### CAPÍTULO 5

### 5. PRUEBAS ESTADÍSTICAS UTILIZADAS PARA EL ANÁLISIS

El quinto capítulo presenta los conceptos y herramientas que serán utilizadas en los diversos análisis: univariado y multivariado, con el objeto de facilitar al lector, la comprensión sobre el estudio que se efectuará en la investigación, a la información recolectada, por medio del cuestionario.

**5.1. Coeficiente de Sesgo**

Es una medida relativa que permite describir la asimetría de los datos alrededor de la media, y se presentan tres casos: cuando el coeficiente de sesgo es negativo la mayor concentración de datos se encuentran hacia la derecha de la media, es decir que está sesgada hacia la izquierda; cuando el coeficiente de sesgo es positivo la mayor concentración de los datos se encuentran hacia la izquierda de la media, es decir que está sesgada a la derecha; y cuando el coeficiente de sesgo es cero la media y la mediana son iguales, es decir que la distribución es simétrica.

El coeficiente de sesgo, se calcula de la siguiente forma:



**5.2. Coeficiente de Kurtosis**

Es una medida relativa, que permite establecer el grado de apuntamiento o achatamiento de la curva de una distribución con respecto a la distribución normal, que, como sabemos, es campaniforme y simétrica.

La mayor o menor concentración de frecuencias en torno a la media y en la zona central de la distribución dará lugar a una distribución más o menos apuntada, distinguiéndose entre:

**Leptocúrticas:** Distribuciones más apuntadas que la normal y su coeficiente es mayor que tres.

**Mesocúrticas:** Distribuciones con apuntamiento normal y su coeficiente es tres.

**Platicúrticas:** Distribuciones menos apuntadas que la normal y su coeficiente es menor a tres.

El coeficiente de kurtosis se lo calcula a través de la relación entre el cuarto momento central y la varianza al cuadrado, como se observa a continuación:



Para determinar el estimado del coeficiente de kurtosis tenemos que:



donde s4 es el estimador de la varianza.

**5.3. Covarianza**

La covarianza es una definición estadística que mide la relación lineal entre dos variables aleatorias Xi y Xj, a mayor valor absoluto de la covarianza corresponde una mayor dependencia lineal entre Xi y Xj, valores positivos indican que cuando Xi crece también lo hace Xj, valores negativos indican que cuando Xi crece Xj decrece.

La covarianza de Xi y Xj, se la estima de la siguiente forma:

 I = 1,2,...,p j = 1,2,...p

donde y son los estimadores de los valores esperados de Xi y Xj respectivamente.

**5.4. Coeficiente de correlación**

El coeficiente de correlación  mide el grado de asociación lineal entre dos variables, tomando valores entre –1 y 1. Valores de  próximos a 1 indicarán fuerte asociación lineal positiva, mientras que los cercanos a –1 señalaran una asociación lineal negativa, y las cantidades de  inmediatos a cero, mostraran que no existe asociación lineal.

Debido a que el presente estudio está basado en una muestra se hablará del estimador , llamado coeficiente de correlación de la muestra, que puede denotarse por  y su cálculo se lo realiza por medio de la siguiente expresión:



Donde:

 es el estimador de la covarianza entre las variables X y Y.

 es el estimador de la varianza de la variable X.

es el estimador de la varianza de la variable Y.

**5.5. Hipótesis Estadística**

Una ***hipótesis estadística*** es una afirmación o conjetura acerca de la distribución de una o más variables aleatorias. Si una hipótesis estadística especifica por completo la distribución, recibe el nombre de ***hipótesis simple***; si no, se conoce como ***hipótesis compuesta***.

Con frecuencia, los investigadores enuncian como sus hipótesis lo contrario de lo que creen es verdad, con la esperanza de que los procesos de demostración los conduzcan a rechazarlas.

Simbólicamente, se utilizará Ho para la hipótesis nula que deseamos probar y H1 para la alternativa.

Una vez realizado el contraste de hipótesis se procede a decidir si se rechaza o no la hipótesis nula planteada, basados en la información que proporciona una muestra aleatoria de tamaño n: X1, X2, X3,...,Xn.

Las partes funcionales de una prueba estadística son el estadístico de la prueba y la región de rechazo asociada. El estadístico de prueba (como un estimador) es una función de las mediciones muéstrales en el cual se fundamenta la decisión estadística. La región de rechazo, especifica los valores del estadístico de la prueba para los cuales se rechaza la hipótesis nula. Si en una muestra particular el valor calculado del estadístico de la prueba se localiza en la región de rechazo, se rechaza la hipótesis nula Ho y se acepta la hipótesis alternativa H1. Si el valor del estadístico de la prueba no cae en la región de rechazo, se acepta Ho.

**5.6. Tablas de contingencia**

Las tablas de contingencia son arreglos matriciales formados por r filas y c columnas, donde las filas indican la cantidad de niveles que posee un determinado factor X (variable aleatoria) y las columnas determinan de la misma manera la cantidad de niveles de otro factor Y. El objetivo principal es determinar si existe una dependencia lineal ó no lineal entre las variables que se consideran de importancia; a continuación se presenta la forma general de una tabla de contingencia y todos sus elementos.

|  |  |  |
| --- | --- | --- |
|  | Factor 1 |  |
| Factor 2 | ***Nivel 1*** | ***Nivel 2*** |  | ***Nivel c*** |  |
| Nivel 1 | X11E11 | X12E12 |  | X1cE1c | X1. |
| ***Nivel 2*** | X21E21 | X22E22 |  | X2cE2c | X2. |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
| ***Nivel r*** | Xr1Er1 | Xr2Er2 |  | XrcErc | Xr. |
|  | X.1 | X.2 |  | X.c |  |

Donde:

**Xij:** es el número de unidades de investigación sometidas al i\_ésimo nivel del factor 2 y el j\_ésimo nivel del factor 1.

**Eij:** es el número esperado de unidades de investigación sometidas al i- ésimo nivel del factor 2 y al j-ésimo nivel del factor 1, esto es:



donde:







Para el análisis de las tablas de contingencia se postula el siguiente contraste de hipótesis:

Ho: Los factores 1 y 2 son independientes

Vs.

H1: ¬Ho

El estadístico de prueba es , pudiéndose probar que ésta es una variable aleatoria  donde: r es el número de niveles del factor 2 y c el número de niveles del factor 1.

Entonces, se rechaza Ho a favor de H1 si  con  de confianza.

**5.7. Análisis de Correspondencias Simples**

Uno de los fines del análisis de correspondencias es describir las relaciones existentes entre dos variables nominales, recogidas en una tabla de correspondencias, sobre un espacio de pocas dimensiones, mientras que al mismo tiempo se describen las relaciones entre las categorías de cada variable. Para cada variable, las distancias sobre un gráfico entre los puntos de categorías reflejan las relaciones entre las modalidades, con las categorías similares representadas próximas unas a otras.

La proyección de los puntos de una variable sobre el vector desde el origen hasta un punto de categoría de la otra variable describe la relación entre ambas variables.

El análisis de las tablas de contingencia a menudo incluye examinar los perfiles de fila y de columna, así como contrastar la independencia a través del estadístico de chi-cuadrado. Sin embargo, el número de perfiles puede ser bastante grande y la prueba de chi-cuadrado no revelará la estructura de la dependencia. El procedimiento Tablas de contingencia ofrece varias medidas y pruebas de asociación pero no puede representar gráficamente ninguna relación entre las variables.

El análisis factorial es una técnica típica para describir las relaciones existentes entre variables en un espacio de pocas dimensiones. Sin embargo, el análisis factorial requiere datos de intervalo y el número de observaciones debe ser cinco veces el número de variables. Por su parte, el análisis de correspondencias asume que las variables son nominales y permite describir las relaciones entre las categorías de cada variable, así como la relación entre las variables. Además, el análisis de correspondencias se puede utilizar para analizar cualquier tabla de medidas de correspondencia que sean positivas.

##### Formulación del problema

Si *n* y *p* son el número de categorías de la primera y la segunda variable, respectivamente, la tabla de contingencia correspondiente al cruce de los valores de las dos variables tendrá *n* filas y *p* columnas. Cada fila puede ser considerada como un punto dotado de masa, en un espacio de p dimensiones. Las coordenadas de cada punto se obtendrán a partir de las frecuencias en las p celdas de la fila correspondiente. Recíprocamente, cada columna puede ser considerada como un punto, dotado de masa, en un espacio de *n* dimensiones. En este segundo caso, las coordenadas de cada punto se obtendrán a partir de las frecuencias en las *n* celdas de la columna correspondiente. A partir de la representación de los *n* puntos-fila o, equivalentemente, de la representación de los *p* puntos-columna, se tratará de extraer un nuevo espacio, de pequeña dimensión, tal que, al proyectar la nube de puntos en dicho espacio, la deformación de las distancias originales entre los puntos sea pequeña.

En otras palabras, a partir de la representación de las filas, se extraerá un nuevo espacio *c*-dimensional (*c* es igual al mínimo entre n y p, menos 1), de tal forma que:

* El primer eje o factor, F1, del nuevo espacio será aquel tal que, de todas las posibles proyecciones de la nube de puntos sobre un único eje, la mínima deformación sea la obtenida con F1.
* El segundo, F2, será aquel tal que, de todas las posibles proyecciones de la nube de puntos sobre un espacio de dos dimensiones generado por el eje F1 y un segundo eje perpendicular a él, la mínima deformación sea la obtenida con F2.
* En términos generales, el s-ésimo eje, Fs, s=2,..., c, c = [mín (n,p)]-1, será aquel tal que, de todas las posibles proyecciones de la nube de puntos sobre un espacio s dimensiones generado por los ejes F1,...,Fs-1 u un s-ésimo eje perpendicular a todos los s-1 anteriores, la mínima deformación sea la obtenida con Fs.

Alternativamente, el espacio factorial *c*-dimensional podría haber sido extraído a partir de la representación de las *p* columnas como *p* puntos dotados de masa en un espacio *n*-dimensional. Es decir el espacio *c*-dimensional permite representar a las categorías de cualquiera de las dos variables. Mediante la proyección sobre dicho espacio, las similitudes entre las categorías estarán perfectamente representadas. Sin embargo, si el número de factores es grande será difícil interpretarlas. En consecuencia, a partir del espacio de *c* dimensiones se tratará de encontrar un subespacio *k-*dimensional tal que, por un lado, k sea pequeño y por otro, se pierda poca información respecto a la similitud entre las categorías. Teniendo en cuenta que, dado el criterio de extracción, los primeros factores son los más importantes, el subespacio k-dimensional elegido será el determinado por los k primeros factores.

##### Análisis de la relación entre las variables

El análisis de correspondencias simples es una técnica para analizar la homogeneidad entre las categorías de cada una de las dos variables respecto a las categorías de la otra. Mediante el estadístico Ji-cuadrado para tablas de contingencia de doble entrada, era posible contrastar la hipótesis nula de que las categorías de una variable eran homogéneas entre sí respecto a las de la otra, y que dicha hipótesis era equivalente a la hipótesis de independencia entre las variables. Por otro lado, en el supuesto caso de dependencia entre las variables, o heterogeneidad de las categorías, se disponía de medidas que permitían medir el grado de dependencia. Sin embargo, dichas medidas no permitían detectar en qué consistían las similitudes entre las categorías de cualquiera de las dos variables o la dependencia entre ellas (en qué celdas de la tabla de frecuencia observada era significativamente mayor o menor que la esperada bajo el supuesto de independencia. En el caso de que la tabla de contingencia sea pequeña, el aspecto mencionado será sencillo de abarcar mediante la observación de la propia tabla de frecuencias; pero a medida que la tabla sea de gran tamaño el problema se complica; es por eso que en el caso de tablas grandes el análisis de correspondencias simples tratará de simplificar el problema mediante la representación de las categorías en un espacio de pequeña dimensión.

##### Extracción del espacio factorial

Se menciona en el planteamiento del problema, que la extracción del espacio factorial se realizaría a partir de la representación de las categorías como puntos dotados de masa.

La *masa*  de cada punto será igual a la frecuencia relativa de observaciones en la categoría correspondiente.

|  |  |  |
| --- | --- | --- |
|  | *Factor 1* |  |
| *Factor 2* | ***Nivel 1*** | ***Nivel 2*** |  | ***Nivel p*** |  |
| Nivel 1 | K11 | K2 |  | K1p | k1. |
| ***Nivel 2*** | k21 | k22 |  | k2p | K2. |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
| ***Nivel n*** | Kn1 | Kn2 |  | Knp | Kn. |
|  | K.1 | K.2 |  | K.p | K |

Donde:

**Kij:** es el número de unidades de investigación sometidas al i\_ésimo nivel del factor 2 y el j\_ésimo nivel del factor 1.





El peso (masa) para cada punto i es para j = 1,...,p; de igual manera se obtiene el peso para cada punto j en el espacio , para i = 1,...,n.

El resultado de la asignación de masas será que, en las direcciones de los ejes del espacio factorial, unas categorías tendrán más influencia que otras: a mayor masa mayor será la importancia relativa de la categoría correspondiente.

Cada masa es una ponderación asignada con la finalidad de que, a la hora de extraer un eje tratando de que la deformación de la nube de puntos sea mínima, las categorías que se presentan con mayor frecuencia, al tener mayor peso, influyan más en la dirección del eje. En otras palabras, se trata de que, en la deformación experimentada por la nube de puntos al reducir el espacio factorial *c*-dimensional al subespacio generado por los k primeros ejes, los puntos más importantes se vean menos afectados.

Teniendo en cuenta que cada punto tiene un peso o ponderación igual a su masa, un estadístico adecuado para medir la dispersión de la nube de puntos será la *inercia*.

La inercia es el promedio de las distancias de los distintos puntos a su centro de gravedad, estando cada distancia ponderada por la masa del punto correspondiente.

La inercia total será la misma tanto si la nube de puntos corresponde a la representación de filas como si corresponde a la de las columnas. Además se verifica que la inercia total es igual al cociente entre el estadístico Ji-cuadrado para la tabla de contingencia y el total de observaciones. Luego para un tamaño muestral fijo, si el estadístico Ji-cuadrado es grande la inercia también lo será (los puntos estarán muy dispersos). En otras palabras, si las variables son muy dependientes, tanto las filas como las columnas serán muy distintas entre sí, mientras que si son independientes son parecidas.

Al proyectar los puntos correspondientes a las categorías de cualquiera de las dos variables en el espacio factorial *c*-dimensional, la nube de puntos no sufre ninguna deformación. En consecuencia considerando los c factores la inercia total de la muestra estará perfectamente representada y, en particular, también lo estará la de cada una de las categorías.

El objetivo primordial es encontrar el valor de k, tal que, al proyectar la nube de puntos en el subespacio correspondiente, permita interpretar las similitudes entre las categorías. El inconveniente que surge en la elección de k es que cuanto menor sea su valor menor será la calidad de representación. Si k es pequeño la solución será fácil de analizar, pero será poco fiable, mientras que si es grande sucederá lo contrario. La situación ideal entonces sería que la parte de inercia atribuible a los k primeros factores, con k pequeño, fuera muy grande. Dicha situación se dará cuando, dentro del conjunto de categorías de una variable, sea posible distinguir un número pequeño de subconjuntos tales que, por un lado, dentro de cada uno de ellos las categorías sean muy homogéneas entre sí y, por otro, cualquier par de categorías correspondientes a distintos subconjuntos sean muy distintas. Sin embargo, si todas las categorías son completamente distintas, el número de subconjuntos será igual al total de categorías. En dicho caso las inercias atribuibles a cada uno de los factores serán muy parecidas. Recíprocamente, si las inercias atribuibles a cada uno de los factores son muy parecidas, las proporciones correspondientes será próximas a (1/c) y la conclusión será que no existen subconjuntos de categorías relacionadas entre sí. En este sentido, un posible criterio para determinar k es conservar aquellos factores tales que la proporción de inercia explicada por cada uno de ellos sea mayor que 1/c, aunque, en general, será necesaria más de una solución para poder interpretar las relaciones entre todas las categorías. En cualquier caso, salvo que se indique lo contrario, el paquete estadístico SPSS proporciona la solución sobre los dos primeros factores.

**Interpretación de los resultados**

Al analizar el gráfico obtenido, los resultados se interpretan de la siguiente manera:

Si dos categorías de una misma variable estén próximas entre sí significa que en las dos columnas correspondientes de la tabla de contingencia de la distribución de frecuencias relativas en las celdas es parecida. Analizar en qué sentido lo son será equivalente a analizar si la frecuencia tiende a concentrarse en determinadas celdas o si, por el contrario, se reparte homogéneamente a lo largo de todas ellas.

La categoría i, esté próxima a una categoría j significa que en la celda (i,j) la frecuencia presenta mayor concentración que la que cabría esperar si las modalidades de la primera variable (puntos fila) fueran homogéneos respecto a las categorías de los puntos columna.

En términos generales, cuanto mayor sea la distancia al origen mayor será la tendencia de la categoría correspondiente a concentrar su frecuencia en determinadas celdas.

El gráfico permite detectar que categorías se parecen (en el caso de que pertenezcan a una misma variable) o están relacionadas entre sí (en el caso de que pertenezcan a variables distintas), pero no en qué grado.

**Herramientas utilizadas en el análisis de correspondencia**

Los diversos análisis de correspondencia se realizaron con el paquete estadístico SPSS 10.0 el cual provee de lo siguiente:

##### Estadísticos y gráficos

Los estadísticos que se obtienen del análisis de correspondencias simple son: medidas de correspondencia (tabla), perfiles de fila y de columna, valores propios, puntuaciones de fila y de columna, inercia, masa y el diagrama de dispersión biespacial.

**Tabla de correspondencias**: Es la tabla de contingencia de las variables de entrada con los totales marginales de fila y columna.

**Inspección de los puntos de fila:**. Para cada categoría de fila, las puntuaciones, la masa, la inercia, la contribución a la inercia de la dimensión y la contribución de la dimensión a la inercia del punto.

**Inspección de los puntos de columna**: Para cada categoría de columna, las puntuaciones, la masa, la inercia, la contribución a la inercia de la dimensión y la contribución de la dimensión a la inercia del punto.

**Diagrama de dispersión biespacial:** Produce una matriz de diagramas conjuntos de los puntos de fila y de columna.

### Consideraciones sobre los datos

Las variables categóricas que se van a analizar se encuentran escaladas a nivel nominal. Para los datos agregados o para una medida de correspondencia distinta de las frecuencias, utilice una variable de ponderación con valores de similitud positivos.

**Supuestos:** El máximo número de dimensiones utilizado en el procedimiento depende del número de categorías activas de fila y de columna y del número de restricciones de igualdad. Si no se utilizan criterios de igualdad y todas las categorías son activas, la dimensionalidad máxima es igual al número de categorías de la variable con menos categorías menos uno.

Por ejemplo, si una variable dispone de cinco categorías y la otra de cuatro, el número máximo de dimensiones es tres. Las categorías suplementarias no son activas. Por ejemplo, si una variable dispone de cinco categorías, dos de las cuales son suplementarias, y la otra variable dispone de cuatro categorías, el número máximo de dimensiones es dos. Considere todos los conjuntos de categorías con restricción de igualdad como una única categoría. Por ejemplo, si una variable dispone de cinco categorías, tres de las cuales tienen restricción de igualdad, dicha variable se debe tratar como si tuviera tres categorías en el momento de calcular la dimensionalidad máxima. Dos de las categorías no tienen restricción y la tercera corresponde a las tres categorías restringidas. Si se especifica un número de dimensiones superior al máximo, se utilizará el valor máximo.

**Procedimientos relacionados**

Si se encuentran implicadas más de dos variables, se debe utilizar el análisis de homogeneidad; pero si se deben escalar las variables de forma ordinal, utilice el análisis de componentes principales mediante escalamiento óptimo.

#### Modelo

Permite especificar el numero de dimensiones, la medida de distancia, el método de estandarización y el método de normalización.

**Dimensiones en la solución**: En el cual se debe especificar el número de dimensiones. En general, seleccione el menor número de dimensiones que necesite para explicar la mayor parte de la variación. El máximo número de dimensiones depende del número de categorías activas utilizadas en el análisis y de las restricciones de igualdad. El máximo número de dimensiones es el menor entre:

* El número de categorías de fila activas menos el número de categorías de fila con restricción de igualdad, más el número de conjuntos de categorías de fila que se han restringido.
* El número de categorías de columna activas menos el número de categorías de columna con restricción de igualdad, más el número de conjuntos de categorías de columna que se han restringido.

Para el presente análisis el valor tomado para la dimensión en la solución es dos.

**Medida de distancia:** Se puede seleccionar la medida de distancia entre las filas y columnas de la tabla de correspondencias, en este apartado la medida utilizada para la investigación en chi-cuadrado.

* **Chi-cuadradro**: Utiliza una distancia ponderada entre los perfiles, donde la ponderación es la masa de las filas o de las columnas. Esta distancia es necesaria para el análisis de correspondencias típico.
* **Euclídea**: Utiliza la raíz cuadrada de la suma de los cuadrados de las diferencias entre los pares de filas y entre los pares de columnas.

**Método de estandarización**: Se eliminan las medias de filas y columnas. Se centran las filas y las columnas. Este método es necesario para el análisis de correspondencias típico.

**Método de normalización**: El método utilizado fue el siguiente:

**Simétrico:** Para cada dimensión, las puntuaciones de fila son la media ponderada de las puntuaciones de columna divididas por el valor propio coincidente y las puntuaciones de columna son la media ponderada de las puntuaciones de fila divididas por el valor propio coincidente. Utilice este método si desea examinar las diferencias o similitudes entre las categorías de las dos variables.

**5.8. Análisis de Homogeneidad (HOMALS)**

El análisis de homogeneidad cuantifica los datos (categóricos) nominales mediante la asignación de valores numéricos a los casos (los objetos) y a las categorías. El análisis de homogeneidad se conoce también por el acrónimo HOMALS, del inglés homogeneity analysis by means of alternating least squares (análisis de homogeneidad mediante mínimos cuadrados alternantes).

La palabra HOMALS, se encuentra compuesta por las abreviaciones HOM, para el análisis de homogeneidad y ALS alternating least. El término es utilizado para una técnica específica de cuantificación óptima múltiple, el programa HOMALS acepta solamente variables nominales múltiples.

El término de homogeneidad también se refiere al hecho, que el análisis será más exitoso cuando las variables son homogéneas, es decir, cuando ellas participan los objetos (casos) dentro de grupos homogéneos.

El objetivo de HOMALS es describir las relaciones entre dos o más variables nominales en un espacio de pocas dimensiones que contiene las categorías de las variables así como los objetos pertenecientes a dichas categorías. Los objetos pertenecientes a la misma categoría se representan cerca los unos de los otros, mientras que los objetos de diferentes categorías se representan alejados los unos de los otros. Cada objeto se encuentra lo más cerca posible de los puntos de categoría para las categorías a las que pertenece dicho objeto.

El análisis de homogeneidad es similar al análisis de correspondencias, pero no está limitado a dos variables. Es por ello que el análisis de homogeneidad se conoce también como el análisis de correspondencias múltiple. También se puede ver el análisis de homogeneidad como un análisis de componentes principales para datos nominales.

El análisis de homogeneidad es más adecuado que el análisis de componentes principales típico cuando puede que no se conserven las relaciones lineales entre las variables, o cuando las variables se miden a nivel nominal. Además, la interpretación del resultado es mucho más sencilla en HOMALS que en otras técnicas categóricas, como pueden ser las tablas de contingencia y los modelos loglineales. Debido a que las categorías de las variables son cuantificadas, se pueden aplicar sobre las cuantificaciones técnicas que requieren datos numéricos, en análisis subsiguientes.

**Herramientas utilizadas en el análisis de homogeneidad**

Los diversos análisis de correspondencia se realizaron con el paquete estadístico SPSS 10.0 el cual provee de lo siguiente:

**Estadísticos y gráficos**: Los estadísticos que se obtienen del análisis de homogeneidad son: frecuencias, autovalores, historial de iteraciones, puntuaciones de objeto, cuantificaciones de categoría, medidas de discriminación Entre las representaciones gráficas que brinda estas: gráficos de las puntuaciones de objeto, gráficos de las cuantificaciones de categoría, gráficos de las medidas de discriminación.

### Consideraciones sobre los datos

**Datos:** Todas las variables son categóricas (nivel de escalamiento óptimo nominal). Utilice enteros para codificar las categorías. Para minimizar los resultados, utilice enteros consecutivos, comenzando por el 1, para codificar cada variable.

**Supuestos:** Todas las variables del análisis tienen cuantificaciones de categoría que pueden diferir para cada dimensión (nominal múltiple). En el análisis, sólo se utiliza un conjunto de variables. El número máximo de dimensiones utilizado en el procedimiento es el más pequeño entre el número total de categorías menos el número de variables sin datos perdidos y el número de casos menos 1. Por ejemplo, si una variable dispone de cinco categorías y la otra de cuatro (sin datos perdidos), el número máximo de dimensiones es siete ((5+4) - 2). Si especifica un número superior al máximo, se utilizará el valor máximo.

**Procedimientos relacionados**

Para dos variables, el Análisis de homogeneidad es análogo al Análisis de correspondencias. Si piensa que las variables poseen propiedades ordinales o numéricas, se deben utilizar Componentes principales mediante escalamiento óptimo. Si hay conjuntos de variables que son de interés, se debe utilizar el Análisis de correlación canónica no lineal.

La idea básica es realizar una escala de N objetos (y proyectarlos en un espacio Euclidiano de dimensiones pequeñas), en el que los objetos con perfiles similares se encuentren relativamente cerca, mientras que los objetos con perfiles diferentes se encuentren relativamente distantes. El énfasis se produce en los aspectos geométricos del problema, los principios que rigen el Análisis de Homogeneidad son:

* Una escala que consiste en variables numéricas es ***homogénea*** si todas las variables en la escala están linealmente relacionadas.
* Una escala que consiste en variables: nominales, ordinales y numéricas es ***homogenizable*** si todas las variables en la escala pueden ser transformadas o cuantificadas de forma tal que el resultado de la escala es homogénea .
* La ***homogeneidad*** de un conjunto de variables (centradas) es medida por el cálculo de la suma de los cuadrados dentro de los objetos y la suma de los cuadrados entre los objetos.
* El análisis de homogeneidad transforma en variables numéricas (es decir, asigna valores numéricos a cada una de las categorías de las variables) a las cantidades de las variables nominales u ordinales, de tal forma que la homogeneidad es maximizada.

**Solución HOMALS.**

Análisis de Homogeneidad es el término utilizado para la técnica especifica de cuantificación óptima múltiple, así como la correspondiente al programa computacional SPSS 10,0; algunas de sus propiedades básicas son:

Las Cuantificaciones de las Categorías y las Puntuaciones de los Objetos son representados en un espacio común.

Las soluciones sucesivas para las Puntuaciones de los Objetos no estén correlacionadas entre ellas, pero esto no implica que las cuantificaciones sucesivas de la misma variable sean no correlacionadas.

Existe una excepción a la regla anterior, si se aplica HOMALS a una situación con solamente dos variables categóricas, las cuantificaciones sucesivas de estas dos variables no serán correlacionadas, para esto, existe otro programa denominado ANACOR.

Una variable binaria (de dos categorías) puede ser cuantificada en una sola vía. Las cuantificaciones sucesivas de una variable son perfectamente correlacionadas. Cuando todas las variables son binarias, los resultados de HOMALS son los mismos que aquellos obtenidos por el clásico Análisis de Componentes Principales, sin importar las cuantificaciones previas escogidas.

Si una variable *K*, categorías , los puntos categóricos van a ser restringidos a un espacio con *( K, - 1)* dimensiones. Una variable con *K,* categorías nunca tiene más de *( K, -1)* cuantificaciones correlacionadas. De hecho, si existen más de *(K, -1)* soluciones HOMALS existirá dependencia lineal entre las cuantificaciones de la variable. Un punto categórico es el centro del objeto que pertenece a la categoría.

Los objetos con patrones idénticos reciben idénticas puntuaciones de objeto En general, la distancia entre dos puntos del objeto está relacionada con la similitud entre los perfiles o patrones. Una variable discrimina mejor a la extensión si sus puntos categóricos están alejados.

Si una medida discriminante es grande, los puntos categóricos están alejados entre ellos en dicha dimensión y las puntuaciones de objeto están cerca de sus puntos categóricos. De esta manera, los gráficos muestran para cada dimensión que variables son efectivas y cuales no.

La solución es expresada en términos de los valores propios, los cuales proporcionan para cada dimensión el valor promedio de las medidas de discriminación.

Si una categoría es solamente aplicada a un objeto entonces la puntuación del objeto y el punto de la categoría coincidirán.

Los punto categóricos con frecuencia marginales bajas estarán localizados lejos del origen del espacio común, mientras que las categorías con frecuencias marginales altas estarán localizadas cerca del origen.

Las cuantificaciones categóricas de cada variable J, poseen una suma ponderada sobre las categorías iguales a cero.

La solución HOMALS permite trabajar con más de dos dimensiones . La solución HOMALS es anidada. Esto significa que si uno requiere una solución HOMALS  dimensional y después una segunda solución tal que es menor que  entonces las primeras dimensiones de la última solución son idénticas que la solución  dimensional, en otras palabras incrementando el número de dimensiones no requiere la revisión de cuantificaciones en las dimensiones previas.

La solución para las subsecuentes dimensiones son ordenadas, esto significa que la primera solución tiene el mayor valor propio absoluto. Se obtiene un buen resultado cuando los valores propios de la solución HOMALS son grandes y son cercanos a uno, dicho resultado implica que las variables diferentes están cerca las una de las otras.

**5.9. Vector Aleatorio**

Sean X1, X2, X3,...,Xp p variables aleatorias sujetas a investigación. Se define un vector p variado **X** ∈Rp, el que está compuesto por las p variables aleatorias como se muestra a continuación:

**=**[ X1, X2, X3,...,Xp]

**5.10. Matriz de Datos**

En la matriz de datos **X**  cada elemento xij representa el i-ésimo ente al cual se le realiza la j-ésima medida, cada columna corresponde a las p mediciones tomada a un ente. Es decir, a n entes se les miden p características:

 = [  **...** ], ∈Rp

X1, X2,…,Xn es una muestra tomada de una población de tamaño N que tiene p variables o características de interés (una población p-variada).

**5.11. Vector de medias**

Sea: = [X1 X2 ... Xp], un vector p variado, es decir, compuesto por p variables aleatorias, se define al estimador de su vector de medias como:



**5.12. Matriz de varianzas y covarianzas**

Sea: Xt=[X1 X2 ... Xp] un vector p variado, se define para éste el estimador de la matriz de varianzas y covarianzas como:





donde sij=sji, i≠j, se tiene entonces que es simétrica y por lo tanto es diagonizable ortogonalmente.

**5.13. Análisis de componentes principales**

Componentes principales es una técnica estadística multivariada que permite la reducción de datos; algebraicamente son una particular combinación lineal de p variables aleatorias observadas X1, X2,...,Xp. En forma geométrica, esta combinación lineal representa la elección de un nuevo sistema de coordenadas obtenidas al rotar el sistema original, con X1, X2,...,Xp como los ejes coordenados. Los nuevos ejes representan la dirección de máxima variabilidad.

Se tiene una muestra tomada de tamaño n, cuyo vector aleatorio p-variado es:



y cada una de las variables que lo componen son variables aleatorias observables. El vector p-variado **X,**  tiene una matriz de estimadores de varianzas y covarianzas, con pares de valores y vectores propios  donde  y



donde:

 para i ≠j

y

 para i = 1, 2, ..., p

Siendo  = norma del vector  y  es el producto interno del vector  consigo mismo.

Tenemos que Y1, Y2,....,Yp son las componentes principales, donde:



son no correlacionadas, ortonormales entre ellas y además tenemos que  Donde 

Entonces, la primera componente principal es la combinación lineal  de máxima varianza, esto es que maximiza la varianza de Y1 sujeta a que la norma del vector a1 sea unitaria.

La segunda componente principal es la combinación lineal  que maximiza la varianza de Y2, sujeta a que la normal del vector a2 sea unitaria y a que .

En general la i-ésima componente principal es la combinación lineal que maximiza la varianza de , sujeta a que la norma del vector ai sea unitaria y a que  para k<i.

El porcentaje total de la varianza contenida por la i-ésima componente principal, ó su explicación viene dado por:



Este método se lo puede aplicar a la matriz de datos originales, pero cuando estos datos no se encuentran en una misma escala se permite que las que tengan escalas mayores absorban los pesos más significativos; para evitar estos problemas, se llevan todas las variables a una misma escala, lo cual consiste en estandarizar los valores de cada una de estas, es decir a cada variable se le resta su respectiva media y se divide para la desviación estándar, como se muestra a continuación:





M



Donde Z1, Z2,..., Zp son los valores estandarizados de las variables X1,X2,...Xp. Esto visto en forma matricial es:



Siendo  es el vector aleatorio p variado estandarizado, X es el vector aleatorio p variado de los datos originales, μ es el vector de medias asociado a X, en nuestro caso por ser muestra se utilizará el estimador de μ que es , y  se define como:



Donde  es el estimador de la desviación estándar de la variable aleatoria Xi, para i = 1,2...p.

Además se puede probar que  tienen las siguientes propiedades:

E[Z] = 0 y 

Las componentes principales de , que es el vector p variado estandarizado las podemos obtener de los vectores propios de la matriz de correlación Σ, en nuestro caso S, asociada a X.

Obteniendo la i-ésima componente principal para la matriz de datos estandarizada de la siguiente forma:

 i = 1,2,...,p

Para poder conocer si es procedente ó no aplicar el método de componentes principales se utilizan el criterio de Bartlett, el cuál basado sobre un supuesto de normalidad sobre las variables aleatorias, implica en estas una independencia al decir que las covarianzas son 0; así la hipótesis se plantea de la siguiente manera:

 ó  para j≠k

vs.

H1: ¬Ho

Donde  y R = matriz de correlación . El estadístico de prueba es:  donde ; y n = tamaño muestra. Este estadístico de prueba es aproximadamente  donde , siendo f los grados de libertad de la distribución ji\_cuadrado y donde p es el número de variables investigadas.

Entonces, se rechaza Ho a favor de H1 si:  con de confianza.

**Determinación del número de componentes principales:**

**Matriz de Varianzas-Covarianzas**

Cuando se lleva a cabo un análisis de componentes principales, se necesita determinar la dimensionalidad real en el espacio en el que caen los datos; es decir, el número de componentes principales que tienen varianzas mayores que cero. Si varios de los eigenvalores de  son ceros o están suficientemente cercanos a cero, entonces la dimensionalidad real de los datos es la del número de eigenvalores diferentes de cero.

Existen dos métodos para ayudarse a elegir el número de componentes principales que usar cuando se está aplicando este análisis a . Los dos se basan en los eigenvalores de . Sea d la dimensionalidad del espacio en el cual se encuentran en realidad los datos obtenidos.

**Método 1:** Suponga que se desea tomar en cuenta 100% de la variabilidad total en las variables originales. En uno de los métodos para estimar d se considera V=(, para valores sucesivos de k = 1,2,..p. Entonces d se estima por el menor de los valores de k en el que, por primera vez, sobrepasa .

**Método 2:** Para estimar d se utiliza una gráfica de los eigenvalores. El gráfico se construye al situar el valor de cada eigenvalor contra el recíproco. Es decir se sitúan las parejas (1,1), (1,2),..., (1,p). Cuando los puntos de la gráfica tienden a nivelarse, estos eigenvalores suelen estar suficientemente cercanos a cero como para que puedan ignorarse. A lo menos es probable que los más pequeños estén midiendo nada más que ruido aleatorio y éste no debe tratar de interpretar. Por tanto, por este método se supone que la dimensionalidad del espacio de datos es la que corresponde al eigenvalor grande más pequeño.

Matriz de Correlación

Los dos métodos descritos anteriormente para determinar la dimensionalidad del espacio en el cual en realidad se encuentran los datos estandarizados también se pueden aplicar cuando se está realizando un análisis de este tipo, sobre una matriz de correlación, y con esta matriz se puede usar un tercer método.

En éste, se buscan eigenvalores que sean mayores que1 y se estima que la dimensionalidad del espacio muestral es el del número de eigenvalores que sean mayores que 1. La razón para comparar los eigenvalores con 1 es que cuando se está realizando el análisis sobre datos estandarizados, la varianza de cada variable estandarizada es igual a 1. La creencia es que si una componente principal no puede explicar más variación que una sola variable por sí misma, entonces es probable que no sea importante, por lo que frecuentemente se ignoran componentes cuyos eigenvalores son menores que 1. Nunca debe considerarse la comparación de los eigenvalores con 1, cuando se analizan los datos en bruto o, lo que es equivalente, la matriz de varianzas-covarianzas de la muestra.