Análisis Multivariado de las variables que conforman las pruebas 240

4.1 Matriz de Correlación 247

4.2 Componentes principales 256

4.3 Tablas de contingencia 270

4.3.1 Variables Sección y Nota de Matemáticas 272

4.3.2 Variables Sección y Nota de Lenguaje 273

4.3.3 Variables Sexo y Nota de Matemáticas 275

4.3.4 Variables Sexo y Nota de Lenguaje 276

4.3.5 Variables edad y Nota de Matemáticas 277

4.3.6 Variables edad y Nota de Lenguaje 279

4.3.7 Variables Nota de matemáticas y Nota de Lenguaje 280

4.3.8 Variables edad y suma de enteros 282

4.4 Correlación canónica 285

4.4.1 Análisis de correlación canónica 289

**Capítulo 4**

# Análisis Multivariado de las variables que conforman las pruebas

En este capítulo se presentará un resumen de la estadística multivariada que se hará sobre las variables que conforman las pruebas, y las variables resultantes, como lo son la nota de lenguaje, la nota de matemáticas y el promedio general.

## 4.1 Definiciones

### Covarianza

La covarianza entre dos variables X1 y X2 mide la relación lineal, a mayor valor absoluto de la covarianza corresponde una mayor dependencia lineal entre X1 y X2, valores positivos indican que cuando X1 crece también X2 crece, valores negativos indican que cuando X1 crece X2 decrece.



En el caso del presente estudio, este valor puede ser estimado por:

siendo

i = 1,2,...,p

j = 1,2,...,p



### Coeficiente de correlación

Debido a la dificultad para utilizar la covarianza como una medida absoluta de la dependencia dado a que su valor depende de la escala de medición, no es sencillo determinar a simple vista si una covarianza en particular es grande o pequeño, para eliminar este problema se estandariza su valor.

donde: xi y xj son las desviaciones estándar de Xi y Xj respectivamente



El cual puede ser estimado por:



Se puede probar que el coeficiente de correlación se encuentra entre –1 ij 1. Entre mas cercano este el valor de ij hacia –1 o hacia 1 mayor será la relación lineal entre las variables.

Ahora procedemos a realizar la demostración de que el coeficiente de correlación está entre –1 y 1, para lo cual usaremos una propiedad del valor esperado de X y de la varianza:





Tenemos que probar que –1ij 1

La varianza de un número por la definición siempre es positiva, por lo tanto:



Para llegar a nuestra demostración necesitamos probar que la var(X+Y) = var(X) + var(Y) + 2cov(X,Y)



Se puede probar de manera similar que:

var(X-Y) = var(X) + var(Y) - 2cov(X,Y)

haciendo uso del primer resultado tenemos:





Tenemos la primera parte de la demostración, ahora procedemos a demostrar que ij 1. Partimos de:





Por lo tanto –1ij 1

### Matriz de datos multivariados

Denominaremos matriz de datos a la matriz **X** que tiene n filas y p columnas, el número n de filas corresponde al total de unidades investigadas y p es el número de variables (características de interés) que se investigan.

∈ Mnxp



Si no se toman en cuenta el número de observaciones realizadas y sólo se considera las p variables de interés lo que resulta se denomina vector aleatorio, así:

**XT** = [ X1 X2 ... Xp]

### Vector de medias

Sea **X** un vector p variado, se define a su vector de medias como:



El cual puede ser estimado por:



Donde: n es el tamaño de la muestra, **X** la matriz de datos y **1n**es un vector de 1 que pertenece a Rn, es decir:



### Matriz de covarianzas

Sea **X** un vector p variado, se define la matriz de covarianzas como:

**Σ** =



Estimada por:



donde:

**S** es el estimador de la matriz de covarianzas **Σ**, n el tamaño de la muestra, **X** la matriz de datos y **1n** es un vector de "unos" perteneciente a R**n**



Cuando se habla de la matriz de varianzas y covarianzas, se refiere a la matriz en cuya diagonal principal se localizan las varianzas de cada una de las variables de interés, y en la posición (i,j) se tiene la covarianza entre la i-ésima y la j-ésima variable, cabe recalcar además que i y j son los números que representan las filas y las columnas (respectivamente). **** es simétrica con dimensiones pxp.

### Matriz de correlación

Sea **S** el estimador de la matriz de covarianzas **Σ** de un vector aleatorio **X**  RP, defínase **V1/2**como la matriz de desviaciones estándar de **X**, como sigue:



es la desviación estándar de la variable aleatoria Xii

**ρ** =



**ρ** = **V-1/2ΣV-1/2**

lo que puede ser estimado por



donde: rij es el estimador del coeficiente de correlación ρij entre la variable Xi y Xj, esta matriz es simétrica y su dimensión es pxp.

## 4.2 Análisis de la matriz de correlación

Con la matriz de correlación (presentada en el Anexo 1) se puede analizar si existe algún tipo de relación lineal ya sea directa (es decir si Xi crece, también crece Xj) o inversa (es decir si Xi crece, decrece Xj), y los resultados son los siguientes:

* Hay una fuerte relación lineal entre la variable X51 (identificación de sujetos) y X52 (identificación de predicados), el coeficiente de correlación existente entre ambas variables es 0.978, es decir que a mayor cantidad de sujetos identificados, también habrá mayor cantidad de predicados identificados.

|  |  |  |
| --- | --- | --- |
|  | X51 | X52 |
| X51 | 1 | 0.978 |
| X52 | 0.978 | 1 |

* Entre las variables X38 (identificación de palabras agudas) y X39 (identificación de palabras graves), el coeficiente de correlación lineal es 0.724, lo cual nos indica que mientras más palabras agudas son identificadas más palabras graves también lo serán.

|  |  |  |
| --- | --- | --- |
|  | X38 | X39 |
| X38 | 1 | 0.724 |
| X39 | 0.724 | 1 |

* Las variables X39 (identificación de palabras graves) y X40 (identificación de palabras esdrújulas), tienen un coeficiente de correlación de 0.730

|  |  |  |
| --- | --- | --- |
|  | X39 | X40 |
| X39 | 1 | 0.730 |
| X40 | 0.730 | 1 |

* Los coeficientes de correlación anteriores son las más altos, pero también existen coeficientes un poco menores, pero igual importantes entre las variables X35 (identificación de diptongos) y X36 (identificación de triptongos) por ejemplo, ambas variables tienen un coeficiente de correlación de 0.696.

|  |  |  |
| --- | --- | --- |
|  | X35 | X36 |
| X35 | 1 | 0.696 |
| X36 | 0.696 | 1 |

* Hay una relación lineal fuerte entre las variables X55 (identificación de frases) y X56 (identificación de oraciones), el coeficiente de correlación entre ambas variables es 0.687, es decir que los estudiantes mientras más frases identifiquen entonces podrán identificar más oraciones.

|  |  |  |
| --- | --- | --- |
|  | X55 | X56 |
| X55 | 1 | 0.687 |
| X55 | 0.687 | 1 |

* Las variables X38 (identificación de palabras agudas) y X40 (identificación de palabras esdrújulas) tienen un coeficiente de correlación de 0.684.

|  |  |  |
| --- | --- | --- |
|  | X38 | X40 |
| X38 | 1 | 0.684 |
| X40 | 0.684 | 1 |

* Las variables X40 (identificación de palabras esdrújulas) y X41 (identificación de palabras sobreesdrújulas), tienen un coeficiente de correlación de 0.614.

|  |  |  |
| --- | --- | --- |
|  | X40 | X41 |
| X40 | 1 | 0.614 |
| X41 | 0.614 | 1 |

* Entre las variables X39 (identificación de palabras graves) y X41 (identificación de palabras sobreesdrújulas), el coeficiente de correlación es 0.586.

|  |  |  |
| --- | --- | --- |
|  | X39 | X41 |
| X39 | 1 | 0.586 |
| X41 | 0.586 | 1 |

* El coeficiente de correlación entre las variables X38 (identificación de palabras agudas) y X41 (identificación de palabras sobreesdrújulas) es de 0.570, esta relación lineal así como la de agudas con graves, de graves con esdrújulas, graves con sobreesdrújulas y la de esdrújulas y sobreesdrújulas, en realidad resultan ser un poco obvias, debido a que este tópico es el primero que se enseña en acentos.

|  |  |  |
| --- | --- | --- |
|  | X38 | X41 |
| X38 | 1 | 0.570 |
| X41 | 0.570 | 1 |

* Entre las variables X53 (reconocimiento de sustantivos) y X54 (reconocimiento de verbos) el coeficiente de correlación lineal es de 0.616, esta relación lineal también es un poco obvia, ya que al identificar los sujetos y predicados es sencillo identificar los núcleos del sujeto y los núcleos del predicado.

|  |  |  |
| --- | --- | --- |
|  | X53 | X54 |
| X53 | 1 | 0.616 |
| X54 | 0.616 | 1 |

* Las variables X5 (suma de quebrados) y X7 (resta de quebrados) tienen un coeficiente de correlación lineal de 0.598, lo cual quiere decir que mientras más sumas de quebrados realicen correctamente los estudiantes, más restas de quebrados realizarán.

|  |  |  |
| --- | --- | --- |
|  | X5 | X7 |
| X5 | 1 | 0.598 |
| X7 | 0.598 | 1 |

* Las variables X9 (multiplicación de quebrados) y X11 (división de quebrados) tienen un coeficiente de correlación de 0.533, lo cual quiere decir que mientras más multiplicaciones de quebrados realicen correctamente los estudiantes, más divisiones de quebrados realizarán correctamente.

|  |  |  |
| --- | --- | --- |
|  | X9 | X11 |
| X9 | 1 | 0.533 |
| X11 | 0.533 | 1 |

* Entre las variables X5 (suma de quebrados) y X34 (nota de matemáticas) el coeficiente de correlación es 0.512, lo cual quiere decir que la nota obtenida en matemáticas tiene una alta relación de lo que se obtenga sumando quebrados.

|  |  |  |
| --- | --- | --- |
|  | X5 | X34 |
| X5 | 1 | 0.512 |
| X34 | 0.512 | 1 |

* Entre las variables X7 (resta de quebrados) y X34 (nota de matemáticas) el coeficiente de correlación es 0.454, lo cual quiere decir que la nota obtenida en matemáticas tiene una alta relación de lo que se obtenga restando quebrados.

|  |  |  |
| --- | --- | --- |
|  | X7 | X34 |
| X7 | 1 | 0.454 |
| X34 | 0.454 | 1 |

* Entre las variables X11 (división de quebrados) y X34 (nota de matemáticas) el coeficiente de correlación es 0.478, lo cual quiere decir la nota obtenida en matemáticas tiene una alta relación de lo que se obtenga dividiendo quebrados.

|  |  |  |
| --- | --- | --- |
|  | X11 | X34 |
| X11 | 1 | 0.478 |
| X34 | 0.478 | 1 |

* Entre las variables X18 (proporcionalidad de interés) y X34 (nota de matemáticas) el coeficiente de correlación lineal es 0.544, lo cual quiere decir la nota obtenida en matemáticas tiene una alta relación de lo que se obtenga en la pregunta acerca de intereses.

|  |  |  |
| --- | --- | --- |
|  | X18 | X34 |
| X18 | 1 | 0.544 |
| X34 | 0.544 | 1 |

* Entre las variables X19 (regla de tres simple) y X34 (nota de matemáticas) el coeficiente de correlación lineal es 0.457, lo cual quiere decir la nota obtenida en matemáticas tiene una alta relación de lo que se obtenga en la pregunta de regla de tres simple.

|  |  |  |
| --- | --- | --- |
|  | X19 | X34 |
| X19 | 1 | 0.457 |
| X34 | 0.457 | 1 |

* El coeficiente de correlación lineal de 0.476 entre las variables X25 (área del triángulo) y X26 (área del círculo), nos indica que mientras mejor realicen el problema del área del triángulo, los estudiantes, entonces mejor realizarán el problema del área del círculo.

|  |  |  |
| --- | --- | --- |
|  | X25 | X26 |
| X25 | 1 | 0.476 |
| X26 | 0.476 | 1 |

* Entre las variables X28 (factorización de trinomio cuadrado perfecto) y X29 (factorización del trinomio de la forma x2+bx+c) el coeficiente de correlación lineal de 0.521, nos indica que mientras mejor factoricen un caso, tanto mejor factorizarán el otro caso.

|  |  |  |
| --- | --- | --- |
|  | X28 | X29 |
| X28 | 1 | 0.521 |
| X29 | 0.521 | 1 |

* La variable X34 (nota obtenida en la prueba de matemáticas) tiene un coeficiente de correlación lineal de 0.512 con la variable X24 (perímetro del cuadrado), lo que quiere decir que la nota de matemáticas depende linealmente del puntaje que se obtenga en el tema del perímetro del cuadrado.

|  |  |  |
| --- | --- | --- |
|  | X24 | X34 |
| X24 | 1 | 0.512 |
| X34 | 0.512 | 1 |

* Las variables X43 (sinónimos) y X44 (antónimos) tienen un coeficiente de correlación de 0.585, lo cual nos indica que mientras más sinónimos se identifiquen los alumnos del décimo año de educación básica, más antónimos serán identificados.

|  |  |  |
| --- | --- | --- |
|  | X43 | X44 |
| X43 | 1 | 0.585 |
| X44 | 0.585 | 1 |

* Entre las variables X45 (palabras aumentativas) y X46 (palabras diminutivas) el coeficiente de correlación lineal es 0.555, lo cual nos indica que mientras más palabras aumentativas se identifiquen, se identificarán más palabras diminutivas.

|  |  |  |
| --- | --- | --- |
|  | X45 | X46 |
| X45 | 1 | 0.555 |
| X46 | 0.555 | 1 |

* Entre las variables X46 (palabras diminutivas) y X48 (palabras simples) el coeficiente de correlación lineal es 0.503, lo cual nos indica que mientras más palabras diminutivas se identifiquen, se identificarán más palabras simples.

|  |  |  |
| --- | --- | --- |
|  | X46 | X48 |
| X46 | 1 | 0.503 |
| X48 | 0.503 | 1 |

* Cabe recalcar que se esperaba fuertes correlaciones lineales entre las operaciones básicas en la prueba de matemáticas, es decir entre sumas, restas, multiplicaciones y divisiones de enteros y quebrados, pero estas no se dieron entre todas (es decir se dio solo entre unas cuantas como las mencionadas en la parte superior).
* La variable X62 (promedio general) está correlacionada linealmente con las variables X34 (nota obtenida en matemáticas) y X61 (nota obtenida en lenguaje) con coeficientes de 0.681 y 0.818 respectivamente.

|  |  |  |  |
| --- | --- | --- | --- |
|  | X34 | X61 | X62 |
| X34 | 1 | 0.143 | 0.681 |
| X61 | 0.143 | 1 | 0.818 |
| X62 | 0.681 | 0.818 | 1 |

* Otra relación lineal fuerte que se esperaba era entre las variables X32 (probabilidad) y X33 (estadística), pero tampoco se dio, la correlación entre ambas es de 0.025. Esto posiblemente se debe a que no se les imparte la materia dentro del programa de estudios como se supone debería hacérselo.

|  |  |  |
| --- | --- | --- |
|  | X32 | X33 |
| X32 | 1 | 0.025 |
| X33 | 0.025 | 1 |

## 4.3 Componentes principales

A un análisis de componentes principales le concierne explicar las varianzas y covarianzas de un conjunto de datos a través de unas pocas combinaciones lineales de las mismas. Sus objetivos generales son la reducción de datos y la interpretación. Un análisis de componentes principales a menudo revela relaciones que no se sospechaba y por lo tanto permite una interpretación que ordinariamente no habría sido posible hacer.

Algebraicamente, las componentes principales son una combinación de las p variables aleatorias X1,X2,...,Xp observables. Geométricamente, estas combinaciones lineales representan la selección de un sistema de coordenadas obtenido rotando el sistema original con X1,X2,...,Xp como los ejes. Las componentes principales dependen de la matriz de varianzas y covarianzas **Σ**, estimada por la matriz **S** (o de la matriz de correlación , estimada por **R**) de X1,X2,...,Xp.

Sea el vector aleatorio **XT**=[X1,X2,...,Xp] que tiene la matriz de varianzas y covarianzas **Σ** (en el caso del presente estudio se usará la matriz de estimadores de varianzas y covarianzas **S**) con valores propios λ1λ2... λp0, y considere las combinaciones lineales:



En forma general: Yi = **aiTX,** donde **aiT**= **eiT**, siendo **ei** el vector propio de ****.

Con:

*Var(Yi)* = **eTiΣei** = λi para *i=1,2,...,p*

*Cov(Yi, Yk)* = **eTiΣek**= 0 para *i≠k*

Se puede observar que las componentes principales están ordenadas de tal forma que entre menor sea el índice que tenga la componente mayor será la varianza de la misma, es decir que:

Var(Yi) > Var(Yk), para todo i<k

La covarianza de Yi con Yk, para i ≠k es cero, ya que son independientes.

Sea el vector aleatorio **XT**=[X1,X2,...,Xp] que tiene la matriz de varianzas y covarianzas **Σ** (estimada por la matriz de estimadores de varianzas y covarianzas **S**) con valores propios λ1λ2... λp0, y Y1=**eT1X**, Y2=**eT2X,** ..., Yp=**eTpX,** las componentes principales, entonces:



Se puede probar que σ11+σ22+...+σpp = tr(****). Con A = ****, se puede escribir **=PΛPT** donde Λ es la matriz diagonal de valores propios y P=[**e1,e2,...,ep**] así que PPT=PTP=I. Usando este resultado se tiene:

*tr(******) = tr*(PΛPT) *= tr*(ΛPPT)*= tr*(Λ) *= λ1+λ2+...+λp*

Así,



Y consecuentemente, la proporción de la varianza total explicada por la k-ésima componente es:



Aplicando componentes principales a la matriz de datos originales, compuesta por todas las variables utilizadas para este estudio, obtenemos los siguientes resultados con la ayuda del software estadístico SPSS. Se obtuvieron 61 componentes principales, los 5 primeros componentes explican el 84.451% y se los presenta a continuación:

**Tabla LXXV**

**Porcentajes de explicación de las componentes principales obtenidos a partir de la matriz de datos originales.**

|  |  |  |  |
| --- | --- | --- | --- |
| **Componente** | **Varianza** | **Porcentaje de explicación** | **Porcentaje acumulado** |
| 1 | 181.936 | 49.933 | 49.933 |
| 2 | 99.936 | 27.428 | 77.361 |
| 3 | 12.325 | 3.383 | 80.744 |
| 4 | 7.129 | 1.957 | 82.700 |
| 5 | 6.378 | 1.751 | 84.451 |

**Primera componente principal**



**Segunda componente principal**

**Tercera componente principal**



**Cuarta componente principal**



**Quinta componente principal**



Debido a que las variables que utilizamos (en los datos originales) no están en la misma escala, surge un problema ya que las variables que están en escalas mayores van a absorber los pesos más significativos como ocurre en las componentes principales calculadas con la matriz de datos, 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 que no es mas que: a cada variable se le resta la media y se divide para la desviación estándar.

*i=1,2,...,p*



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

**Z** =(**V1/2**)**-1**(**X** - **μ**)

Siendo **Z**  RP es el vector aleatorio p variado estandarizado, **X** es el vector aleatorio p variado original, la matriz **V1/2**y el vector de medias definidos al iniciar el capítulo.

Las componentes principales de **Z**  RP, que es el vector p variado estandarizado las podemos obtener de los vectores propios de la matriz de correlación **ρ** (estimado por **R**) asociada a **X**. Obteniendo la i-ésima componente principal para la matriz de datos estandarizada de la siguiente forma:

Yi *=* ***eTiZ*** *i=1,2,...p*

Ahora procedemos a calcular los coeficientes de las componentes principales de la matriz de datos estandarizada, con la ayuda de la matriz de correlación de **X** y la cantidad se elevó de 5 a 20, pero con un porcentaje de explicación menor, solo del 63.63%, las componentes obtenidas se presentan en el ANEXO 7, y además se calcularon las componentes principales rotándolos con varimax, las componentes se presentan en el ANEXO 8, así mismo la cantidad de componentes es 20 y una explicación de 63.63%

Después de analizar las componentes obtenidas se ha decidido trabajar con las componentes obtenidas por la matriz de datos estandarizados y rotadas ya que estas tienen a todas las variables en una misma escala, pero estas explican un 63.63%

Los porcentajes de explicación de cada componente se presentan a continuación:

**Tabla LXXVI**

**Porcentajes de explicación de los componentes principales obtenidos a partir de la matriz de datos estandarizados**

|  |  |  |  |
| --- | --- | --- | --- |
| **Componente** | **Varianza** | **Porcentaje de explicación** | **Porcentaje acumulado** |
| 1 | 6.670 | 10.934 | 10.934 |
| 2 | 4.928 | 8.079 | 19.013 |
| 3 | 2.458 | 4.029 | 23.042 |
| 4 | 2.453 | 4.021 | 27.063 |
| 5 | 2.002 | 3.283 | 30.346 |
| 6 | 1.923 | 3.152 | 33.498 |
| 7 | 1.789 | 2.932 | 36.430 |
| 8 | 1.655 | 2.714 | 39.144 |
| **Tabla LXXVI (Continuación...)** | | | |
| **Componente** | **Varianza** | **Porcentaje de explicación** | **Porcentaje acumulado** |
| 9 | 1.519 | 2.491 | 41.634 |
| 10 | 1.492 | 2.446 | 44.080 |
| 11 | 1.397 | 2.289 | 46.369 |
| 12 | 1.367 | 2.241 | 48.610 |
| 13 | 1.282 | 2.102 | 50.712 |
| 14 | 1.230 | 2.016 | 52.728 |
| 15 | 1.185 | 1.943 | 54.671 |
| 16 | 1.142 | 1.872 | 56.543 |
| 17 | 1.135 | 1.861 | 58.404 |
| 18 | 1.091 | 1.788 | 60.192 |
| 19 | 1.069 | 1.753 | 61.945 |
| 20 | 1.033 | 1.694 | 63.639 |

De acuerdo a los pesos más significativos de cada componente se procedió a darles nombre.

* La primera componente tiene sus pesos más importantes en las variables Z37 (palabras agudas), Z38 (palabras graves), Z39 (palabras esdrújulas), Z40 (palabras sobreesdrújulas), por lo tanto esta componente se llamará "**acentos en las palabras**".
* La segunda componente tiene sus pesos más importantes en las variables Z5 (suma de quebrados), Z7 (resta de quebrados), Z9 (multiplicación de quebrados), Z11 (división de quebrados) y por lo tanto esta componente se llamará "**operaciones con quebrados**".
* La tercera componente tiene sus pesos más importantes en las variables Z44 (palabras aumentativas), Z45 (palabras diminutivas), Z46 (palabras despectivas), Z47 (palabras simples) y por lo tanto esta componente se llamará "**tipos de palabras**"
* La cuarta componente tiene sus pesos más importantes en las variables Z18 (proporcionalidad interés), y Z19 (regla de tres simple) y por lo tanto esta componente se va a llamar "**proporcionalidad**".
* La quinta componente tiene sus pesos más importantes en las variables Z50 (número de sujetos identificados correctamente) y Z51 (número de predicados identificados correctamente), así que esta componente se llamará "**sujetos y predicados**".
* La sexta componente se llamará "**jornada en que funciona el colegio**" debido a que su peso más importante en la variable Z1 (jornada).
* La séptima componente principal se llamará "**factorización**" debido a que los pesos más importantes que tiene son en la variable Z28 (trinomio cuadrado perfecto y diferencia de cuadrados) y Z29 (trinomio de la forma x2+bx+c).
* La octava componente principal se la nombrará "**frases y oraciones**" debido a que las dos variables con más peso en la misma son la Z54 (número de frases correctamente identificadas) y Z55 (número de oraciones correctamente identificadas).
* La novena componente tiene sus pesos más importantes en las variables Z52 (número de núcleos del sujeto correctamente identificados) y Z53 (número de núcleos del predicado correctamente identificados), por lo tanto se procederá a llamarla "**núcleos**".
* La décima componente principal será llamada "**sinónimos y antónimos**" dado que las variables con más peso en la misma son: Z42 (número de sinónimos correctos) y Z43 (número de antónimos correctos).
* La décima primera componente tiene sus pesos más importantes en las variables Z25 (área del triángulo) y Z26 (área del círculo), por lo tanto es se llamará "**áreas**".
* La décima segunda componente se llamará "**diptongos y triptongos**" dado que las variables Z34 (número de diptongos identificados) y Z35 (número de triptongos identificados) son las que más pesos tienen.
* La décima tercera componente será llamada "**sexo del estudiante**" debido a que es la variable Z2 (sexo del estudiante) la que más peso tiene.
* La décima cuarta componente principal será llamada "**conjunto de números**" dado a que la variable Z12 (conjunto de números es la que más peso tiene).
* La décima quinta componente tiene su peso más importante en la variable Z41 (oraciones con correcta semántica) por lo tanto será llamada "**semántica**".
* La décima sexta componente será llamada "**suma de enteros**" debido a que la variable Z4 (suma de enteros) es la que más peso tiene en la misma.
* La décima séptima componente tiene su peso más importante en la variable Z48 (número de palabras con correcto significado) y por lo tanto se la llamará "**vocabulario**".
* La décima octava componente principal tiene su peso más importante en la variable Z6 (resta de enteros), por lo tanto esta será llamada "**resta de enteros**".
* La décima novena componente será llamada "**funciones**" dado que la variable Z23 (número de funciones correctamente resueltas) es la que más peso tiene.
* La vigésima componente tiene su peso más importante en la variable Z13 (valor absoluto) y por lo tanto se la llamará "**valor absoluto**".

Para verificar la validez del método de componentes se efectúa la prueba de Barlett, la cual nos indicará si la matriz es factorizable o no, si la matriz es factorizable entonces los resultados obtenidos en componentes principales son válidos, y caso contrario no son válidos.

La misma se basa en el estadístico ji cuadrado y el contraste de hipótesis planteado es:

H0: La matriz es factorizable

vs

H1: la matriz no es factorizable

Ji-cuadrado 35190.815

df 1653

Valor p .000

Aplicando esta prueba se obtuvo un valor p de 0.000 lo cual nos indica que se rechaza la hipótesis nula, es decir la matriz es no factorizable, por lo tanto los resultados obtenidos en la sección anterior no son recomendables para utilización posterior.

## 4.4 Tablas de contingencia

La tabla de contingencia es un arreglo matricial de r filas y c columnas, donde r es el número de niveles del factor 1 o de la variable Xi y c el número de niveles del factor 2 o de la variable Xj, cada variable debe tener al menos dos niveles los cuales deben ser exhaustivos y mutuamente excluyentes. Las tablas de contingencia permiten determinar la dependencia o independencia de dos variable aleatorias Xi y Xj (llamados factores). A continuación se muestra la forma de una tabla de contingencia:

###### TABLA DE CONTINGENCIA

|  |  |  |  |
| --- | --- | --- | --- |
| Factor 1 | | | |
| Factor 2 | Nivel 1 | **Nivel 2** |  | **Nivel c** |  |
| **Nivel 1** | X11  E11 | X12  E12 |  | X1c  E1c | X1. |
| **Nivel 2** | X21  E21 | X22  E22 |  | X2c  E2c | X2. |
|  |  |  |  |  |  |
|  |  |  |  |  |
| **Nivel r** | Xr1  Er1 | Xr2  Er2 |  | Xrc  Erc | 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.

El contraste de hipótesis planteado es:

H0: Los factores son independientes

vs

H1: no es verdad H0

en base a:

donde



Se puede probar que 2 tiene una distribución 2(r-1)(c-1), esto es ji cuadrado con (r-1)(c-1) grados de libertad.

χ2 ~ χ2α (r-1)(c-1)

bajo estas condiciones rechace H0 a favor de H1 con (1-)100% de confianza si:

χ2 > χ2α (r-1)(c-1)

El análisis precedente se aplicó a las variables que se suponía podían tener algún tipo de dependencia lineal o no lineal, las variables que se analizaron se presentan a continuación, cada una con su respectiva tabla de contingencia.

Al realizar las tablas de contingencia en el software especializado (SPSS o SYSTAT) se advertía que el resultado de ciertas tablas de contingencia era inconsistente (es decir que podía ser erróneo) por cuanto existían en las mismas Xij < 5. Por lo tanto se procedió a disminuir el número de niveles con el propósito de eliminar este problema, los resultados se presentan a continuación

### 4.4.1 Variables: Jornada y Nota de Matemáticas

Factor 1:

1. Notas de Matemáticas desde 2 a 21
2. Notas de Matemáticas desde 22 a 60

Factor 2:

x: Sección Matutina

y: Sección Vespertina

z: Sección Nocturna

El contraste de hipótesis planteado es:

H0: la nota de matemáticas es independiente a la jornada

vs

H1: la nota de matemáticas depende de la jornada

**Tabla LXXVII**

**Tabla de contingencia variables Nota de matemáticas - jornada**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **a** | **b** | **Total** |
| **x** | 192  229.67 | 260  222.32 | 452 |
| **y** | 268  262.19 | 248  253.8 | 516 |
| **z** | 102  70.12 | 36  67.87 | 138 |
|  | 562 | 544 | 1106 |

χ2=42.28

Valor p=0.000 (2 grados de libertad)

Existe evidencia estadística para rechazar la hipótesis nula, es decir que existe en realidad una dependencia entre la jornada en la que funcione el colegio y la nota de matemáticas.

### 4.4.2 Variables: Jornada y Nota de Lenguaje

Factor 1:

1. Notas de Lenguaje desde 15 a 66
2. Notas de Matemáticas desde 67 a 88

Factor 2:

x: Sección Matutina

y: Sección Vespertina

z: Sección Nocturna

El contraste de hipótesis planteado es:

H0: la nota de lenguaje es independiente a la jornada

vs

H1: la nota de lenguaje depende de la jornada

**Tabla LXXVIII**

**Tabla de contingencia variables Nota de lenguaje - jornada**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **a** | **b** | **Total** |
| **x** | 386  358.8 | 66  93.17 | 452 |
| **y** | 382  409.62 | 134  106.37 | 516 |
| **z** | 110  109.55 | 28  28.44 | 138 |
|  | 878 | 228 | 1106 |

χ2=19.03

Valor p=0.00007 (2 grados de libertad)

Existe evidencia estadística para rechazar la hipótesis nula, es decir que existe en realidad dependencia entre la jornada en la que funcione el colegio y la nota de lenguaje.

### 4.4.3 Variables: Sexo y Nota de Matemáticas

Factor 1:

1. Notas de Matemáticas desde 2 a 21
2. Notas de Matemáticas desde 22 a 60

Factor 2:

x: Sexo Femenino

y: Sexo Masculino

El contraste de hipótesis planteado es:

H0: la nota de matemáticas es independiente del sexo del estudiante

vs

H1: la nota de matemáticas depende del sexo del estudiante

**Tabla LXXIX**

**Tabla de contingencia variables Nota de matemáticas - sexo**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **a** | **B** | **Total** |
| **x** | 263  290.65 | 310  282.34 | 573 |
| **y** | 297  269.34 | 234  261.65 | 531 |
|  | 560 | 544 | 1104 |

χ2=11.10

Valor p=0.0008 (1 grado de libertad)

Existe evidencia estadística para rechazar la hipótesis nula, es decir que existe en realidad dependencia entre el sexo del estudiante y la nota de matemáticas, en la tabla LXXVIII se aprecia que las notas más altas en matemáticas son obtenidas por el sexo femenino, mientras que las notas más bajas son obtenidas por el sexo masculino.

### 4.4.4 Variables: Sexo y Nota de Lenguaje

Factor 1:

1. Notas de Lenguaje de 15 a 66
2. Notas de Lenguaje de 67 a 88

Factor 2:

x: Sexo Femenino

y: Sexo Masculino

El contraste de hipótesis planteado es:

H0: la nota de lenguaje es independiente del sexo del estudiante

vs

H1: la nota de lenguaje depende del sexo del estudiante

**Tabla LXXX**

**Tabla de contingencia variables Nota de lenguaje - sexo**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **a** | **b** | **Total** |
| **x** | 433  454.66 | 140  118.33 | 573 |
| **y** | 443  421.33 | 88  109.66 | 531 |
|  | 876 | 228 | 1104 |

χ2=10.391

Valor p=0.0012 (1 grado de libertad)

Existe evidencia estadística para rechazar la hipótesis nula, es decir que existe en realidad una dependencia entre el sexo del estudiante y la nota de lenguaje, en la tabla LXXIX se aprecia que las notas en lenguaje tanto el sexo femenino como el sexo masculino tienen notas que oscilan entre 15 y 66.

### 4.4.5 Variables: edad y Nota de Matemáticas

Factor 1:

1. Notas de Matemáticas de 2 a 21
2. Notas de Matemáticas de 22 a 60

Factor 2:

x: edades entre los 12 y 14 años

y: edades entre los 15 y 18 años

El contraste de hipótesis planteado es:

H0: la nota de matemáticas es independiente de la edad del estudiante

vs

H1: la nota de matemáticas depende de la edad del estudiante

**Tabla LXXXI**

**Tabla de contingencia variables Nota de matemáticas - edad**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **a** | **b** | **Total** |
| **x** | 320  331.60 | 330  318.39 | 650 |
| **y** | 232  220.39 | 200  211.60 | 432 |
|  | 552 | 530 | 1082 |

χ2=2.07

Valor p=0.149 (1 grado de libertad)

Existe evidencia estadística para no rechazar la hipótesis nula, es decir que no existe en realidad dependencia entre la edad del estudiante y la nota de matemáticas, pero al analizar la variable edad se puede apreciar que la variación entre las edades no es mucha, por lo tanto se pone en duda esta conclusión obtenida, ya que la mayoría de alumnos tienen entre 14 y 15 años (la edad promedio que hay en el décimo año de educación básica es de 14.44).

### 4.4.6 Variables: edad y Nota de Lenguaje

Factor 1:

1. Notas de Lenguaje de 15 a 66
2. Notas de Lenguaje de 67 a 88

Factor 2:

x: edades entre los 12 y 14 años

y: edades entre los 15 y 18 años

El contraste de hipótesis planteado es:

H0: la nota de lenguaje es independiente a la edad del estudiante

vs

H1: la nota de lenguaje depende de la edad del estudiante

**Tabla LXXXII**

**Tabla de contingencia variables Nota de lenguaje - edad**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **a** | **b** | **Total** |
| **x** | 494  515.43 | 156  134.56 | 650 |
| **y** | 364  342.56 | 68  89.43 | 432 |
|  | 852 | 224 | 1082 |

χ2=10.78

Valor p=0.00102 (1 grado de libertad)

Existe evidencia estadística para rechazar la hipótesis nula, es decir que existe en realidad una dependencia entre la edad del estudiante y la nota de lenguaje.

### 4.4.7 Variables: Nota de matemáticas y Nota de Lenguaje

Factor 1:

1. Notas de Matemáticas de 2 a 21
2. Notas de Matemáticas de 22 a 60

Factor 2:

x: Notas de Lenguaje de 15 a 66

y: Notas de Lenguaje de 67 a 88

El contraste de hipótesis planteado es:

H0: la nota de lenguaje es independiente de la nota de matemáticas

vs

H1: la nota de lenguaje depende de la nota de matemáticas

**Tabla LXXXIII**

**Tabla de contingencia variables Nota de matemáticas - Nota de lenguaje**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **a** | **b** | **Total** |
| **x** | 468  446.14 | 94  115.85 | 562 |
| **y** | 410  431.85 | 134  112.14 | 544 |
|  | 878 | 228 | 1106 |

χ2=10.558

Valor p=0.00115 (1 grado de libertad)

Existe evidencia estadística para rechazar la hipótesis nula, es decir que existe en realidad una dependencia entre la nota de matemáticas y la nota de lenguaje, en la tabla LXXXII se aprecia que los estudiantes del décimo año de educación básica que obtienen las menores calificaciones en matemáticas, también obtienen las menores calificaciones en lenguaje.

### 4.4.8 Variables: edad y suma de enteros

Factor 1:

1. Suma hasta las decenas
2. Suma hasta las centenas

Factor 2:

x: edades entre los 12 y 14

y: edades entre los 15 y 18

El contraste de hipótesis planteado es:

H0: la capacidad para sumar enteros no depende de la edad del estudiante.

vs

H1: la capacidad para sumar enteros depende de la edad del estudiante.

**Tabla LXXXIV**

**Tabla de contingencia variables edad - suma de enteros**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **a** | **b** | **Total** |
| **x** | 74  69.68 | 576  580.31 | 650 |
| **y** | 42  46.314 | 390  385.6 | 432 |
|  | 878 | 966 | 1082 |

χ2=0.7493

Valor p=0.3866

Existe evidencia estadística para aceptar la hipótesis nula, es decir que no existe en realidad una dependencia entre la suma de enteros y la edad, ya que sin importar la edad saben realizar correctamente la operación suma de enteros, como se aprecia en la tabla LXXXIII.

Dado que las correlaciones lineales de las operaciones básicas fueron bajas, se procedió a hacer tablas de contingencia para poder precisar si las mismas son independientes o no.

Tabla LXXXV

**Valores p obtenidos de las tablas de contingencia**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Factor 1** | **Factor 2** | **Correlación lineal** | **Valor p** | **Independencia** |
| Suma de enteros | Suma de quebrados | 0.121 | 0.000 | Dependientes |
|  | Resta de enteros | 0.138 | 0.000 | Dependientes |
|  | Resta de quebrados | 0.106 | 0.000 | Dependientes |
|  | Multiplicación de enteros | 0.099 | 0.007 | Dependientes |
|  | Multiplicación de quebrados | 0.032 | 0.257 | **Independientes** |
|  | División de enteros | 0.127 | 0.001 | Dependientes |
|  | División de quebrados | 0.089 | 0.005 | Dependientes |
| Suma de quebrados | Resta de enteros | 0.143 | 0.000 | Dependientes |
|  | Resta de quebrados | **0.603** | 0.000 | Dependientes |
|  | Multiplicación de enteros | 0.078 | 0.012 | Dependientes |
| **Tabla LXXXV (continuación)** | | | | |
| **Factor 1** | **Factor 2** | **Correlación lineal** | **Valor p** | **Independencia** |
| Suma de quebrados | Multiplicación de quebrados | 0.275 | 0.000 | Dependientes |
|  | División de enteros | 0.171 | 0.000 | Dependientes |
|  | División de quebrados | 0.339 | 0.000 | Dependientes |
| Resta de enteros | Resta de quebrados | 0.105 | 0.001 | Dependientes |
|  | Multiplicación de enteros | 0.061 | 0.319 | **Independientes** |
|  | Multiplicación de quebrados | 0.100 | 0.002 | Dependientes |
|  | División de enteros | 0.116 | 0.000 | Dependientes |
|  | División de quebrados | 0.125 | 0.000 | Dependientes |
| Resta de quebrados | Multiplicación de enteros | 0.057 | 0.090 | Dependientes |
|  | Multiplicación de quebrados | 0.270 | 0.000 | Dependientes |
|  | División de enteros | 0.175 | 0.000 | Dependientes |
|  | División de quebrados | 0.321 | 0.000 | Dependientes |
| Multiplicación de enteros | Multiplicación de quebrados | 0.132 | 0.000 | Dependientes |
|  | División de enteros | 0.107 | 0.000 | Dependientes |
|  | División de quebrados | 0.159 | 0.000 | Dependientes |
| Multiplicación de quebrados | División de enteros | 0.100 | 0.001 | Dependientes |
|  | División de quebrados | **0.529** | 0.000 | Dependientes |
| División de enteros | División de quebrados | 0.180 | 0.000 | Dependientes |

Se puede apreciar que a pesar de que los coeficientes de correlación lineal son bajos, los conocimientos en las operaciones básicas están relacionados pero no de una forma lineal, excepto la suma de enteros con la multiplicación de quebrados y la resta de enteros con la multiplicación de enteros.

## 4.5 Correlación canónica

La correlación canónica mide la fuerza de la asociación entre dos conjuntos de variables. El primer grupo, de p variables, es representado por el vector aleatorio **X**(1) ∈ RP, el segundo grupo de, de q variables, es representado por el vector aleatorio **X**(2) ∈ Rq. Se supone que p < q.

Para los vectores aleatorios **X**(1) y **X**(2), sean

E(**X**(1)) = **μ**(1)

E(**X**(2)) = **μ**(2)

Cov(**X**(1)) = **Σ**11

Cov(**X**(2)) = **Σ**22

Cov(**X**(1), **X**(2)) = **Σ**12 = **ΣT**21

En ciertas operaciones es conveniente particionar al vector aleatorio **X** en **X**(1) y **X**(2), así:



El vector aleatorio **X** tiene un vector de medias:



Que puede ser estimado por



Y una matriz de covarianzas

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Σ**11 | **Σ**12 |  |
| **Σ** = |  |  | ∈ M(p+q)(p+q) | |
|  | **Σ**21 | **Σ**22 |  |

Que puede ser estimada por



donde



La correlación canónica presenta las asociaciones entre los vectores **X**(1) y **X**(2) en términos de unas pocas escogidas covarianzas (o correlaciones) en lugar de las pq covarianzas en **Σ12** (estimada por **S12**).

Sean

U = **aTX(1)**

V = **bTX(2)**

Para un par de vectores coeficientes a y b, obtenemos

*Var*(U) =**aT***Cov*(**X(1)**)**a** = **aT11a**

*Var*(V) =**bT***Cov* (**X(2)**)**b** = **bT22b**

*Cov*(U,V) =**aT***Cov*(**X(1), X(2)**)**b** = **aT12b**

Se buscará vectores de coeficientes tal que



Para obtener U1, V1 se procede así



Donde



Son los valores propios de



Y los **e1,e2,...,en** son los vectores propios de px1 asociados. Las cantidades



Son además los p valores propios más grandes de la matriz



Con los correspondientes vectores propios (de qx1) **f1,f2,...,fp**.

Se tiene además que:

*Var*(U) = *Var* (V) = 1

*Cov*(Uk,Ul) = *Cov*(Ul,Uk) = 0 para *k≠l*

*Cov*(Vk,Vl) = *Cov*(Vl,Vk) = 0 para *k≠l*

*Cov*(Uk,Vl) = *Cov*(Ul,Vk) = 0 para *k≠l*

Se procedió a calcular las correlaciones canónicas entre los vectores **X**(1) ∈ RP: variables que representan a las preguntas de la prueba de lenguaje y **X**(2) ∈ Rq: variables que representan a las preguntas de la prueba de matemáticas, donde p=26 y q=30.

### 4.5.1 Análisis de correlación canónica

Para obtener las correlaciones canónicas entre los vectores **X(1)** y **X(2)** se utilizó el software estadístico SPSS (Social Purpose Statistical System), llamando a una función "CANCORR" y definiendo los set1 y set2 (**X(1)** y **X(2)** respectivamente)**.** Haciendo el análisis de los resultados obtenidos se observa lo siguiente, el grupo de variables de Lenguaje es de tamaño 26 (p=26) y las variables de matemáticas son 30 (q=30) por lo tanto se obtuvieron 26 pares de variables canónicas y las correlaciones canónicas que se obtuvieron entre los pares de variables son los siguientes:

**Tabla LXXXVI**

**Correlaciones Canónicas obtenidas entre pares de variables**

|  |  |  |  |
| --- | --- | --- | --- |
| **Par** | **Correlación** | **Par** | **Correlación** |
| 1 | 0.479 | 14 | 0.216 |
| 2 | 0.452 | 15 | 0.204 |
| 3 | 0.418 | 16 | 0.189 |
| 4 | 0.392 | 17 | 0.170 |
| 5 | 0.375 | 18 | 0.167 |
| 6 | 0.343 | 19 | 0.145 |
| 7 | 0.342 | 20 | 0.109 |
| 8 | 0.314 | 21 | 0.102 |
| 9 | 0.301 | 22 | 0.072 |
| 10 | 0.278 | 23 | 0.063 |
| 11 | 0.269 | 24 | 0.062 |
| 12 | 0.257 | 25 | 0.054 |
| 13 | 0.219 | 26 | 0.037 |

Se puede apreciar en la tabla que la correlación entre el primer par de variables U1, V1 es de 0.479 lo cual indica que hay una correlación baja entre las variables de Lenguaje y las variables de Matemáticas. Se analizará los cinco primeros pares de variables canónicas.

A continuación se muestra los coeficientes de los 5 pares de variables canónicas, primero se mostrarán las de Lenguaje (U1) y posteriormente las de Matemáticas (V1), luego se procederá a indicar entre que variables está las más fuerte correlación.

**Primer par de variables canónicas**



Cabe recalcar que la Var(U1)=1, Var(V1)=1, Cov(U1,V1)=0.479, es decir que la correlación entre el primer par de variables no es fuerte y principalmente se concentra en las siguientes variables:

**Tabla LXXXVII**

**Variables que más aportan al primer par de variables canónicas**

|  |  |
| --- | --- |
| **Lenguaje** | **Matemáticas** |
| X35(identificación de diptongos) | X15 (Potenciación y radicación) |
| X39 (palabras graves) | X18 (proporción interés) |
| X56 (identificación de oraciones) | X21 (propiedades de conjuntos) |
| X58 (sintaxis) | X22 (lógica matemática) |
| X59 (tildes) | X23 (funciones) |

Las variables de la tabla anterior son las variables que más aportan a las variables U1 y V1 respectivamente, como ya se indicó la correlación entre lenguaje y matemáticas no es fuerte por lo tanto se concluye que no están altamente correlacionadas.

**Segundo par de variables canónicas**



Como se apreció en la tabla LXXXV la correlación entre U2, V2 (Cov(U2,V2)=0.452) y además la Var(U2)=1, Var(V2)=1, lo cual nos indica que la correlación entre el segundo par de variables canónicas no es fuerte, y principalmente se concentra en las siguientes variables:

**Tabla LXXXVIII**

**Variables que más aportan al segundo par de variables canónicas**

|  |  |
| --- | --- |
| **Lenguaje** | **Matemáticas** |
| X38 (palabras agudas) | X5 (suma de quebrados) |
| X41 (palabras sobreesdrújulas) | X14 (relaciones de orden) |

**Tercer par de variables canónicas**



Cabe recalcar que la Var(U3)=1, Var(V3)=1, Cov(U3,V3)=0.418, es decir que la correlación entre el tercer par de variables no es fuerte y principalmente se concentra en las siguientes variables:

**Tabla LXXXIX**

**Variables que más aportan al tercer par de variables canónicas**

|  |  |
| --- | --- |
| **Lenguaje** | **Matemáticas** |
| X60 (lectura comprensiva) | X19 (regla de tres simple) |
|  | X24 (perímetro del cuadrado) |
|  | X27 (teorema de Pitágoras) |

**Cuarto par de variables canónicas**



Cabe recalcar que la Var(U4)=1, Var(V4)=1, Cov(U4,V4)=0.392, es decir que la correlación entre el cuarto par de variables no es fuerte y principalmente se concentra en las siguientes variables:

**Tabla XC**

**Variables que más aportan al cuarto par de variables canónicas**

|  |  |
| --- | --- |
| **Lenguaje** | **Matemáticas** |
| X44 (antónimos) | X11 (división de quebrados) |
| X57 (clasificación de oraciones) | X30 (ecuación con una incógnita) |

**Quinto par de variables canónicas**



Cabe recalcar que la Var(U5)=1, Var(V5)=1, Cov(U5,V5)=0.375, es decir que la correlación entre el quinto par de variables no es fuerte y principalmente se concentra en las siguientes variables:

**Tabla XCI**

**Variables que más aportan al quinto par de variables canónicas**

|  |  |
| --- | --- |
| **Lenguaje** | **Matemáticas** |
| X42 (semántica de oraciones) | X28 (factorización de trinomio cuadrado perfecto y diferencia de cuadrados) |
| X55 (identificación de frases) | X29 (factorización trinomio de la forma x2+bx+c) |

Los demás pares de variables canónicas no se comentan debido a que sus correlaciones son bajas.