CAPITULO 4.

 4.- ANÁLISIS MULTIVARIADO.

 4.1.- Introducción.

En este capítulo se realizará el análisis multivariado para obtener un conocimiento del comportamiento de las variables en conjunto, determinando así sus relaciones y los efectos que ejercen una sobre otras, para ello utilizaremos las siguientes técnicas multivariadas: Correlación, Distribución Conjunta, Tablas de Contingencia, Componentes Principales y Correlación Canónica.

Este capítulo está dividido en 6 secciones: En la sección 4.2 se presenta el marco teórico de las técnicas multivariadas antes mencionadas, en la sección 4.3 se podrá apreciar el cálculo de la estimación de la matriz de correlación para determinar el grado de dependencia lineal entre las variables utilizadas en este estudio, luego en la sección 4.4 se procederá a efectuar un análisis bivariado para presentar por medio de un arreglo bidimensional la proporción de la población que poseen determinadas características, a continuación de este análisis se realizarán tablas de contingencia para establecer si existe algún tipo de dependencia entre las variables estudiadas. En la sección 4.5 se realizará el análisis de componentes principales mediante el cual se obtendrá un número reducido de combinaciones lineales de las variables originales, las cuales proporcionan una mejor interpretación y finalmente en la sección 4.6 se aplicará el método de correlación canónica para así determinar la correlación existente entre el grupo de variables relacionadas con la Opinión que el usuario tenga sobre Internet y el Conocimiento y Uso que este le de a Internet

 4.2.- Marco Teórico

 4.2.1 Matriz de Correlación

Sea  la matriz de datos donde cada (), de la matriz representa la i-ésima observación de la j-ésima variable, así:

 = 

Para este estudio, la matriz de datos es una matriz de n filas que corresponden al número de estudiantes usuarios de Internet (628 estudiantes) y p columnas que son 16 variables que han sido investigadas en dichos estudiantes . Las variables que serán analizadas, se presentan en la sección 4.3.

Se define  ( matriz de correlación) como :  , donde  es la matriz de varianzas y covarianzas de  y Ves la matriz cuya diagonal contiene la desviación estándar de cada variable,  resume los coeficientes de correlación de todos los pares de variables entre las p dadas , ,...., .

= 

=

El coeficiente de correlación  determina el grado de dependencia lineal entre la i-ésima y la j-ésima variable. Si =1 o =-1 determina en ambos casos que existe una perfecta dependencia lineal entre las variables, la diferencia en el signo determina únicamente el sentido de variación si es positivo ambas variables aumentan o disminuyen simultáneamente y si es negativo el sentido de variación es opuesto lo que significa que al aumentar una disminuye la otra. El hecho que el coeficiente de correlación sea igual a cero indica que entre ambas variables no existe relación lineal.

Como en este análisis se está trabajando con una muestra se desconocen los parámetros antes mencionados por lo que se hace necesario estimarlos; así que el estimador insesgado de la varianza poblacional es:



Y el estimador de la covarianza entre dos variables Xj y Xk es:

****

En el anexo 4 se presenta la estimación de la matriz de correlación de las variables utilizadas para el presente análisis.

.

 **4.2.2.- Tablas de Contingencia**

Una tabla de contingencia es un arreglo bidimensional en la que se detalla los factores a ser analizados con igual o diferentes niveles de información que nos permitirá determinar si esos dos factores son independientes al realizar un contraste de hipótesis sobre independencia de los factores.

Sea A un factor con r niveles y B un factor con c niveles, se define el modelo de la tabla de contingencia como se muestra en el Cuadro 4.1:

 **Cuadro 4.1**

***Provincia del Guayas: Internet y su Incidencia en la Educación Universitaria Estatal***

**Modelo de la Tabla de Contingencia**

|  |  |  |
| --- | --- | --- |
|  |  | **FACTOR B** |
|  |  | **Nivel 1** | **Nivel 2** | **…** | **Nivel c** | **Xi.** |
| **FACTOR A** | **Nivel 1** | X11E11 | X12E12 | **…** | X1cE1c |  X1 **.** |
| **Nivel 2** | X21E21 | X22E22 | **…** | X2 cE2 c |  X2 **.** |
|  |  |  |  |  |  |
| **Nivel r** | Xr 1Er 1 | X r 2E r 2 | **…** | Xr cEr c |  Xr **.** |
| **X..j** | X**.**1 | X**.**2 | **…** | x**.**3 |  X**..** =n |

Donde :

  = es el número de observaciones

= es el número de valores observados que simultáneamente poseen la i-ésima característica del factor A y la característica j-ésima del factor B.

= es el número de observaciones esperadas con la i-ésima característica del factor A y la característica j-ésima del factor B y se lo obtiene: 

= al número de observaciones que poseen la característica i-ésima del factor B.

= al número de observaciones que poseen la característica j-ésima del factor A.

Con los valores calculados procedemos a postular el siguiente contraste de hipótesis :

H0: El factor A es independiente del factor B.

 Vs.

H1: No es cierto H0.

siendo el estadístico de prueba utilizado  el cual se distribuye según una ji-cuadrado con (r-1)\*(c-1)grados de libertad, se rechaza la hipótesis nula a favor de la hipótesis alterna con (1-α)100% de confianza si

  > .

 **4.2.3.- Componentes Principales**

El análisis de componentes principales es una técnica estadística multivariada que transforma un conjunto de variables correlacionadas de respuestas en un conjunto menor de variables no correlacionadas llamadas componentes principales.

Algebraicamente es una combinación lineal de las p variables aleatorias observables y geométricamente esta combinación lineal representa la elección de un nuevo sistema de coordenadas obtenidas al rotar el sistema original.

El objetivo principal de utilizar componentes principales es explicar la estructura de la matriz de varianzas y covarianzas de un grupo de p variables aleatorias observables, para lo cual se construyen k variables aleatorias no observables, como combinaciones lineales de las p variables.

Existen dos propósitos fundamentales para la aplicación de componentes principales: El primero es la reducción de datos, la cual se da cuando se selecciona un número k de componentes principales que explican la mayor cantidad de variabilidad de las variables originales de modo que k sea menor al número de variables observables. El segundo propósito es la interpretación de las variables no observables con el fin de identificar alguna representación real de estas.

Sea  un vector aleatorio p variado , no necesariamente normal p variado con una matriz de varianzas y covarianzas  estimada por **S** y con un vector de medias estimado por  , se procede a calcular los valores y vectores propios asociados a la matriz de varianzas y covarianzas estimada para formar las combinaciones lineales de acuerdo al siguiente criterio 



donde

  



Donde **(****a1), (****a2), ........ (**P**aP)** son los valores propios y vectores propios asociados a **S**:

Vale destacar que los vectores propios empleados en las combinaciones lineales son ortogonales  y que las varianzas de las componentes principales son las más altas posibles.

La proporción del total de la variación explicada por la k-ésima componente principal es:

 , k=1, 2,......p

Para poder conocer, si en principio, es procedente o no aplicar Componentes Principales se utiliza el criterio de esfericidad de Bartlett, el cuál basado sobre un supuesto de normalidad de las variables aleatorias en discusión, implica en estas una independencia al decir que las covarianzas entre ellas son 0; así la hipótesis se plantea de la siguiente manera:

*H0 =*

 *Vs.*

*H1: No es verdad H0*

Para verificar este contraste de Hipótesis se debe calcular los estadísticos :



 Donde, ;

siendo, R = matriz de correlación muestral; S = matriz de covarianzas muestral y Sii = .

y donde el estadístico de prueba es:

 donde = n-1;

y, n = tamaño muestra. Este estadístico de prueba es aproximadamente χ2(f); donde: f = p(p-1)/2, siendo f los grados de libertad de la distribución Ji-cuadrado y p es el número de variables investigadas.

Entonces, con (1-α)100% de confianza se rechaza H0 a favor de H1 si: 

El número de componentes principales, dependerá del porcentaje de varianza que se desea que estas expliquen.

**4.2.4.- Correlación Canónica.**

El análisis de correlación canónica busca identificar y cuantificar la asociación lineal entre dos grupos de variables **X(1)** y **X(2)** El primer grupo de variables es representadas por un vector aleatorio p variado. El segundo grupo es representado por un vector aleatorio q variado. Donde q ≥ p.

Para los vectores **X(1)** y **X(2)** se tiene:



Considerando a **X(1)** y a **X(2)** conjuntamente tenemos:



 además, 

Las covarianzas entre pares de variables de diferentes conjuntos, esto es una variable de **X(1)** y una variable de **X(2),** pueden ser resumidas en unas pocas covarianzas cuidadosamente escogidas en lugar de las pq covarianzas contenidas en **∑12**.

Consideremos las siguientes combinaciones lineales, sus varianzas y correlaciones: *U=* ***atX(1)****; V=* ***btX(2)****.*

*Var(U)=* ***at ∑11a****; Var(V)=* ***bt ∑22b****;*

*Cov(U,V)=* ***at ∑12 b***

Buscaremos coeficientes de **a** y **b** tal que maximicen: 

En base a esto definimos:

El primer par de variables canónicas, es el par de combinaciones lineales *U*1, *V*1 que tiene varianza unitaria y que maximiza la correlación entre ambas. El segundo par de variables canónicas, es el par de combinaciones lineales *U*2, *V*2 que tiene varianza unitaria y que maximiza la correlación entre ambas, y además en todos los casos no está correlacionada con el primer par de variables canónicas.

En general podemos definir el k-ésimo par de variables canónicas, como combinaciones lineales *U*k y *V*k, de las variables observables del primero y segundo grupo respectivamente, que tienen varianza unitaria y que maximiza la correlación entre ambas, y además en todos los casos no esta correlacionada con las k-1 pares de variables canónicas previas.

Se denomina a la correlación entre el k-ésimo par de variables canónicas a la k-ésima correlación canónica. Para encontrar los vectores **a** y **b** nos basamos en los siguientes resultados:

Suponga que q ≥ p y que los vectores **X(1)** y **X(2)** tienen: *Cov(****X(1)****)=****∑11****;*

*Cov(X(2))=* ***∑22****; Cov(****X****(1),****X****(2))=* ***∑12*** *=* ***∑21t****.*

Los coeficientes de los vectores **a** y **b**, para la k-ésima combinación lineal son:

*Uk=* ***etk ∑11****-1/2****X(1)****; Vk=* ***f tk ∑22****-1/2****X(2*)**; con:

*Corr(Uk ,Vk)= *

Donde *ρ1\*2≥ ρ 2\*2≥...≥ ρ p\*2* son los valores propios de la matriz que resulta de la multiplicación de:  y **e1**, **e2**,...,**ep** son los vectores propios asociados a ésta, y **f1**,**f2**,...**fp** son los vectores propios de la matriz obtenida del producto .

Las variables canónicas tienen las siguientes propiedades:

 *Var (Uk) = Var (Vk) = 1;*

 *Cov (Uk, Ul) = Corr (Uk, Ul) = 0, ;*

 *Cov (Vk, Vl) = Corr (Vk, Vl) = 0, ;*

 *Cov (Uk, Vl) = Corr (Uk, Vl) = 0, ; para k,l= 1,2,...,p*

En nuestro caso, debido a que los datos obtenidos son basados en una muestra, se trabajará con los estimadores poblacionales para **μ** y , entonces tendremos que: =  y = **S**. Siendo  y **S** dichos estimadores