

Estadística Bayesiana – R

Clase 08 de Abril

Análisis de correspondencia

Variables Categóricas

Las variables categóricas son aquellas que representan diferentes categorías o grupos en los que se pueden clasificar los datos.

Las variables categóricas pueden ser de dos tipos principales:

- **Nominales:** Las categorías no tienen un orden específico. Por ejemplo, el género (masculino, femenino) o el tipo de vehículo (automóvil, camión, moto).
- **Ordinales:** Las categorías tienen un orden específico. Aunque las categorías son distintas, existe una relación de orden entre ellas. Por ejemplo, la escala de calificación (bajo, medio, alto) o el nivel de educación (primaria, secundaria, universitaria).

Análisis de Correspondencia

Tipos de variable

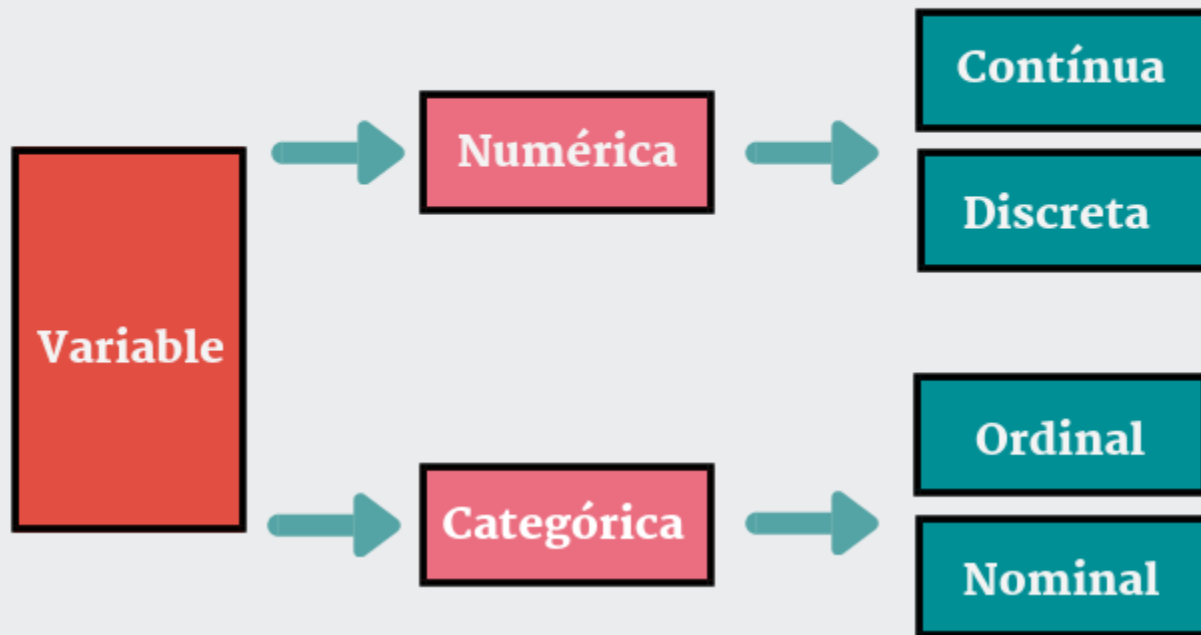


Tabla de contingencia

Las tablas de contingencia de dos variables categóricas son una herramienta fundamental en el análisis estadístico que se utiliza para examinar la relación entre dos variables cualitativas. Estas tablas muestran la distribución conjunta de las dos variables y permiten observar cómo se distribuyen los valores de una variable en función de los distintos niveles de la otra variable.

En una tabla de contingencia, las filas representan los valores de una variable categórica 1, mientras que las columnas representan los valores de la otra variable categórica 2. Dentro de cada celda de la tabla se muestra la frecuencia o el porcentaje de observaciones que caen en esa combinación particular de valores de las dos variables.

Tabla de contingencia







	Pastel	Helados	Donut	Total
Mujer	4	3	6	13
Hombre	5	7	9	21
Total	9	10	15	34

Prueba de independencia chi-cuadrado

La prueba de independencia chi-cuadrado es una técnica estadística utilizada para determinar si existe una relación entre dos variables categóricas en una población. Es especialmente útil cuando se trabaja con tablas de contingencia, que muestran la distribución conjunta de las dos variables.

La prueba de independencia chi-cuadrado se basa en la comparación entre la distribución observada en la tabla de contingencia y la distribución que se esperaría si las dos variables fueran independientes entre sí. La hipótesis nula de la prueba afirma que las dos variables son independientes, mientras que la hipótesis alternativa sostiene que hay alguna asociación entre ellas.

Prueba de independencia chi-cuadrado

El procedimiento típico para llevar a cabo la prueba de independencia chi-cuadrado implica los siguientes pasos:

1. Formular las hipótesis nula (H_0) y alternativa (H_1).
2. Construir una tabla de contingencia que muestre la distribución conjunta de las dos variables categóricas.
3. Calcular la frecuencia esperada para cada celda de la tabla bajo la suposición de independencia.
4. Calcular el estadístico de prueba chi-cuadrado, que mide la discrepancia entre la distribución observada y la esperada.
5. Determinar el valor p asociado al estadístico de prueba.
6. Tomar una decisión sobre las hipótesis nula en función del valor p . Si el valor p es menor que un nivel de significancia predeterminado, se rechaza la hipótesis nula y se concluye que hay una asociación significativa entre las variables.

Prueba de independencia chi-cuadrado

El estadístico de prueba chi-cuadrado se calcula como la suma de los cuadrados de las diferencias entre las frecuencias observadas y esperadas, dividido por las frecuencias esperadas. Si la prueba produce un valor de chi-cuadrado grande, indica una discrepancia significativa entre la distribución observada y la esperada, lo que sugiere una asociación entre las variables

Prueba de independencia chi-cuadrado

Supongamos que estamos investigando si existe una relación entre el género y la preferencia de deportes en una muestra de personas. Tenemos los siguientes datos:

	Futbol	Baloncesto	Tenis	Otro
Hombre	50	30	20	10
Mujer	20	40	30	10

Prueba de independencia chi-cuadrado

Paso 1: Formulación de hipótesis.

- Hipótesis nula (H_0): No hay relación entre el género y la preferencia de deportes.
- Hipótesis alternativa (H_1): Hay una relación entre el género y la preferencia de deportes.

Paso 2: Construcción de la tabla de contingencia.

	Futbol	Baloncesto	Tenis	Otro	Total
Hombre	50	30	20	10	110
Mujer	20	40	30	10	100
Total	70	70	50	20	210

Prueba de independencia chi-cuadrado

Paso 3: Cálculo de las frecuencias esperadas.

Para calcular las frecuencias esperadas, primero necesitamos calcular los totales marginales de cada fila y cada columna.

Ahora, calculamos las frecuencias esperadas para cada celda multiplicando los totales marginales correspondientes. Por ejemplo, la frecuencia esperada para la celda en la fila Hombres y la columna Fútbol sería $(110 * 70) / 210 = 36.67$. Repetimos este proceso para cada celda.

Paso 4: Cálculo del estadístico de prueba chi-cuadrado.

Una vez que tenemos las frecuencias observadas y las frecuencias esperadas, podemos calcular el estadístico de prueba chi-cuadrado utilizando la fórmula:

Prueba de independencia chi-cuadrado

Donde f_o representa la frecuencia observada y f_e representa la frecuencia esperada.

$$\chi^2 = \sum (f_o - f_e)^2 / f_e$$

Análisis de Correspondencia

Análisis de correspondencia simple es una técnica factorial multivariante de interdependencia que tiene como objetivo analizar tablas de contingencia, específicamente analizar la relación entre categorías de dos variables cualitativas.

Análisis de Correspondencia

La idea reproducir información en un número pequeño de factores con la menor pérdida de información posible

En un sentido más práctico queremos explorar y visualizar relaciones entre dos variables categóricas en un espacio bidimensional.

Análisis de Correspondencia

Requisitos para aplicar el método:

- ☐ Tener dos variables categóricas
- ☐ Nivel de medida de dos variables debe ser ordinal o nominal.
Variables cualitativas pueden ser decodificadas, por ejemplo:
edades en rangos de edades
- ☐ Las variables deben estar asociadas

Análisis de Correspondencia

Sea $X_{i \times j}$ una tabla de contingencia

$$\begin{array}{ccc} x_{11} & x_{12} & x_{1\bullet} \\ x_{21} & x_{22} & x_{2\bullet} \\ x_{\bullet 1} & x_{\bullet 2} & x \end{array}$$

Sea $R_{i \times j}$ la matriz donde cada elemento muestra la proporción del elemento x_{ij} respecto a su fila, es decir $r_{ij} = x_{ij}/x_{i\bullet}$.

$$R = \begin{bmatrix} r_{11} & r_{12} \\ r_{21} & r_{22} \end{bmatrix} = \begin{bmatrix} x_{11}/x_{1\bullet} & x_{12}/x_{1\bullet} \\ x_{21}/x_{2\bullet} & x_{22}/x_{2\bullet} \end{bmatrix}$$

Análisis de Correspondencia

El centro de masa R es c^T

$$c^T = (1_{1 \times I} \times X \times 1_{J \times 1})^{-1} (1_{1 \times I} \times X)$$

(Esto me entrega un escalar que es el total de individuos en la tabla de contingencia) *(El total de individuos de cada columna) = centro de masa

-> Es como un vector de promedios

Análisis de Correspondencia

Se puede generar una matriz que me compute las desviaciones respecto a R

R- promedios

$$Y = R - (1_{I \times 1} \times c^T)$$

Análisis de Correspondencia

El centro de masa R es c^T

$$c^T = (1_{1 \times I} \times X \times 1_{J \times 1})^{-1} (1_{1 \times I} \times X)$$

El peso de cada columna refleja la información de cada columna para identificar la fila

$$w_{J \times 1} = 1/c_j$$

Matriz de pesos

$$W_{I \times J} = \text{dig}\{w\}$$

Análisis de Correspondencia

La masa de cada fila es la proporción del reglón con respecto al total de X

$$m = (1_{1 \times I} \times X \times 1_{J \times 1})^{-1} (X \times 1_{J \times 1})$$

(Total de individuos)⁻¹ (Total de individuos en cada fila)

Formamos la matriz masa

$$M_{I \times J} = \text{dig}\{m\}$$

Matriz diagonal con los elementos m calculados antes

Análisis de Correspondencia

La matriz de desviaciones Y se descompone o factoriza mediante descomposición del valor singular tal que

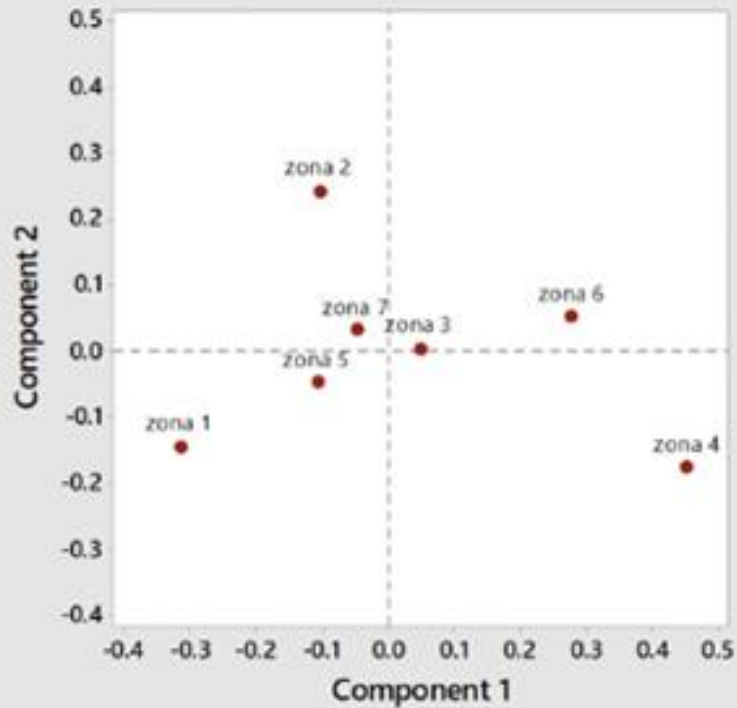
$$Y = P \Delta Q' \text{ con } P'MP = Q'WQ = I \text{ finalmente } F = P \Delta$$

De la matriz F se toman las primeras dos columnas de manera que cada fila se grafica como un punto en el plano. Puntos cercanos significan filas similares, puntos lejanos significan filas diferentes

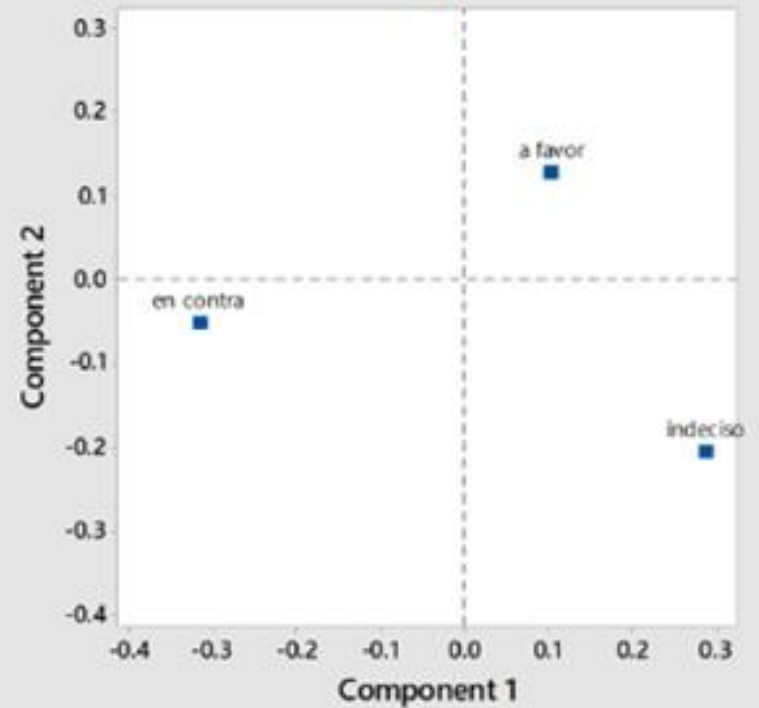
Lo mismo se hace con X' , para hacer un proceso similar al de antes, pero para las columnas.

Análisis de Correspondencia

Row Plot

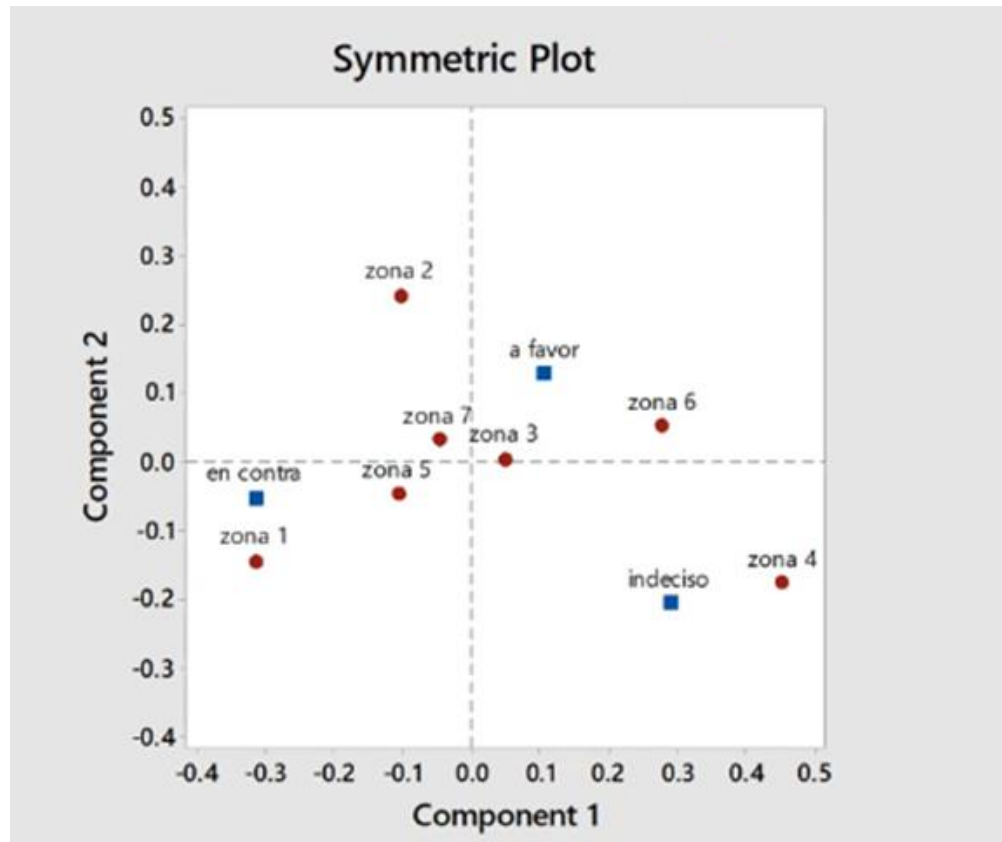


Column Plot



Análisis de Correspondencia

Se empalman ambas graficas para ver cuales quedan juntas



Análisis de Correspondencia

Conclusiones

- ☐ Categorías cercanas están asociadas
- ☐ Mientras más lejos del origen más fuerte la asociación
- ☐ Categorías opuestas, asociadas negativamente

Paso a paso en R

Prueba de chi-cuadrado: Se realiza una prueba de independencia de chi-cuadrado utilizando los datos de la tabla de contingencia para determinar si hay una asociación significativa entre los estratos y la clasificación del interés en billetes de cien.

Análisis de correspondencia simple (ACS): Se realiza un análisis de correspondencia simple utilizando los de la tabla de contingencia y se almacena en la variable ACS.

Gráfica del porcentaje de varianza explicado: Se grafica el porcentaje de varianza explicado por cada eje utilizando un gráfico de screeplot.

Paso a paso en R

Perfiles fila y nube de individuos fila: Se obtienen los perfiles fila y se grafica la nube de individuos fila.

Perfiles columna y nube de individuos columna: Se obtienen los perfiles columna y se grafica la nube de individuos columna.

Representación simultánea: Se grafica la representación simultánea de los perfiles fila y columna