UNA APLICACIÓN DEL ANÁLISIS POR COMPONENTES PRINCIPALES EN LOS INDICADORES DE GESTIÓN DE LAS UNIVERSIDADES PÚBLICAS

YENNY DALEXA BUENO GUERRERO
COD. 1991623

UNIVERSIDAD INDUSTRIAL DE SANTANDER
FACULTAD DE CIENCIAS
ESCUELA DE MATEMÁTICAS
BUCARAMANGA
2004

UNA APLICACIÓN DEL ANÁLISIS POR COMPONENTES PRINCIPALES EN LOS INDICADORES DE GESTIÓN DE LAS UNIVERSIDADES PÚBLICAS

YENNY DALEXA BUENO GUERRERO
COD. 1991623

DIRECTOR:
GABRIEL YÁÑEZ CANAL

UNIVERSIDAD INDUSTRIAL DE SANTANDER
FACULTAD DE CIENCIAS
ESCUELA DE MATEMÁTICAS
BUCARAMANGA
2004

TITULO: UNA APLICACIÓN DEL ANÁLISIS POR COMPONENTES PRINCIPALES EN LOS INDICADORES DE GESTIÓN DE LAS UNIVERSIDADES PÚBLICAS*

AUTOR: YENNY DALEXA BUENO GUERRERO**

PALABRAS CLAVES:

Componentes Principales

Variación

Correlación

DESCRIPCIÓN

El análisis por componentes principales (ACP), es un instrumento sumamente valioso en la investigación científica dado que permite, mediante diversos procedimientos matemáticos y estadísticos, simplificar fenómenos complejos y reducirlos a dimensiones básicas más asequibles a nuestro entendimiento. El ACP también es útil cuando el investigador desea agrupar las unidades experimentales en subgrupos de tipos semejantes; permitiendo establecer con ellos posibles tipologias.

Con el propósito de documentar los resultados del análisis realizado al sistema de indicadores de las Universidades Estatales se presenta este informe estructurado de la siguiente manera:
El primero y el segundo capítulo presenta los conceptos básicos de Álgebra lineal y Estadística, necesarios para la interpretación geométrica y analítica del ACP.
El tercer capítulo abarca el ACP, que consiste en describir la variación producida por la observación de *p* variables aleatorias, en términos de un conjunto de nuevas variables no correlacionadas entre sí (denominadas componentes principales, CP), cada una de las cuales es combinación lineal de las variables originales y se derivan en orden decreciente de importancia, de manera que la primera CP explique tanta variación en los datos originales como sea posible.
Finalmente, a partir de la aplicación de la técnica seleccionada, en el capítulo cuarto se presentan los resultados del análisis realizado a los indicadores y se formulan las respectivas conclusiones de acuerdo a los hallazgos documentados.

---

* Monografía. **FACULTAD DE CIENCIAS, LICENCIATURA EN MATEMATICAS. Director Gabriel Yañez Canal.

TITLE: AN APPLICATION OF THE ANALYSIS FOR PRINCIPAL COMPONENTS IN THE INDICATOR OF ADMINISTRATION OF THE PUBLIC UNIVERSITIES *

AUTHOR: YENNEY DALEXA BUENO GUERRERO**

KEY WORDS:

Principal components

Variation

Correlation

DESCRIPTION

The Analysis for Principal Components (PCA) it is an extremely valuable instrument in the scientific investigation since it allows by means of diverse mathematical and statistical procedures, to simplify complex phenomena and to reduce them to more affordable basic dimensions to our understanding. The PCA is also useful when the investigator wants to contain the experimental units in subgroups of similar types, allowing settling down with them tipologys possible.

With the purpose of documenting the results of the analysis carried out to the system of indicators of the State Universities, this structured report in the following way is presented:

The first and the second chapter present the basic concepts of lineal and statistical Algebra, necessary for the geometric and analytic interpretation of the PCA. The third chapter, embraces the PCA that consists on describing the variation taken place by the observation of p random variables, in terms of a group of new variables non correlated among if (denominated Principal Components, PC), each one of those which is lineal combination of the original variables and they are derived in falling order of importance, so that the first PC explains as much variation in the original data as it is possible. Finally, starting from the application of the selected technique, in the quarter chapter, the results of the analysis are presented carried out to the indicators and the respective conclusions are formulated according to the documented discoveries.

* Monograph. ** Science faculty. LICENCIATE IN MATHEMATICS. Managing Gabriel Yañez Canal

AGRADECIMIENTOS

A **DIOS** por que siempre esta conmigo y me ha permitido culminar una meta mas en mi realización personal.

De todo corazón agradezco a mi madre **MYRIAN** por su apoyo sincero y desmedido.

A mi padre **RAUL**.

A mis hermanos **OSWALDO, OSCAR Y JEFFERSON** por su apoyo y comprensión.

A **JORGE** por estar a mi lado en momentos de felicidad y tristeza, siempre brindando amor para olvidar los sin sabores de la vida.

A mis amigos **LUCHO, HERMAN, SERGIO Y ANDERSON** por su apoyo incondicional en la búsqueda de mi desarrollo profesional.

Al profesor **GABRIEL YÁNEZ**, por su paciencia y colaboración, en la orientación de este trabajo.

Y a todas aquellas personas que influyeron en la elaboración de esta monografía

INTRODUCCIÓN

Al tomar en consideración la importancia de la educación pública en el país y el afán por mejorar su calidad, el Sistema de Universidades Estatales (SUE) definió una serie de indicadores aplicables a la gestión de las universidades públicas con el fin de registrar los eventos significativos de un campo determinado de la gestión institucional y de esa manera tener una visión clara y precisa de la gestión de cada dependencia.

Cada universidad estatal representa una unidad enmarcada dentro de elementos circunstanciales que la hacen particular y diferente de sus homólogas; por tanto, es relevante contar con características medibles que permitan efectuar un seguimiento al desempeño de cada una y realizar las respectivas comparaciones con el propósito de identificar oportunidades de mejora y tomar las acciones correctivas necesarias en pro de un bienestar individual para la universidad y grupal para el sistema educativo nacional.

Con el propósito de documentar los resultados del análisis realizado al sistema de indicadores de las Universidades Estatales, se presenta este informe estructurado en su primera parte con los conceptos básicos de Álgebra Lineal y Estadística, necesarios para la interpretación geométrica y analítica del Análisis por componentes principales, ACP; seguido de un marco teórico con la definición de la técnica utilizada para el análisis de los datos recopilados, ACP.

Finalmente, a partir de la aplicación de la técnica seleccionada, se presentan los resultados del análisis realizado y se formulan las respectivas conclusiones de acuerdo a los hallazgos documentados.

TABLA DE CONTENIDO

1. FUNDAMENTOS TEÓRICOS DE ÁLGEBRA LINEAL

Dado que la teoría matricial es usada intensamente en el Análisis por Componentes Principales, decidimos realizar una sección inicial con los conceptos básicos del álgebra lineal necesarios para realizar dicho análisis. Tales conceptos son presentados a continuación.

1.1 Espacio vectorial real

Un espacio vectorial real **V** es un conjunto de objetos, llamados vectores, junto con dos operaciones llamadas **suma y multiplicación por un escalar** que satisface los diez axiomas enumerados a continuación.

i. Si $x \in V$ y $y \in V$ entonces $x + y \in V$ (cerradura bajo la suma).

ii. Para todo x, y y $z \in V$, $(x + y) + z = x + (y + z)$ (ley asociativa de la suma de vectores).

iii. Existe un vector $0 \in V$ tal que para todo $x \in V$, $x + 0 = 0 + x = x$ (el 0 se llama vector cero o idéntico aditivo).

iv. Si $x \in V$, existe un vector $-x \in V$ tal que $x + (-x) = 0$ ($-x$ se llama inverso aditivo de x).

v. Si x y z están en V, entonces $x + z = z + x$ (ley conmutativa de la suma de vectores).

vi. Si $x \in V$ y α es un escalar, entonces $\alpha x \in V$ (cerradura bajo la multiplicación por un escalar).

vii. Si x y z están en V y α es un escalar, entonces $\alpha(x + z) = \alpha x + \alpha z$ (primera ley distributiva).

viii. Si $x \in V$ y α, β son escalares, entonces $(\alpha + \beta)x = \alpha x + \beta x$ (segunda ley distributiva).

ix. Si $x \in V$ y α, β son escalares, entonces $\alpha(\beta x) = (\alpha\beta)x$ (ley asociativa de la multiplicación por escalares).

x. Para cada vector $x \in V$, $1x = x$.

1.2 Transformaciones lineales

Definición 1.1 *Sea V y W espacios vectoriales reales. Una función $T: V \to W$ se denomina transformación lineal T de V en W, si satisface que para cada u y v en V y cada escalar α,*

$$T(u+v) = Tu + Tv$$

y

$$T(\alpha v) = \alpha T v$$

Ejemplo 1.1 Una transformación lineal de R^2 en R^3

Sea $T: R^2 \to R^3$ definida por $T\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x+y \\ x-y \\ 3y \end{pmatrix}$. *Por ejemplo* $T\begin{pmatrix} 2 \\ -3 \end{pmatrix} = \begin{pmatrix} -1 \\ 5 \\ -9 \end{pmatrix}$.

Entonces $$T\left[\begin{pmatrix} x_1 \\ y_1 \end{pmatrix} + \begin{pmatrix} x_2 \\ y_2 \end{pmatrix}\right] = T\begin{pmatrix} x_1 + x_2 \\ y_1 + y_2 \end{pmatrix} = \begin{pmatrix} x_1 + x_2 + y_1 + y_2 \\ x_1 + x_2 - y_1 - y_2 \\ 3y_1 + 3y_2 \end{pmatrix}$$

$$= \begin{pmatrix} x_1 + y_1 \\ x_1 - y_1 \\ 3y_1 \end{pmatrix} + \begin{pmatrix} x_2 + y_2 \\ x_2 - y_2 \\ 3y_2 \end{pmatrix} = T\begin{pmatrix} x_1 \\ y_1 \end{pmatrix} + T\begin{pmatrix} x_2 \\ y_2 \end{pmatrix}$$

Así, $T\left[\begin{pmatrix} x_1 \\ y_1 \end{pmatrix} + \begin{pmatrix} x_2 \\ y_2 \end{pmatrix}\right] = T\begin{pmatrix} x_1 \\ y_1 \end{pmatrix} + T\begin{pmatrix} x_2 \\ y_2 \end{pmatrix}$

De manera similar,

$$T\left[\alpha\begin{pmatrix}x\\y\end{pmatrix}\right]=T\begin{pmatrix}\alpha x\\\alpha y\end{pmatrix}=\begin{pmatrix}\alpha x+\alpha y\\\alpha x-\alpha y\\3\alpha y\end{pmatrix}=\alpha\begin{pmatrix}x+y\\x-y\\3y\end{pmatrix}=\alpha T\begin{pmatrix}x\\y\end{pmatrix}$$

Así, T es una transformación lineal.

1.3 Independencia lineal

En el estudio del álgebra lineal, una de las ideas centrales es la de dependencia o independencia lineal de los vectores. En esta sección se define el significado de independencia lineal.

Definición 1.2 *Sea* $V_1, V_2, \ldots, V_n$, *n vectores en un espacio vectorial* V. *Entonces se dice que los vectores son linealmente dependientes si existen n escalares* $C_1, C_2, \ldots, C_n$ *no todos cero tales que*

$$C_1V_1 + C_2V_2 + \ldots + C_nV_n = 0$$

si los vectores no son linealmente dependientes, se dice que son linealmente independientes.

Para decirlo de otra manera $V_1, V_2, \ldots, V_n$ *son linealmente independientes si la ecuación* $C_1V_1 + C_2V_2 + \ldots + C_nV_n = 0$ *se cumple sólo para* $C_1 = C_2 = \ldots = C_n = 0$. *Son linealmente dependientes si el vector cero en* V *se puede expresar como un combinación lineal de* $V_1, V_2, \ldots, V_n$ *con coeficientes no todos iguales a cero.*

Ejemplo 1.2 *Dos vectores linealmente independientes en* R^3.

Los vectores $\begin{pmatrix}1\\2\\4\end{pmatrix}$ *y* $\begin{pmatrix}2\\5\\-3\end{pmatrix}$ *son linealmente independientes; si no lo fueran, se tendría*

$\begin{pmatrix} 2 \\ 5 \\ -3 \end{pmatrix} = c\begin{pmatrix} 1 \\ 2 \\ 4 \end{pmatrix} = \begin{pmatrix} c \\ 2c \\ 4c \end{pmatrix}$. *entonces* $c = 2,\ \ 2c = 5,\ \ 4c = -3,$ *lo cual es evidentemente imposible para cualquier número* c.

1.4 Valores y vectores propios

En el análisis por componentes principales es necesario calcular e interpretar tanto los valores propios generados como los vectores propios.

Definición 1.3 S*ea A una matriz cuadrada de dimensión (p x p), es posible encontrar un escalar* λ *y un vector V de dimensión (p x 1), no nulo tal que*

$$A \cdot V = \lambda \cdot V$$

El vector $V \neq 0$ *se llama vector propio de A correspondiente al valor propio* λ.

Ejemplo 1.3 *Valores y vectores propios de una matriz de* 2×2

Sea $A = \begin{bmatrix} 10 & -18 \\ 6 & -11 \end{bmatrix}$ *entonces* $A\begin{bmatrix} 2 \\ 1 \end{bmatrix} = \begin{bmatrix} 10 & -18 \\ 6 & -11 \end{bmatrix}\begin{bmatrix} 2 \\ 1 \end{bmatrix} = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$.

Así, $\lambda_1 = 1$ *es un valor propio de A con el correspondiente vector propio* $v_1 = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$. *De manera similar,* $A\begin{bmatrix} 3 \\ 2 \end{bmatrix} = \begin{bmatrix} 10 & -18 \\ 6 & -11 \end{bmatrix}\begin{bmatrix} 3 \\ 2 \end{bmatrix} = \begin{bmatrix} -6 \\ 4 \end{bmatrix} = -2\begin{bmatrix} 3 \\ 2 \end{bmatrix}$ *de manera que* $\lambda_2 = -2$ *es un valor propio de A con el correspondiente vector propio* $v_2 = \begin{bmatrix} 3 \\ 2 \end{bmatrix}$. *Como se verá enseguida, estos son los únicos valores propio de A.*

Teorema 1.1 *Sea A una matriz de* $n \times n$. *Entonces* λ *es un valor propio de A si y solo si*

$$p(\lambda) = \det(A - \lambda I) = 0 \qquad (1.1)$$

Demostración. Como A es una matriz de $n \times n$, la ecuación $(A - \lambda I)V = 0$ corresponde a un sistema homogéneo de n ecuaciones con las incógnitas $X_1, X_2, \ldots, X_n$. Como se ha puesto que el sistema tiene soluciones no triviales, se concluye que $\det(A - \lambda I) = 0.$ Inversamente, si $\det(A - \lambda I) = 0,$ entonces $(A - \lambda I)V = 0$ tiene soluciones no triviales y λ es el valor propio de A.

Por otro lado, si $\det(A - \lambda I) \neq 0,$ entonces la única solución a $(A - \lambda I)V = 0$ es $V = 0$ de manera que λ no es un valor propio de A.

Definición 1.4 Ecuación y polinomio característico

La ecuación (1.1) *se llama la ecuación característica de A;* $p(\lambda)$ *se llama el polinomio característico de A.*

Teorema 1.2 *Sea A una matriz de* $n \times n$ *y sea* $\lambda_1, \lambda_2, \ldots, \lambda_m$ *valores propios distintos de A (es decir,* $\lambda_i \neq \lambda_j$, *si* $i \neq j$*) con vectores propios correspondientes* $v_1, v_2, \ldots, v_m$. *Entonces* $v_1, v_2, \ldots, v_m$ *son linealmente independientes. Esto es: los vectores propios correspondientes a valores propios distintos son linealmente independientes.*

Demostración

Se hará la demostración por inducción matemática comenzando con m=2, suponga que

$$C_1V_1 + C_2V_2 = 0 \qquad (1.2)$$

Multiplicando ambos lados de (1.2) por A se tiene:

$$A(C_1V_1 + C_2V_2) = C_1AV_1 + C_2AV_2 = 0$$

o sea (como $AV_I = \lambda_I V_I$ para i=1,2)

$$C_1\lambda_1V_1 + C_2\lambda_2V_2 = 0 \qquad (1.3)$$

se multiplica (1.2) por λ_1 y se resta de (1.3) para obtener

$$(C_1\lambda_1V_1 + C_2\lambda_2V_2) - (C_1\lambda_1V_1 + C_2\lambda_1V_2) = 0$$

o sea

$$C_2(\lambda_2 - \lambda_1)V_2 = 0$$

como $V_2 \neq 0$ (por definición de vector propio) y como $\lambda_2 \neq \lambda_1$, se concluye que $C_2 = 0$, entonces sustituyendo $C_2 = 0$ en (1), se ve que $C_1 = 0$, lo que prueba el teorema en el caso m=2. Ahora suponga que el teorema se cumple para m=k. Esto es, se supone que k vectores propios correspondientes a valores propios distintos son linealmente independientes. Ahora se prueba el teorema para m=k+1.

Así, que se supone que

$$C_1V_1 + C_2V_2 + \ldots + C_kV_k + C_{k+1}V_{k+1} = 0 \qquad (1.4)$$

multiplicando ambos lados de (1.4) por A y usando el hecho de que $AV_I = \lambda_I V_I$, se obtiene

$$C_1\lambda_1V_1 + C_2\lambda_2V_2 + \ldots + C_k\lambda_kV_k + C_{k+1}\lambda_{k+1}V_{k+1} = 0 \qquad (1.5)$$

se multiplican ambos lados de (1.4) por λ_{k+1} y se resta de (1.5)

$$C_1(\lambda_1 - \lambda_{k+1})V_1 + C_2(\lambda_2 - \lambda_{k+1})V_2 + \ldots + C_{k+1}(\lambda_k - \lambda_{k+1})V_k = 0$$

pero por la suposición de inducción, $V_1, V_2, \ldots, V_K$ son linealmente independientes.

Así,

$$C_1(\lambda_1 - \lambda_{k+1})V_1 = C_2(\lambda_2 - \lambda_{k+1})V_2 = \ldots = C_{k+1}(\lambda_k - \lambda_{k+1})V_k = 0;$$

y como $\lambda_i \neq \lambda_{k+1}$ para $i = 1,2,\ldots,k,$ se concluye que $C_1 = C_2 = \ldots = C_k = 0$ pero de (1.4) esto significa que $C_{k+1} = 0.$

Por lo tanto el teorema se cumple para m=k+1 y la prueba queda completa..

Procedimiento para calcular valores y vectores propios

i. *Se encuentra* $p(\lambda) = \det(A - \lambda I)$.

ii. *Se encuentra las raíces* $\lambda_1, \lambda_2, \ldots, \lambda_m$ *de* $p(\lambda) = 0$.

iii. *Se resuelve el sistema homogéneo* $(A - \lambda_i I)v = 0$, *correspondiente a cada valor propio* λ_i.

Ejemplo 1.4 *Calculo de valores y vectores propios.*

Sea $A = \begin{bmatrix} 6 & 3 \\ 3 & 4 \end{bmatrix}$ *entonces*

$$\det(A-\lambda I)=\begin{vmatrix}6-\lambda & 3\\ 3 & 4-\lambda\end{vmatrix}=(6-\lambda)(4-\lambda)-9=\lambda^2-10\lambda+24-9=\lambda^2-10\lambda+15=0.$$

Aplicando la nomenclatura tradicional de una ecuación de segundo grado con valores $a=1,\ b=-10$ *y* $c=15$, *resolviendo con respecto a* $\lambda,$ *se obtienen dos raíces que, en este caso, son* $\lambda_1=5-\sqrt{10}$ *y* $\lambda_2=5+\sqrt{10}.$

Para $\lambda_1=5-\sqrt{10}$ *se obtiene* $(A-\lambda I)v=\begin{bmatrix}1+\sqrt{10} & 3\\ 3 & -1+\sqrt{10}\end{bmatrix}\begin{bmatrix}x_1\\ x_2\end{bmatrix}=\begin{bmatrix}0\\ 0\end{bmatrix}.$ *Es claro que para cualquier vector propio correspondiente a* $\lambda_1=5-\sqrt{10}$ *satisface* $(1+\sqrt{10})x_1+3x_2=0$ *y* $3x_1+(-1+\sqrt{10})x_2=0.$

Ahora igualamos las dos ecuaciones

$$(1+\sqrt{10})x_1+3x_2=3x_1+(-1+\sqrt{10}) \qquad (1.6)$$

Es decir que se dispone de una sola ecuación para resolver un sistema que posee dos incógnitas x_1 *y* x_2. *Es posible reducir la ecuación* (1.6) *y expresarla como sigue* $1.623x_1+0.8377x_2=0.$

Para encontrar otra ecuación que permita completar el sistema se establece la condición de que los vectores propios estén normalizados. Esto equivale en términos algebraicos a que la suma de los cuadrados de los elementos del vector debe ser 1.

Así, deberá cumplirse que:

$$x_1^2+x_2^2=1$$

y se completa así el sistema de dos ecuaciones que, para λ_1 *será en este ejemplo*

$$x_1^2+x_2^2=1$$

$$1.623x_1+0.8377x_2=0$$

despejando por el método de sustitución se encuentra que

$$x_1=\frac{-0.8377}{1.1623}x_2=-0.7207x_2$$

$$x_1^2+x_2^2=1$$

$$(-0.7207x_2)^2+x_2^2=1$$

$$0.5187x_2^2 + x_2^2 = 1$$

$$1.5187x_2^2 = 1$$

$$x_2^2 = \frac{1}{1.5187}$$

$$x_2 = \sqrt{\frac{1}{1.5187}} = 0.8114$$

y sustituyendo para encontrara x_1 *se obtiene:*

$$x_1 = -0.7207x_2$$

$$x_1 = -0.7207(0.8114)$$

$$x_1 = -0.5847$$

En forma matricial el vector propio generado por λ_1 *puede ordenarse así:*

$$v_1 = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} -0.5847 \\ 0.8114 \end{bmatrix}$$

de manera similar podemos encontrar el vector propio generado por λ_2 *y se obtiene:*

$$v_2 = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0.8112 \\ 0.5847 \end{bmatrix}$$

Definición 1.5 Matriz triangular superior e inferior

Una matriz cuadrada se llama triangular superior(inferior) si todos sus componentes abajo(arriba) de la diagonal principal son cero.

Ejemplo 1.5 Una matriz triangular superior y una matriz triangular inferior. *La matriz U es triangular superior, mientras que la matriz V es triangular inferior.*

$$U = \begin{bmatrix} 2 & -3 & 5 \\ 0 & 1 & 6 \\ 0 & 0 & 2 \end{bmatrix} \qquad V = \begin{bmatrix} 2 & 0 & 0 \\ -5 & 4 & 0 \\ 6 & 1 & 2 \end{bmatrix}$$

Teorema 1.3 *Los valores propios de una matriz triangular son las componentes diagonales de la matriz.*

Demostración

Si $A = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ 0 & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & a_{nn} \end{bmatrix}$ *entonces*

$$A - \lambda I = \begin{bmatrix} a_{11} - \lambda & a_{12} & \dots & a_{1n} \\ 0 & a_{22} - \lambda & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & a_{nn} - \lambda \end{bmatrix}$$

Como el determinante de una matriz triangular es igual al producto de las componentes de la diagonal, se ve que $\det(A - \lambda I) = (a_{11} - \lambda)(a_{22} - \lambda)\dots(a_{nn} - \lambda)$ *con ceros* $a_{11}, a_{22}, \dots, a_{nn}$.

La demostración para una matriz triangular inferior es idéntica.

Ejemplo 1.6 Valores propios de una matriz triangular.

Sea $A = \begin{bmatrix} 2 & 5 & 6 \\ 0 & -3 & 2 \\ 0 & 0 & 5 \end{bmatrix}$ *entonces*

$$\det(A - \lambda I) = \begin{vmatrix} 2-\lambda & 5 & 6 \\ 0 & -3-\lambda & 2 \\ 0 & 0 & 5-\lambda \end{vmatrix} = (2-\lambda)(-3-\lambda)(5-\lambda) = 0$$

de manera que los valores propios son $\lambda_1 = 2, \lambda_2 = -3$ *y* $\lambda_3 = 5$.

Definición 1.6 Matriz simétrica

La matriz cuadrada A de $n \times n$ *se llama simétrica si* $A^t = A$. *Es decir, las columnas de A son también los renglones de A.*

Ejemplo 1.7 *Las siguientes dos matrices son simétricas*

$$A = \begin{bmatrix} 1 & -4 & 2 \\ -4 & 7 & 5 \\ 2 & 5 & 0 \end{bmatrix} \qquad B = \begin{bmatrix} -1 & 2 & 4 & 6 \\ 2 & 7 & 5 & 3 \\ 4 & 3 & 8 & 0 \\ 6 & 5 & 0 & -4 \end{bmatrix}$$

Definición 1.7 Producto escalar

Sean $a = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix}$ *y* $b = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix}$ *dos vectores. Entonces el producto escalar de a y b, denotado por a.b esta dado por*

$$a.b = a_1 b_1 + a_2 b_2 + \ldots + a_n b_n .$$

El producto escalar se llama con frecuencia producto punto o producto interno.

Definición 1.8 Producto interno en C^n**.**

Sean $X = (x_1, x_2, \ldots, x_n)$ *y* $Y = (y_1, y_2, \ldots, y_n)$ *en* C^n. *Entonces el producto interno de X y Y, denotado por* (X, Y) *esta dado por*

$$(X, Y) = x_1 \overline{y}_1 + x_2 \overline{y}_2 + \ldots + x_n \overline{y}_n$$

Definición 1.9 Espacio con producto interno

Un espacio vectorial complejo V se llama espacio con producto interno si para cada par ordenado de vectores u y v en V, existe un número complejo único (u, v) *llamado producto interno de u y v, talque si u,v y w están en V y* $\alpha \in C$, *entonces*

i. $(v, v) \geq 0.$

ii. $(v, v) = 0$ *si y solo si* $v = 0.$

iii. $(u, v + w) = (u, v) + (u, w).$

iv. $(u + v, w) = (u, w) + (v, w).$

v. $(u,v)=(\overline{v,u})$.

vi. $(\alpha u,v)=\alpha(u,v)$.

vii. $(u,\alpha v)=\overline{\alpha}(u,v)$.

Teorema 1.4 *Sea A una matriz simétrica real de $n\times n$. Entonces los valores propios de A son reales.*

Demostración. Sea λ un valor propio de A con vector propio V; es decir, $A\cdot V=\lambda\cdot V$.

El vector V esta en C^n, y el producto interno en C^n satisface que

$$(\alpha x,y)=\alpha(x,y) \text{ y } (x,\alpha y)=\overline{\alpha}(x,y),\ \alpha\in C^n \qquad (1.7)$$

Entonces

$$(AV,V)=(\lambda V,V)=\lambda(V,V) \qquad (1.8)$$

Mas aún, teniendo en cuenta que $(Ax)\cdot y=x\cdot(A^t y)$ donde $x,y\in R^n$ y $A^t=A$

$$(AV,V)=(V,A^tV)=(V,AV)=(V,\lambda V)=\overline{\lambda}(V,V) \qquad (1.9)$$

Igualando (1.8) y (1.9) se tiene

$$\lambda(V,V)=\overline{\lambda}(V,V) \qquad (1.10)$$

Pero $(V,V)=\|V\|^2\neq 0$, ya que V es un vector propio. Entonces se puede dividir ambos miembros de (1.10) entre (V,V) para obtener

$$\lambda=\overline{\lambda} \qquad (1.11)$$

Si $\lambda=a+bi$, entonces $\overline{\lambda}=a-bi$ y de (1.11) se tiene

$$a+bi=a-bi \qquad (1.12)$$

lo que se cumple solo si $b=0$. Esto muestra que $\lambda=a$; por lo tanto λ es real y la demostración queda completa.

Definición 1.10 Vectores ortogonales

*Se dice que dos vectores **a** y **b** son ortogonales si **a.b=0**.*

Teorema 1.5 *Sea A una matriz simétrica real de* $n \times n$. *Si* λ_1 *y* λ_2 *son valores propios diferentes con vectores propios reales correspondientes* V_1 *y* V_2, *entonces* V_1 *y* V_2 *son ortogonales*.

Demostración. Se calcula

$$AV_1 \cdot V_2 = \lambda_1 V_1 \cdot V_2 = \lambda_1 (V_1 V_2) \qquad (1.13)$$

y

$$AV_1 \cdot V_2 = V_1 \cdot A^t V_2 = V_1 \cdot AV_2 = V_1 \cdot (\lambda_2 V_2) = \lambda_2 \cdot (V_1 V_2) \qquad (1.14)$$

Combinando (1.13) y (1.14), se tiene $\lambda_1 \cdot (V_1 V_2) = \lambda_2 \cdot (V_1 V_2)$ y como $\lambda_1 \neq \lambda_2$, se concluye que $V_1 \cdot V_2 = 0.$ Esto es lo que se quería demostrar.

Definición 1.11 Coseno del ángulo entre dos vectores

Un producto escalar induce, además de la noción de norma, el concepto de ángulo. El ángulo θ *entre dos vectores x e y está definido por su coseno, que por definición, es igual al producto escalar de estos dos vectores divididos por sus normas:*

$$Cos\theta = \frac{x.y}{\|x\|.\|y\|}$$

el coseno está comprendido entre 1 y –1. si vale 1 los vectores son colineales del mismo sentido, si vale –1, son colineales de sentido opuesto. Si vale 0 los dos vectores son ortogonales.

2. CONCEPTOS BASICOS DE ESTADÍSTICA

En el Análisis por Componentes Principales tanto la varianza como la covarianza juegan un papel importante a la hora de crear las matrices de covarianza y correlación, ya que a partir de estas matrices podemos generar las componentes principales.
Por esta razón citamos sus definiciones y algunas de sus propiedades.

Definición 2.1 *Sea X una variable aleatoria con valores posibles* $X_1, X_2, \ldots, X_n$*. Si todos los valores posibles son igualmente probables, la esperanza de la variable X se define por:*

$$E(X) = \overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

Algunas propiedades de la media muestral.

Propiedad 1. *Si c es una constante,*

$$\overline{cX} = c\overline{X}$$

Demostración.

$$\overline{cX} = \frac{cx_1 + cx_2 + \ldots + cx_n}{n} = c\,\frac{(x_1 + x_2 + \ldots + x_n)}{n} = c\overline{X}$$

Propiedad 2. *Si c es una constante,*

$$\overline{X + c} = \overline{X} + c$$

Demostración.

$$\overline{X + c} = \frac{x_1 + c + x_2 + c + \ldots + x_n + c}{n} = \frac{x_1 + x_2 + \ldots + x_n}{n} + \frac{nc}{n} = \overline{X} + c$$

Definición 2.2 *Sea X una variable aleatoria. Definimos la varianza muestral de X, denotada por* $Var(X)$ *como sigue :*

$$Var(X)=E\left[(X-E(X))^2\right]=\frac{\sum\left(X-\overline{X}\right)^2}{n}$$

Algunas propiedades de la varianza muestral.

Propiedad 3. *Si c es una constante,*

$$Var(cX)=c^2Var(X)$$

Demostración.

$$Var(cX)=\frac{\sum\left(cX-\overline{cX}\right)^2}{n}=\frac{\sum\left(cX-c\overline{X}\right)^2}{n}$$

$$=\frac{\sum\left[(cX)^2-2(cX)\left(c\overline{X}\right)+\left(c\overline{X}\right)^2\right]}{n}=\frac{\sum\left(c^2X^2-2c^2X\overline{X}+c^2\overline{X}^2\right)}{n}$$

$$=c^2\frac{\sum\left(X^2-2X\overline{X}+\overline{X}^2\right)}{n}=c^2\frac{\sum\left(X-\overline{X}\right)^2}{n}=c^2Var(X)$$

Propiedad 4. *Si c es una constante,*

$$Var(X+c)=Var(X)$$

Demostración.

$$Var(X+c)=\frac{\sum\left((X+c)-(\overline{c+X})\right)^2}{n}=\frac{\sum\left((X+c)-(\overline{X}+c)\right)^2}{n}$$

$$=\frac{\sum\left(X+c-\overline{X}-c\right)^2}{n}=\frac{\sum\left(X-\overline{X}\right)^2}{n}=Var(X)$$

Teorema 2.1 $Var(X)=E\left(X^2\right)-[E(X)]^2$

Demostración.

$$Var(X)=E\left[(X-E(X))^2\right]=E\left[X^2-2XE(X)+[E(X)]^2\right]$$

$$=E\left(X^2\right)-2E(X)E(X)+[E(X)]^2=E\left(X^2\right)-[E(X)]^2$$

Definición 2.3 *Sean X e Y variables aleatorias. Definimos la covarianza muestral entre la variable X e Y, denotada por* $Cov(X,Y)$ *como:*

$$Cov(X,Y) = E[(X - E(X))(Y - E(Y))] = \frac{\sum (X - \overline{X})(Y - \overline{Y})}{n}$$

Propiedad 5. (de la covarianza muestral) *Si c y d son constantes,*

$$Cov(cX, dY) = cdCov(X,Y)$$

Demostración.

$$Cov(cX, dY) = \frac{\sum (cX - \overline{cX})(dY - \overline{dY})}{n} = \frac{\sum (cX - c\overline{X})(dY - d\overline{Y})}{n}$$

$$= cd\,\frac{\sum (X - \overline{X})(Y - \overline{Y})}{n} = cdCov(X,Y)$$

Proposición 2.1 *Sean* $X_1, X_2, \ldots, X_n, Y_1, Y_2, \ldots, Y_m$ *variables aleatorias entonces*

$$Cov(X_1 + X_2 + \ldots + X_n, Y_1 + Y_2 + \ldots + Y_m) = \sum_{i=1}^{n} \sum_{j=1}^{m} Cov(X_i, Y_j)$$

Demostración

$$Cov(X_1 + X_2 + \ldots + X_n, Y_1 + Y_2 + \ldots + Y_m) =$$

$$= E((X_1 + X_2 + \ldots + X_n) - E(X_1 + X_2 + \ldots + X_n))E((Y_1 + Y_2 + \ldots + Y_m) - E(Y_1 + Y_2 + \ldots + Y_m))$$

$$= E((X_1 + X_2 + \ldots + X_n) - (E(X_1) + E(X_2) + \ldots + E(X_n)))E((Y_1 + Y_2 + \ldots + Y_m)) -$$

$$- (E(Y_1) + E(Y_2) + \ldots + E(Y_m)))$$

$$= E((X_1 - E(X_1)) + (X_2 - E(X_2)) + \ldots + (X_n - E(X_n)))E((Y_1 - E(Y_1)) + (Y_2 - E(Y_2)) + \ldots + (Y_m - E(Y_m)))$$

$$= E((X_1 - E(X_1))(Y_1 - E(Y_1)) + (X_1 - E(X_1))(Y_2 - E(Y_2)) + \ldots + (X_1 - E(X_1))(Y_m - E(Y_m)) +$$

$$+ (X_2 - E(X_2))(Y_1 - E(Y_1)) + \ldots + (X_n - E(X_n))(Y_1 - E(Y_1)) + \ldots + (X_n - E(X_n))(Y_m - E(Y_m)))$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{m} Cov(X_i, X_j)$$

Proposición 2.2 *Sea* $X_1, X_2, \ldots, X_n$ *variables aleatorias entonces*

$$Var(X_1 + X_2 + \ldots + X_n) = \sum_{i=1}^{n} Var(X_i) + 2\sum_{i<j} Cov(X_i, X_j)$$

Demostración

$Var(X_1 + X_2 + \ldots + X_n) = E\left(\left(X_1 + X_2 + \ldots + X_n\right) - E\left(X_1 + X_2 + \ldots + X_n\right)\right)^2 =$

$= E\left(\left(X_1 + X_2 + \ldots + X_n\right) - \left(E\left(X_1\right) + E\left(X_2\right) + \ldots + E\left(X_n\right)\right)\right)^2 =$

$= E\left(\left(X_1 - E\left(X_1\right)\right) + \left(X_2 - E\left(X_2\right)\right) + \ldots + \left(X_n - E\left(X_n\right)\right)\right)^2 =$

$= E\left[\sum_{i=1}^{n}\left(X_i - E\left(X_i\right)\right) + 2\sum_{i<j}\left(X_i - E\left(X_i\right)\right)\left(X_j - E\left(X_j\right)\right)\right] = \sum_{i=1}^{n} Var(X_i) + 2\sum_{i<j} Cov(X_i, X_j)$

Proposición 2.3 *Sea* $\vec{a} = \left(a_1, a_2, \ldots, a_p\right)$ *vector de* p *constantes,* $\vec{X} = \left(X_1, X_2, \ldots X_p\right)$ *y la matriz de covarianza*

$$\Sigma = \begin{bmatrix} Var\left(X_1\right) & Cov\left(X_1, X_2\right) & \ldots & Cov\left(X_1, X_p\right) \\ Cov\left(X_2, X_1\right) & Var\left(X_2\right) & \ldots & Cov\left(\left(X_2, X_p\right)\right. \\ \vdots & \vdots & \ddots & \vdots \\ Cov\left(X_p, X_1\right) & Cov\left(X_p, X_2\right) & \ldots & Var\left(X_p\right) \end{bmatrix},$$

entonces $Var(\vec{a}'\vec{X}) = \vec{a}'\sum\vec{a}$

Demostración.

A continuación se presenta el caso particular $p = 3$

$Var(\vec{a}'\vec{X}) = Var(a_1X_1 + a_2X_2 + a_3X_3)$

$= Var(a_1X_1) + Var(a_2X_2) + Var(a_3X_3) + 2Cov(a_1X_1, a_2X_2) + 2Cov(a_1X_1, a_3X_3) +$

$2Cov(a_2X_2, a_3X_3)$

$= a_1^2 Var(X_1) + a_2^2 Var(X_2) + a_3^2 Var(X_3) + 2a_1a_2Cov(X_1, X_2) + 2a_1a_3Cov(X_1, X_3) +$

$2a_2a_3Cov(X_2, X_3)$

$= a_1\left[a_1Var(X_1) + a_2Cov(X_1, X_2) + a_3Cov(X_1, X_3)\right] + a_2\left[a_1Cov(X_1, X_2) + a_2Var(X_2) +\right.$

$\left.a_3Cov(X_2, X_3)\right] + a_3\left[a_1Cov(X_1, X_3) + a_2Cov(X_2, X_3) + a_3Var(X_3)\right]$

$$=\begin{bmatrix} a_1 & a_2 & a_3 \end{bmatrix}\begin{bmatrix} a_1Var(X_1)+a_2Cov(X_1,X_2)+a_3Cov(X_1,X_3) \\ a_1Cov(X_1,X_2)+a_2Var(X_2)+a_3Cov(X_2,X_3) \\ a_1Cov(X_1,X_3)+a_2Cov(X_2,X_3)+a_3Var(X_3) \end{bmatrix}$$

$$=\begin{bmatrix} a_1 & a_2 & a_3 \end{bmatrix}\begin{bmatrix} Var(X_1) & Cov(X_1,X_2) & Cov(X_1,X_3) \\ Cov(X_1,X_2) & Var(X_2) & Cov(X_2,X_3) \\ Cov(X_1,X_3) & Cov(X_2,X_3) & Var(X_3) \end{bmatrix}\begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix}$$

como $Cov(X_i,X_j)=Cov(X_j,X_i)$ para $i \neq j$ tenemos que

$$=\begin{bmatrix} a_1 & a_2 & a_3 \end{bmatrix}\begin{bmatrix} Var(X_1) & Cov(X_1,X_2) & Cov(X_1,X_3) \\ Cov(X_2,X_1) & Var(X_2) & Cov(X_2,X_3) \\ Cov(X_3,X_1) & Cov(X_3,X_2) & Var(X_3) \end{bmatrix}\begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix}$$

$$=\overrightarrow{a'}\sum\overrightarrow{a'}$$

de manera similar se realiza para cualquier valor de p.

Proposición 2.4 *Si* $\vec{a}=(a_1,a_2,\ldots,a_p)$, $\vec{b}=(b_1,b_2,\ldots,b_p)$ *vectores de* p *constantes y* $\overrightarrow{X}=(X_1,X_2,\ldots X_p)$ *entonces* $Cov(\overrightarrow{a'}\overrightarrow{X},\overrightarrow{b'}\overrightarrow{X})=\overrightarrow{a'}\sum\vec{b}$.

Demostración.

A continuación se presenta el caso particular $p=3$

$$Cov(\overrightarrow{a'}\overrightarrow{X},\overrightarrow{b'}\overrightarrow{X})=Cov(a_1X_1+a_2X_2+a_3X_3,b_1X_1+b_2X_2+b_3X_3)$$

$$=Cov(a_1X_1,b_1X_1)+Cov(a_1X_1,b_2X_2)+Cov(a_1X_1,b_3X_3)+Cov(a_2X_2,b_1X_1)+$$

$$Cov(a_2X_2,b_2X_2)+Cov(a_2X_2,b_3X_3)+Cov(a_3X_3,b_1X_1)+Cov(a_3X_3,b_2X_2)+$$

$$+Cov(a_3X_3,b_3X_3)$$

$$=a_1b_1Cov(X_1,X_1)+a_1b_2Cov(X_1,X_2)+a_1b_3Cov(X_1,X_3)+a_2b_1Cov(X_2,X_1)+$$

$$a_2b_2Cov(X_2,X_2)+a_2b_3Cov(X_2,X_3)+a_3b_1Cov(X_3,X_1)+a_3b_2Cov(X_3,X_2)+$$

$$+a_3b_3Cov(X_3,X_3)$$

$$=a_1b_1Var(X_1)+a_1b_2Cov(X_1,X_2)+a_1b_3Cov(X_1,X_3)+a_2b_1Cov(X_2,X_1)+a_2b_2Var(X_2)+$$

$$a_2b_3Cov(X_2,X_3)+a_3b_1Cov(X_3,X_1)+a_3b_2Cov(X_3,X_2)+a_3b_3Var(X_3)$$

$$= \begin{bmatrix} a_1 & a_2 & a_3 \end{bmatrix} \begin{bmatrix} b_1 Var(X_1) + b_2 Cov(X_1, X_2) + b_3 Cov(X_1, X_3) \\ b_1 Cov(X_1, X_2) + b_2 Var(X_2) + b_3 Cov(X_2, X_3) \\ b_1 Cov(X_1, X_3) + b_2 Cov(X_2, X_3) + b_3 Var(X_3) \end{bmatrix}$$

$$= \begin{bmatrix} a_1 & a_2 & a_3 \end{bmatrix} \begin{bmatrix} Var(X_1) & Cov(X_1, X_2) & Cov(X_1, X_3) \\ Cov(X_1, X_2) & Var(X_2) & Cov(X_2, X_3) \\ Cov(X_1, X_3) & Cov(X_2, X_3) & Var(X_3) \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix}$$

$$= \overrightarrow{a'} \sum \vec{b}$$

de manera similar se realiza para cualquier valor de p.

2.1 Variables aleatorias independientes

Lo que queremos decir intuitivamente es que X e Y son variables aleatorias independientes si el resultado de X, digamos, de ninguna manera influye en el resultado de Y. Esta es una noción extremadamente importante y hay muchas situaciones en que tal suposición se justifica.

Definición 2.4 *Dos variables aleatorias X e Y son independientes si y solo si* $E(XY) = E(X)E(Y)$*, es decir,* $Cov(X, Y) = 0$.

Corolario 2.1 *Sea* $X_1, X_2, \ldots, X_n$ *variables aleatorias tal que* $Cov(X_i, X_j) = 0$ *para* $i \neq j$ *entonces* $Var(X_1 + X_2 + \ldots + X_n) = \sum_{i=1}^{n} Var(X_i)$.

3. ANALISIS POR COMPONENTES PRINCIPALES

El Análisis de Componentes Principales es un Método Estadístico, propuesto por Pearson (1901), y de forma independiente también por Hotelling (1933), que consiste en describir la variación producida por la observación de *p* variables aleatorias X_i, en términos de un conjunto de nuevas variables no correlacionadas entre sí (denominadas Componentes Principales), cada una de las cuales es combinación lineal de las variables originales.

Estas nuevas variables son obtenidas en orden de importancia, de manera que la primera componente principal incorpora la mayor cantidad posible de variación debida a las variables originales; la segunda componente principal se elige de forma que explique la mayor cantidad posible de variación que resta sin explicar por la primera componente principal, sujeta a la condición de ser no correlacionada con la primera componente principal, y así sucesivamente.

El propósito del Análisis de Componentes Principales es ver si las dos o tres primeras componentes principales reúnen ya la mayor parte de la variación producida por las *p* variables originales puesto que, de ser así, considerando sólo estas dos o tres primeras, reduciremos la dimensionalidad de los datos al considerar únicamente dos o tres variables en lugar de *p*, y apenas perderemos información relevante.

Los datos se organizan en tablas rectangulares, donde las columnas representan a las variables, las cuales son de tipo cuantitativo, y las filas representan a los individuos sobre los que estas variables son medidas.

La tabla de datos se visualiza de la siguiente forma:

$$X = \begin{bmatrix} x_{11} & \cdots & x_{1j} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots & & \vdots \\ x_{i1} & \cdots & x_{ij} & \cdots & x_{ip} \\ \vdots & & \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nj} & \cdots & x_{np} \end{bmatrix}$$

Donde x_{ij} es el valor de la variable *j* en el individuo *i*, n es el número de individuos y *p* es el número de variables.

3.1 Objetivos

Los objetivos más importantes del análisis por componentes principales son:

- Generar nuevas variables que puedan expresar la información contenida en el conjunto original de datos.
- Reducir la dimensionalidad del problema que se está estudiando, como paso previo para futuros análisis.
- Eliminar, cuando sea posible , algunas de las variables originales ya sea porque ellas aportan poca información o porque una variable contiene en parte información ya suministrada por otra variable.

Las nuevas variables generadas se denominan **“Componentes Principales”**, las cuales son combinaciones lineales de las variables originales.
El estudio se centra en las componentes que sintetizan la mayor variabilidad de los datos originales. Por inspección de estas componentes es posible detectar relaciones entre los individuos y las variables.

Balance de la semejanza entre individuos
Se trata de responder a preguntas como:

- ¿Qué individuos son los que se asemejan?
- ¿Qué individuos son los que se diferencian?
- ¿Existe grupos homogéneos de individuos?
- ¿Se puede poner en evidencia una tipología entre individuos?

Balance de las relaciones entre variables

Se trata de responder a preguntas como:

- ¿Qué variables son las que aparecen relacionadas positivamente entre sí?
- ¿Cuáles son las que se oponen (relacionadas negativamente)?
- ¿Existen grupos de variables relacionadas entre sí ?
- ¿Se puede poner en evidencia una tipología entre variables?

Definición 3.1 *Al Conjunto de individuos se le denomina nube de individuos* ***N(n).*** *Las n filas son tratadas como n puntos en el espacio vectorial* R^p.

Definición 3.2 *Al Conjunto de variables se le denomina nube de variables* ***N(p).*** *Las p columnas son tratadas como p puntos en el espacio vectorial* R^n.

En la nube de individuos se comparan éstos en términos de sus características y en la nube de variables se obtiene información acerca de la relación entre características consideradas en función de los individuos que se estudian.

Definición 3.3 *La media de la variable j se define por:*

$$\overline{X}_j = \frac{\sum_{i=1}^{n} X_{ij}}{n} \quad , \quad j = 1,2,...,p.$$

Definición 3.4 *En la matriz de datos X el centro de gravedad de la nube de individuos se le define como:*

$$G = \begin{bmatrix} \bar{x}_1 \\ \vdots \\ \bar{x}_p \end{bmatrix} \quad donde \quad \bar{x}_j = \frac{\sum_{i=1}^{n} x_{ij}}{n} \quad , \quad \text{j = 1,2,...,p.}$$

Se observa que $\bar{x}_j$ es la media de la variable j.

Definición 3.5 *La varianza de la variable j se define por:*

$$S_{jj} = \frac{\sum_{i=1}^{n} \left(x_{ij} - \bar{x}_j \right)^2}{n} \quad , \quad j = 1,2,...,p.$$

Definición 3.6 *La covarianza entre la j-ésima y la k-ésima variable está dada por:*

$$S_{jk} = \frac{\sum_{i=1}^{n} \left(x_{ij} - \bar{x}_j \right)\left(x_{ik} - \bar{x}_k \right)}{n} \quad , \quad j,k = 1,2,...,p.$$

Definición 3.7 *La matriz formada por los* s_{jj} *y los* s_{jk} *es la matriz de covarianza. Y se denota como:*

$$\sum = \begin{bmatrix} s_{11} & \cdots & s_{1j} & \cdots & s_{1p} \\ \vdots & \ddots & \vdots & & \vdots \\ s_{i1} & \cdots & s_{ij} & \cdots & s_{ip} \\ \vdots & & \vdots & \ddots & \vdots \\ s_{p1} & \cdots & s_{pj} & \cdots & s_{pp} \end{bmatrix}$$

Definición 3.8 A l*a traza de la matriz* $\sum$*, se le denomina variación total. Y se denota como*

$$Tr \sum = \sum_{j=1}^{p} s_{jj}$$

Es decir, la suma de las varianzas de las variables.

La varianza total será mayor cuanto mayor sea la dispersión de los datos alrededor de la media.

Definición 3.9 *Se define la distancia entre los individuos i y l como:*

$$d^2(i,l) = \sum_{j=1}^{p} \left(X_{ij} - X_{lj} \right)^2$$

Definición 3.10 *A partir de los elementos de la matriz* $\sum$ *es posible calcular la matriz* ***R*** *(matriz de correlación), de igual dimensión que* $\sum$, *y cuyos elementos son los coeficientes de correlación entre la j-ésima y la k-ésima variables:*

$$\rho_{jk} = \frac{s_{jk}}{\sqrt{s_{jj}}\sqrt{s_{kk}}}$$

siendo $-1 \leq \rho_{jk} \leq 1$.

También podrán ser arreglados en una matriz de correlación muestral cuya diagonal principal esta formada por números unos (1) y será simétrica como la matriz de covarianza, por ser $\rho_{jk} = \rho_{kj}$.

$$R = \begin{bmatrix} 1 & \cdots & \rho_{1i} & \cdots & \rho_{1p} \\ \vdots & \ddots & \vdots & & \vdots \\ \rho_{i1} & \cdots & 1 & \cdots & \rho_{kp} \\ \vdots & & \vdots & \ddots & \vdots \\ \rho_{p1} & \cdots & \rho_{pk} & \cdots & 1 \end{bmatrix}$$

3.2 Cómo Generar componentes principales

Algebraicamente las componentes principales son combinaciones lineales particulares de las *p* variables aleatorias $x_1, x_2, ..., x_p$.

Geométricamente las combinaciones representan la selección de un nuevo sistema de coordenadas obtenidas por rotación del sistema original ($x_1, x_2, ..., x_p$).

Los nuevos ejes representan las direcciones con máxima variabilidad. La dirección de la primera componente principal, define el primer eje principal en el cual hay mayor variación estadística. El segundo componente define el eje principal de segunda máxima variación estadística sujeto a ser ortogonal al primero y así sucesivamente.

Como se verá en esta sección, las componentes principales dependen solamente de la matriz de covarianza (o correlación) de $x_1, x_2, ..., x_p$.

Sea la matriz X = ($x_1, x_2, ..., x_p$), con una matriz de covarianza $\sum$, con valores propios $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p$. Considere las combinaciones lineales

$$Y_1 = \alpha_1^{'} X = \alpha_{11} x_1 + \alpha_{12} x_2 + ... + \alpha_{1p} x_p .$$

$$Y_2 = \alpha_2^{'} X = \alpha_{21} x_1 + \alpha_{22} x_2 + ... + \alpha_{2p} x_p .$$

$$\vdots \qquad \vdots \qquad \qquad \vdots$$

$$Y_p = \alpha_p^{'} X = \alpha_{p1} x_1 + \alpha_{p2} x_2 + ... + \alpha_{pp} x_p .$$

donde α_i es un vector de p constantes $\alpha_{i1}, \alpha_{i2}, ..., \alpha_{ip} \quad i = 1,2,\ldots, p.$

La varianza de cada Y_i, está dada por

$$Var(Y_i) = \alpha_i^{'} \sum \alpha_i \, , \; i = 1,2,..., p. \qquad (3.1)$$

$$Cov(Y_i, Y_k) = \alpha_i^{'} \sum \alpha_k \, , \; i,k = 1,2,..., p .$$

Las componentes principales son las combinaciones lineales no correlacionadas $Y_1, Y_2, ..., Y_p$ cuyas variaciones en (3.1) son tan grandes como sea posible. La primer componente principal es la combinación lineal con máxima varianza, es decir máxima $Var(Y_1) = \alpha_1^{'} \sum \alpha_1$. Es claro que $Var(Y_1)$ puede ser incrementado multiplicando por cualquier constante al vector α_1. Para eliminar esta indeterminación se restringe los vectores de coeficientes a que tengan longitud unitaria (i.e. $\alpha_1^{'} \alpha_1 = 1$).

En resumen :

a) Primera componente principal: hallar la combinación lineal $\alpha_1^{'}x$ que maximiza la función $Var(Y_1) = \alpha_1^{'}\sum\alpha_1$ sujeto a $\alpha_1^{'}\alpha_1 = 1$.

b) Segunda componente principal: hallar la combinación lineal $\alpha_2^{'}x$ que maximiza la función $Var(Y_2) = \alpha_2^{'}\sum\alpha_2$ sujeto a $\alpha_2^{'}\alpha_2 = 1$ y $Cov\left[\alpha_1^{'}x, \alpha_2^{'}x\right] = 0$.

c) La i-ésima componente principal: hallar la combinación lineal $\alpha_i^{'}x$ que maximiza la función $Var(Y_i) = \alpha_i^{'}\sum\alpha_i$ sujeto a $\alpha_i^{'}\alpha_i = 1$ y $Cov\left[\alpha_i^{'}x, \alpha_k^{'}x\right] = 0$, $i < k$ $i = 1,2,...p.$

Teorema 3.1 Sea $X = (x_1, x_2, \ldots, x_p)$ una matriz de variables aleatorias con matriz de covarianza $\sum$, y sean los valores–vectores propios $(\lambda_1, \alpha_1), (\lambda_2, \alpha_2), \ldots, (\lambda_p, \alpha_p)$ donde $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p$ y sean $Y_1 = \alpha_1{'}X, Y_2 = \alpha_2{'}X, \ldots, Y_P = \alpha_P{'}X$, *entonces la i-ésima componente principal está dada por*

$$Y_i = \alpha_i^{'}X = \alpha_{i1}x_1 + \alpha_{i2}x_2 + ... + \alpha_{ip}x_p \ , i = 1,2,\ldots, p .$$

Con esta selección se tiene que

$$Var(Y_i) = \alpha_i^{'}\sum\alpha_i = \lambda_i , \ i = 1,2,..., p.$$

$$Cov(Y_i, Y_k) = \alpha_i^{'}\sum\alpha_k = 0 , \ i \neq k .$$

Demostración.

Como α_i es el vector propio correspondiente al valor propio λ_i de la matriz $\sum$, tenemos que $\sum\alpha_i = \lambda_i\alpha_i$ dado que $\alpha_i'\alpha_i = 1$ multiplicamos ambos lados de la igualdad por α_i' y obtenemos $Var(Y_i) = \alpha_i'\sum\alpha_i = \alpha_i'\lambda_i\alpha_i = \lambda_i\alpha_i'\alpha_i = \lambda_i$ para $i = 1,2,\ldots, p.$

Ahora, los vectores propios de $\sum$ son ortogonales si todos los valores propios $\lambda_1, \lambda_2, ..., \lambda_p$ son diferentes.

Si todos los valores propios no son diferentes, los vectores propios correspondientes a los valores propios comunes pueden ser elegidos para ser ortogonales. Por lo tanto, para cualquier par de vectores propios α_i y α_k , $\alpha_i{}'\alpha_k = 0$, para $i \neq k$.

Ya que

$$\alpha_k{}' \sum \alpha_k = \lambda_k$$

$$\sum \alpha_k = \lambda_k \alpha_k$$

multiplicando por $\alpha_i{}'$ ambos lados de la igualdad se obtiene:

$$Cov(Y_i, Y_k) = \alpha_i{}' \sum \alpha_k = \alpha_i{}' \lambda_k \alpha_k = \lambda_k \alpha_i{}' \alpha_k = \lambda_k 0 = 0 \text{, para cualquier } i \neq k.$$

Teorema 3.2 *Sea* $X = (x_1, x_2, \ldots, x_p)$ *una matriz de variables aleatorias con matriz de covarianza* $\sum$*, sean los valores–vectores propios* $(\lambda_1, \alpha_1), (\lambda_2, \alpha_2), \ldots, (\lambda_p, \alpha_p)$ *donde* $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$ *y sean* $Y_1 = \alpha_1{}' X, Y_2 = \alpha_2{}' X, \ldots, Y_P = \alpha_P{}' X$ *las componentes principales. Entonces*

$$s_{11} + s_{22} + \ldots + s_{pp} = \sum_{i=1}^{p} Var(X_i) = \lambda_1 + \lambda_2 + \ldots + \lambda_p = \sum_{i=1}^{p} Var(Y_i)$$

por lo tanto, la varianza total es igual a:

$$s_{11} + s_{22} + \ldots + s_{pp} = \lambda_1 + \lambda_2 + \ldots + \lambda_p.$$

Demostración.

Sea $\lambda = \begin{bmatrix} \lambda_1 & 0 & \ldots & 0 \\ 0 & \lambda_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \lambda_p \end{bmatrix}$ y $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_p)$ donde α_i es el vector propio correspondiente a λ_i.

$\sum_{i=1}^{p} Var(X_i) = \sum_{i=1}^{p} s_{ii} = tr(\sum)$ como $\sum \alpha = \lambda \alpha$ entonces $\sum = \lambda \alpha \alpha'$; es decir,

$$\sum = \left[\begin{array}{ccc} \alpha_{11}^2\lambda_1 + \alpha_{12}^2\lambda_2 + \ldots + \alpha_{1p}^2\lambda_p & \alpha_{11}\alpha_{21}\lambda_1 + \alpha_{12}\alpha_{22}\lambda_2 + \ldots + \alpha_{1p}\alpha_{2p}\lambda_p & \ldots \\ \alpha_{21}\alpha_{11}\lambda_1 + \alpha_{22}\alpha_{12}\lambda_2 + \ldots + \alpha_{2p}\alpha_{1p}\lambda_p & \alpha_{21}^2\lambda_1 + \alpha_{22}^2\lambda_2 + \ldots + \alpha_{2p}^2\lambda_p & \ldots \\ \vdots & \vdots & \vdots \\ \alpha_{p1}\alpha_{11}\lambda_1 + \alpha_{p2}\alpha_{12}\lambda_2 + \ldots + \alpha_{pp}\alpha_{1p}\lambda_p & \alpha_{p1}\alpha_{21}\lambda_1 + \alpha_{p2}\alpha_{22}\lambda_2 + \ldots + \alpha_{pp}\alpha_{2p}\lambda_p & \ldots \end{array}\right.$$

$$\left.\begin{array}{c} \alpha_{11}\alpha_{p1}\lambda_1 + \alpha_{12}\alpha_{p2}\lambda_2 + \ldots + \alpha_{1p}\alpha_{pp}\lambda \\ \alpha_{21}\alpha_{p1}\lambda_1 + \alpha_{22}\alpha_{p2}\lambda_2 + \ldots + \alpha_{2p}\alpha_{pp}\lambda \\ \vdots \\ \alpha_{p1}^2\lambda_1 + \alpha_{p2}^2\lambda_2 + \ldots + \alpha_{pp}^2\lambda_p \end{array}\right]$$

$$tr(\sum) = \left(\alpha_{11}^2\lambda_1 + \alpha_{12}^2\lambda_2 + \ldots + \alpha_{1p}^2\lambda_p\right) + \left(\alpha_{21}^2\lambda_1 + \alpha_{22}^2\lambda_2 + \ldots + \alpha_{2p}^2\lambda_p\right) + \ldots +$$
$$\left(\alpha_{p1}^2\lambda_1 + \alpha_{p2}^2\lambda_2 + \ldots + \alpha_{pp}^2\lambda_p\right)$$

$$tr(\sum) = \left(\alpha_{11}^2 + \alpha_{21}^2 + \ldots + \alpha_{p1}^2\right)\lambda_1 + \left(\alpha_{12}^2 + \alpha_{22}^2 + \ldots + \alpha_{p2}^2\right)\lambda_2 + \ldots + \left(\alpha_{1p}^2 + \alpha_{2p}^2 + \ldots + \alpha_{pp}^2\right)\lambda_p$$

dado que los vectores están normalizados; es decir, $\alpha_i'\alpha_i = 1$ se sigue que

$$tr(\sum) = \lambda_1 + \lambda_2 + \ldots + \lambda_p = \sum_{i=1}^{p} Var(Y_i) \text{ por lo tanto } \sum_{i=1}^{p} Var(X_i) = \sum_{i=1}^{p} Var(Y_i).$$

Teorema 3.3 *Si* $Y_1 = \alpha_1' X, Y_2 = \alpha_2' X, \ldots, Y_P = \alpha_P' X$ *son las componentes principales obtenidas de la matriz de covarianza* $\sum$, *entonces*

$$\rho_{y_i,k_k} = \frac{\alpha_{ik}\sqrt{\lambda_i}}{\sqrt{s_{kk}}}, \quad i,k = 1,2,\ldots,p \qquad (3.2)$$

son los coeficientes de correlación entre las componentes Y_i *y las variables* X_k.

Donde α_{ik} *mide la importancia de la k-ésima variable para la i-ésima componente principal, sin tener en cuenta las otras variables.*

Demostración.

Sea $a_k' = [0,\ldots,0,1,0,\ldots,0]$ talque $X_k = a_k' X$.

$Cov(X_k, Y_i) = Cov(a_k' X, \alpha_i' X) = a_k' \sum \alpha_i$ ya que $\sum \alpha_i = \lambda_i \alpha_i$ entonces

$Cov(X_k, Y_i) = a_k' \lambda_i \alpha_i = \lambda_i \alpha_{ik}$. Como $Var(Y_i) = \lambda_i$ y $Var(X_k) = s_{kk}$

tenemos que $\rho_{Y_i, X_k} = \dfrac{Cov(Y_i, X_k)}{\sqrt{Var(Y_i)}\sqrt{Var(X_k)}} = \dfrac{\lambda_i \alpha_{ik}}{\sqrt{\lambda_i}\sqrt{s_{kk}}} = \dfrac{\alpha_{ik}\sqrt{\lambda_i}}{\sqrt{s_{kk}}}$ para $i, k = 1,2,\ldots, p.$

Aunque las correlaciones de las variables con las componentes principales frecuentemente ayudan a interpretar las componentes, ellas miden solamente la contribución univariada de un X individual a un componente Y. Esto es, ellas indican la importancia de una X en el componente Y en presencia de las otras X. Por esta razón, algunos investigadores recomiendan que sólo los coeficientes α_{ik}, y no las correlaciones deben ser usados para interpretar los componentes principales.

Cada componente principal explica una proporción de la variabilidad total, esa proporción puede calcularse expresando la variación explicada por cada componente o por los *m* primeros componentes en relación a la variación total. Así se obtiene:

$\dfrac{\lambda_k}{\sum_{k=1}^{p}\lambda_k} 100\% =$ porcentaje de variación explicada por la *k*-ésima componente.

$\dfrac{\sum_{k=1}^{m}\lambda_k}{\sum_{k=1}^{p}\lambda_k} 100\% =$ porcentaje de variación explicada por las *m* primeras componentes.

Cuando se consideran todas las componentes, es decir si m = p, el porcentaje de la variación explicada es 100%.

Si la mayoría (por ejemplo 80-90%) de la varianza poblacional total puede ser atribuida a las primeras componentes principales, entonces estas componentes pueden “reemplazar” las p variables originales sin mucha pérdida de información.

Ejemplo 3.1 *Supóngase que las variables aleatorias* x_1, x_2 *y* x_3 *tienen la matriz de covarianza*

$$\sum = \begin{bmatrix} 1 & -2 & 0 \\ -2 & 5 & 0 \\ 0 & 0 & 2 \end{bmatrix},$$ *y sus respectivos valores son:* $\lambda_1 = 2, \lambda_2 = 3 + 2\sqrt{2}$ *y* $\lambda_3 = 3 - 2\sqrt{2}$

Puede observarse que la suma algebraica de las tres soluciones de la ecuación, es decir , la suma algebraica de los valores propios, es igual a la suma de los valores de la diagonal principal de la matriz original. ($\lambda_1 + \lambda_2 + \lambda_3 =$ *2+5,83+0,17 = 8 = 1+5+2).*

Esta propiedad de los valores propios es importante , ya que cuando se calculan a partir de la matriz de covarianza, la suma de los valores propios es igual a las suma de las varianzas de las variables incluidas en la matriz, o sea es la variación total (teorema 3.2). Los vectores propios correspondientes a los valores propios dados anteriormente son: $\alpha_1 = (0,0,1),\ \alpha_2 = (0.383,-0.924,0),\ \alpha_3 = (0.924,0.383,0)$.

Por lo tanto , los componentes principales son:

$$Y_1 = \alpha_1' X = x_3$$
$$Y_2 = \alpha_2' X = 0.383x_1 - 0.924x_2$$
$$Y_3 = \alpha_3' X = 0.924x_1 + 0.383x_2$$

la variable x_3 *es una de las componentes principales, ya que no está correlacionada con las otras dos variables. Además se observa que*

$$\begin{aligned} Var(Y_2) &= Var(0.383x_1 - 0.924x_2) \\ &= (0.383)^2 Var(x_1) + (-0.924)^2 Var(x_2) + 2(0.383)(-0.924)Cov(x_1, x_2) \\ &= 0.147(1) + 0.854(5) - 0.708(-2) \\ &= 5.83 = \lambda_2 \end{aligned}$$

y $\quad Cov(Y_1, Y_2) = Cov(x_3, 0.383x_1 - 0.924x_2)$

$$= 0.383Cov(x_3, x_1) - 0.924Cov(x_3, x_2)$$

$$= 0.383(0) - 0.924(0)$$

$$= 0$$

fácilmente se puede ver que:

$$s_{11} + s_{22} + s_{33} = 1 + 5 + 2 = 8 = 2 + 5.83 + 0.17 = \lambda_1 + \lambda_2 + \lambda_3$$

el porcentaje de la varianza total explicada por la primer componente principal es

$$\frac{\lambda_1}{\lambda_1 + \lambda_2 + \lambda_3} \times 100 = \frac{2}{2 + 5.83 + 0.17} \times 100 = 25\%$$

el porcentaje de las dos primeras componentes reúnen el $\frac{2 + 5.83}{8} \times 100 = 98\%$ *de la varianza poblacional.*

En este caso las dos primeras componentes Y_1 *y* Y_2 *podrían reemplazar la tres variables originales con una pequeña pérdida de información .*

Ahora, si se usa (3.2)*, se obtiene*

$$\rho_{y_2, x_1} = \frac{\alpha_{21}\sqrt{\lambda_2}}{\sqrt{s_{11}}} = \frac{0.383\sqrt{5.83}}{\sqrt{1}} = 0.925$$

$$\rho_{y_2, x_2} = \frac{\alpha_{22}\sqrt{\lambda_2}}{\sqrt{s_{22}}} = \frac{-0.924\sqrt{5.83}}{\sqrt{5}} = -0.998$$

note que la variable x_2*, con coeficiente –0.924, recibe el peso más grande en el componente* Y_2.

Además x_2 *tiene la correlación más grande (en valor absoluto) con* Y_2*. La correlación de* x_1*, con* Y_2 *es 0.925, casi tan grande como* x_2*, indica que las variables tienen casi la misma importancia para la segunda componente principal. Los tamaños relativos de los coeficientes de* x_1 *y* x_2 *sugiere, sin embargo, que* x_2 *contribuye más a la determinación de*

Y_2 que x_1. En este caso, ambos coeficientes son razonablemente grandes y ellos tienen signos opuestos, pudiéndose argumentar que ambas variables ayudan en la interpretación de Y_2.

Finalmente

$$\rho_{y_1,x_1} = \frac{\alpha_{11}\sqrt{\lambda_1}}{\sqrt{s_{11}}} = \frac{0\sqrt{2}}{\sqrt{1}} = 0 \;, \qquad \rho_{y_1,x_2} = \frac{\alpha_{12}\sqrt{\lambda_1}}{\sqrt{s_{22}}} = \frac{0\sqrt{2}}{\sqrt{5}} = 0$$

$$\rho_{y_1,x_3} = \frac{\alpha_{13}\sqrt{\lambda_1}}{\sqrt{s_{33}}} = \frac{1\sqrt{2}}{\sqrt{2}} = 1$$

las correlaciones remanentes pueden ser " desechadas" , ya que el tercer componentes no es importante.

3.3 Uso de la matriz de correlación

Hasta ahora se ha utilizado siempre la matriz de covarianza $\sum$ de los datos originales presentada en la definición 3.7; pero sin embargo puede hacerse uso de la matriz de correlación, por eso daremos algunas propiedades y ventajas para el uso de esta matriz.

Los valores de la diagonal principal de R (la matriz de correlación) son números unos (1), ya que las nuevas variables estandarizadas poseen varianzas unitaria. Esto significa que en el conjunto de datos a partir del cual se generarán las componentes principales se otorga la misma importancia a todas las variables observadas. Esta situación puede o no ser deseable, pero importa destacar que el uso de la matriz de correlación implica una ponderación de las variables originales, otorgándoles a cada una la misma importancia, independientemente de los valores relativos de sus varianzas.

Cuando se utiliza la matriz R, cambian algunas de las propiedades de los valores y vectores propios generados.

Cabe destacar las siguientes propiedades :

- $\sum_{k=1}^{p} Var(Y_k) = trR = p$

ya que la matriz R tiene números 1 en la diagonal principal y es de dimensión p.

Esta característica de los valores y vectores propios determinan que el análisis por componentes principales sea sensible a los cambios de escala. Por este motivo, es muy importante examinar cuidadosamente los datos originales en sus promedios, sus varianzas y covarianzas a fin de poder decidir que tipo de matriz conviene utilizar , qué cambio de escala puede introducirse antes de encontrar $\sum$ o R y qué interpretación deberá darse a las componentes encontradas.

Un mayor argumento para usar matrices de correlación, en lugar de matrices de covarianza, para definir componentes principales es que el resultado de los análisis para diferentes conjuntos de variables aleatorias, son directamente más comparables que por análisis basados en matrices de covarianza.

La gran desventaja del análisis por componentes principales basado en matrices de covarianza es la sensibilidad de las componentes principales para las unidades de medida usadas por cada elemento de X. Si hay gran diferencia entre las varianzas de los elementos de X, entonces esas variables cuya varianza son las más grandes tenderán a dominar las primeras pocas componentes principales.
Eso podría ser enteramente propio si todos los elementos de X son medidos en las mismas unidades, por ejemplo, si todos los elementos de X son anatómicas medidas en un especie particular de animal, todo registrado en centímetros. Aún en tal ejemplo, los argumentos pueden ser representados por el uso de las matrices de correlación .

En la practica, a menudo sucede esto que los diferentes elementos de X son tipos completamente diferentes de medidas. Algunos podrían ser cantidades, pesos, temperaturas, calificación arbitraria en una escala de 5 puntos y así sucesivamente. En tal caso, la estructura de las componentes principales dependerán en la selección de las unidades de medida, como es ilustrado por el siguiente ejemplo artificial.
Supóngase que nosotros tenemos justo dos variables, X_1, X_2 y que X_1 puede igualmente ser en centímetros o en milímetros.

Las matrices de covarianza en los 2 casos son, respectivamente,

$$\sum\nolimits_1 = \begin{pmatrix} 80 & 44 \\ 44 & 80 \end{pmatrix} \quad \text{y} \quad \sum\nolimits_2 = \begin{pmatrix} 8000 & 440 \\ 440 & 80 \end{pmatrix}$$

La primera componente principal es $0.707X_1 + 0.707X_2$ para $\sum_1$ y $0.998X_1 + 0.055X_2$ para $\sum_2$, de manera que un cambio relativamente pequeño en una variable ha tenido el efecto de cambiar a una componente principal, la cual da igual peso a X_1 y X_2 como una componente principal la cual es casi enteramente dominada por X_1. Además, la primera componente principal da cuenta de un 77.5% de la varianza total para $\sum_1$, pero el 99.33% para $\sum_2$.

3.4 Selección del numero de componentes

Cuando se lleva a cabo un análisis de componentes principales, se necesita determinar la dimensión real del espacio en el que caen los datos, es decir, el número de componentes principales a retener. No hay una norma específica, en algunos caso se siguen los siguientes lineamientos:

1. la cantidad de varianza muestral total explicada y
2. los valores propios mayores a 1 o cuyos valores sean superiores al promedio.

Una ayuda visual útil para determinar el número apropiado de componentes principales es un gráfico donde se represente el porcentaje de variación explicada por cada componente (o la magnitud de los valores propios) en las ordenadas y los componentes en orden decreciente en las abscisas. La Figura 1 muestra una situación con 5 componentes principales. En este caso hay una disminución del primero al segundo componente y los últimos tres aportan muy poco; si se consideran los dos primeros componentes se explicará el 80% de la variación original, resultado no muy bajo.

Cabe señalar que los criterios de selección no son fijos y que dependerán de la interpretación que se haga de las componentes principales.

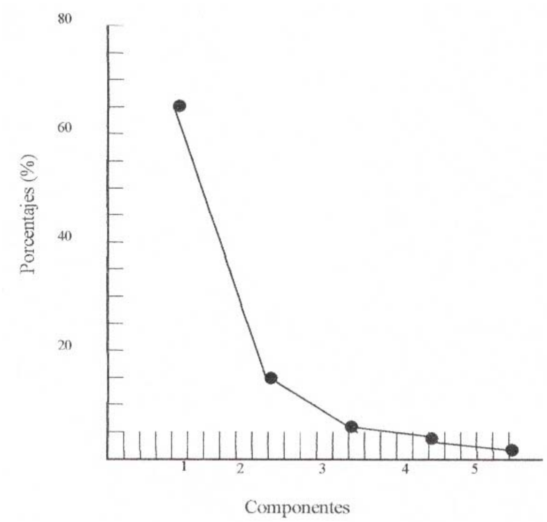


Fig. 3.1: Porcentaje (%) vs. Componentes

3.5 Interpretación de las componentes principales

Las ideas presentadas en esta sección son tomadas del texto “Análisis Factoriales simples y múltiples objetivos, métodos e interpretación” .

No importa cuan complicado sea un método de análisis de datos; una vez concluida la manipulación algebraica, es necesario interpretar correctamente los resultados obtenidos. Es por esta razón que esta sección da a conocer la forma adecuada de interpretar los datos.

Los factores se escogen en el orden decreciente de sus valores propios y pueden ser estudiados separadamente dos a dos con la ayuda de los planos factoriales. Primero se estudia la nube de variables y luego se considera la nube de individuos.

NUBE DE INDIVIDUOS

Interesarse por los individuos supone mirar la tabla como una yuxtaposición de filas. A cada individuo se le asocia una sucesión de *p* números. Bajo este punto de vista, se puede representar a cada individuo como un punto del espacio vectorial *p*-dimensional R^p, en el que cada dimensión representa a una variable.

El conjunto de individuos constituye la nube $N(n)$ cuyo centro de gravedad G coincide con el origen O de los ejes como consecuencia de haber sido previamente centrada; G representa al individuo medio. Estas notaciones están recopiladas en la figura 3.2.

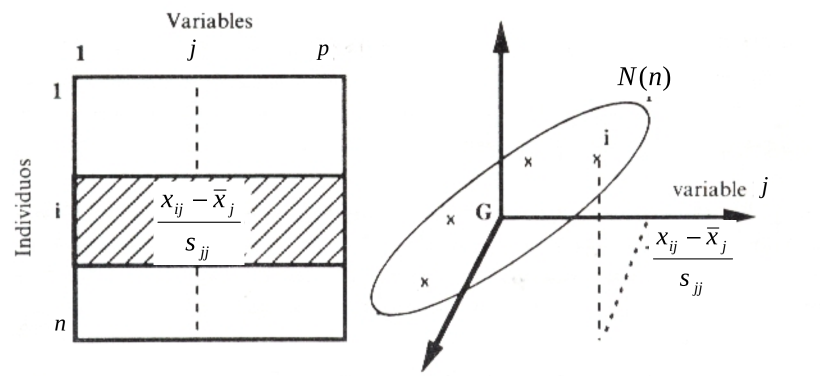


Figura 3.2: La tabla de datos y la nube de individuos asociada en el espacio $\Re^n$. Al haber sido previamente centrada, el origen de los ejes coincide con el centro de la gravedad de la nube.

En el espacio R^p, la noción de semejanza entre dos individuos introducida anteriormente coincide con la distancia euclidea usual.

Como tenemos que suponer que *p* es superior a 3, el estudio directo de la nube $N(n)$ resulta imposible, dada la limitación a tres dimensiones de nuestro sentido visual. De ahí nace precisamente el interés de los métodos factoriales, es general y en este caso particular del ACP, pues aportan imágenes planas que se aproximan lo más posible a una nube de puntos situada en un espacio de gran dimensión.

NUBE DE VARIABLES

Interesarse por las variables supone mirar la tabla como yuxtaposición de columnas. A cada variable se asocia una sucesión de *n* números. Bajo este punto de vista, se puede representar cada variable como un vector del espacio vectorial *n*-dimensional R^n, en el que cada dimensión está representando a un individuo: por ejemplo la variable *j* esta representada por el vector simbolizado igualmente por *j* y cuya componente *i-ésima es* $x_{ij} - \overline{x}_j$.

El conjunto de los puntos extremos de los vectores que representan variables constituyen la nube $N(p)$. Estas nociones están representadas en la figura 3.3.

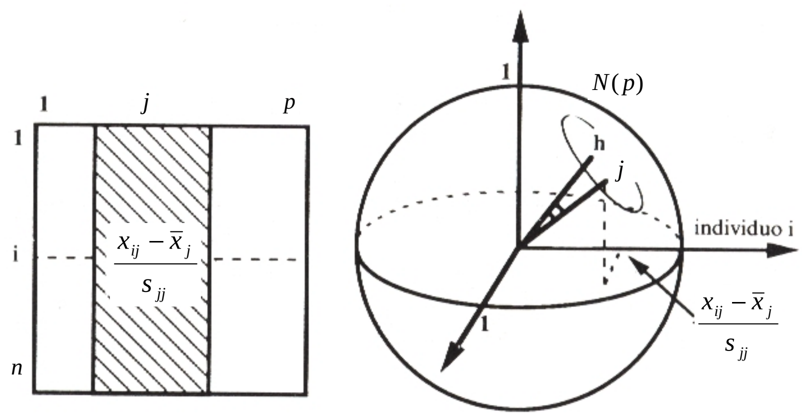


Figura 3.3: La tabla de datos y la nube de variables asociada en el espacio $\Re^n$

La opción de distancia en R^n consiste en afectar a cada dimensión de un coeficiente igual al peso de cada individuo en la nube $N(n)$ de R^p. En el caso general en que los pesos coinciden, la distancia utilizada sólo difiere de la euclidea usual en el coeficiente $1/n$. Con esta distancia, los vectores que representan las variables centradas tienen las siguientes propiedades:

1. Cada vector, que es representación de una variable, tiene como norma la unidad. Así:

$$\|\text{var}iablej\|^2 = \sum_{i=1}^{n} \frac{(x_{ij} - X_j)^2}{n} = 1$$

por ello la nube $N(p)$ está repartida sobre una esfera de radio unidad (se dice también hiperesfera para recordar que R^n es dimensión superior a 3).

2. El coseno del ángulo que forman los vectores que están representando a las dos variables *j* y *k* coinciden con el coeficiente de correlación entre ambas. Así:

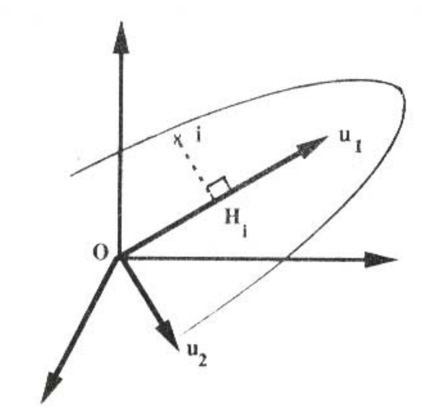
$$Cos\theta = \sum \frac{(x_{ij} - \overline{X}_j)(x_{ik} - \overline{X}_k)}{\sqrt{s_{jj}}\sqrt{s_{kk}}} = \rho_{jk}$$

La interpretación de un coeficiente de correlación como un coseno es una propiedad muy importante porque le da un apoyo geométrico y consiguientemente visual. Esta propiedad requiere el centrado previo, justificando así esta propiedad. Lo que nos interesa son las variables, es decir, mas como vectores que como puntos.
Al ser la longitud de los vectores que representan a las variables igual a la unidad, la coordenada de la proyección de una variable sobre otra se puede tomar como medida del mutuo coeficiente de correlación.

Ajuste de la nube de los individuos

El objetivo es proporcionar representaciones planas aproximadas de la nube $N(n)$ situada en el espacio R^p (figura 1). Prácticamente se busca una sucesión $\{u_k, k = 1,2,\ldots,p\}$ de direcciones privilegiadas de R^p llamadas ejes factoriales, que tomados dos a dos, determinen planos factoriales sobre los que se proyecte la nube $N(n)$. Cada dirección u_s hace máximo la inercia respecto del origen O (que coincide con el centro de gravedad G al haber centrado las variables) de la proyección de $N(n)$ sobre u_s.

Como buscamos una sucesión de direcciones, imponemos a cada nueva dirección la condición de ser ortogonal a las ya halladas (figura 3.4). Se puede demostrar que el plano generado por los dos primeros ejes u_1 y u_2 hace máxima la inercia proyectada sobre este plano. Sucede lo mismo para el subespacio generado por los tres primeros ejes, etc.



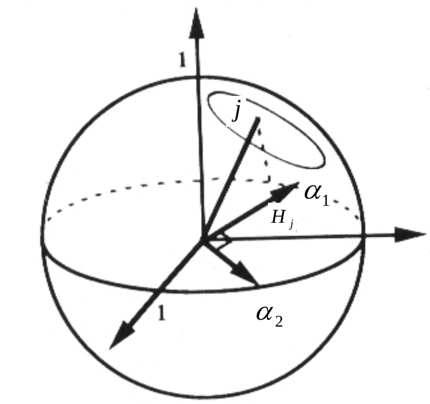
El individuo i se proyecta sobre u_1 *en* H_i

Figura 3.4: El ajuste de la nube de los individuos.

Debido al centrado, el criterio (máxima inercia respecto del centro de gravedad G) permite interpretar los ejes factoriales como direcciones de máximo alargamiento de la nube $N(p)$.

Ajuste de la nube de variables

Para obtener una sucesión de variables $\{v_k, k = 1,2,\ldots, p\}$ y una representación aproximada de las correlaciones entre las variables, el ACP aplica a la nube $N(p)$ de las variables el mismo proceso que a la nube de individuos (figura 3.5)



H_j : proyección sobre α_1 del punto que representa la variable j.

Figura 3.5: El ajuste de la nube de las variables.

El criterio (máxima inercia proyectada) a satisfacer en la elección de ejes es exactamente el mismo que para la nube de individuos. Pero adquiere aquí distintos significado a consecuencia de que esta nube no está centrada (su centro de gravedad no está en el origen) y de que todos sus puntos se hallan situados sobre la esfera unitaria: son los ángulos entre los vectores que representan las variables los que resultan poco deformados por las proyecciones y no las distancias entre los puntos de la nube.

En efecto, el plano $(\alpha_1.,\alpha_2)$ que maximiza la inercia respecto al origen de la nube proyectada, hace máxima la suma de los cosenos cuadrados de los ángulos entre los vectores y su proyección: ajusta los vectores y por ello deforma lo menos posible sus ángulos.

Relación de transición

Se llama relaciones de transición entre los factores F_k y Y_k a la expresión algebraica de las propiedades que se ilustran en las figuras 3.6 y 3.7.

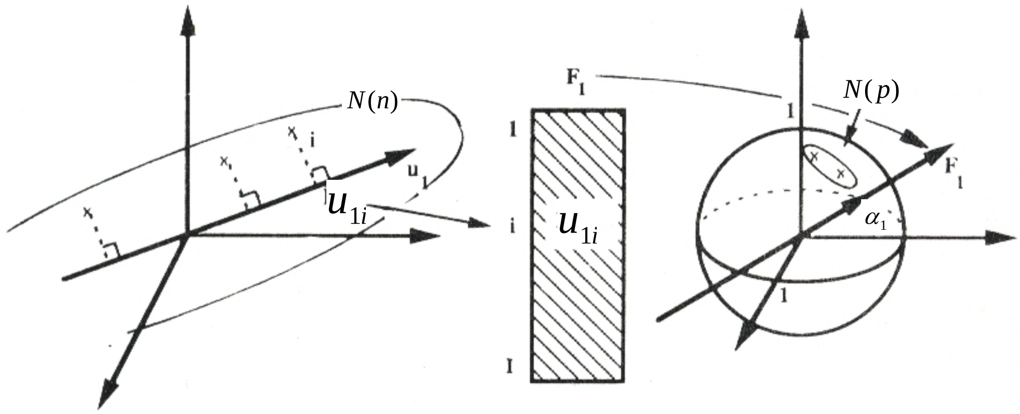


Figura 3.6: Las coordenadas *N(n)* sobre u_1 (primer eje factorial de *N(n)*) constituyen el primer factor sobre los individuos (denotados por F_1). El vector F_1 en $\Re^n$ es colineal a α_1 (primer eje factorial de *N(p)*).

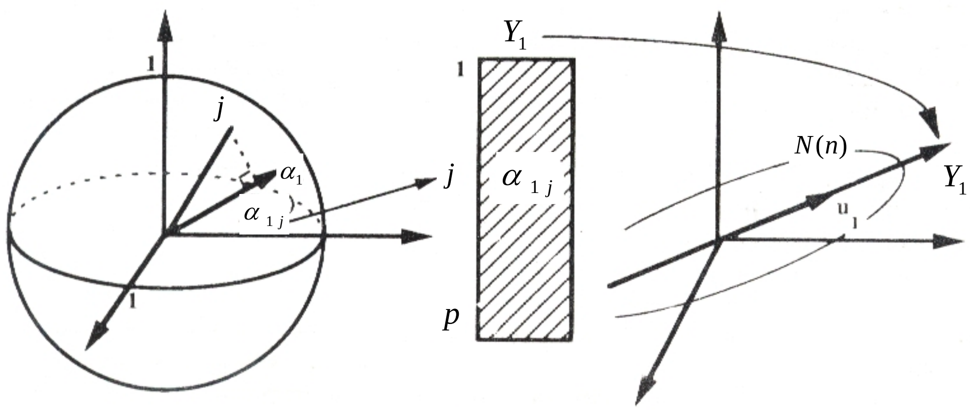


Figura 3.7: Las coordenadas *N(p)* sobre α_1 (primer eje factorial de *N(p)*) constituyen el primer factor (denotados por Y_1). El vector Y_1 en $\Re^p$ es colineal a u_1 (primer eje factorial de *N(n)*).

Se expresa como sigue, llamando λ_k a la inercia proyectada de $N(n)$ (o de $N(p)$) sobre el eje :

$$U_{ki} = \frac{1}{\sqrt{\lambda_k}} X'\alpha_{kj} \qquad \text{y} \qquad \alpha_{ki} = \frac{1}{\sqrt{\lambda_k}} XU_{kj}$$

donde α_{kj} mide la importancia de la *j*-ésima variable en la *k*-ésima componente principal y U_{ki} mide la importancia del *i*-ésimo individuo en la *k*-ésima componente principal.

La primera relación expresa como la proyección U_{ki} de un individuo i, es una combinación lineal de las proyecciones α_{kj} de todas las variables. Así, si se contemplan simultáneamente los dos gráficos, un individuo quedara del lado de las variables para las que presenten valores fuertes, y del lado opuesto de aquellas en que presente valores débiles.

El gráfico de individuos es una representación aproximada de las distancias entre ellos. El de variables se puede considerar como un elemento explicativo de esta representación: dos individuos situados en un mismo extremo de un eje quedan cercanos por tener ambos generalmente valores fuertes en las variables situadas del mismo lado que ellos y generalmente valores débiles en las variables situadas en el lado opuesto.

Recíprocamente, el gráfico de individuos puede ser utilizado como ayuda a la interpretación del gráfico de variables: si dos variables están muy correlacionadas positivamente, resultan situadas al mismo lado del eje. Sobre el eje correspondiente de la nube de individuos, aquellos que tienen valores altos en ambas variables se situaran normalmente del mismo lado que ellas y los que tengan en ellas valores débiles se situaran normalmente del lado opuesto. Los individuos extremos en esas variables quedaran normalmente lejos del origen. Así son localizados con facilidad aquellos individuos particulares que están causando, ellos solos, correlaciones fuertes.

La necesidad de una interpretación conjunta de las representaciones de individuos y de variables lleva a algunos usuarios a superponer ambas. Es importante subrayar que la justificación de esta representación simultanea de individuos y variables es esencialmente pragmática: la representación de las variables ayuda a la interpretación de la de los individuos y recíprocamente.

No obstante, plantea el problema de representar sobre un mismo grafico puntos de diferentes naturaleza que tienen lugar en espacios diferentes. Esta dificultad no es meramente de principio: la presencia simultánea de individuos y de variables sobre un mismo plano genera proximidades entre individuos y variables que, a su vez, pueden sugerir ideas que no se verifican en los datos. Pero si se mantienen presentes las indicaciones que siguen se puede usar sin peligro la presentación simultanea en ACP:

- ❖ Las fórmulas de transición relaciona la coordenada de un individuo sobre un eje con el conjunto de las coordenadas de todas las variables en el eje se ese mismo rango. No obstante se puede interpretar la posición de un individuo en función de la de una variables (y recíprocamente)
- ❖ Las variables son, fundamentalmente vectores mejor que simples puntos. La importancia no radica en la proximidad entre un individuo y un conjunto de puntos que representan variables, sino en el alejamiento del individuo en la dirección de este conjunto de variables.

CALIDAD DE REPRESNTACION DEL PUNTO X_i (INDIVIDUO i) POR SU PROYECCION SOBRE EL EJE U_k

La calidad de representación del punto X_i viene dada por:

$$Cos^2{}_k(i) = \frac{U_{ki}^2}{d^2(X_i, g)}$$

Donde $d^2(X_i, g)$ es la distancia al cuadrado del punto al centro de gravedad y U_{ki} mide la importancia del *i*-ésimo individuo para la *k*-ésima componente.

Gracias a la ortogonalidad de los ejes factoriales la calidad de representación del punto; en el plano factorial viene dada por

$$\sum_{k=1}^{2} Cos^2{}_k(i), \text{ además } \sum_{k=1}^{p} Cos^2{}_k(i) = 1$$

CONTRIBUCION DEL PUNTO X_i A LA INERCIA EXPLICADA POR U_k

$$CTR_k(i) = \frac{P_i U_{ki}^2}{\lambda_k}$$

Esta expresión permite identificar los puntos X_i que contribuyen mas en la formación del eje U_k. Los puntos para los cuales esta contribución es fuerte son los que fijan la posición del eje.

4. APLICACIÓN DEL ANALISIS POR COMPONENTES PRINCIPALES EN LOS INDICADORES DE GESTION DE LAS UNIVERSIDADES PUBLICAS

La construcción de los indicadores surge como respuesta a la necesidad de presentar propuestas sobre temas fundamentales como las políticas de largo plazo para el sector, el fortalecimiento de la financiación de la educación pública por parte del Estado - como proyecto de desarrollo y de Nación -, y la consolidación de la autonomía universitaria a través de la existencia de métodos claros de autorregulación estatal y de rendición de cuentas a la sociedad por parte de las instituciones educativas.

Con estas premisas, el Sistema Universitario Estatal (SUE) encomendó a la Subcomisión Técnica[1] formular un modelo que diera cuenta de la especificidad de sus instituciones, y propusiera una manera de analizar la gestión universitaria a través de indicadores construidos colectivamente por un grupo de universidades representativas y luego validada por el conjunto de Universidades en el ámbito nacional.

En el transcurso del trabajo de la Subcomisión Técnica se ha evidenciado que la construcción de un modelo de indicadores para las universidades públicas está ligada a la formulación y el desarrollo de políticas sectoriales que estén en consonancia con el contexto contemporáneo de la educación superior y en las que se contemple el desarrollo institucional de las universidades en conjunto, de manera sistémica y atendiendo a premisas académicas en torno a la calidad de la educación.

Ello implica trascender la perspectiva eminentemente financiera, que centra su interés en medidas de desempeño ligadas a la distribución de transferencia del presupuesto nacional; por lo tanto a nombre de la equidad, se ha optado por la vía de la “redistribución” de recursos entre las instituciones del SUE.

La ley del Plan Nacional de Desarrollo plantea que un porcentaje del 4% de los recursos presupuestales aportados por la Nación a las Universidades Estatales en el año 2004 deben ser incorporados a la bolsa común para que después sean distribuidos a las universidades de acuerdo a su eficiencia, la cual es evaluada gracias a los indicadores de gestión.

Las variables que se seleccionaron permiten modelar a todas las Universidades y hacer análisis generales sobre el desempeño de las Universidades públicas, contando siempre con referentes internacionales para evitar los juicios de valor y para la determinación de

[1] La Subcomisión Técnica está integrada por los representantes de las oficinas de Planeación de las universidades Nacional de Colombia, de Antioquia, del Valle, Industrial de Santander y Pedagógica Nacional, Con el acompañamiento y apoyo del Instituto Colombiano para el Fomento de la Educación Superior , Icfes.

umbrales razonables para cada indicador. Cada variable incluida tiene una capacidad de explicación de lo que sucede, de los resultados que corresponden y corresponderían a determinada capacidad académica disponible. La estrategia estadística de evaluación es dada a través del análisis por componentes principales, el cual a partir de las variables construye compuestos ICAD e IR.

El ICAD, Índice de Capacidades Académicas Disponibles en un solo número, sintetizan el comportamiento de distintas variables e indicadores y ofrece una idea integral de las dimensiones de capacidad de las instituciones. Es una medida de tamaño y capacidad de talentos humanos y otros recursos implícitos en el desarrollo de la misión, sean estos económicos, organizacionales o de infraestructura.

Como en toda organización, las instituciones educativas y específicamente las universitarias, deben orientar el ejercicio de su misión de forma tal que exista una relación razonable entre las capacidades académicas de las cuales ella dispone y sus logros, sus resultados, su efecto en la sociedad. Una buena gestión debe asegurar el uso óptimo de los recursos para obtener más y mejores resultados manteniendo o mejorando los atributos que distinguen y construyen su calidad, su pertinencia y la manera como cada institución contribuye a la construcción de equidad en la sociedad.

INDICADORES

La Subcomisión Técnica del SUE ha definido 75 indicadores.
Para este estudio hemos tomado solo 14 indicadores los cuales consideramos significativos, además fueron los analizados por la Comisión de estadísticos que creo el Subcomité para el análisis de la información.

Los indicadores de capacidad que vamos a utilizar son:

Profesores: Número de docentes en tiempos completos equivalentes incluyendo cátedra, discriminados por nivel de formación.
Para el cálculo de este indicador se suma el número de Profesores Tiempo Completo, el número de Profesores de Medio Tiempo dividido en dos y el total de horas cátedra semanal dividas por cuarenta. El último sumando asume que cuarenta horas cátedra equivalen a un profesor de tiempo completo. Los profesores se ponderan de acuerdo a los puntos salariales que por títulos contempla el Decreto 1279 y se halla el número total equivalente a profesores con título de doctorado.

Recursos: Aportes del Estado ejecutados en el año e ingresos propios de la Universidad.
Para el cálculo de este indicador se incluyen los aportes del Estado, tanto para funcionamiento como para inversión (Vía Nación, Departamento, Distrito o Municipio), y los Recursos Propios de la institución, menos transferencias (Contraloría, ICFES, etc.), servicios a la deuda, cesantías retroactivas y presupuestos financiados con aportes de la

Nación destinados a colegios adscritos a las Universidades. No se incluyen los recursos obtenidos por vía estampilla.

Los indicadores de resultados que vamos a utilizar son:

Programas: Número de programas académicos ofrecidos por la institución.
Para el cálculo de este indicador es necesario desagregar la información por programas de pregrado, especialización, maestría, doctorado y programas tecnológicos, y por modalidades y áreas de conocimiento, con el fin de ponderar los diferentes programas siguiendo la metodología Australiana[2]. Además se toma el equivalente de que un programa a distancia es igual a 0.4 de un programa presencial.

GruposInv: Grupos de investigación reconocidos por COLCIENCIAS.
Para el cálculo de este indicador se toma el total de grupos de investigación de la institución reconocidos por Colciencias.

Acreditados: Programas académicos de pregrado con acreditación de calidad.
Para el cálculo de este indicador se toma el total de programas de pregrado con acreditación de calidad.

LibrosTex: Productividad y Calidad de la producción bibliográfica derivada de los procesos de docencia.
Para el cálculo de este indicador se toma el total del puntaje obtenido por los libros de textos elaborados por los docentes en la institución en el año.

Revistas: Número de revistas producidas por la institución indexadas en Colciencias.
Para el cálculo de este indicador se toma el número de revistas indexadas en Colciencias publicadas por la institución.

Primíparos: Número de matriculados por primera vez en primer curso en todos los niveles de formación modalidades de enseñanza y por áreas del conocimiento.
Para el cálculo de este indicador se adopta el número total de estudiantes nuevos efectivamente matriculados en primer nivel, en el periodo de referencias en programas curriculares conducentes a título. Se toman el total del año, sumando los matriculados en primer curso en los dos semestres, desagregados por nivel y modalidades de formación. Con el fin de ponderar los estudiantes siguiendo la metodología australiana y tomando el equivalente de que un programa a distancia es igual a 0.4 de un programa presencial.

Estudiantes: Población estudiantil total, discriminada por nivel de formación y modalidades de enseñanza.
Para el cálculo de este indicador es necesario desagregar la información por programas de pregrado, especialización, maestría, doctorado y programas tecnológicos, y por

[2] Este índice que pondera las diferentes carreras por los costos que implican, se ha validado con muy buenos resultados en diversos países. Por esta razón lo incorporó el Subcomité Técnico del SUE.

modalidades y áreas de conocimiento, con el fin de ponderar los estudiantes en los diferentes programas siguiendo la metodología Australiana y tomando el equivalente de que un programa a distancia es igual a 0.4 de un programa presencial.

Proyectos: Capacidad de los investigadores para desarrollar proyectos.
Para el cálculo de este indicador se toma el total de proyectos de investigación activos, aprobados interna y externamente, en ejecución en el año.

Artículos: Difusión de los resultados de la investigación en revistas nacionales y del extranjero indexadas.
Para el cálculo de este indicador se toma el puntaje total por artículos elaborados por investigadores de la Universidad publicados en revistas nacionales y del extranjero indexadas.

LibrosInv: Difusión de los resultados de la investigación en libros.
Para el cálculo de este indicador se toma el puntaje total por libros elaborados por investigadores de la universidad como producto de la investigación.

Graduados: Número de graduados en todos los niveles de formación, modalidades de enseñanza y áreas de conocimiento.
Para el cálculo de este indicador es necesario desagregar la información por programas de pregrado, especialización, maestría, doctorado y programas tecnológicos, y por modalidades y áreas de conocimiento, con el fin de ponderar los graduados en los diferentes programas siguiendo la metodología Australiana.

Extensión: Participación de los ingresos generados por extensión.
Para el cálculo de este indicador se toma los ingresos generados por extensión. La extensión comprende los programas de educación permanente, cursos, seminarios y demás programas destinados a la difusión de los conocimientos, al intercambio de experiencias, comprende además, las actividades de servicio tendientes a procurar el bienestar general y a satisfacción de las necesidades de la sociedad.

UNIVERSIDADES

En el país existen 32 universidades públicas de las cuales 15 son Nacionales y 17 son Territoriales.

Universidades Nacionales
La Universidad del Cesar, Córdoba, Amazonia, Mayor de Cundinamarca, Llanos, Militar, Pedagógica y Tecnológica de Colombia (UPTC), Surcolombiana, Pedagógica Nacional, Nacional de Colombia, Caldas, Cauca, Chocó, Tecnológica de Pereira y La Universidad del Pacifico.

Universidades Territoriales
La Universidad de Antioquia, Atlántico, Cartagena, Guajira, Magdalena, Sucre, Cundinamarca, Distrital, Universidad Industrial de Santander (UIS), Francisco de Paula Santander sede de Cúcuta (FPS-Cúcuta), Francisco de Paula Santander sede de Ocaña (FPS-Ocaña), Pamplona, Nariño, Quindío, Tolima, Valle, Naval.

Para efectos del análisis que aquí presentamos, excluimos dos universidades que tienen características que no las hacen comparables con las demás. Una es la Universidad del **Pacifico** recientemente fundada hasta el punto que todavía no posee egresados. La otra es la Universidad **Naval** que tiene una estructura distinta a las demás universidades: está adscrita al Ministerio de Defensa y forma profesionales militares con fines muy específicos asociados a las necesidades de ese ministerio.

Las variables consideradas y sus valores para cada una de las universidades se encuentran en la tabla 4.1.

Tabla 4.1. Indicadores de gestión (las variables Recursos y Extensión están dadas en millones de pesos)

Nº	Universidades	Recursos	Profesores	LibrosInv	LibrosTex	Artículos	Programas	Primíparos	Estudiantes
1	FPS-Ocaña	4492	38,35	0	0	0	19,475	1410,65	6463,6
2	Sucre	10395	79,851	45	45	12	20,4	1345,48	4776,82
3	Guajira	12161	113,463	0	20	0	27,635	2715,46	7022,595
4	Amazonia	13930	116,228	0	0	0	35,625	1961,88	6371,49
5	Mayor de Cundinamarca	15699	155,336	6,5	0	0	20,615	2063,24	6764,43
6	Llanos	17687	142,303	0	7,41	30	39,075	2189,1	7011,475
7	Cundinamarca	20516	200,641	0	0	0	46,275	4259,625	13649,725
8	Cesar	21174	236,918	0	20	0	38,6	4971,1	13343,9
9	Chocó	25429	162,873	0	0	0	44,32	2523,5	8591,48
10	FPS-Cúcuta	25729	243,124	0	0	0	60,445	5708,425	19426,96
11	Pamplona	27153	359,101	0	31	949	157,54	9439,665	21451,915
12	Tolima	29472	233,571	26,4	11,3	77,2	66,35	3950,32	18104
13	Surcolombiana	30109	235,814	0	0	0	69,325	2799,56	9050,64
14	Magdalena	32034	117,218	60	60	79,5	69,475	4128,365	11407,56
15	Cartagena	35829	242,554	80	0	128,5	152,47	4018,27	13002,86
16	Pedagógica Nacional	37515	187,252	279,88	48,32	38	62	2505,1	8429,1
17	Militar	39322	177,124	0	0	0	264,14	4888,02	12807,68
18	Quindío	39661	268,421	1	12	64	60,215	3912,58	11709,84
19	Nariño	43804	222,871	0	0	0	115,85	4904,5	14553,9
20	Caldas	48766	357,619	60,3	179,2	123	145,43	2272,645	15582,465
21	Tecnológica de Pereira	50834	332,782	161	322	210,4	83,1	5445	14230,15
22	Atlántico	59997	331,922	0	0	0	98,2	3001,7	17882,5
23	Cauca	60445	546,126	140	221	234	211,545	5590,015	19071,475
24	Córdoba	64385	196,243	0	38,3	6	52,4	8418,775	20173,15
25	Distrital	68463	538,15	29	5	212	169,3	12657,15	38048,95
26	UPTC	77147	675,352	0	4	76	123,45	7722,03	30203,87
27	UIS	86383	460,542	20,1	36,1	497,9	209,56	8198,42	32522,015
28	Valle	133601	727,677	0	5,8	45,2	470,69	14161,545	44095,355
29	Antioquia	214532	1333,04	519	920	2488	495,61	14959,875	56214,405
30	Nacional de Colombia	382045	1887,79	5	0	647	977,5	23710,5	92482,8

Nº	Universidades	Graduados	Revistas	GruposInv	Proyectos	Acreditados	Extensión
1	FPS-Ocaña	622,95	0	0	3	0	49,438
2	Sucre	968	0	0	38	0	159,22
3	Guajira	954,6	0	0	62	0	48,2
4	Amazonia	925,425	0	0	33	1	26
5	Mayor de Cundinamarca	1613,275	0	0	25	0	339
6	Llanos	624,95	0	1	17	0	950
7	Cundinamarca	1072,2	0	0	5	0	209
8	Cesar	911,4	0	1	15	0	125
9	Chocó	1283,1	0	3	31	0	169,172
10	FPS-Cúcuta	1817,875	0	1	25	0	1844,231
11	Pamplona	3981,675	0	3	35	0	4337,59
12	Tolima	4481,05	0	3	36	1	526,629
13	Surcolombiana	1466,2	0	1	55	0	2266,04
14	Magdalena	959,15	0	6	41	0	285,389
15	Cartagena	1869,125	0	2	77	1	2870,042
16	Pedagógica Nacional	601,1	0	11	59	0	2335,324
17	Militar	1934,1	0	0	38	2	1222
18	Quindío	3674,75	1	4	37	0	1722,627
19	Nariño	1962,675	1	4	36	2	880
20	Caldas	2874,075	0	10	89	6	1640,928
21	Tecnológica de Pereira	943,5	1	14	95	7	4443
22	Atlántico	2315,1	0	5	99	0	1002,843
23	Cauca	1804,3	0	34	42	4	1162,714
24	Córdoba	849,45	0	2	26	0	720,369
25	Distrital	4542,15	2	12	106	0	0
26	UPTC	4517,2	0	4	315	4	1598
27	UIS	4463,4	2	28	159	17	16333,8
28	Valle	5065,525	3	70	451	14	19124,4
29	Antioquia	6309,7	11	105	923	29	28533,345
30	Nacional de Colombia	11740,8	15	101	732	0	72418

4.1 Análisis Univariado

En esta sección se ha realizado el análisis univariado para las 14 variables de la Tabla 4.1.
La información que se presenta fue obtenida utilizando el paquete estadístico STATGRAPHICS. En las gráficas las universidades se encuentran ubicadas sobre el eje X (de acuerdo a la numeración establecida en la Tabla 4.1 que responde al orden inducido por la variable Recursos) y el valor correspondiente de las universidades sobre estas variables se encuentran en el eje Y.

1. VARIABLE: RECURSOS

En la Tabla 4.2 aparecen los valores de esta variable para cada una de las universidades.

Tabla 4.2. Recursos (millones de pesos)

Nº	Universidades	Recursos
1	FPS-Ocaña	4492
2	Sucre	10395
3	Guajira	12161
4	Amazonia	13930
5	Mayor de Cundinamarca	15699
6	Llanos	17687
7	Cundinamarca	20516
8	Cesar	21174
9	Chocó	25429
10	FPS-Cúcuta	25729
11	Pamplona	27153
12	Tolima	29472
13	Surcolombiana	30109
14	Magdalena	32034
15	Cartagena	35829
16	Pedagógica Nacional	37515
17	Militar	39322
18	Quindío	39661
19	Nariño	43804
20	Caldas	48766
21	Tecnológica de Pereira	50834
22	Atlántico	59997
23	Cauca	60445
24	Córdoba	64385
25	Distrital	68463
26	UPTC	77147
27	UIS	86383
28	Valle	133601
29	Antioquia	214532
30	Nacional de Colombia	382045

Los estadísticos fundamentales asociados a esta variable se presentan en la Tabla 4.3.

Tabla 4.3. Estadísticos básicos de la variable Recursos

Identificación de la variable	Media	Desviación típica	Mínimo	Máximo	Mediana
Recursos	57623,6	72836,3	4492	382045	36672

Como se observa en la tabla 4.3 y lo ratifica los gráficos 4.1 y 4.2, la distribución de los recursos es muy asimétrica sesgada a la derecha (la media es mucho más grande que la mediana). Es decir, existen unas pocas universidades que tienen presupuestos mucho más grandes que el resto. El diagrama de caja que se muestra en el gráfico 4.2 evidencia tres valores atípicos que corresponden con las universidades Nacional, Antioquia y Valle.

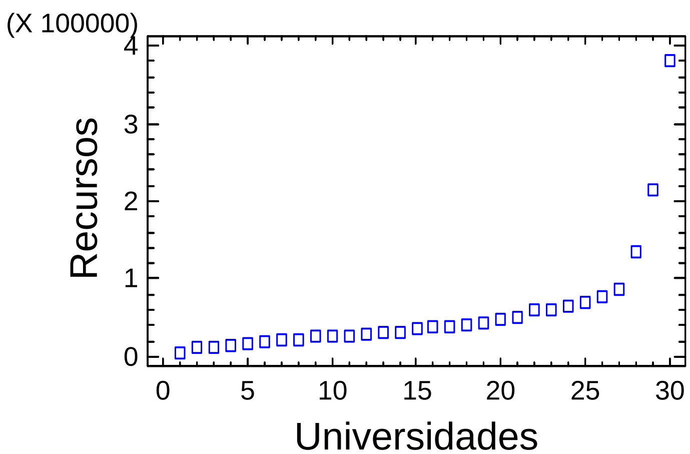


Gráfico 4.1. Recursos

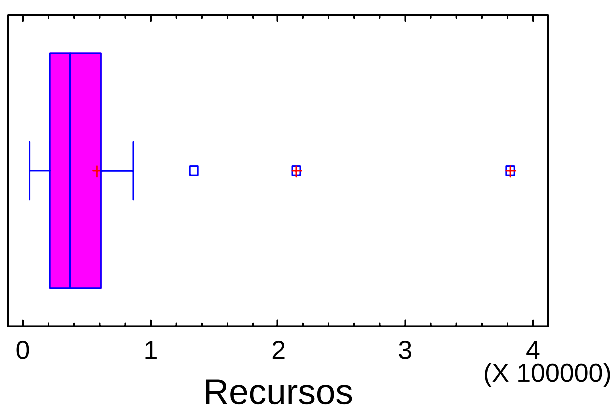


Gráfico 4.2. Diagrama de Caja de la variable Recursos

Incluso entre estas universidades grandes existen apreciables diferencias. La Nacional, por ejemplo, posee un presupuesto que equivale al 178% del presupuesto de la universidad de Antioquia, y ésta, a su vez, tiene un presupuesto que equivale al 160% del presupuesto de la universidad del Valle.

Con ánimos comparativos sencillos dividimos las universidades en varios grupos teniendo en cuenta su tamaño de acuerdo a la cantidad de recursos que poseen. Esta división, juntando las universidades de presupuestos similares, nos permitirá realizar un mejor análisis de las variables de resultados respondiendo al principio básico: "A mayor presupuesto mejores resultados".

Proponemos seis grupos que se caracterizan porque la relación entre los valores máximo y mínimo de los recursos de las universidades en cada grupo es cercano a dos. Los grupos y las universidades que los componen son los siguientes:

Universidad muy Grande
Nacional de Colombia: 382.045

Universidades Grandes
Antioquia: 214.532; Valle: 133.601.

Universidades Medias
UIS, UPTC, Distrital, Córdoba, Cauca, Atlántico, Tecnológica de Pereira, Caldas, Nariño.
Valor mínimo: 43.804
Valor máximo: 86.383

Universidades Pequeñas
Quindío, Militar, Pedagógica Nacional, Cartagena, Magdalena, Surcolombiana, Tolima, Pamplona, FPS-Cúcuta, Chocó.
Valor mínimo: 25.429
Valor máximo: 39.661

Universidades muy Pequeñas
Cesar, Cundinamarca, Llanos, Mayor de Cundinamarca, Amazonia, Guajira, Sucre.
Valor mínimo: 10.395
Valor máximo: 21.174

Universidad muy muy Pequeña
FPS-Ocaña: 4.492

La FPS- Ocaña que surgió como una sede de la FPS-Cúcuta adquirió el estatus de universidad para efectos fiscales, lo que hace que su presupuesto sea tan pequeño comparado con cualquier otra universidad del país.

2. VARIABLE: PROFESORES

En la Tabla 4.4 aparecen los valores de esta variable para cada una de las universidades.

Tabla 4.4. Profesores

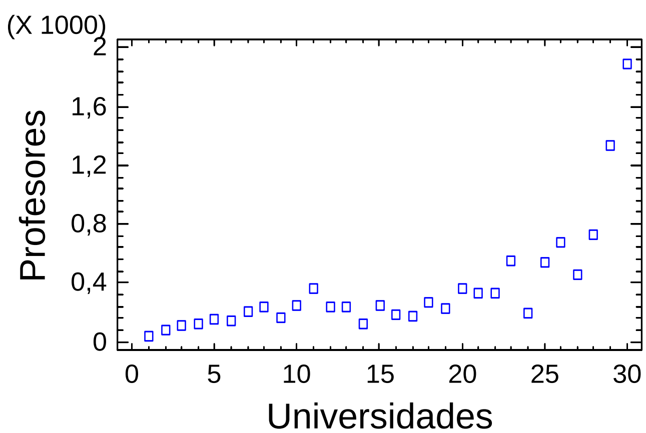
Nº	Universidades	Profesores
Universidad muy muy pequeña		
1	FPS-Ocaña	38,35
Universidades muy pequeñas		
2	Sucre	79,851
3	Guajira	113,463
4	Amazonia	116,228
5	Mayor de Cundinamarca	155,336
6	Llanos	142,303
7	Cundinamarca	200,641
8	Cesar	236,918
Universidades pequeñas		
9	Chocó	162,873
10	FPS-Cúcuta	243,124
11	Pamplona	359,101
12	Tolima	233,571
13	Surcolombiana	235,814
14	Magdalena	117,218
15	Cartagena	242,554
16	Pedagógica Nacional	187,252
17	Militar	177,124
18	Quindío	268,421
Universidades medianas		
19	Nariño	222,871
20	Caldas	357,619
21	Tecnológica de Pereira	332,782
22	Atlántico	331,922
23	Cauca	546,126
24	Córdoba	196,243
25	Distrital	538,15
26	UPTC	675,352
27	UIS	460,542
Universidades grandes		
28	Valle	727,677
29	Antioquia	1333,04
Universidad muy grande		
30	Nacional de Colombia	1887,79

Los estadísticos fundamentales asociados a esta variable se presentan en la Tabla 4.5.

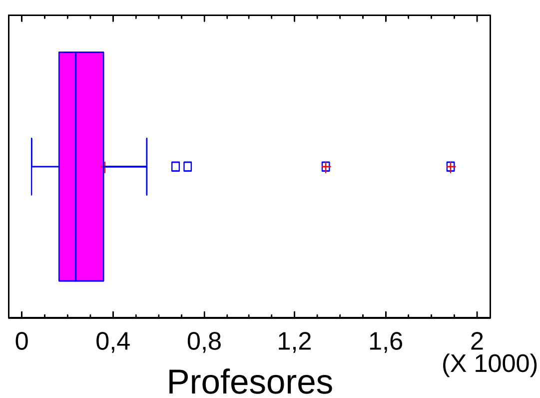
Tabla 4.5. Estadísticos básicos de la variable Profesores

Identificación de la variable	Media	Desviación típica	Mínimo	Máximo	Mediana	Correlación entre Profesores y Recursos
Profesores	364	378,6	38,4	1887,8	236,4	0,97

Como se observa en la Tabla 4.5 y lo ratifica los gráficos 4.3 y 4.4, la distribución del número de profesores es muy asimétrica sesgada a la derecha (la media es mucho más grande que la mediana). Es decir, existen unas pocas universidades que tienen el número de profesores mucho más grande que el resto. El diagrama de caja que se muestra en el gráfico 4.4 evidencia cuatro valores atípicos que corresponden con las universidades Nacional, Antioquia, Valle y la UPTC.



Grafica 4.3 Profesores



Gráfica 4.4 Diagrama de Caja de la variable Profesores

En la gráfica 4.3 se observa una estructura similar a la variable Recursos, excepto para dos grupos de Universidades. El primero conformado por la UIS, Córdoba, Magdalena y Chocó que poseen un número bajo de profesores que no es proporcional a sus recursos. En el caso de la UIS se conoce que este número es el reflejo de una política institucional que ha fomentado el crecimiento de los profesores de cátedra. El segundo grupo conformado por las universidades de Pamplona y Cauca que poseen un alto número de profesores que no responde a la proporción de las universidades de los grupos a los cuales pertenecen.

Como caso especial está la UPTC cuyo número de profesores es un valor atípico de acuerdo al Gráfico 4.4. Donde se observa que tiene casi el mismo número de profesores de la universidad del Valle que posee más del doble de sus recursos. Conjeturamos que puede ser debido a una situación contraria a la de la UIS, esto es, que la política oficial de esta universidad es la vinculación de profesores de planta.

3. VARIABLE: GRUPOSINV

En la Tabla 4.6 aparecen los valores de esta variable para cada una de las universidades.

Tabla 4.6. GruposInv

Nº	Universidades	GruposInv
Universidad muy muy pequeña		
1	FPS-Ocaña	0
Universidades muy pequeñas		
2	Sucre	0
3	Guajira	0
4	Amazonia	0
5	Mayor de Cundinamarca	0
6	Llanos	1
7	Cundinamarca	0
8	Cesar	1
Universidades pequeñas		
9	Chocó	3
10	FPS-Cúcuta	1
11	Pamplona	3
12	Tolima	3
13	Surcolombiana	1
14	Magdalena	6
15	Cartagena	2
16	Pedagógica Nacional	11
17	Militar	0
18	Quindío	4

Tabla 4.6. Continuación GruposInv

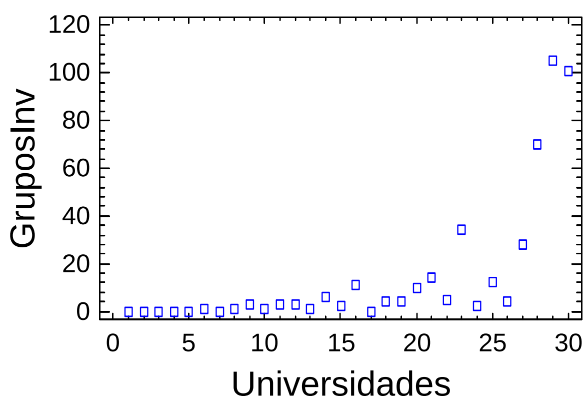
Universidades medianas		
19	Nariño	4
20	Caldas	10
21	Tecnológica de Pereira	14
22	Atlántico	5
23	Cauca	34
24	Córdoba	2
25	Distrital	12
26	UPTC	4
27	UIS	28
Universidades grandes		
28	Valle	70
29	Antioquia	105
Universidad muy grande		
30	Nacional de Colombia	101

Los estadísticos fundamentales asociados a esta variable se presentan en la Tabla 4.7.

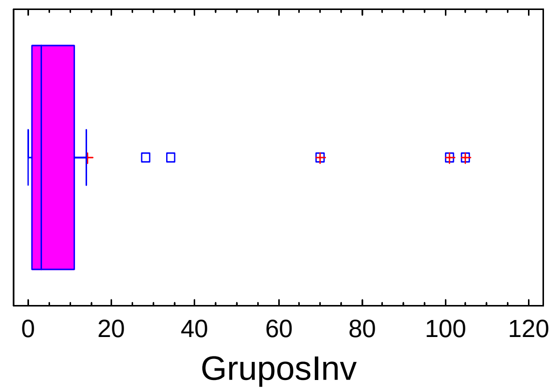
Tabla 4.7. Estadísticos básicos de la variable GruposInv

Identificación de la variable	Media	Desviación típica	Mínimo	Máximo	mediana	Correlación entre GruposInv y Recursos
GruposInv	14,2	27,5	0	105	3	0,91

Como se observa en la Tabla 4.7 y lo ratifica los gráficos 4.5 y 4.6, la distribución de los GruposInv es muy asimétrica sesgada a la derecha. Es decir, existen una pocas universidades que tienen mucho más grupos de investigación que el resto. El diagrama de caja que se muestra en el gráfico 4.6 evidencia cinco valores atípicos que corresponden con las universidades Antioquia, Nacional, Valle, Cauca y la UIS.



Gráfica 4.5. GruposInv



Gráfica 4.6 Diagrama de Caja de la variable GruposInv

Como se puede observar en las gráficas el grupo de las universidades grandes y muy grande obtienen valores altos; pero cabe destacar el caso de la Nacional, que posee menos grupos de investigación que la de Antioquia recibiendo mucho más presupuesto que esta.
En el grupo de las universidades medianas y pequeñas se destacan las universidades Cauca (34), UIS (28), Pedagógica (11) y Magdalena (6) por tener valores altos que no corresponden a la proporción de las universidades de los grupos a los cuales pertenecen; los valores de las universidades Cauca y UIS son tan extraños que aparecen en la grafica 4.6 como valores atípicos.
Para el grupos de las universidades muy pequeñas y muy muy pequeña las proporciones se mantienen.

4. VARIABLE: REVISTAS

En la Tabla 4.8 aparecen los valores de esta variable para cada una de las universidades.

Tabla 4.8. Revistas

Nº	Universidades	Revistas
Universidad muy muy pequeña		
1	FPS-Ocaña	0
Universidades muy pequeñas		
2	Sucre	0
3	Guajira	0
4	Amazonia	0
5	Mayor de Cundinamarca	0
6	Llanos	0
7	Cundinamarca	0
8	Cesar	0

Tabla 4.8. Continuación Revistas

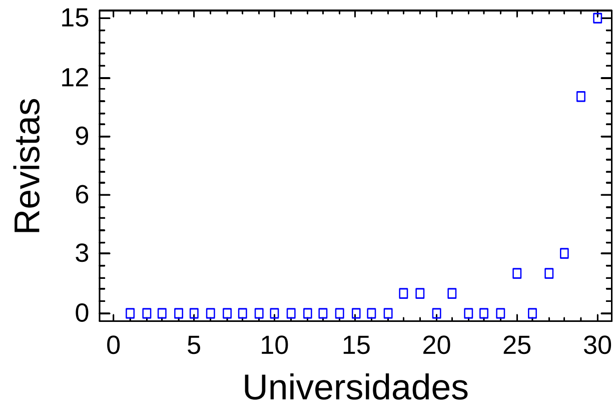
Universidades pequeñas		
9	Chocó	0
10	FPS-Cúcuta	0
11	Pamplona	0
12	Tolima	0
13	Surcolombiana	0
14	Magdalena	0
15	Cartagena	0
16	Pedagógica Nacional	0
17	Militar	0
18	Quindío	1
Universidades medianas		
19	Nariño	1
20	Caldas	0
21	Tecnológica de Pereira	1
22	Atlántico	0
23	Cauca	0
24	Córdoba	0
25	Distrital	2
26	UPTC	0
27	UIS	2
Universidades grandes		
28	Valle	3
29	Antioquia	11
Universidad muy grande		
30	Nacional de Colombia	15

Los estadísticos fundamentales asociados a esta variable se presentan en la Tabla 4.9.

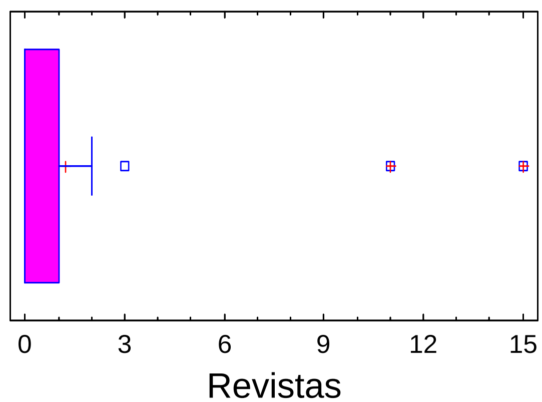
Tabla 4.9. Estadísticos básicos de la variable Revistas

Identificación de la variable	Media	Desviación típica	Mínimo	Máximo	mediana	Correlación entre Revistas y Recursos
Revistas	1,2	3,3	0	15	0	0,96

Como se observa en la Tabla 4.9 y lo ratifican los gráficos 4.7 y 4.8, la distribución de las Revistas es muy asimétrica sesgada a la derecha. Es decir, existen unas pocas universidades que tienen mucho más número de Revistas que el resto. El diagrama de caja que se muestra en la gráfica 4.8 evidencia tres valores atípicos que corresponden a las universidades Nacional, Antioquia y Valle.



Gráfica 4.7. Revistas



Gráfica 4.8. Diagrama de Caja de la variable Revistas

En cuanto a los grupos formado en la variable Recursos la proporción del número de Revistas se mantiene; sin embargo podría decirse que el número de revistas se centra en las universidades Nacional y Antioquia que poseen el 72% de la producción total de Revistas de todas las universidades.

Además el 73% de las universidades públicas del país no poseen revistas indexadas por Colciencias, lo cual nos deja un gran interrogante de lo que sucede con estas universidades.

5. VARIABLE: PROYECTOS

En la Tabla 4.10 aparecen los valores de esta variable para cada una de las universidades.

Tabla 4.10. Proyectos

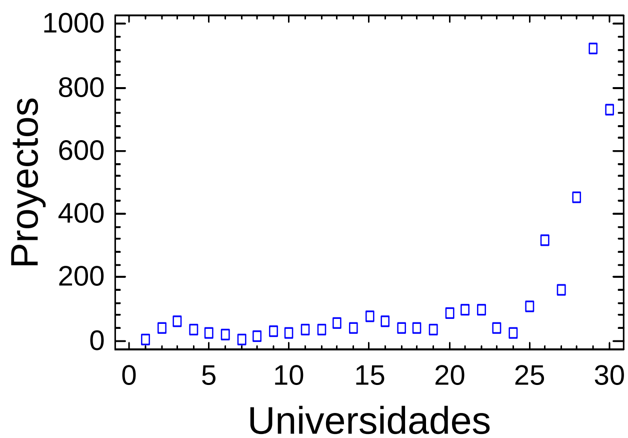
Nº	Universidades	Proyectos
Universidad muy muy pequeña		
1	FPS-Ocaña	3
Universidades muy pequeñas		
2	Sucre	38
3	Guajira	62
4	Amazonia	33
5	Mayor de Cundinamarca	25
6	Llanos	17
7	Cundinamarca	5
8	Cesar	15
Universidades pequeñas		
9	Chocó	31
10	FPS-Cúcuta	25
11	Pamplona	35
12	Tolima	36
13	Surcolombiana	55
14	Magdalena	41
15	Cartagena	77
16	Pedagógica Nacional	59
17	Militar	38
18	Quindío	37
Universidades medianas		
19	Nariño	36
20	Caldas	89
21	Tecnológica de Pereira	95
22	Atlántico	99
23	Cauca	42
24	Córdoba	26
25	Distrital	106
26	UPTC	315
27	UIS	159
Universidades grandes		
28	Valle	451
29	Antioquia	923
Universidad muy grande		
30	Nacional de Colombia	732

Los estadísticos básicos fundamentales asociados a esta variable se presentan en la Tabla 4.11.

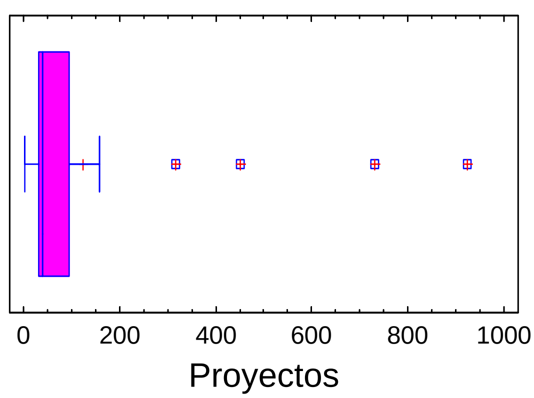
Tabla 4.11. Estadísticos básicos de la variable Proyectos

Identificación de la variable	Media	Desviación típica	Mínimo	Máximo	mediana	Correlación entre Proyectos y Recursos
Proyectos	123,5	210,4	3	923	39,5	0,89

Como se observa en la Tabla 4.11 y lo ratifican los gráficos 4.9 y 4.10, la distribución del número de proyectos es muy asimétrica sesgada a la derecha. Es decir, existen unas pocas universidades que tienen mucho más Proyectos de investigación que el resto.
El diagrama de caja que se muestra en la gráfica 4.10 evidencia cuatro valores atípicos que corresponden a las universidades Antioquia, Nacional, Valle y la UPTC.



Gráfica 4.9. Proyectos



Gráfica 4.10. Diagrama de caja de la variable Proyectos

De acuerdo a las gráficas las universidades grandes y muy grande siguen siendo universidades representativas en esta variable, ya que sus valores son muy altos con respecto a las demás universidades; pero es importante destacar que la universidad de Antioquia supera una vez más a la Nacional.

En el grupo de las universidades medianas, pequeñas y muy pequeñas se destacan las universidades UPTC (315), Cartagena (77) y la Guajira (62) respectivamente por tener valores altos que no responden a la proporción de las universidades de los grupos a los cuales pertenecen; el valor de la universidad UPTC es tan extraño que en la gráfica 4.10 aparece como valor atípico. Lo contrario sucede con la universidad de Córdoba (26); aunque pertenece al grupo de las universidades medianas su número de proyectos es muy bajo en comparación con las demás universidades de su grupo.

Es importante mencionar que la UPTC dio un gran salto en comparación de la UIS, debido a que el número de proyectos es casi el doble de la UIS, recibiendo menos presupuesto que ésta. Recordemos que esta última relación es la misma que se presenta con el número de profesores: entre mayor número de profesores de planta más proyectos de investigación.

6. VARIABLE: ARTÍCULOS

En la Tabla 4.12 aparecen los valores de esta variable para cada una de las universidades.

Tabla 4.12. Artículos

Nº	Universidades	Artículos
Universidad muy muy pequeña		
1	FPS-Ocaña	0
Universidades muy pequeñas		
2	Sucre	12
3	Guajira	0
4	Amazonia	0
5	Mayor de Cundinamarca	0
6	Llanos	30
7	Cundinamarca	0
8	Cesar	0
Universidades pequeñas		
9	Chocó	0
10	FPS-Cúcuta	0
11	Pamplona	949
12	Tolima	77,2
13	Surcolombiana	0
14	Magdalena	79,5
15	Cartagena	128,5
16	Pedagógica Nacional	38
17	Militar	0
18	Quindío	64

Tabla 4.12. Continuación Artículos

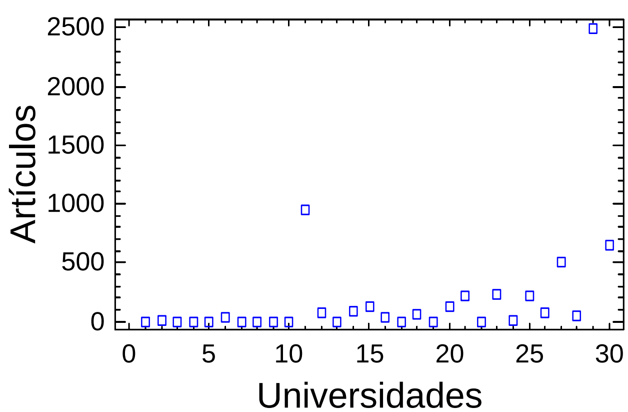
Universidades medianas		
19	Nariño	0
20	Caldas	123
21	Tecnológica de Pereira	210,4
22	Atlántico	0
23	Cauca	234
24	Córdoba	6
25	Distrital	212
26	UPTC	76
27	UIS	497,9
Universidades grandes		
28	Valle	45,2
29	Antioquia	2488
Universidad muy grande		
30	Nacional de Colombia	647

Los estadísticos fundamentales asociados a esta variable se presentan en la Tabla 4.13.

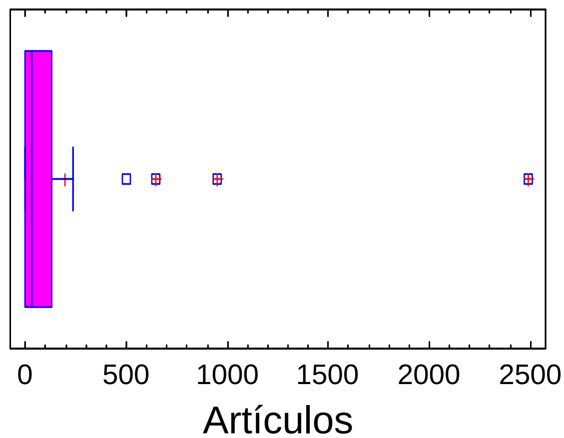
Tabla 4.13. Estadísticos básicos de la variable Artículos

Identificación de la variable	Media	Desviación Típica	Mínimo	Máximo	Mediana	Correlación entre Artículos y Recursos
Artículos	197,3	475,8	0	2488	34	0,56

Como se observa en la Tabla 4.13 y lo ratifican los gráficos 4.11 y 4.12, la distribución de los artículos es muy asimétrica sesgada a la derecha. Es decir, existen unas pocas universidades que tienen mucho más Puntos por Artículos que el resto de las universidades. El diagrama de caja que se muestra en la gráfica 4.12 evidencia cuatro valores atípicos que corresponden a las universidades Antioquia, Pamplona, Nacional y la UIS.



Gráfica 4.11. Artículos



Gráfica 4.12. Diagrama de caja de la variable Artículos

Como se puede observar en el grupo de las universidades grandes se destacan la universidad del Valle (45,2) por tener un valor bajo que no es proporcional al valor de la de Antioquia, además se observa que la universidad de Antioquia supera una vez más a la Nacional.

En el grupo de las universidades medianas se destacan las universidades UIS (497,9), Córdoba (6), Atlántico (0) y Nariño (0); la primera por tener un valor alto que no corresponde a la proporción de las universidades del grupo al que pertenece, su valor es tan extraño que en la gráfica 4.12 aparece como valor atípico; las tres ultimas universidades antes mencionadas se destacan por tener valores bajos.

Para el grupo de las universidades pequeñas y muy pequeñas podemos destacar las universidades de Pamplona (949), Llanos (30) y Sucre (12); por que poseen valores altos que no responde a la proporción de las universidades de los grupos a los cuales pertenece, el valor de Pamplona es tan extraño que aparece en la grafica 4.12 como valor atípico.

El gráfico 4.11 resalta los valores extraños que para esta variable tienen las universidades Valle, Nacional y Pamplona. Las dos primeras por su bajo puntaje y la última por su altísimo puntaje. Los puntajes bajos se pueden explicar por la demora que ha existido en algunas universidades en la implementación del decreto 1279 y en la indexación de revistas en Colciencias. El caso de la Universidad de Pamplona es muy especial y hasta “sospechoso” porque como se observa no se correlaciona con el número de grupos de investigación que posee: 3. Situaciones como ésta son un llamado a la implementación de una política de control por parte del ministerio de educación que permita la transparencia en el manejo de la información por parte de las universidades.

7. VARIABLE: LIBROSTEX

En la Tabla 4.14 aparecen los valores de esta variable para cada una de las universidades.

Tabla 4.14 LibrosTex

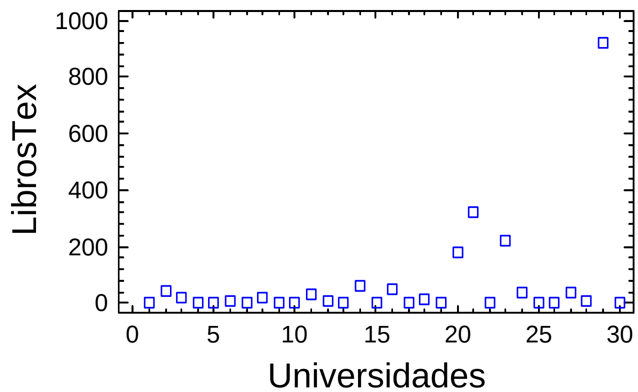
Nº	Universidades	LibrosTex
Universidad muy muy pequeña		
1	FPS-Ocaña	0
Universidades muy pequeñas		
2	Sucre	45
3	Guajira	20
4	Amazonia	0
5	Mayor de Cundinamarca	0
6	Llanos	7,41
7	Cundinamarca	0
8	Cesar	20
Universidades pequeñas		
9	Chocó	0
10	FPS-Cúcuta	0
11	Pamplona	31
12	Tolima	11,3
13	Surcolombiana	0
14	Magdalena	60
15	Cartagena	0
16	Pedagógica Nacional	48,32
17	Militar	0
18	Quindío	12
Universidades medianas		
19	Nariño	0
20	Caldas	179,2
21	Tecnológica de Pereira	322
22	Atlántico	0
23	Cauca	221
24	Córdoba	38,3
25	Distrital	5
26	UPTC	4
27	UIS	36,1
Universidades grandes		
28	Valle	5,8
29	Antioquia	920
Universidad grande		
30	Nacional de Colombia	0

Los estadísticos fundamentales asociados a esta variable se presentan en la Tabla 4.15.

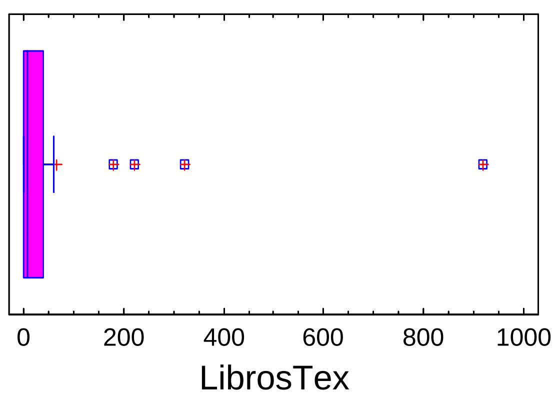
Tabla 4.15. Estadísticos básicos de la variable LibrosTex

Identificación de la variable	Media	Desviación Típica	Mínimo	Máximo	Mediana	Correlación entre LibrosTex y Recursos
LibrosTex	66,2	174,4	0	920	6,6	0,35

Como se observa en la Tabla 4.15 y lo ratifican los gráficos 4.13 y 4.14, la distribución de los LibrosTex es muy asimétrica sesgada a la derecha. Es decir, existen unas pocas universidades que tienen mucho más LibrosTex que el resto de las universidades.



Gráfica 4.13. LibrosTex



Gráfica 4.14. Diagrama de caja de la variable LibrosTex

Lo primero que se puede apreciar de las gráficas es que la Nacional no produce libros derivados de los procesos docentes; sin indagar a fondo es algo que valdría la pena averiguar. Además las universidades Caldas, Tecnológica de Pereira y Cauca tienen valores

que en el conjunto de las universidades son valores extraños o atípicos (Ver el diagrama de caja, Gráfico 4.14).
La universidad de Antioquia posee un valor muy grande, casi el triple de LibrosTex de la Tecnológica, siendo la Tecnológica una universidad con un valor relativamente grande, además sobrepasa la universidad Nacional una vez más.
También llama la atención que 12 universidades (40%) no produjeron en el 2003 ningún libro texto.

8. VARIABLE: LIBROSINV

En la Tabla 4.16 aparecen los valores de esta variable para cada una de las universidades.

Tabla 4.16. LibrosInv

Nº	Universidades	LibrosInv
Universidad muy muy pequeña		
1	FPS-Ocaña	0
Universidades muy pequeñas		
2	Sucre	45
3	Guajira	0
4	Amazonia	0
5	Mayor de Cundinamarca	6,5
6	Llanos	0
7	Cundinamarca	0
8	Cesar	0
Universidades pequeñas		
9	Chocó	0
10	FPS-Cúcuta	0
11	Pamplona	0
12	Tolima	26,4
13	Surcolombiana	0
14	Magdalena	60
15	Cartagena	80
16	Pedagógica Nacional	279,88
17	Militar	0
18	Quindío	1
Universidades medianas		
19	Nariño	0
20	Caldas	60,3
21	Tecnológica de Pereira	161
22	Atlántico	0
23	Cauca	140
24	Córdoba	0
25	Distrital	29
26	UPTC	0
27	UIS	20,1

Tabla 4.16. Continuación LibrosInv

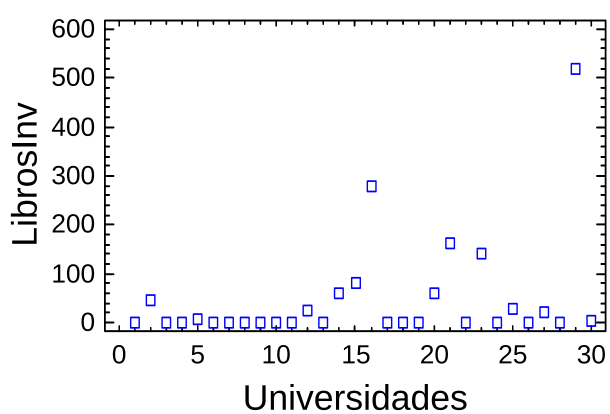
Universidades grandes		
28	Valle	0
29	Antioquia	519
Universidad muy grande		
30	Nacional de Colombia	5

Los estadísticos fundamentales asociados a esta variable se presentan en la Tabla 4.17.

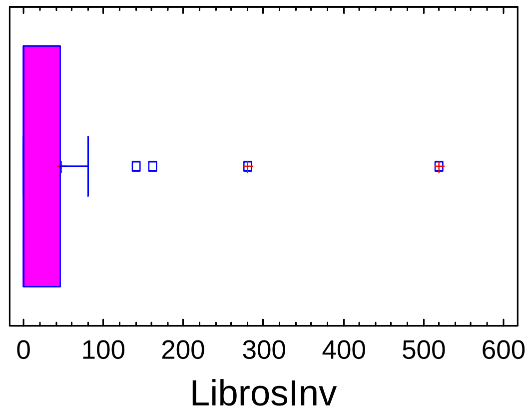
Tabla 4.17. Estadísticos básicos de la variable LibrosInv

Identificación de la variable	Media	Desviación Típica	Mínimo	Máximo	Mediana	Correlación entre LibrosInv y Recursos
LibrosInv	47,8	106,8	0	519	0	0,3

Como se observa en la Tabla 4.17 y lo ratifican los gráficos 4.15 y 4.16, la distribución de LibrosInv es muy asimétrica sesgada a la derecha. Es decir, existen unas pocas universidades que tienen mucho más LibrosInv que el resto de las universidades. El diagrama de caja que se muestra en la gráfica 4.16 evidencia cuatro valores atípicos que corresponden a las universidades Antioquia, Pedagógica, Tecnológica y Cauca.



Gráfica 4.15. LibrosInv



Gráfica 4.16. Diagrama de caja de la variable LibrosInv

En el grupo de las universidades grandes cabe destacar que la universidad del Valle (0) no produce Libros derivados de la investigación y por el contrario la de Antioquia (519) supera a la Nacional.
En el grupo de las universidades medianas y pequeñas podemos destacar las universidades Tecnológica (161), Cauca (140) y Pedagógica (279,88), por que poseen valores altos que no responden a la proporción de las universidades de los grupos a los cuales pertenecen, sus valores son tan extraños que aparecen en la gráfica 4.16 como valores atípicos.

Para el grupo de las universidades muy pequeñas se destaca la universidad del Sucre (45), ya que tiene más grupos de investigación que la Nacional recibiendo está mucho más presupuesto que la del Sucre.

De las gráficas se podría deducir que las universidades no apoyan la producción de textos, ya que más del 50% de las universidades públicas del país no producen libros derivados de la investigación. Escandaloso, por decir lo menos, es la producción de la Nacional: sólo 5 libros como resultado de investigación y ningún libro de texto.

En sentido contrario, llama la atención la Pedagógica, universidad de tamaño medio, que tiene una producción grande de libros resultados de investigación. De hecho, esta universidad cuenta con su propia librería abierta al público donde la gran mayoría de textos son de su propia autoría. En esta misma dirección exitosa se destacan las universidades Tecnológica de Pereira y Cauca, universidades que también se destacan en la publicación de libros de texto.

9. VARIABLE: PROGRAMAS

En la Tabla 4.18 aparecen los valores de esta variable para cada una de las universidades.

Tabla 4.18. Programas

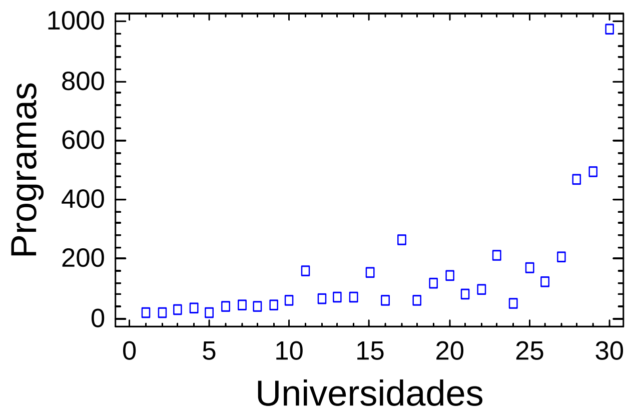
Nº	Universidades	Programas
Universidad muy muy pequeña		
1	FPS-Ocaña	19,475
Universidades muy pequeñas		
2	Sucre	20,4
3	Guajira	27,635
4	Amazonia	35,625
5	Mayor de Cundinamarca	20,615
6	Llanos	39,075
7	Cundinamarca	46,275
8	Cesar	38,6
Universidades pequeñas		
9	Chocó	44,32
10	FPS-Cúcuta	60,445
11	Pamplona	157,54
12	Tolima	66,35
13	Surcolombiana	69,325
14	Magdalena	69,475
15	Cartagena	152,47
16	Pedagógica Nacional	62
17	Militar	264,14
18	Quindío	60,215
Universidades medianas		
19	Nariño	115,85
20	Caldas	145,43
21	Tecnológica de Pereira	83,1
22	Atlántico	98,2
23	Cauca	211,545
24	Córdoba	52,4
25	Distrital	169,3
26	UPTC	123,45
27	UIS	209,56
Universidades grandes		
28	Valle	470,69
29	Antioquia	495,61
Universidad muy grande		
30	Nacional de Colombia	977,5

Los estadísticos fundamentales asociados a esta variable se presentan en la Tabla 4.19.

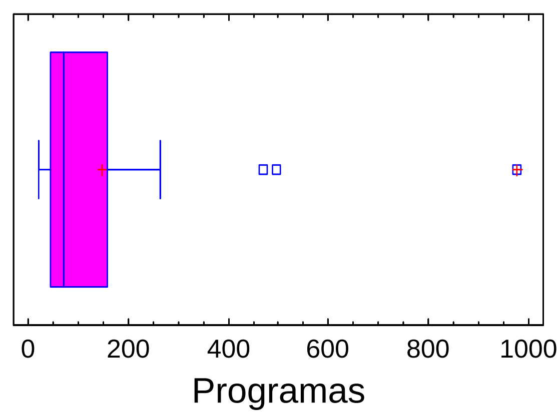
Tabla 4.19. Estadísticos básicos de la variable Programas

Identificación de las variables	Media	Desviación típica	Mínimo	Máximo	mediana	Correlación entre Programas y Recursos
Programas	146,9	192,8	19,5	977,5	69,4	0,96

Como se observa en la Tabla 4.19 y lo ratifican los gráficos 4.17 y 4.18, la distribución del número de Programas es muy asimétrica sesgada a la derecha. Es decir, existen unas pocas universidades que tienen mucho más Programas que el resto de las universidades. El diagrama de caja que se muestra en la gráfica 4.18 evidencia tres valores atípicos que corresponden a las universidades Nacional, Antioquia y Valle.



Gráfica 4.17. Programas



Gráfica 4.18. Diagrama de caja de la variable Programas

En el grupo de las universidades grandes y muy grande las universidades mantienen la proporción de sus programas de acuerdo a sus Recursos. Por lo tanto se observa que esta variable responde al principio básico asumido: el tamaño de la universidad determina el número de programas. De hecho el coeficiente de correlación entre esta variable y la variable Recursos es de 0.96.

En el grupo de las universidades medianas en términos negativos se destaca Córdoba quien ofrece una cantidad de programas muy inferior a la de las universidades del grupo al cual pertenece.

Para el grupo de las universidades pequeñas podemos destacar las universidades de Pamplona, Cartagena y Militar porque tienen una cantidad de programas superior al que se pudiera esperar por sus recursos. En el caso de la de Pamplona se explica por la gran cantidad de programas nuevos que tiene.

10. VARIABLE: PRIMÍPAROS

En la Tabla 4.20 aparecen los valores de esta variable para cada una de las universidades.

Tabla 4.20. Primíparos

Nº	Universidades	Primíparos
Universidad muy muy pequeña		
1	FPS-Ocaña	1410,65
Universidades muy pequeñas		
2	Sucre	1345,48
3	Guajira	2715,46
4	Amazonia	1961,88
5	Mayor de Cundinamarca	2063,24
6	Llanos	2189,1
7	Cundinamarca	4259,625
8	Cesar	4971,1
Universidades pequeñas		
9	Chocó	2523,5
10	FPS-Cúcuta	5708,425
11	Pamplona	9439,665
12	Tolima	3950,32
13	Surcolombiana	2799,56
14	Magdalena	4128,365
15	Cartagena	4018,27
16	Pedagógica Nacional	2505,1
17	Militar	4888,02
18	Quindío	3912,58

Tabla 4.20. Continuación Primíparos

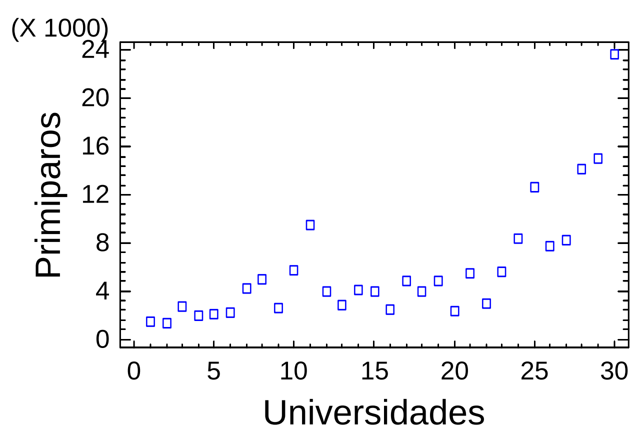
Universidades medianas		
19	Nariño	4904,5
20	Caldas	2272,645
21	Tecnológica de Pereira	5445
22	Atlántico	3001,7
23	Cauca	5590,015
24	Córdoba	8418,775
25	Distrital	12657,15
26	UPTC	7722,03
27	UIS	8198,42
Universidades grandes		
28	Valle	14161,545
29	Antioquia	14959,875
Universidades muy grandes		
30	Nacional de Colombia	23710,5

Los estadísticos fundamentales asociados a esta variable se presentan en la Tabla 4.21.

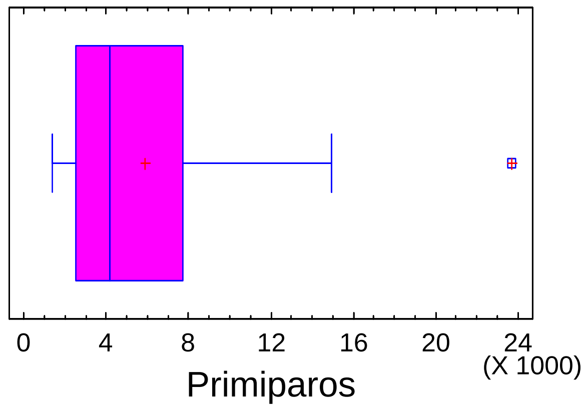
Tabla 4.21. Estadísticos básicos de la variable Primíparos

Identificación de las variables	Media	Desviación típica	Mínimo	Máximo	mediana	Correlación entre Primíparos y Recursos
Primíparos	5861,1	4865,2	1345,5	23710,5	4194	0,9

Como se observa en la Tabla 4.21 y lo ratifican los gráficos 4.19 y 4.20, la distribución del número de Primíparos es muy asimétrica sesgada a la derecha. Es decir, existen unas pocas universidades que tienen mucho más Primíparos que el resto de las universidades. El diagrama de caja que se muestra en la gráfica 4.20 evidencia un valor atípico que corresponde a la universidad Nacional de Colombia.



Gráfica 4.19. Primíparos



Gráfica 4.20. Diagrama de caja de la variable Primíparos

En el grupo de las universidades grandes la proporción del número de Primíparos se mantiene de acuerdo a sus Recursos. Por lo tanto se observa que esta variable responde al principio básico asumido: el tamaño de la universidad determina el número de Primíparos. De hecho el coeficiente de correlación entre esta variable y la variable Recursos es de 0.9.

En el grupo de las universidades medianas se destaca la universidad Distrital (12657,1), Atlántico (3001,7) y Caldas (2272,7); la primera posee un valor alto que no corresponde a la proporción de las universidades del grupo al cual pertenece, pero lo contrario sucede con las dos ultimas que poseen valores bajos en comparación con las universidades de su grupo.

Algo que vale la pena destacar del grupo de las universidades pequeñas, es que a pesar de recibir menos presupuesto que las universidades de Caldas, Atlántico, UPTC y la UIS, la universidad de Pamplona tiene mucho más primíparos; esto se debe a que la universidad ofrece mucha educación a distancia en todo el país.

Para los grupos de universidades muy pequeñas y muy muy pequeñas la proporción se mantiene.

11. VARIABLE: ESTUDIANTES

En la Tabla 4.22 aparecen los valores de esta variable para cada una de las universidades.

Tabla 4.22. Estudiantes

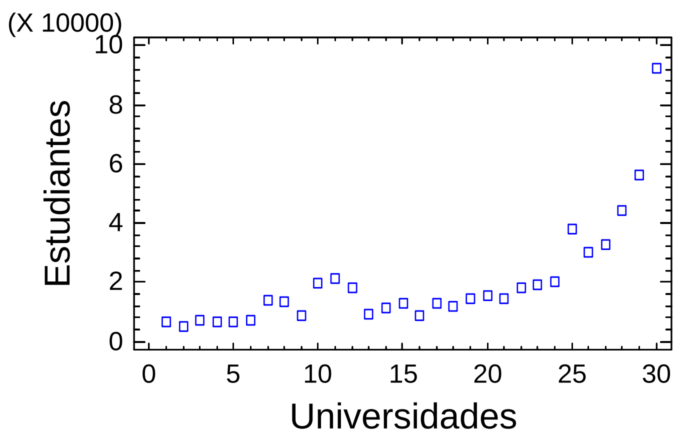
Nº	Universidades	Estudiantes
Universidad muy muy pequeña		
1	FPS-Ocaña	6463,6
Universidades muy pequeñas		
2	Sucre	4776,82
3	Guajira	7022,595
4	Amazonia	6371,49
5	Mayor de Cundinamarca	6764,43
6	Llanos	7011,475
7	Cundinamarca	13649,725
8	Cesar	13343,9
Universidades pequeñas		
9	Chocó	8591,48
10	FPS-Cúcuta	19426,96
11	Pamplona	21451,915
12	Tolima	18104
13	Surcolombiana	9050,64
14	Magdalena	11407,56
15	Cartagena	13002,86
16	Pedagógica Nacional	8429,1
17	Militar	12807,68
18	Quindío	11709,84
Universidades medianas		
19	Nariño	14553,9
20	Caldas	15582,465
21	Tecnológica de Pereira	14230,15
22	Atlántico	17882,5
23	Cauca	19071,475
24	Córdoba	20173,15
25	Distrital	38048,95
26	UPTC	30203,87
27	UIS	32522,015
Universidades grandes		
28	Valle	44095,355
29	Antioquia	56214,405
Universidad muy grande		
30	Nacional de Colombia	92482,8

Los estadísticos fundamentales asociados a esta variable se presentan en la Tabla 4.23.

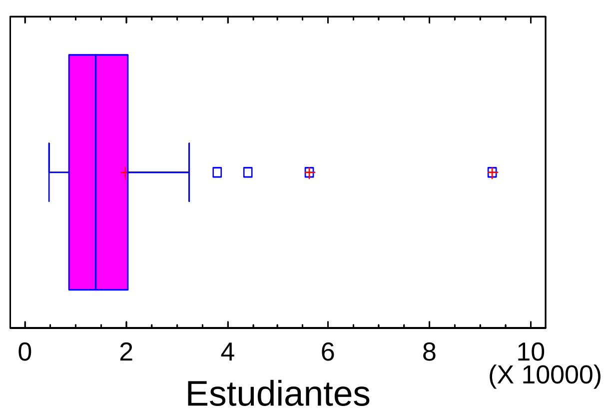
Tabla 4.23. Estadísticos básicos de la variable Estudiantes

Identificación de las variables	Media	Desviación típica	Mínimo	Máximo	mediana	Correlación entre Estudiantes y Recursos
Estudiantes	19814,9	17935,2	4776,8	92482,8	13940	0,96

Como se observa en la Tabla 4.23 y lo ratifican los gráficos 4.21 y 4.22, la distribución del número de Estudiantes es muy asimétrica sesgada a la derecha. Es decir, existen unas pocas universidades que tienen mucho más Estudiantes que el resto de las universidades. El diagrama de caja que se muestra en la gráfica 4.22 evidencia cuatro valores atípicos que corresponden a las universidades Nacional, Antioquia, Valle y Distrital.



Gráfica 4.21. Estudiantes



Gráfica 4.22. Diagrama de caja de la variable Estudiantes

En el grupo de las universidades grandes y muy grande se mantiene la proporción en el número de estudiantes. Por lo tanto se observa que esta variable responde al principio básico asumido: el tamaño de la universidad determina el número de Primíparos. De hecho el coeficiente de correlación entre esta variable y la variable Recursos es de 0.96.

En el grupo de las universidades medianas podemos destacar la universidad Distrital (38048,95), ya que su valor no corresponde a la proporción de las universidades del grupo al cual pertenece; su valor es tan extraño que en la gráfica 4.22 aparece como valor atípico.
Para el grupo de las universidades pequeñas, muy pequeñas y muy muy pequeña la proporción se mantiene.
Cabe destacar que la Distrital en comparación con la UPTC y la UIS posee un número superior de estudiantes, a pesar de recibir menos presupuesto.

12. VARIABLE: GRADUADOS

En la Tabla 4.24 aparecen los valores de esta variable para cada una de las universidades.

Tabla 4.24. Graduados

Nº	Universidades	Graduados
Universidad muy muy pequeña		
1	FPS-Ocaña	622,95
Universidades muy pequeñas		
2	Sucre	968
3	Guajira	954,6
4	Amazonia	925,425
5	Mayor de Cundinamarca	1613,275
6	Llanos	624,95
7	Cundinamarca	1072,2
8	Cesar	911,4
Universidades pequeñas		
9	Chocó	1283,1
10	FPS-Cúcuta	1817,875
11	Pamplona	3981,675
12	Tolima	4481,05
13	Surcolombiana	1466,2
14	Magdalena	959,15
15	Cartagena	1869,125
16	Pedagógica Nacional	601,1
17	Militar	1934,1
18	Quindío	3674,75

Tabla 4.24. Continuación Graduados

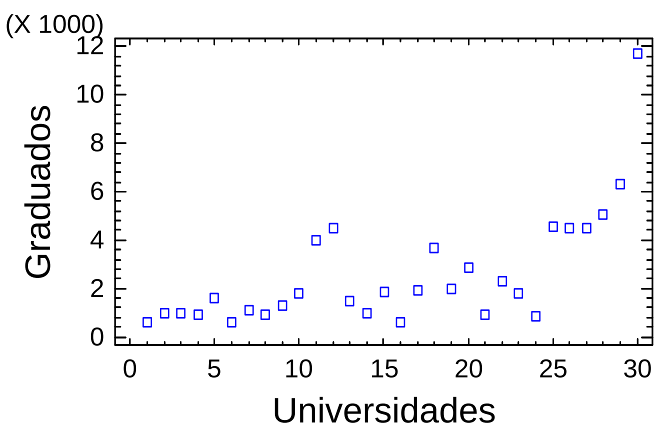
Universidades medianas		
19	Nariño	1962,675
20	Caldas	2874,075
21	Tecnológica de Pereira	943,5
22	Atlántico	2315,1
23	Cauca	1804,3
24	Córdoba	849,45
25	Distrital	4542,15
26	UPTC	4517,2
27	UIS	4463,4
Universidades grandes		
28	Valle	5065,525
29	Antioquia	6309,7
Universidad grande		
30	Nacional de Colombia	11740,8

Los estadísticos fundamentales asociados a esta variable se presentan en la Tabla 4.25.

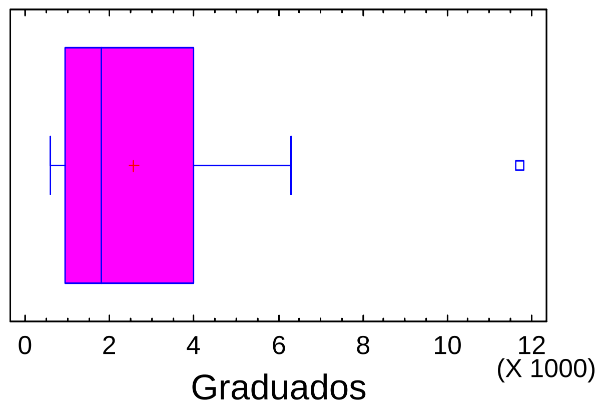
Tabla 4.25. Estadísticos básicos de la variable Graduados

Identificación de las variables	Media	Desviación típica	Mínimo	Máximo	mediana	Correlación entre Graduados y Recursos
Graduados	2571,6	2319	601,1	11740,8	1811,1	0,89

Como se observa en la Tabla 4.25 y lo ratifican los gráficos 4.23 y 4.24, la distribución del número de Graduados es muy asimétrica sesgada a la derecha. Es decir, existen unas pocas universidades que tienen mucho más Graduados que el resto de las universidades. El diagrama de caja que se muestra en la gráfica 4.24 evidencia un valor atípico que corresponde a la universidad Nacional.



Gráfica 4.23. Graduados



Gráfica 4.24. Diagrama de caja de la variable Graduados

En el grupo de las universidades grandes y muy grandes se mantiene al proporción del número de Graduados. Por lo tanto se observa que esta variable responde al principio básico asumido: el tamaño de la universidad determina el número de Graduados. De hecho el coeficiente de correlación entre esta variable y la variable Recursos es de 0.89.

En el grupo de universidades medianas se destacan la de Cauca (1804,3), Tecnológica (943,5) y la de Córdoba (849,45); por que poseen valores bajos que no corresponden con la proporción de las universidades del grupo al cual pertenecen.

Para el grupo de las universidades pequeñas se destacan dos grupos: el primero esta conformado por Pamplona (3981,7) y Tolima (4481,1), ya que poseen valores altos que no corresponden a la proporción de las universidades del grupo al cual pertenece. Y el segundo es la Pedagógica (601,1) que posee un valor muy bajo.
Los grupos de las universidades muy pequeñas y muy muy pequeña mantienen su proporción.

13. VARIABLE: ACREDITADOS

En la Tabla 4.26 aparecen los valores de esta variable para cada una de las universidades.

Tabla 4.26. Acreditados

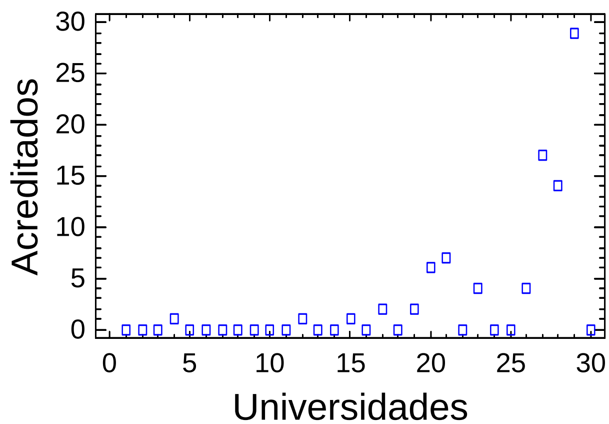
Nº	Universidades	Acreditados
Universidad muy muy pequeña		
1	FPS-Ocaña	0
Universidades muy pequeñas		
2	Sucre	0
3	Guajira	0
4	Amazonia	1
5	Mayor de Cundinamarca	0
6	Llanos	0
7	Cundinamarca	0
8	Cesar	0
Universidades pequeñas		
9	Chocó	0
10	FPS-Cúcuta	0
11	Pamplona	0
12	Tolima	1
13	Surcolombiana	0
14	Magdalena	0
15	Cartagena	1
16	Pedagógica Nacional	0
17	Militar	2
18	Quindío	0
Universidades medianas		
19	Nariño	2
20	Caldas	6
21	Tecnológica de Pereira	7
22	Atlántico	0
23	Cauca	4
24	Córdoba	0
25	Distrital	0
26	UPTC	4
27	UIS	17
Universidades grandes		
28	Valle	14
29	Antioquia	29
Universidad muy grande		
30	Nacional de Colombia	0

Los estadísticos fundamentales asociados a esta variable se presentan en la Tabla 4.27.

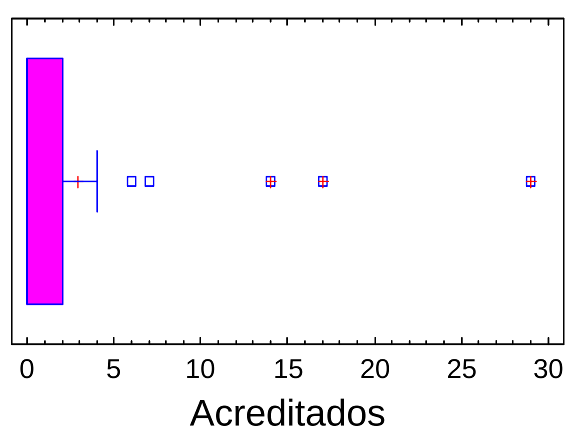
Tabla 4.27. Estadísticos básicos de la variable Acreditados

Identificación de las variables	Media	Desviación típica	Mínimo	Máximo	mediana	Correlación entre Acreditados y Recursos
Acreditados	2,9	6,3	0	29	0	0,43

El diagrama de caja que se muestra en la gráfica 4.26 evidencia cinco valores atípicos que corresponden a las universidades Antioquia, UIS, Valle, Tecnológica y Caldas.



Gráfica 4.25. Acreditados



Gráfica 4.26. Diagrama de caja de la variable Acreditados

En la gráfica 4.25 llama la atención que la universidad Nacional siendo una universidad muy grande no tenga ningún programa acreditado.
En el grupo de la universidades grandes cabe destacar que la universidad del Valle (14) posee muy pocos programas acreditados en comparación con la universidad de Antioquia, y esta a su vez supera una vez más a la Nacional.
Para el grupo de las universidades medianas se destacan la UIS (17), Tecnologica (7) y la del Caldas (6), por que poseen valores altos que no corresponden a la proporción de las universidades del grupo al cual pertenecen; sus valores son tan extraños que aparecen en la Gráfica 4.26 como valores atípicos.

Claramente se puede observar que el 60% de las universidades públicas del país no se están preocupando por la acreditación de sus programas.
Es importante mencionar que las universidades Caldas, Tecnológica, Cauca y UPTC obtiene mayor número de programas acreditados en comparación de la Nacional que recibe un mejor presupuesto.

14. VARIABLE: EXTENSIÓN

En la Tabla 4.28 aparecen los valores de esta variable para cada una de las universidades.

Tabla 4.28. Extensión

Nº	Universidades	Extensión
Universidad muy muy pequeña		
1	FPS-Ocaña	49,438
Universidades muy pequeñas		
2	Sucre	159,22
3	Guajira	48,2
4	Amazonia	26
5	Mayor de Cundinamarca	339
6	Llanos	950
7	Cundinamarca	209
8	Cesar	125
Universidades pequeñas		
9	Chocó	169,172
10	FPS-Cúcuta	1844,231
11	Pamplona	4337,59
12	Tolima	526,629
13	Surcolombiana	2266,04
14	Magdalena	285,389
15	Cartagena	2870,042
16	Pedagógica Nacional	2335,324
17	Militar	1222
18	Quindío	1722,627

Tabla 4.28. Continuación Extensión

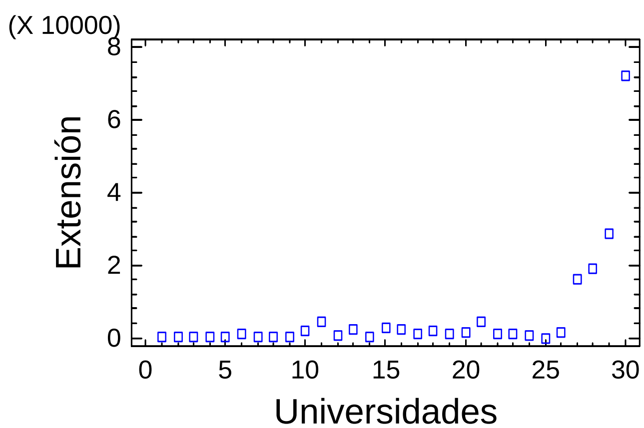
Universidad medianas		
19	Nariño	880
20	Caldas	1640,928
21	Tecnológica de Pereira	4443
22	Atlántico	1002,843
23	Cauca	1162,714
24	Córdoba	720,369
25	Distrital	0
26	UPTC	1598
27	UIS	16333,8
Universidades grandes		
28	Valle	19124,4
29	Antioquia	28533,345
Universidad muy grande		
30	Nacional de Colombia	72418

Los estadísticos fundamentales asociados a esta variable se presentan en la Tabla 4.29.

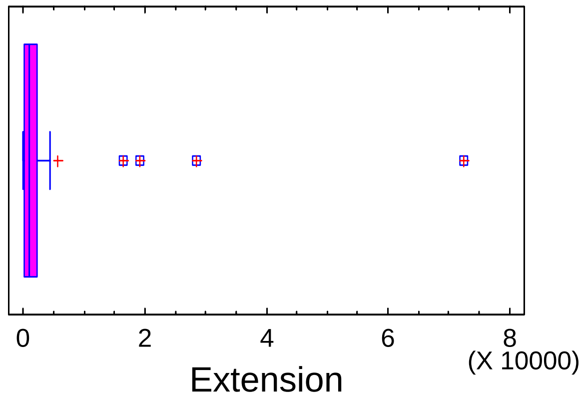
Tabla 4.29. Estadísticos básicos de la variable Extensión

Identificación de las variables	Media	Desviación típica	Mínimo	Máximo	mediana	Correlación entre Extensión y Recursos
Extensión	5578,1	13944,2	0	72418	1082,8	0,96

El diagrama de caja que se muestra en la gráfica 4.28 evidencia cuatro valores atípicos que corresponden a las universidades Nacional, Antioquia, Valle y la UIS.



Gráfica 4.27. Extensión



Gráfica 4.28. Diagrama de caja de la variable Extensión

En el grupo de las universidades grandes y muy grandes se observa que tienen valores muy altos en comparación con el resto de los grupos.
En el grupo de las universidades medianas se destaca la Distrital (0) ya que no posee ningún ingreso por extensión, lo cual nos deja un gran interrogante.
Para el resto de los grupos podría decirse que sus valores son muy bajos que no corresponden a la proporción de las universidades de los grupos a los cuales pertenecen.
A pesar de que la Extensión es una misión de cualquier institución, esta no esta recibiendo la importancia que se merece; pero es importante decir que las universidades con buen presupuesto están interactuando con la sociedad mucho mejor.

4.2 Análisis por Componentes Principales

Con ayuda del paquete estadístico SPAD[3] 4.5 hemos realizado el Análisis de Componentes Principales (ACP) a la información suministrada en la Tabla 4.1.
La matriz de correlaciones aparece en la Tabla 4.2.1.

Realizando un primer análisis de la matriz de correlaciones, que será confirmado por un análisis posterior de ACP, se observa que las variables se agrupan de una manera natural de acuerdo a su significado.

Es así que de una parte están las variables Recursos y Profesores que son Indicadores de Capacidad que tienen una correlación de 0.97.

Entre las variables de Resultados se observan los siguientes grupos:
Grupo 1, conformado por las variables LibrosInv, LibrosTex y Artículos que responden a la producción intelectual de los profesores. A este grupo lo identificamos como Producción.

Grupo 2, conformado por las variables Programas, Primíparos, Estudiantes y Graduados que responden a la misión de Formación o Docencia contemplada en la Misión de la Universidad. A este grupo lo identificamos como Formación.

Grupo 3, conformado por Revistas, GruposInv y Proyectos que se asocian con la actividad investigativa que realizan las Universidades. A este grupo lo identificamos como Investigación.

El ACP para las variables de resultados produce el gráfico 4.29 que muestra precisamente los grupos que acabamos de construir. Como también se observa, la variable Acreditados (número de programas acreditados) aparece en el grupo de Producción, sin embargo, como su significado no se relaciona directamente con el de la producción intelectual de los profesores, optamos por excluirla de este grupo y darle un tratamiento individual. Igualmente la variable Extensión que aparece en el grupo de Formación, y cuya esencia se relaciona con la misión de Extensión, decidimos dejarla aparte y formar el grupo de Extensión conformado por ella sola.

[3] Système Portable pour l’Analyse de Donnée

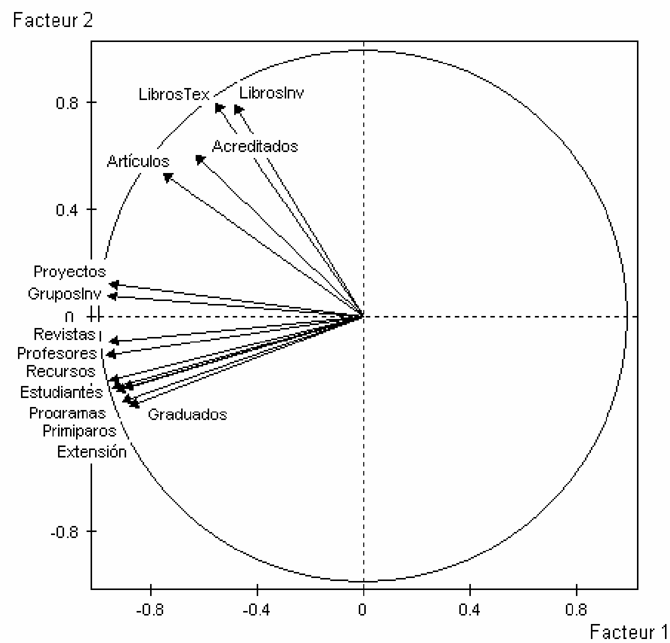


Gráfico 4.29. Grupos de Variables

A continuación realizaremos un ACP para cada uno de los grupos, que nos va a permitir clasificar las universidades en relación a ellos.

A) ACP para las variables ICAD: Recursos y Profesores

Estas variables, como fue mencionado en el análisis univariado, tienen una alta correlación: 0.97 lo que nos permite predecir que una sola componente principal dará cuenta de un alto porcentaje de la variación total de ambas variables.

Tabla 4.30. Valores propios variables ICAD

Número	Valor propio	Porcentaje	Porcentaje acumulado
1	1,9706	98,53	98,53
2	0,0294	1,47	100,00

En la tabla 4.30 se observa que, efectivamente, el primer valor propio explica el 98.53% de la variación total. Por esta razón solo trabajaremos con el primer factor o componente.

Tabla 4.2.1 Matriz de correlación de las variables

	GruposInv	Acreditados	Revistas	Graduados	Artículos	Extensión	Profesores
GruposInv	1,00						
Acreditados	0,69	1,00					
Revistas	0,91	0,48	1,00				
Graduados	0,78	0,38	0,84	1,00			
Artículos	0,71	0,75	0,68	0,53	1,00		
Extensión	0,86	0,36	0,94	0,86	0,50	1,00	
Profesores	0,91	0,49	0,94	0,91	0,64	0,91	1,00
LibrosTex	0,59	0,78	0,48	0,21	0,84	0,25	0,44
Proyectos	0,94	0,70	0,90	0,79	0,76	0,82	0,92
LibrosInv	0,53	0,66	0,41	0,13	0,74	0,21	0,37
Programas	0,90	0,42	0,91	0,88	0,52	0,94	0,93
Primíparos	0,84	0,42	0,85	0,88	0,56	0,85	0,91
Estudiantes	0,88	0,45	0,91	0,94	0,57	0,91	0,97
Recursos	0,91	0,43	0,96	0,89	0,56	0,96	0,97

4.2.1 Continuación de la Matriz de correlación de las variables

	LibrosTex	Proyectos	LibrosInv	Programas	Primíparos	Estudiantes	Recursos
GruposInv							
Acreditados							
Revistas							
Graduados							
Artículos							
Extensión							
Profesores							
LibrosTex	1,00						
Proyectos	0,62	1,00					
LibrosInv	0,89	0,54	1,00				
Programas	0,29	0,84	0,24	1,00			
Primíparos	0,29	0,81	0,21	0,90	1,00		
Estudiantes	0,31	0,87	0,24	0,94	0,97	1,00	
Recursos	0,35	0,89	0,30	0,96	0,90	0,96	1,00

La Tabla 4.31 da los coeficientes que acompañan a las combinaciones lineales de las variables Recursos y Profesores para definir los nuevos ejes, es decir, las componentes principales.

Tabla 4.31. **Ejes Unitarios**

Identificación de la variable	Eje 1
Profesores	0,71
Recursos	0,71

Es decir el eje 1 está definido por la combinación lineal: $0.71 * \text{Profesores} + 0.71 * \text{Recursos}$, lo que se puede interpretar como un índice que da el tamaño de las Universidades. De esta forma el ACP nos resume la información dada por dos variables en una sola.
Veamos ahora la posición que adoptan las universidades en este plano factorial y confrontémosla con la observada en el análisis unidimensional.

El gráfico 4.29 muestra la posición relativa de las universidades en el plano factorial generado. La posición de las universidades grandes (Antioquia y Valle) y muy grande (Nacional) son exactamente las mismas que percibimos en el análisis univariado.

Al excluir las universidades grandes se genera el gráfico 4.30. Este gráfico nos permite agrupar las universidades en Medianas, pequeñas, muy pequeñas y muy muy pequeña. Estos nuevos grupos no son exactamente iguales a los producidos en el análisis univariado, existen algunas variaciones producidas por la variable Profesores.

Con ayuda de las gráficas 4.29 y 4.30 y la Tabla 4.32, los grupos de acuerdo al tamaño de las universidades, quedan constituidos en la siguiente forma, donde el orden se asocia con el tamaño de la universidad:

Universidad muy grande: Nacional de Colombia

Universidades grandes: Antioquia y Valle

Universidades medianas: UPTC, UIS, Distrital, Cauca, Atlántico, Caldas, Tecnológica de Pereira y Córdoba.

Universidades pequeñas: Pamplona, Quindío, Nariño, Cartagena, Surcolombiana, Tolima, Pedagógica Nacional, Militar y FPS-Cúcuta.

Universidades muy pequeñas: Cesar, Cundinamarca, , Chocó, Magdalena, Llanos, Mayor de Cundinamarca, Amazonia, Guajira y Sucre.

Universidad muy muy pequeña: FPS-Ocaña.

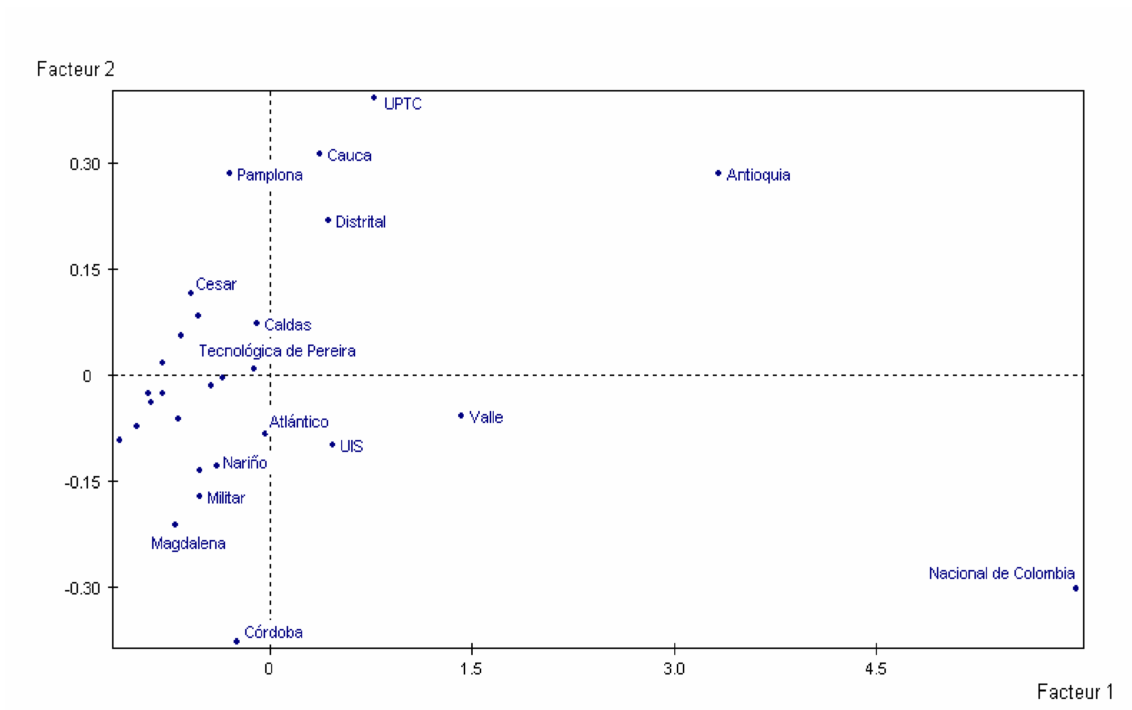


Gráfico 4.29. Gráfico de las Universidades

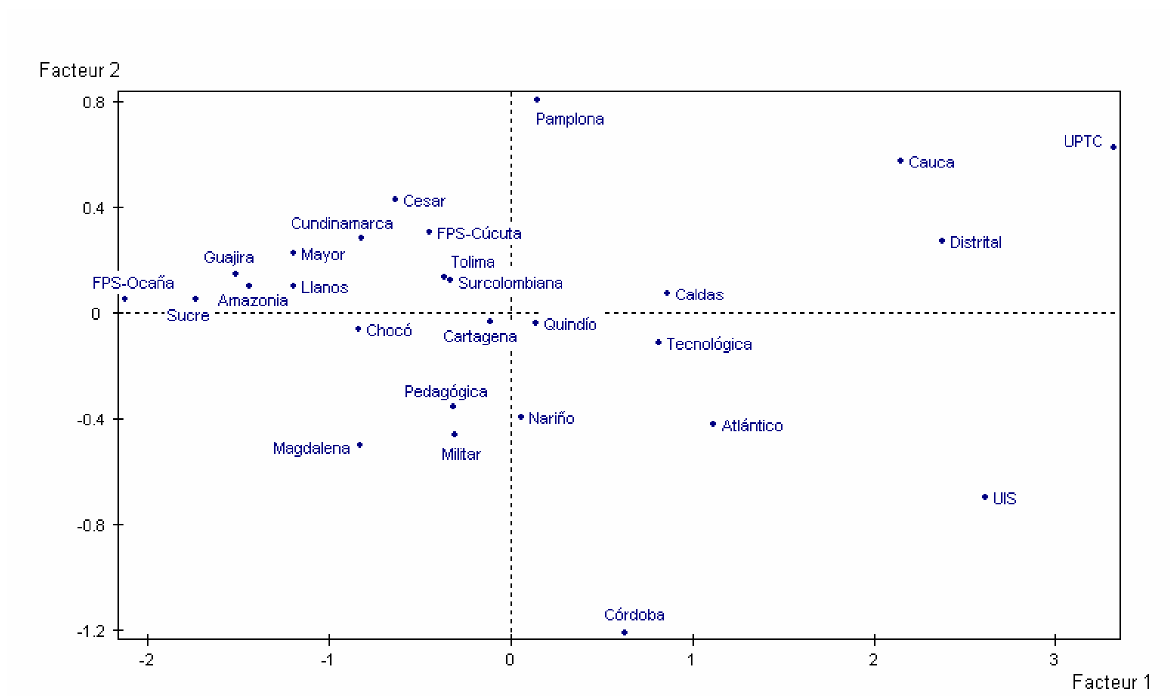


Gráfico 4.30. Gráfico de las universidades excluyendo la Nacional, Antioquia y Valle

Tabla 4.32. Coordenadas de las universidades sobre el índice de Capacidad

Identificación	**Eje 1**
Nacional de Colombia	6,00
Antioquia	3,33
Valle	1,42
UPTC	0,77
UIS	0,46
Distrital	0,43
Cauca	0,37
Atlántico	-0.04
Caldas	-0.10
Tecnológica de Pereira	-0.12
Córdoba	-0.25
Pamplona	-0.30
Quindío	-0.35
Nariño	-0.40
Cartagena	-0.44
Surcolombiana	-0.51
Tolima	-0.52
Pedagógica Nacional	-0.53
Militar	-0.53
FPS-Cúcuta	-0.54
Cesar	-0.59
Cundinamarca	-0.67
Choco	-0.69
Magdalena	-0.71
Llanos	-0.80
Mayor de Cundinamarca	-0.80
Amazonia	-0.89
Guajira	-0.91
Sucre	-0.99
FPS-Ocaña	-1.12

El grupo de las universidades grandes y muy grande en términos de Capacidad son de igual forma universidades grandes, lo cual ratifica lo realizado por el análisis univariado.
En el grupo de las universidades medianas lo único que ha cambiado es la posición que ocupan con respecto al índice de Capacidad dentro del mismo grupo.
En el grupo de las universidades pequeñas se destaca la universidad de Nariño ya que por su bajo número de profesores pasa de ser una universidad mediana a una universidad pequeña.

B) ACP para las variables de Producción: Artículos, LibrosTex y LibrosInv

En la Tabla 4.33 se puede apreciar la alta correlación que existen entre las variables de Producción lo que nos permite predecir que una sola componente principal dará cuenta de un alto porcentaje de la variación total de las tres variables.

Tabla 4.33 Matriz de correlaciones de las variables de producción

	Artículos	**LibrosTex**	**LibrosInv**
Artículos	1,00		
LibrosTex	0,84	1,00	
LibrosInv	0,74	0,89	1,00

Al realizar el ACP se obtienen los siguientes valores propios (Ver Tabla 4.34)

Tabla 4.34. Tabla de valores propios

Número	**Valor propio**	**Porcentaje**	**Porcentaje acumulado**
1	2,6493	88,31	88,31
2	0,2627	8,76	97,07
3	0,0880	2,93	100,00

En la Tabla 4.34 se observa que, efectivamente, el primer valor propio explica el 88.31% de la variación total. Por esta razón solo trabajaremos con el primer factor o componente.

La tabla 4.35 da los coeficientes que acompañan a las combinaciones lineales de las variables Artículos, LibrosTex y LibrosInv para definir los nuevos ejes, es decir, las componentes principales.

Tabla 4.35 Ejes unitarios

Identificación de la variable	**Eje 1**
Artículos	-0,56
LibrosTex	-0,60
LibrosInv	-0,57

Nuevamente tenemos que el eje 1 es una combinación lineal con coeficientes de tamaño parecido y del mismo signo, lo que nos permite interpretar este eje como el total de la Producción intelectual de los profesores. Específicamente el Eje 1 es la combinación lineal:
$-0.56 * \textit{Artículos} - 0.60 * \text{LibrosTex} - 0.57 * \text{LibrosInv}$

No obstante el eje esta definido con coordenadas negativas, es decir el aumento esta de derecha a izquierda pero para mayor formalidad decidimos cambiar la dirección de los ejes de tal forma que el aumento sea de izquierda a derecha.

Como el factor 1 determina la producción de textos producidos por los docentes de la institución podemos decir que las universidades que se ubican del lado positivo del eje son universidades con una alta producción intelectual de los profesores y las universidades ubicadas del lado negativo son universidades con poca producción intelectual.

La posición relativa de las universidades se presenta en el gráfico 4.31.

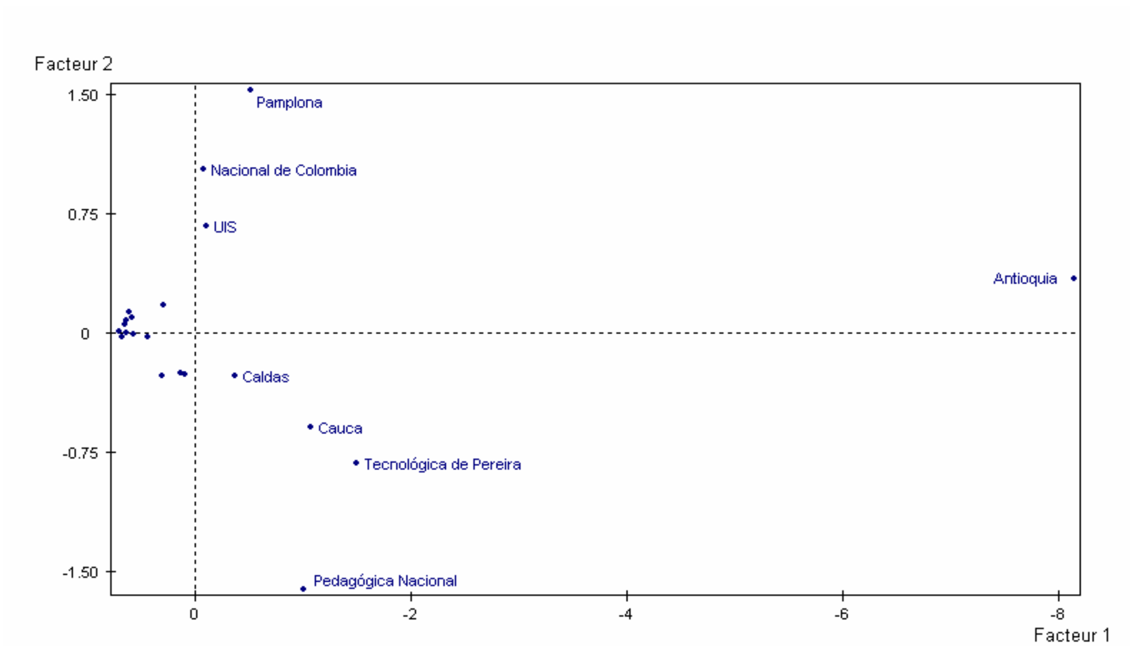


Gráfico 4.31. Universidades

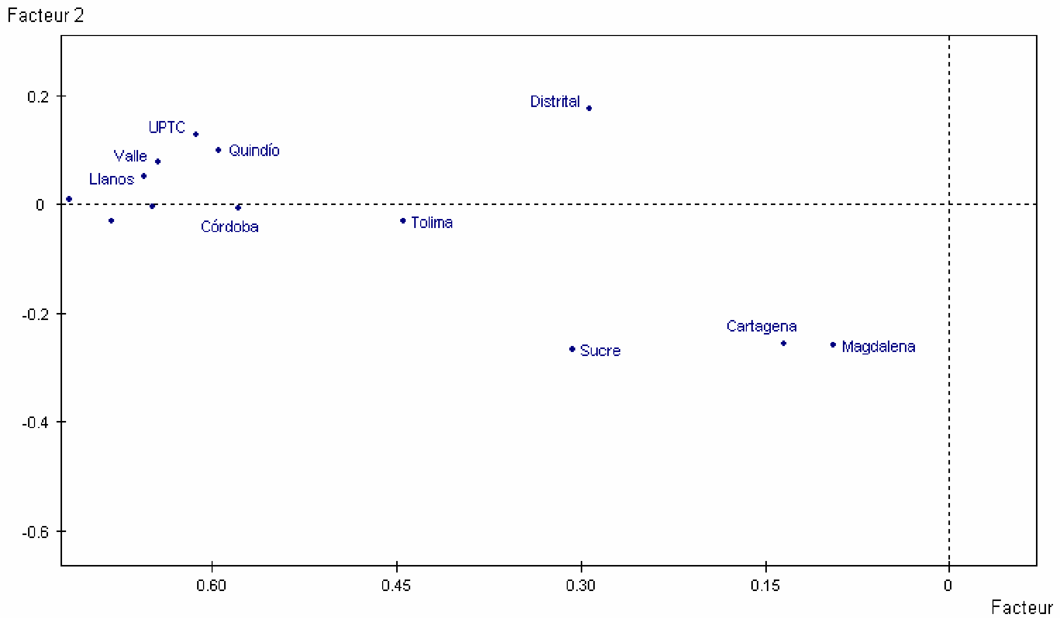


Gráfico 4.31. Zoom Nº1 sobre el Gráfico de las Universidades

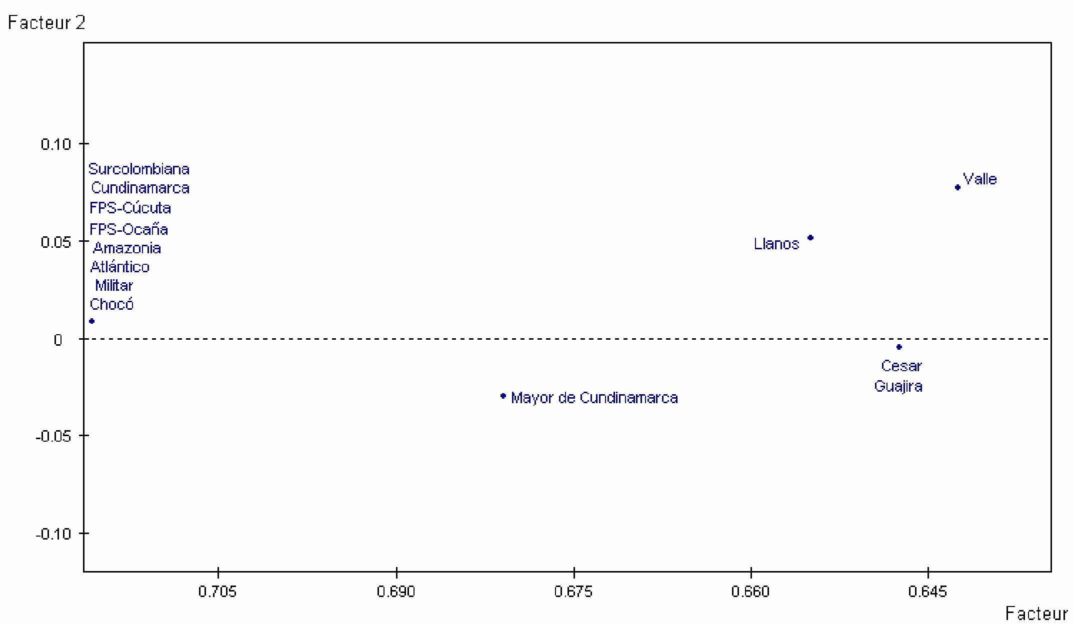


Gráfico 4.31. Zoom Nº2 sobre el Gráfico de las Universidades

Tabla 4.36. Coordenadas de las universidades sobre el índice de Producción

Identificación	Eje 1
Antioquia	8,15
Tecnológica de Pereira	1,50
Cauca	1,07
Pedagógica Nacional	1,00
Pamplona	0,51
Caldas	0,37
UIS	0,10
Nacional de Colombia	0,08
Magdalena	-0,09
Cartagena	-0,13
Distrital	-0,29
Sucre	-0,31
Tolima	-0,44
Córdoba	-0,58
Quindío	-0,59
UPTC	-0,61
Valle	-0,64
Cesar	-0,65
Guajira	-0,65
Llanos	-0,65
Mayor de Cundinamarca	-0,68
Atlántico	-0,72
Amazonia	-0,72
Cundinamarca	-0,72
Militar	-0,72
Surcolombiana	-0,72
FPS-Cúcuta	-0,72
FPS-Ocaña	-0,72
Chocó	-0,72
Nariño	-0,72

Con ayuda de las gráficas 4.31, los Zoom Nº1 y Nº2, y la Tabla 4.36, los grupos de acuerdo a la producción intelectual de los profesores de las universidades, quedan constituidos en la siguiente forma, donde el orden se asocia con la producción intelectual de los profesores de la universidad.

Estos nuevos grupos no son exactamente iguales a los producidos en el análisis univariado, existen algunas variaciones.

Universidad muy grande: Antioquia.

Universidades grandes: Tecnológica, Cauca y Pedagógica.

Universidades medianas: Pamplona, Caldas, UIS y Nacional de Colombia.

Universidades pequeñas: Magdalena, Cartagena, Distrital, Sucre, Tolima.

Universidades muy pequeñas: Córdoba, Quindío, UPTC, Valle, Cesar, Guajira, Llanos.

Universidades muy muy pequeñas: Mayor de Cundinamarca, Atlántico, Amazonia, Cundinamarca, Militar, Surcolombiana, FPS-Cúcuta, FPS-Ocaña, Chocó y Nariño.

Como se esperaba la universidad de Antioquia se considera una universidad grande, lo cual ratifica lo realizado en el análisis univariado.
El grupo de las universidades grandes se destaca ya que las universidades Tecnológica, Cauca y Pedagógica pasaron de ser universidades medianas a ser universidades grandes en el índice de Producción.

En el grupo de las universidades medianas se destacan las universidades Nacional y Pamplona, por que después de ser universidades muy grande y pequeña respectivamente ahora son consideran medianas.

En el grupo de las universidades pequeñas podemos destacar las universidades Distrital y Sucre por que pasaron de ser universidades medianas y muy pequeña (casi en el último lugar) respectivamente a ser universidades pequeñas.

Algo que llama la atención dentro del grupo de las universidades pequeñas es la universidad del Valle por que después de ser una universidad grande ahora es pequeña.

C) ACP para las variables de Formación: Graduados, Programas, Primíparos y Estudiantes.

En la Tabla 4.37 se puede apreciar la alta correlación que existen entre las variables de Formación lo que nos permite predecir que una sola componente principal dará cuenta de un alto porcentaje de la variación total de las cuatro variables.

Tabla 4.37. Matriz de Correlaciones

	Graduados	Programas	Primíparos	Estudiantes
Graduados	1,00			
Programas	0,88	1,00		
Primíparos	0,88	0,90	1,00	
Estudiantes	0,94	0,94	0,97	1,00

En la Tabla 4.38 se observa que , efectivamente, el primer valor propio explica el 93.69% de la variación total. Por esta razón solo trabajaremos con el primer factor o componente.

Tabla 4.38. Valores propios de la variables de Formación

Número	Valor propio	Porcentaje	Porcentaje acumulado
1	3,7477	93,69	93,69
2	0,1267	3,17	96,86
3	0,1077	2,69	99,55
4	0,0179	0,45	100,00

La tabla 4.39 da los coeficientes que acompañan a las combinaciones lineales de las variables Graduados, Programas, Primíparos y Estudiantes para definir los nuevos ejes, es decir, las componentes principales.

Tabla 4.39. Ejes unitarios

Identificación de la variable	Eje 1
Graduados	-0,49
Programas	-0,50
Primíparos	-0,50
Estudiantes	-0,51

Nuevamente tenemos que el eje 1 es una combinación lineal con coeficientes de tamaño parecido y del mismo signo, lo que nos permite interpretar este eje como el total de la Formación. Específicamente el Eje 1 es la combinación lineal:

$$-0.49 * Graduados - 0.50 * \text{Programas} - 0.50 * \text{Primíparos} - 0.51 * \text{Estudiantes}$$

No obstante el eje esta definido con coordenadas negativas, es decir el aumento esta de derecha a izquierda pero para mayor formalidad decidimos cambiar la dirección de los ejes de tal forma que el aumento sea de izquierda a derecha.

La posición relativa de las universidades se presenta en el gráfico 4.32.

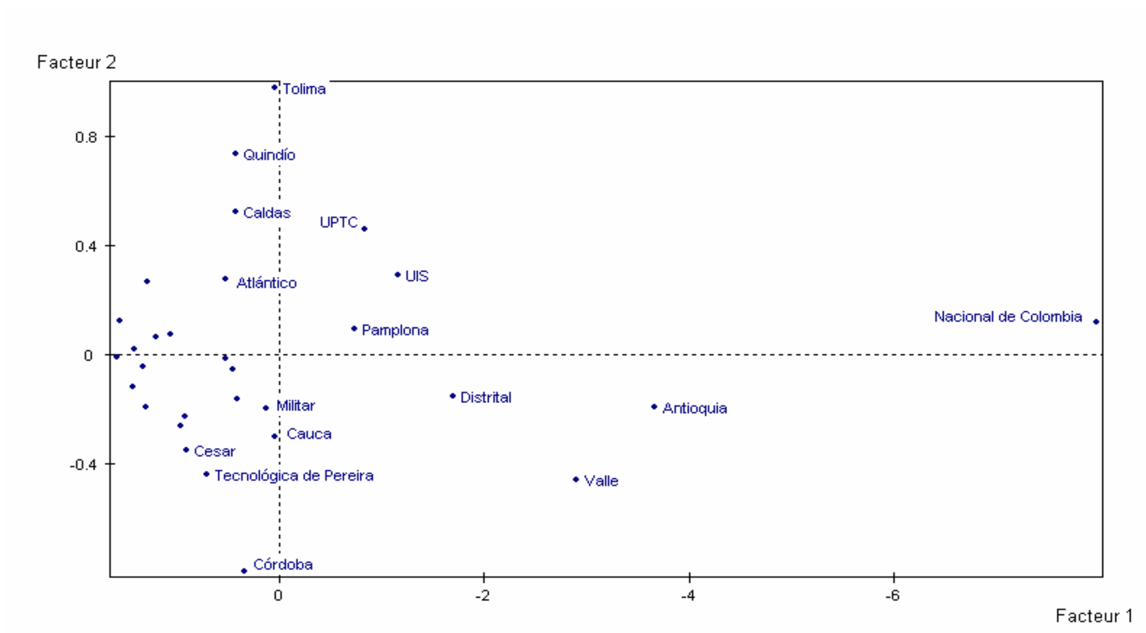


Gráfico 4.32. Grafico de las Universidades sobre el índice de Formación

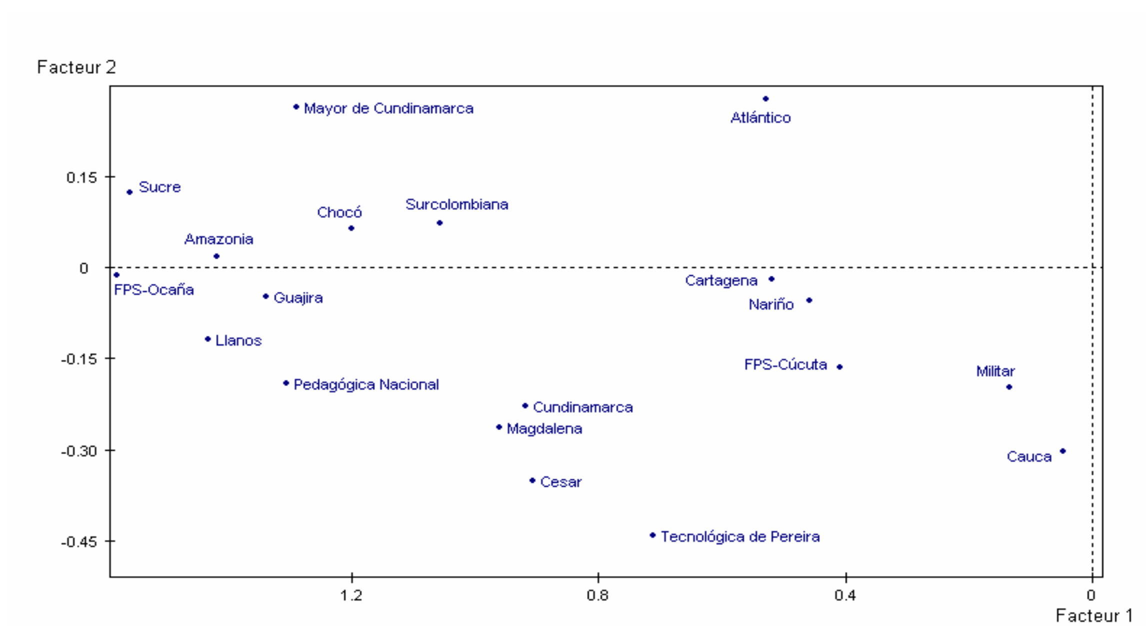


Gráfico 4.32. Zoom del gráfico de las Universidades en el índice de Formación

Tabla 4.40. Coordenadas de las universidades sobre el índice de Formación

Identificación	Eje 1
Nacional de Colombia	7,99
Antioquia	3,66
Valle	2,91
Distrital	1,69
UIS	1,17
UPTC	0,84
Pamplona	0,74
Cauca	-0,05
Tolima	-0,05
Militar	-0,13
Córdoba	-0,34
FPS-Cúcuta	-0,41
Quindío	-0,42
Caldas	-0,43
Nariño	-0,46
Cartagena	-0,52
Atlántico	-0,53
Tecnológica de Pereira	-0,71
Cesar	-0,91
Cundinamarca	-0,92
Magdalena	-0,96
Surcolombiana	-1,06
Chocó	-1,20
Mayor de Cundinamarca	-1,29
Pedagógica Nacional	-1,31
Guajira	-1,34
Amazonia	-1,42
Llanos	-1,43
Sucre	-1,56
FPS-Ocaña	-1,58

Con ayuda de las gráficas 4.32, el Zoom y la Tabla 4.40, los grupos de acuerdo a la Formación de las universidades, quedan constituidos en la siguiente forma, donde el orden se asocia con la Formación estudiantil.

Estos nuevos grupos no son exactamente iguales a los producidos en el análisis univariado, existen algunas variaciones.

Universidad muy grande: Nacional de Colombia.

Universidades grandes: Antioquia, Valle.

Universidades medianas: Distrital, UIS, UPTC y Pamplona.

Universidades pequeñas: Cauca, Tolima, Militar y Córdoba.

Universidades muy pequeñas: FPS-Cúcuta, Quindío, Caldas, Nariño, Cartagena, Atlántico, Tecnológica, Cesar, Cundinamarca, Magdalena.

Universidades muy muy pequeñas: Surcolombiana, Chocó, Mayor de Cundinamarca, Pedagógica, Guajira, Amazonia, Llanos, Sucre y FPS-Ocaña.

Las universidades Valle, Antioquia y Nacional consideradas universidades grandes y muy grande respectivamente por el índice de Capacidad son universidades que por sus niveles de Formación son de igual forma universidades grandes en este índice.

La universidad de Pamplona es una universidad pequeña que en términos de su Formación estudiantil se encuentra ubicada en el grupo de las universidades medianas.

Para el resto de los grupos podría decirse que se esta cumpliendo con el principio básico: "A mayor índice de Capacidad mejor es su índice de Formación".

D) ACP para las variables de Investigación: GruposInv, Revistas y Proyectos.

En la Tabla 4.41 se puede apreciar la alta correlación que existen entre las variables de Investigación lo que nos permite predecir que una sola componente principal dará cuenta de un alto porcentaje de la variación total de las tres variables.

Tabla 4.41.Matriz de correlaciones

	GruposInv	Revistas	Proyectos
GruposInv	1,00		
Revistas	0,91	1,00	
Proyectos	0,94	0,90	1,00

En la Tabla 4.42 se observa que , efectivamente, el primer valor propio explica el 94.53% de la variación total. Por esta razón solo trabajaremos con el primer factor o componente.

Tabla 4.42 valores propios

Número	Valor propio	Porcentaje	Porcentaje acumulado
1	2.8359	94,53	94,53
2	0,1013	3,38	97,91
3	0,0628	2,09	100,00

La Tabla 4.43 da los coeficientes que acompañan a las combinaciones lineales de las variables GruposInv, Revistas y Proyectos para definir los nuevos ejes, es decir, las componentes principales.

Tabla 4.43.Ejes unitarios

Identificación de la variable	**Eje 1**
GruposInv	-0,58
Revistas	-0,57
Proyectos	-0,58

Una vez más tenemos que el eje 1 es una combinación lineal con coeficientes de tamaño parecido y del mismo signo, lo que nos permite interpretar este eje como la difusión de los resultados de Investigación de la universidad. Específicamente el Eje 1 es la combinación lineal: $-0.58 * GruposInv - 0.57 * \text{Revistas} - 0.58 * \text{Proyectos}$.

No obstante el eje esta definido con coordenadas negativas, es decir el aumento esta de derecha a izquierda pero para mayor formalidad decidimos cambiar la dirección de los ejes de tal forma que el aumento sea de izquierda a derecha.

La posición relativa de las universidades se presenta en el gráfico 4.32.

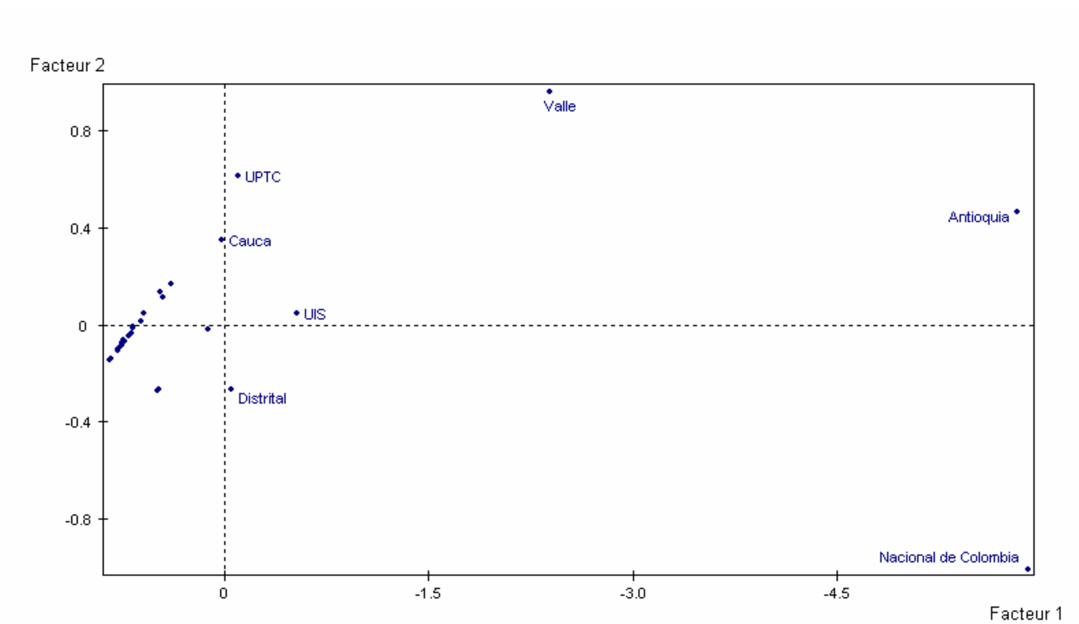


Gráfico 4.33. Gráfico de las universidades sobre el índice de Investigación

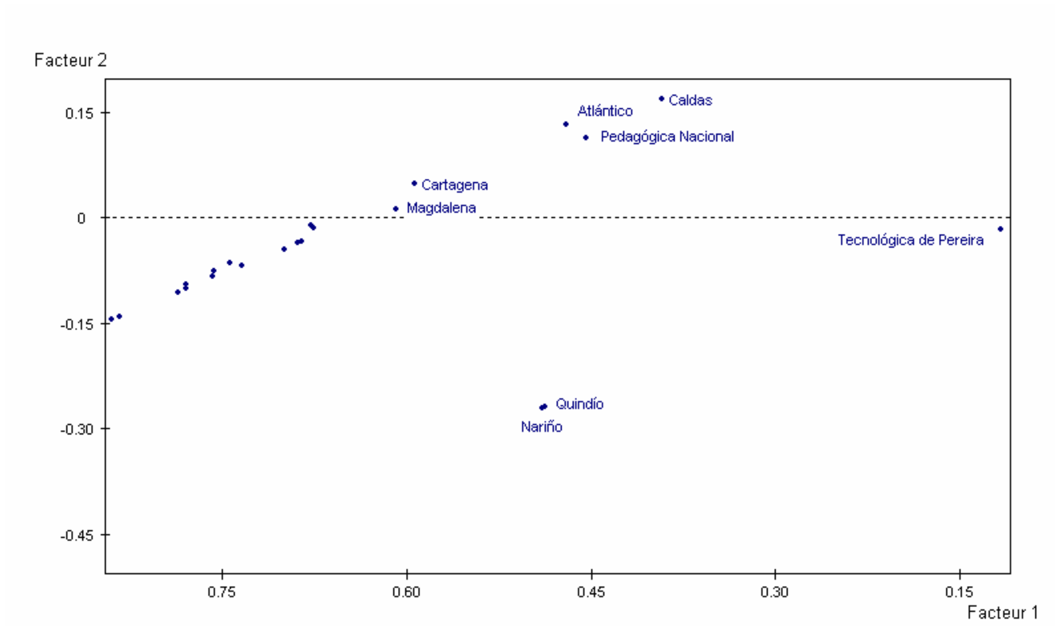


Gráfico 4.33. Zoom Nº1 de las universidades sobre el índice de Investigación.

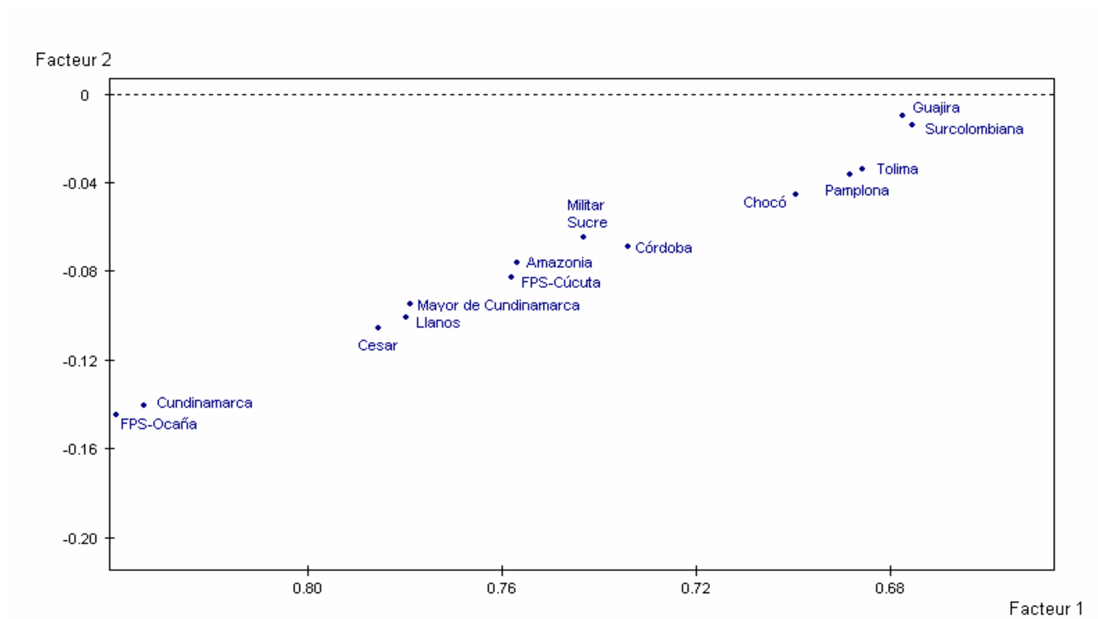


Gráfico 4.33. Zoom Nº2 de las universidades sobre el índice de Investigación

Tabla 4.44. Coordenadas de las universidades sobre el índice de investigación.

Identificación	Eje 1
Nacional de Colombia	5,91
Antioquia	5,82
Valle	2,39
UIS	0,53
UPTC	0,10
Distrital	0,05
Cauca	0,02
Tecnológica de Pereira	-0,12
Caldas	-0,39
Pedagógica Nacional	-0,45
Atlántico	-0,47
Nariño	-0,49
Quindío	-0,49
Cartagena	-0,59
Magdalena	-0,61
Guajira	-0,68
Surcolombiana	-0,68
Pamplona	-0,69
Tolima	-0,69
Chocó	-0,70
Córdoba	-0,73
Sucre	-0,74
Militar	-0,74
Amazonia	-0,76
FPS-Cúcuta	-0,76
Mayor de Cundinamarca	-0,78
Llanos	-0,78
Cesar	-0,79
Cundinamarca	-0,83
FPS-Ocaña	-0,84

Con ayuda de las gráficas 4.33, los Zoom Nº1 y Nº2, y la Tabla 4.44, los grupos de acuerdo a la Difusión de los resultados de Investigación de las universidades, quedan constituidos en la siguiente forma, donde el orden se asocia con la Formación estudiantil.

Estos nuevos grupos no son exactamente iguales a los producidos en el análisis univariado, existen algunas variaciones.

Universidad muy grande: Nacional de Colombia, Antioquia.

Universidades grandes: Valle.

Universidades medianas: UIS, UPTC y Distrital.

Universidades pequeñas: Tecnológica, Caldas, Pedagógica, Atlántico, Nariño y Quindío.

Universidades muy pequeñas: Cartagena, Magdalena, Guajira, Surcolombiana, Pamplona, Tolima y Chocó.

Universidades muy muy pequeñas: Córdoba, sucre, Militar, Amazonia, FPS-Cúcuta, Mayor de Cundinamarca, Llanos, Cesar, Cundinamarca y FPS-Ocaña.

Las universidades Valle, Antioquia y Nacional consideradas universidades grandes y muy grande respectivamente por el índice de Capacidad son nuevamente universidades grandes, ya que su valor en la difusión de los resultados de investigación es muy representativo.

Para el resto de los grupos existen pequeñas variaciones que no vale resaltar sin embargo estos grupos no han cambiado mucho en comparación con los ya formados por el índice de Capacidad y el análisis univariado, de hecho podría afirmarse que se esta cumpliendo con un principio básico: “A mayor índice de Capacidad mejor es su índice de Investigación”.

5. CONCLUSIONES

Los indicadores de gestión que se consideraron en este trabajo fueron clasificados en cuatro grandes grupos: “Variables de Capacidad”, “Variables de Producción”, “Variables de Investigación” y “Variables de formación”. La variable de ingresos de extensión y la de programas acreditados se analizaron independientemente.

En el grupo de las Variables de Capacidad se encuentran: Los recursos económicos de la universidad bien como aportes del gobierno o propios, y el número de profesores.
En el grupo de las Variables de Producción se encuentran: los puntos salariales adjudicados por producción de libros de texto, libros de investigación y artículos publicados en revistas nacionales e internacionales indexadas.
En el grupo de las Variables de Investigación se encuentran: el número de Revistas, Grupos y Proyectos de investigación reconocidos por Colciencias.
En el grupo de las Variables de Formación se encuentran: el número de Programas, Primíparos, Estudiantes y Graduados de todos los niveles de formación, modalidades de enseñanza y áreas del conocimiento.

El análisis de componentes principales nos permitió obtener índices únicos para estos grupos de variables. Estos índices los identificamos en forma similar, esto es, índice de Capacidad, índice de Producción, índice de Investigación e índice de Formación.

Realizamos una clasificación de las universidades de acuerdo al índice de Capacidad y lo tomamos como referencia para comparar los resultados de ellas en cada uno de los demás grupos. La clasificación es la siguiente:

1. Universidad muy grande: Nacional de Colombia
2. Universidades grandes: Antioquia y Valle
3. Universidades medianas: UPTC, UIS, Distrital, Cauca, Atlántico, Caldas, Tecnológica de Pereira y Córdoba.
4. Universidades pequeñas: Pamplona, Quindío, Nariño, Cartagena, Surcolombiana, Tolima, Pedagógica Nacional, Militar y FPS-Cúcuta.
5. Universidades muy pequeñas: Cesar, Cundinamarca, , Chocó, Magdalena, Llanos, Mayor de Cundinamarca, Amazonia, Guajira y Sucre.
6. Y por último la Universidad muy muy pequeña: FPS-Ocaña.

Al comparar los resultados del índice de Capacidad y el índice de Producción podemos observar cosas tan particulares como:
Se esperaba la universidad de Antioquia se considera una universidad grande, lo cual ratifica lo realizado en el análisis univariado.

El grupo de las universidades grandes se destaca ya que las universidades Tecnológica, Cauca y Pedagógica pasaron de ser universidades medianas a ser universidades grandes en el índice de Producción.
En el grupo de las universidades medianas se destacan las universidades Nacional y Pamplona, por que después de ser universidades muy grande y pequeña respectivamente ahora son consideran medianas.
En el grupo de las universidades pequeñas podemos destacar las universidades Distrital y Sucre por que pasaron de ser universidades mediana y muy pequeña (casi en el último lugar) respectivamente a ser universidades pequeñas.
Algo que llama la atención dentro del grupo de las universidades pequeñas es la universidad del Valle por que después de ser una universidad grande ahora es pequeña.

Al comparar los resultados del índice de Capacidad y el índice de Investigación podemos observar:
Las universidades Valle, Antioquia y Nacional consideradas universidades grandes y muy grande respectivamente por el índice de Capacidad son nuevamente universidades grandes, ya que su valor en la difusión de los resultados de investigación es muy representativo.
Para el resto de los grupos existen pequeñas variaciones que no vale resaltar sin embargo estos grupos no han cambiado mucho en comparación con los ya formados por el índice de Capacidad y el análisis univariado, de hecho podría afirmarse que se esta cumpliendo con un principio básico: “A mayor índice de Capacidad mejor es su índice de Investigación”.

Al comparar los resultados del índice de Capacidad y el índice de de Formación podemos observar que:
Las universidades Valle, Antioquia y Nacional consideradas universidades grandes y muy grande respectivamente por el índice de Capacidad son universidades que por sus niveles de Formación son de igual forma universidades grandes en este índice.
La universidad de Pamplona es una universidad pequeña que en términos de su Formación estudiantil se encuentra ubicada en el grupo de las universidades medianas.
Para el resto de los grupos podría decirse que se esta cumpliendo con el principio básico: “A mayor índice de Capacidad mejor es su índice de Formación”.

En cuanto a la variable de programas acreditados es importante mencionar que la universidad de Antioquia es una de las universidades con más programas acreditados, sin embargo la cifra de ceros en esta variable es alarmante, lo cual permite a las universidades ponerse al día con la acreditación de calidad de sus programas.

Por último La universidad de Antioquia podría considerarse una universidad muy completa, ya que su valor en todas los índices definidos gracias al ACP es alto.

REFERENCIAS BIBLIOGRÁFICAS

- SISTEMA DE UNIVERSIDADES ESTATALES, SUE, (2001). “Indicadores de gestión para las universidades públicas”, Bogotá.

- SISTEMA DE UNIVERSIDADES ESTATALES, SUE, (2003). “Indicadores de gestión para las universidades públicas”, Bogotá.

- ESCOFIER, BRIGITTE y PAGÉS, JEROME, (1992). “Análisis factoriales simples y múltiples objetivos, métodos e interpretación”, Bilbao.

- GARCIA, HERNÁN, (1998). “Análisis Multivariado de Datos”, San Juan de Pasto.

- JOLLIFFE, T. (1986) “Principal Component Analysis”. Springer- Verlag.

- SECRETARIA GENERAL DE LA ORGANIZACIÓN DE LOS ESTADOS AMERICANOS, (1986). “Análisis Multivariado: Método de Componentes Principales”. Washington, D.C.