24-12-2024

TAREA III: ANÁLISIS DE CORRESPONDENCIAS

TÉCNICAS ESTADÍSTICAS EN CIENCIA DE DATOS

GRUPO A

XINYUAN ZHENG
xinyuanz@ucm.es

ÍNDICE

ANÁLISIS DE CORRESPONDENCIAS SIMPLES

Para realizar este trabajo, se ha utilizado un dataset proveniente de INE, denominado “Porcentaje de viviendas, por tamaño de la vivienda (nº de personas que la habitan) y sistema de calefacción disponible”, proveniente de un proyecto de “Encuesta de Hogares y Medio Ambiente 2008”.

```
/*Mostrar dataset original*/
proc print data= series.calefaccion;
run;
```

Obs	Viviendas	Electrica	Gas	Gasoleo	Otros
1	Vivienda con 1 persona	17.9	29.9	8.9	10.5
2	Vivienda con 2 personas	19.2	32.3	11.7	12.5
3	Vivienda con 3 personas	18.9	32.9	11.6	13.0
4	Vivienda con 4 o más personas	18.4	33.4	13.5	14.4

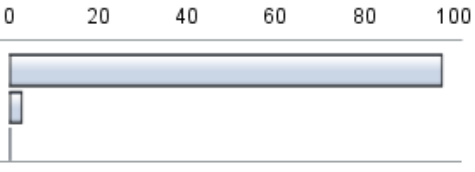
Esta dataset filtrado, está compuesto por 2 variables y 4 categorías para cada una; en la cual las categorías de la variable columna están relacionadas con el tipo de calefacción usado en cada vivienda(eléctrica, gas, gasóleo y otros), mientras que las categorías de la variable fila están relacionadas con el tamaño de la vivienda (1persona, 2 persona, 3 personas y 4 o más personas). A partir de esta dataset, se realizará esta tarea de análisis de correspondencia simple.

Antes de comenzar con el trabajo, se verifica previamente que el test de Chi-Cuadrado y el p-value nos devuelva una relación de dependencias entre las 2 variables.

```
/* Comprobar si el dataset se puede aplicar el analisis de correspondencias*/
proc corresp data=series.calefaccion outc=Resultados_prueba chi2p all;
        var Electrica Gasoleo Gas Otros;
   id Viviendas;
   ods output RowProfiles=PerfilFila;
   ods output ColProfiles=PerfilColumna;
   ods output CellChiSq=Aportaciones;
run;
```

Inertia and Chi-Square Decomposition

Singular Value	Principal Inertia	Chi-Square	Percent	Cumulative Percent
0.04832	0.00233	0.6980	97.34	97.34
0.00795	0.00006	0.0189	2.64	99.97
0.00082	0.00000	0.0002	0.03	100.00
	0.00240	0.7171	100.00	



Pr > Chisq = .9999 DF = 9

En este caso, se observa que el p-value es de 0,99 siendo este superior al 0,05; por lo tanto, no se puede rechazar la Ho de la independencia entre las variables, ya que no hay evidencia significativa en la cual se pueda afirmar que existe dependencia entre las variables.

Este problema de un valor de p-value extremadamente alto, se puede deber a que en un análisis de correspondencias, es mejor trabajar con los datos en frecuencias absolutas para el cálculo de Chi-Cuadrado y sus contribuciones. Sin embargo, los datos recogidos en el dataset se refieren a la proporción de las viviendas que utilizan según el tipo de calefacción en sus hogares, por lo tanto están en porcentajes, generando así unos resultados poco relevantes para el análisis de correspondencias.

Para solucionar este problema, se ha accedido a la documentación de la metodología usado en la "Encuesta de Hogares y Medio Ambiente 2008" para conocer el número total de la muestra usada. Dentro de ella, se ha encontrado el número de la muestra efectiva final de 26689 viviendas.[1]

```
/*Transformar los valores de proporcion a frecuencias absolutas de viviendas */
data series.calefaccion;
   set series.calefaccion;
   Electrica = round(Electrica * 26689 / 100);
   Gasoleo = round(Gasoleo * 26689 / 100);
   Gas = round(Gas * 26689 / 100);
   Otros = round(Otros * 26689 / 100);
run;
proc print data= series.calefaccion;
run;
```

Obs	Viviendas	Electrica	Gas	Gasoleo	Otros
1	Vivienda con 1 persona	4777	7980	2375	2802
2	Vivienda con 2 personas	5124	8621	3123	3336
3	Vivienda con 3 personas	5044	8781	3096	3470
4	Vivienda con 4 o más personas	4911	8914	3603	3843

[1] Muestra efectiva total de la "Encuesta de Hogares y Medio Ambiente 2008" en el anexo.

Por lo tanto, a partir de este número de muestra total de 26689 viviendas, se procede a transformar estos datos de porcentaje de viviendas, en valores de frecuencias absolutas sin decimales.

```
proc corresp data=series.calefaccion outc=Resultados_prueba chi2p all;
        var Electrica Gasoleo Gas Otros;
   id Viviendas;
   ods output RowProfiles=PerfilFila;
   ods output ColProfiles=PerfilColumna;
   ods output CellChiSq=Aportaciones;
run;
```

Inertia and Chi-Square Decomposition

Singular Value	Principal Inertia	Chi-Square	Percent	Cumulative Percent
0.04832	0.00233	186.316	97.33	97.33
0.00796	0.00006	5.056	2.64	99.97
0.00083	0.00000	0.055	0.03	100.00
	0.00240	191.427	100.00	

Pr > Chisq < .0001 DF = 9

Con este cambio, el resultado de p-value ya es inferior a 0,05 (0,0001) por lo que se puede rechazar la Ho de independencia , siendo ambos variables dependientes. Por lo que ya está listo para realizar el análisis de correspondencia.

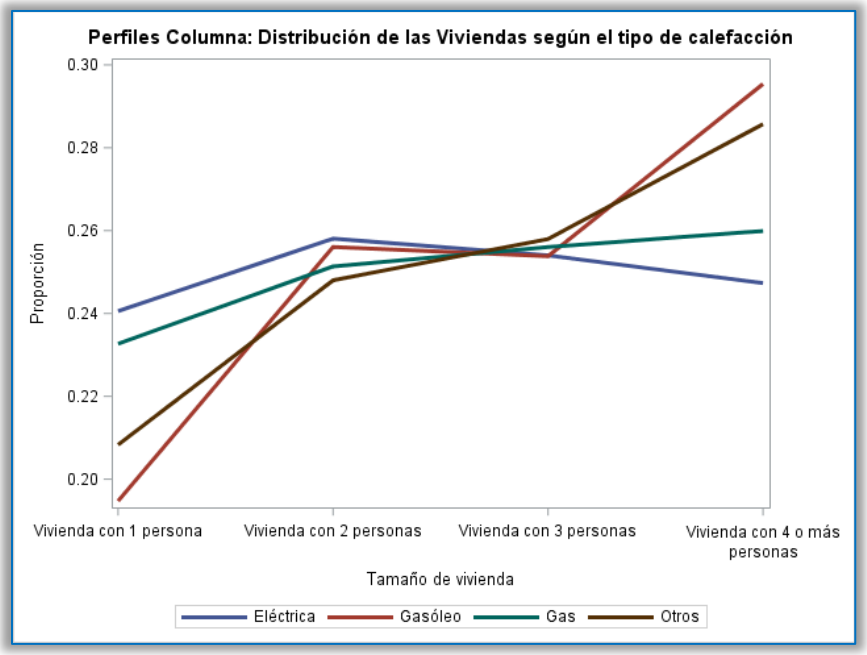
1. **Realizar un PROC CORRESP. Guardar los perfiles fila, perfiles columna y la tabla de las contribuciones al estadístico Chicuadrado en ficheros con la opción ODS para construir los gráficos de líneas y un mapa de calor. (0.5)**

```
proc corresp data=series.calefaccion outc=ResultadosACS chi2p all;
        var Electrica Gasoleo Gas Otros;
   id Viviendas;
   ods output RowProfiles=PerfilFila;
   ods output ColProfiles=PerfilColumna;
   ods output CellChiSq=Aportaciones;
run;
```

2. Representar los gráficos de líneas de los perfiles columna y perfiles fila acompañados de las tablas de dichos perfiles que aparecen en la salida del proc corresp. Comentar lo más destacado de ambos gráficos (2)

```
/* Perfil columna */
proc sgplot data=PerfilColumna;
  series x=Label y=Electrica / lineattrs=(thickness=3) legendlabel="Eléctrica";
  series x=Label y=Gasoleo / lineattrs=(thickness=3) legendlabel="Gasóleo";
  series x=Label y=Gas / lineattrs=(thickness=3) legendlabel="Gas";
  series x=Label y=Otros / lineattrs=(thickness=3) legendlabel="Otros";
  yaxis label='Proporción';
  xaxis label='Tamaño de vivienda';
  title "Perfiles Columna: Distribución de las Viviendas según el tipo de calefacción";
run;

proc print data=PerfilColumna;
run;
```



Obs	Label	Electrica	Gasoleo	Gas	Otros
1	Vivienda con 1 persona	0.240582	0.194720	0.232680	0.208312
2	Vivienda con 2 personas	0.258058	0.256047	0.251370	0.248011
3	Vivienda con 3 personas	0.254029	0.253833	0.256036	0.257973
4	Vivienda con 4 o más personas	0.247331	0.295401	0.259914	0.285704

En esta gráfica de “PerfilColumna” se puede observar como las viviendas unipersonales tienen una proporción más alta en el uso de calefacción eléctrica con un 24% aprox. mientras que la calefacción tipo gasóleo representa menos del 20%, aunque al final la diferencia entre el uso de uno u otro no muestra una diferencia muy grande. Para las viviendas de 2 o 3 personas, la elección de un tipo de calefacción u otro, casi no se percibe diferencia, todos los tipos de calefacción representan el 25% aproximada para cada uno, generando un punto de corte de los 4 tipos de calefacción. A partir de las viviendas con 4 o más personas, el uso del tipo de calefacción se reinvierte, ahora la calefacción por gasóleo toma preferencia con un 30% aprox. de las viviendas, mientras que la calefacción eléctrica se convierte en el menos usado con 24% aprox.

Por lo tanto, se puede resumir como que a medida que se incremente el número de personas de una vivienda, la elección de tecnología de calefacción alternativa al de eléctrica toma más peso, especialmente las calefacciones basado en gasóleo.

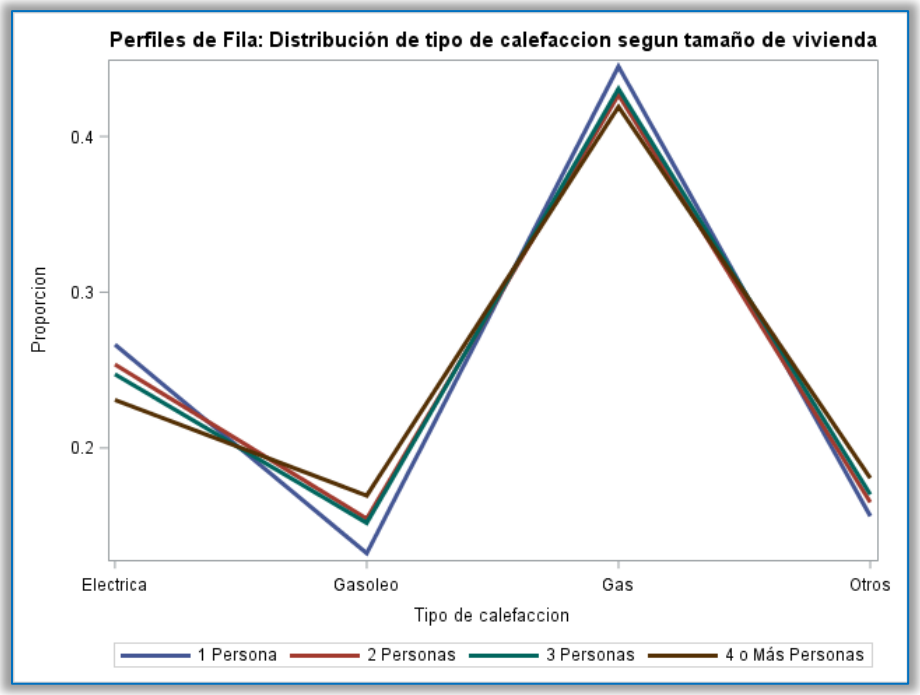
```
/* Perfil fila */

proc transpose data=PerfilFila out=PerfilFilaT;
id Label;
run;

proc contents data=PerfilFilaT;
run;

proc sgplot data=PerfilFilaT;
   series x=_NAME_ y=Vivienda_con_1_persona / lineattrs=(thickness=3) legendlabel="1
Persona";
   series x=_NAME_ y=Vivienda_con_2_personas / lineattrs=(thickness=3) legendlabel="2
Personas";
   series x=_NAME_ y=Vivienda_con_3_personas / lineattrs=(thickness=3) legendlabel="3
Personas";
   series x=_NAME_ y=Vivienda_con_4_o_m_s_personas / lineattrs=(thickness=3)
legendlabel="4 o Más Personas";
   yaxis label='Proporcion';
   xaxis label='Tipo de calefaccion';
   title "Perfiles de Fila: Distribución de tipo de calefaccion segun tamaño de vivienda";
run;

proc print data=PerfilFila;
run;
```



Obs	Label	Electrica	Gasoleo	Gas	Otros
1	Vivienda con 1 persona	0.266366	0.132430	0.444965	0.156240
2	Vivienda con 2 personas	0.253613	0.154573	0.426698	0.165116
3	Vivienda con 3 personas	0.247364	0.151832	0.430631	0.170173
4	Vivienda con 4 o más personas	0.230878	0.169386	0.419068	0.180669

Revisando la gráfica de “PerfilFila” se puede observar que el pico más alto se encuentra en la calefacción mediante “gas” , siendo elegido por más del 40% de las viviendas de todos los tamaños. Sin embargo, la calefacción por “gasóleo” es el menos elegido por las viviendas, con aprox. el 15% de las viviendas para todos los tamaños.

Por lo tanto, la calefacción por “gas” se sitúa como el método de calefacción más común entre las viviendas encuestados, siendo usado casi por la mitad de las viviendas encuestados para todos los tamaños de hogares. Mientras que la calefacción por “gasóleo” es el menos usando entre las viviendas encuestados.

3. ¿Cómo se calculan los valores de la frecuencia esperada para el cálculo del estadístico Chicuadrado? Poner el ejemplo de uno. (0.5)

```
proc corresp data=series.calefaccion outc=ResultadosACS chi2p all;
        var Electrica Gasoleo Gas Otros;
   id Viviendas;
   ods output RowProfiles=PerfilFila;
   ods output ColProfiles=PerfilColumna;
   ods output CellChiSq=Aportaciones;
run;
```

Contingency Table					
	Calefacción eléctrica:Total	**Calefacción por gasóleo : Total**	**Calefacción de gas : Total**	**Otros sistemas de calefacción: Total**	**Sum**
Vivienda con 1 persona	4777	2375	7980	2802	17934
Vivienda con 2 personas	5124	3123	8621	3336	20204
Vivienda con 3 personas	5044	3096	8781	3470	20391
Vivienda con 4 o más personas	4911	3603	8914	3843	21271
Sum	19856	12197	34296	13451	79800

Estableciendo como ejemplo, calcular la frecuencia esperada de las viviendas con “1 persona” con la calefacción por “electricidad”, el método de cálculo manual será de la siguiente manera:

Frecuencia marginal de la fila (Vivienda con 1 persona):

$$F_{i.} = 17934$$

Frecuencia marginal de la columna (Calefacción eléctrica):

$$F_{.j} = 19856$$

Total general (N):

$$N = 79800$$

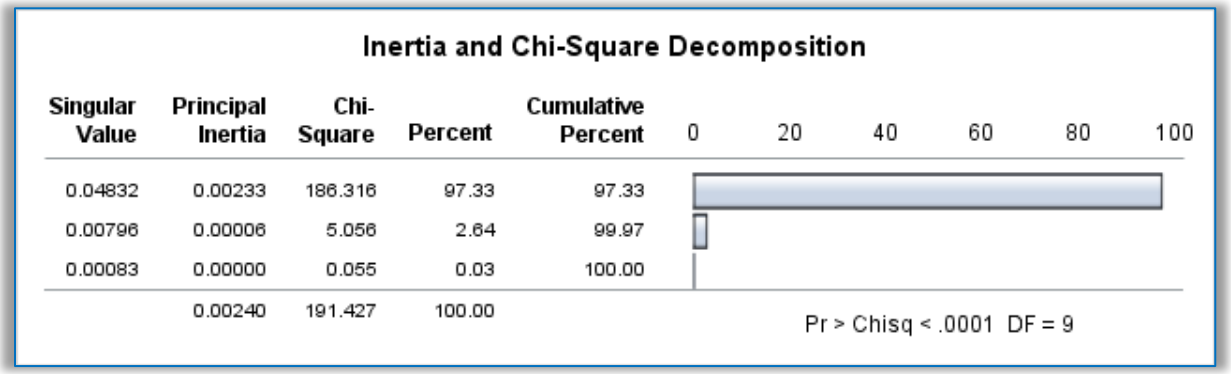
Frecuencia esperada:

$$E_{ij} = \frac{(17934 * 19856)}{79800} \approx 4462{,}37$$

La frecuencia esperada para la vivienda de “1 persona” y calefacción por “electricidad” es de 4462,37 aprox. coincidiendo con la frecuencia esperada calculado por SAS, mostrado a continuación:

Chi-Square Statistic Expected Values				
	Calefacción eléctrica:Total	**Calefacción por gasóleo : Total**	**Calefacción de gas : Total**	**Otros sistemas de calefacción: Total**
Vivienda con 1 persona	4462.37	2741.12	7707.57	3022.94
Vivienda con 2 personas	5027.20	3088.07	8683.16	3405.56
Vivienda con 3 personas	5073.73	3116.65	8763.53	3437.08
Vivienda con 4 o más personas	5292.69	3251.16	9141.73	3585.42

4. ¿Cuánto vale el estadístico Chicuadrado? ¿Qué nos dice este estadístico sobre la independencia entre las variables estudiadas? (0.5)



El valor de estadístico Chi-Cuadrado es de 191,427, con un grado de liberta de 9 que da lugar a un p-value de 0,0001.

Al tener un p-value inferior a 0,05, se rechaza con la Ho de la independencia entre las variables, siendo en este caso ambas variables dependientes.

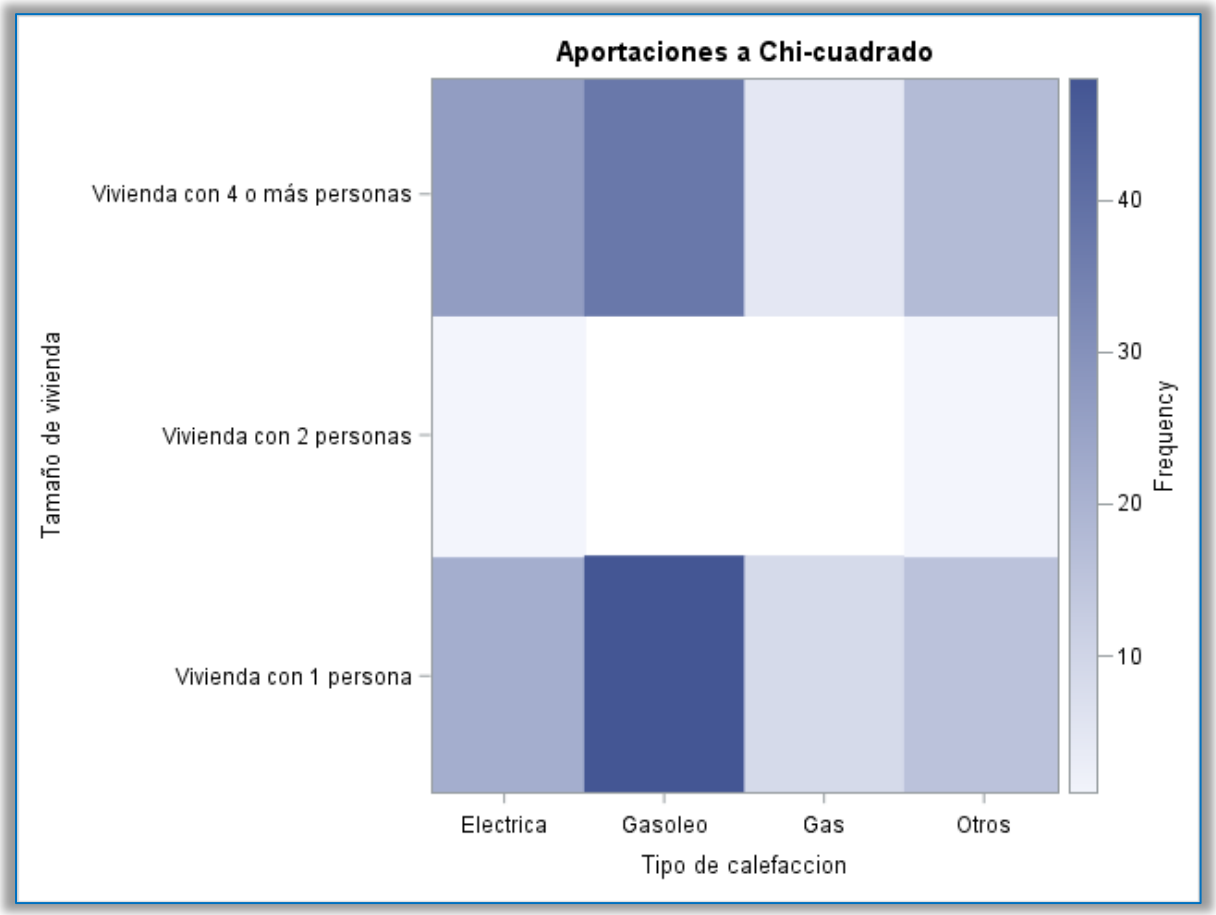
5. **Representar el mapa de calor de las contribuciones al estadístico Chicuadrado y la tabla correspondiente ¿Que combinaciones de categorías aportan más al estadístico Chicuadrado? (1)**

```
data Aportaciones2(drop=Sum);
   set Aportaciones;
   if Label = "Sum" then delete;
run;

data Aportaciones3(keep=filas col ff);
array vector{4} Electrica Gasoleo Gas Otros;
set Aportaciones2;
a=0;
do aux = '  Electrica', 'Gasoleo', 'Gas', 'Otros';
  a = a + 1;
  filas = aux;
  col = Label;
  ff = vector{a};
  output;
end;
run;

/* Mapa de calor */
proc sgplot data=Aportaciones3;
   heatmap x=filas y=col /freq=ff  colormodel=TwoColorRamp;
   title "Aportaciones a Chi-cuadrado";
   xaxis label="Tipo de calefaccion";
   yaxis label="Tamaño de vivienda";
run;
title '';

proc print data=Aportaciones;
run;
```



Obs	Label	Electrica	Gasoleo	Gas	Otros	Sum
1	Vivienda con 1 persona	22.183	48.900	9.629	16.147	96.859
2	Vivienda con 2 personas	1.864	0.395	0.445	1.421	4.125
3	Vivienda con 3 personas	0.174	0.137	0.035	0.315	0.661
4	Vivienda con 4 o más personas	27.527	38.077	5.673	18.505	89.782
5	Sum	51.748	87.508	15.782	36.389	191.427

La combinación de la categoría “Vivienda con 1 persona” y “Gasóleo” presenta la mayor contribución al estadístico Chi-Cuadrado con un valor de 48,900. Esto significa que estas 2 categorías están más asociadas entre ellos de lo que se esperaría bajo la hipótesis de independencia.

Debido a la poca contribución de la categoría “Vivienda con 3 personas” al Chi-Cuadrado, ésta no queda reflejada en el grafico de mapa de calor, siendo este el motivo de su ausencia en la gráfica generado en SAS.

6. **¿Qué porcentaje de la inercia queda explicado con los dos primeros autovalores? ¿Cuánto valen dichos autovalores? (0.5)**

```
proc corresp data=series.calefaccion outc=ResultadosACS chi2p all;
        var Electrica Gasoleo Gas Otros;
   id Viviendas;
   ods output RowProfiles=PerfilFila;
   ods output ColProfiles=PerfilColumna;
   ods output CellChiSq=Aportaciones;
run;
```

Inertia and Chi-Square Decomposition

Singular Value	Principal Inertia	Chi-Square	Percent	Cumulative Percent
0.04832	0.00233	186.316	97.33	97.33
0.00796	0.00006	5.056	2.64	99.97
0.00083	0.00000	0.055	0.03	100.00
	0.00240	191.427	100.00	

0 20 40 60 80 100

Pr > Chisq < .0001 DF = 9

Los dos primeros autovalores explican el 99,97% de la inercia total. Dentro de ella, la primera dimensión contribuye el 97,33%, lo que corresponde a un valor de 0,00233 de la inercia, mientras que la segunda dimensión explica el 2,64%, con un valor correspondiente de 0,00006 de la inercia.

7. **Para los perfiles fila contestar a las siguientes preguntas acompañadas de la tabla correspondiente. (1.5)**

```
proc corresp data=series.calefaccion outc=ResultadosACS dim=2 chi2p all;
        var Electrica Gasoleo Gas Otros;
   id Viviendas;
   ods output RowProfiles=PerfilFila;
   ods output ColProfiles=PerfilColumna;
   ods output CellChiSq=Aportaciones;
run;
```

Para continuar con el análisis, establecemos el numero de la dimensión con dim=2, mostrado en este código.

7.1. ¿Qué categorías explican la dim 1? ¿Cuáles la dim 2?

Squared Cosines for the Row Points		
	Dim1	**Dim2**
Vivienda con 1 persona	0.9956	0.0043
Vivienda con 2 personas	0.0908	0.9092
Vivienda con 3 personas	0.0372	0.9106
Vivienda con 4 o más personas	0.9967	0.0032

Para la dim1, las categorías “Vivienda con 1 persona” y “Vivienda con 4 o más personas” son las categorías que tienen la mayor proporción de su variabilidad explicada en la Dim1, con un valor de 99,56% (Vivienda de 1 personas) y 99,67%(Vivienda con 4 o más personas).

En cuanto a la dim2, las categorías “Vivienda con 2 persona” y “Vivienda con 3 personas” son las categorías que tienen la mayor proporción de su variabilidad explicada en la Dim2, con un valor de 90,92% (Vivienda de 2 personas) y 91,06%(Vivienda con 3 personas).

7.2. ¿Qué categoría es la que queda peor explicada en las dos dimensiones? ¿Qué porcentaje de Inercia explica dicha categoría? ¿es una categoría poco frecuente?

Summary Statistics for the Row Points			
	Quality	**Mass**	**Inertia**
Vivienda con 1 persona	0.9999	0.2247	0.5060
Vivienda con 2 personas	1.0000	0.2532	0.0215
Vivienda con 3 personas	0.9478	0.2555	0.0035
Vivienda con 4 o más personas	0.9999	0.2666	0.4690

La categoría peor explicada en las 2 dimensiones es “Vivienda con 3 personas”, donde un 94,78% de su variabilidad total está explicado en el espacio factorial definido por las 2 dimensiones, representando el 0,35% de la inercia total explicado.

Si se observa la columna de “mass” podemos ver que “Vivienda con 3 personas” representa el 25,55% de las observaciones totales, por lo tanto no es una categoría poco frecuente.

7.3. ¿Qué relación hay entre el índice Quality y los cosenos al cuadrado?

Ambos mantienen una relación directa. El valor calculado por SAS en su columna de “Quality” es el resultado de la suma de los cosenos al cuadrado de cada categoría en las 2 dimensiones.

8. Para los perfiles columna contestar a las siguientes preguntas acompañadas de la tabla correspondiente (1.5)

```
proc corresp data=series.calefaccion outc=ResultadosACS dim=2 chi2p all;
        var Electrica Gasoleo Gas Otros;
   id Viviendas;
   ods output RowProfiles=PerfilFila;
   ods output ColProfiles=PerfilColumna;
   ods output CellChiSq=Aportaciones;
run;
```

8.1. ¿Qué categoría explica más proporción de la inercia de la dimensión 1?

Partial Contributions to Inertia for the Column Points		
	Dim1	**Dim2**
Calefacción eléctrica:Total	0.2698	0.2894
Calefacción por gasóleo : Total	0.4617	0.2930
Calefacción de gas : Total	0.0796	0.1849
Otros sistemas de calefacción: Total	0.1889	0.2328

La categoría que más contribuye en la proporción de la inercia total explicada en la dimensión 1 es “Calefacción de gas” con un 46,17%.

Si queremos encontrar las categorías que más contribuyen hasta alcanzar el 80% de la inercia total explicada en la dimensión 1, tendríamos que añadir las categorías de “Calefacción eléctrica” con un 26,98% y “Otros sistemas de calefacción” con un 18,89%.

8.2. ¿Qué categorías explican la dim 1? ¿Cuáles la dim 2?

Squared Cosines for the Column Points		
	Dim1	**Dim2**
Calefacción eléctrica:Total	0.9715	0.0283
Calefacción por gasóleo : Total	0.9830	0.0169
Calefacción de gas : Total	0.9397	0.0592
Otros sistemas de calefacción: Total	0.9670	0.0323

En este caso, las 4 categorías tienen la mayor proporción de su variabilidad explicada en la Dimensión 1, en la cual la "Calefacción eléctrica" consigue una proporción de 97,15%, "Calefacción por gasóleo" un 98,30%, "Calefacción de gas" un 93,97% y "Otros sistemas de calefacción" un 96,70%.

Mientras tanto, la proporción de variabilidad explicada en la Dimensión 2 solo alcanza el más alto con "Calefacción de gas" con apenas 5,92% y para el resto de las categorías, una proporción inferior al 5%.

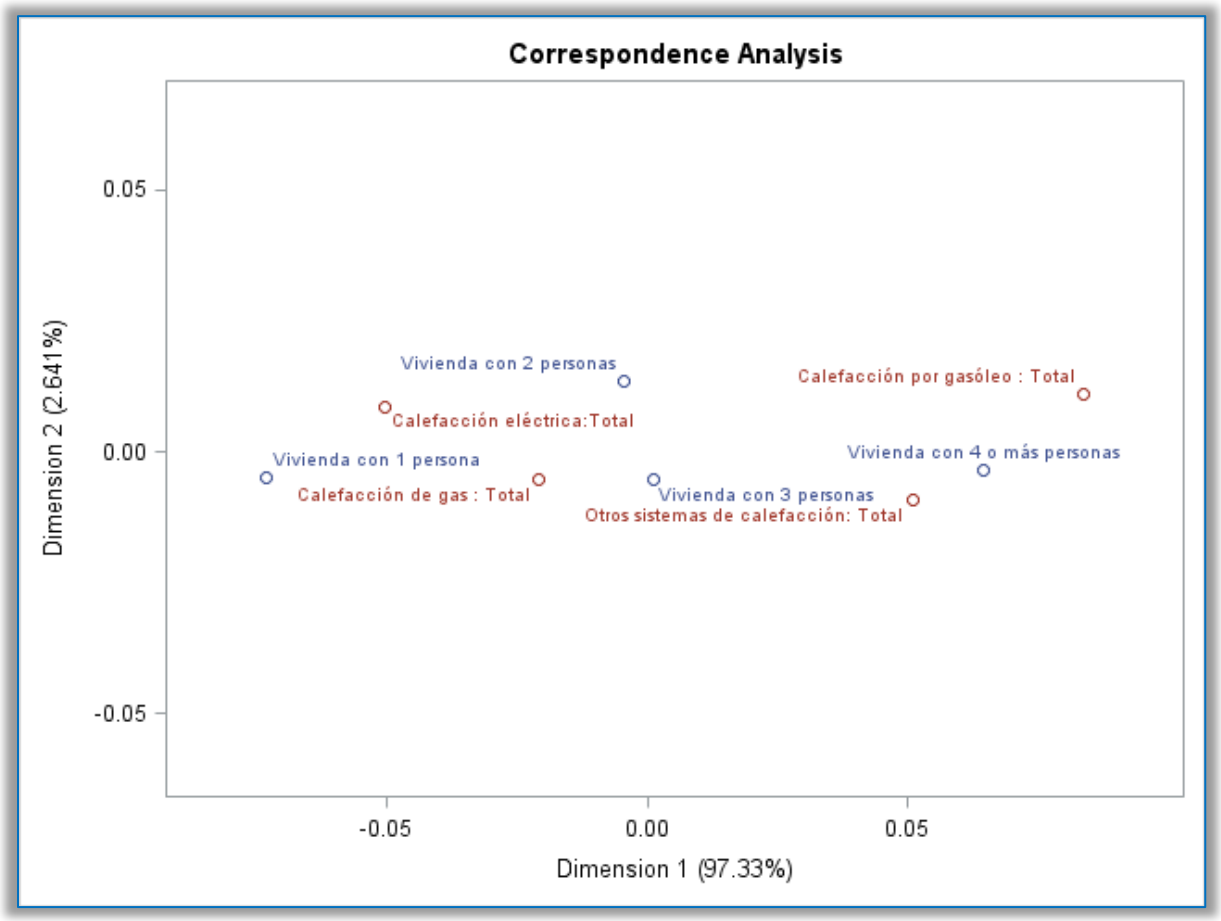
8.3. ¿Qué categoría es la que explica más Inercia?

Summary Statistics for the Column Points			
	Quality	**Mass**	**Inertia**
Calefacción eléctrica:Total	0.9998	0.2488	0.2703
Calefacción por gasóleo : Total	0.9999	0.1528	0.4571
Calefacción de gas : Total	0.9989	0.4298	0.0824
Otros sistemas de calefacción: Total	0.9994	0.1686	0.1901

La categoría que mayor inercia explica es "Calefacción por gasóleo" con un valor de 0,4571.

9. Comentar el gráfico conjunto que representa los perfiles fila y columna en el plano factorial. ¿Cómo se relacionan las categorías de las dos variables? (2)

```
proc corresp data=series.calefaccion outc=ResultadosACS dim=2 chi2p all;
        var Electrica Gasoleo Gas Otros;
   id Viviendas;
   ods output RowProfiles=PerfilFila;
   ods output ColProfiles=PerfilColumna;
   ods output CellChiSq=Aportaciones;
run;
```



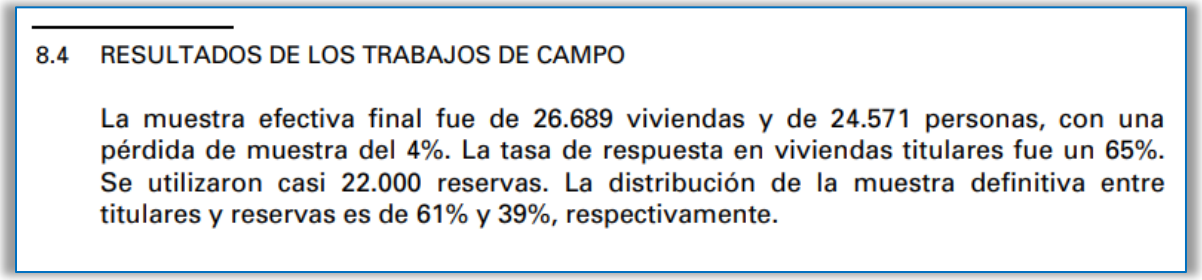
En este plano factorial de las dimensiones 1 y 2, se puede observar que la dimensión 1 obtiene un mayor peso al conseguir explicar el 97,33% de la inercia total, mientras que la dimensión 2 solo consigue explicar el 2,641% de la inercia total.

Si observamos las categorías, se puede percibir que no hay un agrupamiento entre las categorías de las 2 variables de forma muy clara. Se puede observar algunas relaciones algo más cercana, como por ejemplo que la “Vivienda con 4 o más personas” está más cercano a “Otros sistemas de calefacción” y la categoría “Vivienda con 3 personas” está más cercano con “calefacción con gas”.

ANEXO

1. Muestra efectiva total de la “Encuesta de Hogares y Medio Ambiente 2008”

Enlace: https://www.ine.es/metodologia/t25/t2530500.pdf



8.4 RESULTADOS DE LOS TRABAJOS DE CAMPO

La muestra efectiva final fue de 26.689 viviendas y de 24.571 personas, con una pérdida de muestra del 4%. La tasa de respuesta en viviendas titulares fue un 65%. Se utilizaron casi 22.000 reservas. La distribución de la muestra definitiva entre titulares y reservas es de 61% y 39%, respectivamente.