frecuencia no son fiables.

5.7. Análisis de Condición Múltiple

En los datos no solamente lingüísticos sino también naturales y sociales en general, hay veces que no una sola condición sino una combinación de condiciones se relaciona con el efecto determinado. En esta sección consideramos un método que permite analizar, por ejemplo, el uso de una forma lingüística en relación con las variables cronológicas, geográficas, sociológicas, estilísticas, etc.

5.7.1. Lista de condición múltiple

Por el método que llamamos «Análisis de condición múltiple»,

analizamos la relación entre las múltiples condiciones (c1, c2, ..., cp) con un efecto (E) en distintas combinaciones de condición:

M	c1	c2	c3	c4	E
d1	A	C	F	I	X
d2	A	D	F	J	X
d3	A	D	G	K	Y
d4	B	D	H	L	Z
d5	B	E	H	M	Z

(1) Condición simple

En primer lugar preparamos listas que muestran la relación entre la(s) condición(es) y el efecto:

MA.c1	X	Y	Z
A	d1	d3	
A	d2		
B			d4
B			d5

c2	X	Y	Z
C	d1		
D	d2	d3	d4
E			d5

c3	X	Y	Z
F	d1		
F	d2		
G		d3	
H			d4
H			d5

c4	X	Y	Z
I	d1		
J	d2		
K		d3	
L			d4
M			d5

Por estas listas sabemos la relación, por ejemplo, entre la condición A y efecto X con los dos datos d1 y d2. Lo mismo se puede observar en distintas condiciones y combinación de condiciones:

(2) Condición doble

MA.c1+c2	X	Y	Z
A + C	d1		
A + D	d2	d3	
B + D			d4
B + E			d5

MA.c1+c3	X	Y	Z
A + F	d1		
A + F	d2		
A + G		d3	
B + H			d4
B + H			d5

MA.c1+c4	X	Y	Z
A + I	d1		
A + J	d2		
A + K		d3	
B + L			d4
B + M			d5

(…)

(3) Condición múltiple

MA.c1+c2+c3+c4	X	Y	Z
A + C + F + I	d1		
A + D + F + J	d2		
A + D + G + K		d3	

B + D + H + L	d4
B + E + H + M	d5

5.7.2. Frecuencia de condición múltiple

Calculamos las frecuencias de datos correspondientes a la relación de condiciones y efectos:

(1) Frecuencia de condición simple

MA.f.c1	X	Y	Z
A	2	1	
B			2

c2	X	Y	Z
C	1		
D	1	1	1
E			1

c3	X	Y	Z
F	2		
G		1	
H			2

c4	X	Y	Z
I	1		
J	1		
K		1	
L			1
M			1

Las cifras de las columnas X, Y, Z son frecuencias absolutas de efectos causados con condiciones correspondientes: A, B, ... M.

(2) Frecuencia de condición doble

Calculamos las frecuencias correspondientes a las combinaciones de dos condiciones relacionadas con los tres efectos:

MA.f.c1+c2	X	Y	Z
A + C	1		
A + D	1	1	
B + D			1
B + E			1

MA.f.c1+c3	X	Y	Z
A + F	2		
A + G		1	
B + H			2

MA.f.c1+c4	X	Y	Z
A + I	1		
A + J	1		
A + K		1	
B + L			1
B + M			1

También calculamos las combinaciones: c2+c3, c2+c4, C3+c4.

(3) Frecuencia de condición triple

Calculamos los casos de la combinaciones de tres condiciones:

MA.f.c1+c2+c3	X	Y	Z
A + C + F	1		
A + D + F	1		
A + D + G		1	
B + D + H			1
B + E + H			1

MA.f.c1+c2+c4	X	Y	Z
A + C + I	1		
A + D + J	1		
A + D + K		1	
B + D + L			1
B + E + M			1

MA.f.c1+c3+c4	X	Y	Z
A + F + I	1		
A + F + J	1		
A + G + K		1	
B + H + L			1
B + H + M			1

MA.f.c2+c3+c4	X	Y	Z
C + F + I	1		
D + F + J	1		
D + G + K		1	
D + H + L			1
E + H + M			1

De la misma manera calculamos la combinación de cuatro condidiones.

MA.f.c1+c2+c3+c4	X	Y	Z
A + C + F + I	1		
A + D + F + J	1		
A + D + G + K		1	
B + D + H + L			1
B + E + H + M			1

Todas las cifras de esta sección son frecuencias absolutas, de modo que necesitamos un método para realizar su relativización.

5.7.3. Coeficiente de condición múltiple

Proponemos pensar en un «Coeficiente de condición múltiple» (CCM) que sirve para calcular el grado de relación existente entre las condiciones y el efecto.

(1) Coeficiente de condición simple

Para el CCM de condición simple, utilizamos el coeficiente de Jaccard con prominencia, que hemos tratado anteriormente.

Condición (c)	Efecto (e)	Peso (p)	Frec. (f)	Grupo
+1 (Sí)	+1 (Sí)	p1:(+1)(+1) = +1	f 1	a
+1 (Sí)	−1 (No)	p2:(+1)(−1) = −1	f 2	b
−1 (No)	+1 (Sí)	p3:(−1)(+1) = −1	f 3	c
−1 (No)	−1 (No)	p4:(−1)(−1) = +1	f 4	d

Calculemos por ejemplo el caso de la condición A con el efecto X: en la tabla siguiente:

MA.f.c1	X	Y	Z
A	2	1	
B			2

MA.t.c1	X	Y	Z
A	.857	.600	
B			1.000

El Grupo «a» es el de condición «+» y efecto «+», que tiene la Frec[uencia] 2. La frecuencia del Grupo «b», de condición «+» y efecto «−» es 1 en A:Y. No se encuentran casos del Grupo «c», de condición «−» y efecto «+». El grupo «d», de condición «−» y efecto «−» corresponde a B:Z = 2. Sin embargo, la fórmula de Jaccard (J), a / (a + b + c), no considera «d»:

$$J = 2 / (2 + 1 + 0) = 2 / 3 \fallingdotseq .666$$

La fórmula de CCM de condición simple, $CCM^{(1)}$, es:

$$CCM^{(1)} = p1*f1*(N+P-2) / [p1*f1*(N+P-2)+abs(p2)*f2 + abs(p3)*f3]$$

: (N: número de filas, P: número de columnas, abs: valor absoluto)

La aplicamos a la distribución siguiente:

MA.f.c1	X	Y	Z
A	f1: 2 (a)	f2: 1 (b)	
B			f4: 2 (d)

MA.t.c1	X	Y	Z
A	.857	.600	
B			1.000

$$CCM^{(1)} = 1*2*(2+3-2) / [1*2*(2+3-2)+1*1+0]$$
$$= 6 / 7 \fallingdotseq .857$$

El $CCM^{(1)}$ de A:Y es:

MA.f.c1	X	Y	Z
A	f2: 2 (b)	f1: 1 (a)	
B			f4: 2 (d)

MA.t.c1	X	Y	Z
A	.857	.600	
B			1.000

$$CCM^{(1)} = 1*1*(2+3-2) / [1*(2+3-2)*2+1*2+0] = 3 / 5 = .600$$

(2) Coeficiente de condición doble

En la condición doble, sumamos los valores de las dos condiciones. Para obtener el peso multiplicamos la suma de condiciones por el efecto. Cuando la suma de condiciones presenta la cifra 0, no pertenece a ningún grupo, es decir, no participa en el cálculo del coeficiente.

Con.(c1)	Con.(c2)	Efecto (e)	Peso (p)	Frec. (f)	Grupo
+1	+1	+1	p1:(+1+1)(+1) = +2	f 1	a
+1	+1	−1	p2:(+1+1)(−1) = −2	f 2	b
+1	−1	+1	p3:(+1−1)(+1) = 0	f 3	—
+1	−1	−1	p4:(+1−1)(−1) = 0	f 4	—
−1	+1	+1	p5:(−1+1)(+1) = 0	f 5	—
−1	+1	−1	p6:(−1+1)(−1) = 0	f 6	—
−1	−1	+1	p7:(−1−1)(+1) = −2	f 7	c
−1	−1	−1	p8:(−1−1)(−1) = +2	f 8	d

Los pesos son de +2 ó −2, de modo que resulta igual si los representamos con +1 y −1. Sin embargo, mantenemos las mismas cifras para no romper la generalidad que mantiene con las fórmulas de condición triple y de cuadruple, que trataremos seguidamente.

MA.f.c1+c2	X	Y	Z
A + C	1		
A + D	1	1	
B + D			1
B + E			1

MA.t.c1+c2	X	Y	Z
A + C	1.000		
A + D	.833	.833	
B + D			1.000
B + E			1.000

La fórmula de $CCM^{(2)}$ es:

$$CCM^{(2)} = p1*f1*(N+P-2) / [p1*f1*(N+P-2)+p2*f2+p7*f7]$$

(*Los pesos (p) son valores absolutos.)

MA.f.c1+c2	X	Y	Z
A + C	f1:1 (a)		
A + D	f3: 1 (-)	f4: 1 (-)	
B + D			f8: 1 (d)
B + E			f8: 1 (d)

MA.t.c1+c2	X	Y	Z
A + C	1.000		
A + D	.833	.833	
B + D			1.000
B + E			1.000

Por ejemplo el $CCM^{(2)}$ de [A+C]:X es:

$$CCM^{(2)} = 2*1*(4+3-2) / [2*(4+3-2) + 2*0 + 2*0]$$
$$= 10 / 10 = 1.000$$

donde la fila de [A+D] resulta 0 por ofrecer c1=+1, c2=−1, en comparación con [A+C] y no pertenece a ningún Grupo y no participa en el cálculo de CCM.

Calculamos el $CCM^{(2)}$ de [A+D]:X en la tabla siguiente:

MA.f.c1+c2	X	Y	Z
A + C	f3:1 (-)		
A + D	f1: 1 (a)	f2: 1 (b)	
B + D			f8: 1 (d)
B + E			f8: 1 (d)

MA.t.c1+c2	X	Y	Z
A + C	1.000		
A + D	.833	.833	
B + D			1.000
B + E			1.000

$$CCM^{(2)} = p1*f1*(N+P-2) / [p1*f1*(N+P-2)+abs(p2)*f2+abs(p7)*f7]$$
$$= 2*1*(4+3-2) / [2*1*(4+3-2) + 2*1 + 0]$$
$$= 10 / 12 = .833$$

(3) Coeficiente de condición triple

En el cálculo de Coeficiente de condición triple, se multiplica el número de condiciones. No obstante el método es similar.

c1	c2	c3	Efecto (e)	Peso (p)	Frec. (f)	Grupo
+1	+1	+1	+1	p1:(+1+1+1)(+1) = +3	f 1	a
+1	+1	+1	−1	p2:(+1+1+1)(−1) = −3	f 2	b
+1	+1	−1	+1	p3:(+1+1−1)(+1) = +1	f 3	a
+1	+1	−1	−1	p4:(+1+1−1)(−1) = −1	f 4	b
+1	−1	+1	+1	p5:(+1−1+1)(+1) = +1	f 5	a
+1	−1	+1	−1	p6:(+1−1+1)(−1) = −1	f 6	b
+1	−1	−1	+1	p7:(+1−1−1)(+1) = −1	f 7	c
+1	−1	−1	−1	p8:(+1−1−1)(−1) = +1	f 8	d
−1	+1	+1	+1	p9:(−1+1+1)(+1) = +1	f 9	a
−1	+1	+1	−1	p10:(−1+1+1)(−1) = +1	f 10	b
−1	+1	−1	+1	p11:(−1+1−1)(+1) = −1	f 11	c
−1	+1	−1	−1	p12:(−1+1−1)(−1) = +1	f 12	d
−1	−1	+1	+1	p13:(−1−1+1)(+1) = −1	f 13	c
−1	−1	+1	−1	p14:(−1−1+1)(−1) = +1	f 14	d
−1	−1	−1	+1	p15:(−1−1−1)(+1) = −3	f 15	c
−1	−1	−1	−1	p16:(−1−1−1)(−1) = +3	f 16	d

Todos los casos de f1 a f16, llevan peso, puesto que las condiciones presentan valores positivos o negativos, y no neutros (0). Según el signo (+/−) de la suma de as condiciones y el de efecto, se determina el Grupo. Los $CCM^{(3)}$ son:

MA.f.c1+c2+c3	X	Y	Z
A + C + F	1		
A + D + F	1		
A + D + G		1	
B + D + H			1
B + E + H			1

MA.t.	X	Y	Z
A + C + F	1.000		
A + D + F	.960		
A + D + G		.947	
B + D + H			1.000
B + E + H			1.000

Veamos, por ejemplo, el $CCM^{(3)}$ de [A+C+F]:X:

MA.f.c1+c2+c3	X	Y	Z
A + C + F	f1: 1 (a)		
A + D + F	f5: 1 (a)		
A + D + G		f8: 1 (d)	
B + D + H			f16: 1 (d)
B + E + H			f16: 1 (d)

MA.t.	X	Y	Z
A + C + F	1.000		
A + D + F	.960		
A + D + G		.947	
B + D + H			1.000
B + E + H			1.000

$CCM^{(3)}$ = (p1*f1+p3*f3+p5*f5+p9*f9)*(N+P–2)
/ (p1*f1+p3*f3+p5*f5+p9*f9)*(N+P–2)
+ (p2*f2+p4*f4+p6f6+p10*)
+ (p7*f7+p11*f11+p13*f13+p15+f15)]
= (3*1+1*1)*(5+3－2)
/ (3*1+1*1)*(5+3－2)
+ 0
+ 0
= 24 / 24 = 1

El $CCM^{(3)}$ de [A+D+F]:X es:

MA.f.c1+c2+c3	X	Y	Z
A + C + F	f1: 1 (a)		
A + D + F	f5: 1 (a)		
A + D + G		f4: 1 (b)	
B + D + H			f16: 1 (d)
B + E + H			f16: 1 (d)

MA.t.	X	Y	Z
A + C + F	1.000		
A + D + F	.960		
A + D + G		.947	
B + D + H			1.000
B + E + H			1.000

$CCM^{(3)}$ = (p1*f1+p3*f3+p5*f5+p9*f9)*(N+P–2)
/ (p1*f1+p3*f3+p5*f5+p9*f9)*(N+P–2)
+ (p2*f2+p4*f4+p6f6+p10*)
+ (p7*f7+p11*f11+p13*f13+p15+f15)]
= (3*1+1*1)*(5+3－2)
/ (3*1+1*1)*(5+3－2)
+ 1*1

+ 0
= 24 / (24 + 1) = .960

Aquí la condición [A+D+G], con respecto a [A+D+F] ofrece los valores de [+, +, −] y pertenece a f4, Grupo «b» y aumenta levemente el denominador.

El programa produce los resultados de los cálculos hasta la condición cuadruple. En la práctica aún en la matriz de datos dotada de más de 4 columnas con todas las combinaciones de hasta 4 condiciones es suficiente. Lo mismo que en el Análisis multivariantes que calcula los valores propios y vectores propios, en el Análisis de condición múltiple, ante los datos resultados de más de cinco condiciones posibles, el análisis es difícil de realizar.

(*) Experimento de relativización en el Análisis de condición múltiple

Con frecuencia los datos lingüísticos se presentan en una tabla de datos textuales como en la siguiente:

ID	1.Rasgo	Entorno	2.A25	3.Tipo	Letra
2	/y/	#V_V	a0925	1.V	i
3	/y/	#V_V	a0925	1.V	i
4	/y/	#V_V	a0950	1.V	i
4	/i/	#V_V	a0950	1.V	i
5	/y/	#V_V	a0975	1.V	i
5	/y/	#V_V	a0975	1.V	i
5	/y/	#V_V	a0975	1.V	i
9	/y/	#V_V	a1225	4.Gc	y
9	/y/	#V_V	a1225	4.Gc	y

donde no se sabe si los datos están correctamente relativizados. Las frecuencias de cada ítems dependen de la manera de recolección de los materiales.

Vamos a realizar un experimento con dos matrices de datos, que preparamos, por ejemplos. La matriz M1 contiene 5 filas:

M1	c1	c2	c3	c4	E
d1	A	C	F	I	X
d2	A	D	F	J	X
d3	A	D	G	K	Y
d4	B	D	H	L	Z
d5	B	E	H	M	Z

La matriz siguiente M2 lleva cinco filas mas/, repetidas la última de la M1:

M2	c1	c2	c3	c4	E
d1	A	C	F	I	X
d2	A	D	F	J	X
d3	A	D	G	K	Y
d4	B	D	H	L	Z
d5	B	E	H	M	Z
d6	B	E	H	M	Z
d7	B	E	H	M	Z
d8	B	E	H	M	Z
d9	B	E	H	M	Z
d10	B	E	H	M	Z

En M2 se han aumentado los datos y el aumento de los casos lógicamente se refleja en la casilla de M2.f, B:Z = 7:

M1.f	X	Y	Z
A	2	1	
B			2

M1.c	X	Y	Z
A	.857	.600	
B			1.000

M2.f	X	Y	Z
A	2	1	
B			7

M2.c	X	Y	Z
A	.857	.600	
B			1.000

Por esta razón los analistas tienen que tener cuidado al tratar los datos de dimensión distinta. El aumento de frecuencia puede deberse al aumento de materiales tratados y, por consiguiente, se hace alguna relativización en forma de porcentaje o de por mil palabras, etc.

No obstante, al observar las tablas (derechas) de Coeficientes de condición simple, el aumento de la frecuencia no afecta a la cifra de B:Z que presenta 1.000 tanto en M1 como en M2. La causa de esta invariabilidad es que la celda B:Z no compite con ninguna otra ni en la fila ni en la columna, lo que llamamos «distribución exclusiva». Esto se comprueba fácilmente en la fórmula de Jaccard a / (a + b + c), donde b y c son cero en el caso de la distribución exclusiva.

Tampoco las celdas de A:X ni A:Y están afectadas, puesto que en las filas aumentadas en M2 no se encuentras A:X ni A:Y.

La situación cambia en los casos siguientes:

MA1.f.	X	Y	Z
C	1		
D	1	1	1
E			1

MA1.c.	X	Y	Z
C	.800		
D	.571	.667	.571
E			.800

MA2..f.	X	Y	Z
C	1		
D	1	1	1
E			6

MA2.c.	X	Y	Z
C	.800		
D	.571	.667	.333
E			.960

En la celda de E:Z, el CCM ha aumentado de acuerdo con el cambio de la frecuencia absoluta. Esto se debe a que la celda E:Z no presenta la distribución exclusiva, sino la «distribución competitiva» con D:Z, de modo que el aumento de E:Z produce la proporción inclinada a E, lo que es también lógico y natural. No obstante no presenta un cambio tan drástico como de 1 a 6 en caso de frecuencia absoluta, sino un cambio controlado dentro del rango de 0 a 1 (.800 →.960).

En los datos en disribución exclusiva el Coeficiente de condición múltiple (CCM) no recibe el efecto del cambio de frecuencia absoluta. En los datos en distribución competitiva, tampoco presenta unos cambios grandes, que pueden ser debidos al número grande de los materiales. Por esta razón, es recomendable ver sus valores así relativizados de CCM. Las frecuencias absolutas, por otra parte, muestran el estado inicial de cifras reales, que se pierden de vista en los valores relativizados. Conviene utilizar los dos en la práctica del análisis.

(#) Letra <j> con valor fonológico palatal

La tabla siguiente es un extracto de los resultados de Análisis de condición múltiple de los documentos del Norte de España del siglo X a XIII. Hemos calculado los casos de letras <i>, <j>, <y> con distintos valores fonológicos y los años de emisión:

MA.f.Fonema+a25	<i>	<j>	<y>
/i/ + 1200	1465	451	1
/i/ + 1175	805	122	2
/i/ + 1225	333	111	18
/ly/ + 1200	201	56	
/ʒ/ + 1250	21	52	
/i/ + 1150	386	42	1
/ʒ/ + 1225	4	37	

MA.t.Fonema+a25	<i>	<j>	<y>
/i/ + 1200	.988	.922	.020
/i/ + 1225	.955	.896	.547
/i/ + 1175	.984	.834	.070
/ly/ + 1200	.695	.784	
/i/ + 1150	.967	.741	.061
/ʒ/ + 1250	.142	.710	
/i/ + 1275	.831	.684	.853

/ly/ + 1225	15	26		/ʒ/ + 1225	.031	.667	
/i/ + 1250	233	21	70	/ʒ/ + 1200	.122	.659	
/ʒ/ + 1275	1	21		/i/ + 1250	.941	.640	.890
/ly/ + 1175	116	20		/ly/ + 1225	.117	.585	
/ʒ/ + 1200	13	20		/ʒ/ + 1275	.008	.498	
/i/ + 1275	70	18	33	/ly/ + 1175	.530	.483	

Por estas tablas sabemos que a partir del siglo XIII, empieza a usarse <j>, en lugar de <i> para representar el fonema palatal /ʒ/, por ejemplo, *fijo, ojo, concejo*. El Coeficiente de condición múltiple (doble) nos permite interpretar la distribución de frecuencias, que es difícil de hacerlo en forma de frecuencias absolutas. Y después de interpretar la cronología de los datos en cuestión, también conviene observar sus frecuencias reales.

5.8. Análisis de generalidad y peculiaridad

A partir de los datos en forma de matriz, calculamos los grados de «generalidad» y de «particularidad» para ver qué variables (v1-v5) los poseen con respecto a otros.

G	v1	v2	v3	v4	v5
d1	1	1	1	0	0
d2	1	1	0	1	1
d3	0	1	0	1	0
d4	1	0	1	1	1

H	Suma
d1	3
d2	4
d3	2
d4	4

Para medir el grado de la «generalidad» de las variables, no sirven la suma vertical, aunque admitamos que es algo para ver la fuerza de cada variable:

V	v1	v2	v3	v4	v5
Suma	3	3	2	3	2

La suma vertical no nos sirve mucho porque no toma en consideración la comunalidad que posee cada valor positivo (1) en las filas correspondientes. Seguidamente buscamos una alternativa.

En primer lugar consideramos los valores que posee cada punto en relación con otros puntos de la misma fila. Por ejemplo, el valor 1 de [v1:d1] es distinto al de [v1:d2], puesto que el primero es 1 al lado de 2, mientras que el segundo es 1 al lado de 3, de modo que consideramos que el 1 de segundo caso posee más generalidad que el primero, es decir, sigue la moda de la distribución.

Por consiguiente, aprovechamos la suma horizontal para representar la «generalidad» del valor positivo (1) de manera siguiente:

H = sumH(G) - G

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

Ahora los puntos positivos (1) se han convertido en los valores de la suma horizontal (3, 4, 2, 4) menos la matriz G. De esta manera los mismos valores han cobrado valores que representan el grado de comunalidad en cada fila. Por ejemplo, el 1 de [v1, d1] posee el valor comunal de otros 2, mientras que el 1 de [v1, d2] posee el valor comunal de otros 3. Cuanto más el valor de comunalidad, mayor es el grado de «generalidad». Según nuestro sentido común, el ciudadano que vota al partido ganador de elección, igual que mayoría de los votantes, representa el mayor «generalidad» en la votación.

G.a.val.	v1	v2	v3	v4	v5
Sum	8	6	5	7	6

H2	Suma
Suma	32

Ahora procedamos a sumar los valores de manera vertical: (8, 6, 5, 7, 6). Este vector horizontal representa de alguna manera el grado de «generalidad»: «Generalidad por valor absoluto» (F.a.val.).

Conviene calcular su valor relativo para evaluar el mismo grado dentro de la escala de 0 a 1. Por consiguiente dividimos el vector horizontal (8, 6, 5, 7, 6) por la suma del mismo vector (32):

r.freq.G.	v1	v2	v3	v4	v5
M*H/H2	.250	.188	.156	.219	.188

De esta manera conseguimos el grado de generalidad por valor relativo (G.r.val.)

Para medir el grado de particularidad (P.r.val.), a partir del de generalidad (G.r.val.), procedamos al cálculo de sustración:

P.r.val. = 1 - G.r.val.

r.freq.P.	v1	v2	v3	v4	v5
1-M*H/H2	.750	.813	.844	.781	.813

Notamos que el grado de «generalidad» es siempre considerablemente

menor que el de «particularidad», no por la característica misma de los datos, sino más bien simplemente por el número de variables. Hemos visto que entre las cinco variables (v1:5), se divide la suma vertical correspondiente a cada columna. Nos imaginamos que entre veinte variables, se reduce todavía más sin llegar casi al primer dígito (0.1). Y consecuentemente el valor del grado de particularidad se eleva de manera descomunal.

Para buscar la fórmula que no recibe la influencia de la cantidad de variables, proponemos una operación que denominamos «frecuencia relativa prominente» (f.r.p.). Consiste en elevar el valor de la frecuencia absoluta de la variable en cuestión, multiplicado por el número de variables en comparación. Para ver la operación concreta, reproducimos el cuadro anteriormente usado:

a.freq.G..	v1	v2	v3	v4	v5
Suma	8	6	5	7	6

H2	Suma
Suma	32

Por ejemplo, 8 entre 32, es 8 / 32 = .250, que es la frecuencia relativa. Ahora bien, consideramos que el valor 8 es difícil de evaluar dentro de la totalidad tan grande como de cinco miembros, es decir, no es directamente comparable con los 4 restantes, puesto que se trata de un valor contra 4 valores. De ahí viene el valor siempre considerablemente reducido en la frecuencia relativa, como acabamos de ver: (.250, .188, .156, .219, .188). Para subsanar esta reducción, y para igualar la condición de comparación, multiplicamos el valor en cuestión por el número de los restantes: 8 * (5 - 1) = 32, que es ahora comparable con los 4 restantes (6 + 5 + 7 + 6 = 24). Para obtener la «frecuencia relativa prominente» calculamos el valor de 32 entre las dos cifras (32 + 24 = 56): es decir, 32 / (32 + 24) = 32 / 56 = .571. La fórmula de la frecuencia relativa prominente (f.r.p.) es:

$$f.r.p = [f.a * (n-1)] / [f.a * (n-1) + (s.h. - f.a)]$$

$$= [8 * (5 - 1)] / [8 * (5 - 1) + (32 - 8)] = .571$$

donde, f.a. es frecuencia absoluta, n es número de variables, s.h. es suma horizontal. El resultado del valor relativo prominente (f.r.p.) es:

G.p.r.val.	v1	v2	v3	v4	v5
Suma	.571	.480	.426	.528	.480

que representa nuestro grado de «generalidad por valor relativo prominente» (G.p.r.val.). Y su correspondiente grado de peculiaridad por frecuencia relativa prominente (p.f.r.p.: p.r.freq.P.) es:

P.p.r.val. = 1 - G.p.r.val.

p.r.freq.P.	v1	v2	v3	v4	v5
Suma	.429	.520	.574	.472	.520

Resumimos todas estas operaciones de manera siguiente:

Generality	v1	v2	v3	v4	v5
Sum	3	3	2	3	2
G.a.val.	8	6	5	7	6
G.r.val.	.250	.188	.156	.219	.188
P.r.val.	.750	.813	.844	.781	.813
G.p.r.val.	.571	.480	.426	.528	.480
P.p.r.val.	.429	.520	.574	.472	.520

Para ver los grados de «generalidad» - «particularidad» de los datos individuales, en lugar de las variables, transponemos la matriz y le aplicamos el mismo programa:

Tr(M)	d1	d2	d3	d4
v1	1	1	0	1
v2	1	1	1	0
v3	1	0	0	1
v4	0	1	1	1
v5	0	1	0	1

Generality	d1	d2	d3	d4
Sum	3	4	2	4
G.a.val.	5	7	4	6
G.r.val.	.227	.318	.182	.273
P.r.val.	.773	.682	.818	.727
G.p.r.val.	.469	.583	.400	.529
P.p.r.val.	.531	.417	.600	.471

(*) «Generalidad / Peculiaridad» de matrices frecuenciales y decimales

El mismo cálculo de «generalidad / peculiadad» es viable con tales datos frecuenciales como los siguientes:

F	v1	v2	v3	v4	v5
d1	2	3	4	0	0
d2	1	2	0	5	2
d3	0	4	0	4	0
d4	1	0	3	1	1

G.a.val. = Fnp * [sumH(Fnp) - Inp]

Generality	v1	v2	v3	v4	v5
Sum	4	9	7	10	3
G.a.val.	28	50	29	46	21
G.r.val.	.161	.287	.167	.264	.121
P.r.val.	.839	.713	.833	.736	.879
G.p.r.val.	.434	.617	.444	.590	.354
P.p.r.val.	.566	.383	.556	.410	.646

También es posible con los datos decimales, con tal de que se traten de los datos no negativos (igual o más de 0):

RSv	v1	v2	v3	v4	v5
d1	.500	.333	.571	.000	.000
d2	.250	.222	.000	.500	.667
d3	.000	.444	.000	.400	.000
d4	.250	.000	.429	.100	.333

Generality	v1	v2	v3	v4	v5
Sum	1.000	1.000	1.000	1.000	1.000
G.a.val.	1.015	.850	.769	.848	.908
G.r.val.	.231	.194	.175	.193	.207
P.r.val.	.769	.806	.825	.807	.793
G.p.r.val.	.546	.490	.459	.489	.510
P.p.r.val.	.454	.510	.541	.511	.490

5.9. Análisis por ejes selectivos

Para situar los atributos y casos en un espacio bidimensional, se utilizan los valores estadísticos obtenidos en análisis multivariantes tales como el Análisis de Componentes Principales, Análisis de Correspondencia, Análisis Factorial, etc. Para entender estos métodos se necesitan preparaciones matemáticas de alto nivel. En cambio, en el método que llamamos «Análisis por Ejes Selectivos» es suficiente saber lo que es matriz simétrica de correlación, asociación o proximidad. En lo siguiente explicamos el método de observar las relaciones entre los atributos, lo que es aplicable también para la observación de los casos. Finalmente procedemos al método de observar las relaciones entre atributos y casos en un mismo plano bidimensional.

5.9.1. Análisis por ejes selectivos de matriz de correlación

La tabla inferior izquierda (M) es una matriz de datos y la tabla derecha es su matriz de correlación (Co(M)):

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

Co(M)	X:A	Y:B	C	D	E
A	1.000	.787	.944	-.480	.436
B	.787	1.000	.945	-.092	.896
C	.944	.945	1.000	-.331	.709
D	-.480	-.092	-.331	1.000	.140
E	.436	.896	.709	.140	1.000

En la matriz de correlacion Co(M), están calculados todos los posibles coeficientes de correlacion entre 5 variables A, B, C, D, E[33]. Al fijarnos en las columnas de A y B, nos damos cuenta de que la columna A representa las relaciones entre A de un lado y A, B, C, D, E de otro; y la columna B representa las relaciones entre B de un lado y A, B, C, D, E de otro. De modo que utilizando estas dos columnas como las dos coordenadas, X e Y, podemos situar los 5 puntos con coordenadas correspondientes: A(1.000, .787), B(.787, 1.000), C(.944, .945), D(-.480, -.092), E(.436, .896):

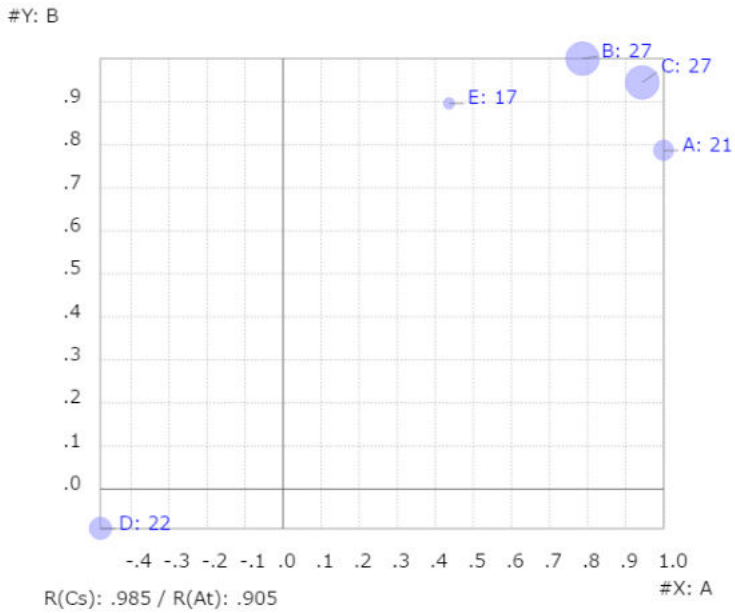


Fig.1

Este grafico muestra las relaciones de todos los atributos con los dos atributos A y B. La localizacion de A (1.000, .787) presenta el valor 1.000 en el

[33] En la matriz de datos (M) solo contamos con 4 casos de modo que estos coeficientes no poseen una significancia suficiente. Aquí utilizamos esta matriz para la facilidad de explilcación y entendimiento.

eje X por tratarse de autocorrelacion, mientras que posee el valor .787 en el eje Y que corresponde a la columna B de la matriz, lo que impide colocar el punto en su propio eje X:A. Por lo tanto el eje X (horizontal) representa el grado de relación entre A y cada atributo; y el eje Y (vertical) el grado de relación entre B y cada atributo. Si el analista adopta dos puntos de vista significativos para su proposito de estudio, el gráfico así dibujado sirve para su análisis, puesto que la situación de los atributos se interpreta en relación con los dos atributos seleccionados.

Ademas de las combinaciones de la pareja A:B, hay multitud de las posibles combinaciones: A:C, A:D, A:E, B:C, B:D, ..., etc[34]. Pensamos que al seleccionar los dos atributos que presentan el valor mínimo absoluto del coeficiente (el valor más cercano a 0), los puntos se colocan con mayor dispersión dentro del espacio del grafico, lo que garantiza la mayor facilidad de caracterización de los atributos. Si se unieran en unos lugares cercanos a un punto o en una línea determinado, no se revelarían sus características con facilidad. Para buscar el valor mínimo absoluto de correlación, tratamos la matriz simétrica de correlación Co(M) como si fuera una matriz de datos para obtener su matriz de correlación Co(Co(M))), que se presenta de manera siguiente:

Co(Co(M))	A	B	C	D	E
A	1.000	.905	.981	-.993	.549
B	.905	1.000	.970	-.867	.853
C	.981	.970	1.000	-.961	.700
D	-.993	-.867	-.961	1.000	-.486
E	.549	.853	.700	-.486	1.000

donde encontramos el valor mínimo absoluto |-.486| = .486 en la pareja D:E. Ahora bien, dibujamos de nuevo el grafico de dispersión con las dos columnas D y E de la matriz Co(M):

Co(M)	X:A	Y:B	C	D	E
A	1.000	.787	.944	-.480	.436
B	.787	1.000	.945	-.092	.896
C	.944	.945	1.000	-.331	.709
D	-.480	-.092	-.331	1.000	.140
E	.436	.896	.709	.140	1.000

[34] Hay p*(p-1)/2 = 5*4 = 10 parajas posibles, donde p = numero de atributos.

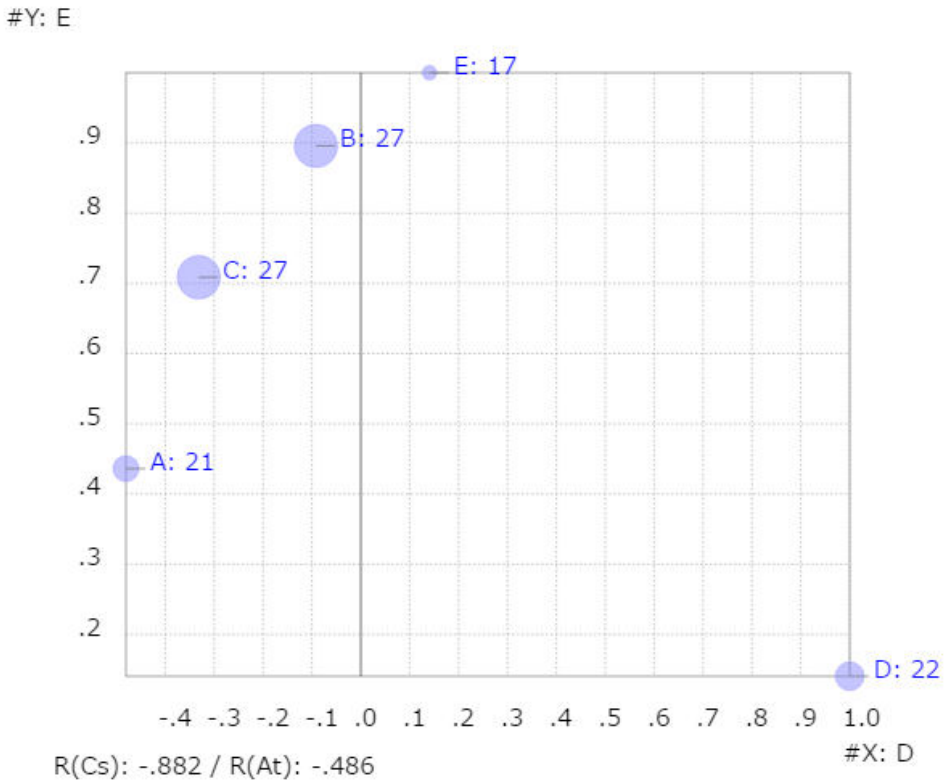


Fig.2

En los dos gráficos, observamos que los puntos se dividen en los dos grupos: {A, B, C, E} y {D}. Ahora preparamos una nueva columna A.B.C.E con los valores promedios de los cuatro atributos:

M'	A	B	C	D	E	A.B.C.E.
h1	10	19	14	7	12	13.75
h2	11	7	10	0	1	7.25
h3	0	0	1	12	1	.50
h4	0	1	2	3	3	1.50

La matriz de correlacion de M' es:

Co(M')	A	B	C	X:D	E	Y:A.B.C.E.
A	1.000	.787	.944	-.480	.436	.867
B	.787	1.000	.945	-.092	.896	.989
C	.944	.945	1.000	-.331	.709	.983
D	-.480	-.092	-.331	1.000	.140	-.207
E	.436	.896	.709	.140	1.000	.826
A.B.C.E.	.867	.989	.983	-.207	.826	1.000

de la que obtenemos el gráfico siguiente:

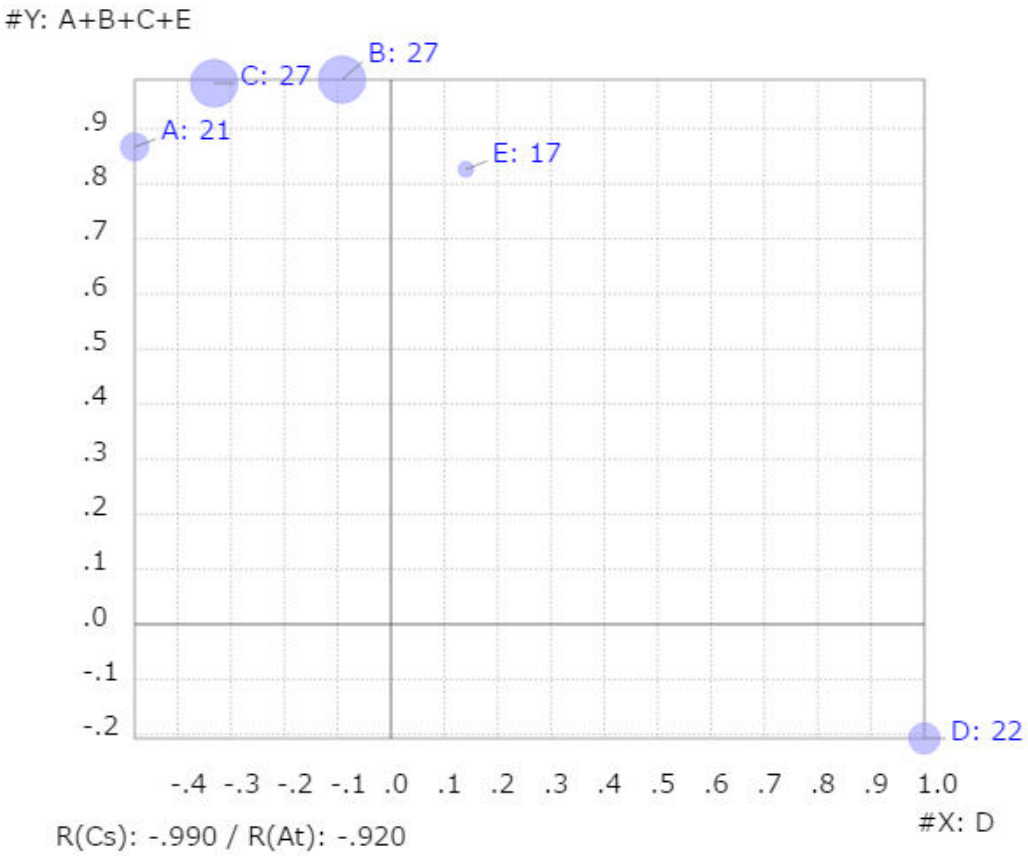


Fig.3

La Fig.3 se parece a la Fig.2. En la Fig.3, sin embargo, notamos que E no representa el grupo A.B.C.E, sino que más bien B y C lo hacen[35].

5.9.2. Análisis por ejes selectivos de matriz de proximidad

Para realizar un análisis por ejes selectivos de correlación, necesitamos por lo menos 3 casos[36]. En cambio si utilizamos el coeficiente de proximidad en lugar del de correlación, podemos calcular la matriz incluso con un solo caso. Para analizar un matriz de casos reducidos, conviene utilizar coeficientes de Proximidad.

Las tablas siguientes son la matriz de datos y su matriz simétrica de coeficientes de Proximidad Regular (PR):

[35] Aqui dividimos los atributos en dos grupos observando la dispersión de atributos. En análisis con más atributos en múltiples relaciones, recurrimos al análisis de agrupamiento no jerárquico en dos grupos.

[36] Cuando contamos con un solo caso, los coeficientes de correlacion son 0.000 en todas las pareja, y con dos casos, los coeficientes de correlacion son necesariamente 1 o -1. En realidad con 3 casos su coeficiente de correlacion no posee significancia suficiente.

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

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

Al seleccionar los atributos A y B para los dos ejes X e Y obtenemos las dos coordenadas y su gráfico correspondiente:

RP(A, B)	X: A	Y: B
A	1.000	.845
B	.845	1.000
C	.958	.949
D	.331	.444
E	.697	.841

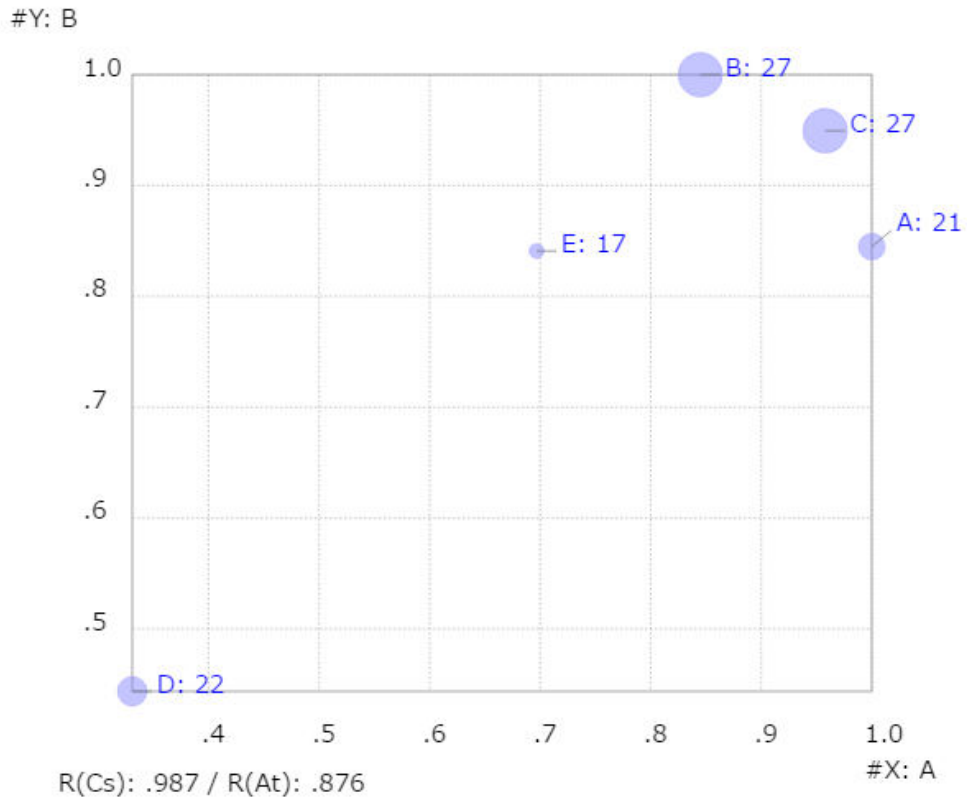


Fig.4

Para ver sus características de cada atributo en un espacio de la mayor dispersión, calculamos la matriz de correlación de RP, donde encontramos su valor absoluto mínimo (.355) en la pareja A:E:

Co(RP)	A	B	C	D	E

A	1.000	.876	.978	-.980	.355
B	.876	1.000	.956	-.898	.661
C	.978	.956	1.000	-.970	.498
D	-.980	-.898	-.970	1.000	-.430
E	.355	.661	.498	-.430	1.000

Volviendo a la matriz de RP obtenemos las dos coordenadas X:Y en la pareja A:E:

RP(A, E)	X: A	Y: E
A	1.000	.697
B	.845	.841
C	.958	.811
D	.331	.588
E	.697	1.000

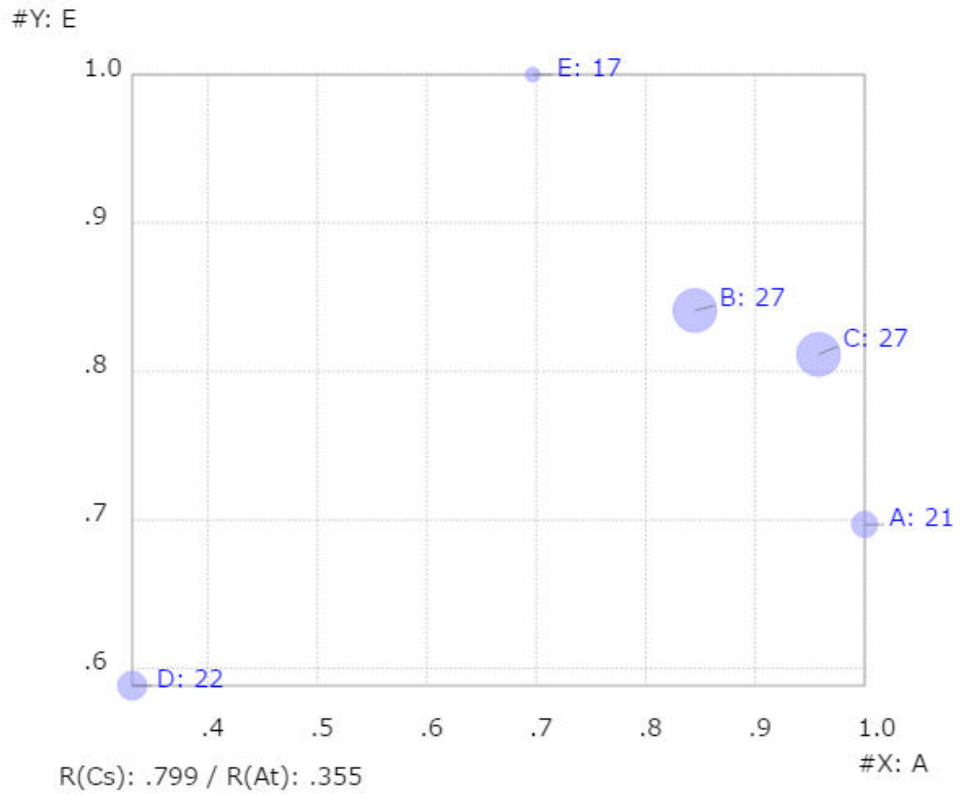


Fig.5

La Fig.5 donde analizamos los atributos con los ejes A:E se presenta muy diferente de la Fig.4, debido a que la matriz de la Proximidad calcula las posiciones de la columna en cuanto a la distancia que hay entre los dos ejes.

La Fig.4 y la Fig.5 son diferentes de las figuras de los gráficos del análisis por ejes selectivos por la correlación. La relación que presenta el coeficiente de correlación muestra la tendencia de movimiento de los puntos, mientras que el coeficiente de proximidad calcula la cercanía de los puntos. De

modo que el analista debe escoger el coeficiente adecuado para su propósito.

5.9.3. Analisis por ejes selectivos de los casos

Es posible analizar los casos por ejes selectivos transponiendo la matriz de datos. La tabla siguiente izquierda de la matriz de datos, la central es su matriz transpuesta y la derecha es la matriz simétrica de correlación de los datos transpuestos:

M	A	B	C	D	E
i1	10	19	14	7	12
i2	11	7	10	0	1
i3	0	0	1	12	1
i4	0	1	2	3	3

T(M)	i1	i2	i3	i4
A	10	11	0	0
B	19	7	0	1
C	14	10	1	2
D	7	0	12	3
E	12	1	1	3

Co(T(M))	X:i1	Y:i2	i3	i4
i1	1.000	.387	-.683	-.323
i2	.387	1.000	-.670	-.840
i3	-.683	-.670	1.000	.586
i4	-.323	-.840	.586	1.000

El gráfico siguiente es el resultado del análisis de los casos por ejes selectivos. Por este resultado podemos dividir los casos en dos grupos: i1-i2 y i3-i4:

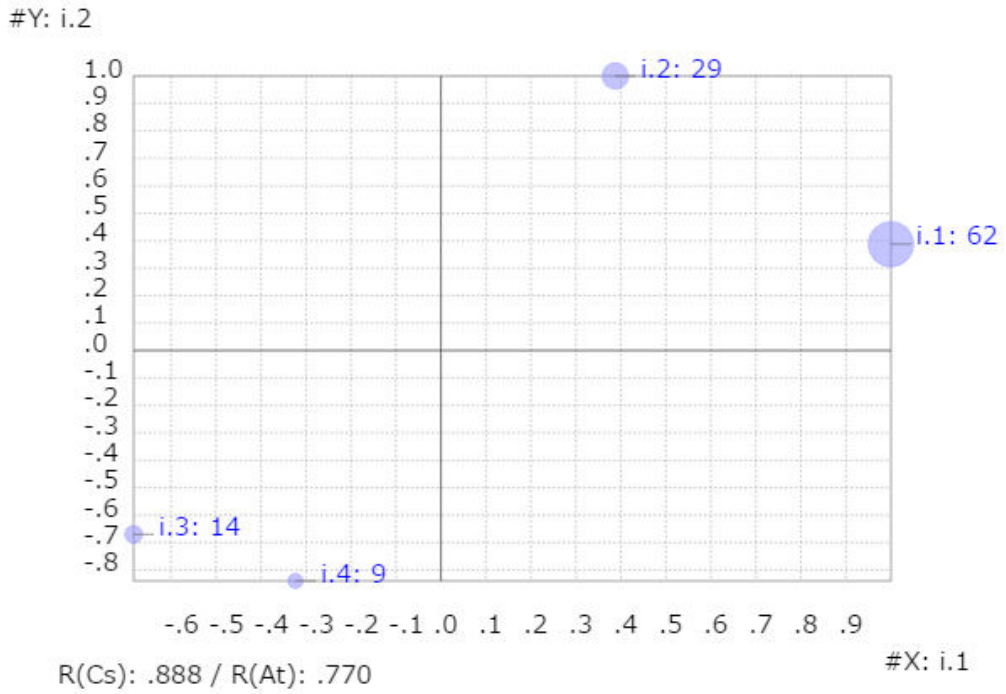


Fig. 6

5.9.4. Análisis por ejes selectivos de los atributos y casos

Desarrollamos un método de situar los atributos y los casos en el mismo

plano. Las tablas siguientes son la matriz de datos y su matriz simétrica de coeficientes de correlacion. Sn1 es la suma de la fila y S1p es la suma de la columna:

Mnp	A	B	C	D	E	Sn1
h1	10	19	14	7	12	62
h2	11	7	10	0	1	29
h3	0	0	1	12	1	14
h4	0	1	2	3	3	9
S1p	21	27	27	22	17	114

Co(M)	A	:B	C	D	E
A	1.000	.787	.944	-.480	.436
B	.787	1.000	.945	-.092	.896
C	.944	.945	1.000	-.331	.709
D	-.480	-.092	-.331	1.000	.140
E	.436	.896	.709	.140	1.000

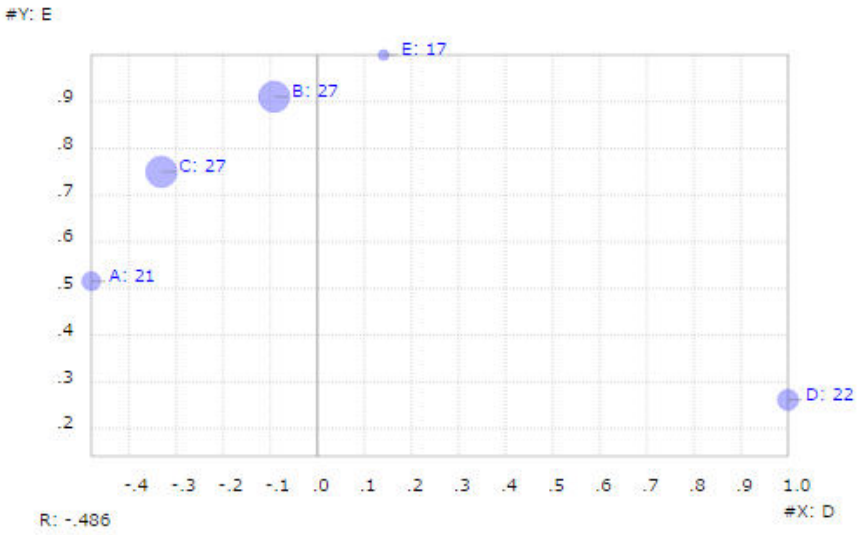


Fig.7

Por ejemplo, h1 posee los valores (10, 19, ..., 12) correspondientes a los atributos (A, B, ..., E). Utilizando este vector para obtener las dos coordenadas en forma de media ponderada.

La tabla siguiente (Ap2) es extracto de Co(X) de las dos columnas D:E.

Ap2	X: D	Y: E
A	-.480	.436
B	-.092	.896
C	-.331	.709
D	1.000	.140
E	.140	1.000

Estas dos columnas representan las coordenadas de los atributos A, B, ..., E, de modo que para determinar las coordenadas de h1 calculamos las medias ponderadas de la manera siguiente[37]:

[37] Debemos este método al profesor Hiroshi Kurata de la Universidad de Tokio (2017-2-11).

X(h1) = (10*-.480 + 19*-.092 + 14*-.331 + 7*1.000 + 12*.140) / 62 = -.040
Y(h1) = (10*.436 + 19*.896 + 14*.709 + 7*.140 + 12*1.000) / 62 = .714

De esta manera conseguimos las coordenas medias de h1 (-.040, .714). De la misma manera las coordenadas de h2, h3, h4 son:

Cn2	X:D	Y:E
h.1	-.040	.714
h.2	-.314	.661
h.3	.844	.242
h.4	.296	.637

El grafico que muestra las posiciones de atributos y los casos es el siguiente:

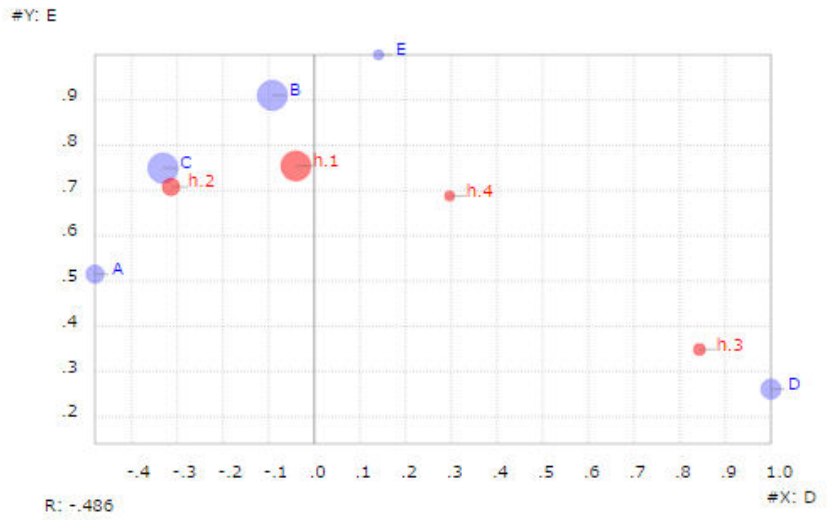


Fig.8

Por ejemplo el caso h1 está cerca del atributo B y h2 es media de A, B, C. En la matriz de datos, h2 tiene los valores (11, 7, 10, 0, 1), de modo que aparentemente debería estar cerca de A (=11). Sin embargo, considerando no solamente A, C, sino en la totalidad de los atributos, sus medias ponderadas se calculan (.910, .897) que determinan su posición, más cercana a C.

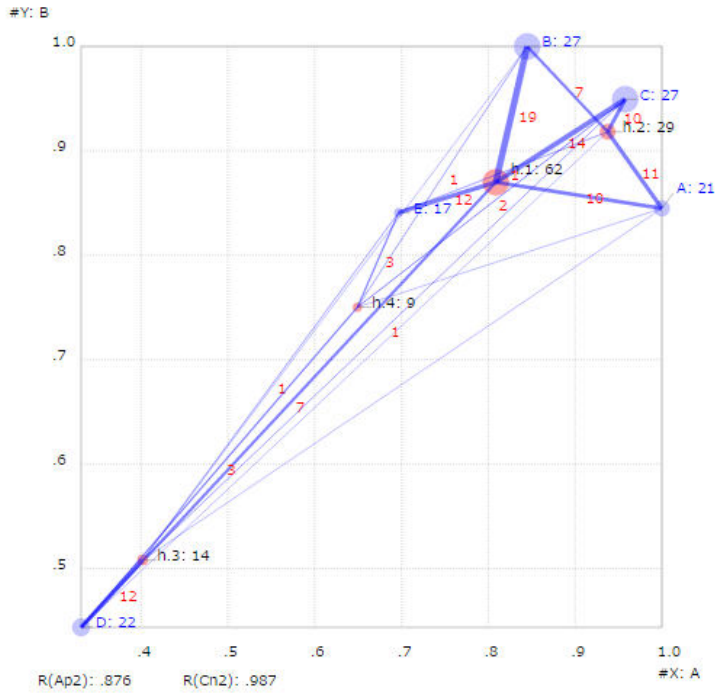
5.9.5. Selección de ejes

Para seleccionar los dos ejes podemos pensar en varias maneras. En lo siguiente las explicamos con la matriz simétrica de coeficientes de Proximidad Regular (PR).

M	A	B	C	D	E
h.1	10	19	14	7	12
h.2	11	7	10	0	1
h.3	0	0	1	12	1
h.	0	1	2	3	3

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

(1) Selección libre simple: Por ejemplo seleccionamos libremente A:B para ver las relaciones de los atributos con respecto a los atributos A y B:



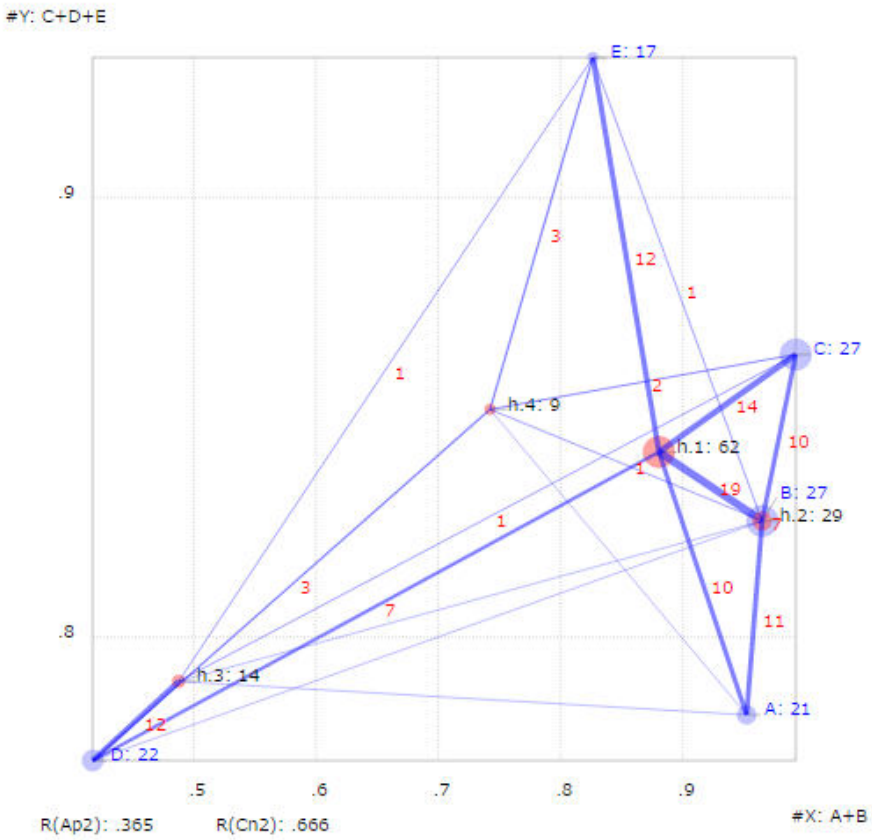
(2) Correlación mínima simple: Como observamos anteriormente la correlación de PR se pone mínima al seleccionar A:E.

(3) Selección libre multiple: Es posible seleccionar varios atributos para un eje, por ejemplo: X:A+B, Y:C. Al seleccionar múltiples atributos, el vector del eje seleccionado va a ser media de los miembros correspondientes:

Cn2	X: A+B	Y: C
h.1	14.500	14.000
h.2	9.000	10.000
h.3	.000	1.000
h.4	.500	2.000

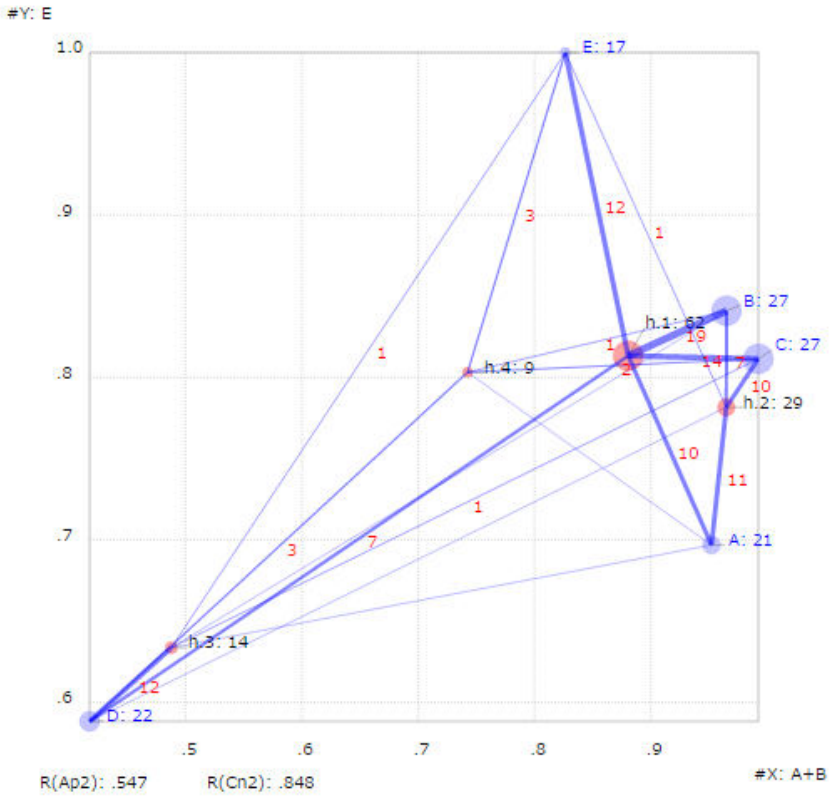
La selección siguiente representa X:A+B, Y:C+D+E:

Cn2	X: A+B	Y: C+D+E
h.1	14.500	11.000
h.2	9.000	3.667
h.3	.000	4.667
h.4	.500	2.667



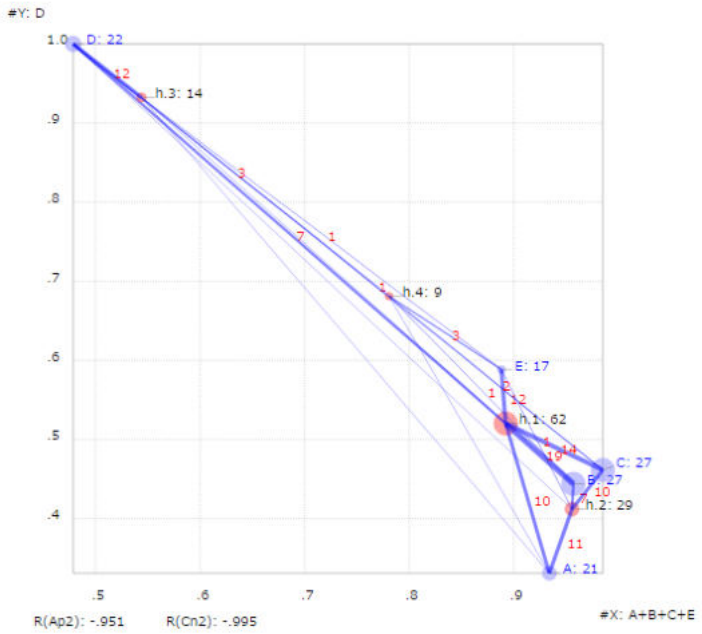
(4) Selección del eje restante en relación mínima: Para un eje seleccionamos uno o varios atributos que es nuestro punto de vista y el eje restante va a ser el que muestra la correlacion absoluta minima, por ejemplo:

Cn2	A+B	E
h.1	14.500	12.000
h.2	9.000	1.000
h.3	.000	1.000
h.4	.500	3.000



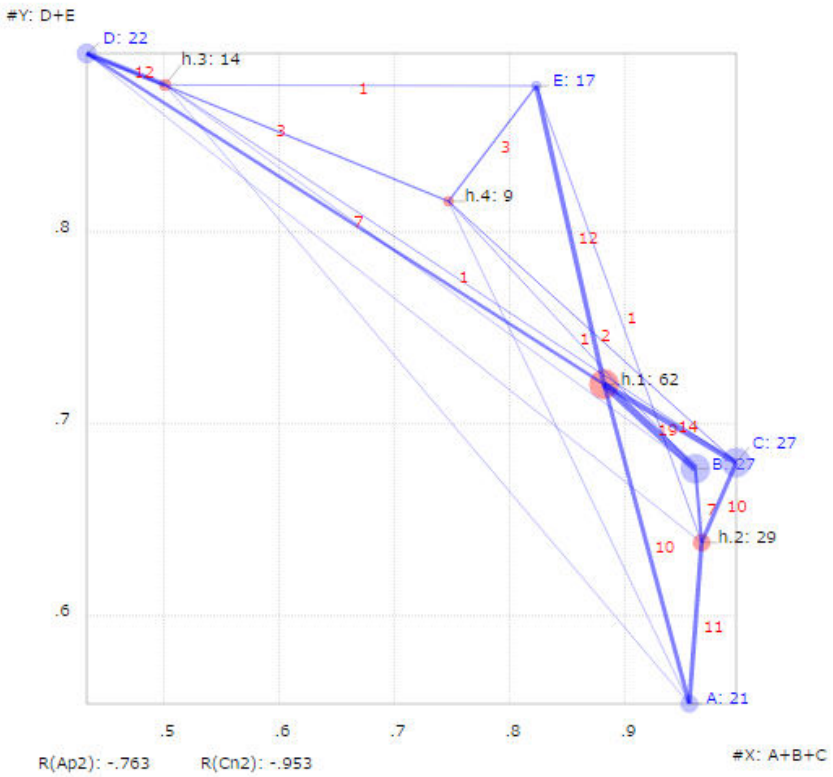
(5) Selección por agrupamiento: Utilizamos el análisis de agrupamiento no jerárquico para formar dos grupos compuestos, por ejemplo:

Cn2	A+B+C+E	D
h.1	13.750	7.000
h.2	7.250	.000
h.3	.500	12.000
h.4	1.500	3.000



(6) Selección residual: Al seleccionar atributo(s) en un eje simple o compuesto, el segundo eje va a ser el resto de los atributos:

Cn2	A+B+C	D+E
h.1	14.333	9.500
h.2	9.333	.500
h.3	.333	6.500
h.4	1.000	3.000



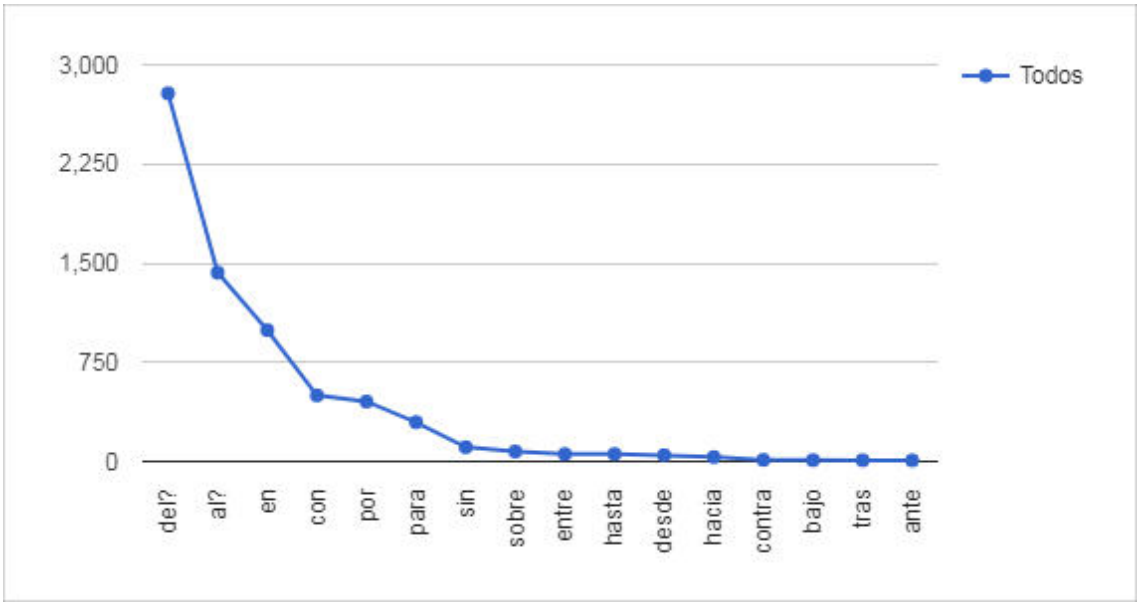
5.10. Análisis de Pareto

5.10.1. Índice de Pareto

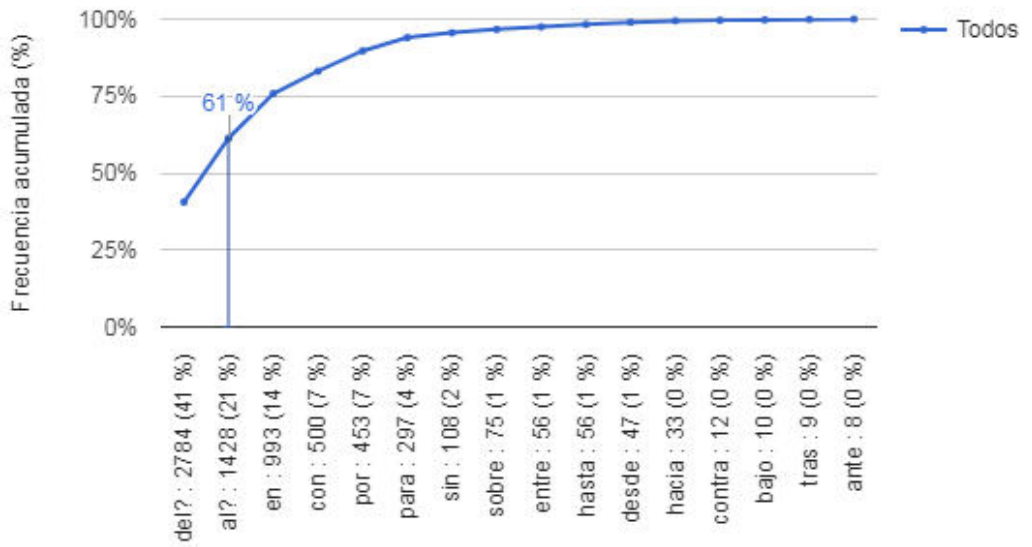
Partimos de la distribución de frecuencia de las preposiciones que hay en *Marianela* (1878) de Benito Pérez Galdós, que presentamos ordenada de manera descendente[38]:

FA	Todos
del?	2 784
al?	1 428
en	993
con	500
por	453
para	297
sin	108
sobre	75
entre	56
hasta	56
desde	47
hacia	33
contra	12
bajo	10
tras	9
ante	8

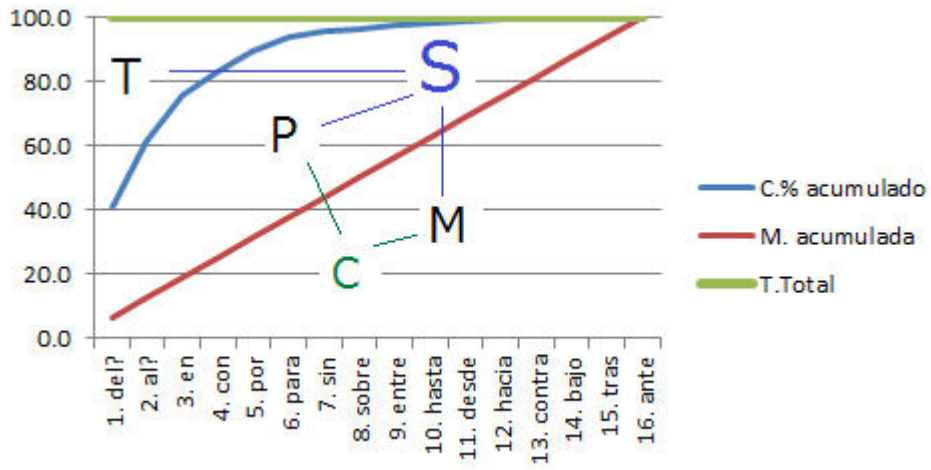
[38] Las expresiones regulares, <del?> y <al?>, significan que se han sumado las frecuencias tanto de <de> y <a> como de <del y <al>.



El gráfico de Pareto consiste en calcular las frecuencias acumuladas de las frecuencias ordenadas de manera descendente y convertirlas en el porcentaje. En el siguiente gráfico, el punto de 61% indica que los dos primeros elementos, concretamente, las dos preposiciones <de> y <a> ocupan más de la mitad (61 %) dentro de la frecuencia total de preposiciones:

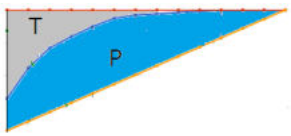


Para proceder a formular el Índice de Pareto, que indica el grado de concentración de los primeros elementos, elaboramos manualmente, la siguiente figura:

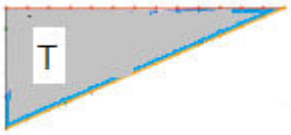


Hemos dividido el espacio total del diagrama (S: suma) en tres partes: T (Tope), P (Pareto) y M (Media), utilizando la línea horizontal de tope, la curva de Pareto y la frecuencia acumulada de media. Nuestra idea para evaluar el grado de concentración, que denominamos Índice de Pareto (IP), consiste en la proporción que ocupa P dentro del triángulo superior (T + P):

IP = P / (T + P)



Cuanto más espacio ocupa el P en el triángulo superior, más elevado sería el grado de concentración. El mínimo de IP se consigue cuando el espacio P es nulo, es decir, la curva se identifica con la línea recta (hipotenusa) de media; y su máximo se presenta cuando el espacio ocupa todo el triángulo, que ocurre cuando se produce la concentración única en el primer elemento:



Mínimo: IP = 0 (P = 0) / Máximo: IP = 1 (T = 0)

La suma (S) reúne las tres partes: T, P, M. Otro conjunto de espacio es C, que suma P y M:

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

De todos estos espacios, se calculan directamente S, C y M. Por otra parte, el espacio T se obtiene por la suma (S) de 100 (%) menos el espacio de

Pareto (P) y el de media (M):

$$T = S - P - M = S - C$$

El espacio P se obtiene por la sustracción de C por M:

$$P = C - M$$

Por consiguiente, el Índice de Pareto (IP) es:

$$IP = P / (T + P) = (C - M) / (S - C + C - M) = (C - M) / (S - M)$$

lo que se entiende inmediatamente en el gráfico anterior. Para comprobar el cálculo paso por paso, veamos la tabla siguiente:

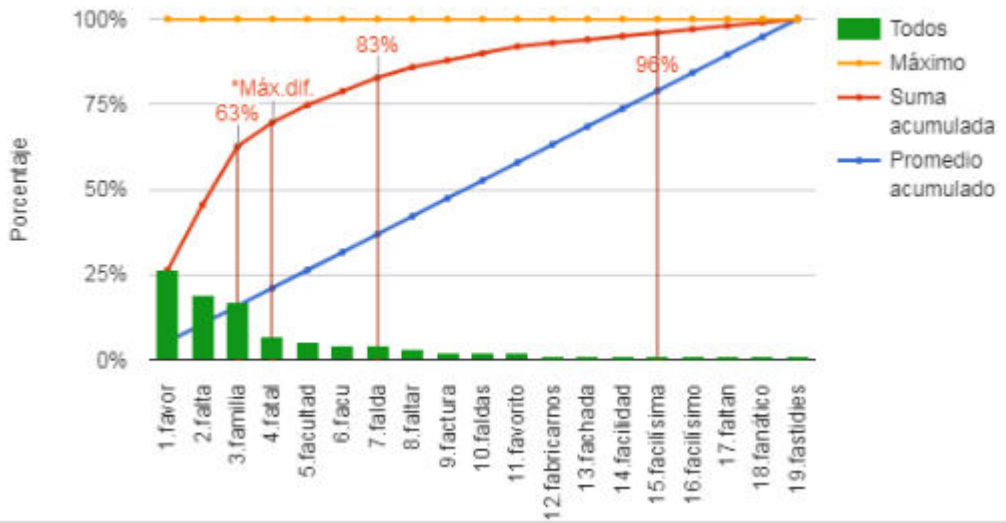
Prep.	Frec.	%	C.% acumulado	Media	M. acumulada	T.Total
1. del?	2 784	40.5	40.5	6.3	6.3	100.0
2. al?	1 428	20.8	61.3	6.3	12.5	100.0
3. en	993	14.5	75.8	6.3	18.8	100.0
4. con	500	7.3	83.1	6.3	25.0	100.0
5. por	453	6.6	89.6	6.3	31.3	100.0
6. para	297	4.3	94.0	6.3	37.5	100.0
7. sin	108	1.6	95.5	6.3	43.8	100.0
8. sobre	75	1.1	96.6	6.3	50.0	100.0
9. entre	56	0.8	97.5	6.3	56.3	100.0
10. hasta	56	0.8	98.3	6.3	62.5	100.0
11. desde	47	0.7	99.0	6.3	68.8	100.0
12. hacia	33	0.5	99.4	6.3	75.0	100.0
13. contra	12	0.2	99.6	6.3	81.3	100.0
14. bajo	10	0.1	99.8	6.3	87.5	100.0
15. tras	9	0.1	99.9	6.3	93.8	100.0
16. ante	8	0.1	100.0	6.3	100.0	100.0
Suma	6 869	100.0	1429.8	100.0	850.0	1 600.0

P:	s(C) - s(M)	1429.8 - 850 = 580
T + P:	s(T) - s(M)	1600 - 850 = 750
IP:	P / (T + P)	580 / 750 = 0.773

5.10.2. Análisis de Pareto simple

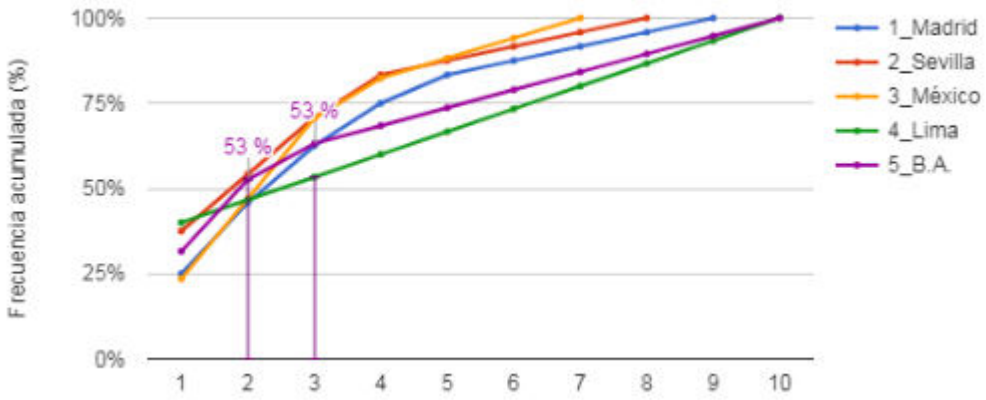
Pareto	Total
Índice de Pareto	.631
M: Punto de mitad (50%)	3
N: Número de datos	19
M / N (%)	16 %
A: Acumulación en mitad	62
T: Frecuencia total	99
A / T (%)	63 %
Punto de 50%	3
Punto de 80%	7
Punto de 95%	15

Rango	Total
1.favor	26 (26 %)
2.falta	19 (19 %)
3.familia	17 (17 %)
4.fatal	7 (7 %)
5.facultad	5 (5 %)
6.facu	4 (4 %)
7.falda	4 (4 %)
8.faltar	3 (3 %)
9.factura	2 (2 %)
10.faldas	2 (2 %)
11.favorito	2 (2 %)
12.fabricarnos	1 (1 %)
13.fachada	1 (1 %)
14.facilidad	1 (1 %)
15.facilísima	1 (1 %)
16.facilísimo	1 (1 %)
17.faltan	1 (1 %)
18.fanático	1 (1 %)
19.fastidies	1 (1 %)



5.10.3. Análisis de Pareto múltiple

Pareto	1_Madrid	2_Sevilla	3_México	4_Lima	5_B.A.
Índice de Pareto	.833	.876	.871	.778	.805
M: Punto de mitad (50%)	3	2	3	3	2
N: Número de datos	9	8	7	10	10
M / N (%)	33 %	25 %	43 %	30 %	20 %
A: Acumulación en mitad	15	13	12	8	10
T: Frecuencia total	24	24	17	15	19
A / T (%)	63 %	54 %	71 %	53 %	53 %
Punto de 50%	3	2	3	3	2
Punto de 80%	5	4	4	8	7
Punto de 95%	8	7	7	10	10



5.11. Diversidad de frecuencia

5.11.1. Índice de Tipo (IT)

Para formular una escala "externa y general" que sirva de medida de diversidad léxica, proponemos utilizar la conocida ecuación de Zipf (1936: 44-48, 1949: 22-27), que explica la relación general que existe entre el Rango (R) y la Frecuencia (F) en las frecuencias de palabras en los textos grandes:

R(ango) * F(recuencia) = C(onstante)

Esta fórmula muestra la proporción inversa que se observa entre el Rango y la Frecuencia. Por ejemplo, si la palabra más frecuente (rango 1, artículo definido *el*) presenta la frecuencia de 50, la segunda palabra (rango 2) en teoría ofrecería la de 25 (50 / 2), la tercera (rango 3), 16 (50 / 3) y la cuarta (rango 4), 12 (50 / 4), y así sucesivamente. En la práctica, observamos unos residuos de error grandes, lo cual, sin embargo, no implica necesariamente su invalidez de la fórmula R * F = C como una tendencia general. Por consiguiente, no la rechazamos tajantemente por los residuos que hay en su aplicación, sino más bien la adoptamos para medir la diferencia que se presenta entre los valores teóricos derivados de la fórmula de Zipf y los observados en la práctica en el

texto objeto, que creemos que sirven para evaluar las características propias de los textos. Por medio de la fórmula de Zipf, R* F = C, podremos comparar los distintos números de tipos (Tipo) basándonos en la misma escala externa y general.

Pero antes de proceder a explicar el método de utilización de la fórmula, conviene confirmar en la siguiente tabla una propiedad única y útil de la misma fórmula:

R	F = C / R	Int(F)	R * F = C
R(1): 1	F(1) = 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
R(n): 50	F(n): 1.00	1	50
51	0.98	0	50

Tabla 1.3. Rango (R), Frecuencia (F) y Constante (C) en la fo/rmula de Zipf (1)

La columna R representa el Rango (1, 2, 3, ...); la de F = C / R, el resultado de la división C / R según la fórmula de Zipf, la de Int(F) indica la parte del número entero de la misma división y la de R * F = C muestra que la multiplicación de R * F igual a C, que es constante.

El punto importante que destacamos en la tabla es que la Frecuencia del Rango 1 (F(1) = 50), es decir, el Máximo (M) coincide con el lugar del último Rango (R(n) = 50), el que equivale al número de tipos (Tipo = 50), coincidente con el C(onstante):

F(1) = M = R(n) = Tipo = C

Por ejemplo, si la Frecuencia de la palabra del Rango 1 (F(1) = Máximo: M) es 50, el último rango, R(n), es decir, el número de Tipo es necesariamente 50, que coincide con el Constante, que es 50. De esta manera, al saber la máxima frecuencia (F(1) = M), podemos determinar el último rango R(n), el número de Tipo y la cantidad de Constante, al mismo tiempo.

En realidad, es difícil imaginar inmediatamente que el Máximo, por ejemplo, la frecuencia del artículo *el*, coincida con el número de palabras diferentes (Tipo: *el, que, y*, ...). Sin embargo, todos los experimentos, realizados bajo la condición de que la distribución de frecuencia obedezca a la fórmula de Zipf, muestran que el máximo M equivale al número de Tipo. Veamos sus

razones prácticas y teóricas.

Podemos entender fácilmente que la máxima frecuencia, F(1) = M, coincide necesariamente con el último rango, R(n). Imaginemos que vamos a descender el rango, uno por uno (1, 2, 3, ...), del denominador R dentro de la fórmula de Zipf, F = C / R. En la tabla observamos que el resultado de la división llega a 1 cuando el rango R es 50, puesto que al pasar el rango a 51, el resultado de la división va a ser menos de 1. Como la frecuencia debe ser un número entero, sin decimales, la serie tiene que terminar en el mismo rango 50.

Esta coincidencia entre el máximo M y el último rango R(n), es decir, el número de Tipos no se limita al Máximo de 50, sino se presenta en todos los valores del Máximo, 1, 2, ..., 500, 9000, ... Como esta es la propiedad importante de la fórmula de Zipf, creemos conveniente presentar una demostración matemática para asegurarnos de su validez universal:

$R * F = C$ ← Fórmula de Zipf

$R(1) * F(1) = C$ ← R(1): rango 1, F(1): frecuencia de la palabra del rango 1

$1 * F1 = C$ ← R(1) = 1

$1 * M = C$ ← F(1) = M: máxima frecuencia

$M = C$ ← Máxima frecuencia = constante

$F = C / R$ ← Fórmula de Zipf: R * F = C

$F = M / R$ ← M = C

$F = M / R \geq 1$ ← La frecuencia debe ser igual o más de 1.

$M \geq R$ ← Multiplicar ambos lados por R, que es positivo

$R \leq M$ ← Significa que el rango es igual o menos de M.

$R(n) = M$ ← Significa que el último rango = frecuencia máxima.

$Tipo = M$ ← R(n) = Tipo: número de tipos

Por lo tanto, el número de Tipos que hay en la distribución de frecuencia que obedece a la fórmula de Zipf, Tipo.z. resulta ser necesariamente la frecuencia máxima, M:

$$Tipo.z. = M$$

Este número de Tipos calculado de acuerdo con la fórmula de Zipf, Tipo.z., equivale a la máxima frecuencia, de modo que podemos utilizarlo para suponer la cantidad teórica de tipos a partir de la frecuencia máxima del texto. Pensamos utilizar la fórmula del Índice de Tipo (IT) siguiente:

$$IT = Tipo / (Tipo + Tipo.z.) = Tipo / (Tipo + M)$$

donde Tipo es cantidad de palabras tipo del texto, Tipo.z. es cantidad de tipos

diferentes de acuerdo con la fórmula de Zipf y M es la frecuencia máxima del texto, concretamente la frecuencia de la forma del artículo *el*. El Índice de tipo (IT) posee las siguientes propiedades:

IT → 0	cuando Tipo → 0
IT = 0.5	cuando Tipo = Tipo.z
IT → 1	cuando Tipo → ∞ y/o Tipo.z. → 0 (M → 0)

Como veremos más adelante (1.4.), la frecuencia máxima M es generalmente proporcional a la cantidad Total de palabras en texto. Por esta razón, es fiable para utilizar como un elemento de la función que deriva el Índice de tipo (IT): Tipo / (Tipo + M).

5.11.2. Índice de Hápax (IH)

Otro índice de diversidad léxica puede ser el número de las palabras de la frecuencia única, las palabras que aparecen solo una vez en el texto. Son las palabras estadísticamente "raras" y, por esta razón, su número representa la diversidad léxica, es decir, cuanto más se presenta el número de las palabras raras, más alto sería el grado de diversidad léxica (Baker *et al*., 2006: 81). Al contrario, cuanto mayor es el número de las palabras repetidas, más bajo sería el grado de diversidad léxica.

En la lingüística de corpus, las palabras de la frecuencia única se llama *hápax legomenon* (gr. 'dicho una vez'), abreviado en *hápax*[39], de modo que utilizamos la mayúscula Hápax para representar el número de hápax dentro del texto. Para conocer el número de hápax dentro de un texto, hay que ir contando las palabras de la frecuencia única, lo que se lleva a cabo en el sistema analizador automático, LYNEAL. En cambio, si las frecuencias de palabras obedecen a la mencionada fórmula de Zipf, R * F = C, el mismo número Hápax se deriva inmediatamente de la frecuencia máxima M, de la manera parecida a la de Tipo.z., que acabamos de ver en la sección anterior. Veámoslo a continuación.

La siguiente tabla representa la relación que hay entre el rango (R) y la frecuencia (F) de acuerdo con la fórmula de Zipf. Es la misma tabla que la anterior. Únicamente se han agregado las cifras que se encuentran alrededor del rango 25: R(25), el punto medio de rangos:

[39] Registrado con acento en la primera sílaba en el Diccionario de la lengua española de la Real Academia Española (2014): "m. Ling. En lexicografía o en crítica textual, voz registrada una sola vez en una lengua, en un autor o en un texto."

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

Tabla 1.4. Rango (R), Frecuencia (F) y Constante (C) en la fórmula de Zipf (2)

En esta tabla observamos que las palabras del rango 26 al 50, en total 25, son de la frecuencia única (hápax). De ahí intuimos que la frecuencia única equivale a la mitad de la frecuencia máxima (M), lo que se comprueba en varios experimentos con distintas cantidades de la frecuencia máxima: 50, 80, 100, 500, ... El resultado es siempre el mismo: el número de hápax (Hápax) es la mitad de la frecuencia máxima (M)[40]. H = M / 2. Es lógico si nos fijamos en el punto del rango 25, R(25), donde la frecuencia es 50 / 25 = 2, según la fórmula de Zipf, C / R = F. Al sobrepasar el rango 25, empieza la frecuencia 1 en número entero. Desde 26 hasta 50, los casos suman a 25, mitad del máximo (M). A pesar de poseer la evidencia empírica, para conseguir la validez universal de este cálculo, probamos a demostrarlo matemáticamente:

$F = C / R$ ← Fórmula de Zipf: $R * F = C$

$F = M / R$ ← M = C (demostrado anteriormente)

$F = M / R \geq 2$ ← La frecuencia de no hápax debe ser igual o más de 2.

$M \geq 2\ R$ ← Multiplicamos ambos lados por R, que es positivo.

$R \leq M / 2$ ← Dividimos ambos lados por 2 e intercambiamos los dos lados.

Hápax.z. = $Rn - M / 2 = M - M / 2 = M / 2$

← Hápax.z. debe ser fuera del ámbito de $R \leq M / 2$

Por consiguiente[41]:

[40] Cuando el máximo es impar, por ejemplo, 51, el número de hápax va a ser valor redondeado por exceso: 26.

[41] Cuando la máxima frecuencia (M) es impar, Hápax.z. se presenta con un decimal 0.5, que parece conveniente redondear por exceso. Sin embargo, el propósito de aquí no es obtener el Hápax.z. mismo, sino el Índice de Hápax (IH)

Hápax.z = M / 2

Conseguido el valor de Hápax según la fórmula de Zipf, Hápax.z., lo utilizamos para calcular el segundo índice de diversidad léxica, en forma del «Índice de Hápax» (IH):

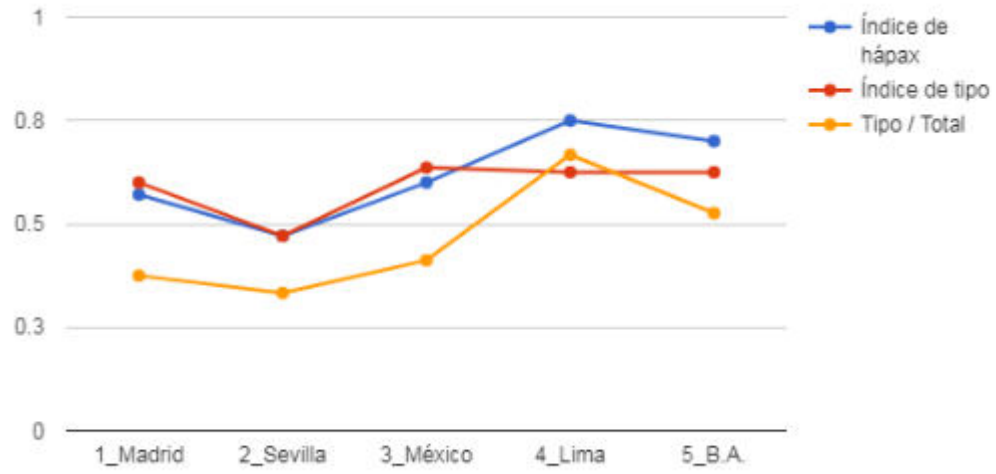
IH = Hápax / (Hápax + Hápax.z.) = Hápax / (Hápax + M / 2)

De la fórmula arriba expuesta, se deduce que el Índice de Hápax (IH) vacila en el rango de [0, 1):

IH → 0 cuando Hápax → 0
IH = 0.5 cuando Hápax = Hápax.z.
IH → 1 cuando Hápax → ∞ y/o Hápax.z.→0 (M→0)

El rango no está limitado en 1, puesto que Hápax no puede ser infinito; ni Hápax.z. ni M pueden ser cero (0). Se tratan de las aproximaciones al infinito y al cero (0).

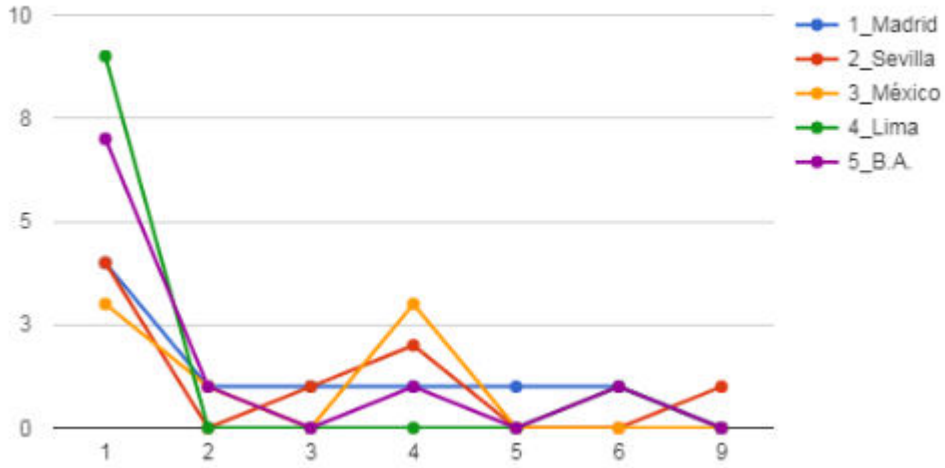
→	Diversidad de frecuencia	1_Madrid	2_Sevilla	3_México	4_Lima	5_B.A.	*Resetear*
1	Total	24	24	17	15	19	Total
2	Tipo	9	8	7	10	10	Tipo
3	Máximo	6	9	4	6	6	Máximo
4	Hápax	4	4	3	9	7	Hápax
5	Tipo / Total	0.375	0.333	0.412	0.667	0.526	Tipo / Total
6	Índice de tipo	0.600	0.471	0.636	0.625	0.625	Índice de tipo
7	Índice de hápax	0.571	0.471	0.600	0.750	0.700	Índice de hápax
Resetear		1_Madrid	2_Sevilla	3_México	4_Lima	5_B.A.	



de manera precisa, lo dejamos sin redondear.

5.11.3. Espectro de frecuencia: Número

FA	1_Madrid	2_Sevilla	3_México	4_Lima	5_B.A.
1	4	4	3	9	7
2	1	0	1	0	1
3	1	1	0	0	0
4	1	2	3	0	1
5	1	0	0	0	0
6	1	0	0	1	1
9	0	1	0	0	0

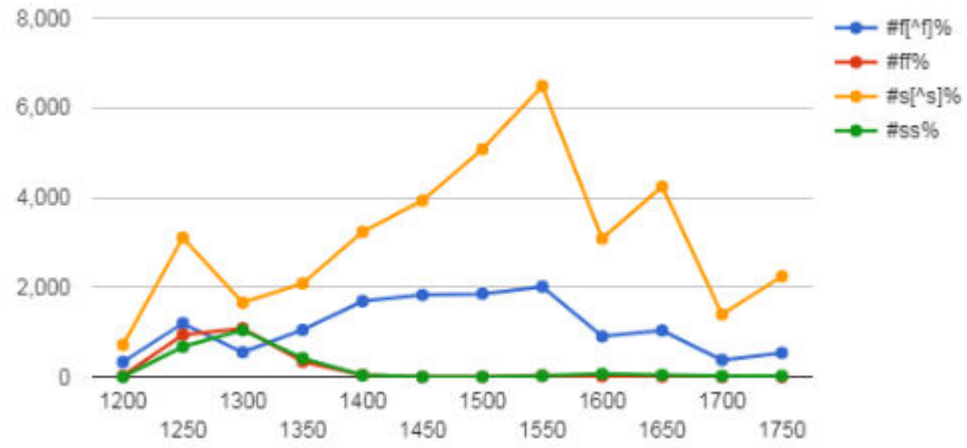


5.11.4. Espectro de frecuencia: Letra

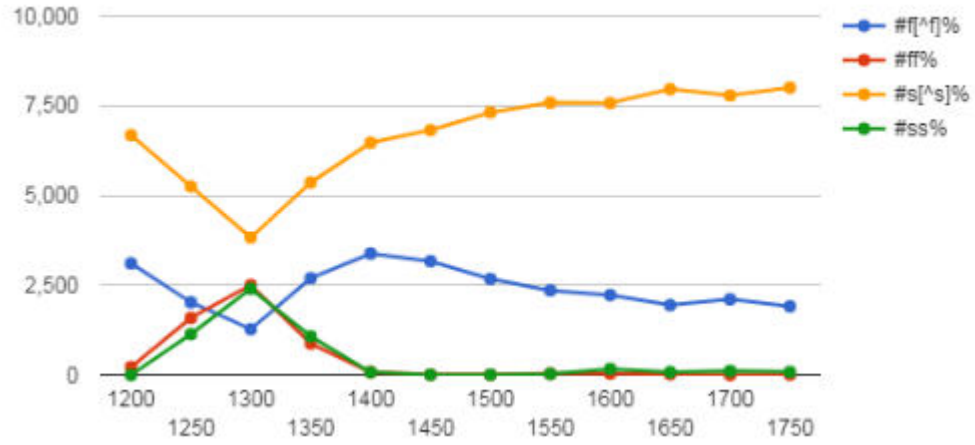
FA	1_Madrid	2_Sevilla	3_México	4_Lima	5_B.A.
1	facilísima, facultad, falda, faldas	fachada, facultad, falda, faltar	falda, faldas, faltar	facilísimo, factura, falda, falta, faltan, faltar, familia, fastidies, favorito	fabricarnos, facilidad, factura, facultad, fanático, fatal, favorito
2	facu	-	facultad	-	facu
3	fatal	fatal	-	-	-
4	familia	familia, favor	falta, familia, favor	-	familia
5	falta	-	-	-	-
6	favor	-	-	favor	favor
9	-	falta	-	-	-

5.12. Distribución de ocurrencia

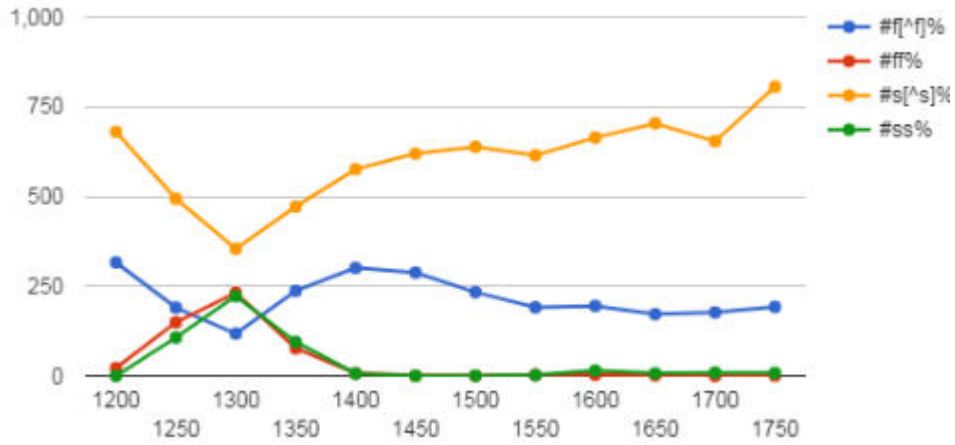
→	FA	1200	1250	1300	1350	1400	1450	1500	1550	1600	1650	1700	1750
1	#f[^f]%	332	1195	551	1048	1689	1826	1854	2013	905	1036	377	535
2	#ff%	23	938	1083	340	40	5	5	29	15	10	0	2
3	#s[^s]%	714	3101	1656	2085	3233	3930	5076	6482	3083	4242	1394	2242
4	#ss%	0	671	1039	418	37	2	4	24	68	42	20	23



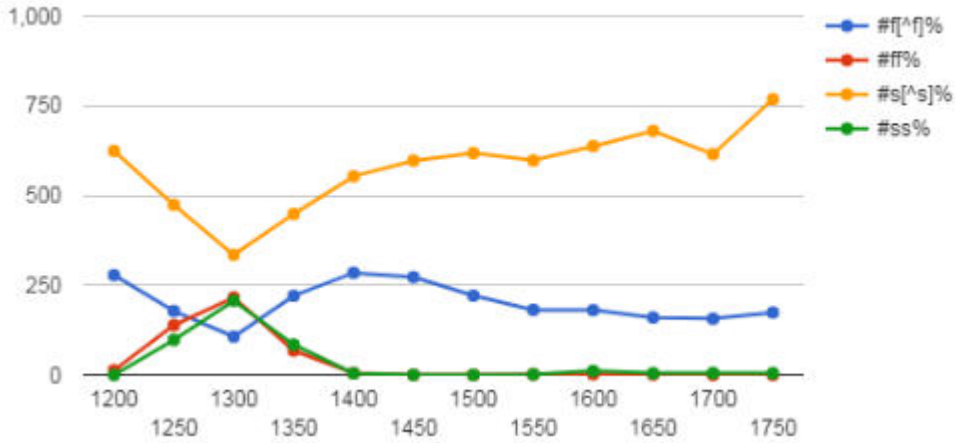
→	FR.Atr.:10000	1200	1250	1300	1350	1400	1450	1500	1550	1600	1650
1	#f[^f]%	3105.7	2023.7	1272.8	2693.4	3378.7	3168.5	2671.9	2354.9	2223.0	1943.7
2	#ff%	215.2	1588.5	2501.7	873.8	80.0	8.7	7.2	33.9	36.8	18.8
3	#s[^s]%	6679.1	5251.5	3825.4	5358.5	6467.3	6819.4	7315.2	7583.1	7573.1	7958.7
4	#ss%	0.0	1136.3	2400.1	1074.3	74.0	3.5	5.8	28.1	167.0	78.8



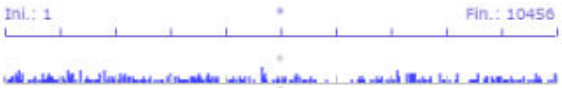
→	FN.Pl.:10000	1200	1250	1300	1350	1400	1450	1500	1550	1600	1650	1700	1750
1	#f[^f]%	316.2	190.2	117.8	236.9	300.9	287.7	233.1	190.9	194.9	171.9	176.8	192.3
2	#ff%	21.9	149.3	231.6	76.9	7.1	0.8	0.6	2.7	3.2	1.7	0.0	0.7
3	#s[^s]%	680.0	493.7	354.2	471.4	575.9	619.2	638.1	614.6	664.0	703.9	653.8	805.7
4	#ss%	0.0	106.8	222.2	94.5	6.6	0.3	0.5	2.3	14.6	7.0	9.4	8.3



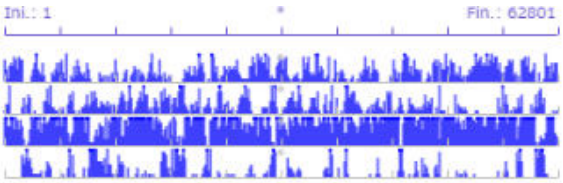
→	FP.PI.:10000	1200	1250	1300	1350	1400	1450	1500	1550	1600	1650	1700	1750
1	#f[^f]%	277.9	177.8	106.6	220.4	284.4	272.5	220.8	181.2	180.3	159.9	156.6	173.7
2	#ff%	12.7	138.3	215.8	67.6	4.8	0.2	0.2	1.7	1.6	0.7	0.0	0.1
3	#s[^s]%	624.1	473.8	334.6	448.3	553.3	597.1	618.1	597.5	637.4	679.9	615.0	768.2
4	#ss%	0.0	97.5	206.7	84.2	4.3	0.0	0.1	1.3	10.9	4.7	5.2	4.8



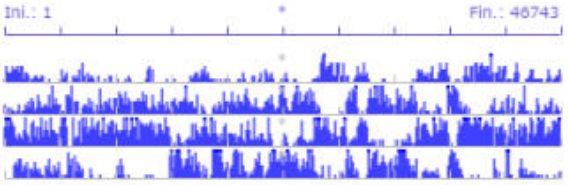
Grupo	Forma	Frec.	Disp.	Unif.	Uso
1200	#f[^f]%	332	.912	.585	242.5
1200	#ff%	23	.639	.391	11.5
1200	#s[^s]%	714	.941	.609	540.5
1200	#ss%	***	***	***	***



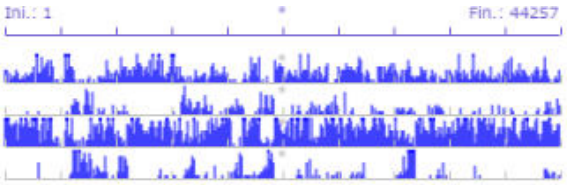
Grupo	Forma	Frec.	Disp.	Unif.	Uso
1250	#f[^f]%	1195	.909	.529	828.7
1250	#ff%	938	.885	.489	617.1
1250	#s[^s]%	3101	.968	.5922	347.5
1250	#ss%	671	.833	.332	352.9



Grupo	Forma	Frec.	Disp.	Unif.	Uso
1300	#f[^f]%	551	.856	.469	349.1
1300	#ff%	1083	.934	.550	776.2
1300	#s[^s]%	1656	.913	.553	1176.7
1300	#ss%	1039	.833	.457	641.1



Grupo	Forma	Frec.	Disp.	Unif.	Uso
1350	#f[^f]%	1048	.925	.562	755.6
1350	#ff%	340	.777	.389	186.9
1350	#s[^s]%	2085	.954	.586	1558.9
1350	#ss%	418	.727	.336	206.6



Grupo	Forma	Frec.	Disp.	Unif.	Uso
1400	#f[^f]%	1689	.974	.619	1311.5
1400	#ff%	40	.719	.452	22.8
1400	#s[^s]%	3233	.970	.604	2474.6
1400	#ss%	37	.392	.245	11.5