Benemérita Universidad
Autónoma de Puebla
Facultad de Ciencias físico-matemáticas
licenciatura en matemáticas

Introducción al análisis de componentes principales

TESIS

QUE PARA OBTENER EL TITULO DE:

LICENCIADO EN MATEMÁTICAS

PRESENTA:

FABIOLA BLANCO INFANSON

DIRECTOR DE TESIS:

DR. FERNANDO VELASCO LUNA

PUEBLA,PUE., MAYO 2021

Dedicado a

La memoria de mi padre y amigo Fortino Blanco Aparicio

II

Agradecimientos

Quiero empezar agradeciendo a mi asesor el Dr. Velasco, el cual ha sido mi profesor en múltiples cursos de la licenciatura y que acepto ser mi asesor, siempre le estaré agradecida por todo lo que me ha enseñado, la paciencia tenida y la oportunidad que me brindo al poder trabajar con usted.
En segundo lugar, quisiera agradecer a mis sinodales el Dr. Victor, el Dr. Tajonar y el Dr. Hugo por todas las observaciones realizadas, y todo el tiempo invertido en revisar este trabajo.
Por último pero no menos importante quiero agradecer a todas las personas que han sido mi apoyo en estos años, empezando con mis padres, a los cuales doy gracias por la oportunidad brindada, por la paciencia tenida a lo largo de estos años, por participar en este loco sueño, y por quererme tanto, a mi mamá Eva por todo el apoyo brindado, aunque desgraciadamente no llego a ver la culminación, estoy segura que ella seria la primera en estar orgullosa de este logro.
Quiero agradecer a Misael porque nunca me dejo sola en ningún curso de la licenciatura y la amistad que logramos formar, es una de las mejores experiencias que me llevó de toda mi estadía en la universidad, además de un montón de anécdotas que hicieron que la universidad fuera una época

memorable, a Ángel Michel por ser mi mejor amigo y mi apoyo emocional todos estos años, por aguantarme cuando solo hablaba de la universidad o de la tesis, por estar allí en los momentos buenos y malos, me alegra mucho haberte conocido, haces que mi vida sea más divertida y no puedo darte las suficientes gracias, por todo el apoyo que me has dado.

A Juan y Rodolfo por no dejar que desistiera en los últimos cursos y por la ayuda brindada en la relización de esta tesis, gracias, y a todos los demás que estuvieron allí en alguna parte de este largo camino que fue la licenciatura, ya sea al inicio o final les doy gracias por mejorar esta experiencia, por las risas compartidas, las tareas repartidas, y los momentos vividos, muchas gracias.

V

Introducción al análisis de componentes principales

Índice general

Introducción al análisis de componentes principales

Fabiola Blanco Infanson

fecha

Introducción

0.1. Marco contextual

En la literatura tradicional de Estadística podemos encontrar métodos desarrollados en una variable,sin embargo en la vida real, se tiene que los eventos por lo general implican, varias características de interés, es decir, varias variables aleatorias. Por ejemplo un investigador de mercados podría querer identificar las características de los individuos que le permitirían determinar, si es probable que determinada persona compre un producto específico, o en el caso de un agrónomo podría interesarse en la resistencia de nuevas variedades de trigo y la resistencia que tienen estas a la sequía y los insectos. En ocasiones los investigadores tienen o recolectan información de un gran número de variables, lo cual dificulta su posible interpretación para dar solución al problema que ocupa su interés. Por lo regular cuando se tiene un gran número de variables estás se encuentran relacionadas por lo cual es deseable reducir el número de variables pero sin perder la información de las originales. Una técnica estadística para reducir el número de variables originales a un conjunto menor, es la técnica de componentes principales (ACP) por sus siglas en español (Análisis de Componentes Principales).

Podemos decir que los objetivos del análisis de componentes principales son 1) reducir dimensionalidad de los datos, y 2) encontrar nuevas variables importantes subyacentes.

Aunque autores como Araneo D.[1], Tenko Raykov [6], entre otros, confirma que el objetivo 1 no siempre se llega a obtener, se obtiene la relación entre las variables, es decir la dimensionalidad de los datos y en el caso del objetivo 2 aunque no se consiga, que las nuevas variables sean significativas es decir que tengan una interpretación, estás todavía pueden ser útiles, por diversos motivos, como el cribado de datos y verificación de las agrupaciones.

En las ciencias del comportamiento, sociales y educativas, el ACP tiene una historia relativamente larga de aplicaciones en el desarrollo de pruebas objetivas para medir habilidades específicas, motivación, personalidad, inteligencias y otras construcciones relacionadas o dimensiones latentes (no observables, ocultas). En estas aplicaciones, se comienza típicamente con un gran conjunto de medidas destinadas a evaluar esas construcciones. Luego se emplea la técnica ACP en los datos obtenidos con el objetivo de reducir su extensión a unos pocos componentes significativos que representan medidas "puras"de las dimensiones o variables latentes subyacentes.

El ACP no sólo logra la reducción de variables, sino que también el resultado puede ser usado en aplicaciones de otros métodos estadísticos multivariados (análisis de varianza o análisis de regresión).

En conclusión este método brinda la posibilidad de analizar la dimensión de conjuntos de datos y la relación que hay entre ellos, lo que nos brinda un mundo de posibilidades, ya que es una técnica exploratoria muy útil, en la aplicación de otros métodos estadísticos, y en algunos casos ayudará a la

reducción de la dimensión del problema a costa de una pequeña pérdida de información. Aunque hay una tendencia a interpretar las nuevas variables recién creadas, esto no siempre sucede, son pocos casos pero hay que recordar que aunque esto no suceda el análisis de componentes principales es muy útil.

0.2. Antecedentes

El análisis de componentes principales es una técnica estadística de análisis multivariado que se emplea para extraer información relevante de un conjunto inicial de variables correlacionadas transformándolas en variables no correlacionadas, con el objeto de identificar patrones y estructuras. Esta técnica es utilizada por diversos investigadores de múltiples áreas de estudio.

Bajo M. en [3], utiliza la técnica de ACP para determinar los principales factores de riesgo de la curva de rendimientos, con un énfasis especial en la gestión activa de carteras de renta fija, donde proporciona el enfoque del gestor de carteras o *practitioner* mediante el análisis se pudo reducir el estudio de un número elevado de parámetros, (como son los tipos de interés) a un conjunto reducido de componentes que representan los principales factores de riesgo a los que se enfrenta el gestor.

Además que el ACP reveló que la decisión de apuesta por un mercado alcista o bajista de tipos es mucho más importante, en términos de retorno, que aquellas decisiones de valor relativo sobre distintas zonas de la ETTI (Estructura Temporal de Interés) como pueden ser posiciones en pendiente (flattening o steepening) o estrategias de inversión basadas en el análisis de

la curvatura (butterflies).

Mesa-Ramos L.,*et al* [5] utilizaron la técnica para el análisis del proceso de fermentación de un anticuerpo monoclonal, en el Centro de Inmunología Molecular (La Habana, Cuba) se produce un anticuerpo monoclonal terapéutico que ha encontrado una efectiva aplicación en el tratamiento de pacientes aquejados de cáncer de cabeza y cuello. Dada la gran variabilidad que ha tenido la concentración de este anticuerpo en la etapa de fermentación industrial de la planta donde es producido, se hizo necesaria la aplicación de una técnica de análisis multivariante como el Análisis de Componentes Principales, con el fin de reducir la dimensionalidad de los datos y de explicar las principales fuentes de variabilidad del proceso.

Ávila H.,*et al*[2] aplicaron el análisis de componentes principales con el objeto de conocer las interrelaciones entre las variables analizadas que determinan el grado de alteración del agua de la Laguna de Coyuca de Benítez, en Guerrero, México. El ACP arrojó que las variables Temperatura, pH y Oxígeno disuelto presentaron mayores interrelaciones en este sistema lentico.

Se espera que el presente trabajo pueda ayudar a los estudiantes de licenciatura a entender mejor la técnica de análisis de componentes principales.

0.3. Objetivos de la tesis

Objetivo General

Presentar una introducción de la técnica de análisis de componentes principales, dándole un enfoque más sencillo y entendible para los estudiantes de licenciatura, usando softwares estadísticos.

Objetivos particulares

- Presentar la parte teórica de la técnica de componentes principales
- Aplicar la técnica de componentes principales a una base de datos
- Presentar la forma en que se implementa en R y SPSS

Este trabajo de tesis se divide en cuatro capítulos, al inicio se presenta la introducción del tema, en el primer capítulo se analiza la teoría que está detrás de la técnica estadística, y en el segundo y tercer capítulo se analizará la aplicación de la técnica a través de los software estadísticos en un ejemplo donde la base de datos, se obtuvo de la literatura y el otro ejemplo analizado es de una base de datos de calificaciones de las primeras materias que se cursan en las diversas carreras de la facultad de ciencias Físico-Matemáticas de la Benemérita Universidad Autónoma de Puebla.

Capítulo 1

La teoría de componentes principales

1.1. Descripción de la técnica

El análisis de componentes principales es una técnica estadística cuyo objetivo principal es transformar un conjunto de p variables correlacionadas a otras nuevas variables cuyo número es menor que p y están incorrelacionadas. En la mayoría de los textos se manejan dos objetivos del análisis de componentes principales, el anteriormente expuesto y el segundo es identificar nuevas variables significativas subyacentes. Aunque siempre se identificarán nuevas variables, no se puede garantizar que sean significativas, es decir tengan una interpretación, sin embargo aunque no sean significativas, estas variables serán útiles para diversas cosas, como cribado de datos y verificación de agrupaciones.

1.2. Preliminares

Definición 1.1 (Eigenvalor y eigenvector)**.** *Sean A una matriz $p \times p$ (con $p>1$) y un vector $\bar{x}$ diferente de cero. Decimos que $\bar{x}$ es un eigenvector de A, si cumple que $A\bar{x}$ es un múltiplo escalar de $\bar{x}$, es decir, $A\bar{x} = \lambda\bar{x}$ para algún escalar $\lambda \in \mathbb{R}$.*
El escalar λ se denomina eigenvalor de A y decimos que $\bar{x}$ es un vector correspondiente a λ.

Definición 1.2. *Los eigenvalores de una matriz cuadrada A son las soluciones de la ecuación $|A - \lambda I_p| = 0$, es decir, se nulifica el determinante de la matriz $A - \lambda I_p$.*

Proposición 1.1. *Todos los eigenvalores de una matriz simétrica son números reales, y sus vectores propios pueden elegirse para contener solo elementos reales. En adición, los eigenvectores correspondientes a diferentes valores son ortogonales uno a otro, y los eigenvectores pertenecientes a eigenvalores iguales pueden ser escogidos tal que sean ortogonales uno a otro.*

Proposición 1.2 (Descomposición espectral)**.** *Si $\lambda_1, ..., \lambda_p$ y $\bar{e}_1, ..., \bar{e}_p$ son los valores propios correspondientes ($\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p$) y los vectores auto asociados de la matriz simétrica A, entonces esta matriz puede ser representada como la suma siguiente: $A = \lambda_1\bar{e}_1\bar{e}_1^T + \lambda_2\bar{e}_2\bar{e}_2^T + ... + \lambda_p\bar{e}_p\bar{e}_p^T$*

Teorema 1.1. *Teorema de Rouché Frobenius*

Dado un sistema de ecuaciones lineales con m ecuaciones y con n incógnitas:

$$\begin{aligned}
&a_{11}x_1 + a_{12}x_2 + ... + a_{1n}x_n = b_1,\\
&a_{21}x_1 + a_{22}x_2 + ... + a_{2n}x_n = b_2,\\
&.....\\
&a_{m1}x_1 + a_{m2}x_2 + ... + a_{mn}x_n = b_m.
\end{aligned}$$

Se llama matriz del sistema a la matriz, A, formada por los coeficientes de las incógnitas

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} & ... & a_{1n} \\ a_{21} & a_{22} & a_{23} & ... & a_{2n} \\ ... & ... & ... & ... & ... \\ a_{m1} & a_{m2} & a_{m3} & ... & a_{mn} \end{pmatrix}$$

y matriz ampliada A', a la matriz del sistema ampliado con los términos independientes

$$A' = \left(\begin{array}{cccc|c} a_{11}x_1 & a_{12}x_2 & ... & a_{1n}x_n & b_1 \\ a_{21}x_1 & a_{22}x_2 & ... & a_{2n}x_n & b_2 \\ ... & ... & ... & ... & ... \\ a_{m1}x_1 & a_{m2}x_2 & ... & a_{mn}x_n & b_m \end{array}\right)$$

Un sistema de ecuaciones lineales es compatible si y solo si el rango de la matriz de los coeficientes es igual al rango de la matriz ampliada.

1.3. Cálculo de los componentes principales

Sea una matriz X con $n \times p$ datos

$$X = \begin{pmatrix} x_{11} & x_{12} & ... & x_{1p} \\ x_{21} & x_{22} & ... & x_{2p} \\ ... & ... & ... & ... \\ x_{n1} & x_{n2} & ... & x_{np} \end{pmatrix},$$

de una población con matriz de covarianza $\sum$ o (correlación) para las p variables involucradas.

Por notación nos referiremos a las variables de este conjunto como elementos del vector aleatorio $\bar{x} = (x_1, x_2, ..., x_p)$, con el ACP intentaremos construir las siguientes combinaciones lineales:

$$\begin{aligned} y_1 &= \bar{a_1'}\bar{x} = a_{11}x_1 + a_{12}x_2 + ... + a_{1p}x_p, \\ y_2 &= \bar{a_2'}\bar{x} = a_{21}x_1 + a_{22}x_2 + ... + a_{2p}x_p, \\ &\vdots \\ y_p &= \bar{a_p'}\bar{x} = a_{p1}x_1 + a_{p2}x_2 + ... + a_{pp}x_p. \end{aligned} \tag{1}$$

donde $\bar{a_1}, a_2, \bar{...}, \bar{a_p}$ son desconocidos, los px1 vectores contienen los pesos de las combinaciones lineales. Necesitamos que la primera componente principal y_1 explique la mayor varianza posible, que resultara de las combinaciones lineales de las variables iniciales en $\bar{x}$, y ahora se necesita imponer una restricción para la magnitud de los elementos de $\bar{a_1}$, ya que, si no se hace, la var(y_1) aumentará desmedidamente, lo cual impide encontrar un vector $\bar{a_1}$.

La elección que por lo general se hace es asumir que la longitud de $\bar{a_1}$ es 1,

esto es $||\bar{a_1}|| = 1$ una vez que se encuentre la solución y se encuentre este vector, óptimo de pesos $\bar{a_1}$, la combinación lineal $\bar{a_1}'\bar{x}$ es llamado, la primera componente principal. El siguiente paso es encontrar un segundo vector $\bar{a_2}$ tal que la segunda componente $y_2 = \bar{a_2}'\bar{x}$ tenga las siguientes propiedades:

1. Debe de explicar la mayor varianza posible de los datos restantes (considerando de nuevo la restricción $||\bar{a_2}|| = 1$) y

2. No debe estar correlacionado con la primera componente principal.

Se busca que no tengan correlación, ya que buscamos terminar con el menor número de combinaciones lineales que expliquen la porción mas grande posible de las varianzas de las variables iniciales.
Una vez encontrado este conjunto de valores $\bar{a_2}$, las combinaciones lineales $\bar{a_2}'\bar{x}$ se denominará la segunda componente principal. Podemos continuar con este comportamiento hasta que encontremos p vectores $\bar{a_1}, ..., \bar{a_p}$ que aparecen en el lado derecho de las ecuaciones (1), por construcción las sucesivas componentes principales no están correlacionadas y formalmente hay tantas componentes principales como variables estudiadas.

1.3.1. Proceso de extracción de factores

Extracción de la primera componente

Se quiere elegir $\bar{a_1}$ de tal manera que y_1 tenga una varianza máxima, sujeta a la restricción $||\bar{a_1}|| = 1$ se tiene que:

$$Var(y_1) = Var(\bar{a_1}'\bar{x}) = \bar{a_1}'\Sigma\bar{a_1}. \tag{2}$$

Como se sabe se debe de imponer la restricción $\bar{a_1}'\bar{a_1} = 1$ a la maximización de $\bar{a_1}'\Sigma\bar{a_1}$ porque de lo contrario, se podría maximizar la varianza sin límite aumentando el módulo del vector $\bar{a_1}$, lo cual no ayudaría a la solución de la maximización.
Se introduce esta restricción a través del multiplicador de Langrage:

$$M(a_1) = \bar{a_1}'\Sigma\bar{a_1} - \lambda(\bar{a_1}'\bar{a}_1 - 1).$$

Se maximiza esta expresión derivando respecto a los componentes de $\bar{a_1}$ e igualando a 0. Entonces

$$\frac{\partial M}{\partial \bar{a_1}} = 2\Sigma\bar{a_1} - 2\lambda\bar{a_1} = 0 \Rightarrow (\Sigma - \lambda I)\bar{a_1} = 0.$$

Al ser esto un sistema de ecuaciones lineales, se tiene por el teorema de Roché-Frobenius,para que el sistema tenga una solución distinta de $\bar{0}$, la matriz $(\Sigma - \lambda I)$ tiene que ser singular:

$$|\Sigma - \lambda I| = 0.$$

Esto implica que, λ es un eigenvalor de Σ. La matriz de covarianzas Σ es de orden p y si además es definida positiva, tendrá p valores propios distintos, $\lambda_1, \lambda_2, \lambda_3, ..., \lambda_p$ tales que, por ejemplo, $\lambda_1 > \lambda_2 > \lambda_3 > .... > \lambda_p$. Desarrollando la expresión anterior, se obtiene que:

$$(\Sigma - \lambda I)\bar{a_1} = 0 \Longrightarrow \Sigma\bar{a_1} - \lambda I\bar{a_1} = 0$$

$$\Longrightarrow \Sigma\bar{a_1} = \lambda I\bar{a_1} \tag{3}$$

Esto implica que $\bar{a_1}$ es un eigenvector de la matriz Σ, y λ su correspondiente eigenvalor. Para determinar que valor propio de Σ es la solución de la ecuación (3), se multiplica por la izquierda por $\bar{a_1}'$ esta ecuación,

$$\bar{a_1}'\Sigma\bar{a_1} = \lambda\bar{a_1}'\bar{a_1} = \lambda$$

se concluye por (2), que λ es la varianza de y_1.

Dado que esta es la cantidad que se quiere maximizar, λ será el mayor eigenvalor de la matriz Σ. Su vector asociado, $\bar{a_1}$, define los coeficientes de cada variable en el primera componente principal.

Extracción de la segunda componente

La segunda componente se obtendrá utilizando un razonamiento similar al anterior. Se tiene que $Var(y_2) = \bar{a_2}'\Sigma\bar{a_2}$, aún mas, dado que se desea que y_1 y y_2, sean no correlacionadas, se verifica que $Cov(y_2, y_1) = 0$. Es decir:

$$Cov(y_2, y_1) = Cov(\bar{a_2}'\bar{x}, \bar{a_1}'\bar{x}) =$$

$$= \bar{a_2}' \cdot E[(\bar{x} - \mu)(\bar{x} - \mu)'] \cdot \bar{a_1} =$$

$$= \bar{a_2}'\Sigma\bar{a_1} \tag{4}$$

Entonces $\bar{a_2}'\Sigma\bar{a_1}$ debe ser necesariamente igual a cero.

Anteriormente se obtuvo que $\Sigma\bar{a_1} = \lambda\bar{a_1}$ y esto sustituyéndolo en la ecuación (4), se obtiene que:

$$\bar{a_2}'\Sigma\bar{a_1} = \lambda\bar{a_2}'\bar{a_1} = 0 \Rightarrow$$

$$\Rightarrow \bar{a_2}'\bar{a_1} = 0 \Rightarrow \bar{a_2}' \perp \bar{a_1}$$

Se tiene que maximizar $\bar{a_2}'\Sigma\bar{a_2}$, considerando las siguientes restricciones:

$$\bar{a_2}'\bar{a_2} = 1 \tag{5}$$

$$\bar{a_2}'\bar{a_1} = 0 \tag{6}$$

Usando los multiplicadores de Langrange, se obtiene la función

$$M(\bar{a_2}) = \bar{a_2}'\Sigma\bar{a_2} - \lambda(\bar{a_2}'\bar{a_2} - 1) - \gamma\bar{a_2}'\bar{a_1}$$

Se deriva respecto a $\bar{a_2}$ y se iguala a cero

$$\frac{\partial M}{\partial \bar{a_2}} = 2\Sigma\bar{a_2} - 2\lambda\bar{a_2} - \gamma\bar{a_1} = 0 \tag{7}$$

Se sabe que $\bar{a_1}'\bar{a_2} = \bar{a_2}'\bar{a_1} = 0$ y que $\bar{a_1}'\bar{a_1} = 1$, si se multiplica la ecuación (7) por $\bar{a_1}'$, se obtiene

$$2\bar{a_1}'\Sigma\bar{a_2} - \gamma = 0$$

Al despejar γ se tiene que

$$\gamma = 2\bar{a_1}'\Sigma\bar{a_2} = 2\bar{a_2}'\Sigma\bar{a_1}$$

Dado que $Cov(y_2, y_1) = 0$.

Finalmente, al sustituir la ecuación (7) se obtiene que:

$$\tfrac{\partial M}{\partial \bar{a_2}} = 2\Sigma\bar{a_2} - 2\lambda\bar{a_2} = (\Sigma - \lambda I)\bar{a_2} = 0$$

Siguiendo el razonamiento de la primera componente principal tenemos que solo será un sistema homogéneo si $|\Sigma - \lambda I| = 0$.

Si en la ecuación $(\Sigma - \lambda I)\bar{a_2} = 0$ se multiplica a la derecha $\bar{a_2}'$, se tiene que:

$$\bar{a_2}'(\Sigma - \lambda I)\bar{a_2} = 0 \Rightarrow \bar{a_2}'\Sigma\bar{a_2} = \lambda$$

Por lo tanto, para maximizar la $Var(y_2)$, se ha de tomar el segundo eigenvalor más grande de la matriz Σ,con su respectivo eigenvector asociado $\bar{a_2}$.

Los razonamientos anteriormente expuestos, se pueden extender al componente i-ésimo, al cual le correspondería el i-ésimo eigenvalor.

En realidad no se ha reducido la complejidad de los datos, lo que haremos es redistribuir la varianza de las variables originales del vector $\bar{x}$ en las varianzas derivadas de las medidas del vector $\bar{y}$. Sin embargo, en la práctica la mayoría de veces es posible explicar una cantidad suficientemente grande de varianza de las variables originales con solo unas componentes principales.

Se retomará los preliminares enunciados en el inicio del capítulo, ya que resulta que las combinaciones lineales de las variables originales que proporcionan la primera componente principal, tienen pesos que son exactamente los elementos del eigenvector (normalizado), perteneciente al mayor eigenvalor, λ_1, de $\sum$,donde $\sum$ es la matriz de covarianza (o correlación) de la población analizada. De manera similar, la combinación lineal de las variables originales que produce la segunda componente principal tiene pesos que son los elementos del segundo eigenvector, correspondiente al segundo eigenvalor mas grande, λ_2, de $\sum$ y así sucesivamente hasta el k eigenvalor, λ_k, de $\sum$, $(1 \leq k \leq p)$.

Así, lo que se necesita hacer para obtener las componentes principales es descubrir en orden los eigenvalores y eigenvectores (normalizados) de la matriz analizada. Se debe recordar que estos eigenvectores cumplen que son ortogonales entre si (Proposición 1.1).

En la práctica no se conoce la matriz de covarianza $\sum$ o correlación ρ para proceder con el ACP, generalmente se estima esta matriz de población con la matriz empírica de covarianza matriz S o la matriz empírica de correlación (R), estás se estiman a partir de una muestra dada.

Si se lleva a cabo este procedimiento, produce las componentes principales para un conjunto dado de variables analizadas. En su totalidad, estas P componentes representan la varianza de todo el conjunto de datos, y por tanto no representa ninguna reducción de información.

1.4. Determinación del número de componentes principales

Se necesita una regla que permita decidir cuando dejar de extraer los componentes principales.

Dicha regla se puede obtener de la siguiente manera,

$$Var(Y_i) = \lambda_i$$

donde λ_i es el i-ésimo eigenvalor de la matriz analizada S $o(R; i = 1, ..., p)$, es decir, el i-ésimo eigenvalor representa la varianza del i-ésimo componente principal. Aún mas se puede mostrar que:

$$Var(y_1)+Var(y_2)+...+Var(y_p) = Var(x_1)+Var(x_2)+...+Var(x_p) = \lambda_1+\lambda_2+...+\lambda_p.$$

Las últimas ecuaciones indican que la suma de las varianzas de todos las componentes principales, es igual a la suma de las varianzas de todas las variables observadas, lo que es igual a la suma de los eigenvalores. Esto es, la varianza del conjunto de datos original, es la suma de todos los eigenvalores, está se redistribuye a través de los componentes principales. Por tanto, la proporción

$$r_1 = \frac{\lambda_1}{\lambda_1 + \lambda_2 + ... + \lambda_p}$$

puede ser usada para representar la proporción de varianza explicada por la primera componente principal.

Generalizando, se puede considerar la proporción

$$r_k = \frac{\lambda_1 + \lambda_2 + ... + \lambda_k}{\lambda_1 + \lambda_2 + ... + \lambda_p}$$

como la proporción de varianza explicada por la k-ésima componente principal.

1.4.1. Método 1: Análisis de la matriz de covarianza y correlación.

Hay dos reglas para determinar el número de componetes principales:

Regla 1:

Cuando se analiza la matriz de covarianza del conjunto de variables, y se establece una proporción deseable de la varianza original explicada antes de examinar los datos, entonces se siguen extrayendo componentes principales hasta que el valor de r_k excede esta proporción especificada por primera vez.

Por ejemplo, en la mayoría de los casos se especifica que la proporción debe ser al menos del 80 % (véase [6]).

Regla 2:

En caso de que esta proporción deseable no se pueda especificar antes de mirar los datos, se pueden seguir extrayendo componentes principales hasta que se encuentre uno cuyo eigenvalor sea menor que el eigenvalor promedio de la matriz de covarianza analizada. Entonces todos los componentes extraídos, antes de este punto pueden verse como los que se deben conservar posteriormente.

Análisis de la matriz de correlación

Cuando la matriz de correlación es analizada, la Regla 1 no cambia, sin embargo, la Regla 2 se modifica de la siguiente forma: Se seguiran extrayendo componentes principales siempre que sus eigenvalores sean mayores que 1. Notamos que la última regla es de hecho idéntica a la Regla 2 mencionada anteriormente en el presente caso, ya que las variaciones de todas las variables son 1 aquí (porque se analiza la matriz de correlación) y, por lo tanto, su promedio también es 1. Esta versión es referida como criterio de valor propio de Kaiser (véase [6]).

1.4.2. Método 2: Diagrama de scree de catell

Para este método se utiliza una gráfica de SCREE de los eigenvalores, este método fue popularizado por Cattell (1966). Una gráfica SCREE se construye al situar el valor de cada eigenvalor contra el recíproco. Cuando los puntos de la gráfica tienden a nivelarse, estos eigenvalores están suficientemente cercanos a cero como para que se puedan ignorar. Es probable que los más pequeños estén midiendo nada más que ruido aleatorio y éste no se debe tratar de interpretar. Por tanto, este método descarta a las componentes que ya no aportan significativamente a la varianza. En la Figura 1.1, se muestra un ejemplo de una gráfica SCREE. Esta gráfica SCREE sugeriría que la dimensionalidad real del espacio en que se encuentran los datos es 2 y las componentes a utilizar también son dos.

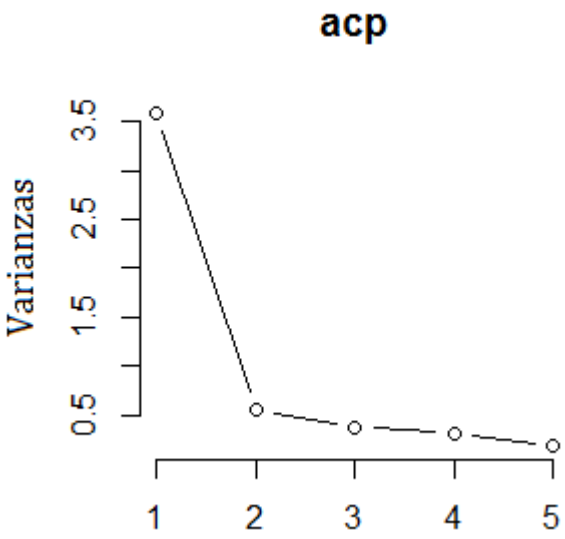


Figura 1.1: Gráfica SCREE

En la práctica los investigadores por lo general realizan y utilizan ambos criterios aquí presentados para tomar una decisión.

Capítulo 2

Medidas de pájaros

2.1. Aspectos generales

En este capítulo se abordará el caso que se obtuvo de la literatura analizada (véase [4]) , en este caso tenemos que los datos presentados, son las medidas de 49 pájaros que estuvieron en una tormenta el 1 de febrero de 1898, donde se consideran las siguientes variables.

Tabla 2.1: Medidas

Variable	Descripción	Valores
x_1	longitud total	(152,165)
x_2	extensión alar	(230,250)
x_3	longitud de pico y cabeza	(30.1,33.1)
x_4	longitud de humerus	17.2-19.8
x_5	longitud de la quilla del esternón	(19.0-23.1)

2.2. Análisis estadístico

2.2.1. Análisis preliminar

Se realizarán gráficas de cajas y bigotes de las variables longitud total, extensión alar, longitud de pico y cabeza, longitud de humerus y longitud de la quilla del esternón, se analizarán con el fin de observar el comportamiento de las variables en ambos casos.

2.2.2. Análisis definitivo

Se realizará un análisis de componentes principales, con el objetivo de reducir la dimensión del estudio, para generar nuevas variables que nos ayuden a clarificar, cuales fueron los factores que influyeron a la sobrevivencia de algunos pájaros y el fallecimiento de otros.

2.3. Resultados

2.3.1. Análisis preliminar

Análisis univariado

La tabla 2.2 presenta las variables junto con su media aritmética y su desviación estándar, en la cual se observa que los datos de las variables se encuentran en promedio cerca de su media.

Tabla 2.2: Media y desviación estándar de las variables

	Longitud Total	Extensión Alar	Longitud de Pico y Cabeza	Longitud de Humerus	Longitud de la Quilla del Esternón
Media	158	241	31	18	21
Desviación Estándar	3.6542	5.0678	0.7947	0.5937	0.9913

En el análisis preliminar se obtuvieron las siguientes gráficas (Figura 2.1, Figura 2.2, Figura 2.3, Figura 2.4 y Figura 2.5) donde los primeros 21 pájaros sobrevivieron y los otros 28 pájaros perecieron.

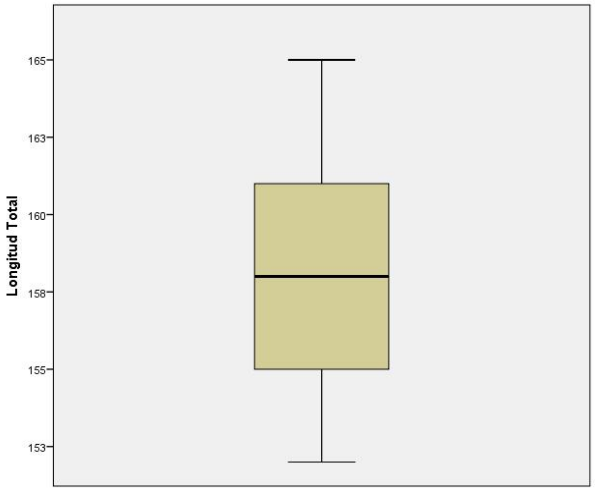


Figura 2.1: Longitud Total

Se nota que la variable de longitud total tiene una distribución asimétrica positiva, se puede observar en la gráfica 2.1 que se encuentra una mayor dispersión en los datos cuyos valores oscilan entre 161 y 165.

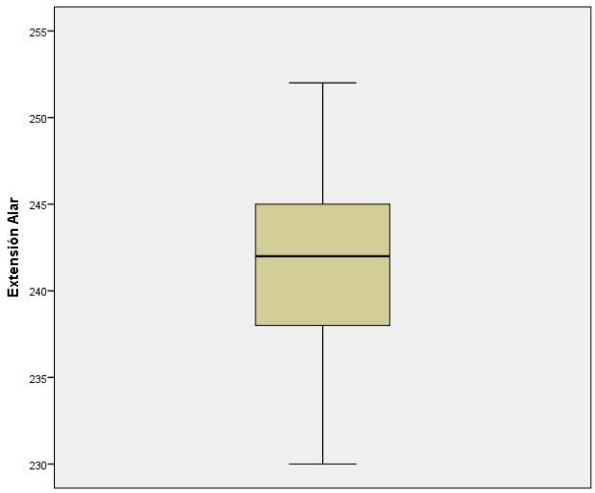


Figura 2.2: Extensión Alar

La variable extensión alar tiene una distribución sesgada a la izquierda, la menor dispersión de los datos se encuentra entre las medidas de 242 y 245.

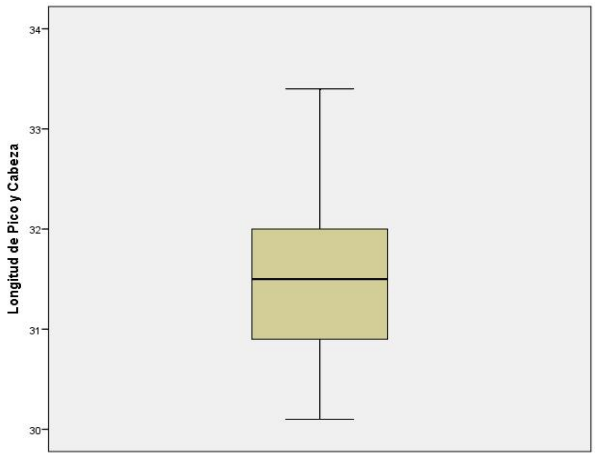


Figura 2.3: Longitud de Pico y Cabeza

La variable longitud de pico y cabeza se observa que tiene una distribución asimétrica negativa, y la mayor dispersión se encuentra en el intervalo

(32,33.10).

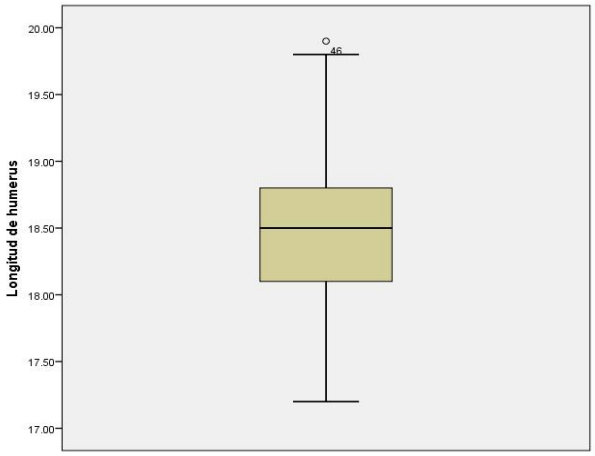


Figura 2.4: Longitud de Humerus

En la variable longitud de humerus es en la única donde se observa un punto atípico, el cuál es el pájaro número 46, y se puede observar bastante variabilidad después del valor 18.8.

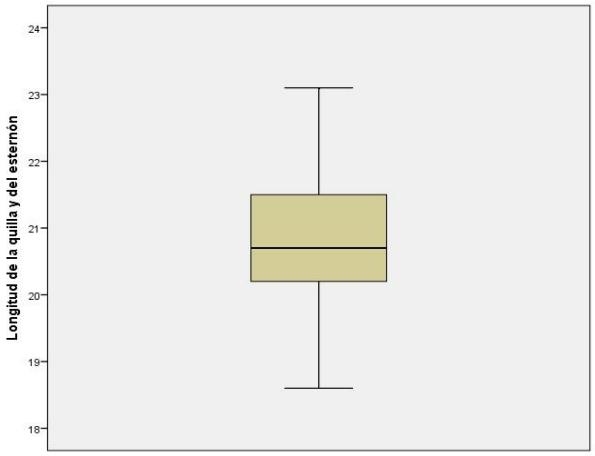


Figura 2.5: Longitud de la Quilla del Esternón

Los datos se concentran en los valores de 20.2-20.7, y se puede ver que

los datos se mantienen dispersos en los extremos.

Análisis bivariado

A continuación se presentan los gráficos que se obtuvieron por la comparación entre los sobrevivientes y no sobrevivientes, en cada variable.

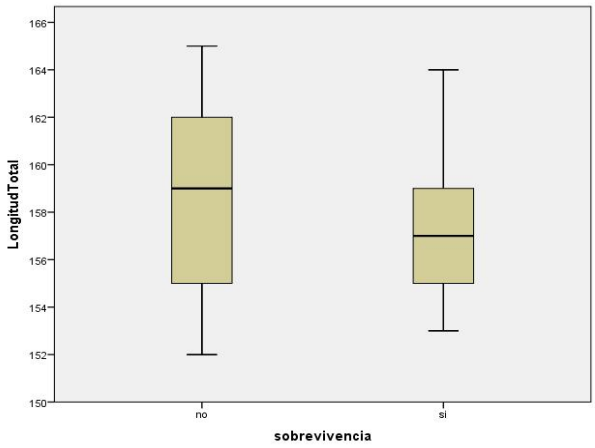


Figura 2.6: Longitud Total: Sobrevivencia

En la gráfica (Figura 2.6), se observa que los pájaros que no sobrevivieron, presentan una distribución simétrica sin embargo se observa que la mayoría de los pájaros que sobrevivieron, fueron los que, eran más pequeños, es decir, tenían una longitud menor.

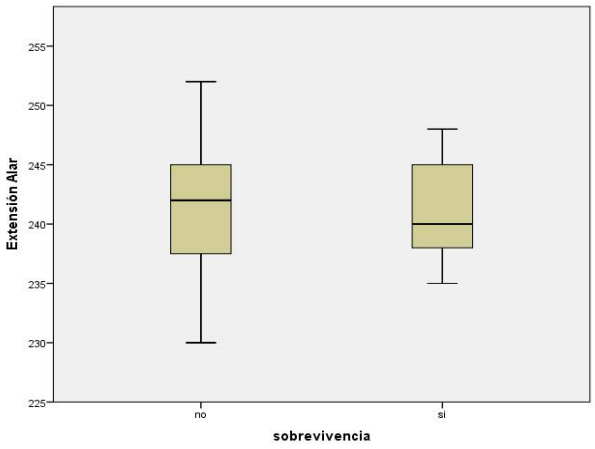


Figura 2.7: Extensión Alar: Sobrevivencia

En el caso de la variable Extensión Alar (Figura 2.7), la sobrevivencia presenta una distribución sesgada a la derecha mientras que los no sobrevivientes presentan una distribución sesgada a la izquierda.

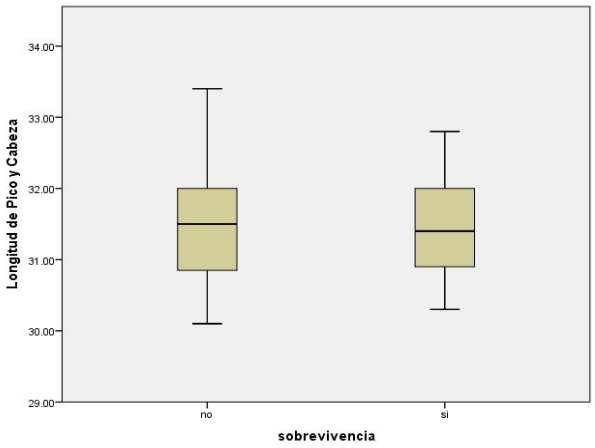


Figura 2.8: Longitud de Pico y Cabeza: Sobrevivencia

En la gráfica de la variable Longitud de Pico y Cabeza (Figura 2.8), los

sobrevivientes muestran simetría, sin embargo, los no sobrevivientes tienen una distribución sesgada a la derecha.

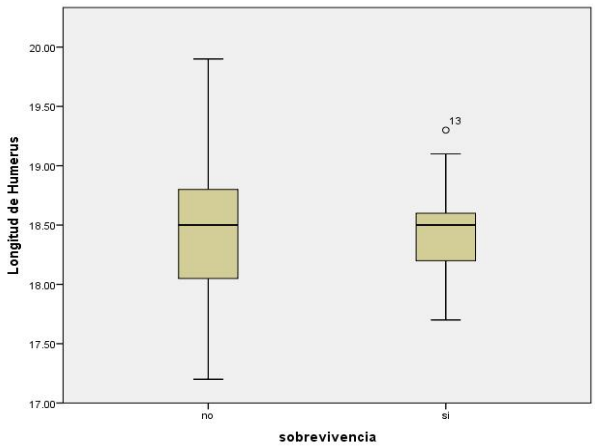


Figura 2.9: Longitud de Humerus: Sobrevivencia

En el gráfico comparativo (Figura 2.9) en el caso de los no sobrevivientes también se aprecian valores mas grandes y mayor variabilidad que en el caso de los sobrevivientes,más sin embargo, en el caso de los sobrevivientes,se presenta un punto atípico, el cuál es el pájaro número 13.

Como puede observarse en el gráfico comparativo (figura 2.10) los pájaros que no son sobrevivientes presentan una mayor variación en las medidas de la quilla además que alcanza medidas más grandes que en el caso de los sobrevivientes.

En la (Figura 2.11) se observa que hay una correlación positiva y además que los pájaros con medidas más extremas tanto en su extensión alar como en la longitud total, son los que no sobrevivieron.

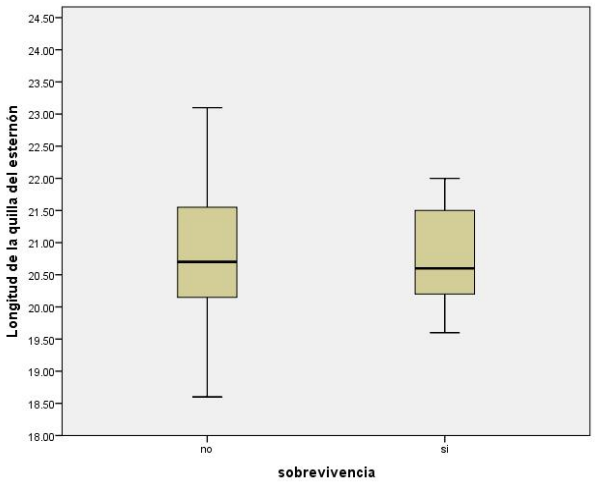


Figura 2.10: Longitud de la Quilla del Esternón: Sobrevivencia

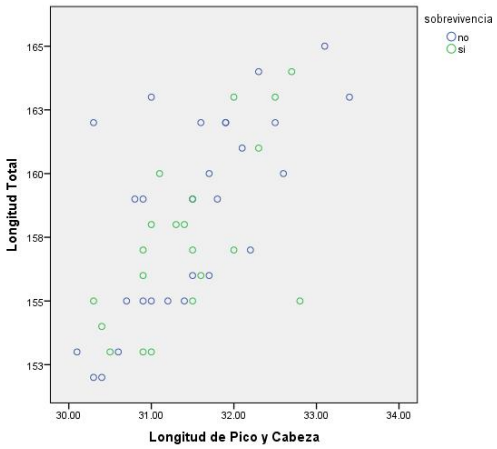


Figura 2.12: Diagrama de dispersión Longitud Total vs Longitud de Pico y Cabeza

En el diagrama de dispersión (Figura **??**) aunque hay una correlación positiva entre las variables se observa que es baja.

Se aprecian en el diagrama de dispersión (Figura 2.13) que la correlación

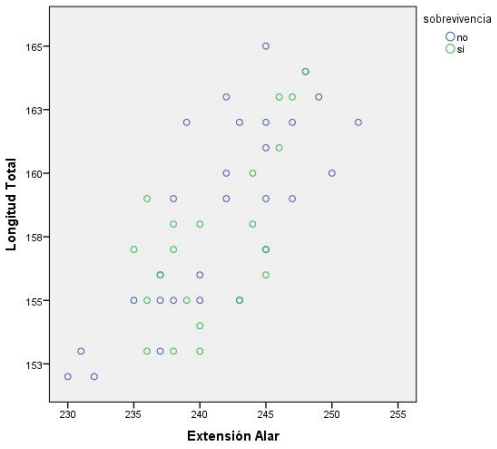


Figura 2.11: Diagrama de dispersión Longitud Total vs Extensión Alar

es baja aunque sigue siendo positiva y los pájaros con medidas mas pequeñas o grandes, respecto a la longitud total y la longitud de pico y cabeza, son los que no sobrevivieron.

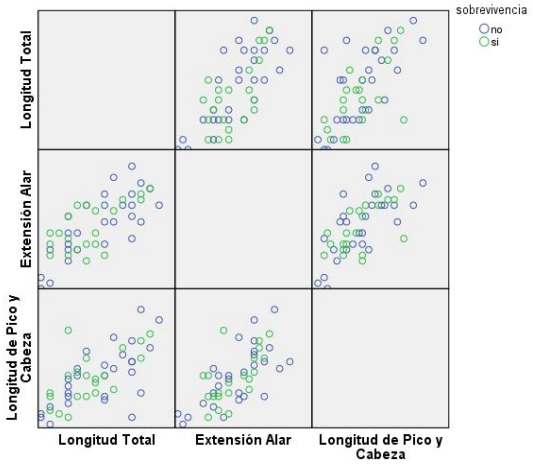


Figura 2.14: Matriz de dispersión

Se puede observar que en el caso de extensión alar vs longitud de pico

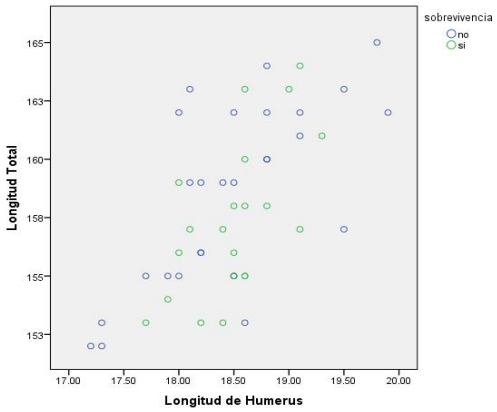


Figura 2.13: Diagrama de dispersión Longitud total vs Longitud de pico y cabeza

y cabeza hay una correlación positiva y los pájaros que no sobrevivieron muestran una mayor dispersión en sus medidas.

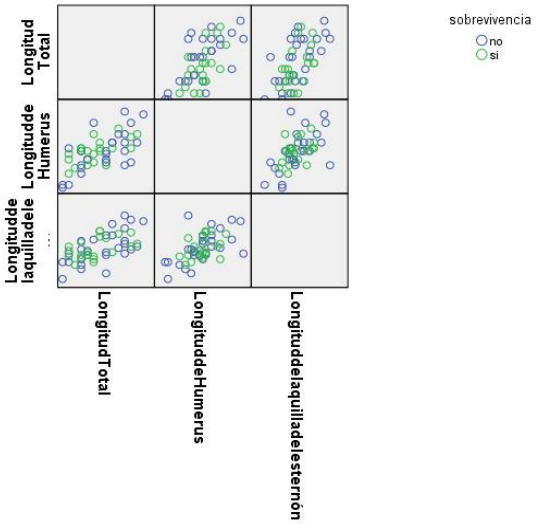


Figura 2.15: Matriz de dispersión

En todos los casos se observa una correlación positiva, sin embargo, los

pájaros que no sobrevivieron presentan medidas con mayor dispersión, y también se aprecia que los pájaros que tuvieron medidas muy grandes como muy pequeñas, no lograron sobrevivir.

En la Tabla 2.3 se muestra la correlación entre las variables.

Tabla 2.3: Tabla de correlaciones

	L.Total	E. Alar	L. de Pico y Cabeza	L. de Humerus	L. de la Quilla del Esternón
L.Total	1	.735	.662	.644	.605
E. Alar	.735	1	.674	.762	.529
L. de Pico y Cabeza	.662	.674	1	.741	.526
L. de Humerus	.644	.762	.741	1	.564
L. de la Quilla del Esternón	.605	.529	.526	.564	1

Se observa en primer lugar que todas las correlaciones son estadísticamente significativas, también se nota que las variables que están más correlacionadas son: Longitud de Humerus con Extensión Alar, Longitud Total con Extensión Alar, además, de Longitud de Humerus con Longitud de Pico y Cabeza.

De lo anterior se tiene que Longitud de Humerus se encuentra correlacionado con Extensión Alar y Longitud de Pico y Cabeza. Y viendo la correlación entre estas dos variables, se observa que es de 0.624.

Lo que nos dice que estas 3 variables podrían formar una componente.

2.3.2. Análisis definitivo

Cuando se realiza el análisis de componentes principales, algunos autores [7] recomiendan examinar en un principio la prueba de KMO y la prueba de esfericidad de Bartlett.

El valor obtenido en el test de KMO nos da información sobre la pertinencia del análisis. Nos dice si la correlación entre las variables es fuerte. Dado que el valor es de 0.846 se tiene que es aplicable, ya que es mayor de 0.6.

Tabla 2.4: Prueba de esfericidad de Bartlet

Prueba de esfericidad de Bartlet	Aprox. chi cuadrado	144.327
	gl	10
	Sig	.000

La Prueba de esfericidad de Bartlett, parte de la hipotesis nula de que las variables no están correlacionadas entre sí.

De acuerdo a lo anterior se rechaza la hipótesis nula, lo cual nos lleva a que las variables se encuentran correlacionadas y tiene sentido aplicar el análisis de componentes principales.

Tabla 2.5: Comunalidades

	Extracción
Longitud Total	.749
Extensión Alar	.817
Longitud de Pico y Cabeza	.776
Longitud de Humerus	.816
Longitud de la quilla del esternón	.974

Las comunalidades nos dan información acerca de la proporción de la varianza de cada una de las variables originales es explicada por las componentes principales. Se observa que el modelo es capaz de reproducir más del 70 % de la variabilidad de cada variable, y en el caso de la variable de longitud de la quilla del esternón es capaz de reproducir el 97 %.

Tabla 2.6: Varianza total

Componente	Total	% de la varianza	% acumulado
1	3.588	71.752	71.752
2	.544	10.884	82.635
3	.374	7.481	90.117
4	.307	6.150	96.267
5	.187	3.733	100.00

El porcentaje de la varianza representa la proporción de varianza explicada por la primera componente principal, en este caso, se observa que la primera componente explica un 71.75 % lo cual es más de la mitad, lo que implica que se conseguirá una reducción significativa.

Al comparar la primera componente con los demás se nota que la primera componente es por mucho la más importante, dado que la diferencia es enorme.

Tabla 2.7: Matriz de componentes

	Componente				
	1	2	3	4	5
Longitud de Humerus	.882	-.195	.227	-.252	-.262
Extensión Alar	.880	-.204	-.226	-.256	.258
Longitud Total	.863		-.409	.230	-.176
Longitud de Pico y Cabeza	.855	-.215	.293	.348	.128
Longitud de la quilla del esternón	.748	.644	.135		

Interpretando los componentes principales se observa la Matriz de componentes. Los coeficientes son las cargas factoriales que expresan la magnitud de la correlación entre la variable y el componente principal.

En la tabla 2.7 se aprecia que en la primer componente hay muy poca diferencia entre los coeficientes de las variables, la variable con el coeficiente mayor longitud de humerus con .882 y la de menor es longitud de la quilla del esternón la cual tiene .748, así que se observa que la diferencia es sólo de

0.134, al ver que la diferencia es tan reducida se puede decir que el 71.75 % de la varianza de los datos está relacionada con los tamaños de los pájaros.

En la segunda componente principal se ve que el coeficiente de longitud total es tan pequeño que no afecta a esta, se puede notar que cuando las variables de longitud de humerus, extensión alar y longitud de pico y cabeza obtengan un coeficiente alto entonces la longitud de la quilla obtendrá un coeficiente bajo y así la segundo componente podría obtener una mayor proporción.

Cuando se revisa la tercer componente, se aprecia que la longitud del humerus, del pico y cabeza y de la quilla del esternón no presentan una diferencia muy amplia, sin embargo, contrastan con la extensión alar y la longitud total, lo cual como los otros componentes nos muestra que hay una diferencia de forma entre los pájaros analizados.

A continuación se presenta en gráfico de sedimentación.

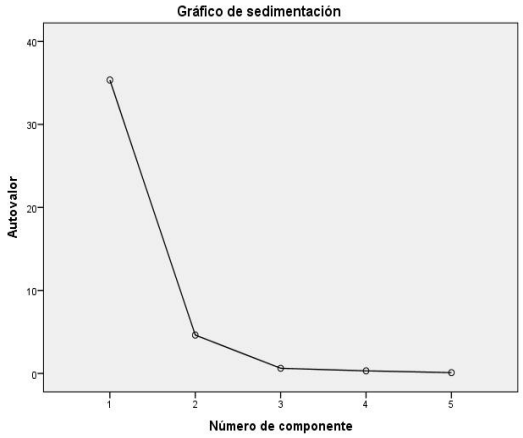


Figura 2.16: Gráfico de sedimentación en SPSS

Gráfico de sedimentación (2.16) Se observa que se forma el codo con los dos primeros componentes principales, así que se puede determinar que solo se necesita ocupar las dos primeras componentes principales.

Tabla 2.8: Varianza explicada en R studio

	y_1	y_2	y_3	y_4	y_5
Desviación Estándar	1.8941	0.7377	0.61161	0.5545	0.43205
Proporción de la varianza	0.7175	0.1088	0.07481	0.0615	0.0373
Varianza acumulada	0.7175	0.8264	0.90117	0.9627	100.00

R studio En R studio, se decidió estandarizar para que las medidas tengan el mismo peso, se puede observar que se obtuvieron los mismos resultados que en SPSS, lo cual se puede comprobar en la proporción de varianza, y la varianza acumulada, lo que confirma que solamente se necesita tomar las dos primeras componentes principales.

Interpretación de resultados

En la tabla 2.9 se puede apreciar los coeficientes obtenidos con el ACP

Tabla 2.9: Análisis de los coeficientes de las componentes en R studio

	y_1	y_2	y_3	y_4	y_5
Longitud total	0.4556	-0.0850	0.6688	0.4140	-0.4077
Extensión alar	0.4647	0.2769	0.3691	-0.4622	0.5977
Longitud de Pico y cabeza	0.4512	0.2910	-0.4788	0.6283	0.2960
Longitud de humerus	0.4656	0.2648	-0.3719	-0.4546	-0.6066
Longitud de la quilla del esternón	0.3949	-0.8724	0.2206	-0.1155	0.1440

Con esto se tiene que el primer componente

$$y_1 = 0{,}455x_1 + 0{,}464x_2 + 0{,}451x_3 + 0{,}465x_4 + 0{,}394x_5$$

Se observa que los coeficientes de x son casi iguales, la diferencia es poca entre ellos, lo que nos indica que aproximadamente el 71.75 % de la variación en los datos está relacionada con las diferencias de tamaño entre los gorriones, esto se puede afirmar ya que como se ve en la tabla 2.8, el primer componente explica el 71.75 %.

La segunda componente es de la forma:

$$y_2 = -0{,}085x_1 + 0{,}276x_2 + 0{,}291x_3 + 0{,}264x_4 - 0{,}872x_5$$

en este componente se puede ver un contraste entre las variables extensión alar, longitud de pico y cabeza y longitud de humerus con longitud de la quilla del esternón. Se puede apreciar que y_2 será mayor si extensión alar, longitud de pico y cabeza y longitud de humerus son mayores y si extensión alar, longitud de pico y cabeza y longitud de humerus son menores, es decir, tienen coeficientes más bajos entonces y_2 también será menor.

Por todo esto se puede decir que y_2 representa una diferencia de forma entre los gorriones. Como longitud total tiene un coeficiente muy bajo entonces eso nos indica que no tiene relevancia en la segunda componente principal.

Después de estos resultados, al considerar los dos componentes principales se puede graficar, para poder visualizar de mejor manera los resultados (Véase Figura 2.17 y Figura 2.18).

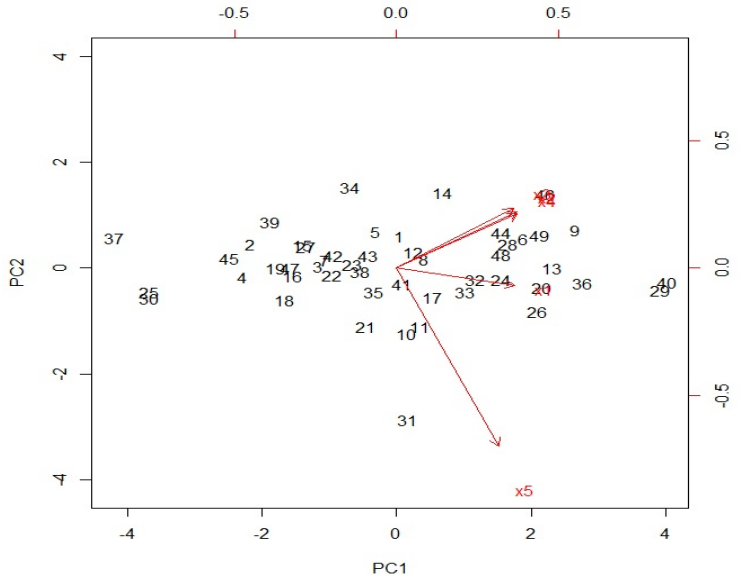


Figura 2.17: Gráfica de sobrevivencia : pc1 vs pc2

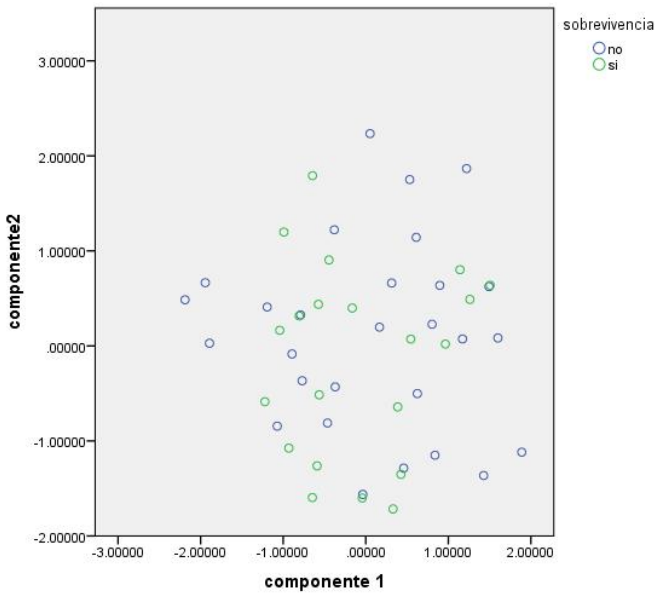


Figura 2.18: Gráfica de sobrevivencia : pc1 vs pc2

Se aprecia en Figura 2.17 y Figura 2.18 que las aves con valores extremos para la primera componente principal, no pudieron sobrevivir, y todo parece indicar que la tendencia se mantiene en la segunda componente principal.

Capítulo 3

Calificaciones

3.1. Aspectos generales

En este segundo caso, la información fue obtenida de los alumnos de las cinco licenciaturas, que se encuentran en la Facultad de Ciencias Físico-Matemáticas, las cuales son Matemáticas, Matemáticas aplicadas, Física, Física aplicada y Actuaría. El periodo del cual se obtuvo fue en primavera 2019, en total 66 alumnos fueron encuestados, donde todos se encontraban cursando su sexto semestre o algún semestre superior, se trabajarón con las variables expuestas en la Tabla 3.1.

Tabla 3.1: Caso calificaciones de alumnos.

Variable	Descripción	Valores
licenciatura	se refiere a la licenciatura que estudia el alumno	Matemáticas, Matemáticas aplicadas, Física, Física aplicada y Actuaría
calMB	es la calificación del curso de matemáticas básicas	6, 7, 8, 9 y 10
calDIF	es la calificación del curso de cálculo diferencial	6, 7, 8, 9 y 10
calIN	es la calificación del curso de cálculo integral	6, 7, 8, 9 y 10
calTE	es la calificación del curso de teoría de ecuaciones	6, 7, 8, 9 y 10
calAL	es la calificación del curso de álgebra lineal	6, 7, 8, 9 y 10
calDIFV	es la calificación del curso de cálculo diferencial en varias variables	6, 7, 8, 9 y 10
calINV	es la calificación del curso de cálculo integral en varias variables	6, 7, 8, 9 y 10

3.2. Análisis estadístico

3.2.1. Análisis preliminar

Se realizaran gráficas de cajas y bigotes y analizarán las variables de licenciatura, Matemáticas básicas (calMB), Teoría de ecuaciones (calTE), Cálculo integral(calIN), Cálculo diferencial (calDIF), Álgebra lineal(calAL), Cálculo diferencial en varias variables (calDIFV) y Cálculo integral en varias variables (calINV) respecto a las licenciaturas, todo esto se analizará con el fin de observar el comportamiento de las variables.

3.2.2. Análisis definitivo

Se realizará un análisis de componentes principales, con el objetivo de reducir la dimensionalidad del estudio, y analizar si las nuevas variables pueden ser interpretadas y presentar información acerca del comportamiento de las variables.

3.3. Resultados

3.3.1. Análisis preliminar

A continuación se presentan las estadísticas descriptivas básicas de las variables bajo estudio.

En la tabla 3.2 se observa que las calificaciones que la mayoría de los alumnos obtuvo son 8 y 9, además que las calificaciones de los cursos en promedio son parecidas.

Tabla 3.2: Estadísticas Descriptivas

	Media	Desviación Estándar	Moda	Mediana
Matemáticas Básicas	8.24	1.32	9	8
Teoría de Ecuaciones	8.27	1.11	8	8
Álgebra Lineal	8.09	1.29	9	8
Cálculo Diferencial	8.21	1.04	8	8
Cálculo Integral	8.18	1.11	9	8
Cálculo Diferencial en varias variables	8.48	1.23	9	9
Cálculo Integral en varias variables	8.65	1.12	8	9

A continuación se presentan los gráficos de cajas y bigotes de las variables respecto a las licenciaturas.

En matemáticas básicas (Figura 3.1) al hacer el análisis respecto a las licenciaturas, en Física se observa un dato atípico que es el alumno número 65 que obtuvo 10, mientras que la mayoría de los datos se encuentra en un intervalo de 6 a 8 aunque en Física aplicada observamos simetría, en Matemáticas aplicadas y Matemáticas se observa que no hay mucha variabilidad ya que se concentra en las calificaciones 9 y 10.

Se observa en teoría de ecuaciones (Figura 3.2) que en el caso de Actuaría

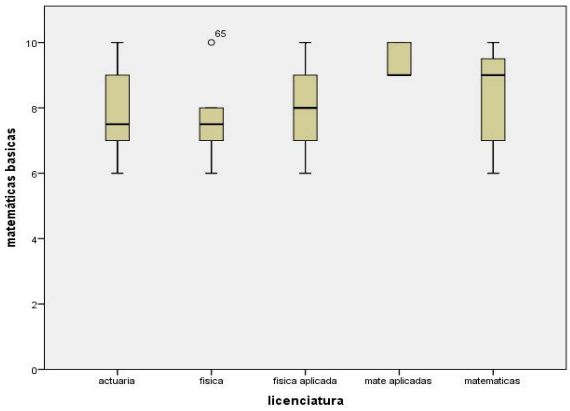


Figura 3.1: matemáticas básicas: Comparación

se presenta un punto atípico el estudiante número 17, Física presenta una distribución simétrica y Matemáticas es donde se presenta mayor variabilidad.

Nótese que en álgebra lineal (Figura 3.3) que en física aplicada su rango va de 7 a 9, y en el caso de Matemáticas aplicadas se observa que las calificaciones se concentran en 9 y 10, lo cual sucede también en Matemáticas pero con la diferencia que hay alumnos que obtuvieron 6 lo cual provoca que la dispersión sea mayor.

Observando la gráfica de cálculo diferencial (Figura 3.4) se aprecia que las licenciaturas de Física y Actuaría presentan una asimetría positiva y cada uno presenta un punto atípico, donde este representa un alumno que obtuvo 6, en oposición tanto Física aplicada como Matemáticas aplicadas presentan una asimetría negativa.

Al ver las gráficas de caja de cálculo integral (Figura3.5) se tiene que las licenciaturas de Actuaría, Matemáticas y Matemáticas aplicadas presentan asimetría negativa, en caso contrario con Física ya que esta presenta asimetría

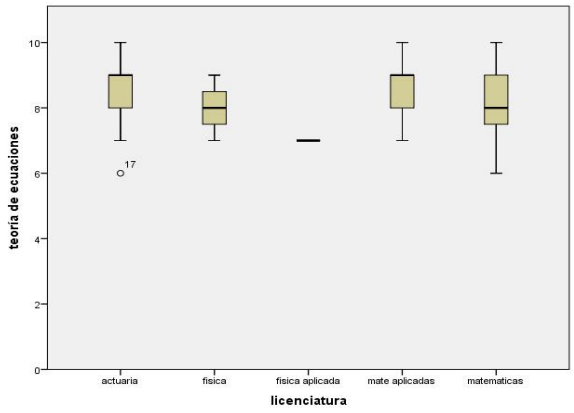


Figura 3.2: teoría de ecuaciones: Comparación

positiva.

En cálculo diferencial en varias variables (Figura 3.6) las licenciaturas de física aplicada, matemáticas y matemáticas aplicadas presentan una asimetría negativa.

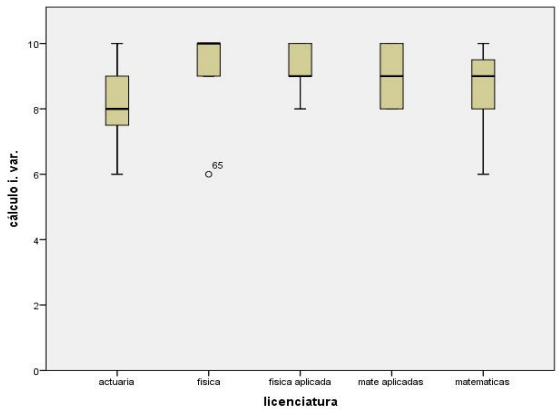


Figura 3.7: cálculo integral en varias variables

En cálculo integral en varias variables (Figura 3.7), Física tiene un dato atípico, que es un alumno que obtuvo 6 mientras que los demás datos se

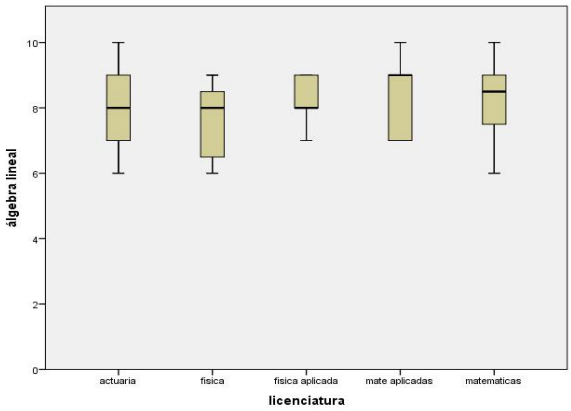


Figura 3.3: álgebra lineal: Comparación

concentran en las calificaciones 9 y 10, muy parecido a lo que se observa en Física aplicada donde las calificaciones también se concentran en 9 y 10, en el caso de Matemáticas aplicadas se observa que es simétrica.

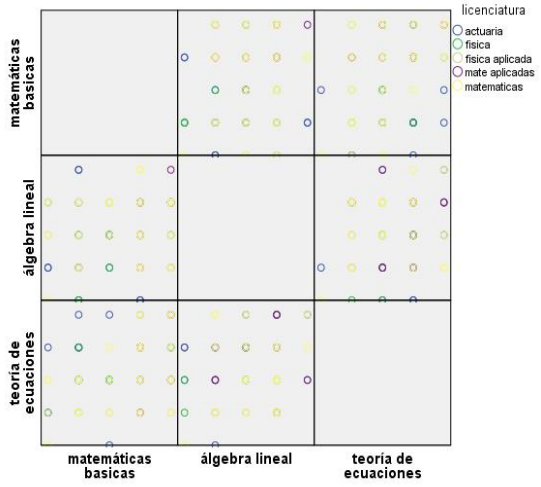


Figura 3.8: Matriz de dispersión

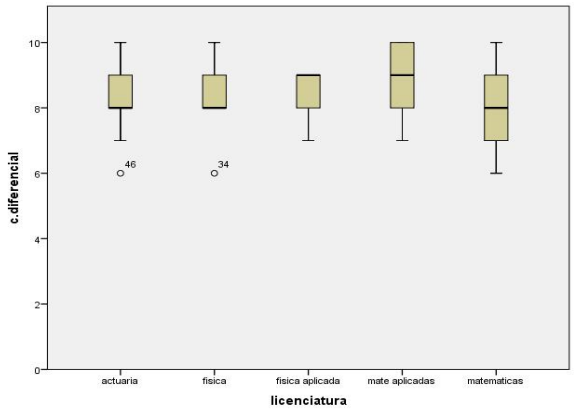


Figura 3.4: cálculo diferencial:Comparación

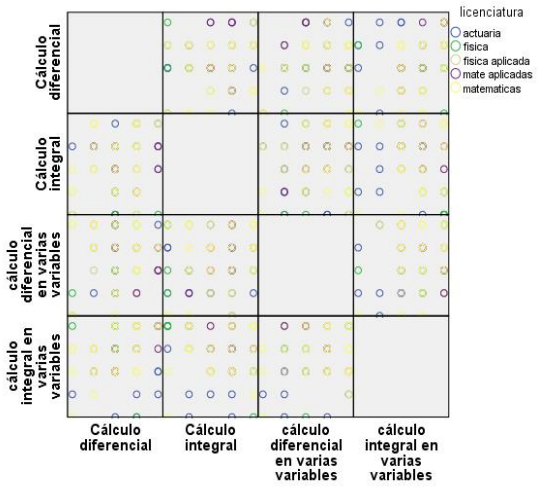


Figura 3.9: Matriz de dispersión

Se observa en los gráficos (figuras 3.8 y 3.9) que en general, las correlaciones que se presentan son muy débiles, ya que en los gráficos se presenta una gran dispersión, por ejemplo, cálculo integral en varias variables con cálculo diferencial en varias variables o matemáticas básicas con álgebra lineal.

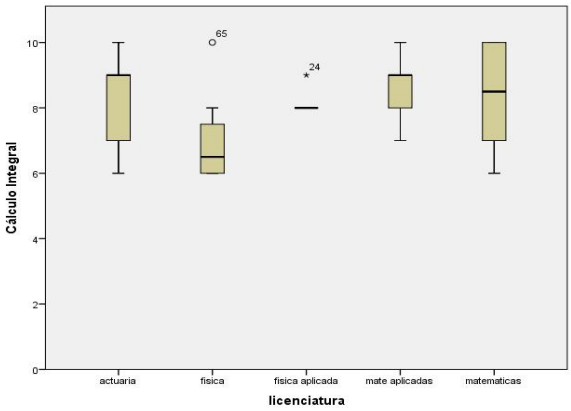


Figura 3.5: cálculo integral:Comparación

A continuación se presentan la tabla de las correlaciones de las variables.

Tabla 3.3: Tabla de correlaciones

	calMB	calDIF	calIN	calTE	calAL	calDIFV	calINV
calMB	1.000	.297	.349	.329	.203	.087	.037
calDIF	.297	1.000	.153	.055	.255	.137	.194
calIN	.349	.153	1.000	.212	.390	.117	-.239
calTE	.329	.055	.212	1.000	.176	.039	-.127
calAL	.203	.255	.390	.176	1.000	.101	.013
calDIFV	.087	.137	.117	.039	.101	1.000	.190
calINV	.037	.194	-.239	-.127	.013	.190	1.000

Se aprecia que aunque hay una correlación entre las variables, es muy débil, incluso hay correlaciones que se encuentran cercanas a 0.

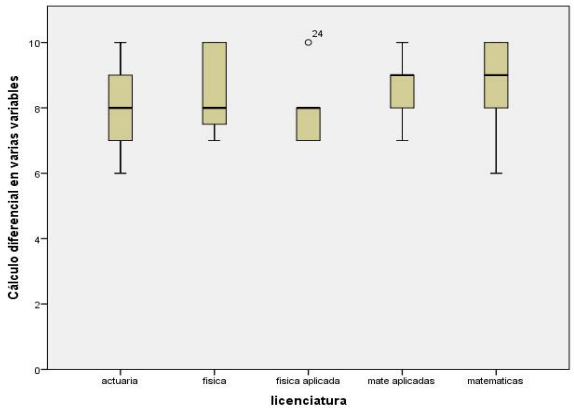


Figura 3.6: cálculo diferencial en varias variables

3.3.2. Análisis definitivo

En base al análisis preliminar, se puede afirmar que la correlación entre las variables es baja, sin embargo, se realizarán la prueba de KMO y la prueba de esfericidad de Bartlett, para determinar si hay la suficiente correlación para aplicar la técnica de análisis de componentes principales.

El valor obtenido en el test de KMO es 0.595 lo cual es menor que 0.6, aunque es muy cercano, así que es una relación baja.

Tabla 3.4: Prueba de esfericidad de Bartlet

Prueba de esfericidad de Bartlet	Aprox. chi cuadrado	49.086
	gl	21
	Sig.	.000

Dado que el p-value es menor que 0.05 todavía es aplicable el análisis de componentes principales.

Tabla 3.5: Comunalidades

	Extracción
calMB	.577
calTE	.251
calAL	.345
calDIF	.410
calIN	.689
calDIFV	.448
calINV´	.632

Se observa en la tabla 3.5 que ninguna varianza de las variables originales es explicada, en al menos un 70 %, de hecho la varianza mayormente explicada es la de cálculo integral 68.9 %, además de está, sólo cálculo integral en varias variables y matemáticas básicas son explicada más del 50 %, y las demás están por debajo de este porcentaje.

Se obtuvo lo siguiente al análizar la matriz de covarianza. En la tabla 3.6 la primera componente principal solamente representa un 30.9 % y dado que esta será la de mayor representación, se puede notar que se necesitarán varias componentes, para que se cumpla la regla del 80 %.

En este caso se necesitarían las primeras 5 componentes principales para poder trabajar en un análisis, y dado que son 7 variables no hay una reducción significativa.

Gráfico de sedimentación Se utilizará el método del codo para visualizar, si se debe considerar las 5 variables que nos sugiere el análisis anterior.

Tabla 3.6: Varianza de los componentes

	y_1	y_2	y_3	y_4	y_5	y_6	y_7
Desviación Estándar	1.743	1.357	1.178	1.088	0.954	0.905	0.801
Proporción de la varianza	0.309	0.187	0.141	0.120	0.092	0.083	0.053
Proporción Acumulada	0.309	0.496	0.637	0.758	0.851	0.934	1.000

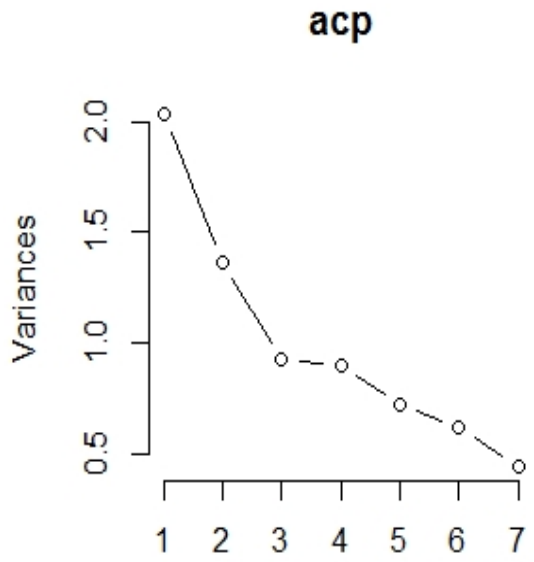


Figura 3.10: Gráfico de Sedimentación: R studio

Como se aprecia en la Figura 3.10, con este criterio, se tendría que considerar las 4 primeras variables, aunque esto sólo representaría un 75.8 %.

Interpretación de resultados

Tabla 3.7: Análisis de los coeficientes de las componentes

	y_1	y_2	y_3	y_4	y_5	y_6	y_7
calMB	.702	.023	.437	-.017	-.306	.353	-.311
calDIF	.522	.482	.077	-.430	-.254	-.479	.095
calIN	.706	-.317	-.360	.016	-.196	.246	.415
calTE	.521	-.314	.530	.341	.378	-.254	.160
calAL	.647	.027	-.418	-.246	.527	.022	-.257
calDIFV	.283	.491	-.315	.728	-.138	-.137	-.110
calINV	-.044	.831	.206	-.037	.288	.346	.245

En la tabla 3.7, se observa que en la primera componente hay una relación entre matemáticas básicas, cálculo diferencial e integral, teoría de ecuaciones y álgebra lineal, es decir las primeras 5 materias tienen una correlación fuerte y positiva, cálculo diferencial e integral no se consideran, dado que sus coeficientes son muy bajos, recordando que esta componente solo explica el 30.9 % de la varianza.

La segunda componente es dominada por las materias del área de análisis matemático, ya que las variables de matemáticas básicas y teoría de ecuaciones no se toman en cuenta, donde las únicas variables que tiene una carga negativa son cálculo integral y teoría de ecuaciones, aunque las variables mas representativas son cálculo integral en varias variables, cálculo diferencial en varias variables y cálculo diferencial.

Las demás componentes no muestran una tendencia clara como las pri-

meras dos componentes pero eso es esperado, debido a que representan el 14 % o menos de la varianza, entre menos represente, la interpretación será más complicada.

Dado que las componentes no tienen una proporción suficientemente grande, no se puede dar una interpretación correcta de las componentes.

Se concluye que en este caso el análisis de componentes principales no puede brindar una mayor interpretación dado que las variables originales no están tan correlacionadas, para lograr una interpretación más significativa.

Capítulo 4

Conclusión

Como se pudo observar, en los ejemplos expuestos, la técnica de análisis de componentes principales, para poder ser aplicada , se tiene que considerar la prueba de Barttlet y la medida de adecuación muestral de Kaiser-Meyer-Olkin, sin embargo esta última, generalmente se evalúa de manera algo subjetiva, en caso contrario, la prueba de Barttlet es muy importante, para determinar si se aplica o no, el ACP, ya que en caso de que la prueba no sea rechazada, no tendrá caso seguir con el análisis.

Al realizar el análisis se puede concluir que se debe tener cuidado cuando se escoja, en que matriz se va a realizar el ACP, en la matriz de correlación o covarianza, ya que se pueden obtener resultados diferentes que afectarán tanto a la elección de cuantos componentes se deben considerar como a la interpretación de estos, para hacer una elección adecuada, se debe considerar si tienen unidades de medida de valores muy diferentes y en gran medida arbitrarios, es mejor usar la matriz de correlación, como sucedió en el caso analizado de las medidas de los pájaros, donde había medidas de valores

muy diferentes, ya que la estandarización de las medidas aseguró que todas tuvieran el mismo peso en el análisis. En caso contrario, que se decida que las diferencias se deben conservar en vez de suprimir o como sucede, en el caso de estudio de las calificaciones, tengan las mismas unidades de medidas se puede usar la matriz de covarianza.

Es importante destacar que cuando se usa el software estadístico R, se debe especificar con los parámetros, `scale` y `center` la estandarización de los datos, dado que, si no se lleva a cabo esto, el comando `prcomp ()` que es el que se utiliza, realizará el análisis en la matriz de covarianzas.

Por último, es importante hacer hincapié que como se pudo observar en los casos presentados, aunque a veces hay una reducción significativa en las variables y una interpretación de estás, esto no necesariamente sucederá siempre, ya que, hay algunas veces en los que no se podrá obtener esto y aunque haya una reducción, puede que esta no sea útil, para lo que el investigador necesite o no haya una interpretación de las nuevas variables resultantes del análisis.

Apéndice

En esta sección se explicará cómo realizar la técnica de Análisis de Componentes Principales en los softwares SPSS y R studio.

4.1. SPSS

Los datos fueron guardados en una hoja de excel, asi que antes de empezar, importaremos los datos. Seleccionamos el menú **Archivo** → **Importar datos** → **Excel**

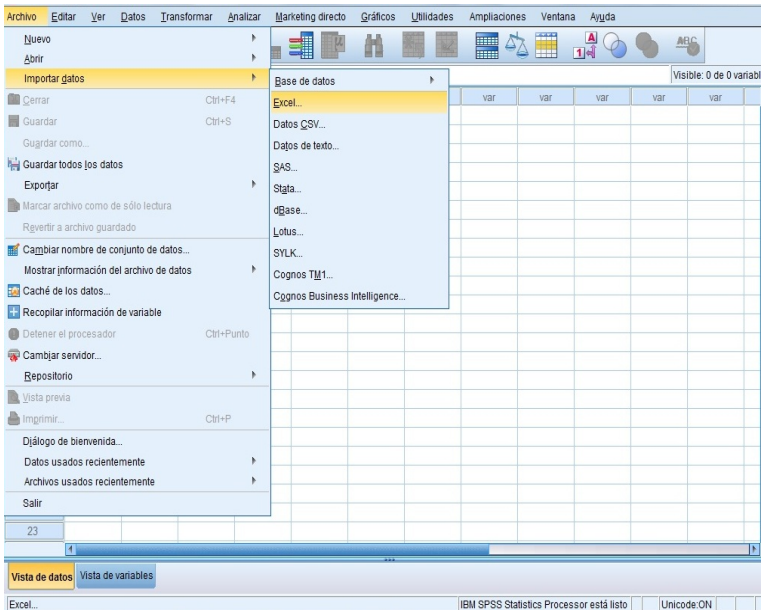
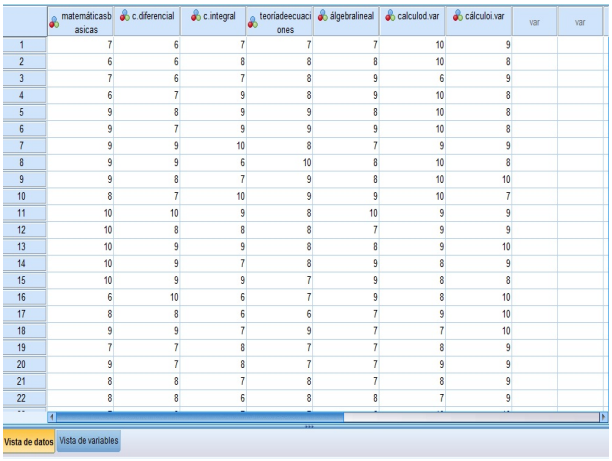


Figura 4.1: Importar datos de Excel

Después seleccionamos nuestro archivo y aparecerá en nuestro entorno de trabajo



	matemáticasbasicas	c.diferencial	c.integral	teoriadeecuaciones	álgebralineal	calculod.var	cálculoi.var	var	var
1	7	6	7	7	7	10	9		
2	6	6	8	8	8	10	8		
3	7	6	7	8	9	6	9		
4	6	7	9	8	9	10	8		
5	9	8	9	9	8	10	8		
6	9	7	9	9	9	10	8		
7	9	9	10	8	7	9	9		
8	9	9	6	10	8	10	8		
9	9	8	7	9	8	10	10		
10	8	7	10	9	9	10	7		
11	10	10	9	8	10	9	9		
12	10	8	8	8	7	9	9		
13	10	9	9	8	8	9	10		
14	10	9	7	8	9	8	9		
15	10	9	9	7	9	8	8		
16	6	10	6	7	9	8	10		
17	8	8	6	6	7	9	10		
18	9	9	7	9	7	7	10		
19	7	7	8	7	7	8	9		
20	9	7	8	7	7	9	9		
21	8	8	7	8	7	8	9		
22	8	8	6	8	8	7	9		

Vista de datos Vista de variables

Figura 4.2: Entorno de trabajo

Se selecciona **Analizar→ Reducción de Dimensiones→ Factor**

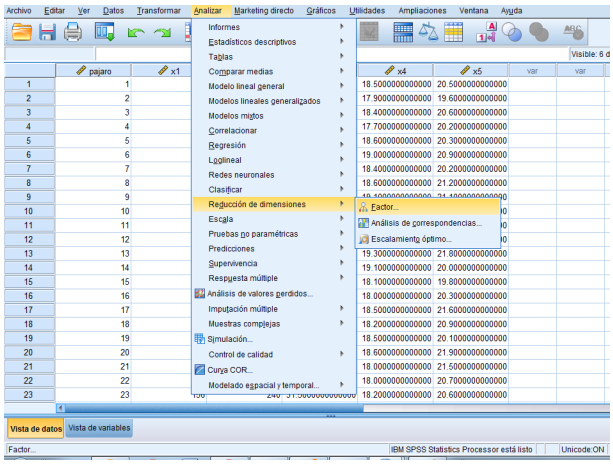


Figura 4.3: Analizar: Dimensiones

Una vez seleccionado esto, aparecerán las variables que tenemos en nuestro entorno del trabajo, así que escogeremos las variables que consideraremos en el análisis.

Figura 4.4: Variables

Después de esto iremos a la opción de extracción, donde señalaremos los parámetros necesarios, escogeremos la matriz de correlación.

También se seleccionaran la solución factorial sin rotar y el gráfico de sedimentación, para que podamos utilizar el criterio de Diagrama de Catell. Escogeremos extraer número fijo de factores, pondremos 5, para que podamos visualizar todos los factores.

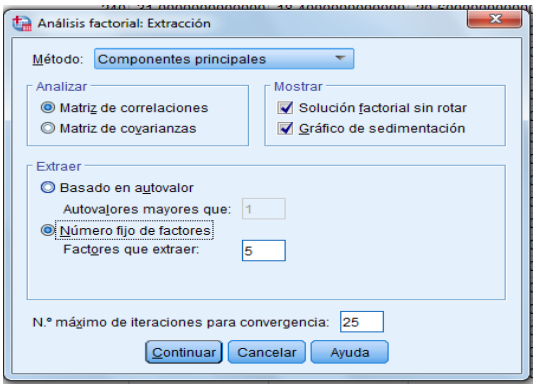
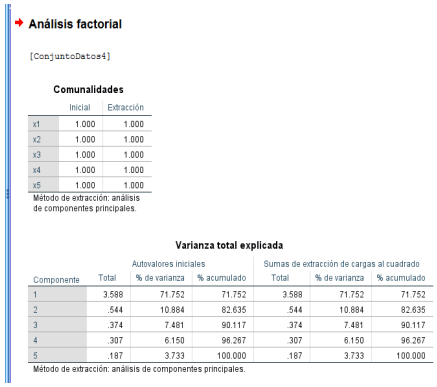


Figura 4.5: Extracción

Seleccionaremos en continuar y después Aceptar y así obtendremos los siguientes resultados.



Análisis factorial

[ConjuntoDatos4]

Comunalidades

	Inicial	Extracción
x1	1.000	1.000
x2	1.000	1.000
x3	1.000	1.000
x4	1.000	1.000
x5	1.000	1.000

Método de extracción: análisis de componentes principales.

Varianza total explicada

Componente	Autovalores iniciales			Sumas de extracción de cargas al cuadrado		
	Total	% de varianza	% acumulado	Total	% de varianza	% acumulado
1	3.588	71.752	71.752	3.588	71.752	71.752
2	.544	10.884	82.635	.544	10.884	82.635
3	.374	7.481	90.117	.374	7.481	90.117
4	.307	6.150	96.267	.307	6.150	96.267
5	.187	3.733	100.000	.187	3.733	100.000

Método de extracción: análisis de componentes principales.

Figura 4.6: Resultados

Matriz de componente[a]

	Componente				
	1	2	3	4	5
x1	.863	.063	-.409	.230	-.176
x2	.880	-.204	-.226	-.256	.258
x3	.855	-.215	.293	.348	.128
x4	.882	-.195	.227	-.252	-.262
x5	.748	.644	.135	-.064	.062

Método de extracción: análisis de componentes principales.

a. 5 componentes extraídos.

Figura 4.7: Resultados

4.2. R studio

En R studio, importamos los datos de Excel y a continuación se muestra el código usado.

```
#Importaremos los datos de una hoja de Excel.
> library(readxl)
> cal <- read_excel("tesis/cal.xls")
> View(cal)
> apply(cal,2,var) # Aplicamos varianza por columnas

matematicas basicas  c.diferencial  c.integral
1.755711             1.246620       1.689510
teoriadeecuaciones algebralineal calculo d.var
1.093706             1.253147          1.515152
calculo i. var.
1.276690
> acp<-prcomp(cal, center = TRUE, scale = TRUE)
```

```
> print(acp)
Standard deviations (1, .., p=7):
[1] 1.4269712 1.1681789 0.9614179 0.9455169
0.8487761 0.7864191 0.6647621

Rotation (n x k) = (7 x 7):
PC1          PC2          PC3          PC4
matematicas basicas 0.49199707 -0.01948380
 0.45464283 -0.01837131
c.diferencial 0.36586889 -0.41278162
 0.07991562 -0.45530037
c.integral 0.49487687  0.27114383 -0.37409125
 0.01671916
teoria de ecuaciones  0.36501003  0.26873609
  0.55130363  0.36047426
algebra lineal  0.45357084 -0.02287619
 -0.43522070 -0.26019788
calculo d. var.       0.19815301 -0.42012744
 -0.32803547  0.77000586
calculo i. var.      -0.03085906 -0.71166233
 0.21449733 -0.03912663
PC5          PC6          PC7
matematicas basicas   0.3608921  0.44950195  0.4671473
c.diferencial         0.2998064 -0.60964689 -0.1432242
c.integral            0.2314643  0.31245036 -0.6246224
teoria de ecuaciones -0.4456690 -0.32237333 -0.2410929
algebra lineal       -0.6212192  0.02850719  0.3871747
```

```
calculo d. var.        0.1623129 -0.17362242  0.1649904
calculo i. var.       -0.3397318  0.44019482 -0.3685855
> plot(acp, type="l")#Se obtendra el diagrama de catell
> summary(acp)
Importance of components:
PC1    PC2    PC3    PC4    PC5     PC6     PC7
Standard deviation     1.4270 1.1682 0.9614 0.9455
 0.8488 0.78642 0.66476
Proportion of Variance 0.2909 0.1950 0.1321 0.1277
 0.1029 0.08835 0.06313
Cumulative Proportion  0.2909 0.4858 0.6179 0.7456
 0.8485 0.93687 1.00000
}
}
```

En el ejemplo podemos notar que utilizamos la función `prcomp` para calcular los componentes principales, se decidió estandarizar los datos, asi que por eso se escribieron los parámetros `center = TRUE, scale=TRUE`, además que al utilizar `summary`, se visualiza la proporción de la varianza, y nos ayuda a utilizar el criterio de selección de componentes cuando se llegue al 80 %.

Bibliografía

[1] Araneo, D. Introducción al análisis de componentes principales.

[2] Avila, H., Garcia, S., et al. Análisis de componentes principales, como herramienta para interrelaciones entre variables fisicoquímicas y biológicas en un ecosistema léntico de guerrero, méxico.

[3] Bajo, M. Aplicaciones prácticas del análisis de componentes principales en gestión de carteras de renta fija (i). determinación de los principales factores de riesgo de la curva de de rendimientos. *Análisis financiero*, 124 (2014), 20–38.

[4] Mainly, B., and Navarro, J. *Métodos estadísticos multivariados: un manual*. CRC press.

[5] Mesa, L., Gozá, O., Uranga, M., Toledo, A., and Gálvez, Y. Aplicación del análisis de componentes principales en el proceso de fermentación de un anticuerpo monoclonal. *VacciMonitor 27*, 1 (2018), 8–15.

[6] Raykov, T., and Marcoulides, G. *An introduction to applied multivariate analysis.* Routledge.

[7] Ruiz, M., and Pardo, A. *Análisis de datos con SPSS 13 Base*. 2005.