SABENCIA

Sociedá asturiana de les ciencies

ESTAYA MATEMÁTIQUES

XUGANDO COLES SOPES DE LLETRES

R. FERNÁNDEZ[1] & X. MARTINO[2]

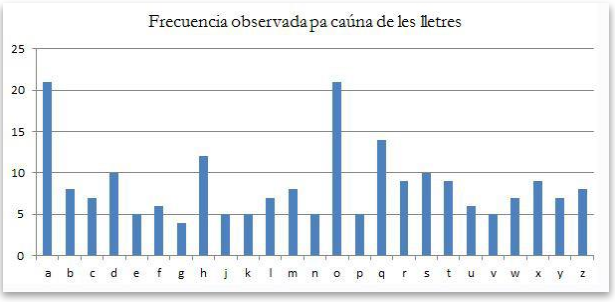
1. DEPARTAMENTU METEOROLOXÍA SABENCIA
2.DEPARTAMENTU FILOLOXÍA SABENCIA

Toos comimos daqué vegada sopes de lletres. Cuando en casa hai neños ye dalgo qu'hai que tener na alacena. La verdá, que ye bien entretenío comer sopes de lletres con neños, puedes formar pallabres mentanto comes, y buscar el to nome ente'l caldu.

Yá de grandín dime cuenta en que depende la garfellada unes lletres apaecíen más qu'otres, anque siempre camenté qu'eses diferencies yeren coses del azar; poro vi qu'eso nun yera asina siempre. Una tardiquina, sentáu na antoxana casa, xugando a les cocinetes cola mio fía.

Garráramos unos puñaos de fideos de sopes de lletres crudos pa xugar, y cayí na cuenta que na distribución de les lletres, la A y la O apaecíen con más frecuencia, o eso paecía, ¿sedría por azar, o ye que'l fabricante metía más "as" y "os"? Nun quedaba otra qu'usar l'análisis estadísticu pa retrucar la entruga.

Garré un puñáu de lletres (215exautamente), y cunté cuántes vegaes apaecía cada lletra, ves el resultáu na gráfica.



Como ves la A y la O apaecieron munches más vegaes (n=21) que les otres, depués apaecíen otres con más de 10 repeticiones como la D, H, Q o la S.

Toos sabemos qu'el azar ye caprichosu y munches vegaes engáñanos, polo que tenía qu'asegurame qu'estes diferencies nun yeren causa d'elli, por suerte, problemes como estos soceden en toles estayes de la investigación, asina que cuantayá qu'hai una ferramienta perútil pa caltriar cuándo un socesu ye o non productu del azar, ye l'análisis de frecuencies.

Nesti casu la frecuencia ye'l númberu de vegaes qu'apaez una lletra, y lo que queremos saber ye si eses lletres qu'apaecen más son productu del azar o non.

Podemos entamar reflexonando sobre cuála sedría la distribución de frecuencies si toles lletres salieren el mesmu númberu de vegaes; pa ello calculamos la frecuencia esperada, que ye'l númberu de lletres total (n=215) ente'l númberu de lletres qu'apaecieron

(n=26), o sía 8,27 vegaes, más o menos 8 o 9 vegaes, los datos que tomamos del puñáu del sopa nun s'asemeyen muncho a esto pol azar, poro l'azar a partir de ciertu númberu d'elementos déxase sentir menos, y la distribución aseméyase más a la realidá; les 215 lletres son bastantes p'averamos a lo que socede (yá te cuntaré porqué n'otra entrada), entós les frecuencies son productu del azar y hai diferencia na distribución de delles lletres. Por ciertu, nel análisis quité la lletra I, por que yera perpequeña, y de xugar con elles perdiéranse munches, asina que pa nun sesgar el resultáu quitéles.

Sabemos la distribución esperada, sabemos la observada, porqué nun restamos una de la otra, si les diferencies son grandes algo tendrá qu'haber, ¿non? Nesto ye no que se sofita la prueba de chi cuadráu o de bondá d'axuste, anque p'agrandar estes diferencies eleven al cuadráu la resta y dividen ente la frecuencia esperada:

$$\chi^2 = \frac{(fesp - fobs)^2}{fesp}$$

Asina vemos qu'el resultáu d'unes lletres y otres destrémase enforma, los qu'idearon esti métodu, vieron que si sumaben tolos resultaos d'aplicar la fórmula pa caúna de les lletres, y lo comparaben nuna tabla con resultaos teóricos estadísticos pa dellos porcentaxes d'error podíen saber si les diferencies qu'apaecen son frutu del azar nesi porcentaxe o non. Polo xeneral nun s'almite munchu error, y aveza aceutase que los resultaos nun son frutu del azar a partir del 5%, anque en munchos ámbitos nun s'aceuta hasta'l 1%, esto quier dicir que namás qu'hai una posibilidá ente 100 de qu'una distribución de frecuencies seya por azar, paezme bastante poco, más cuando tenemos tamaños de muestra grandes.

Pal exemplu de les sopes de lletres la suma de tolos resultaos ye 62,53, esti númberu nun diz nada solu, pero cuando lu compares col resultáu de la tabla de resultaos teóricos pa esi númberu de variables (26 lletres) y con 1% d'error (ver apéndiz), vi que yera perdestremáu, 44,31. Cuando'l resultáu ye más altu que'l teóricu sabemos que la distribución nun se debe al azar, sinón qu'hai daqué que produz esa diferencia, quiciabes el fabricante mete más lletres A y O por daqué motivu, como por aforrar en producción.

Apaez otra entruga agora, yeren solo les lletres A y O les qu'apaecen más veces de lo que cabría esperar, o aquelles qu'apaecíen más de 10 vegaes tamién lo faen de forma fortuita. Ye fácil de saber, quitamos les lletres A y O, y volvemos a facer l'análisis, esta vegada con un total 173 lletres y 24 diferentes. El resultáu 23,02, qu'al comparar cola tabla, non solo ye menor al nivel del 1%, sinón que ye menor con un cachu qu'al nivel del 5%, polo xeneral con errores mayores yá nun s'aceuta una hipótesis de qué hai diferencies sentibles ente una y otra distribución, asina que paecía ser que les lletres A y O yeren les culpables d'esa diferencia.

¿Qué socedió cuándo la neña y la muyer fueron pala camina, y quedé yo solu na cocina? Púnxime a cuntar más puñaos, con mesmos resultaos (la I apaecía igual que toes), la marca de sopes metía más lletres A y O qu'otres. Col problema fináu colé pala cama, y a la vera la rapacina dormime tranquilu.

La cosa foi que díes depués asoleyé los resultaos nel mio blogue, y nel facebook de mio entamó un alderique perinteresante sobre'l tema. La verdá que nun m'esperaba que diba adicar tan bien el tema, y que diera en milenta comentarios na muria de facebook, nel mio corréu y en persona, toos sobre si podría tar venceyada la distribución de frecuencies de les lletres de les sopes cola distribución d'usu de les lletres d'una llingua (la marca de sopa yera española, poro sedría'l castellanu). Polo que púnxime a tratar d'atalantar si había o non rellación tala.

Pa tal fin tuvi a la gueta de les frecuencies d'usu de delles llingües del nueso arrodiu, coles que facer comparanza cola frecuencia qu'atopé cuntando los fideos de les sopes de lletres, cuntando que'l país d'orixe del productu fore dalgunu qu'usaren daqué llingua de les que m'informé. Puedes ver na tabla les diferentes frecuencies, a golpe de vista paezse más l'usu ente toles llingües (a lo cabero toes son derivaes, direuta o indireutamente, de la llingua indoeuropea) qu'a la distribución de frecuencies de los fideos.

	castellano	inglés	francés	alemán	sopa de lletres
a	12,53	8,17	7,63	6,51	9,76
b	1,42	1,49	3,26	3,06	3,72
c	4,68	2,78	3,26	3,06	3,25
d	5,86	4,25	3,67	5,08	4,65
e	13,68	12,7	14,71	17,4	2,32
f	0,69	2,22	1,07	1,66	2,79
g	1,01	2,01	0,87	3,01	1,86
h	0,7	6,02	0,74	4,76	5,58
i	6,25	6,97	7,53	7,55	0,93
j	0,44	0,15	0,54	0,27	2,32
k	0,01	0,77	0,05	1,21	2,32
l	4,97	4,02	5,46	3,44	3,25
m	3,15	2,4	2,97	2,53	3,72
n	6,71	6,75	7,02	9,78	2,32
o	8,68	7,51	5,38	2,51	9,77
p	2,51	1,93	3,02	0,79	2,32
q	0,88	0,09	1,36	0,02	6,51
r	6,87	5,99	6,55	7	4,19
s	7,98	6,33	7,95	7,27	4,65
t	4,63	9,06	7,24	6,15	4,19
u	3,93	2,76	6,31	4,35	2,79
v	0,9	0,98	1,63	0,67	2,32
w	0,02	2,36	0,11	1,89	3,25
x	0,22	0,15	0,39	0,03	4,18
y	0,9	1,97	0,31	0,04	3,25
z	0,52	0,07	0,14	1,13	3,72

Distribución de frecuencies de les lletres en distintes llingües y nes sopes de lletres

Poro, como sabes, en ciencia (y na vida en xeneral) nun sirven conxetures sobre lo que nos paez a nós, hai que sofitase en daqué, pa ello, y anque hai preseos más concretos pa ello, polo que decidí usar ún de los que más uso pa entrugues d'esta triba, porque da resultaos nos que podemos enfotar y ye bien cenciellu d'usar, el coeficiente de correllación (que trai cualesquier programa de fueya de cálculu).

Esti coeficiente va dicinos cuánto d'asemeyáu ye una y otra distribución de frecuencia pa caún de los pares de llingües y sopa posibles.

El valor de la correllación ta siempre ente 1 y -1, siendo 1 una correllación direuta, nesti casu les dos llingües tienen la mesma distribución de frecuencies, y -1 negativa, y que 0 sedría que nun s'asemeyen nada, los valores entemedies indiquen el cuánto d'asemeyaes son los pares de llingües, polo xeneral (anque depende de la muestra) si ta perbaxo el 0,5 nun hai venceyu o ye débil, a partir d'ehí tarán más venceyaes a midida que nos averamos al 1 (o -1 nel casu de facelo de forma negativa), anque como nel análisis anterior, hai tables nes que se dan los valores d'aceutación de la hipótesis pa caún de los niveles d'error (ver apéndiz).

Lo qu'atopé foi la tabla que vien darréu, ye lo qu'apaecía a lo primero, tán toes pervenceyaes unes con otres menos la distribución de los fideos:

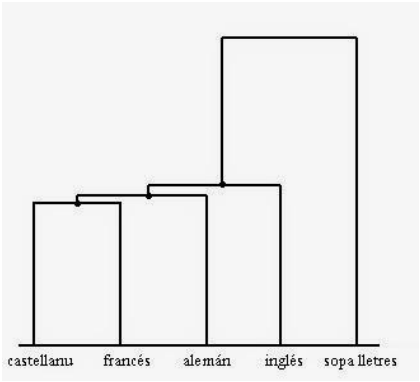
	castellanu	inglés	francés	alemán	sopa de lletres
castellanu	1				
inglés	0,86	1			
francés	0,91	0,87	1		
alemán	0,81	0,89	0,9	1	
sopa de lletres	0,37	0,26	0,09	-0,05	1

Coeficiente de correllación pa caún de los pares de llingües y los fideos

Nun nos tenía que garrar de sustu, qu'el castellanu y el francés seyan les más venceyaes (siempre falando de la distribución d'usu de les lletres, y non de la llingua mesma, cuidáu con esto), siguíu del inglés y l'alemán (nun miré l'italianu, pero de xuru que taba más averáu al castellanu). No que cinca a los fideos de les sopes, nun s'asemeyaba a nenguna llingua, la más averada'l castellanu, pero con un valor perbaxo (0,37), perpoco pa cavilar que tienen la mesma distribución, les otres peor tovía, l'alemán hasta con valores negativos.

Pa comprobar el resultáu fici un análisis de conglomeraos (una triba d'análisis multivariante) que busca atopar los elementos más asemeyaos y rescampla les diferencies ente los grupos, esplicate cómo se fai nesta entrada diba ser enguedeyame muncho, asina que prométote falar d'ello n'otra entrada con un exemplu ilustrativu, porque pal científicu amateur ye perinteresante conocer esti preséu. Bono a lo que diba, fici esti análisis, con resultaos perasemeyaos a los del de correllación, puedes ver el dendrograma resultante, equí a la vera.

Nelli vemos que cuánto más separtaes, n'altor, tán les distribuciones de frecuencies de dos llingües más s'estremen éstes, y que, como ves, tán toes apiñaes alredor d'un mesmu valor, menos la distribución de les sopes de lletres, que ta un cachu grande perriba de toles demás, hasta del castellanu. Entós si la distribución de frecuencies d'apaición de cada lletra na sopa nun sigue patrones de nenguna llingua, ¿qué criteriu sigue la marca? Paezme que sedrá un misteriu que nun vamos poder resolver col métodu científicu, o quiciabes sí...



Dendrograma que fici a lo cabero del análisis de conglomeraos