Análisis de datos numéricos para la filología digital: Problemas de cuantificación y algunas propuestas de solución

Hiroto Ueda (Universidad de Tokio)

uedahiroto@jcom.home.ne.jp

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

(Google: hiroto ueda)

1. Problema de la frecuencia absoluta

¡No se pueden comparar las frecuencias absolutas, porque las bases de suma (horizontal) y de total (vertical) son distintas!

→¿Solución: frecuencia relativa?

«Corpus Histórico del Español Norteño» (CORHEN)

Fig. 1: Palabras con <i>, <j> (CORHEN)

Palabras	**925**	**950**	**975**	**1000**	**1050**	**1075**	**1100**	**1125**	**1150**	**1175**	**1200**	**1225**	**1250**	**1275**	**Suma**
i	306	177	132	672	1,267	653	876	1,579	1,090	2,958	5,338	1,624	1,819	1,175	19,666
j	17	2	2	19	57	16	21	24	76	180	592	293	144	127	1,570
Total	323	179	134	691	1,324	669	897	1,603	1,166	3,138	5,930	1,917	1,963	1,302	21,236

Fig. 2: Relativización interna

Palabras	925	950	975	1000	1050	1075	1100	1125	1150	1175	1200	1225	1250	1275	Hor.
con <i>	.016	.009	.007	.034	.064	.033	.045	.080	.055	.150	.271	.083	.092	.060	
con <j>	.011	.001	.001	.012	.036	.010	.013	.015	.048	.115	.377	.187	.092	.081	
Palabras	925	950	975	1000	1050	1075	1100	1125	1150	1175	1200	1225	1250	1275	Ver.
con <i>	.947	.989	.985	.973	.957	.976	.977	.985	.935	.943	.900	.847	.927	.902	
con <j>	.053	.011	.015	.027	.043	.024	.023	.015	.065	.057	.100	.153	.073	.098	

Frecuencia relativa interna horizontal
= Frecuencia absoluta / Suma (horizontal)

Frecuencia relativa interna vertical
= Frecuencia absoluta / Total (vertical)

Relativización externa: se realiza basándose en las frecuencias que existen fuera del cuadro.

Fig. 3: Grafías <i, I, j, J, y, Y> (por mil letras)

Letra	1.V	2.Gq	3.Ga	4.Gc	a0925	a0950	a0975	a1000	a1050	a1075	a1100	a1125	a1150	a1175	a1200	a1225	a1250	a1275
i	163.9	192.6	181.0	146.1	154.3	172.5	172.2	150.3	149.8	185.4	198.2	197.4	200.4	191.6	189.9	165.5	150.5	136.0
I	26.3	0.9	0.7	0.4	30.9	36.9	24.8	36.5	28.7	14.1	13.3	2.3	1.2	0.9	0.6	0.4	0.5	
j	**4.4**	**5.5**	**12.7**	**11.7**	**5.9**	**1.8**	**2.5**	**3.6**	**5.6**	**3.7**	**4.0**	**1.2**	**5.5**	**7.8**	**14.0**	**22.0**	**7.7**	**13.1**
J	**0.7**	**2.3**	**3.7**	**2.5**	**2.4**			**1.0**	**0.8**	**0.2**		**1.2**	**5.8**	**2.0**	**3.8**	**5.2**	**4.3**	**1.0**
y	0.2	5.7	5.8	29.3					0.7			0.1	0.6	1.4	0.6	9.4	30.1	33.5
Y	0.1	0.1	0.1	0.5		0.9					0.2					0.4	0.6	0.1

Fig. 4: Frecuencia relativa de palabras con <i> / <j> (por mil palabras)

Palabras con i / j	1.V	2.Gq	3.Ga	4.Gc	a0925	a0950	a0975	a1000	a1050	a1075	a1100	a1125	a1150	a1175	a1200	a1225	a1250	a1275
i	968.0	967.9	924.5	922.8	954.5	986.9	982.6	972.9	955.6	975.9	975.1	992.4	967.0	953.4	919.4	872.3	950.7	908.7
j	32.0	32.1	75.5	77.2	45.5	13.1	17.4	27.1	44.4	24.1	24.9	7.6	33.0	46.6	80.6	127.7	49.3	91.3

2. Problema de la frecuencia relativa

¡No es lo mismo uno contra uno que uno contra muchos!

→ ¿Solución: «Frecuencia prominente»?

Fig. 5: Frecuencia relativa (horizontal)

Palabras	925	950	975	1000	1050	1075	1100	1125	1150	1175	1200	1225	1250	1275	Hor.
con <i>	.016	.009	.007	.034	.064	.033	.045	.080	.055	.150	.271	.083	.092	.060	
con <j>	.011	.001	.001	.012	.036	.010	.013	.015	.048	.115	.377	.187	.092	.081	

El valor relativo entre muchos no se destaca. Veánse <i> en 925 (306 / 19,666 = .016). En cambio, la frecuencia relativa vertical es sumamente grande (306 / 323 = .947):

Fig. 6: Frecuencia relativa (vertical)

Palabras	925	950	975	1000	1050	1075	1100	1125	1150	1175	1200	1225	1250	1275	Ver.
con <i>	.947	.989	.985	.973	.957	.976	.977	.985	.935	.943	.900	.847	.927	.902	
con <j>	.053	.011	.015	.027	.043	.024	.023	.015	.065	.057	.100	.153	.073	.098	

Para hacer la comparación equitativa libre de la dimensión de cuadro, proponemos utilizar la «Frecuencia prominente» (F.p.):

$$F.p. = f * (n - 1) / [f * (n - 1) + (s - f)]$$

donde *f* es frecuencia absoluta, *n* es el número de miembros, *s* es suma. Por ejemplo, el valor F.p. de

<i>:a925 se calcula: F.p. (<i>:a925) = 306 * 13 / 306 * 13 + 19666 - 306 = .170. De esta manera, el valor de 306 se multiplica por el número del resto (13) para ser igual de condición de comparación. Esta cifra multiplicada se compara con totalidad del resto, que corresponde a la suma menos la frecuencia del valor en cuestión (<i>:a925). El cuadro entero de F.p. es:

Fig. 6: Frecuencia prominente (horizontal)

Palabras	925	950	975	1000	1050	1075	1100	1125	1150	1175	1200	1225	1250	1275	Hor.prm
con <i>	.170	.106	.081	.315	.472	.309	.377	.532	.433	.697	.829	.539	.570	.452	
con <j>	.125	.016	.016	.137	.329	.118	.150	.168	.398	.627	.887	.749	.568	.534	

La misma operación se puede realizar verticalmente:

Fig. 7: Frecuencia prominente (vertical)

Palabras	925	950	975	1000	1050	1075	1100	1125	1150	1175	1200	1225	1250	1275	Ver.prm
con <i>	.947	.989	.985	.973	.957	.976	.977	.985	.935	.943	.900	.847	.927	.902	
con <j>	.053	.011	.015	.027	.043	.024	.023	.015	.065	.057	.100	.153	.073	.098	

que resulta ser el mismo cuadro de la frecuencia relativa vertical, puesto que se trata de una comparación entre dos miembros, es decir, se compara uno contra otro, sin necesidad de multiplicar con el número del otro, o se multiplica por el número 1, que es el número del resto.

3. Problema del cuadro bidimensional

Los cuadros bidimensionales ofrecen dos puntos de vista: el horizontal y el vertical. Para saber las visicitudes históricas de una letra, se toma en consideración el horizontal, y para saber la proporción de una letra en un mismo año, se toma el vertical. Sin embargo, para conocer el valor que cobra una letra determinada en un año determinado dentro de la totalidad del cuadro, no podemos elegir uno de los dos, sino que tenemos que buscar otra solución para considerar los dos.

→ ¿Solución: «Promedio fraccional»?

Para tomar en consideración tanto el valor relativo horizontal como el vertical al mismo tiempo, buscamos la manera de calcular un promedio de los dos. Creemos que un promedio normal no es adecuado, puesto que se trata de dos valores relativizados, ratios obtenidos de división. Por ejemplo, el promedio de la velocidad de unos movimientos o la densidad de sal del agua salada no se puede calcular sumando las velocidades o densidades dividiendo por el número de objetos, puesto que la base de división puede ser distinta. El promedio de la velocidad de caminata de un día que una persona ha recorrido 6 kilómetros en una hora (6 km/hora) y la de

otro día que ha recorrido 16 kilómetros en dos horas (8 km/hora) no es (6 + 8) / 2 = 7 km/h, sino (6 + 16) / (1 + 2) = 7.3 km/h. Esta manera de obtener el promedio de las dos fracciones la denomamos un «Promedio fraccional» (P.f.).

P.f. = (suma de numeradores) / (suma de denominadores)

Calculando el promedio de la frecuencia relativa horizontal y la vertical, utilizando el promedio fraccional obtenemos el cuadro siguiente que demuestra el «Valor matricial» de frecuencia:

Fig. 8: Valor matricial de frecuencia relativa

Palabras	925	950	975	1000	1050	1075	1100	1125	1150	1175	1200	1225	1250	1275	Matr.
con <i>	.031	.018	.013	.066	.121	.064	.085	.148	.105	.259	.417	.150	.168	.112	
con <j>	.018	.002	.002	.017	.039	.014	.017	.015	.056	.076	.158	.168	.082	.088	

Lo mismo puede hacerse con las frecuencias prominentes:

Fig. 9: Valor matricial de frecuencia prominente

Palabras	925	950	975	1000	1050	1075	1100	1125	1150	1175	1200	1225	1250	1275	Matr.prm
con <i>	.181	.113	.086	.331	.490	.325	.395	.550	.450	.710	.834	.554	.586	.469	
con <j>	.113	.016	.016	.107	.223	.092	.108	.097	.292	.367	.568	.586	.383	.404	

4. Problema de la(s) causa(s) y el efecto

Estamos ante un cuadro que ofrece multitud de las posibles causas del único efecto. En nuestro caso la selección de una letra dentro de las tres posibilidades: <i>, <j> y <y>. Se han enumerado el fonema, el entorno gráfico, el año y el tipo de letra. ¿De qué manera se destacan las causas más importantes para la aparición de la letra <j>? →¿Solución: «Condiciones combinadas»?

Proponemos efectuar la búsqueda de todas las posibles combinaciones de causas y su frecuencia de correspondencias con el efecto:

Fig. 10: Condición simple

Prm.Sg.Cnd.	i.cnt	y.cnt	j.cnt	i.prm	y.prm	j.prm	i.mlt	y.mlt	j.mlt
1:/i/	8895	198	553	.974	.074	.191	8660.831	14.643	105.551
1:/j/	**157**	**1**	**275**	**.064**	**.004**	**.604**	**10.007**	**.004**	**166.209**
1:/y/	54	468	9	.022	.877	.026	1.197	410.542	.234
2:#V_V	**448**	**470**	**304**	**.399**	**.874**	**.746**	**178.926**	**410.758**	**226.708**
2:#_[^p]	957	24	19	.621	.172	.129	593.881	4.125	2.447
2:[^p]_#	**2217**	**113**	**251**	**.811**	**.344**	**.547**	**1797.061**	**38.828**	**137.172**
2:[^p]_i	58			.082			4.776		
2:[^p]_m	459		1	.426		.011	195.670		.011
2:[^p]_n	1137	13	9	.666	.092	.060	756.985	1.194	.539
2:[^p]_u	521	42	16	.458	.336	.139	238.450	14.112	2.229
2:[p]_[p]	698		17	.537		.136	374.834		2.304
2:i_[^p]	**37**		**89**	**.054**		**.613**	**1.981**		**54.601**
2:m_[^p]	654		4	.520		.036	339.997		.145
2:n_[^p]	573		13	.484		.115	277.439		1.498
2:t_V	**479**	**1**	**100**	**.434**	**.011**	**.535**	**208.123**	**.011**	**53.496**
2:u_[^p]	868	4	14	.595	.035	.104	516.853	.140	1.451

Fig. 11: Condición doble

Prm.Db.Cnd.	i.cnt	y.cnt	j.cnt	i.prm	y.prm	j.prm	i.mlt	y.mlt	j.mlt
1:/i/ + 2:#V_V	237	1	20	.299	.017	.233	70.865	.017	4.655
1:/i/ + 2:#_[^p]	957	24	19	.651	.192	.145	623.451	4.601	2.747
1:/i/ + 2:[^p]_#	**2217**	**113**	**251**	**.830**	**.374**	**.579**	**1840.642**	**42.299**	**145.415**
1:/i/ + 2:[^p]_i	58			.093			5.395		
1:/i/ + 2:[^p]_m	459		1	.459		.012	210.786		.012
1:/i/ + 2:[^p]_n	1137	13	9	.695	.104	.068	789.990	1.347	.611
1:/i/ + 2:[^p]_u	521	42	16	.491	.366	.156	255.790	15.389	2.498
1:/i/ + 2:[p]_[p]	698		17	.570		.152	397.860		2.583
1:/i/ + 2:i_[^p]	**37**		**89**	**.061**		**.645**	**2.247**		**57.373**
1:/i/ + 2:m_[^p]	654		4	.553		.041	361.705		.165
1:/i/ + 2:n_[^p]	573		13	.518		.130	296.560		1.685
1:/i/ + 2:t_V	**479**	**1**	**100**	**.468**	**.013**	**.568**	**223.954**	**.013**	**56.798**
1:/i/ + 2:u_[^p]	868	4	14	.627	.040	.117	544.382	.159	1.634
1:/i/ + 3:a0925	143		17	.371		.395	53.018		6.717
1:/i/ + 3:a0950	92		2	.274		.074	25.215		.148

Fig. 12: Condición triple

Prm.Tp.Cnd.	i.cnt	y.cnt	j.cnt	i.prm	y.prm	j.prm	i.mlt	y.mlt	j.mlt
1:/j/ + 3:a1075 + 4:1.V	3			.023			.069		
1:/j/ + 3:a1100 + 4:1.V	1			.008			.008		
1:/j/ + 3:a1100 + 4:2.Gq	1			.008			.008		
1:/j/ + 3:a1125 + 4:2.Gq	9			.066			.591		
1:/j/ + 3:a1150 + 4:2.Gq	2			.015			.031		
1:/j/ + 3:a1150 + 4:3.Ga			7			.375			2.622
1:/j/ + 3:a1175 + 4:2.Gq	2		1	.015		.078	.031		.078
1:/j/ + 3:a1175 + 4:3.Ga	6		2	.045		.144	.268		.289
1:/j/ + 3:a1200 + 4:2.Gq	10		2	.072		.144	.724		.288
1:/j/ + 3:a1200 + 4:3.Ga	23		18	.152		.603	3.499		10.851
1:/j/ + 3:a1200 + 4:4.Gc	4		4	.030		.253	.121		1.013
1:/j/ + 3:a1225 + 4:2.Gq	3		19	.023		.622	.068		11.812
1:/j/ + 3:a1225 + 4:3.Ga	**6**		**64**	**.044**		**.854**	**.267**		**54.634**
1:/j/ + 3:a1250 + 4:2.Gq	8		8	.059		.404	.470		3.234
1:/j/ + 3:a1250 + 4:3.Ga	**30**		**82**	**.189**		**.881**	**5.661**		**72.257**
1:/j/ + 3:a1250 + 4:4.Gc	19		24	.129		.672	2.450		16.126
1:/j/ + 3:a1275 + 4:2.Gq			10			.462			4.619
1:/j/ + 3:a1275 + 4:3.Ga	2		7	.015		.374	.031		2.618
1:/j/ + 3:a1275 + 4:4.Gc	28	1	27	.179	.090	.696	5.017	.090	18.780

* Programa: NUMEROS.xlsm → Página web