CAPITULO 3

3. ANÁLISIS MULTIVARIANTE: ANÁLISIS DE DATOS CATEGÓRICOS Y CUANTITATIVOS.

* 1. Análisis Factorial de Correspondencias: Concepto de Correspondencias

El análisis de correspondencias es un método multivariante factorial de reducción de la dimensión de una tabla de casos-variables con datos cualitativos con el fin de obtener un número reducido de factores, cuya posterior interpretación permitirá un estudio más simple del problema investigado. El trabajar con variables *cualitativas* o variables *cualitativas categorizadas* confiere a esta prueba factorial una característica diferencial:

No se utilizan como datos de partida mediciones individuales, sino frecuencias de una tabla; es decir, número de individuos contenidos en cada casilla. El análisis factorial es de aplicación, incluso con sólo dos caracteres o variables cualitativas (*análisis de correspondencia simple*), cada una de las cuales puede presentar varias *modalidades* o *categorías*. El método se generaliza cuando el número de variables o caracteres cualitativos es mayor de dos (*análisis de correspondencia múltiple*).

El tratamiento conjunto de dos caracteres o variables cualitativas a través de la prueba de asociación o independencia de la χ2 da información sobre la relación significativa o no entre ambas, sin aclarar qué categorías o modalidades estaban implicadas. En cambio, el análisis de correspondencias extrae relaciones entre *categorías*  y define similaridades o desimilaridades entre ellas, lo que permite su agrupamiento si detecta que se corresponden. Lo cual queda plasmado en un espacio dimensional de escasas variables sintéticas o factores que pueden ser interpretados o nombrados y que, además deben condensar el máximo posible de información.

Las dimensiones que definen el espacio en que se representan las categorías se obtienen como factores cuantitativos, por lo que el análisis de correspondencias es un método de extracción de variables ficticias cuantitativas a partir de las variables cualitativas originales, al definir aquéllas las relaciones entre las categorías de éstas. Esto puede permitir la aplicación posterior de otras pruebas multivariantes cuantitativas (regresión, clusters, …). Una posibilidad propia de este análisis es la inclusión a posteriori de una nueva categoría de alguna de las variables (*categoría suplementaria)* que, no habiendo participado en el cálculo, interese representar para su comparación con las originales. La abundancia y vistosidad de los resultados obtenidos hacen de esta prueba un fuente de hipótesis de trabajo para continuar la investigación.

El carácter cualitativo de las variables también obliga a un proceso distinto. Si se trata de estudios de similaridad o desimilaridad entre categorías, se cuantificará la diferencia o distancia entre ellas. En una tabla de frecuencias, cada categoría de una variable está formada por un conjunto de individuos distribuidos en cada una de las categorías de la otra variable. El proceso para hallar la distancia entre dos categorías de una variable, es utilizado en Estadística para el cálculo de desajuste de dos distribuciones, por medio de las diferencias (desajustes) cuadráticas (para evitar relacionar diferencias positivas con negativas) relativas (es menos clara una diferencia de dos individuos en 4% que un 2%). La suma de estas diferencias cuadráticas relativas entre las frecuencias de ambas distribuciones es el conocido concepto de la χ2. Así, el análisis de correspondencia, puede considerarse como un análisis de componentes principales aplicado a las variables cualitativas, que al no poder utilizar correlaciones, se basa en la distancia no euclídea de la χ2 .

* + 1. **Análisis de Correspondencias Simples.**

El análisis factorial de correspondencias simples está particularmente adaptado para tratar tablas de contingencia, representando los efectivos existentes es las múltiples modalidades (*categorías*) combinadas de dos caracteres (*variables cualitativas*). Al cruzar en una tabla de contingencia el carácter I con modalidades i=1,…,n (filas), con carácter J con modalidades j=i,…,p (columnas), se puede representar el número de unidades estadísticas que se pertenecen simultáneamente a la modalidad *i* del carácter *I* y a la modalidad *j* del carácter *J* mediante *kij*. En este caso, la distinción entre observaciones y variables en el cuadro de doble entrada e artificial, pero, por similitud con componentes principales, suele hablarse a veces de individuos u observaciones cuando nos referimos al conjunto de modalidades del carácter *I* (filas), y de variables cuando nos referimos al conjunto de modalidades del carácter *J* (columnas), tal como apreciamos en la siguiente tabla:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| J  I | 1 | 2… | j… | p |
| 1 |  |  |  |  |
| 2 |  |  |  |  |
| .  .  i  .  . |  | … | .  .  kij  .  . |  |
| n |  |  |  |  |

Los objetivos del análisis factorial de correspondencias son similares a los de componentes principales. Dichos objetivos son:

* El estudio de las relaciones existentes en el interior del conjunto de modalidades I y el estudio de las relaciones existentes en el interior del conjunto de modalidades del carácter J.
* El estudio de las relaciones existentes entre las modalidades del carácter I y las modalidades del carácter J.

La tabla de datos (kij) es una matriz K de orden (n,p) donde kij representa la frecuencia absoluta de asociaciones entre los elementos *i* y *j*; es decir, el número de veces que se presentan simultáneamente las modalidades *i* y *j* de los caracteres I y J.

Designamos:

Ki.== efectivo total de la fila *i*

K.j== efectivo total de la columna *j*

Kij== efectivo total de población

El método buscado para el análisis factorial de correspondencia simple deberá ser simétrico con relación a las líneas y columnas de K (para estudiar las relaciones en el interior de los conjuntos I y J) y deberá permitir comparar las distribuciones de frecuencias de las dos características (para estudiar las relaciones entre los conjuntos I y J).

Para comparar dos líneas entre sí (filas o columnas) en una tabla de contingencia, no interesan los valores brutos sino los porcentajes o distribuciones condicionadas. En una tabla de contingencia, el análisis buscado debe trabajar no con los valores brutos kij sino con perfiles o porcentajes. No interesa poner de manifiesto las diferencias absolutas que existen entre dos líneas, sino que los elementos i,i’ (j,j’) se consideran semejantes si presentan la misma distribución condicionada.

* + 1. **Formación de las nubes**

En Rp tomaremos la nube *n* puntos *i* (n filas de la tabla de perfiles de las variables *i*) cuyas coordenadas son 

En Rn tomaremos la nube *p* puntos *j* (p filas de la tabla de perfiles de las variables *j*) cuyas coordenadas son 

Las transformaciones realizadas son idénticas en los espacios Rp y Rn, pero pueden llevar a transformaciones analíticas diferentes. Los nuevos datos en Rn no son la transpuesta de la matriz en Rp; lo cual conduce a realizar *dos análisis factoriales diferentes*, *uno en cada espacio*. Pero existen relaciones entre los factores que permitirán reducir los cálculos a una sola factorización facilitando además la interpretación.

Se trabajará con la *tabla de contingencia en frecuencias relativas* fij=con k=. Tendremos el esquema:

Perfil de las líneas en Rp

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 1 | 2… | j… | p |  | 1 | 2… | j… | p |
| 1 |  |  |  |  | 1 |  |  |  |  |
| 2 |  |  |  |  | 2 |  |  |  |  |
| .  .  i  .  . |  | … | .  .  kij  .  . |  | .  .  i  .  . |  | … | .  .  fij/fi.  .  . |  |
| n |  |  |  |  | n |  |  |  |  |

Perfil de las líneas en Rn

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 1 | 2… | j… | p |
| 1 |  |  |  |  |
| 2 |  |  |  |  |
| .  .  i  .  . |  | … | .  .  fij/f.j  .  . |  |
| n |  |  |  |  |

 fi.= f.j=

 

El análisis de correspondencias trabaja con perfiles, pero no olvida las diferencias entre los efectivos de cada línea o columna, sino que les asigna un peso proporcional a su importancia en el total. En Rp cada punto *i*  está afectado por un peso fi. y en Rn cada punto *j* está afectado por un peso f.j con lo que, de esta forma, se evita que al trabajar con perfiles se privilegie a las clases de efectivos pequeños.

* + 1. **Definición de distancias.**

El hecho de trabajar con perfiles, en vez de con los valores absolutos iniciales no lleva a utilizar las distancias ji-cuadrado (distancia entre distribuciones) en lugar de la euclídea. Partiendo de la definición de distancia Chi-Cuadrado en el análisis de correspondencias la distancia entre los individuos (punto fila) *i* e *i’* en Rp vendrá definida como:





La única diferencia entre esta distancia y la euclídea es la ponderación, lo que evita que pequeñas diferencias entre las componentes de las líneas influyan mucho en la distancia. El uso de la distancia Ji-Cuadrado estabiliza los datos, hasta el punto de que, por el principio de la equivalencia distribucional, dos líneas (filas o columnas) con el mismo perfil pueden ser sustituidas por una sola afectada por una masa igual a la suma de las masas, sin que se alteren las distancias entre los demás pares de puntos en Rp o Rn.

* 1. Concepto de Correspondencias múltiples.

El análisis de correspondencia múltiple, es un método generalizable al caso de un número de variables o caracteres cualitativos mayor de dos; es decir, generalizable al análisis de correspondencia simple.

Cuando el número de caracteres es mayor que dos (en lugar de tener los caracteres I, J, tenemos los caracteres J1, J2,…, JQ) ya no se puede hablar de tabla de contingencia y la representación tabulada de los datos se complica. No obstante, el análisis en correspondencias múltiples permite estudiar las relaciones entre las modalidades de todas las características cualitativas consideradas.

En el análisis de correspondencias múltiples se ordenan los datos en una tabla Z denominada *tabla disyuntiva completa* que consta de un conjunto de individuos I=1,…,i,…,n (en filas), un conjunto de variables o caracteres cualitativos J1, …,Jk,…, JQ (en columnas) y un conjunto de modalidades excluyentes 1,…,mk para cada carácter cualitativo. El número total de modalidades será entonces:

J=

La tabla disyuntiva completa Z de dimensiones IxJ tiene el siguiente aspecto:

J

JQ

Jk

Jl

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 1………ml | ….. | 1……mk |  | 1……mQ |
| 1 |  |  |  |  |  |
| .  I  . |  |  |  |  |  |
| I | Z1 | … | Zk | … | ZQ |
| .  . |  |  |  |  |  |
| n |  |  |  |  |  |

Z=Zl…Zk…ZQ

El elemento zij de la tabla toma el valor 0 o 1 según que el individuo i haya elegido (esté afectado por) la modalidad j o no. Por consiguiente cada rectángulo de la tabla disyuntiva completada puede considerarse, aunque no lo sea, como una tabla de contingencia cuyos elementos son 0 o 1. La tabla disyuntiva completa Z consta entonces de Q subtablas yuxtapuestas, con la finalidad de obtener una representación simultánea de todas las modalidades (columnas) de todos los individuos (filas). Si las modalidades son excluyentes, cada subtabla tiene un único 1 en cada una de sus fialas.

Si conservamos la notación que hemos manejado hasta ahora tenemos que:

Zij= kij = 0 ó 1

Ki..== Q = número de modalidades (cada subtabla tiene un único 1 en cada fila.

K.j== número de individuos que poseen modalidad j

fij/fi.=kij/kj.=1/Q= inverso del número de modalidades (0 si el individuo no elige j

* + 1. Obtención de los factores: Tabla de Burt.

Para obtener los factores es necesario diagonalizar la matriz V=D-1B/Q donde B=Z’Z es la tabla de Burtz, matriz simétrica formada por Q2 bloques, de modo que sus bloques de la diagonal Z’kZk cuyos elementos son tablas diagonales que cruzan una variable con ella misma, siendo los elementos de la diagonal los efectivos de cada modalidad k.j. Los bloques fuera de la diagonal son tablas de contingencia obtenidas cruzando las tablas de características de dos en dos Z’kZk cuyos elementos son las frecuencias de asociación de las dos modalidades correspondientes. La matriz D es una matriz diagonal cuyos elementos diagonales son los de la matriz de Burtz, siendo nulos el resto de los elementos. El aspecto de la tabla de Burt es el siguiente:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | J1 | J2 | . . . | JQ |
| J1 | 0 | C12 | . . . | C1Q |
| J2 | C21 | 0 | . . . | C2Q |
| .  .  . | .  .  . | .  .  . | .  .  . | .  .  . |
| JQ | CQ1 | CQ2 | . . . | 0 |

Las fórmulas de transición que permiten representar simultáneamente los puntos línea y los puntos columna sobre los mismos gráficos relacionando así los resultados en los dos subespacios, tomarán ahora las siguientes expresiones:





Si tenemos en cuenta que kij = 1 cuando el individuo i posee la modalidad j y cero cuando no, la proyección de un punto individuo *i* sobre el eje α, Fα(i), es el baricentro (salvo un coeficiente de dilatación 1/√λα) de las proyecciones de los puntos modalidades sobre el eje Gα(j). Todas las modalidades están afectadas del mismo peso 1/Q. Análogamente, la proyección de un punto modalidad j sobre el eje α, Gα(j), es el baricentro (salvo un coeficiente de dilatación 1/√λα) de las proyecciones de los puntos individuos que poseen esa modalidad sobre el eje Fα(i), todos ellos afectados del mismo peso k.j.

El centro de gravedad de la nube de puntos de cada variable N(j) en análisis factorial de correspondencias (ACM) es √fi ., que en este caso puede equipararse a una distribución uniforme 1/√n, ya que:

Ki..== Q ⇒ =nQ ⇒ fi.=1/n

El centro de gravedad de las modalidades de cada variable, cada una ponderada por su peso, es el mismo que el de la nube de modalidades N(J), es decir, 1/√n, ya que el centro de gravedad de la subtabla IxJk se obtiene a partir de su distribución marginal. Como sólo recoge una variable, la suma de cada línea es 1 y el total de la tabla es n, de dónde fi=1/n.

Como el análisis factorial de correspondencia es centrado y en el centro de gravedad de las modalidades de una variable coincide con el conjunto J, y con el origen, las modalidades de cada variable están centradas en torno al origen, no pudiendo tener todas el mismo signo.

Al igual que en cualquier Análisis Factorial de Correspondencias, se calculan las *ayudas a la interpretación para cada fila y columna*, definiendo la contribución de una variable Jk al factor α, como la suma de las contribuciones de las modalidades de la variable:

CTAα(Jk)= 

La parte de inercia debida a una modalidad *j* es mayor cuanto menor sea el efectivo de esa modalidad. Si G representa el centro de gravedad, la inercia debida a la modalidad j viene dada por:

I(j)=f.jd2(G,j)=f.j

Por lo tanto, es aconsejable eliminar las modalidades elegidas muy pocas veces, construyendo otra modalidad uniéndola a la más próxima.

La parte de inercia debida a una variable es función creciente del número de la modalidades de respuesta que tiene, ya que la inercia de una variable es la suma de las inercias de sus modalidades:



Si una variable tiene un número de modalidades demasiado grande, al igual que en el caso de que su efectivo sea muy pequeño, conviene reagrupar las modalidades en un número que sea razonable y mantenga el sentido, para evitar así influencias extremas.

La inercia total es la suma de las inercias de todas las modalidades:



J/Q es el número medio de modalidades por variable cualitativa o carácter. En consecuencia, la inercia total sólo depende del número de modalidades y del de preguntas.

Si el número de variables es dos, y cada una tiene dos modalidades, los resultados se pueden analizar tanto por Análisis Factorial de Correspondencias (AFC), como por Análisis de Correspondencia Múltiple (ACM). En el primer caso obtendríamos un único factor que recoge el 100% de la inercia total. Esta inercia dependerá del grado de relación que exista entre las modalidades, de modo que, si están poco relacionadas, la inercia será próxima a cero, y si están muy relacionadas, la inercia tenderá a un valor alto.

Si la misma información la analizamos mediante análisis de correspondencias múltiples, obtendremos siempre la misma inercia (J/Q-1=1), pero obtendremos dos ejes. En el caso en que exista mucha relación entre las variables, el primer eje recogerá gran parte de la inercia (casi 1) y el segundo muy poca, mientras que en el caso del total independencia entre las dos variables ambos factores recogerán la misma cantidad de inercia, es decir ½ cada uno.

* 1. Teoría del Muestreo

**3.3.1 Definiciones básicas**

### Universo

Es el conjunto de unidades o elementos, claramente definido para el que se calculan las estimaciones. Los entes que constituyen el universo tienen características.

### Variable aleatoria

Una variable aleatoria X es una función, cuyo dominio es el espacio muestral (Ω, δ), cuyo espacio de llegada es un conjunto de números reales X:Ω→R.

### Población

La población es una característica medible X de un universo, se pueden definir tantas poblaciones como características medibles tenga un universo, bajo estas condiciones X es una variable aleatoria.

# Población Marco

Es el conjunto de unidades a partir del cual se selecciona la muestra.

## Marco

Es un listado de unidades; en un sentido amplio, incluye toda la información que puede ser utilizada en los procesos de estratificación, selección y estimación.

## Muestra

Es un subconjunto de una población. Una muestra de probabilidad, es una muestra en la que cada elemento de la población tiene una probabilidad conocida de selección.

## Unidad de muestreo

Son las unidades que se seleccionan de una muestra. La elección de la unidad de muestreo más eficiente es una consideración importante dentro del diseño de una encuesta.

* + 1. **Tipos de Muestreo empleado en el estudio**
       1. **Muestreo Aleatorio Simple**

El muestreo aleatorio simple es un método en el cual, todas las unidades de la población tienen la misma probabilidad de ser seleccionadas de una población de tamaño **N**, para formar parte de una muestra de tamaño **n**. Este método es también conocido como Muestreo Aleatorio simple con reposición.

Para calcular el tamaño de la muestra se debe considerar: cierta característica medible e importante dentro de la población, el grado de confianza y la precisión a estimar. Una vez fijado el error máximo admisible, que representa la precisión mínima a elegir de los resultados, y el coeficiente de seguridad o confianza, se necesita conocer además la variabilidad de la población, puesto que, cuanto más dispersos estén los valores de las variables en el estudio, más arriesgado será obtener una muestra de tamaño pequeño.

Sabiendo que el error muestral no es más que el valor absoluto de la diferencia entre un estimador y su parámetro respectivo; tenemos:

e=|| **3.3.1**

Dicho error nos permitirá obtener el tamaño de la muestra, para estimar la media poblacional a partir del Teorema del Límite Central; de este modo nuestro , tenemos así:

** 3.3.2**

Reemplazando 3.3.1 en 3.3.2, tenemos:

 **3.3.3**

Reemplazando la desviación estándar de la media muestral en términos de la cuasivarianza poblacional en la ecuación 3.3.3; tenemos:

 **3.3.4**

A partir de ello, obtenemos el tamaño de la muestra, en función del error e, el tamaño de la población N y la cuasivarianza s2 y con (1-α)100% de confianza. Así:



 **3.3.5**

# 3.3.2.2 Muestreo Estratificado.-

En el muestreo estratificado, la población de ***N*** unidades se dividen primero en subpoblaciones de *N1, N2, N3, ..., NL,* unidades, respectivamente. Estas subpoblaciones, no se traslapan y en su conjunto comprenden a toda la población, por lo tanto,

N1 + N2 + N3 + ... + NL = N

Las subpoblaciones se denominan ***estratos***. Para obtener todo beneficio de la estratificación, los valores de los *Nh* deben ser conocidos. Una vez determinados los estratos, se extraen una muestra de cada uno. Las extracciones deben hacerse independientemente en los diferentes estratos. Los tamaños de muestras dentro de los estratos se denotan con *n1, n2, ..., nL,*  respectivamente.

Si se toma una muestra aleatoria simple en cada estrato, el procedimiento total se describe como ***muestreo aleatorio estratificado***.

La estratificación es una técnica común, y una de sus razones de aplicación es cuando los datos deseados deben tener una precisión conocida en algunas en algunas subdivisiones de la población, par lo que se aconseja tratar cada subdivisión como una “población” por derecho propio.

La teoría del muestreo estratificado se ocupa de las propiedades de las estimaciones de una muestra estratificada y de la mejor elección para los tamaños de muestras *nh* que deben dar la precisión máxima.

En la estratificación se deben agrupar elementos con similares características a fin de que la varianza dentro de cada estrato sea pequeña; al mismo tiempo es deseable que las medias de los distintos estratos sean lo más diferentes posibles.

En el muestreo estratificado las probabilidades de selección de un grupo al otro pueden ser iguales o diferentes. No es necesario que todos los elementos tengan una misma probabilidad de selección aunque se debe conocer la probabilidad que corresponde a cada uno. Por lo general todos los elementos que forman parte de un estrato dado tienen probabilidades de selección iguales.

* + - 1. **Notación**

A continuación se muestra la notación empleada para el muestreo estratificado. El subíndice h denota el estrato, e i la unidad dentro del estrato. Todos los símbolos siguientes se refieren al estrato h.

Nh número total de unidades

nh número de unidades en la muestra

yhi valor obtenido para la i-ésima unidad

Wh ponderación del estrato

fh fracción de muestreo en el estrato

 media verdadera

 media de la muestra

 varianza verdadera

* + - 1. **Afijación en los estratos**

La definición del muestreo estratificado no especifica para la muestra en cada estrato un tamaño determinado. Se puede seleccionar la muestra de modo que en cada estrato tenga el mismo tamaño o distribuir el tamaño total en alguna otra forma. En tanto se seleccione al menos un elemento por estrato se satisface la especificación de una muestra estratificada. A su vez, con dos elementos por estrato, se puede estimar la media como su error. Por lo general el tamaño total de la muestra es mucho mayor que dos elementos por estrato. Por lo tanto surge la necesidad de establecer un criterio para afijar el tamaño total de la muestra en los estratos.

Para determinar la distribución de la muestra entre los distintos estratos, existen dos criterios principales. El primero es la conveniencia, es decir, elegir un procedimiento que sea fácil de aplicar y simple para tabular. Este criterio nos conduce, por lo general, al muestreo proporcional. El segundo criterio es la exactitud: elegir un procedimiento que proporcione el error estándar más pequeño. Esto nos lleva al uso de la afijación óptima.

* + - 1. **Muestreo Estratificado Proporcional**

En el muestreo estratificado es muy común seleccionar en cada estrato la misma proporción de elementos. Según este procedimiento, para seleccionar el diez por ciento de una cierta población, tomaríamos una muestra del diez por ciento de cada estrato.

En este caso, dado que las tasas de muestreo son iguales en todos los estratos, el número de elementos tomados en cada estrato para la muestra, variará de un estrato a otro dependiendo del tamaño de los estratos.

Dentro de cada estrato el tamaño de la muestra será proporcional a la población total del estrato. De ello obtenemos:

 **3.3.6**

De dónde n se calcula a partir de:



De ello:

 **3.3.7**

Así:

 **3.3.8**

Para una población infinita sería:

 **3.3.9**

* **Zα/2:**Es el nivel de confianza seleccionado
* **p:** es la proporción de una categoría de la variable
* **e:** es el error de diseño
* **N:** es el tamaño de la población
* **pxq**: es el estimador de la varianza poblacional

Tratándose de las características de la población en las que se está interesado (es decir, X y ), se puede preparar estimaciones con una muestra estratificada proporcional tan fácil como en una muestra simple al azar con la misma fórmula:

** 3.3.10**

La suma se refiere a todos los elementos muestrales sin considerar los estratos. Se tiene además:

** 3.3.11**

El procedimiento de ponderación simple, hace que el muestreo proporcional sea muy conveniente dada que los resultados son fáciles de tabular. No es necesario tabular cada estrato separadamente, sino que, se pueden sumar en forma conjunta todos los datos muestrales y luego aplicar un cierto factor que podría ser .

Se dice que una muestra con esta característica está *autoponderada.*

* + - 1. **Afijación Óptima**

El muestreo estratificado con afijación óptima, consiste en dejar que la tasa de muestreo en cada estrato cambiara con la cantidad de variabilidad de cada estrato, es decir; hacer la tasa de muestreo en un estrato dado, proporcional a la desviación estándar es dicho estrato. De esta forma, el número de elementos a extraer para la muestra en cada estrato, dependerá sólo del número total de elementos a extraerse de los mismos, sino también de la desviación estándar de la característica que se va a medir. Para esta **afijación óptima,** el número de elementos que se selecciona en un estrato está dado por la fórmula:

 **3.3.12**

Con una afijación óptima el error estándar de la media se reduce a:

** 3.3.13**

Para aplicar este tipo de afijación es necesario conocer los valores de en el universo. Si no se conocen, se pueden estimar dentro de cada estrato usando .