PASOS PARA SU APLICACIÓN

ANÁLISIS DE CORRESPONDENCIA SIMPLE

1. Verificar supuestos

Esta técnica ***NO*** requiere de verificar supuestos de adecuación de datos. El uso de información no-numérica por medio de una tabla de frecuencia representa en una buena manera las relaciones lineales y no-lineales de las variables analizadas.

2. Obtener perfiles y matriz de correspondencia

Se genera los *perfiles* de los datos representados en la tabla de contingencia.

Perfil marginal:

Describe la distribución marginal (frecuencia marginal) de las variables X y Y respectivamente:

	x_1	x_2 ……	x_r	**Total**
$X=x_r$	$\frac{n_{1\bullet}}{n}$	$\frac{n_{2\bullet}}{n}$ ……	$\frac{n_{r\bullet}}{n}$	**1**

	y_1	y_2 ……	y_c	**Total**
$Y=y_c$	$\frac{n_{\bullet 1}}{n}$	$\frac{n_{\bullet 2}}{n}$ ……	$\frac{n_{\bullet c}}{n}$	**1**

2. Obtener matriz de frecuencia relativa (O)

	y_1	…	y_c	***Perfil marginal fila***
x_1	o_{11}	…	o_{1c}	$o_{1\bullet}$
⋮	⋮		⋮	⋮
x_r	o_{r1}	…	O_{rc}	$o_{r\bullet}$
Perfil marginal columna	$o_{\bullet 1}$	…	$o_{\bullet c}$	1

Donde: $o_{rc} = \frac{n_{rc}}{n}$

Tabla en donde se expresan las frecuencias de la tabla de contingencia en términos de porcentaje o proporción.

Caso: La delincuencia

Siguiendo con el ejemplo anterior, se analiza el sexo (Male, Female), culpabilidad del delincuente (Si, No) y el tipo de delito:

Cinco categorías de cargos:

ID Impaired Driving (daños al manejar)
TU Theft Under $1000 (robo menor a $1000)
MI Mischief (cargos menores)
PN Possession of Narcotics (posesión de narcóticos)
OT Others (otros)

	ID	*TU*	*MI*	*PN*	*OT*	*TOTAL*
No-Male	8	11	5	7	12	*43*
No-Female	5	15	3	1	6	*30*
Si-Male	105	32	11	23	37	*208*
Si-Female	32	57	6	2	25	*122*
TOTAL	*150*	*115*	*25*	*33*	*80*	*403*

Caso: La delincuencia

A partir de las frecuencias contenidas en la tabla de contingencia se procede a calcular su correspondiente matriz de frecuencias relativas.

Matriz de frecuencia relativa:

	ID	TU	MI	PN	OT	Row masses
No-Male	0.020	0.027	0.012	0.017	0.030	*0.107*
No-Female	0.012	0.037	0.007	0.002	0.015	*0.074*
Si-Male	0.261	0.079	0.027	0.057	0.092	*0.516*
Si-Female	0.079	0.141	0.015	0.005	0.062	*0.303*
Col. masses	*0.372*	*0.285*	*0.062*	*0.082*	*0.199*	*1*

Perfil marginal de columnas

Perfil marginal de filas

Caso: La delincuencia

Perfiles fila (r): distribución condicionada de la variable Y para las categorías de la variable X.

$\mathbf{r}_2$ *(perfil del renglón 2)*

	y_1	y_2	……	y_c	***Total***
x_1	$n_{11}/n_{1\bullet}$	$n_{12}/n_{1\bullet}$	……	$n_{1c}/n_{1\bullet}$	1
x_2	$n_{21}/n_{2\bullet}$	$n_{22}/n_{2\bullet}$	……	$n_{2c}/n_{2\bullet}$	1
⋮	⋮	⋮		⋮	⋮
x_r	$n_{r1}/n_{r\bullet}$	$n_{r2}/n_{r\bullet}$	……	$n_{rc}/n_{r\bullet}$	1
Perfil marginal columna	$n_{\bullet 1}/n$	$n_{\bullet 2}/n$	……	$n_{\bullet c}/n$	1

Caso: La delincuencia

Perfiles columna (c): distribución condicionada de la variable X para las categorías de la variable *Y*.

$\mathbf{c}_2$ (*perfil de la columna 2*)

	y_1	y_2	……	y_c	***Perfil marginal fila***
x_1	$n_{11}/n_{\bullet 1}$	$n_{12}/n_{\bullet 2}$	……	$n_{1c}/n_{\bullet c}$	$n_{1\bullet}/n$
x_2	$n_{21}/n_{\bullet 1}$	$n_{22}/n_{\bullet 2}$	……	$n_{2c}/n_{\bullet c}$	$n_{2\bullet}/n$
⋮	⋮			⋮	⋮
x_r	$n_{r1}/n_{\bullet 1}$	$n_{r2}/n_{\bullet 2}$	……	$n_{rc}/n_{\bullet c}$	$n_{r\bullet}/n$
Total	*1*	*1*	……	*1*	*1*

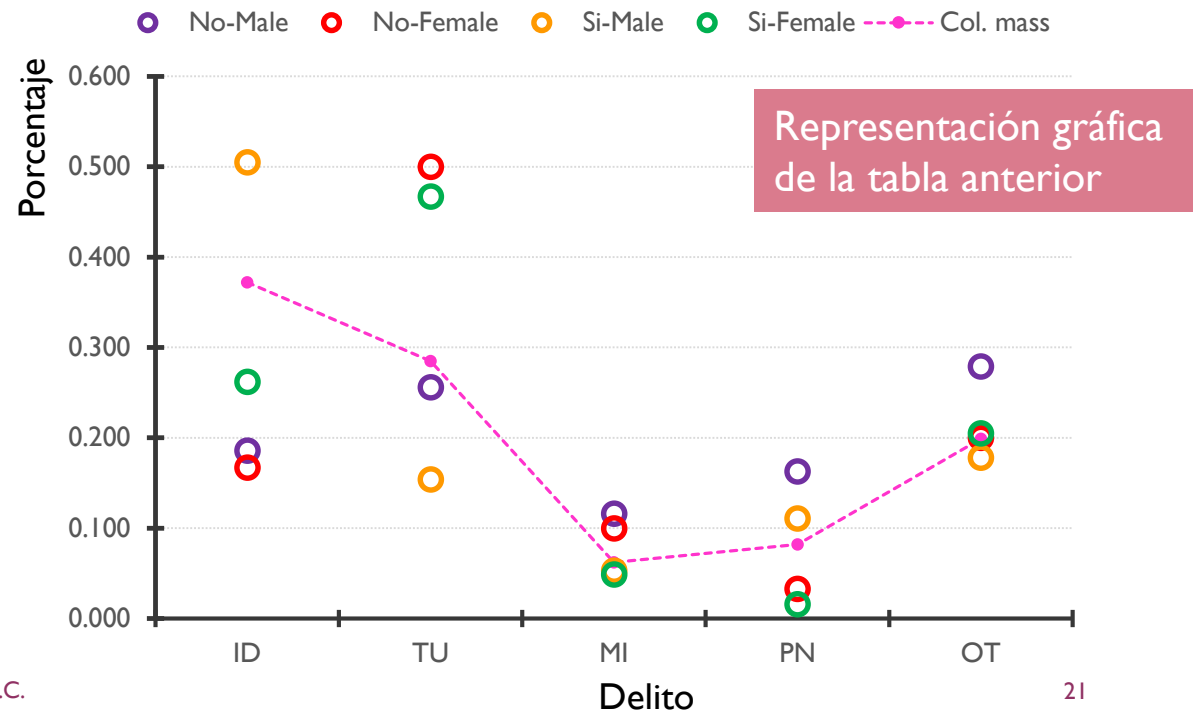
Caso: La delincuencia

$\mathbf{r}_3$ (*perfil del renglón 3*)

	ID	TU	MI	PN	OT	TOTAL
No-Male	0.186	0.256	0.116	0.163	0.279	1.000
No-Female	0.167	0.500	0.100	0.033	0.200	1.000
Si-Male	0.505	0.154	0.053	0.111	0.178	1.000
Si-Female	0.262	0.467	0.049	0.016	0.205	1.000
Col. mass	*0.372*	*0.285*	*0.062*	*0.082*	*0.199*	*1.000*

Interpretación:

De los hombres declarados culpables ($\mathbf{r}_3$): el 50.5% cometió daños al manejar, 15.4% fue por robo menor a $1000, 5.3% por cargos menores, 11.1% por narcóticos y 17.8% por otros delitos.



Representación gráfica de la tabla anterior

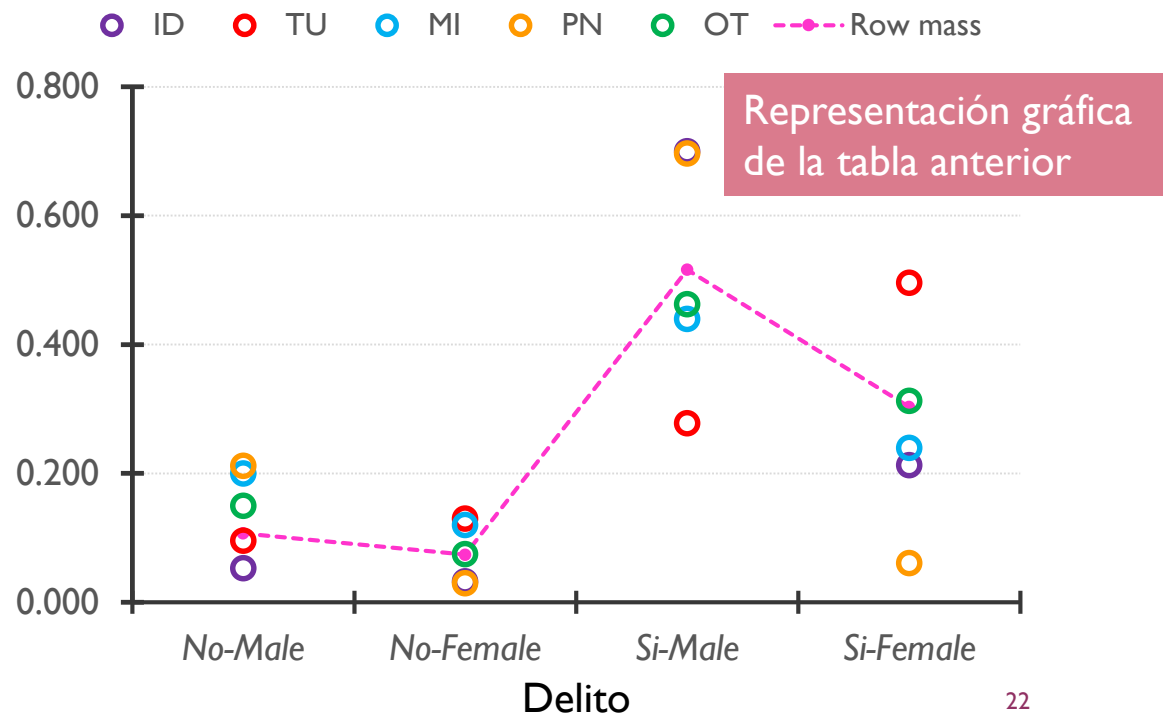
Caso: La delincuencia

$\mathbf{c}_2$ (*perfil de la columna 2*)

	ID	TU	MI	PN	OT	Row mass
No-Male	0.053	0.096	0.200	0.212	0.150	*0.107*
No-Female	0.033	0.130	0.120	0.030	0.075	*0.074*
Si-Male	0.700	0.278	0.440	0.697	0.463	*0.516*
Si-Female	0.213	0.496	0.240	0.061	0.313	*0.303*
TOTAL	1.000	1.000	1.000	1.000	1.000	*1.000*

Interpretación:

De los casos de robo menor a \$1000 ($\mathbf{c}_2$): el 9.6% fueron cometidos por hombres que no fueron considerados culpables, 13.0% fueron mujeres no consideradas culpables, 27.8% fueron hechos hombres que si resultaron culpables mientas que 49.6% son mujeres encontradas culpables.



Representación gráfica de la tabla anterior

Diagrama de mosaico

La *superficie* de un elemento del mosaico es *proporcional a los efectos* contenidos dentro de las celdas de la tabla de contingencia.

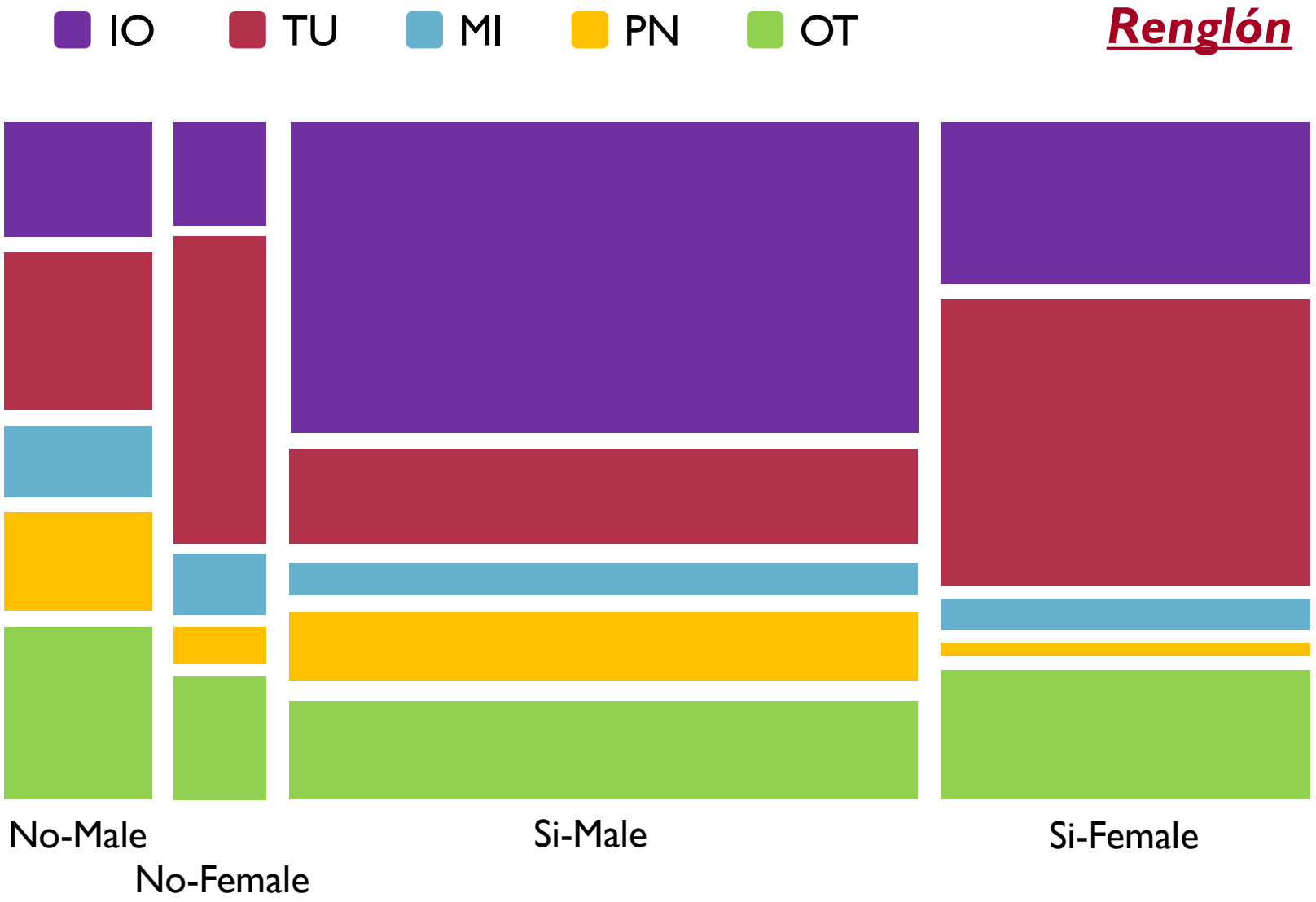
Con R

Sobre el eje horizontal, la distribución marginal de la variable X es utilizada para determinar el largo de los mosaicos. Para cada categoría x_r se reparte de forma proporcional el perfil del fila (R_r).

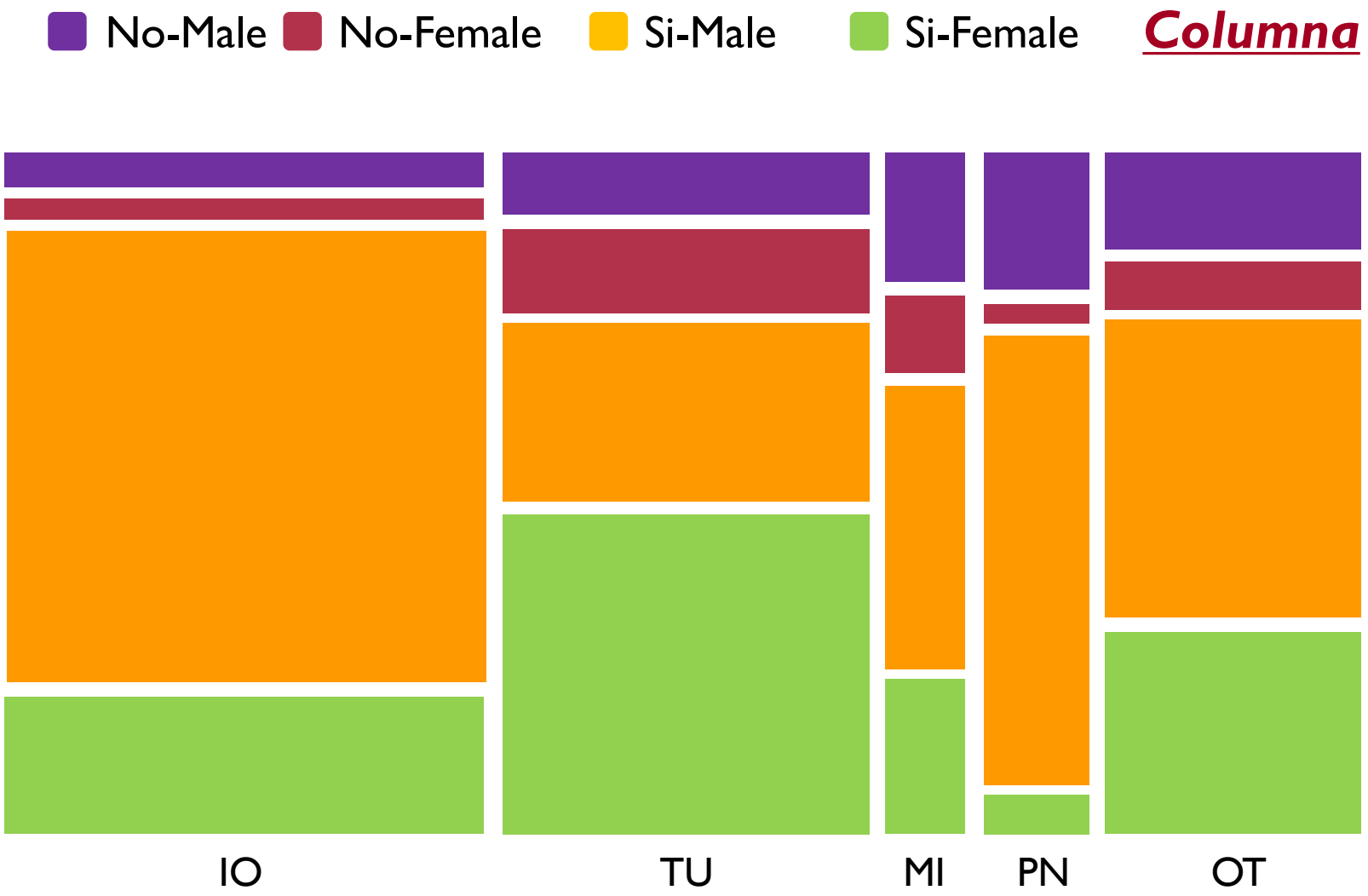
Con C

Sobre el eje horizontal, la distribución marginal de la variable Y es utilizada para determinar el largo de los mosaicos. Para cada categoría y_c se reparte de forma proporcional el perfil del columna (C_c).

Caso: La delincuencia



Caso: La delincuencia



3. Probar la Independencia de las variables

Propósito:

Estudiar la desviación de las *frecuencias observadas* de las *frecuencias esperadas* bajo el supuesto de no relación entre las variables X e Y (independencia). (***supuesto de homogeneidad***).

El test de hipótesis es χ^2 de Pearson. Compara los perfiles de fila y columna con los perfiles marginales correspondientes.

Prueba:

H_0: Ambas variables son independientes (*homogéneas*)

H_1: Existe una relación de dependencia

La matriz de diferencias es una medida de la desviación del modelo de independencia. Esta matriz se obtiene de la diferencia existente entre los valores de la tabla de contingencia (frecuencia observada) y la tabla de valores esperados bajo el supuesto de no existencia de dependencia entre las variables.

Lo que se busca es rechazar estadísticamente la hipótesis nula, lo que indicaría la existencia de asociación entre las variables categóricas.

Caso: La delincuencia – *Matriz de valor esperado (E)*

La matriz de valor esperado del ejemplo es:

	ID	TU	MI	PN	OT	Total
No-Male	16.00	12.27	2.667	3.521	8.536	*43*
No-Female	11.17	8.561	1.861	2.457	5.955	*30*
Si-Male	77.42	59.35	12.9	17.03	41.29	*208*
Si-Female	45.41	34.81	7.568	9.99	24.22	*122*
Total	*150*	*115*	*25*	*33*	*80*	*403*

Donde los valores de la matriz se calculan a partir de los valores de la tabla de contingencia:

$$e_{rc} = \frac{n_{r\bullet}\ n_{\bullet c}}{n} = (Perfil\ marginal\ r)n_{\bullet c} = (Perfil\ marginal\ c)n_{r\bullet}$$

La tabla indica que, de no existir dependencia o asociación entre categorías de las variables, se debieron presentar 16 casos de hombres no culpables en el delito de daños al manejar, o que las mujeres culpables por narcóticos debieron ser casi 10.

Caso: La delincuencia – *Independencia*

El estadístico se calcula: $G^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(n_{ij} - e_{ij})^2}{e_{ij}} = n \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(o_{ij} - o_{i\bullet}o_{\bullet j})^2}{o_{i\bullet}o_{\bullet j}}$

Valores de la matriz de frecuencia relativa (O)

Ejemplo:

→ *No-male y PN*: $\frac{(7-3.521)^2}{3.521} = 3.4372$

→ *Si-female y TU*: $\frac{(57-34.81)^2}{34.81} = 14.139$

	ID	TU	MI	PN	OT	Total
No-Male	4.0037	0.1315	2.0396	3.4372	1.4057	*11.018*
No-Female	3.4051	4.8434	0.697	0.8636	0.0003	*9.8096*
Si-Male	9.8256	12.607	0.2807	2.091	0.4458	*25.250*
Si-Female	3.9598	14.139	0.325	6.3905	0.0252	*24.839*
Total	*21.194*	*31.721*	*3.3423*	*12.782*	*1.8771*	***70.917***

G^2

Criterio de rechazo de H_0: $G^2 \geq \chi^2_{\alpha;(r-1)(c-1)}$ o p-valor < α

Inercia

La inercia total (φ), o *inercia*, *es* una medida de la *varianza total de la tabla* independiente de su tamaño.

$$\varphi = G^2/n$$

Geométricamente, mide lo "*lejos*" que se hallan los perfiles fila (o columna) de su perfil medio, es decir, si las variables categóricas son independientes habrá poca inercia y si son dependientes (están relacionadas) habrá mucha inercia (mucha dispersión). En consecuencia, la inercia total *representa* la magnitud de *la desviación de independencia que necesita ser explicada*.

La inercia total para la tabla de contingencia de caso "*delincuencia*" es:

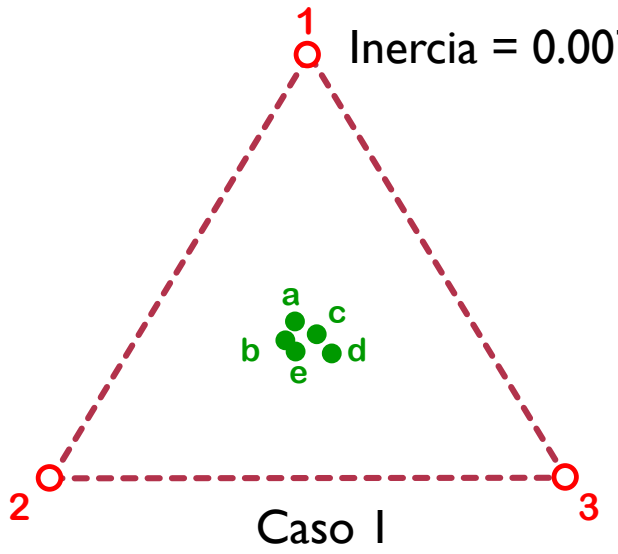
$$G^2/n = 70.916/403 = \mathbf{0.17957}$$

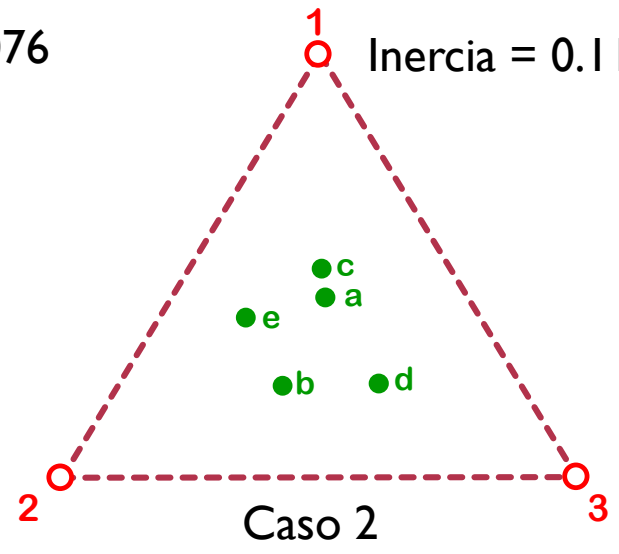
La ***Descomposición en Valores Singulares*** (*DVS*) puede usarse para repartir la inercia total en varias dimensiones facilitando su representación gráfica mediante un bi-plot.

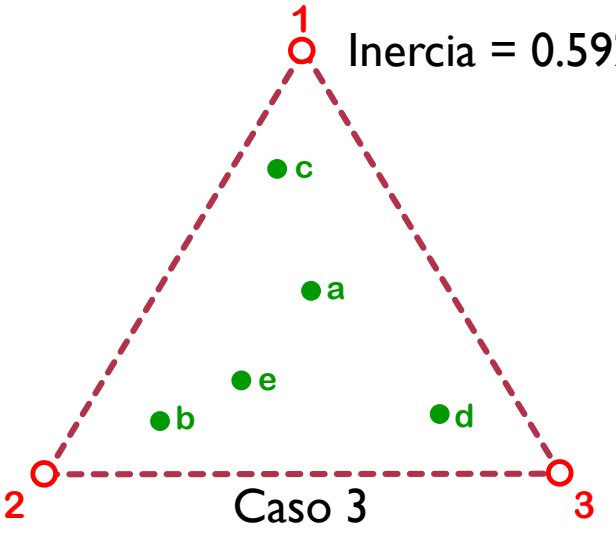
Inercia

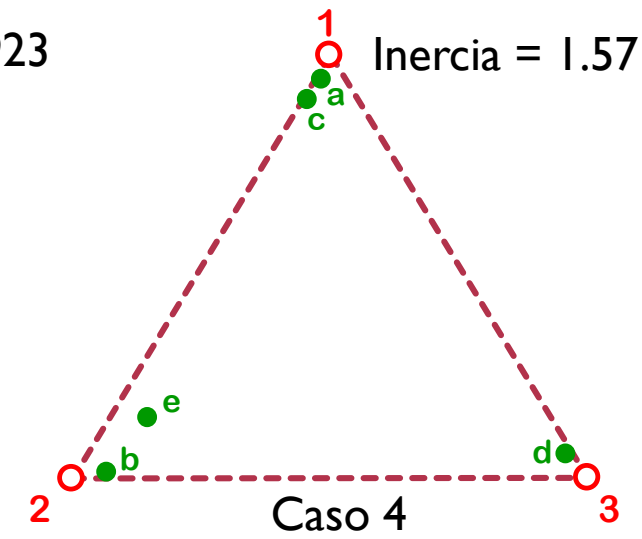
A mayor inercia total, mayor es la asociación entre filas y columnas.

Observar la dispersión de los puntos verdes en el espacio de perfiles. A valores pequeños de inercia existe poca dispersión, mientras que a valores grandes la dispersión es mayor. Además mientras más grande el valor de la inercia más evidente la asociación entre categorías de las variables (rojo y verde).









4. Realizar la grafica Bi-plot

Descomposición en Valores Singulares Generalizada

La dimensión K es el rango de la matriz que se está descomponiendo y es $min[(r-1),(c-1)]$. Y $\mu_1,\mu_2,\ldots,\mu_k$ son los elementos de la matriz diagonal $D_m(K\ x\ K)$.

Los vectores a_k, $k=1,2,\ldots,K$ se llaman los ejes principales de las columnas de ***(0-rc')***. Los vectores b_k, se llaman los ejes principales de los renglones de ***(0-rc')***. Los elementos diagonales $\mu_1,\mu_2,\ldots,\mu_k$ de D_μ se llaman los valores singulares (eigenvalores) de ***(0-rc')***.

De esta forma, la *inercia total* se puede escribir de la siguiente manera:

$$tr\left[D_r^{-1}(O-rc')D_c^{-1}(O-rc')'\right]=\sum_{k=1}^{K}\mu_k^2$$

En el caso de ejemplo, la *inercia total (0.17957)* puede distribuirse en las tres dimensiones *(K=3)* como μ_1^2=0.14191 (80.65%), μ_2^2=0.03286 (18.67%) y μ_3^2=0.00120 (0.68%)

4. Realizar la grafica Bi-plot

Como se aprecia el Análisis de Correspondencia es una técnica de reducción de dimensiones que permite visualizar una nube de puntos multidimensional a un espacio de dos dimensiones. Por lo que en realidad está técnica es el equivalente del Análisis de Componentes Principales pero para variables de tipo cualitativo categórico.

Recordando que, al igual que en el Análisis de Componentes Principales, se buscan las combinaciones lineales de los valores nominales de cada variable categórica que representen el peso que cada uno de ellos tiene en las nuevas dimensiones. Así mismo, habrá una proporción de la varianza explicada (inercia) por las dos dimensiones lo que dará, con cierta confianza, el grado de aproximación entre las categorías de las variables analizadas.

Será el gráfico Bi-plot el que permita hacer la representación gráfica de la reducción dimensional y visualizar la asociación entre categorías.

4. Realizar la grafica Bi-plot

Coordenadas estimadas para los perfiles por columna

99.32%

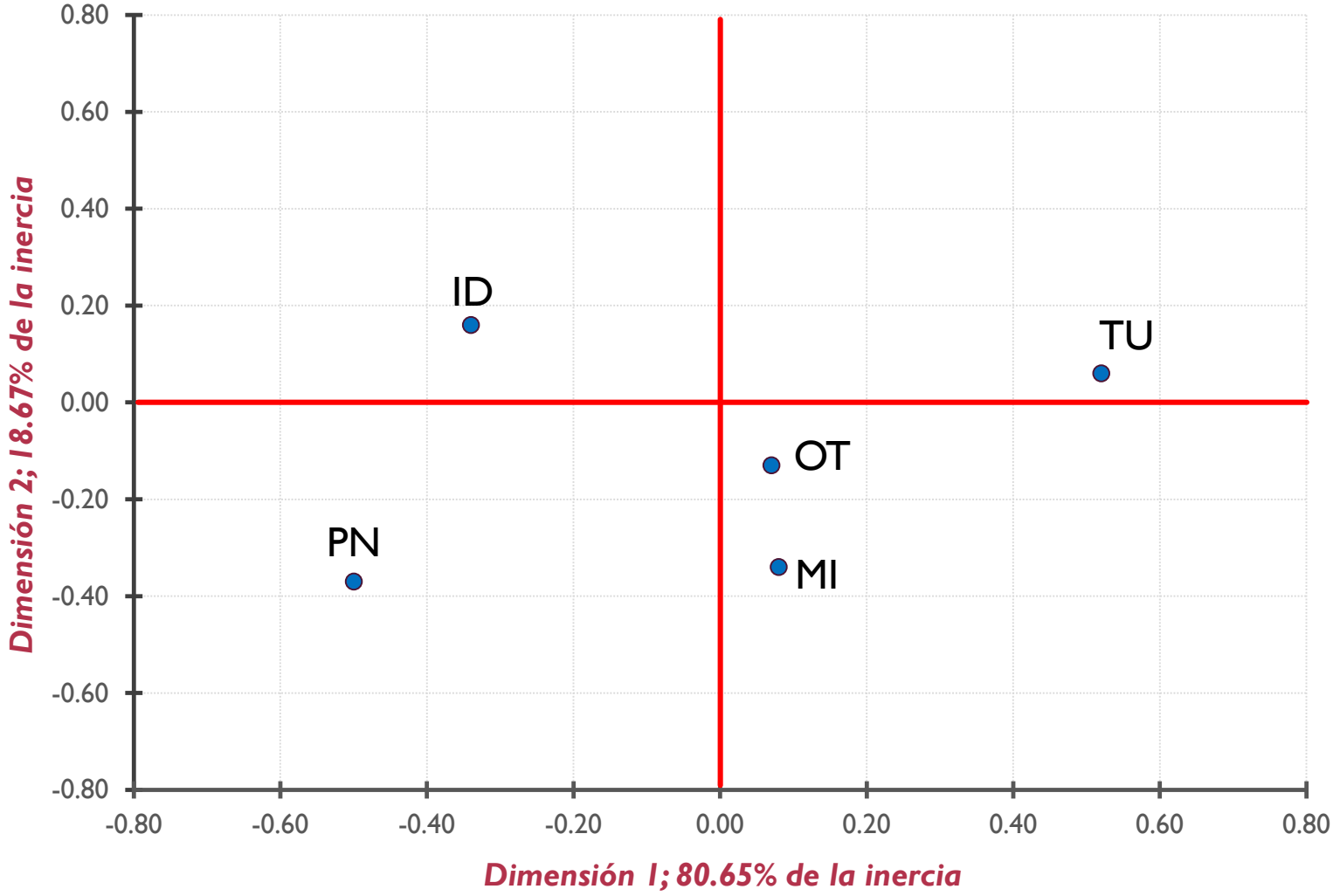
Dimensiones

Perfiles columna	1	2	3
Impaired Driving (ID)	*-0.34*	*0.16*	*0.01*
Theft under $1000 (TU)	*0.52*	*0.06*	*0.00*
Mischief (MI)	*0.08*	*-0.34*	*0.11*
Possession of Narcotics (PN)	*-0.50*	*-0.37*	*-0.01*
Other (OT)	*0.07*	*-0.13*	*-0.05*

a_1 a_2 a_3

4. Realizar la grafica Bi-plot

Representación de las coordenadas para los perfiles por columna



4. Realizar la grafica Bi-plot

Coordenadas estimadas para los perfiles por renglón

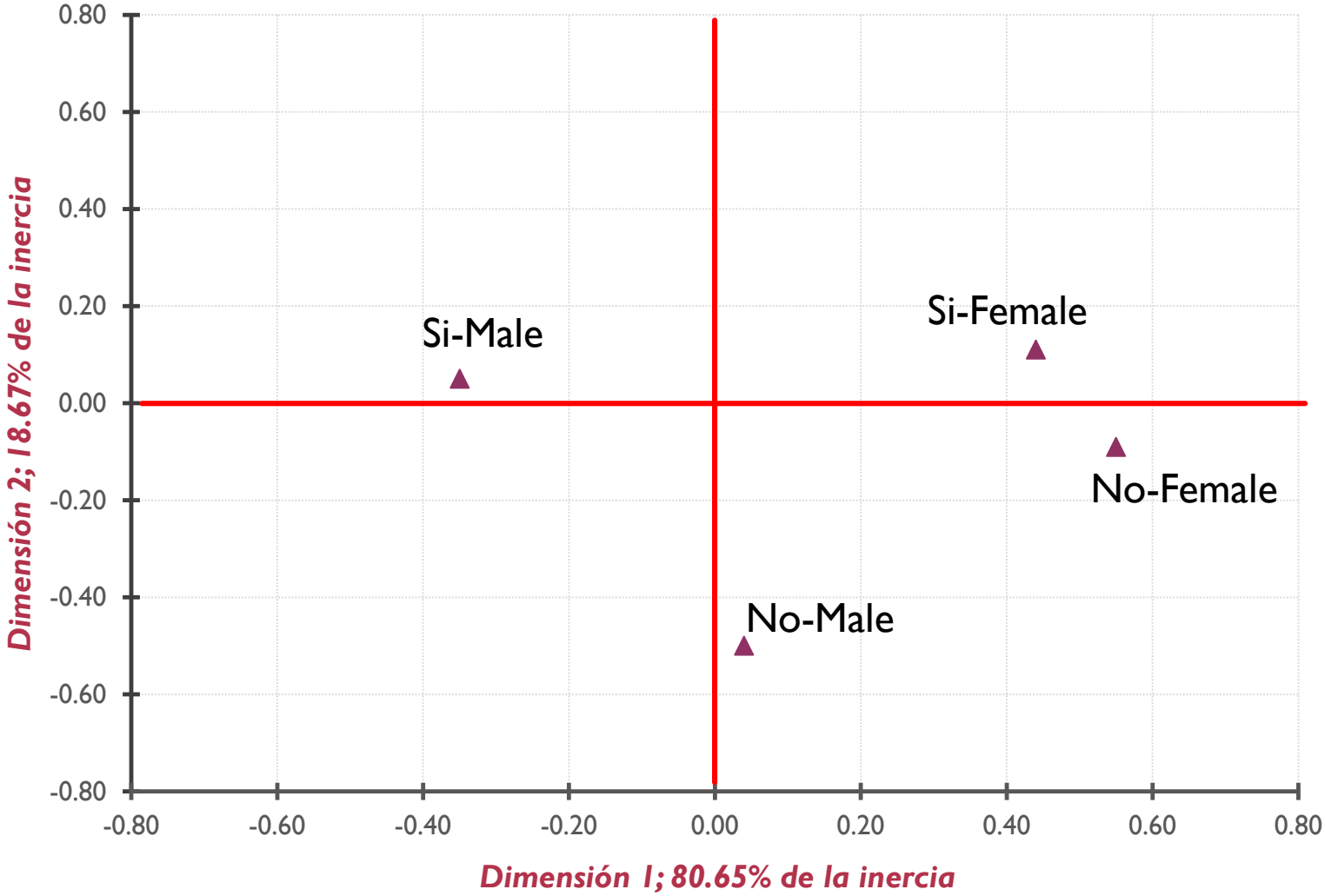
99.32%

Perfiles columna	***Dimensiones*** 1	2	3
No- Male	*0.04*	*-0.50*	*-0.03*
No - Female	*0.55*	*-0.09*	*0.11*
Si - Male	*-0.35*	*0.05*	*0.01*
Si - Female	*0.44*	*0.11*	*-0.03*

b_1 b_2 b_3

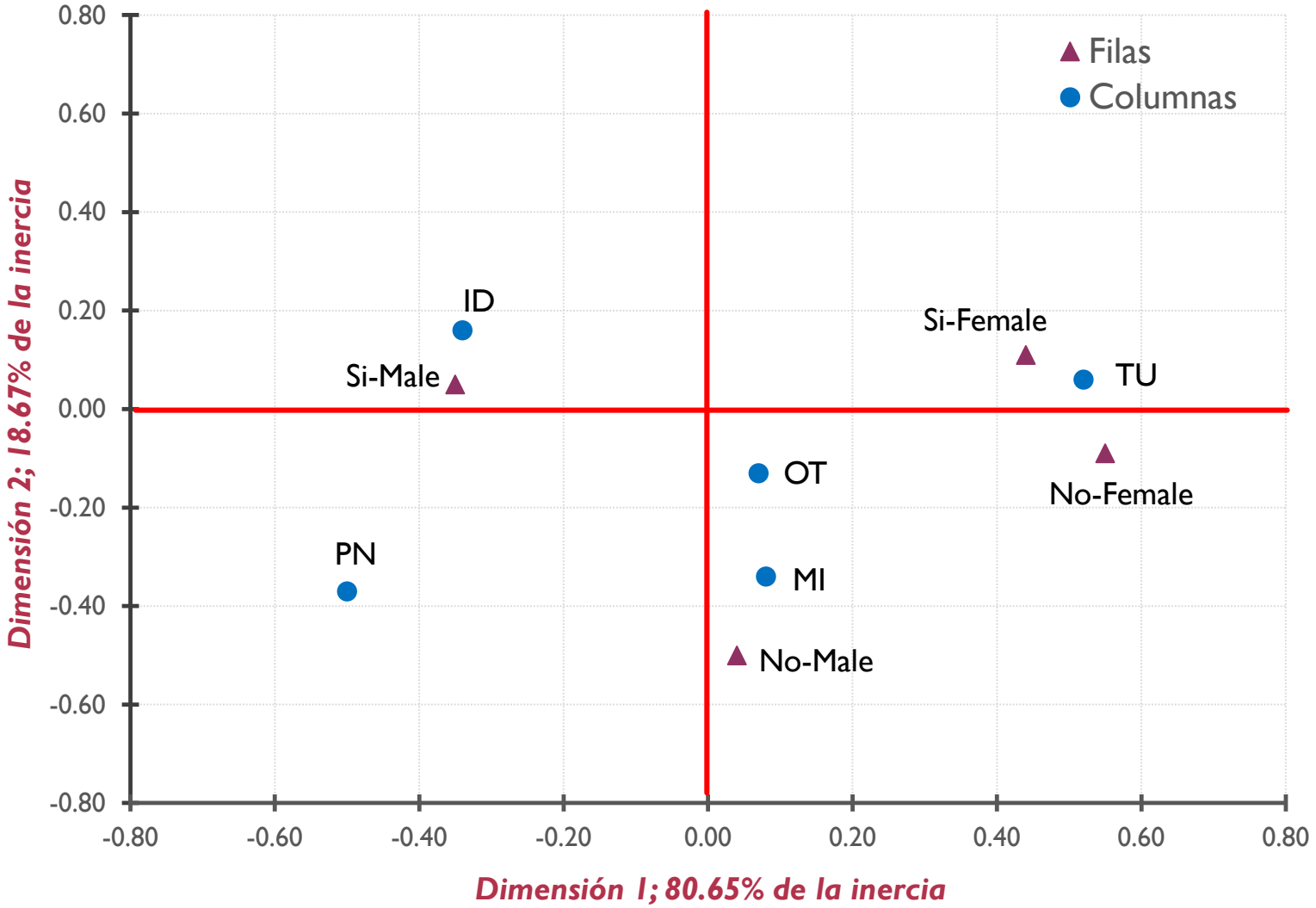
4. Realizar la grafica Bi-plot

Representación de las coordenadas para los perfiles por renglón



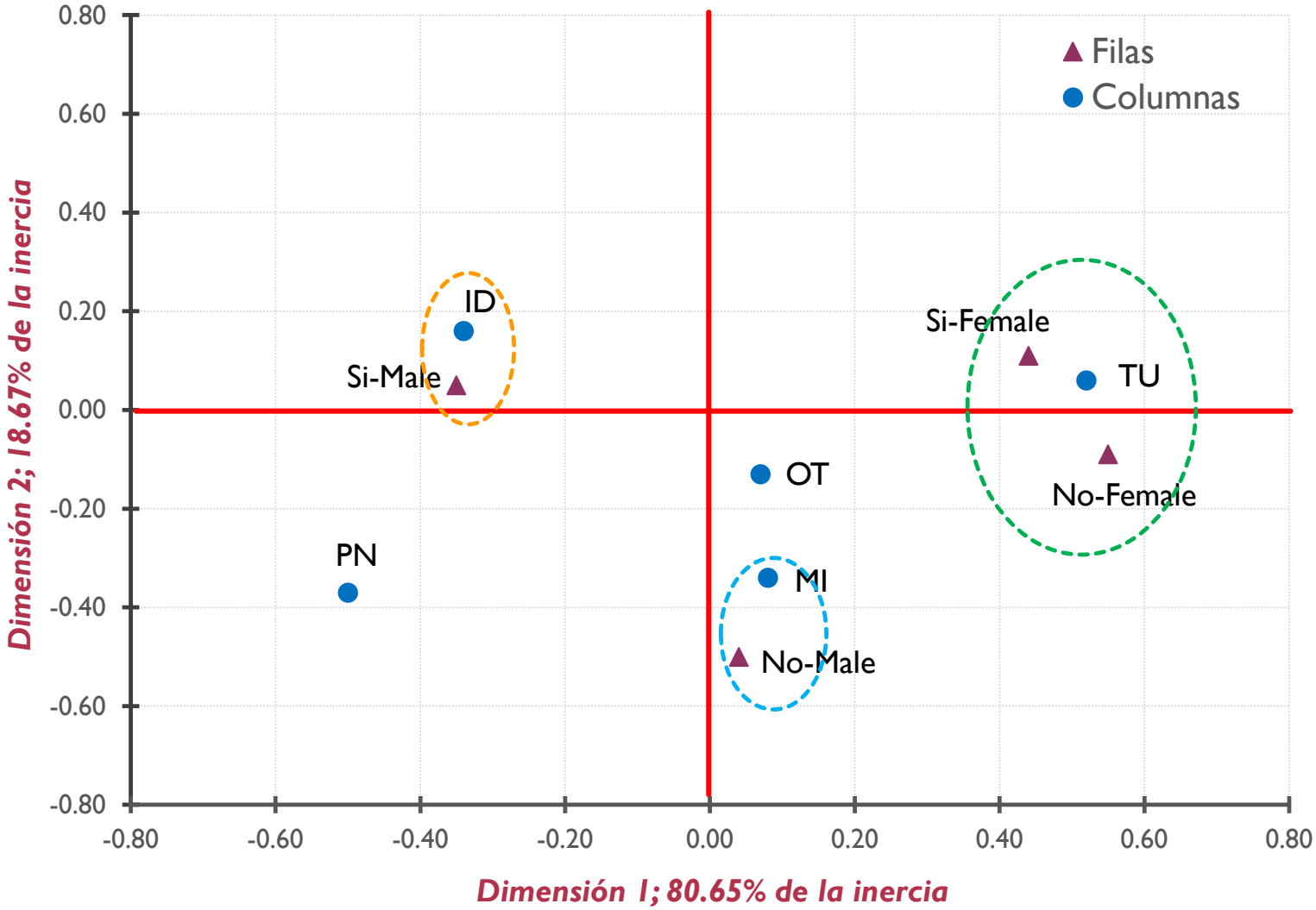
4. Realizar la grafica Bi-plot

Representación conjunta de las variables



4. Realizar la grafica Bi-plot

Asociación de categorías entre las dos variables bajo estudio



4. Realizar la grafica Bi-plot

Del grafico anterior se identifica que:

- Se observa la existencia de ciertas asociaciones, por ejemplo los robos menores a $1000 (TU) son cometidos principalmente por Mujeres.
- Las infracciones menores (MI) son mayoritariamente cometidos por hombres a los cuales después de su proceso no resultan culpables de cometer dicha infracción.
- Por otro lado, los hombres que después del juicio obtienen una sentencia de culpabilidad están relacionados con delitos por daños al manejar (ID).
- De igual forma, los ejes de las dimensiones proporcionan información relevante sobre la estructura de los datos, por ejemplo es muy evidente que la segunda dimensión permite hacer una separación entre los casos en que el acusado es hallado culpable de los casos en que resulta exonerado el imputado.

Resumen

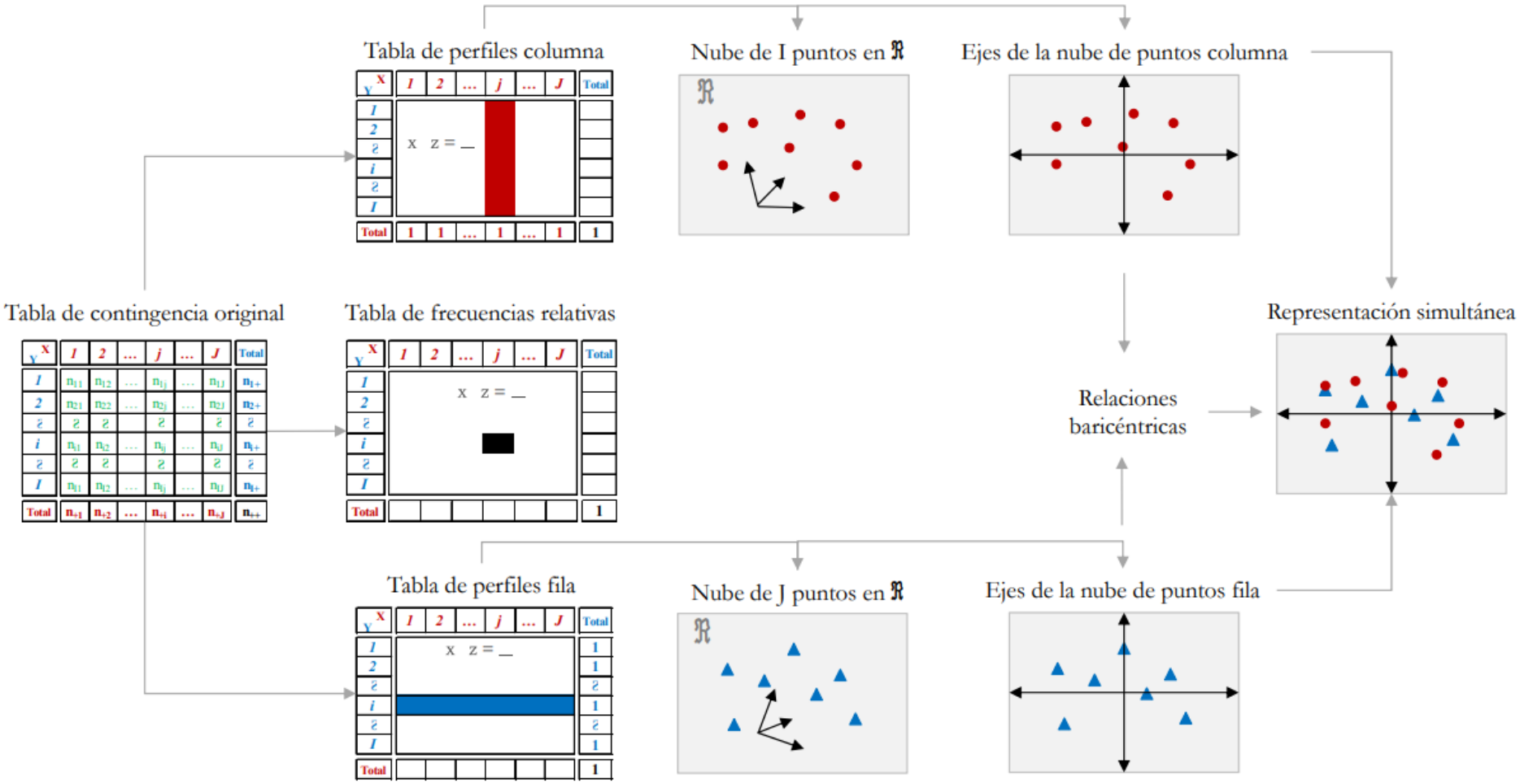


Tabla de perfiles columna
Nube de I puntos en ℜ
Ejes de la nube de puntos columna
Tabla de contingencia original
Tabla de frecuencias relativas
Representación simultánea
Relaciones baricéntricas
Tabla de perfiles fila
Nube de J puntos en ℜ
Ejes de la nube de puntos fila