UCM - Mineria de Datos

ANÁLISIS DE CORRESPONDENCIAS SIMPLES

COMPLEMENTOS DE FORMACIÓN EN TÉCNICAS DE MINERÍA DE DATOS

CAIO FERNANDES MORENO

1. Evaluación de la idoneidad de los datos y sus categorías. Si fuera necesario, incluir categorías suplementarias o agruparlas.

Los datos que se van analizar son: población por sexo, edad (grupos quinquenales) y relación entre el lugar de nacimiento y el de residencia (estudiar en qué medida la población nacida en un lugar reside posteriormente en el mismo (en función de provincia, edad, sexo, etc.)

Enlace para descargar los datos:
http://www.ine.es/jaxi/tabla.do?path=/t20/e244/avance/p01/l1/&file=03003.px&type=pcaxis&L=1

Los datos originales han sido agrupados en tramos de 10 anos, como se puede ver en la tabla abajo:

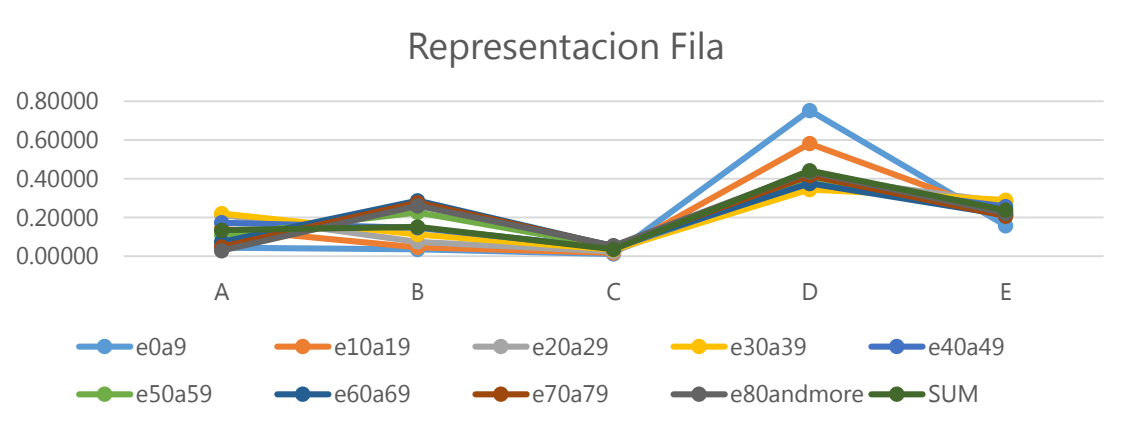
Categoria	e0a9	e10a19	e20a29	e30a39	e40a49	e50a59	e60a69	e70a79	e80andmore
A	214093	623184	1195793	1758321	1287814	643953	343717	169926	70476
B	170776	203497	413168	896341	1091057	1341937	1329901	963722	639246
C	55882	76019	131650	268129	299415	281844	237161	177623	132179
D	3663554	2560996	2298799	2746638	2860643	2271738	1742175	1449338	1069695
E	765735	938636	1563644	2313411	1908292	1391196	987638	721652	545313

He cambiado para letras las categorias y abajo tenemos el significado de cada categoria.

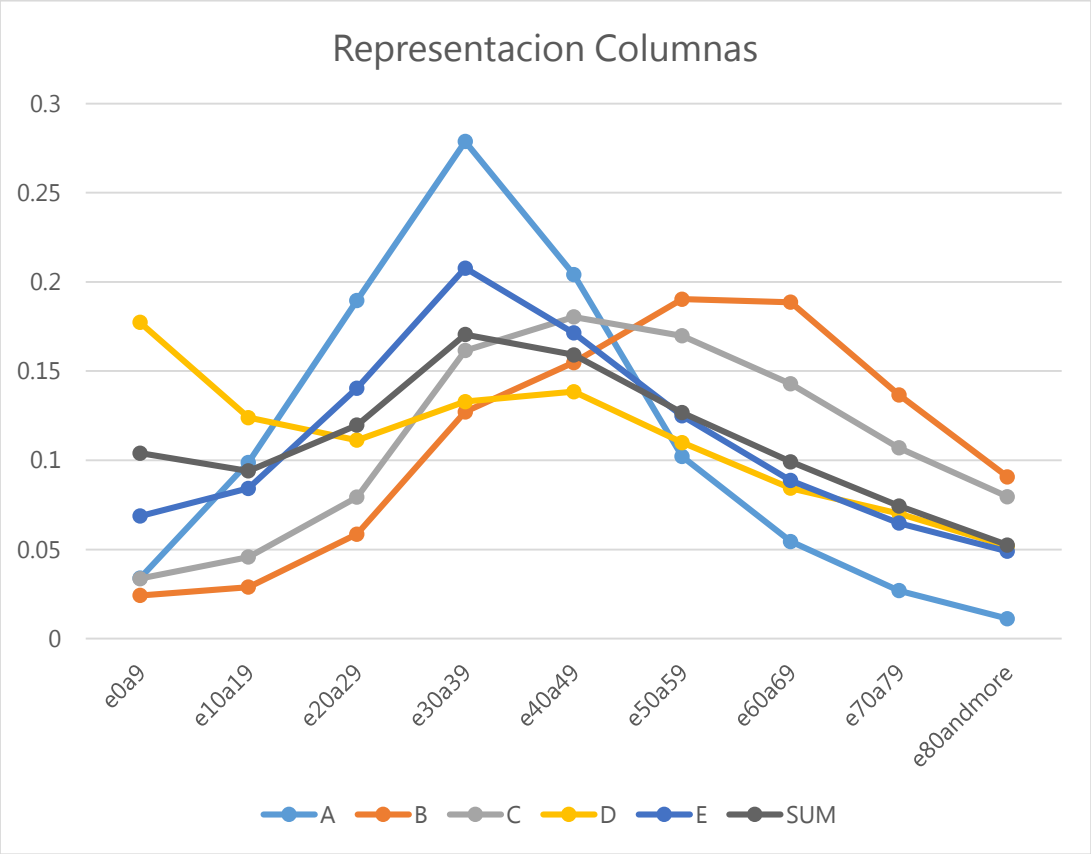
A = Born abroad
B = In a different Autonomous Community
C = Same Autonomous Community. Different province
D = Same Autonomous Community. Same province. Same municipality
E = Misma Comunidad Autónoma. Same province. Different municipality

Ejemplo: 214093 espanoles que en el censo de 2011 tenian entre 0 a 9 anos y han nacido fuera de Espana.

2. Representación de los perfiles fila y columna y evaluación grafica de la hipótesis de independencia, así como de las posibles relaciones entre categorías (hipótesis a confirmar posteriormente con las técnicas correspondientes).



Categoria	A	B	C	D	E	SUM
e0a9	0.04396	0.03507	0.01147	0.75226	0.15723	**1**
e10a19	0.14156	0.04622	0.01727	0.58174	0.21321	**1**
e20a29	0.21342	0.07374	0.02350	0.41028	0.27907	**1**
e30a39	0.22026	0.11228	0.03359	0.34407	0.28980	**1**
e40a49	0.17293	0.14651	0.04020	0.38412	0.25624	**1**
e50a59	0.10858	0.22627	0.04752	0.38305	0.23458	**1**
e60a69	0.07407	0.28658	0.05111	0.37542	0.21283	**1**
e70a79	0.04880	0.27675	0.05101	0.41621	0.20724	**1**
e80andmore	0.02868	0.26018	0.05380	0.43538	0.22195	**1**
SUM	**0.13473**	**0.15058**	**0.03546**	**0.44138**	**0.23786**	**1**



Representacion Columnas

Categoria	A	B	C	D	E	SUM
e0a9	0.033943808	0.024224766	0.033665843	0.177295256	0.068765105	**0.104025304**
e10a19	0.098803969	0.028866276	0.045797282	0.123937696	0.08429209	**0.094034941**
e20a29	0.189589422	0.058608341	0.079311911	0.111248847	0.140419524	**0.119682671**
e30a39	0.278776562	0.12714697	0.16153303	0.132921717	0.207750659	**0.170515511**
e40a49	0.204179078	0.154767651	0.180381131	0.138438913	0.171369861	**0.159074552**
e50a59	0.102096832	0.190355259	0.169795566	0.109939248	0.12493322	**0.12668059**
e60a69	0.054495308	0.188647939	0.142876507	0.084311399	0.088692604	**0.099124236**
e70a79	0.026941262	0.13670504	0.107008125	0.070139747	0.064806331	**0.074381989**
e80andmore	0.01117376	0.090677758	0.079630605	0.051767177	0.048970605	**0.052480207**
SUM	**1**	**1**	**1**	**1**	**1**	**1**

Como podemos observar en los gráficos podríamos pronosticar que las variables son dependientes.

3. Realización del contraste para la independencia de las variables (incluyendo un análisis básico de la tabla de aportaciones a la X^2).

Dx^2

Categoria	A	B	C	D	E	SUM
e0a9	297789.6767	431558.7174	78992.97376	1066395.499	133088.6136	**2007825.481**
e10a19	1525.500416	318387.2283	41073.87148	196489.173	11240.69781	**568716.471**
e20a29	257542.4742	219711.8068	22603.98576	12280.70554	40009.6558	**552148.6281**
e30a39	433533.3352	77759.41428	785.4365839	171267.0837	90542.54115	**773887.8109**
e40a49	80664.3141	822.0460244	4737.063513	55314.70457	10582.508	**152120.6362**
e50a59	30090.42915	225626.7346	24357.27315	45716.8155	268.3922558	**326059.6446**
e60a69	126734.5329	569985.0593	32055.59201	45740.62696	12224.61066	**786740.4218**
e70a79	190843.2433	368126.3278	23754.50309	4999.521062	13727.13211	**601450.7274**
e80andmore	205060.5133	195993.9865	23315.209	200.1819602	2613.548756	**427183.4395**
SUM	**1623784.019**	**2407971.321**	**251675.9083**	**1598404.311**	**314297.7002**	**6196133.26**

Con el codigo SAS:

```
proc corresp data=nacidos all chi2p print=both;
var E0A9 E10A19 E20A29 E30A39 E40A49 E50A59 E60A69 E70A79 E80ANDMO;
id CATEGORI;
run;
```

Tenemos:

Contributions to the Total Chi-Square Statistic										
Percents	**E0A9**	**E10A19**	**E20A29**	**E30A39**	**E40A49**	**E50A59**	**E60A69**	**E70A79**	**E80ANDMORE**	**Sum**
A	4.806	0.025	4.157	6.997	1.302	0.486	2.045	3.080	3.309	26.206
B	6.965	5.138	3.546	1.255	0.013	3.641	9.199	5.941	3.163	38.862
C	1.275	0.663	0.365	0.013	0.076	0.393	0.517	0.383	0.376	4.062
D	17.211	3.171	0.198	2.764	0.893	0.738	0.738	0.081	0.003	25.797
E	2.148	0.181	0.646	1.461	0.171	0.004	0.197	0.222	0.042	5.072
Sum	32.404	9.179	8.911	12.490	2.455	5.262	12.697	9.707	6.894	100.000

4. Selección del número de componentes a retener (calculadora de la distribución X^2 aquí). Comprobación de que el número de componentes retenido es suficiente a partir de las calidades.

Inertia and Chi-Square Decomposition

Singular Value	Principal Inertia	Chi-Square	Percent	Cumulative Percent	11 22 33 44 55 ----+----+----+----+----+---
0.27235	0.07417	3472467	56.04	56.04	**********
0.24027	0.05773	2702727	43.62	99.66	**********
0.01793	0.00032	15057	0.24	99.91	
0.01121	0.00013	5883	0.09	100.00	
Total	0.13235	6196133	100.00		

Degrees of Freedom = 32

Pr > ChiSq < .0001

Hipotesis de independencia

El p-valor es 0.0001 de esta forma se rechaza la hipotesis nula o sea la hipotesis de independencia y se puede decir que las variables son dependentes, hay dependencias entre las variables.

En el material se encuentra el texto *"la tabla nos muestra el p-valor asociado al test de Pearson. Intuitivamente, el p-valor es la probabilidad de "equivocarse" al rechazar la hipotesis de independencia dados los datos. En este caso es menor que 0,0001 por lo que podemos rechazar la hipotesis de independencia."*

Ahora que hemos verificado la dependencia de las variables estudiadas se procederá a determinar el número de factores a retener, teniendo en cuenta de minimizar la perdida de información recogida en las variables originales, de tal modo que nuestro modelo sea fiable.

Entonces realizaremos una evaluación con distintas reglas.

1. **Criterio de practicidad:**

   Quedo con dos dimensiones que explican 99.66% pero tambien por el consejo visto en el material del curso:

   > *"A nivel práctico el mejor criterio es tomar solo los dos o tres primeros ejes significativos siempre que estos expliquen una variabilidad aceptable (por encima de 60%) ya que esta técnica es primordialmente gráfica."*

   Sumando 56.04 + 43.63 = 99.66
   Entonces con dos dimensiones se explica 99.66%, entonces eso es suficiente para decirmos quedar con 2.

2. **Criterio de inercias principales:** Dimensiones con inercias mayores que **I/min(r-1,c-1)**

   Filas = Rows = 5
   Columnas = Columns = 9

   Min { r-1, c-1} = Min(5-1), (9-1) = Min(4, 8) = 4 El minimo entre 4 y 8 es 4.

   Total Principal Inertia dividido por Min { r-1, c-1}
   0.13235 / 4 = 0.033, por lo que deberiamos retener solo aquellos factores cuya inercia sea mayor que esa cantidad, es decir, solo el primero y el segundo.

   En la dimension 1 tenemos el Principal Inertia como 0.07 y este valor es maior que 0.033 que es el valor calculado en la formula acima. Lo mismo calculo se hace con las otras dimensiones.
   0.07 es maior que 0.033

0.05 es maior que 0.033
0.00032 es menor que 0.033

Entonces se quedo con solo los dos dimensiones.

3. **Principio de Carlier**: Este principio propone retener el menor valor m de factores que verifique:
$n(\lambda_{m+1} + \cdots + \lambda_r) < (r-1)(c-1)$

Teste Carlier
Grado de Libertad = 32

m = 1 = 2702727 + 15057 > 32
m = 2 = 15057 < 32

4. **Contraste de suficiencia:**

propone considerar **m** factores en los cuales:
   - $P[n(\lambda_1 + \cdots + \lambda_m) > \chi^2_{(r+c-3)+(r+c-5)+\cdots+(r+c-2m-1)}] < \alpha$, debe ser significativo
   - $P[n(\lambda_{m+1} + \cdots + \lambda_r) > \chi^2_{(r+c-2m-3)+\cdots+(|r-c|+1)}] > \alpha$, Debe ser no significativo

Conclusión: De estos 4 criterios que se puede utilizar he probado 3, los dos primeros nos dicen que debemos quedarnos con 2 dimensiones.

Los otros dos se sabe que dejan de ser efectivos en muestras grandes.
De esta forma me quedo con 2 dimensiones pues explican en conjunto el 99.66% de la inercia total.

Codigo SAS utilizado:

```
proc corresp data=nacidos all chi2p print=both dim=2;
var E0A9 E10A19 E20A29 E30A39 E40A49 E50A59 E60A69 E70A79 E80ANDMO;
id CATEGORI;
run;
```

5. Interpretación de las componentes/ejes a partir de las contribuciones parciales, las coordenadas y los cosenos al cuadrado.

Row Coordinates		
	Dim1	**Dim2**
A	-0.1743	0.4758
B	0.5793	-0.0753
C	0.3827	0.0392
D	-0.1801	-0.2119
E	0.0092	0.1654

Summary Statistics for the Row Points			
	Quality	**Mass**	**Inertia**
A	0.9975	0.1347	0.2621
B	0.9992	0.1506	0.3886
C	0.9760	0.0355	0.0406
D	0.9998	0.4414	0.2580
E	0.9726	0.2379	0.0507

Como se puede ver las variables A, B, C, D y E tienen muy buena calidad entre 0.97 a 0.99.

Partial Contributions to Inertia for the Row Points		
	Dim1	**Dim2**
A	0.0552	0.5284
B	0.6814	0.0148
C	0.0700	0.0009
D	0.1931	0.3432

Partial Contributions to Inertia for the Row Points		
	Dim1	Dim2
E	0.0003	0.1127

Indices of the Coordinates That Contribute Most to Inertia for the Row Points			
	Dim1	Dim2	Best
A	0	2	2
B	1	0	1
C	0	0	1
D	2	2	2
E	0	0	2

En esta tabla se puede ver que las variables B y D quedan en la Dim1. Las variables A y E quedan con la Dim2.

Squared Cosines for the Row Points		
	Dim1	Dim2
A	0.1180	0.8794
B	0.9826	0.0166
C	0.9658	0.0101
D	0.4195	0.5803
E	0.0030	0.9696

La dimensión 1 explica los grupos B, C y D y la dimensión 2 explica el grupo A, D y E.

Column Coordinates		
	Dim1	Dim2
E0A9	-0.4297	-0.4771
E10A19	-0.3456	-0.0975
E20A29	-0.2087	0.2338
E30A39	-0.0727	0.3027
E40A49	0.0120	0.1408
E50A59	0.2332	-0.0243

Column Coordinates		
	Dim1	**Dim2**
E60A69	0.3929	-0.1192
E70A79	0.3609	-0.2060
E80ANDMORE	0.3302	-0.2470

Summary Statistics for the Column Points			
	Quality	**Mass**	**Inertia**
E0A9	1.0000	0.1040	0.3240
E10A19	0.9981	0.0940	0.0918
E20A29	0.9964	0.1197	0.0891
E30A39	0.9994	0.1705	0.1249
E40A49	0.9783	0.1591	0.0246
E50A59	0.9997	0.1267	0.0526
E60A69	0.9944	0.0991	0.1270
E70A79	0.9997	0.0744	0.0971
E80ANDMORE	0.9781	0.0525	0.0689

Hay muy buena calidad también en las variables de tramo de edad.

Partial Contributions to Inertia for the Column Points		
	Dim1	**Dim2**
E0A9	0.2589	0.4102
E10A19	0.1514	0.0155
E20A29	0.0703	0.1133
E30A39	0.0122	0.2705
E40A49	0.0003	0.0547
E50A59	0.0929	0.0013
E60A69	0.2063	0.0244
E70A79	0.1306	0.0547
E80ANDMORE	0.0772	0.0555

Indices of the Coordinates That Contribute Most to Inertia for the Column Points			
	Dim1	**Dim2**	**Best**
E0A9	2	2	2
E10A19	1	0	1
E20A29	0	2	2

Indices of the Coordinates That Contribute Most to Inertia for the Column Points			
	Dim1	Dim2	Best
E30A39	0	2	2
E40A49	0	0	2
E50A59	1	0	1
E60A69	1	0	1
E70A79	1	0	1
E80ANDMORE	0	1	1

Se puede ver algunas variables en la dim1 y otras en la dim2, también algunas que están en las dos.

Squared Cosines for the Column Points		
	Dim1	Dim2
E0A9	0.4478	0.5521
E10A19	0.9245	0.0736
E20A29	0.4419	0.5545
E30A39	0.0546	0.9448
E40A49	0.0071	0.9712
E50A59	0.9889	0.0108
E60A69	0.9105	0.0839
E70A79	0.7540	0.2458
E80ANDMORE	0.6272	0.3509

En el caso de los tramos de edad vemos algo un poco distinto de lo que ha pasado con los grupos.

Se puede ver por ejemplo que la dimensión 1 he explicada muy bien los tramos de edad de 10 a 19, 50 a 59, 60 a 69 y 70 a 79.

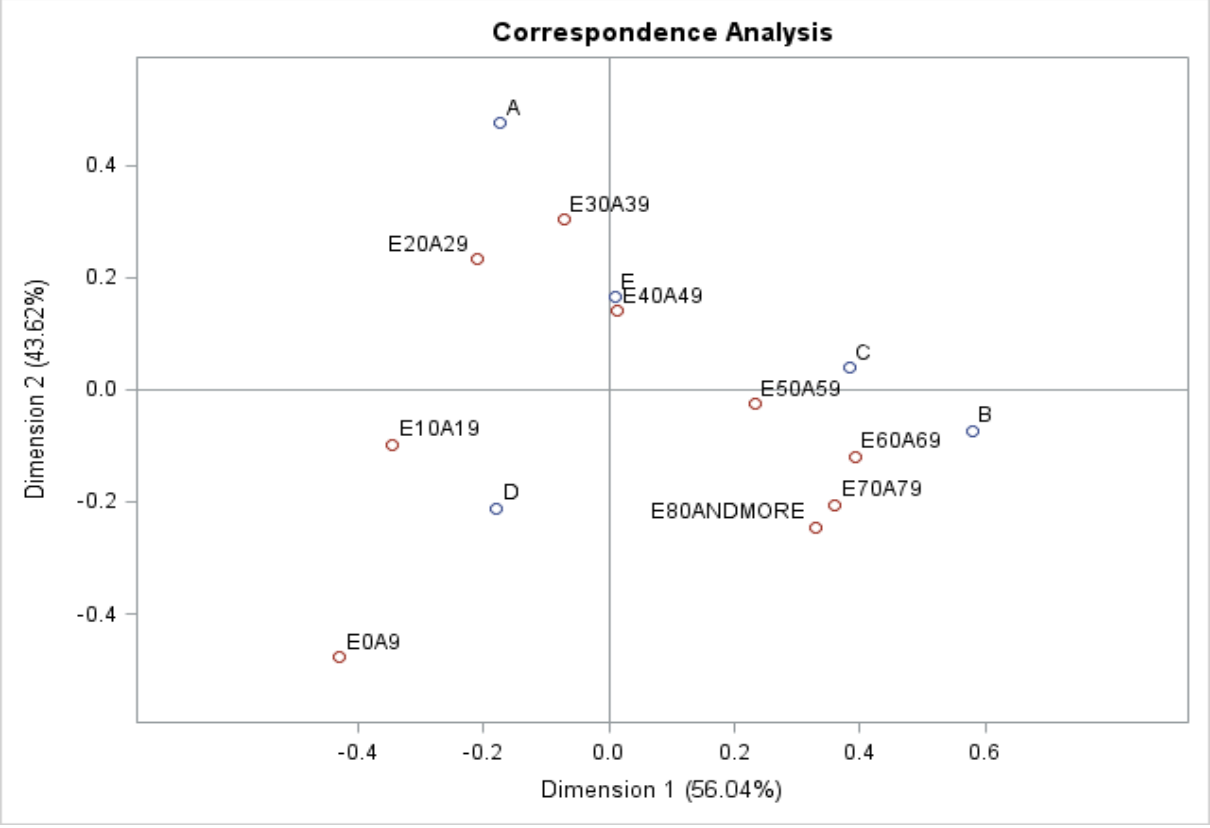
La dimensión 2 he explicada por el opuesto y hay casos como de 0 a 9, 20 a 29 donde se queda mitad para cada dimensión.

Los con edad de 80 o más, está casi en el medio de las dos dimensiones, siendo pendiente de la dimensión 1 pero también se mueve un poco para la dimensión 2.

Se pude notar una polarización entre personas jóvenes y personas mayores.

6. Análisis del gráfico obteniendo, incluyendo las relaciones entre categorías (tanto de una misma variable, como entre las dos).

```
proc corresp data=nacidos all chi2p print=both dim=2;
var E0A9 E10A19 E20A29 E30A39 E40A49 E50A59 E60A69 E70A79 E80ANDMO;
id CATEGORI;
run;
```



En 2011 durante el censo se puede ver que las personas que tenían de 0 a 9 años de edad son parte del grupo D, o sea "Same autonomous Community. Same province. Same municipality".

Eso significa que hasta los 9 años mucha gente no se cambia de la ciudad que ha nacido.

Las personas de 60 a 69, 70 a 79 y también los de 80 años o son parte del grupo B, que son los que viven en una comunidad autónoma distinta de la que ha nacido (In a different Autonomous Community).

Los del grupo A son los de 20 a 29 años y 30 a 39 años, el grupo A son los Born abroad.

Los del grupo E son los de 40 a 49 años, el grupo E son Misma Comunidad Autonoma. Same province. Diffferent municipality.

Los del grupo C son los de 50 a 59 años, el grupo C son los "Same Autonomous Community. Different province".

Código SAS Total utilizado en el trabajo:

```
PROC IMPORT OUT= UCM.Nacidos
            DATAFILE= "C:\Users\win\Desktop\trabajo\DatosCaio-sumados.xls"
            DBMS=EXCEL5 REPLACE;
     GETNAMES=YES;
RUN;

libname ucm 'C:\Users\win\Desktop\trabajo\';

data nacidos;
set ucm.nacidos;
run;

proc print data=ucm.nacidos (obs=100);


proc contents data=ucm.nacidos out=sa;
data;set sa;if _n_=1 then put 'LISTA DE VARIABLES CONTINUAS';if type=1 then
put name @@;run;
data;set sa;if _n_=1 then put 'LISTA DE VARIABLES CATEGÓRICAS';if type=2 then
put name @@;run;


proc corresp data=nacidos all chi2p print=both;
var E0A9 E10A19 E20A29 E30A39 E40A49 E50A59 E60A69 E70A79 E80ANDMO;
id CATEGORI;
run;

proc corresp data=nacidos all chi2p print=both dim=2;
```

```
var E0A9 E10A19 E20A29 E30A39 E40A49 E50A59 E60A69 E70A79 E80ANDMO;
id CATEGORI;
run;


proc corresp data=nacidos all chi2p print=both dim=2;
var E0A9 E10A19 E20A29 E30A39 E40A49 E50A59 E60A69 E70A79 E80ANDMO;
id CATEGORI;
supplementary E0A9;
run;



proc template;
      define statgraph Stat.Corresp.Graphics.Configuration;
         dynamic xVar yVar head legend;
         begingraph;
            entrytitle HEAD;
            layout overlayequated / equatetype=fit xaxisopts=(offsetmin=0.1
               offsetmax=0.1) yaxisopts=(offsetmin=0.1 offsetmax=0.1);

               referenceline x=0;
               referenceline y=0;

               scatterplot y=YVAR x=XVAR / group=GROUP index=INDEX
                  datalabel=LABEL datalabelattrs=GRAPHVALUETEXT
                  name="Type" tip=(y x datalabel group)
                  tiplabel=(group="Point");
               if (LEGEND)
                  discretelegend "Type";
               endif;
            endlayout;
         endgraph;
      end;
   run;
```