### A picture containing text Description automatically generatedUniversidad de Granada

### Escuela Internacional de Posgrado

### Máster en Estadística Aplicada

### Materia: Técnicas Estadísticas Multivariantes.

### Alumno: Francisco Javier Márquez Rosales

# **Análisis de Correspondencias:**

# **Ejercicios:**

Noviembre, 2022

Ejercicio de Análisis de Correspondencias

Para la evaluación del tema 6 se deben realizar las siguientes actividades y subirlas a Prado en un archivo pdf preferentemente. La fecha límite es el 23 de noviembre a las 23:55 hora española.

Se pide:

1. Realizar un resumen esquemático de 2 o 3 páginas por una cara sintetizando los resultados principales del tema.

2. Siguiendo los pasos del primer ejemplo dado en el documento (Tema 2: Aplicación en R, sección 2.1: Análisis de Correspondencias Simples), buscar un par de variables categóricas de interés que estén relacionadas y aplicar el Análisis de Correspondencias Simples en R.

**Realizar un informe que contenga las salidas del programa R y el análisis razonado de los resultados obtenidos.**

**Nota: Considerar dos variables categóricas con al menos 3 categorías cada una de ellas.**

3. Realizar un Análisis de Correspondencias Múltiples utilizando al fichero de datos que aquí aparece (datos2.csv). • **Anuncios**: Busca empleo a través de anuncios

• **INEM**: Busca empleo en las oficinas del INEM

• **ETTs**: Busca empleo en las empresas de trabajo temporal

• **Contactos**: Busca empleo a través de contactos personales

• **Oposiciones**: Se está preparando oposiciones

• **Autónomo**: Ha emprendido un negocio

• **Otra**

**Realizar un informe que contenga las salidas del programa R y el análisis razonado de los resultados obtenidos.**

Los datos del fichero datos2.csv corresponden a un grupo de egresados universitarios a los que se les ha preguntado por su forma de buscar empleo tras acabar la Universidad. En este fichero de datos tenemos las siguientes variables:

Todas estas variables son dicotómicas, con categorías Si y No cada una de ellas.

**1. Realizar un resumen esquemático de 2 o 3 páginas por una cara sintetizando los resultados principales del tema.**

Análisis de Correspondencias

El Análisis de Correspondencias es una técnica multivariante de análisis de datos descriptiva desde un enfoque gráfico. Desarrollada por Jean-Paul Benzecri, esta técnica se enfoca en variables cualitativas, nominales u ordinales dispuestas en tablas de contingencia. El objetivo principal es mostrar las relaciones de dependencia de dichas variables de forma gráfica.

Este análisis es conceptualmente similar al análisis de componentes principales, excepto por que los datos se escalan de modo que filas y columnas se tratan de modo equivalente.

En las tablas de contingencia, las categorías de una variable vendrán dadas en las filas y las de la otra vendrán dadas en las columnas. En el centro de la tabla aparecen las frecuencias, kij, es decir, el número de observaciones que presentan la modalidad i de la variable I, y la modalidad j de la variable J.

Diagram

Description automatically generated with medium confidence

Obtenemos la matriz de correspondencias dividiendo cada valor kij por el valor total K, obteniendo la siguiente tabla:

Table

Description automatically generated

Se consideran las frecuencias relativas condicionadas de filas y columnas, para entender cómo se reparten los individuos de las categorías I en las distintas categorías de la variable J, introduciendo así el concepto de perfiles-fila perfiles-columna. Los cuales a nivel tabular se presentan así:

Perfiles Fila

Diagram, schematic

Description automatically generated

Perfiles Columna

Diagram, table, Excel

Description automatically generated

Estos perfiles forman dos nubes de puntos, que pueden ser vistos como coordenadas de un espacio multidimensional:

n puntos del espacio Rp

A picture containing text, clock

Description automatically generated

p puntos del espacio Rn

A picture containing text, clock, watch, gauge

Description automatically generated

2. Siguiendo los pasos del primer ejemplo dado en el documento (Tema 2: Aplicación en R, sección 2.1: Análisis de Correspondencias Simples), buscar un par de variables categóricas de interés que estén relacionadas y aplicar el Análisis de Correspondencias Simples en R.

**Realizar un informe que contenga las salidas del programa R y el análisis razonado de los resultados obtenidos.**

**Nota: Considerar dos variables categóricas con al menos 3 categorías cada una de ellas.**

**Solución:**

Para este ejercicio vamos a utilizar los datos ‘Burdeaux’ que son las opiniones de 200 jueces en una cata a ciegas de cinco tipos diferentes de vinos tinto de la zona de Burdeos, en el suroeste de Francia. Esta data viene incluida en el paquete ‘Ade4’.

**library**("ade4")

**library**("factoextra")

data(bordeaux)

head(bordeaux)

## excellent good mediocre boring

## Cru\_Bourgeois 45 126 24 5

## Grand\_Cru\_classe 87 93 19 1

## Vin\_de\_table 0 0 52 148

## Bordeaux\_d\_origine 36 68 74 22

## Vin\_de\_marque 0 30 111 59

Tal como fue solicitado, vemos que la variable opinión cuenta con 4 categorias y la variable marca cuenta con 5 marcas de vinos distintos.

chisq.test(bordeaux)

##

## Pearson's Chi-squared test

##

## data: bordeaux

## X-squared = 731.85, df = 12, p-value < 2.2e-16

El resultado del test chi-cuadrado nos sugiere rechazar la hipótesis de independencia de las variables (p<0.05) por lo que es válido aplicar un análisis de correspondencias para estudar su posible asociación

Aplicamos ahora el análisis de correspondencias

db <- dudi.coa(bordeaux, scan = FALSE)

db

## Duality diagramm

## class: coa dudi

## $call: dudi.coa(df = bordeaux, scannf = FALSE)

##

## $nf: 2 axis-components saved

## $rank: 3

## eigen values: 0.5906 0.1102 0.03109

## vector length mode content

## 1 $cw 4 numeric column weights

## 2 $lw 5 numeric row weights

## 3 $eig 3 numeric eigen values

##

## data.frame nrow ncol content

## 1 $tab 5 4 modified array

## 2 $li 5 2 row coordinates

## 3 $l1 5 2 row normed scores

## 4 $co 4 2 column coordinates

## 5 $c1 4 2 column normed scores

## other elements: N

El resultadio nos muestran valores para posible tres factores a considerar, el resultado de los valoires de inercia en los tres casos (eigen values): 0.5906 0.1102 0.03109, nos inidca que el primer eje tiene significativamente más poder clasificatorio que los otros dos. El tercero es muy bajo y puede ser descartado.

Ahora obtenemos gráficamente estos resultados:

fviz\_screeplot(db, addlabels=TRUE,ylim=c(0,70))

Chart

Description automatically generated Como se mencionó antes parece adecuado considerar sólo los dos primeros ejes para la explicación de la varianza total.

A continuación se obtienen los pesos de las filas y las columnas:

db$lw

## Cru\_Bourgeois Grand\_Cru\_classe Vin\_de\_table Bordeaux\_d\_origine

## 0.2 0.2 0.2 0.2

## Vin\_de\_marque

## 0.2

db$cw

## excellent good mediocre boring

## 0.168 0.317 0.280 0.235

De las marcas de vinos, todos parecen tener la misma influencia mientras que de las opiniones la de mayor influencia fue la de ‘Bueno’ (good)

Generamos ahora gráficos que nos permitan describir como se relacionan las variables

En este primer gráfico, en este las filas y las columnas se representan en el mismo espacio utilizando las coordenadas principales.

fviz\_ca\_biplot(db)

Chart, histogram

Description automatically generated

Dado que la distancia entre los puntos, azules filas rojos columnas, es la que nos permite concluir sobre las similitudes en este grafico. vemos claramente como

* Grand\_Cru\_classe se relaciona con la mayoria de las opiniones ‘excelente’.
* Cru\_Bourgeois se relaciona con la mayoria de las opiniones ‘bueno’.
* Vin\_de\_table se relaciona con la mayoria de las opiniones ‘aburrido’.
* Vin\_de\_marque se relaciona con la mayoria de las opiniones ‘mediocre’.

Graficamos ahora los perfiles fila en el espacio de las columnas y El segundo gráfico representa los perfiles columna en el espacio de las filas

fviz\_ca\_biplot(db, map ="rowprincipal", arrow = c(TRUE, TRUE))

Chart, line chart

Description automatically generated

fviz\_ca\_biplot(db, map ="colprincipal", arrow = c(TRUE, TRUE))

Chart, line chart

Description automatically generated

En estos gráficos, si el ángulo entre dos flechas es agudo, entonces existe una fuerte asociación entre la fila y la columna correspondientes. Deducimos entonces, sugeridas en el análisis anterior, las siguientes altas correlaciónes:

* Grand\_Cru\_classe y ‘excelente’.
* Cru\_Bourgeois con ‘bueno’.
* Vin\_de\_table con ‘aburrido’.
* Vin\_de\_marque con ‘mediocre’.

3. Realizar un Análisis de Correspondencias Múltiples utilizando al fichero de datos que aquí aparece (datos2.csv). • **Anuncios**: Busca empleo a través de anuncios

• **INEM**: Busca empleo en las oficinas del INEM

• **ETTs**: Busca empleo en las empresas de trabajo temporal

• **Contactos**: Busca empleo a través de contactos personales

• **Oposiciones**: Se está preparando oposiciones

• **Autónomo**: Ha emprendido un negocio

• **Otra**

**Realizar un informe que contenga las salidas del programa R y el análisis razonado de los resultados obtenidos.**

Los datos del fichero datos2.csv corresponden a un grupo de egresados universitarios a los que se les ha preguntado por su forma de buscar empleo tras acabar la Universidad. En este fichero de datos tenemos las siguientes variables:

Todas estas variables son dicotómicas, con categorías Si y No cada una de ellas.