CAPÍTULO IV

IV. ANÁLISIS ESTADÍSTICO MULTIVARIADO

4.1 INTRODUCCIÓN

Una vez concluido el análisis univariado, es necesario realizar un estudio más elaborado, en la que se pueda combinar varias variables en estudio, de tal manera que permita examinar la interacción y la influencia existente entre las variables.

Las técnicas multivariada a utilizar en esta investigación son:

a. Correlación lineal

b. Tabla de contingencia

c. Componentes principales

4.2 DEFINICIONES BÁSICAS

Las técnicas de análisis multivariado están organizados con base en un esquema que divide los procedimientos en **métodos de dependencia,** éstos especifican una o más variables como si se hubiesen pronosticado mediante un conjunto de variables independientes (el análisis de varianza es un ejemplo de este tipo de análisis), y **métodos de interdependencia**, éstas tratan de explicar la interrelación entre todas las variables tomadas como un conjunto que interesa al investigador (el análisis de componentes principales es un ejemplo de este tipo de procedimiento).

Para poder utilizar estas técnicas es necesario representar la información disponible por medio de una tabla de datos rectangular que corresponden a las observaciones disponibles (n filas), y p columnas, correspondientes al número de características medidas.

TABLA XCIII
MATRIZ DE DATOS

		Variables				
		1	**2**	**3**	…	**P**
Individuos	**1**	X_{11}	X_{12}	X_{13}	…	X_{1p}
	2	X_{21}	X_{22}	X_{23}	…	X_{2p}
	3	X_{31}	X_{32}	X_{33}	…	X_{3p}
	⋮	⋮	⋮	⋮	⋱	⋮
	n	X_{n1}	X_{n2}	X_{n3}	…	X_{np}

4.2.1 Tabla de contingencia

En muchas ocasiones, una muestra (de tamaño n) tomada de una población pueden clasificarse de acuerdo con dos criterios diferentes, por lo tanto, es interesante saber si los métodos de clasificación son estadísticamente independientes. Supóngase que el primer criterio tiene **r** niveles, y que el segundo tiene **c** niveles. Sea f_{ij} la frecuencia observada para el nivel **i** del primer método de clasificación y el nivel **j** del segundo criterio. En general, los datos aparecerán como se muestran en la tabla XCIV, usualmente se la conoce como **tabla de contingencia.**

TABLA XCIV
DISEÑO DE UNA TABLA DE CONTINGENCIA

		COLUMNAS				
		1	**2**	**3**	…	**c**
RENGLONES	**1**	f_{11}	f_{12}	f_{13}	…	f_{1c}
	2	f_{21}	f_{22}	f_{23}	…	f_{2c}
	3	f_{31}	f_{32}	f_{33}	…	f_{3c}
	⋮	⋮	⋮	⋮	⋱	⋮
	r	f_{r1}	f_{r2}	f_{r3}	…	f_{rc}

El interés recae en probar la hipótesis de que los métodos de clasificación renglón – columnas son independientes. Si se rechaza esta hipótesis, entonces se concluye que existe alguna relación entre los dos criterios de clasificación.

El procedimiento de prueba puede obtenerse por medio de un estadistivo valido para n grande. En general tenemos, su $\mathbf{p_{ij}}$ es la probabilidad de que un elemento seleccionado al azar caiga en la ij-ésima celda, dado que las dos clasificaciones son independientes. Entonces $\mathbf{p_{ij} = u_i\, v_j}$, donde $\mathbf{u_i}$ es la probabilidad de que un elemento seleccionado al azar pertenezca a la renglón de la clase i, y $\mathbf{v_j}$ es la probabilidad de que un elemento pertenezca a la columna de la clase i. Ahora, si se supone independencia, los estimadores de $\mathbf{u_i}$ y $\mathbf{v_j}$ son:

$$\hat{u}_i = \frac{1}{n}\sum_{j=1}^{c} f_{ij}$$

$$\hat{v}_j = \frac{1}{n}\sum_{i=1}^{r} f_{ij}$$

Por lo tanto, la frecuencia esperada de cada celda es

$$e_{ij} = n\,\hat{u}_i\,\hat{v}_j = \frac{1}{n}\sum_{j=1}^{c} f_{ij}\sum_{i=1}^{r} f_{ij}$$

Entonces, para n grande, el estadístico

$$\chi_o^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(f_{ij}-e_{ij})^2}{e_{ij}}$$

tiene una distribución aproximada ji-cuadrada con (r-1)(c-1) grados de libertad. Por consiguiente, la hipótesis de independencia debe rechazarse si el valor observado del estadístico de prueba χ^2_0 es mayor que $\chi^2_{\alpha\,,\,(r-1)(c-1)}$.

4.2.2 Componentes principales

Es una técnica multivariada de interdependencia, que estudia p variables de investigación, a través de las cuales se generarán k variables latentes, k<p, que contengan aproximadamente tanta información como las p variables originales, donde los objetivos generales de este análisis son la reducción de los datos y la interpretación.

Sea $\mathbf{X^t} = [X_1, X_2, ..., X_p]$ el vector aleatorio p-variado con media $\boldsymbol{\mu}$ y matriz de varianza y covarianza ($\boldsymbol{\Sigma}$), supongamos además que los valores propios de $\boldsymbol{\Sigma}$ son $\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \cdots \geq \lambda_p \geq 0$ y sus respectivos vectores propios ortonormales $\boldsymbol{\beta_1}, \boldsymbol{\beta_2}, \boldsymbol{\beta_3}, \cdots, \boldsymbol{\beta_p}$; definamos p variables no observadas denotadas por $Y_1, Y_2, \cdots, Y_p$ como una combinación lineal de $X_1, X_2, X_3, \cdots, X_p$, tal que:

$$
\begin{aligned}
Y_1 &= \beta_{11}X_1 + \beta_{12}X_2 + \ldots + \beta_{1p}X_p \\
Y_2 &= \beta_{21}X_1 + \beta_{22}X_2 + \ldots + \beta_{2p}X_p \\
&\vdots \\
Y_p &= \beta_{p1}X_1 + \beta_{p2}X_2 + \ldots + \beta_{pp}X_p
\end{aligned}
$$

En síntesis, tenemos:

$$
\begin{aligned}
Y_i &= \beta_{1i}X_i + \beta_{2i}X_i + \ldots + \beta_{pi}X_i = \boldsymbol{\beta_i^t X} \qquad \boldsymbol{\beta_i, X} \in R^p \\
&= \begin{bmatrix} \boldsymbol{\beta_1} & \boldsymbol{\beta_2} & \cdots & \boldsymbol{\beta_p} \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_P \end{bmatrix}
\end{aligned}
$$

De resultados podríamos sabe que:

$$E[Y_i] = \beta_i^t \mu \qquad i = 1,2,\ldots,p$$
$$Var(Y_i) = \beta_i^t \Sigma \beta_i = \lambda_i$$
$$Cov(Y_i, Y_j) = \beta_i^t \Sigma \beta_j = 0 \qquad i \neq j$$

Las componentes principales de $\mathbf{X}$ son aquellas combinaciones lineales Y_1, Y_2, $\cdots$, Y_p que son no correlacionadas entre sí y cuyas varianzas son tan grandes como sea posible.

Primera componente principal = combinación lineal $\beta^t_1 \mathbf{X}$ que maximiza
Var ($\beta^t_1 \mathbf{X}$) sujeto a:
$\langle \beta_1 , \beta_1 \rangle = 1$

Segunda componente principal = combinación lineal $\beta^t_2 \mathbf{X}$ que maximiza
Var ($\beta^t_2 \mathbf{X}$) sujeto a:
$\langle \beta_2 , \beta_2 \rangle = 1$ y
Cov ($\beta^t_1 \mathbf{X}$, $\beta^t_2 \mathbf{X}$)=0
Var ($\beta^t_2 \mathbf{X}$) > Var ($\beta^t_1 \mathbf{X}$)

I – ésima componente principal = combinación lineal $\beta^t_i \mathbf{X}$ que maximiza
Var ($\beta^t_i \mathbf{X}$) sujeto a:
$\langle \beta_i , \beta_i \rangle = 1$
Cov ($\beta^t_k \mathbf{X}$, $\beta^t_i \mathbf{X}$)=0 ; k<i
Var ($\beta^t_k \mathbf{X}$) > Var ($\beta^t_i \mathbf{X}$)

Se desea obtener la mayor proporción del total de varianza de la población, explicada por las k-ésima componentes principales, donde el valor individual de su aporte está dado por:

$$\frac{\lambda_k}{\lambda_1 + \lambda_2 + \cdots + \lambda_p} \quad \text{para } k = 1,2,\cdots,p$$

4.3 ANÁLISIS DE CORRELACIÓN LINEAL APLICADAS A LAS CARACTERÍSTICAS DE ESTUDIO

En la tabla XCV se muestra la matriz de correlación entre las variables de estudio para el año lectivo 1998-99, el propósito es determinar que variables están relacionadas la una con la otra, es decir que dependientes son entre sí. La mayoría de los coeficientes son bajas (valor absoluto menor a 0.30), tomaremos como valores significativos a los coeficientes cuyos valores absolutos sean mayores a 0.30, es decir tienen una correlación aceptable en el intervalo [0.3 ; 0.6) y correlación alta [0.6 ; 1.0), en la matriz de correlación (tabla XCV) los valores con **negrillas** tienen correlaciones significativas. La más notable relación de dependencia lineal está entre el número de aulas que hay en la institución educativa (X_{10}) y el número de alumnos, sean estos promovidos (X_{12}), no promovidos (X_{13}) y desertores (X_{14}), en este caso los estimadores de coeficientes de correlación son iguales a 0.742, 0.636, 0.573 respectivamente.

La variable X_5 (sexo) está correlacionada significativamente con X_{12} y X_{13}, que corresponden respectivamente al número de estudiantes promovidos y reprobados.

Definición de las variables de la tabla XCV:

X_1 :	Cantón	X_8 :	Tenencia del edificio
X_2 :	Parroquia	X_9 :	Tipo de construcción
X_3 :	Zona (urbana o rural)	X_{10} :	Número de aulas
X_4 :	Jornada	X_{11} :	Personal del colegio
X_5 :	Clasificación por sexo	X_{12} :	Número de estud. promovidos
X_6 :	Tipo de establecimiento	X_{13} :	Número de estud. no promovidos
X_7 :	Propietario del edificio	X_{14} :	Número de estudiantes desertores

En la tabla XCV, observamos que entre la clasificación por sexo del colegio ($\mathbf{X_5}$) y los estudiantes reprobados de la institución ($\mathbf{X_{13}}$), tienen un coeficiente de correlación significativamente alto (**– 0.401**), de esto podemos concluir que posee una relación inversa entre la variables, es decir, los alumnos de los colegios fiscales masculinos tienden a tener más alumnos reprobados que en los femeninos y mixtos. La variable $\mathbf{X_5}$ también está correlacionada significativamente con $\mathbf{X_{10}}$ (número de aulas en los colegios), con un coeficiente de **–0.382,** esta tiene una dependencia inversa, es decir, los colegios fiscales masculinos tienen más aulas que las demás instituciones educativas fiscales en la provincia del Guayas.

Con respecto a las variables $\mathbf{X_{12}}$ (número de estudiantes aprobados) y $\mathbf{X_{13}}$ (número de estudiantes reprobados) están correlacionadas significativamente (**0.784**), tienen una dependencia directa entre ella, es decir, los alumnos aprobados y reprobados aumentan en la mismas proporción en los colegios fiscales de la provincia del Guayas.

4.4 ANÁLISIS DE INDEPENDENCIA DE LAS CARACTERÍSTICAS EN ESTUDIO UTILIZANDO TABLA DE CONTINGENCIA

En esta sección determinaremos que características interaccionan, lo que desaseamos probar si existe independencia entre dos variables aleatorias. En la tabla XCVI se muestra un resumen de todas las pruebas de independencias realizadas para las variables (valor p de la prueba), los valores resaltados en negrillas indican los casos en que se rechazó la hipótesis de independencia (H_O).

H_O : La variable X_i es independiente de la variable X_j **;** para $i \neq j$
Vs.
H_a : $\neg H_O$

Veamos algunos casos:

X_6 (Tipo de establecimiento) vs. X_5 (Sexo): El tipo del establecimiento del nivel medio depende de su sexo, la mayor cantidad de instituciones son mixtas e instituciones regulares (informática, ciencias humanísticas, etc) llegando a obtener 85.6% del total provincial.

X_7 (Propietario) vs. X_8 (Tenencia): El propietario del edificio del establecimiento del nivel medio depende de su tenencia, la mayor cantidad de instituciones son propiedad del estado y cuya tenencia es propia, cuyo porcentaje es del 72.6% del total provincial.

En la tabla XCVI, se muestran todos los valores p de la prueba de independencia utilizando tablas de contingencias, tales que los indicados con **negrillas** indican que hay dependencia entre la variable del i-ésimo renglón, con la j-ésima columna.

X_6 (Tipo de establecimiento) vs. X_9 (Tipo de construcción): El tipo de colegio fiscal es independiente del tipo de construcción del mismo, la mayoría de estas instituciones educativas del nivel medio de la provincia del Guayas son regulares y están construidas de hormigón armado, teniendo como porcentaje 71.8% del total provincial.

X_{12} (Estudiantes promovidos) vs. X_{13} (estudiantes reprobados)

H_O : El número de estudiantes promovidos es independiente de los estudiantes reprobados
Vs.
H_a : $\neg H_O$

Estadístico de prueba χ^2	**Grados de libertad**	**Valor p**
191.475	30	1.92E-25

En este caso, el valor p es suficientemente pequeño, nos permite concluir que existe evidencia estadística para rechazar la hipótesis nula (Ho), es decir, las variables X_{12} y X_{13} son dependientes. La tabla XCVII muestra la tabla de contingencia asociada a la prueba de independencia.

Tabla XCVII
Tabla de contingencia para la prueba de independencia entre las variables X_{12} vs X_{13}

			X_{13}: Estudiantes no promovidos							TOTAL
			[0 , 43)	[43 , 86)	[86 , 129)	[129, 172)	[172, 215)	[215, 258)	[258, 301)	$f._j$ $e._j$ %
X_{12}: Estudiantes promovidos	[0, 358)	f_{1j}	160	10	1	0	0	0	0	171
		e_{1j}	126.5	20.3	11.4	4.4	1.3	5.1	1.9	171
		%	59.5	3.7	0.4	0.0	0.0	0.0	0.0	63.6
	[358, 716)	f_{2j}	28	9	8	2	0	1	0	48
		e_{2j}	35.5	5.7	3.2	1.2	0.4	1.4	0.5	48
		%	10.4	3.3	3.0	0.7	0.0	0.4	0.0	17.8
	[716, 1074)	f_{3j}	8	7	7	2	1	2	1	28
		e_{3j}	20.7	3.3	1.9	0.7	0.2	0.8	0.3	28
		%	3.0	2.6	2.6	0.7	0.4	0.7	0.4	10.4
	[1074, 1432)	f_{4j}	0	3	1	2	0	1	0	7
		e_{4j}	5.2	0.8	0.5	0.2	0.1	0.2	0.1	7
		%	0.0	1.1	0.4	0.7	0.0	0.4	0.0	2.6
	[1432, 1790)	f_{5j}	2	1	1	1	1	3	1	10
		e_{5j}	7.4	1.2	0.7	0.3	0.1	0.3	0.1	10
		%	0.7	0.4	0.4	0.4	0.4	1.1	0.4	3.7
	[1790, 2148)	f_{6j}	1	2	0	0	0	1	1	5
		e_{6j}	3.7	0.6	0.3	0.1	0.0	0.1	0.1	5
		%	0.4	0.7	0.0	0.0	0.0	0.4	0.4	1.9
	TOTAL	$f_{i.}$	199	32	18	7	2	8	3	269
		$e_{i.}$	199	32	18	7	2	8	3	269
		%	74.0	11.9	6.7	2.6	0.7	3.0	1.1	100

Fuente: Dirección nacional de estudios (Base de dato año lect. 1998–99)

Analizando la tabla XCVII, observamos que en 59.5% de colegios fiscales de la provincia del Guayas, los estudiantes aprobaron entre [0 ; 358) y reprobaron [0; 43). Hay un pequeño porcentaje de instituciones educativas del nivel medio fiscales que tienen una gran cantidad de alumnos matriculados que representan 0.4% del total provincial, del cual aprobaron entre [1790;2148) y reprobaron entre [258 ; 301) estudiantes.