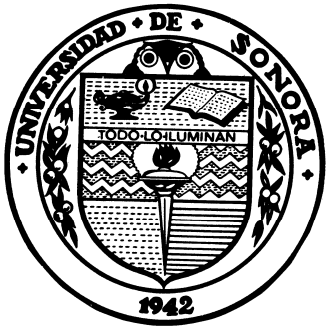



# Universidad de Sonora

División de Ciencias Exactas y Naturales

Departamento de Matemáticas


# Morfometría Geométrica, una nueva rama de la Estadística: Aplicaciones en Bioestadística

Tesis Profesional

Que para obtener el título de:

Lic. En Matemáticas


Presenta

Cinthia Yazmin Cuen Romero



Hermosillo, Sonora, marzo de 2016

**M.C. Alejandrina Bautista Jacobo**
Departamento de Matemáticas – Universidad de Sonora

**Dr. Martín Eduardo Frías Armenta**
Departamento de Matemáticas – Universidad de Sonora

**M.C. Pedro Ignacio Loera Burnes**
Departamento de Matemáticas – Universidad de Sonora

**cDr. Francisco Javier Cuen Romero**
Departamento de Geología – Universidad de Sonora

## Agradecimientos

Agradezco a la Universidad de Sonora y en especial al Departamento de Matemáticas, por permitirme formar parte del Programa Licenciatura en Matemáticas y por todos los conocimientos adquiridos a lo largo de mi formación profesional.

A mi directora de tesis M.C. Alejandrina Bautista Jacobo por su gran apoyo y paciencia para concluir este trabajo. Al Dr. Martín Eduardo Frías Armenta por su ayuda e interés. Al M.C. Pedro Ignacio Loera Burnes y cDr. Francisco Javier Cuen Romero por la atenta lectura y revisión del escrito.

Con todo mi afecto para mis padres: Alba Rosalía y Jesús. A mi esposo Gerardo y hermanos Francisco y Jesús.

## Índice

## 1. Resumen

La estadística moderna ha encontrado dentro de la Morfometría geométrica un nuevo campo de aplicaciones, basadas principalmente en la descripción cuantitativa de las formas biológicas combinadas con análisis estadísticos, con el fin de describir patrones de variación. Es así como dentro del amplio campo de la biología combinada con las técnicas de análisis de la estadística es posible realizar estudios enfocados a salud pública, genómica de poblaciones y genética, ecología, bioensayos, afinidad de especies y de linajes, evaluación forense, entre otras. El hecho de que la Morfometría geométrica trabaje con datos naturales la convierte en una poderosa herramienta para la evaluación objetiva de la variación de la forma. En este trabajo se muestran algunas de las aplicaciones de la Morfometría geométrica como una rama de la estadística, basada en el análisis de Procrustes, el cual permite comparar dos puntos homólogos provenientes de dos variantes de la misma entidad, es decir, es posible ajustar un arreglo sobre otro ya prestablecido por medio de una transformación de matrices, permitiendo realizar comparaciones sin perder las distancias de las configuraciones originales y minimizando la suma de cuadrados entre estos puntos homólogos.

Con base en este análisis de las aplicaciones de la Morfometría geométrica se observa que la principal utilidad dentro del campo de la Biología se encuentra enfocada hacia la búsqueda de nuevos taxones y diferenciación de especies. En el campo de la Paleontología en la comparación de estructuras y finalmente, en el campo de la Antropología, el entendimiento del dimorfismo sexual principalmente en el área del cráneo.

## 2. Introducción

La estadística es una ciencia de aplicación práctica casi universal en todos los campos científicos, por ejemplo, en las ciencias naturales, se emplea con frecuencia en la descripción de modelos termodinámicos complejos (mecánica estadística), en física cuántica, en mecánica de fluidos o en la teoría cinética de los gases, entre otros muchos campos. En las ciencias médicas, permite establecer pautas sobre la evolución de las enfermedades y los enfermos, los índices de mortalidad asociados a procesos morbosos, el grado de eficacia de un medicamento, etcétera.

La estadística moderna surge de la confluencia de dos disciplinas que evolucionaron de manera independiente: la aritmética de estado (estadísticas) y el cálculo de probabilidades.

La mayoría de las civilizaciones antiguas recogían datos sobre los impuestos recaudados, el número de soldados reclutados, bajas en batalla, censos, etc. En el siglo XVII John Graunt fue el primero en realizar tablas de mortalidad y estudios demográficos.

En el siglo XVIII De Moivre comprobó que la distribución binomial podía aproximarse a la normal cuando el número de casos era grande. Este autor junto a Laplace fueron los primeros en aplicar el cálculo de probabilidades a los datos demográficos, contribuyendo a unificar la estadística y el cálculo de probabilidades en una sola disciplina.

En el siglo XIX una de las figuras más relevantes en el campo de la física y de la estadística fue Gauss. Este científico hizo magníficos estudios sobre la curva normal, a la cual también se le conoce como campana de Gauss. Otras figuras relevantes del siglo XIX en el campo de la estadística fueron Newcomb, que realizó importantes estudios en relación a la estimación de parámetros, y K. Pearson, que trabajó, entre otros temas, sobre correlación y regresión de variables.

A partir de 1970 la estadística ha cobrado una gran dimensión entre otras razones por la generalización del uso de ordenadores, los que ha permitido utilizar técnicas estadísticas que, aunque conocidas desde hace tiempo, se aplicaban en pocas ocasiones debido a la dificultada de los cálculos.

Los softwares estadísticos han permitido gran avance en la enseñanza de la estadística, sobre todo para las investigaciones de grandes poblaciones y múltiples variables, estos ordenadores nos permiten manejar estas bases de datos con mayor facilidad y rapidez.

Cuando las aplicaciones son en el campo de la biología o Ciencias de la salud, se tiene un campo de la estadística llamada Bioestadística. Como los objetos de estudio de la Biología son muy variados, tales como la medicina, las ciencias agropecuarias, entre otros, la Bioestadística ha debido ampliar su campo para de esta manera incluir cualquier modelo cuantitativo, no solamente estadístico y que entonces pueda ser empleado para responder a las necesidades oportunas.

Algunas aplicaciones más destacadas en el ámbito de la biología son:

1. **Salud pública:** La estadística permite analizar situaciones en las que los componentes aleatorios contribuyen de forma importante en la variabilidad de los datos obtenidos. En salud pública los componentes aleatorios se deben, entre otros aspectos, al conocimiento o a la imposibilidad de medir algunos determinantes de los estados de salud y enfermedad, así como a la variabilidad en las respuestas por los pacientes, similares entre sí, que son sometidos al mismo tratamiento.

   La extensión de los conocimientos y aptitudes de carácter estadístico que necesitan adquirir los profesionales de la salud pública son importantes, porque el conocimiento de los principios y métodos estadísticos y la competencia en su aplicación se necesitan para el ejercicio eficaz de la salud pública, y adicionalmente para la comprensión e interpretación de los datos sanitarios; a fin de discriminar entre opiniones arbitrarias o

discrecionales, con respecto a las verdaderamente evaluadas en un contexto científico.

2. **Genómica y poblaciones genéticas:** En biología, la forma en que los padres transmiten su información a sus hijos, o genética, es una materia que utiliza mucho la estadística y probabilidad. Es el caso de los estudios de Mendel, por ejemplo, quién se dedicó a estudiar el comportamiento de ciertas plantas a las que cruzó y determinó cómo se relacionaban genéticamente los padres con los hijos, hablando de Genotipo y Fenotipo y para ello utilizó la probabilidad de que los descendientes heredaran caracteres de los progenitores. También se utiliza para medir la estabilidad genética en variedades o genotipos de una especie determinada.
3. **Ecología:** En este ámbito destaca la aplicación de la estadística en la colonialidad y aprendizaje en aves. Se parte del estudio de un modelo reciente en física estadística, el problema del juego de la minoría ("Minority Game"). Este modelo sencillo estudia el proceso de aprendizaje y coordinación de un conjunto de agentes con el mismo objetivo, de forma que la eficiencia global se optimiza según la cantidad de información que hay en el sistema. A través de simulaciones numéricas se han examinado las propiedades del sistema en sus distintos regímenes, así como su robustez frente a distintas condiciones iniciales. Esto se ha estudiado tanto en el modelo original, como en diferentes variaciones posibles, que han permitido obtener una perspectiva diferente del problema, aplicable así al problema de la colonialidad en aves. El comportamiento colonial en aves durante la época de cría, es decir, la estrategia de anidar en grupo frente a la de anidar en solitario, es un problema abierto en Ecología del Comportamiento. Gracias a un modelo basado en los anteriores estudios se ha podido realizar una aproximación numérica de la influencia de uno de los factores principales (predación) en el proceso de aprendizaje y en la evolución de este comportamiento. Los resultados obtenidos

reproducen fenomenológicamente los datos observados en estudios de campo, dando así una posible explicación a la colonialidad en aves como resultante de la presión de prelación.

4. **Bioensayos:** Un bioensayo es un proceso experimental mediante el cual se determinan las características y la fuerza de una sustancia potencialmente tóxica o de un desecho metabolito, a través del estudio de sus efectos sobre organismos cuidadosamente escogidos y bajo condiciones específicas de laboratorio. La bioestadística utiliza la experiencia operacional para adecuar los controles ambientales y bioensayos a la experiencia acumulada.

La colaboración de la bioestadística ha sido clave en el desarrollo de nuevos fármacos, en el entendimiento de enfermedades crónicas como el cáncer y el SIDA, siendo estos son algunos de los miles de ejemplos posibles.

En la Biología, entre los estudios más importantes que se llevan a cabo es la descripción anatómica y morfológica de los seres vivos con el objetivo de identificar diferencias intra e interespecíficas. Con base en lo anterior, se puede inferir que la morfometría posee un papel central en las ciencias biológicas y matemáticas, debido a la manera en que las formas varían y su covarianza con otras variables, es decir, el grado de variación conjunta de dos o más variables consideradas aleatorias.

El entendimiento de las estructuras de los organismos y el funcionamiento de sus sistemas en un contexto netamente comparativo ha sido un elemento principal en la evolución del conocimiento biológico, analizando las pautas de esta ciencia es posible observar que la clasificación taxonómica así como la biodiversidad están basadas en la descripción morfológica de los taxones. (Adams, *et al*., 2004).

A principios del siglo XX la biología comenzó la transición de una ciencia descriptiva a una ciencia cuantitativa, iniciándose así la disciplina denominada morfometría, la cual es relativamente reciente. La descripción cuantitativa de las

formas biológicas combinada con análisis estadísticos, describiendo patrones de variación de formas entre grupos, es el campo de acción de la morfometría.

Según Adams *et al.*, (2004) los estudios morfológicos incluyen datos cuantitativos para uno o más tratamientos medibles, los cuales son resumidos como valores medios y posteriormente son comparados con otros grupos; sin embargo el desarrollo de métodos estadísticos más avanzados, como el coeficiente de correlación, análisis de varianza y el análisis de componentes principales han provisto un rigor cuantitativo mayor.

Existen diversas definiciones para la morfometría, entendiéndose como el estudio de la variación de la forma y su covariación con otras variables (Brookstein, 1991). De igual manera su posición dentro de las ciencias ocupa un lugar entre la biología y las matemáticas, siendo aceptada como una rama de la estadística, la cual combina herramientas de la geometría, gráficos por computadora y técnicas biométricas para el análisis multivariado de la variación de las formas biológicas (Brookstein, 1996).

En 1993 surge una nueva disciplina dentro de la Morfometría, la cual capturaba la geometría de las estructuras morfológicas de interés y preservaba la información a través de los análisis estadísticos, esta nueva disciplina se denomina morfometría geométrica (Rohlf & Marcus, 1993). Es así como se diferencia entre morfometría tradicional y morfometría geométrica. Actualmente se considera a la morfometría geométrica como una ciencia madura, debido principalmente al gran entendimiento de las bases teóricas de la geometría y la metodología geométrica. (Adams, *et al*., 2004)

## 3. Objetivos

### 3.1. Objetivo general

El principal objetivo que persigue este trabajo es presentar las aplicaciones de la morfometría como una rama de la estadística con respecto a la biología.

## 4. Morfometría

El análisis de las formas es parte fundamental de la gran mayoría de la investigación biológica. A medida que el campo de la estadística se ha desarrollado también ha aumentado la sofisticación del análisis de estos tipos de datos. (Adams, *et al*., 2004). Actualmente se distinguen los disciplinas principales, la Morfometría tradicional y la Morfometría Geométrica, esta última incluye técnicas de análisis más avanzadas como de reseñará más adelante.

### 4.1. Morfometría Tradicional

En sus inicios, la morfometría utilizaba variables lineales, por ejemplo medidas, distancias, ángulos o proporciones, entre otras. A partir de estas variables se obtenía un conjunto de datos que posteriormente eran analizados por métodos estadísticos multivariados. Los resultados se expresaban como un conjunto de coeficientes y gráficas, sin embargo, estás eran de tamaño y forma variable, siendo difíciles de interpretar, este enfoque se le llama actualmente morfometría tradicional. (López, 2015).

La mayoría de los estudios que se hicieron por medio de morfometría tradicional corresponden a la descripción de la estructura de tejidos, células, órganos, dimensiones, formas y la relación que guardan entre estos (Toro Ibacache, *et al*., 2010), sin embargo, estas estructuras no podían ser analizadas cuantitativamente. La Morfometría tradicional, al mezclarse con técnicas estadísticas más avanzadas, dio origen a la Morfometría Geométrica.

### 4.2. Morfometría Geométrica

La morfometría geométrica ha sido descrita como: “una fusión empírica de la geometría con la biología” (Brookstein, 1982; López, 2015), debido a que analiza la forma de los organismos o de algunas de sus estructuras considerando el espacio geométrico y empleando métodos estadísticos multivariados (López, 2015). Lo

anterior implica que los objetos no son analizados en función de su dimensión, sino de la relación espacial que existen en sus partes. (Toro Ibacache, *et al*., 2010).

Uno de los conceptos fundamentales en morfometría geométrica es la forma, la cual es una propiedad geométrica de un objeto y que no toma en cuenta la escala, rotación y traslación. El tipo de estudios en los que se analizan los cambios en la forma con respecto al tamaño de los organismos a través de su ciclo de vida se conocen como estudios de alometría, por lo tanto se entiende que la forma se asocia al tamaño (Anzelmo, *et al*., 2012; López, 2015). Considerando la forma, la morfometría geométrica utiliza dos fuentes de información para los análisis, uno es la homología biológica y la localización geométrica (López, 2015). La homología biológica hace referencia a la correspondencia biológica de determinadas estructuras o partes entre individuos (Brookstein, 1986), mientras que la localización geométrica se refiere a la configuración espacial en dos o tres dimensiones de estas estructuras o partes. (Klingenberg & Monteiro, 2005).

Para la localización de estas estructuras homologas, en morfometría se utilizan principalmente dos variables *outlines* (contornos) y los *landmarks* (hitos), que son loci anatómicos que no alteran su posición topológica en relación a otros hitos, por lo tanto proveen una cobertura adecuada de la forma y pueden ser ubicados de manera sencilla y repetidamente entre un organismo y otro (Zelditch, *et al*., 2004). Brookstein (1991) estableció que existen tres tipos de *landmarks*: tipo I son yuxtaposiciones discretas de tejido, tipo II son zonas de máxima o mínima curvatura y tipo III son puntos extremos (López, 2015).

Según (Toro Ibacache, *et al*., 2010) La geometría de las estructuras se lleva a cabo en tres etapas principales, las cuales se mencionan a continuación:

1. La primera parte del análisis morfométrico consiste en la obtención de la muestra, la cual proporcionará los datos que se utilizarán en el estudio.
2. El segundo paso es la obtención de la información que describirá la forma del organismo o de la parte a estudiar (*shape*).

3. Finalmente se elabora un análisis exploratorio que confirme la covariación de la forma con factores considerados como casuales.

### 4.2.1. Obtención de los datos

Los datos que son utilizados para realizar un análisis morfométrico corresponden a un conjunto de hitos los cuales son representativos de una forma específica. La definición que provee Toro Ibacache *et al.,* (2010) para un hito es: el punto en el espacio que posee un nombre y coordenadas cartesiana (x, y) en formas bidimensionales, y (x, y, z) en tridimensionales, las cuales cumplen con la función de describir su posición en el espacio.

El hecho de que los hitos posean coordenadas cartesianas es muy importante, debido a que estás posteriormente son sometidas al análisis. Se deduce entonces que la planificación de la posición que ocuparán los hitos es preponderante, de la correcta planificación depende la calidad de la información que se obtendrá, así como la representatividad estadística del análisis morfométrico.

Para Toro Ibacache *et al.,* (2010), algunos criterios importantes a considerar son: Homología, Consistencia en la posición relativa, Cobertura adecuada de la forma, Repetibilidad, Coplanaridad, Hitos tipo 1, Hitos tipo 2, Hitos tipo 3, Pseudo-hitos, Semi-hitos.

La Homología se entiende como la manera en que los hitos están definidos, ya que las estructuras consideras homologas son aquellas cuya semejanza subyacente son el resultado de derivar a partir de un estructura ancestral común.

La Cobertura adecuada de la forma se refiere a que los hitos debe de recrear la forma que se está estudiando, por ejemplo, una cantidad insuficiente de hitos puede provocar la pérdida de información, y una cantidad exagerada lleva a resultados que de manera estadística no son confiables.

La repetibilidad se lleva a cabo con el fin de evitar un efecto significativo del error del observador, por lo tanto el hito debe de ser fácil de localizar y debe de encontrarse claramente definido.

La coplanaridad debe de ser considerada, puesto que puede llevar a un error de resultados, debido a que pueden existir variaciones considerables en la forma.

Los Hitos tipo 1 corresponden a los hitos localizados en la intersección de tres estructuras, centro de estructuras muy pequeñas, intersecciones de curvas. Con estos hitos se puede observar de manera más precisa el efecto de procesos biológicos como el crecimiento.

Los hitos tipo 2 corresponden a hitos ubicados en curvaturas máximas, donde existe aplicación de fuerzas biomecánicas.

Los hitos tipo 3 corresponden a hitos extremos cuya definición está dada por estructuras distantes.

Pseudo-hitos, corresponden a constructos definidos por términos matemáticos y anatómicos, como puntos tangentes de curvas.

Semi-hitos corresponden a puntos localizados en una curva de acuerdo a la posición de otros hitos o estructuras y que pueden ser desplazados levemente. (Toro Ibacache, et al., 2010)

### 4.2.2. Obtención de la Información de la forma (shape)

Como segunda etapa, se realiza una serie de procedimientos geométricos y estadísticos basados en la definición de la forma, durante esta etapa tiene importancia preponderante el entendimiento de la teoría de la forma y el análisis morfométrico, el espacio de la configuraciones, el espacio pre-Kendalliano pre-shape space, el espacio de la forma y la visualización de los cambios morfológicos, obtención de variables dependientes y función de placa delgada.

### 4.2.3. Análisis exploratorios y confirmatorios de covariación de la forma y factores casuales

Durante esta etapa se aplican diversos estudios, como lo son análisis exploratorios, análisis de componentes principales, análisis de varianzas canónicas, análisis de la deformación relativa, análisis para la prueba de hipótesis, regresiones, entre otros.

### 4.2.4. Análisis de Procrustes

El Análisis de Procrustes describe un conjunto de herramientas matemáticas que permiten comparar dos configuraciones de puntos homólogos provenientes de dos variantes de la misma entidad: dos individuos dentro de una población, dos especies, entre otros. El objetivo principal de esta técnica es determinar si representaciones alternativas de la misma $n$ puntos exhiben diferentes relaciones internas entre ellos. (Torcida & Pérez, 2012).

Los arreglos geométricos que se obtienen mediante varias escalas o coordenadas son una manera sencilla de representar la estructura y relación de un conjunto de elementos o factores, los cuales poseen en común una serie de atributos. La palabra Procrustes fue utilizada por primera vez por (Hurley & Cattell, 1962), para describir la correcta organización de estos arreglos, en relación a la palabra de origen griego Prokroústês: el ajustador. De manera inicial esta metodología fue usada para ajustar un arreglo sobre otro ya preestablecido, es decir, para adecuar una configuración como una transformación de matriz, la cual se denomina matriz objetivo. Lo anterior indica que la matriz transformada debe coincidir lo más posible con la matriz objetivo, a este proceso se le conoce como transformación procrusteana. (Zuliani, 2012).

Utilizando el criterio de rotar una matriz para lograr ajustarla a otra, es posible rotar varias matrices con respecto a una matriz central en común, de esta manera se logra un análisis de Procrustes generalizado.

Los procedimientos para realizar un análisis de Procrustes consisten en los siguientes: normalización, rotación, flexión y escalamiento de los datos bajo dos criterios principales, 1) que se mantengan las distancias entre individuos en las configuraciones originales, y que se minimice la suma de cuadrados entre puntos homólogos, es decir, lo que corresponden al mismo elemento.

Como ilustra (Zuliani, 2012), supóngase que cada matriz de partida está representada por $X_k\left(k=1,\ldots,q\right)$ con $n$ filas y $P_k$ columnas donde la $i$-ésima fila de las coordenadas de un punto (individuo) $P_i^{(k)}$ referido a $P_k$ ejes, el escalamiento, rotación y traslación pueden expresarse algebraicamente por la transformación:

$$X_k \rightarrow \rho_k X_k H_k + T_k$$

En la cual la matriz ortogonal de rotación $H_k$, el factor de escala $\rho_k$ y la matriz de traslación $T_k$ se hallarán de forma que se minimice:

$$S_r = \sum_{i=1}^{n}\sum_{k=1}^{q} d^2\left(P_i^{(k)}, G_i\right)$$

Donde d(A,B) es la distancia euclídea entre el par de puntos A y B, y $G_i$ es el centroide de los $q$ puntos homólogos $P_i^{(k)}(k=1,\ldots,q)$.

## 5. Conceptos básicos y fundamentos matemáticos

En esta sección se exponen los conceptos básicos y fundamentales en los que se sustenta este trabajo.

### 5.1. Conceptos básicos de estadística

Estadística: Parte de la ciencia que estudia las herramientas que nos permiten recopilar, ordenar, organizar y analizar un conjunto de datos provenientes de una muestra tomada aleatoriamente de una población con el fin de predecir, comparar y generalizar estos resultados hacia la población muestreada.

La estadística se divide en dos área: estadística descriptiva y estadística inferencial.

Estadística Descriptiva: Conjunto de técnicas ó procedimientos que nos permiten recolectar, organizar, analizar y presentar la información de una muestra en tablas y/o gráficas.

Estadística Inferencial o inductiva: Es el conjunto de métodos que nos permiten predecir, comparar y generalizar los resultados hacia la población muestreada.

Bioestadística: La estadística es una herramienta que se utiliza en muchas áreas como: la educación, psicología, sociología o economía, por mencionar algunas. Cuando las técnicas de estadística se usan para analizar información procedente de las ciencias biológicas o de la salud se le da el nombre de bioestadística.

Población: Es el conjunto de personas, animales u objetos de interés para un estudio particular.

La población puede ser finita o infinita. Una población finita es aquella que tiene un número limitado de valores, como por ejemplo, el número la cantidad de personas que asisten a un hospital en un día determinado, ó la cantidad de plaquetas en sangre que se encuentran en un paciente. La mayoría de las poblaciones se consideran finitas. Una población infinita es aquella que tiene un número ilimitado de elementos. Cuando una población es muy grande y es muy difícil llegar a contar el número de elemento que tiene, se pueden considerar ésta como población infinita. De hecho, existen fórmulas matemáticas para determinar el tamaño de muestra considerando poblaciones finitas e infinitas

Individuos o elementos: Cada una de las personas u objetos que contienen cierta información que se desea estudiar. Por ejemplo, un paciente, un niño, un mes, una célula.

Tamaño de la población: Es el número de individuos, animales u objetos que forman la población de interés.

Muestra: Es un subgrupo tomado de la población.

Muestra representativa: Es aquella que contiene el tamaño y las características en las mismas proporciones que la población. Por ejemplo, si se selecciona una muestra de una población formada por un 60% de hombres y el 40% de mujeres, la muestra debe contemplar a hombres y mujeres en los mismos porcentajes.

Parámetro: Medida de resumen calculado tomando en cuenta todos los elementos de una población. Un parámetro puede ser un promedio, un porcentaje, una desviación estándar, una mediana, etcétera.

Estadístico: Medida de resumen calculado tomando en cuenta todos los elementos de una muestra.

## 5.2. Fundamentos matemáticos

La gran mayoría de la información contenida en este capítulo se refiere a los fundamentos matemáticos – estadísticos de la Morfometría geométrica y fue tomada de (Cuadras, 2014).

### 5.2.1. Datos Multivariantes

El análisis multivariante (AM) es la parte de la estadística y del análisis de datos que estudia, analiza, representa e interpreta los datos que resultan de observar más de una variable estadística sobre una muestra de individuos. Las variables observables son homogéneas y correlacionadas, sin que alguna predomine sobre las demás. La información estadística en AM es de carácter multidimensional, por lo tanto la geometría, el cálculo matricial y las distribuciones multivariantes juegan un papel fundamental.

La información multivariante es una matriz de datos, pero a menudo, en AM la información de entrada consiste en matrices de distancias o similaridades, que miden el grado de discrepancia entre los individuos. Comenzaremos con las técnicas que se basan en matrices de datos $n \times p$, siendo $n$ el número de individuos y $p$ el de variables.

### 5.2.2. Matrices de datos

Supongamos que sobre los individuos $\omega_1,\ldots,\omega_n$ se han observado las variables $X_1,\ldots,X_p$. Sea $x_{ij}=X_j(\omega_i)$ la observación de la variable $X_j$ sobre el individuo $\omega_i$. La matriz de datos multivariantes es

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

Las filas de X se identifican con los individuos y las columnas de X con las variables. Indicaremos:

1. $x_i$ la fila *i*-ésima de X, que operaremos como un vector columna.
2. $X_j$ la columna *j*-ésima de X.
3. $\overline{\mathrm{x}}=(\overline{x}_1,\ldots,\overline{x}_j,\ldots,\overline{x}_p)'$ el vector columna de las medias de las variables, siendo

$$\overline{x}_j=\frac{1}{n}\sum_{i=1}^{n}x_{ij.}$$

4. La matriz simétrica $p\times p$ de covarianzas muestrales

$$\mathrm{S}=\begin{pmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{p1} & s_{p2} & \cdots & s_{pp} \end{pmatrix}$$

Siendo

$$s_{jj'}=\frac{1}{n}\sum_{i=1}^{n}(x_{ij}-\overline{x}_j)(x_{ij'}-\overline{x}_{j'})$$

La covarianza entre las variables $j, j'$. Naturalmente, $\overline{\mathrm{x}}$ y S son medidas multivariantes de tendencia central y dispersión, respectivamente.

5. La matriz simétrica $p \times p$ de correlaciones muestrales

$$\mathrm{R} = \begin{pmatrix} 1 & r_{12} & \dots & r_{1p} \\ r_{21} & 1 & \dots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \cdots & 1 \end{pmatrix}$$

Siendo $r_{jj'} = \mathrm{cor}(X_j, X_{j'})$ el coeficiente de correlación (muestral) entre las variables $X_j, X_{j'}$. Es coeficiente viene dado por

$$r_{jj'} = \frac{s_{jj'}}{s_j s_{j'}},$$

Donde $s_j, s_{j'}$ son las desviaciones típicas.

### 5.2.3. Matriz de centrado

Si $1 = (1,\dots,1)'$ es el vector columna de unos de orden $n \times 1$, y J=$11'$ es la matriz $n \times n$ de unos, ciertas características multivariantes se expresan mejor a partir de la matriz de centrado H, definida como

$$\mathrm{H} = \mathrm{I} - \frac{1}{n}\mathrm{J}.$$

Propiedades:

1. Simétrica: H′=H.
2. Idempotente: $\mathrm{H}^2 = \mathrm{H}$.
3. Los valores propios de H son cero o uno: $\mathrm{H_v} = \lambda \mathrm{v}$ implica $\lambda = 0$ ó 1.
4. 1 es vector propio de valor propio cero: H1=0, $1'\mathrm{H}=0'$.
5. El rango de H es $n-1$, es decir, rango$(\mathrm{H}) = n-1$.

### 5.2.4. Medias, covarianzas y correlaciones

Sean $\mathrm{X} = (x_{ij})$ la matriz de datos. La matriz de datos centrados se obtiene restando a cada variable su media: $\overline{\mathrm{X}} = (x_{ij} - \overline{x}_j)$. Esta matriz, así como el vector de

medias, las matrices de covarianzas y correlaciones, tienen expresiones matriciales simples.

1. $\bar{\mathrm{x}}'=\frac{1}{n}1'\mathrm{X}$.
2. Matriz de datos centrados:

   $\bar{\mathrm{X}}=\mathrm{X}-1\bar{\mathrm{x}}'=\mathrm{HX}$.
3. Matriz de covarianzas:

   $\mathrm{s}=\frac{1}{n}\bar{\mathrm{X}}'\mathrm{X}=\frac{1}{n}\mathrm{X}'\mathrm{HX}$.
4. Matriz de correlaciones:

   $\mathrm{R}=\mathrm{D}^{-1}\mathrm{S}\mathrm{D}^{-1}$, $\mathrm{S}=\mathrm{DRD}$,

   Siendo D la matriz diagonal con las desviaciones típicas de las variables.

### 5.2.5. VARIABLES COMPUESTAS

Algunos métodos de AM consisten en obtener e interpretar combinaciones lineales adecuadas de las variables observables. Una variable compuesta $Y$ es una combinación lineal de las variables observables con coeficientes $a=(a_1,\ldots,a_p)'$

$$Y=a_1X_1+\cdots+a_pX_p.$$

Si $X=[X_1,\ldots,X_p]$ es la matriz de datos, también podemos escribir

$$Y=\mathrm{Xa}$$

Si $Z=b_1X_1+\cdots+b_pX_p=\mathrm{Xb}$ es otra variable compuesta, se verifica:

1. $\bar{Y}=\bar{\mathrm{x}}'\mathrm{a}$, $\bar{Z}=\bar{\mathrm{x}}'\mathrm{b}$.
2. $\mathrm{var}(Y)=\mathrm{a}'\mathrm{Sa}$, $\mathrm{var}(Z)=\mathrm{b}'\mathrm{Sb}$.
3. $\mathrm{cov}(Y,Z)=\mathrm{a}'\mathrm{Sb}$.

Ciertas variables compuestas reciben diferentes nombres según la técnica multivariante: componentes principales, variables canónicas, funciones

discriminantes, etc. Uno de los objetivos del Análisis Multivariante es encontrar variables compuestas adecuadas que expliquen aspectos relevantes de los datos.

### 5.2.6. Transformaciones lineales

Sea T una matriz $p \times q$. Una transformación lineal de la matriz de datos es

$Y=XT.$

Las columnas $Y_1,\ldots,Y_q$ de Y son las variables transformadas.

Propiedades:

1. $\bar{y}'=\bar{x}'T,$ donde $\bar{y}$ es el vector (columna) de medias de Y.
2. $S_Y = T'ST,$ donde $S_Y$ es la matriz de covarianzas de Y.

   Demostración:

   $$\bar{y}' = \frac{1}{n}1'Y = \frac{1}{n}1'XT = \bar{x}'T.\ S_Y = \frac{1}{n}Y'HY = \frac{1}{n}T'X'HXT = T'ST.$$

### 5.2.7. Teorema de la dimensión

La matriz de covarianzas S es (semi) definida, positiva, puesto que:

$$a'Sa = \frac{1}{n}a'X'HXa = \frac{1}{n}a'X'HHXa = b'b \geq 0, \text{ siendo } b = n^{-1/2}HXa.$$

El rango *r* = rango (S) determina la dimensión del espacio vectorial generado por las variables observables, es decir, el número de variables linealmente independientes es igual al rango de S.

**Teorema 5.1:** Si *r* = *rango* (S) $\leq p$ hay r variables linealmente independientes y las otras $p-r$ son combinación lineal de estas *r* variables.

Demostración: podemos ordenar las *p* variables de manera que la matriz de covarianzas $S_r$ de $X_1,\ldots,X_r$ sea no singular

$$S_r = \begin{pmatrix} s_{11} & \cdots & s_{1r} \\ \vdots & \ddots & \vdots \\ s_{r1} & \cdots & s_{rr} \end{pmatrix}$$

Sea $X_j, J > r$. La fila $(s_{j1},\ldots,s_{jr})$ será combinación lineal de las filas de $S_{r.}$ Luego las covarianzas $s_{jl},\ldots,s_{jr}$ entre $X_j$ y $X_1,\ldots,X_r$ verifican:

$$s_{jj} = \sum_{i=1}^{r} a_i s_{ji}, \quad s_{ji} = \sum_{i'=1}^{r} a_{i'} s_{ii'.}$$

Entonces

$$\text{var}(X_j - \textstyle\sum_{i=1}^{r} a_1 X_i) \;= s_{jj} + \sum_{i,i'=1}^{r} a_i a_{i'} s_{ii'} - 2\sum_{i=1}^{r} a_i s_{ji}$$

$$= \sum_{i=1}^{r} a_i s_{ji} + \sum_{i=1}^{r} a_i (\sum_{i'=1}^{r} a_{i'} s_{ii'}) - 2\sum_{i=1}^{r} a_i s_{ji}$$

$$= \sum_{i=1}^{r} a_i s_{ji} + \sum_{i=1}^{r} a_i s_{ji} - 2\sum_{i=1}^{r} a_i s_{ji}$$

$$= 0.$$

Por lo tanto

$$X_j - \sum_{i=1}^{r} a_i X_i = c \Rightarrow X_j = c + \sum_{i=1}^{r} a_i X_i$$

Donde *c* es una constante.

Corolario: si todas las variables tienen varianza positiva (es decir, ninguna se reduce a una constante), y *r=rango* (R) $\leq p,$ hay r variables linealmente independientes y las otras $p-r$ son combinación lineal de estas r variables.

### 5.2.8. Medidas globales de variabilidad y dependencia

Una medida de la variabilidad global de las $p$ variables debe ser función de la matriz de covarianzas S. Sean $\lambda_1,\ldots,\lambda_p$ los valores propios de S. Las siguientes medidas tienen especial interés en AM.

a) Varianza generalizada:

$$|\mathrm{S}| = \lambda_1 \times \ldots \times \lambda_p .$$

b) Variación total:

$$\mathrm{tr}(\mathrm{S}) = \lambda_1 + \ldots + \lambda_p$$

Una medida de dependencia global debe ser función de la matriz de correlaciones **R.** Un coeficiente de dependencia es

$$\eta^2 = 1 - |\mathrm{R}|,$$

que verifica:

1. $0 \leq \eta^2 \leq 1.$

2. $\eta^2 = 0$ si y sólo si las $p$ variables están incorrelacionadas.

3. $\eta^2 = 1$ si y sólo si hay relaciones lineales entre las variables.

Demostración:

1. Sean $\lambda_1, \ldots, \lambda_p$ los valores propios de **R**. Si $g$ y $a$ son las medias geométrica y aritmética de $p$ números positivos, se verifica $g \leq a$. Entonces, de

   $\mathrm{tr}(\mathrm{R}) = p,$

   $$|\mathrm{R}|^{1/p} = \left(\lambda_1 \times \ldots \times \lambda_p\right)^{1/p} \leq \left(\lambda_1 + \ldots + \lambda_p\right)/p = 1,$$

   y por lo tanto $0 \leq |\mathrm{R}| \leq 1.$

2. **R**=**I** (matriz de identidad) si y sólo si las $p$ variables están incorrelacionadas, luego $1 - |\mathrm{I}| = 0.$

3. Si $\eta^2 = 1$, es decir $|\mathrm{R}| = 0$, entonces $\mathrm{rango}(\mathrm{R}) < p$ y por lo tanto existen relaciones lineales entre las variables.

### 5.2.9. DISTANCIAS

Algunos métodos de AM están basados en criterios geométricos y en la noción de distancia entre individuos y entre poblaciones. Si $X=\begin{pmatrix} x_1' \\ \vdots \\ x_n' \end{pmatrix}$ es una matriz de datos, con matriz de covarianzas S, las tres definiciones más importantes de distancias entre las filas $x_i' = (x_{i1,\cdots,}x_{ip})$, $x_j' = (x_{j1,\cdots,}x_{jp})$ de X son:

1. Distancias euclídea:

$$d_E(i,j) = \sqrt{\sum_{h=1}^{p}\left(x_{ih} - x_{jh}\right)^2}.$$

2. Distancia de K. Pearson

$$d_{p(i,j)} = \sqrt{\sum_{h=1}^{p}\left(x_{ih} - x_{jh}\right)^2 \Big/ s_{hh,}}$$

donde $s_{hh}$ es la covarianza de la variable $x_{h.}$

3. Distancia de Mahalanobis:

$$d_M(i,j) = \sqrt{\left(x_i - x_j\right)' S^{-1}\left(x_i - x_j\right)}.$$

Observaciones

Un cambio de escala de una variable $X_j$ es una transformación $Y_j = \alpha X_{j,}$ donde $\alpha$ es una constante. Comparando las tres distancias, se concluye que $d_M$ es muy adecuada en AM debido a que verifica:

a) $d_E$ supone implícitamente que las variables están incorrelacionadas y no es invariante por cambios de escala.

b) $d_p$ también supone que las variables están incorrelacionadas pero es invariante por cambios de escala.

c) $d_M$ tiene en cuenta las correlaciones entre las variables y es invariante por transformaciones lineales no singulares de las variables, en particular cambios de escala.

Las distancias $d_E$ y $d_p$ son casos particulares de $d_M$ cuando la matriz de covarianzas es la identidad $\mathrm{I}_p$ y $\mathrm{diag}(\mathrm{S})$, respectivamente. En efecto:

$$d_E(i,j)^2 = \left(x_i - x_j\right)' \left(x_i - x_j\right),$$

$$d_p(i,j)^2 = \left(x_i - x_j\right)' \left[\mathrm{diag}(\mathrm{S})\right]^{-1} \left(x_i - x_j\right).$$

La distancia de Mahalanobis (al cuadrado) puede tener otras versiones:

1. Distancia de una observación $x_i$ al vector de medias $\bar{\mathrm{x}}$ de X:

$$\left(\mathrm{x}_i - \bar{\mathrm{x}}\right)' \mathrm{S}^{-1} \left(\mathrm{x}_i - \bar{\mathrm{x}}\right)$$

2. Distancia entre dos poblaciones representadas por dos matrices de datos $\mathrm{X}_{n1\times p}$ , $\mathrm{Y}_{n2\times p}$ :

$$\left(\bar{\mathrm{x}} - \bar{\mathrm{y}}\right)' \mathrm{S}^{-1} \left(\bar{\mathrm{x}} - \bar{\mathrm{y}}\right),$$

donde $\bar{\mathrm{x}},\bar{\mathrm{y}}$ son los vectores de medias y

$$\mathrm{S}=\left(n_1\mathrm{S}_1 + n_2\mathrm{S}_2\right)/\left(n_1 + n_2\right)$$

es la media ponderada de las correspondientes matrices de covarianzas.

### 5.2.10. Algunos aspectos del cálculo matricial

#### 5.2.10.1. Descomposición singular

Sea A una matriz de orden $m \times n$ con $m \geq n$. Se llama descomposición en valores singulares de A a

$$\mathrm{A}=\mathrm{U}\mathrm{D}_S\mathrm{V}'$$

Donde U es una matriz $m \times n$ cuyas columnas son vectores ortonormales, $\mathrm{D}_s$ es una matriz diagonal $n \times n$ con los valores singulares

$s_1 \geq \cdots \geq s_r \geq s_r + 1 = \cdots = s_n = 0,$

Y V es una matriz $n \times n$ ortogonal. Se verifica:

1. El rango de **A** es el número $r$ de los valores singulares positivos.
2. **U** contiene los vectores propios (unitarios) de $AA'$, SIENDO $U'U=I_n$ .
3. **V** contiene los vectores propios (unitarios) de $AA'$, siendo $V'V=VV'=I_n$ .
4. Si $m = n$ y **A** es simétrica, entonces U=V y $A=UD_sU'$ es la descomposición espectral de **A**. Los valores singulares son los valores propios de **A.**

#### 5.2.10.2. INVERSA GENERALIZADA

Si **A** es una matriz cuadrada de orden $n \times n$ no singular, es decir, $\text{rango}(A) = n$, existe la matriz inversa $A^{-1}$ tal que

$$AA^{-1} = A^{-1}A = I_n$$

Si el $\text{rango}(A) = r < n$, o A no es matriz cuadrada, la inversa no existe, pero existe la inversa generalizada o $g$-inversa $A^-$.

Sea **A** un matriz de orden $m \times n$ con $m \geq n$. Se llama inversa generalizada de **A** o $g$-inversa, a una matriz $A^-$ que verifica:

$$AA^-A = A$$

La $g$-inversa no es única, pero si $A^-$ verifica además:

$A^-AA^- = A^-$, $(AA^-)' = AA^-$ $(A^-A)' = A^-A$,

Entonces la g-inversa $A^-$ es única.

Sea rango $(A) = r$ y $A=UD_SV'$ la descomposición singular de A, con

$D_S = \text{diag}(s_1, \ldots, s_r, 0, \ldots, 0)$.

Entonces

$\mathrm{D}_s^- = \mathrm{diag}(s_1^{-1},\ldots,s_r^{-1},0,\ldots,0)$.

Y la matriz $m \times n$

$\mathrm{A}^- = \mathrm{V}\mathrm{D}_s^-\mathrm{U}'$

Es una g-inversa de A. En efecto,

$\mathrm{A}\mathrm{A}^-\mathrm{A} = \mathrm{U}\mathrm{D}_S\mathrm{V}'\mathrm{V}\mathrm{D}_S^-\mathrm{U}'\mathrm{U}\mathrm{D}_S\mathrm{V}' = \mathrm{A}.$

### 5.2.10.3. Aproximación matricial de rango inferior

Sea $\mathrm{A} = (a_{ij})$ un matriz de orden $m \times n$ con $m \geq n$ y con rango $r$. Supongamos que deseamos aproximar $\mathrm{A}$ por otra matriz $\mathrm{A}^* = (a_{ij}^*)$, del mismo orden $m \times n$ pero de rango $k < r$, de modo que

$$\mathrm{tr}\left[\left(\mathrm{A}-\mathrm{A}^*\right)'\left(\mathrm{A}-\mathrm{A}^*\right)\right] = \sum_{i=1}^{m}\sum_{j=1}^{n}\left(a_{ij}-a_{ij}^*\right)^2 = \text{mínimo}.$$

Si $\mathrm{A}=\mathrm{U}\mathrm{D}_S\mathrm{V}'$ es la descomposición en valore singulares de A, entonces la solución viene dada por

$\mathrm{A}^* = \mathrm{U}\mathrm{D}_S^*\mathrm{V}',$

Donde $D_S^*$ es diagonal con los $k$ primeros valores singulares de A, siendo nulos los restantes valores, es decir:

$\mathrm{D}_S^* = \mathrm{diag}\left(s_1,\ldots,s_k,0,\ldots,0\right).$

El mínimo es la suma de los cuadrados de los valores singulares eliminados, es decir, $\mathrm{tr}\left[\left(\mathrm{D}_s-\mathrm{D}_s^*\right)^2\right]$. Esta es la llamada aproximación de Eckart-Young. Por ejemplo, si

$$\mathrm{A} = \begin{pmatrix} 1 & 3 & 2 \\ 2 & 0 & 1 \\ 4 & 5 & 6 \\ 3 & 2 & 1 \end{pmatrix}$$

Entonces

$$\mathrm{A} = \begin{pmatrix} 0.35 & -0.42 & 0.52 \\ 0.16 & 0.61 & -0.41 \\ 0.86 & -0.19 & -0.38 \\ 0.33 & 0.63 & 0.63 \end{pmatrix} \begin{pmatrix} 10.14 & 0 & 0 \\ 0 & 2.295 & 0 \\ 0 & 0 & 1.388 \end{pmatrix} \begin{pmatrix} -0.50 & -0.59 & -0.62 \\ 0.86 & -0.40 & -0.31 \\ 0.06 & 0.70 & -0.71 \end{pmatrix}$$

Y la aproximación de rango 2 es

$$\mathrm{A}^{*} = \begin{pmatrix} 0.945 & 2.480 & 2.534 \\ 2.015 & 0.397 & 0.587 \\ 3.984 & 5.320 & 5.628 \\ 2.936 & 1.386 & 1.652 \end{pmatrix}$$

Siendo (redondeando a dos decimales)

$$\mathrm{A}^{*} = \begin{pmatrix} 0.35 & -0.42 & 0.52 \\ 0.16 & 0.61 & -0.41 \\ 0.86 & -0.19 & -0.38 \\ 0.33 & 0.63 & 0.63 \end{pmatrix} \begin{pmatrix} 10.14 & 0 & 0 \\ 0 & 2.29 & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} -0.50 & -0.59 & -0.62 \\ 0.86 & -0.40 & -0.31 \\ 0.06 & 0.70 & -0.71 \end{pmatrix}$$

El valor mínimo es $1.388^2 = 1.926,$ el cuadrado del valor singular eliminado. En particular, si B es matriz simétrica semidefinida positiva de rango $r$ y $\mathrm{B} = \mathrm{TD}_{\lambda}\mathrm{T}'$ es la descomposición espectral (con los valores propios ordenados de mayor a menor), entonces la mejor aproximación de rango $k < r$ es la matriz

$$\mathrm{B}^{*} = \mathrm{TD}_{\lambda}^{*}\mathrm{T}',$$

Donde $\mathrm{D}_{\lambda}^{*}$ contiene los $k$ valores propios de B.

### 5.2.10.4. Transformación de Procrustes

Sea A un matriz de orden $m \times n$ con $m \geq n$. Sea B otra matriz del mismo orden y escala (misma media y varianza para las columnas). Supongamos que queremos transformar A en AT, siendo T matriz $n \times n$ ortogonal, de modo que AT sea lo más próxima posible a B, es decir $\mathrm{tr}\left[(\mathrm{AT}-\mathrm{B})'(\mathrm{AT}-\mathrm{B})\right] = \text{mínimo}$. Si obtenemos la descomposición en valores singulares

$$\mathrm{A}'\mathrm{B} = \mathrm{U}\mathrm{D}_s\mathrm{V}',$$

Entonces la solución es

$$\mathrm{T} = \mathrm{U}\mathrm{V}'.$$

Se conoce AT como la transformación Procrustes.

En el caso general, sean X,Y dos matrices $n \times p$, con $n \geq p$, y vectores (filas) de medias $\overline{\mathrm{x}},\overline{\mathrm{y}}$. Deseamos aproximar X a Y mediante contracción, traslación y rotación. Consideremos la transformación

$$\mathrm{Y}^* = b\mathrm{X}\mathrm{T} + 1\mathrm{c},$$

Donde $b$ es una constante escalar, T es matriz $p \times p$ ortogonal, 1 es el vector $n \times 1$ de unos y c es un vector (fila) $1 \times p$ de constantes. Se trata de encontrar $b, \mathrm{T}, \mathrm{c}$, de modo que $\mathrm{Y}^*$ sea lo más próximo posible a Y en el sentido de que $\mathrm{tr}\left[\left(\mathrm{Y}-\mathrm{Y}^*\right)'\left(\mathrm{Y}-\mathrm{Y}^*\right)\right] = \text{mínimo}$. Es decir, para cada par de columnas $x_j, y_j$ se desea hallar el vector

$$y_j^* = b\mathrm{T}'\mathrm{x}_j + c_j 1$$

Lo más próximo a posible a $y_j$.

Si $\overline{\mathrm{X}},\overline{\mathrm{Y}}$ son las matrices centradas, obtenemos primero la descomposición singular

$$\mathrm{X}'\mathrm{Y} = \mathrm{U}\mathrm{D}_S\mathrm{V}'.$$

Indicando $\mathrm{M}^{1/2} = \mathrm{F}\Delta^{1/2}\mathrm{F}'$, siendo $\mathrm{M}=\mathrm{F}\Delta\mathrm{F}'$ la descomposición espectral de la matriz simétrica $\mathrm{M} = \overline{\mathrm{X}}'\overline{\mathrm{Y}}\overline{\mathrm{Y}}'\overline{\mathrm{X}}$, la solución es

$$b = \mathrm{tr}\left(\overline{\mathrm{X}}'\overline{\mathrm{Y}}\overline{\mathrm{Y}}'\overline{\mathrm{X}}\right)^{1/2} / \mathrm{tr}\left(\overline{\mathrm{X}}'\overline{\mathrm{X}}\right),\ \ \mathrm{T}=\mathrm{U}\mathrm{V}',\ \ \mathrm{c}=\overline{\mathrm{y}} - b\overline{\mathrm{x}}\mathrm{T}.$$

Una medida del grado de relación lineal entre X e Y, llamada coeficiente Procrustes, y que toma valores entre 0 y 1, es

$$P_{XY}^2 = \left[\mathrm{tr}\left(\overline{\mathrm{X}}'\overline{\mathrm{Y}}\overline{\mathrm{Y}}'\overline{\mathrm{X}}\right)^{1/2}\right]^2 / \mathrm{tr}\left(\overline{\mathrm{X}}'\overline{\mathrm{X}}\right)\mathrm{tr}\left(\overline{\mathrm{Y}}'\overline{\mathrm{Y}}\right).$$

Este coeficiente se puede expresar también en términos de matrices de covarianzas, pero no es invariante por transformaciones lineales aplicadas por separado a X y a Y.

Si $p = 1$ el análisis procrustes equivale a la regresión lineal $y^* = bx + \overline{y} - b\overline{x}$, siendo $b = s_{xy} / s_x^2$ y $P_{XY} = s_{xy} / \left(s_x s_y\right)$ los coeficiente de regresión y correlación ordinarios.

### 5.3. ANÁLISIS CANÓNICO DE POBLACIONES

El Análisis de componentes principales permite representar los individuos de una población mediante una única matriz de datos, sin embargo, cuando se tienen varias matrices de datos como resultado de observar variables sobre varias poblaciones es necesario la utilización de Análisis Canónico de Poblaciones (CANP).

Supóngase que de la observación de $p$ variables cuantitativa $x_{1,}...,x_p$ sobre $g$ poblaciones se obtiene $g$ matrices de datos

$$x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_g \end{pmatrix} \begin{matrix} n_1 \times p \\ n_2 \times p \\ \vdots \\ n_g \times p \end{matrix}$$

Donde $x_i$ es la matriz $n_i \times p$ de la población *i*. Sean $\overline{x}_1', \overline{x}_2', ... \overline{x}_g'$ los vectores (fila) de las medias de cada población. $x$ es de orden $n \times p_1$, siendo ahora $n = \sum_{i=1}^{g} n_i$. Se indica

$$\overline{x} = \begin{pmatrix} \overline{x}_1' - \overline{x}' \\ \overline{x}_2' - \overline{x}' \\ \vdots \\ \overline{x}_g' - \overline{x}' \end{pmatrix}$$

La matriz $g \times p$ con las medias de las $g$ poblaciones. Se tienen dos maneras de cuantificar matricialmente la dispersión entre las poblaciones:

- La matriz de dispersión no ponderada entre grupos

$$\mathrm{A} = \overline{x}'\overline{x} = \sum_{i=1}^{g} (\overline{x}_i - \overline{x})(\overline{x}_i - \overline{x})'$$

- La matriz de dispersión ponderada entre grupos

$$\mathrm{B} = \sum_{i=1}^{g} n_i (\overline{x}_i - \overline{x})(\overline{x}_i - \overline{x})'$$

La matriz A es proporcional a una matriz de covarianzas tomando como datos sólo las medias de las poblaciones. La matriz B participa, juntamente con W (matriz de dispersión dentro de grupos) en el test de comparación de medias de $g$ poblaciones. Aquí se trabajará con la matriz A, si bien los resultados serían parecidos si utilizáramos la matriz B. También se hará uso de la matriz de covarianzas.

$$\mathrm{S} = \frac{1}{n-g} \sum_{i=1}^{g} n_i \mathrm{s}_i$$

Entonces $A = \overline{x}'\overline{x}$ juega el papel de matriz de covarianzas "entre" las poblaciones, S juega el papel de matriz de covarianzas "dentro" de las poblaciones.

### 5.4. Variables canónicas

Definición: Sean $V=\left[v_1,...,v_p\right]$ los vectores propios de $A = \overline{x}'\overline{x}$ respecto de S con valores propios $\lambda > \cdots > \lambda_p$ es decir,

$$Av_i = \lambda_i S_i v_i \ ,$$

Normalizados según

$$v_i' S_i v_i = 1$$

Los vectores $v_1,...,v_p$ son los vectores canónicos y las variables canónicas son las variables compuestas

$$Y_i = Xv_i$$

Si $v_i = \left(v_{1i},...v_{pi}\right)'$ y $x = \left[x_1,...,x_p\right]$, la variable canónica $Y_i$ es la variable compuesta

$$Y_i = Xv_i = v_{1i}X_1 + \cdots + v_{pi}X_p$$

Que tiene *S*- varianza 1 y *A*- varianza $\lambda_i$, es decir:

$$var_A(Y_1) = v_i' A v_i = \lambda_{i,} \qquad var_s(Y_i) = v_i' S_i v_i = 1$$

Se trabajará con $p$ variables canónicas, pero de hecho el número efectivo es $k = \text{mín}\{p, g-1\}$

Las variables canónicas verifican

1. Son incorrelacionadas dos a dos respecto a A y también respecto a S

$cov_A(Y_{i,}Y_j) = cov_s(Y_{i,}Y_j) = 0$ si $i \neq j$

2. Las A- varianzas son respectivamente máximas:

$$\mathrm{var}_A(Y_1) = \lambda_1 > \cdots > \mathrm{var}_A(Y_p) = \lambda_p$$

En el sentido de que $Y_1$ es la variable con máxima varianza entre grupos, condicionada a varianza 1 dentro de grupos, $Y_2$ es la variable con máxima varianza entre grupos, condicionada a estar incorrelacionada con $Y_1$ y tener varianza 1 dentro de grupos.

Demostración: supóngase $\lambda_1 > \cdots > \lambda_p > 0$ Probemos que las variables compuestas $Y_i = \mathrm{Xt}_{i,}\ i = 1,...,\mathrm{p}$ están incorrelacionadas:

$$\begin{aligned}\mathrm{cov}_A(Y_i, Y_j) &= \mathrm{t}_i'\mathrm{At}_j = \mathrm{t}_i'\mathrm{S}\lambda_j\mathrm{t}_j = \lambda_j\mathrm{t}_i'\mathrm{St}_{j,}\\ \mathrm{cov}_A(\mathrm{Y}_j, \mathrm{Y}_i) &= \mathrm{t}_j'\mathrm{At}_i = \mathrm{t}_j'\mathrm{S}\lambda_j\mathrm{t}_i = \lambda_i\mathrm{t}_j'\mathrm{St}_{i,}\end{aligned}$$

Restando $(\lambda_j - \lambda_i)\mathrm{t}_i'\mathrm{st}_j = 0 \Rightarrow \mathrm{t}_i'\mathrm{st}_j = 0 \Rightarrow \mathrm{cov}_A(\mathrm{Y}_i, \mathrm{Y}_j) = \lambda_j\mathrm{t}_i'\mathrm{St}_j = \mathrm{cov}_A(Y_i, Y_j) = 0,$ si $i \neq j.$ Además, de $\mathrm{t}_i'\mathrm{St}_j = 1$:

$$\mathrm{var}_A(Y_i) = \lambda_i\mathrm{t}_i'\mathrm{St}_j = \lambda_i.$$

Sea ahora $Y = \sum_{i=1}^{p} a_i x_i = \sum_{i=1}^{p} \alpha_i Y_i$ una variable compuesta tal que $\mathrm{var}_s(Y) = \sum_{i=1}^{p} \alpha_i^2\ \mathrm{var}_s(Y_i) = \sum_{i=1}^{p} \alpha_i^2 = 1.$ entonces $\mathrm{var}_A(Y)$ es:

$\mathrm{var}_A(\sum_{i=1}^{p} \alpha_i Y_i) = \sum_{i=1}^{p} \alpha_i^2\ \mathrm{var}_A(Y_i) = \sum_{i=1}^{p} \alpha_i^2\lambda_i \leq (\sum_{i=1}^{p} \alpha_i^2) = \mathrm{var}_A(Y_1),$ que prueba que $Y_1$ tiene máxima varianza entre grupos. Considérese a continuación las variables $Y$ incorrelacionadas con $Y_1$, que se pueden expresar como:

$$Y = \sum_{i=2}^{p} \beta_i Y_i \text{ condicionado a } \sum_{i=2}^{p} \beta_i^2 = 1.$$

Entonces $\mathrm{var}_A(Y)$ es:

$$\text{var}_A\left(\sum_{i=2}^{p}\beta_i Y_i\right)=\sum_{i=2}^{p}\beta_i^2\,\text{var}_A(Y_i)=\sum_{i=2}^{p}\beta_i^2\lambda_i\le\left(\sum_{i=2}^{p}\beta_i^2\right)\lambda_2=\text{var}_A(Y_2),$$ y por lo tanto $Y_2$ está incorrelacionada con $Y_1$ y tiene varianza máxima. La demostración para $Y_3,...,Y_p$ es análoga.

### 5.5. Distancia de Mahalanobis y transformación canónica

La distancia de Mahalanobis para dos poblaciones se considera una medida de la diferencia entre las medias de las poblaciones, pero teniendo en cuenta las covarianzas. Definición: Considérese muestras multivariantes de *g* poblaciones con vectores de medias $\bar{x}_1,\bar{x}_2,...,\bar{x}_g$ y matriz de covarianzas (común) S. La distancia al cuadrado de Mahalanobis entre las poblaciones *i, j* es

$$M^2(i,j)=(\bar{x}_i-\bar{x}_j)'S^{-1}(\bar{x}_i-\bar{x}_j).$$

Si $\bar{x}$ es la matriz centrada con los vectores de medias y $V=\left[v_1,...,v_p\right]$ es la matriz con los vectores canónicos (vectores propios de $A=\bar{x}'\bar{x}$ respecto de S), la transformación canónica es

$$Y=\bar{x}v$$

La matriz Y de orden $g\times p$ contiene las coordenadas canónicas de las g poblaciones.

**Teorema 5.2.** La distancia de Mahalanobis entre cada par de poblaciones $i,j$ coincide con la distancia euclídea entre las filas $i,j$ de la matriz de coordenadas canónicas Y. Si $y_i=\bar{x}_i V$ entonces

$$d_E^2(i,j)=(y_i-y_j)'(y_i-y_j)=(\bar{x}_i-\bar{x}_j)'S^{-1}((\bar{x}_i-\bar{x}_j).\quad(5.1)$$

Demostración: Basta probar que los productos escalares coinciden

$$y_i y_j'=\bar{x}_i S^{-1}\bar{x}_j'\Leftrightarrow\bar{X}S^{-1}\bar{X}'=YY'.\qquad(5.2)$$

Sea $\Lambda = \mathrm{diag}(\lambda_1, ..., \lambda_p)$ la matriz diagonal con los valores propios de $A = \bar{X}'\bar{X}$ respecto de S. Entonces

$$AV=SV\Lambda \text{ con } V'SV=I_{P,}$$

y la transformación canónica es $Y = \bar{X}V$.

$AV=SV\Lambda$ es $\bar{X}'\bar{X}V=SV\Lambda$, luego $S^{-1}\bar{X}'\bar{X}V=V\Lambda$ y premultiplicando por $\bar{X}$ tenemos $\bar{X}S^{-1}\bar{X}'\bar{X}V=\bar{X}V\Lambda$, es decir,

$$\bar{X}S^{-1}\bar{X}'Y=Y\Lambda$$

Con lo cual $Y$ contiene los vectores propios de $\bar{X}S^{-1}\bar{X}'$, luego cumple la descomposición espectral

$$\bar{X}S^{-1}\bar{X}' = Y\Lambda Y'$$

Suponiendo $Y$ ortogonal. Tomando $Y\Lambda^{1/2}$ que indicamos también por Y, obtenemos finalmente $\bar{X}S^{-1}\bar{X}'=YY'$.

### 5.6. REPRESENTACIÓN CANÓNICA

La representación de las *g* poblaciones mediante las filas de $\bar{X}$ con la métrica de Mahalanobis es bastante complicada: la dimensión puede ser grande y los ejes son oblicuos. En cambio, la representación mediante las coordenadas canónicas Y con la métrica euclídea se realiza a lo largo de ejes ortogonales. Si, además, tomamos las *m* primeras coordenadas canónicas (usualmente m=2), la presentación es totalmente factible y es óptima en dimensión reducida, en el sentido que maximiza la variabilidad geométrica.

**Teorema 5.3:** la variabilidad geométrica de las distancias de Mahalanobis entre las poblaciones es proporcional a la suma de los valores propios:

$$V_M\left(\bar{X}\right) = \frac{1}{2g^2}\sum_{i,j=1}^{g} M(i,j)^2 = \frac{1}{g}\sum_{i=1}^{p}\lambda_i. \quad (5.3)$$

Si $\mathrm{Y}=\bar{\mathrm{X}}\mathrm{V}$, donde V, de orden $p\times m$ es la matriz de la transformación canónica en dimensión m y

$$\delta_{ij}^2(m)=(\mathrm{y}_1-\mathrm{y}_j)(\mathrm{y}_i-\mathrm{y}_j)'=\sum_{h=1}^{m}(\mathrm{y}_{ih}-\mathrm{y}_{jh})^2$$

*Es la distancia euclídea (al cuadrado) entre dos filas de Y, la variabilidad geométrica en dimensión* $m\le p$ *es*

$$V_\delta(\mathrm{Y})_m=\frac{1}{2g^2}\sum_{i,j=1}^{g}\delta_{ij}^2(m)=\frac{1}{g}\sum_{i=1}^{m}\lambda_i,$$

*Y esta cantidad es máxima entre todas las transformaciones lineales posibles en dimensión m.*

Demostración:

De $\frac{1}{2n^2}\sum_{i,j=1}^{n}(x_i-x_j)^2=s^2$ y de la ecuación (5.1) se tiene que:

$$V_M(\mathrm{X})=\frac{1}{2g^2}\sum_{i,j=1}^{g}M(i,j)^2=\frac{1}{2g^2}\sum_{i,j=1}^{g}\sum_{h=1}^{p}(y_{ih}-y_{jh})^2=s_1^2+\cdots+s_p^2$$

Donde $s_j^2=\left(\sum_{i=1}^{g}y_{ij}^2\right)/g$ representa la varianza ordinaria de la columna $Y_J$ de Y. Esa suma de varianzas es

$$\mathrm{tr}(\frac{1}{g}\mathrm{Y}'\mathrm{Y})=\frac{1}{g}\mathrm{tr}(\mathrm{V}'\bar{\mathrm{X}}'\bar{\mathrm{X}}\mathrm{V})=\frac{1}{g}\mathrm{tr}\ (\mathrm{V}'\mathrm{A}\mathrm{V})=\frac{1}{g}\mathrm{tr}(\Lambda)$$

Lo que prueba (5.3)

Sea ahora $\tilde{\mathrm{Y}}=\bar{\mathrm{X}}\mathrm{T}$ otra transformación de $\bar{\mathrm{X}}$ tal que $\mathrm{T}'\mathrm{S}\mathrm{T}=\mathrm{I}$. Indicando $\mathrm{T}=\left[\mathrm{t}_1\ldots,\mathrm{t}_p\right]$, la A-varianza de la primera columna $\tilde{Y}_1$ de $\tilde{\mathrm{Y}}$ es $\mathrm{t}_1'\mathrm{A}\mathrm{t}_1\le\mathrm{v}_1'\mathrm{A}\mathrm{v}_1=\lambda 1$. Es decir, la varianza ordinaria $s^2(\tilde{\mathrm{Y}}_1)=g^{-1}\tilde{Y}_1'\,\tilde{Y}_1=g^{-1}\mathrm{t}_1'\bar{\mathrm{X}}'\bar{\mathrm{X}}\mathrm{t}_1$ es máxima para $Y_1=\bar{\mathrm{X}}\mathrm{v}_1$, primera

columna de Y. Análogamente se demuestre para las demás columnas (segunda, tercera, etc.,. coordenadas canónicas). Se tiene pues

$$V_\delta(\tilde{\mathrm{Y}})_m = \sum_{k=1}^{m} s^2(\tilde{Y}_k) = \frac{1}{g}\sum_{k=1}^{m} \mathrm{var}_A(\tilde{Y}_k) \le V_\delta(\mathrm{Y})_m = \frac{1}{g}\sum_{k=1}^{m} \lambda_k .$$

El porcentaje de variabilidad geométrica explicada por las *m* primeras coordenadas canónicas es

$$P_m = 100\frac{V(\mathrm{Y})_m}{V_M(\bar{\mathrm{X}})} = 100\frac{\lambda_1 + \cdots + \lambda_m}{\lambda_1 + \cdots + \lambda_p}$$

### 5.7. Análisis de la varianza (Anova)

El análisis de la varianza comprende un conjunto de técnicas estadísticas que permiten analizar cómo operan diversos factores, estudiados simultáneamente en un diseño factorial, sobre una variable respuesta.

### 5.8. Diseño de un factor

Supóngase que las observaciones de una variable Y solamente dependen de un factor con *k* niveles:

$$\begin{array}{lcccc} \text{Nivel 1} & y_{11} & y_{12} & \cdots & y_{1n_1} \\ \text{Nivel 2} & y_{21} & y_{22} & \cdots & y_{2n_2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \text{Nivel k} & y_{k1} & y_{k2} & \cdots & y_{kn_k} \end{array}$$

Si escribimos $\mu_i = \mu + \alpha_i$, en el modelo $H = \begin{pmatrix} 010\cdots0 \\ 001\cdots0 \\ \vdots\vdots\vdots\ddots\vdots \\ 000\cdots1 \end{pmatrix}$, rango (H) $= m$ se tiene

$$y_{ih} = \mu_i + e_{ih}, \quad i = 1,\ldots,k; \quad h = 1,\ldots,n_i,$$

Donde $\mu$, es la media de la variable en el nivel $i$. Indicamos:

Media nivel *i*: $y_i = (1/n_i)\sum_h y_{ih}$

Media general: $\overline{y} = (1/n)\sum_i \sum_h y_{ih}$

No. Total de observaciones: $n = n_1 + \ldots + n_k$

Indiquemos también:

Suma de cuadrados entre grupos: $Q_E = \sum_i n_i (y_i - \overline{y})^2$

Suma de cuadrados dentro de grupos: $Q_D \sum_i \sum_h (y_{ih} - y_i)^2$

Suma de cuadrados total: $Q_T = \sum_i \sum_h (y_{ih} - \overline{y})^2$

Se verifica la relación fundamental

$Q_T = Q_E + Q_D$

Las estimaciones LS de las medias $\mu_i$ son

$\hat{\mu}_i = y_i., \quad i = 1, \ldots, k,$

Y la suma de cuadrados residual es $R_0^2 = Q_{D.}$

La hipótesis nula de mayor interés es al que establece que no existen diferencias entre los niveles de los factores

$H_0 : \mu_1 = \cdots = \mu_k.$

Se trata de una hipótesis demostrable de rango $k-1$. Bajo $H_0$ solamente existe una media $\mu$ y su estimación es $\hat{\mu} = \overline{y}$. Entonces la suma de cuadrados residual es $R_1^2 = Q_T$ y además se verifica

$R_1^2 - R_0^2 = Q_E.$

Por tanto, como consecuencia del teorema se tiene que:

1. $Q_D / (n-k)$ es un estimador centrado de $\sigma^2$ y $Q_D / \sigma^2 \sim X_{n-k}^2$.

2. Si $H_0$ es cierta, $Q_E/(k-1)$ es también estimador centrado de $\sigma^2$ y

$$\frac{Q_T}{\sigma^2} \sim X^2_{n-1}, \quad \frac{Q_E}{\sigma^2} \sim X^2_{k-1}.$$

3. Si $H_0$ es cierta, los estadísticos $Q_E$ y $Q_D$ son estocásticamente independientes. Consecuencia inmediata es que, si $H_0$ es cierta, entonces el estadístico

$$F =$$

$$F = \frac{Q_E/(k-1)}{D_D/(n-k)} \sim F^{k-1}_{n-k}.$$

### 5.9. DISEÑO DE DOS FACTORES

Supóngase que las observaciones de una variable *Y* dependen de dos factores A, B, denominados factores fila y columna, con *a* y *b* niveles $A_1,\ldots,A_a$ y $B_1,\ldots,B_b$, y que disponemos de una observación para cada combinación de los niveles de los factores:

| | $B_1$ | $B_2$ | $\cdots$ | $B_b$ | |
|---|---|---|---|---|---|
| $A_1$ | $y_{11}$ | $y_{12}$ | $\cdots$ | $y_{1b}$ | $y_{1.}$ |
| $A_2$ | $y_{21}$ | $y_{22}$ | $\cdots$ | $y_{2b}$ | $y_{2.}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\ddots$ | $\vdots$ | $\vdots$ |
| $A_a$ | $y_{a1}$ | $y_{a2}$ | $\cdots$ | $y_{ab}$ | $y_{a.}$ |
| | $y_{.1}$ | $y_{.2}$ | $\cdots$ | $y_{.b}$ | $y_{..}$ |

Siendo

$$y_i = \frac{1}{b}\sum_{j=1}^{b} y_{ij,} \quad y.j = \frac{1}{a}\sum_{i=1}^{a} y_{ij}, \quad y.. = \overline{y} = \frac{1}{ab}\sum_{i=1}^{a}\sum_{j=1}^{b} y_{ij},$$

Las medias por filas por columnas y general. Supongamos que los datos se ajustan al modelo $y_{ij} = \mu + \alpha_i + \beta_j + e_{ij}$, bajo la condición de $\sum_{i=1}^{a} \alpha_i = \sum_{j=1}^{b} \beta_j = 0$,

donde $\mu$ es la media general, $\alpha_i$ es el efecto del nivel $\mathrm{A}_i$ del factor fila, $\beta_j$ es el efecto del nivel $\mathrm{B}_j$ del factor columna. El rango del diseño y los logs g.l. del residuo son

$$r = 1 + (a-1) + (b-1) = a + b - 1,\ n - r = ab - (a+b-1) = (a-1)(b-1)$$

Las estimaciones de los parámetros son

$$\hat{\mu} = \overline{y},\ \ \hat{\alpha}_i = y_{i,} - \overline{y},\ \ \ \hat{\beta}_j = y_{.j} - \overline{y},$$

Y la expresión de la desviación aleatoria es

$$\hat{e}_{ij} = y_{ij} - \hat{\mu} - \hat{\alpha}_i - \hat{\beta}_j = (y_{ij} - y_{i.} - y_{.j} + \overline{y}).$$

La suma de cuadrados residual del modelo es

$$R_0^2 = \sum_{i=1}^{a}\sum_{j=1}^{b}(y_{ij} - y_{i.} - y_{.j} + \overline{y})^2.$$

También consideramos las siguientes cantidades, donde SC significa "suma de cuadrados".

SC entre filas: $Q_A \ = b\sum_i (y_{i.} - \overline{y})^2$

SC entre columnas: $Q_B \ = a\sum_j (y_{.j} - \overline{y})^2$

SC residual: $Q_R \ = \sum_{i,j} (y_{ij} - y_{i.} - y_{.j} + \overline{y})^2$

SC total: $Q_T \ = \sum_{i,j} (y_{ij} - \overline{y})^2$

Se verifica la siguiente identidad:

$$Q_T = Q_A + Q_B + Q_R.$$

En el modelo de dos factores, las hipótesis de interés son:

$H_0^A:\ \alpha_1 = \cdots = \alpha_a = 0$ (no hay efecto fila)

$H_0^B:\ \beta_1 = \cdots = \beta_b = 0$ (no hay efecto columna)

Ambas hipótesis son demostrables. Supóngase $H_0^B$ cierta. Entonces el modelo se transforma en $y_{ij} = \mu + \alpha_i + e_{ij}$, es decir, actúa solamente un factor, y por tanto

$$R_1^2 = \sum_{i=1}^{a}\sum_{j=1}^{b}(y_{ij} - y_{i.})^2.$$

Ahora bien, desarrollando $(y_{ij} - y_{i.})^2 = ((y_{.j} - \overline{y}) + (y_{ij} - y_{i.} + y_{.j} + \overline{y}))^2$ resulta

$$R_1^2 = Q_B + Q_{R.}$$

Análogamente, si $H_0^A$ es cierta, obtendríamos $R_1^2 = Q_A + Q_{R.}$ se verifica:

1. $Q_R / (a-1)(b-1)$ es un estimador centrado de $\sigma^2$ y $Q_R / \sigma^2 \sim X^2_{(a-1)(b-1).}$
2. Si $H_0^A$ es cierta, $Q_A / (a-1)$ es también estimador centrado de $\sigma^2$, $Q_A / \sigma^2 \sim X^2_{(a-1)}$ y los estadísticos $Q_A$ y $Q_R$ son estocásticamente independientes.
3. Si $H_0^B$ es cierta, $Q_B / (b-1)$ es también estimador centrado de $\sigma^2$, $Q_B / \sigma^2 \sim X^2_{(b-1)}$ y los estadísticos $Q_B$ y $Q_R$ son estocásticamente independientes.

Por lo tanto se tiene que para decir $H_0^A$ se utilizará el estadístico

$$F_A = \frac{Q_A}{Q_R}\frac{(a-1)(b-1)}{(a-1)} \sim F^{a-1}_{(a-1)(b-1)},$$

Y para decidir $H_0^B$ utilizaremos

$$F_B = \frac{Q_B}{Q_R}\frac{(a-1)(b-1)}{(b-1)} \sim F^{b-1}_{(a-1)(b-1)}$$

### 5.10. Diseño de factores con interacción

Supongamos que las observaciones de una variable *Y* dependen de dos factores A,B, denominados factores fila y columna, con *a* y *b* niveles $A_1,\ldots,A_a$ y $B_1,\ldots,B_b$, y

que disponemos de *c* observaciones (réplicas) para cada combinación de los niveles de los factores:

$$\begin{array}{cccccc} & \mathrm{B}_1 & \mathrm{B}_2 & \cdots & \mathrm{B}_b & \\ \mathrm{A}_1 & y_{111},\cdots,y_{11c} & y_{121},\cdots,y_{12c} & \cdots & y_{1b1},\cdots,y_{1bc} & y_{1\cdot\cdot} \\ \mathrm{A}_2 & y_{211},\cdots,y_{21c} & y_{221},\cdots,y_{22c} & \cdots & y_{2b1},\cdots,y_{2bc} & y_{2\cdot\cdot} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathrm{A}_a & y_{a11},\cdots,y_{a1c} & y_{a22},\cdots,y_{a2c} & \cdots & y_{ab1}\cdots,y_{abc} & y_{a\cdot\cdot} \\ & y_{\cdot 1\cdot} & y_{\cdot 2\cdot} & \cdots & y_{\cdot b\cdot} & y_{\cdots} \end{array}$$

Siendo

$$y_{i\cdot\cdot}=\frac{1}{bc}\sum_{j,h=1}^{b.c} y_{ijh}, \qquad y_{\cdot j\cdot}=\frac{1}{ac}\sum_{i,h=1}^{a,c} y_{ijh},$$

$$y_{ij\cdot}=\frac{1}{c}\sum_{h=1}^{c} y_{ijh}, \qquad \overline{y}=y_{\cdots}=\frac{1}{abc}\sum_{i,j,h=1}^{a,b,c} y_{ij\cdot}$$

El modelo lineal del diseño de dos factores con interacción es

$$y_{ijh}=\mu+\alpha_i+\beta_j+\gamma_{ij}+e_{ijh},$$

$$i=1,\ldots,a;\ j=1,\ldots,b;\ h=1,\ldots,c,$$

Siendo $\mu$ la media general, $\alpha_i$ el efecto del nivel $A_i$ del factor fila, $\beta_j$ el efecto del nivel $\mathrm{B}_j$ del factor columna, $\gamma_{ij}$ la interacción entre los niveles $\mathrm{A}_i,\mathrm{B}_j$. El parámetro $\gamma_{ij}$ mide la desviación del modelo aditivo $E\left(y_{ijh}\right)=\mu+\alpha_i+\beta_j$ y solamente es posible estimar si hay $c>1$ réplicas. Se suponen las restricciones

$$\sum_{i=1}^{a}\alpha_i=\sum_{j=1}^{b}\beta_j=\sum_{i=1}^{a}\gamma_{ij}=\sum_{j=1}^{b}\gamma_{ij}=0.$$

Así el número de parámetros independientes del modelo es

$$1+(a-1)+(b-1)+(a-1)(b-1)=ab$$

Y los g. 1. del residuo son $abc-ab=ab(c-1)$.

Las estimaciones de los parámetros son

$$\hat{\mu} = \overline{y}, \quad \hat{\alpha}_i = y_{i\cdot\cdot} - \overline{y}, \quad \hat{\beta}_j = y_{\cdot j\cdot} - \overline{y}, \quad \hat{\gamma}_{ij} = y_{ij\cdot} - y_{i\cdot\cdot} - y_{\cdot j\cdot} + \overline{y},$$

Y la expresión de la desviación aleatoria es

$$\hat{e}_{ijh} = y_{ijh} - \hat{\mu} - \hat{\alpha}_i - \hat{\beta}_j - \hat{\gamma}_{ij} = (y_{ij} - \overline{y}).$$

La suma de los cuadrados residual del modelo es

$$R_0^2 = \sum_{i,j,h=1}^{a,b,c} (y_{ijh} - y_{i\cdot\cdot})^2.$$

También debemos de considerar las siguientes cantidades, donde SC significa "suma de cuadrados":

SC entre filas: $Q_A = bc\sum_i (y_{i\cdot\cdot} - \overline{y})^2$

SC entre columnas: $Q_B = ac\sum_j (y_{\cdot j\cdot} - \overline{y})^2$

SC de la interacción: $Q_{AB} = c\sum_{i,j} (y_{ij\cdot} - y_{i\cdot\cdot} - y_{\cdot j\cdot} + \overline{y})^2$

SC residual: $Q_R = \sum_{i,jh} (y_{ijh} - y_{i\cdot\cdot})^2$

SC total: $Q_T = \sum_{i\cdot j} (y_{ijh} - \overline{y})^2$

Se verifica la siguiente identidad

$$Q_T = Q_A + Q_B + Q_{AB} + Q_R.$$

Las hipótesis de interés son:

$H_0^A$ : $\alpha_1 = \cdots = \alpha_a = 0$ (no hay efecto fila)

$H_0^B$ : $\beta_1 = \cdots = \beta_b = 0$ (no hay efecto columna)

$H_0^{AB}$ : $\gamma_{11} = \cdots = \gamma_{ab} = 0$ (no hay interacción)

Como en los casos anteriores, podemos ver que la aceptación o el rechazo de cada hipótesis se decide mediante el test F:

$$F_A = \frac{Q_A}{Q_R}\frac{ab(c-1)}{a-1} \sim F^{a-1}_{ab(c-1)}$$

$$F_B = \frac{Q_B}{Q_R}\frac{ab(c-1)}{b-1} \sim F^{b-1}_{ab(c-1)}$$

$$F_{AB} = \frac{Q_{AB}}{Q_R}\frac{ab(c-1)}{(a-1)(b-1)} \sim F^{(a-1)(b-1)}_{ab(c-1)}$$

## 5.11. Diseños multifactoriales

Los diseños de dos factores se generalizan a un número mayor de factores. Cada factor representa una causa de variabilidad que actúa sobre la variable observable. Si por ejemplo, hay 3 factores A, B, C, las observaciones son $y_{ijkh,}$ donde *i* indica el nivel *i*-ésimo de A, *j* indica el nivel *j*-ésimo de B, *k* indica el nivel *k*-ésimo de C, y *h* indica la réplica *h* para la combinación *ijk* de los tres factores, que pueden interactuar. Un modelo típico es

$$y_{ijkh} = \mu + \alpha_i^A + \alpha_j^B + \alpha_k^C + \alpha_{ij}^{AB} + \alpha_{ik}^{AC} + \alpha_{jk}^{BC} + \alpha_{ijk}^{ABC} + e_{ijkh,}$$

Siendo

$\mu$ = media general,

$\alpha_i^A, \alpha_j^B, \alpha_k^C$ = efectos principales de A,B,C,

$\alpha_{ij}^{AB}, \alpha_{ik}^{AC}, \alpha_{jk}^{BC}$ = interacciones entre A y B, A y C, B y C,

$\alpha_{ijk}^{ABC}$ = interacción entre A, B y C,

$e_{ijkh}$ = desviación aleatoria $N(0, \sigma^2)$.

Son hipótesis de interés: $H_0^A : \alpha_i^A = 0$ (el efecto principal de A no es significativo), $H_0^{AB} : \alpha_i^{AB} = 0$ (la interacción entre A y B no es significativa), etc. Los contrastes para aceptar o no estas hipótesis se obtienen descomponiendo la variabilidad total en sumas de cuadrados:

$$\sum_{i,j,k,h} (y_{ikjh} - \overline{y})^2 = A + B + C + AB + AC + BC + ABC + R,$$

Donde R es el residuo. Si lo factores tienen *a, b, c* niveles, respectivamente, y hay *d* réplicas para cada combinación de los niveles, entonces *A* tiene $(a-1)$ g.i., *AB* tiene $(a-1)(b-1)$ g.i. si interpretamos las réplicas como un factor *D,* el residuo es

$$R = D + AD + BD + CD + ABD + ACD + BCD + ABCD$$

Con

$$q = (d-1) + (a-1)(d-1) + \cdots + (a-1)(b-1)(c-1)(d-1) = abc(d-1)$$

g.i. Entonces calcularemos los coeficientes F

$$F = \frac{A/(a-1)}{R/q}, \qquad F = \frac{AB/(a-1)(b-1)}{R/q},$$

Que sirven para aceptar o rechazar $H_0^A$ y $H_0^{AB}$ , respectivamente.

En determinadas situaciones experimentales puede suceder que algunos factores no interactúen. Entonces las sumas de cuadrados correspondientes se suman al residuo. Por ejemplo, si C no interactúa con A,B, el modelo es

$$\sum_{i,j,k,h} (y_{ikjh} - \overline{y})^2 = A + B + C + AB + R',$$

Donde $R' = AC + BC + ABC + R$ es el nuevo residuo con g.l.

$$q' = (a-1)(c-1) + (b-1)(c-1) + (a-1)(b-1)(c-1) + q.$$

Los cocientes F para las hipótesis anteriores son ahora

$$F = \frac{A/(a-1)}{R'/q'}, \quad F = \frac{AB/(a-1)(b-1)}{R'/q'}.$$

## 6. Aplicaciones

La Morfometría es la descripción cuantitativa, análisis e interpretación de las formas, así como la variación de las formas biológicas (Rohlf, 1990). En este sentido, la morfometría geométrica ha encontrado una amplia gama de aplicaciones en diferentes ciencias como lo es biología, paleontología y paleobiología, antropología, entre otras, campos donde las mediciones y formas de las estructuras son fundamentales.

A continuación se mencionan las principales aplicaciones.

### 6.1. Aplicación I – Biología

Dentro del amplio campo de la biología las aplicaciones de la morfometría geométrica se basan principalmente en la búsqueda de nuevos taxones y en la diferenciación de especies y/o subespecies con base en características específicas, que a veces no son perceptibles a simple vista.

(Douglas, *et al*., 2001) a través de un estudio de morfometría geométrica estudiaron la distribución de dos especies incluidas dentro del género *Gyla* (*G. robusta* y *G. cypha*) en la parte alta de la cuenca del Río Colorado. La metodología consistió principalmente en el análisis de 215 imágenes de *G. robusta* y 148 imágenes de *G. cypha,* obteniendo que las dos especies presentan la misma distribución en el área. El análisis morfométrico se llevó a cabo principalmente por medio de coordenadas utilizando puntos (*landmarks*), donde el tamaño central del organismo fue utilizado como medida del tamaño del cuerpo. Las diferencias de las formas fueron evaluadas entre las poblaciones utilizando ANOVA y análisis de diferencia canónica. El resultado del análisis morfométrico indica que este es estadísticamente similar a los que se basan en derivados de distancias.

Por otra parte (García & Sánchez-González, 2013) estudian la morfometría geométrica craneal de especies de roedores arborícolas neotropicales (Rodentia: Cricetidae: *Rhipidomys*) en el área de Sierra Aroa, Yaracuy, Venezuela. La

metodología consistió en seleccionar puntos anatómicos de referencia en los cráneos y mandíbulas por medio de un software. Posteriormente se hicieron medidas lineales. El estudio demostró la existencia de tres especies y tres subespecies, habiendo algunas diferencias entre estas, así como la posible existencia de un nuevo taxón aún no descrito dentro de la especie *Rhipidomis venustus.*

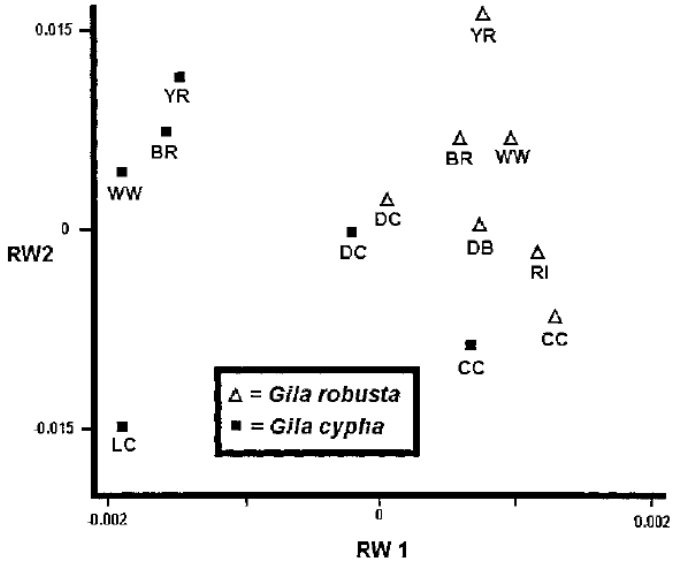


Figura 1.- Variación de la forma en 13 poblaciones de *G. cypha* y *G. robusta*, utilizando la media y configuración tangente. Tomado de (Douglas, *et al*., 2001)

(Calle, *et al*., 2008) Realizan un estudio de discriminación por medio de morfometría geométrica en once especies hembra del mosquito *Anopheles* (*Nyssorhynchus*) presente en cinco áreas representativas de Colombia. La metodología consistió en capturar los ejemplares cuando estos aterrizaban sobre el cebo humano protegido o las hembras recién emergidas procedentes de series o isofamilias. Se realizó la configuración geométrica seleccionando 12 puntos sobre imágenes en las alas, esto en los límites de las manchas basales y subcostales, así como en la intersección de las venas y en el borde distal del ala, los puntos anatómicos fueron convertidos en puntos en el plano bidimensional. La matriz de coordenadas que representa las configuraciones geométricas de las alas se procesó usando el análisis generalizado de Procrustes el cual se basa en tres pasos iterativos que consisten en:

1. Las configuraciones se ajustan a un tamaño único.
2. Las configuraciones ajustadas se trasladan una sobre otra de tal manera que coincidan sus centros de gravedad (centroides).
3. Se rotan hasta minimizar las distancias entre los puntos correspondientes, utilizando el criterio matemático de mínimos cuadrados.

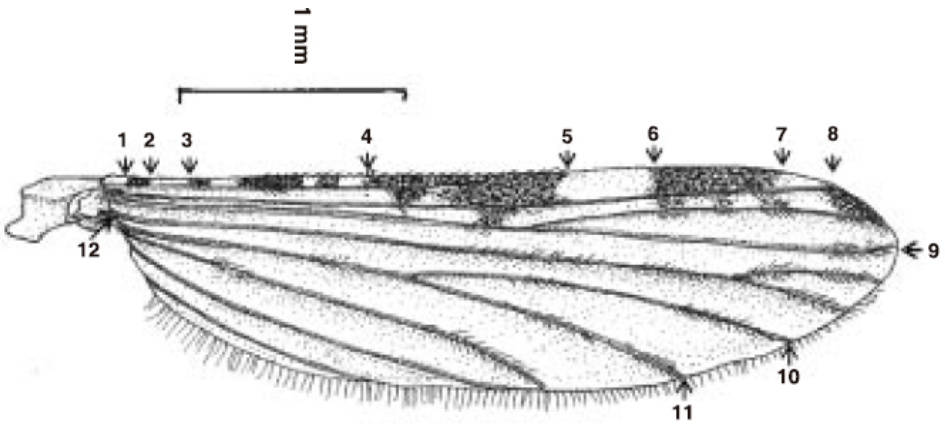


Figura 2.- Detalle de los puntos anatómicos considerados para realizar el estudio de diferenciación de especies hembras del mosquito *Anopheles.* Tomado de Calle, *et al.,* (2008).

Posteriormente se obtienen las variables que incorporan las desviaciones de cada configuración respecto a la de referencia es decir contiene toda la información de la conformación del ala. El resultado obtenido da a conocer que la subdivisión del subgénero *Nyssorhynchus* en secciones no es correlacionable con la forma del ala, llevando a la correcta identificación de tres especies, las cuales son difíciles de identificar en estadio adulto. (Calle, *et al*., 2008).

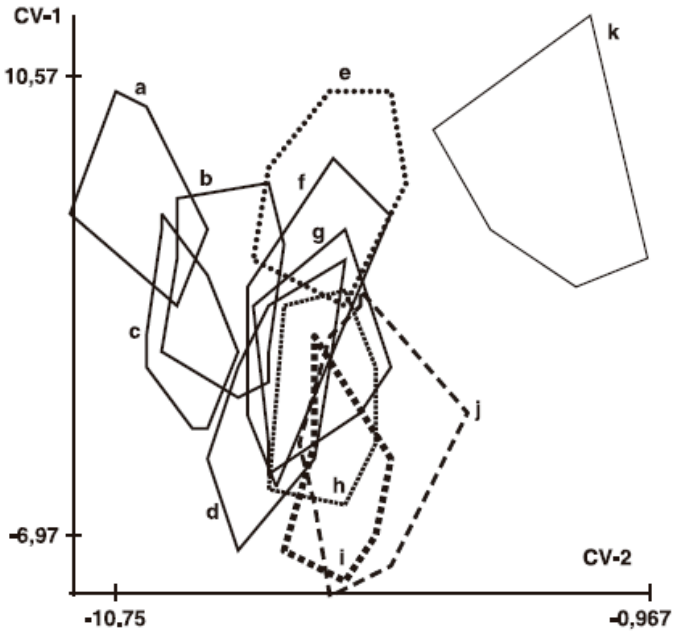


Figura 3.- Detalle de la distribución de especies del género *Anopheles*, derivadas del análisis morfométrico. (a) *A. darlingi*, (b) *A. braziliensis*, (c) *A. triannulatus*, (d) *A. marajoara*, (e) *A. nunetzovari*, (f) *A. albimanus*, (g) *A. aquasalis*, (h) *A. strodei*, (i) *A. benarrochi*, (j) *A. oswaldoi* y (k) *A. rangeli*. Tomado de Calle, *et al.*, (2008).

### 6.2. Aplicación II – Paleontología y Paleobiología

Una de las principales limitaciones que se encuentran en el estudio de los organismos fósiles es la carencia de tejidos blandos, por lo tanto, la mayoría de los estudios paleontológicos tienen que ser basados en el entendimiento de las formas o estructuras duras.

La utilidad de la morfometría geométrica comprende una amplia gama de aplicaciones, entre las que se mencionan investigación sobre grupos para los cuales su taxonomía se basa principalmente en la forma de estos (Baltanás & Danielopol, 2011; Lawing & Polly, 2010; Webster, 2010).

(Baltanás & Danielopol, 2011) Realizan una investigación en la búsqueda de ostrácodos (Arthropoda – Crustacea), destacando que el reconocimiento de las especies de este grupo se realiza principalmente con base en las características

morfológicas en combinación con otras fuentes de información, como lo es análisis molecular, morfología tradicional con filogenia, ecología y biogeografía, sin embargo, algunos de estos métodos se encuentran limitados, puesto que los fósiles no conservan tejidos blandos.

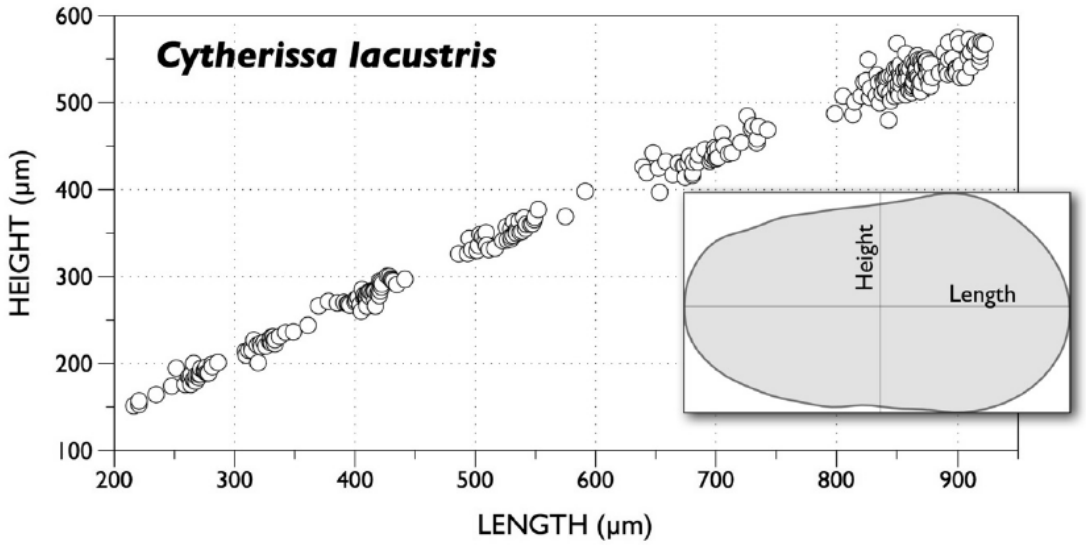


Figura 4.- Detalle de la longitud contra la altura en una muestra de especies *Cytherissa lacustris.* Tomado de Baltanás & Danielopol (2011).

### 6.3. Aplicación III – Antropología.

La aplicación más frecuente de la morfometría geométrica dentro del campo de la antropología forense está relacionada con la determinación de la afinidad de poblaciones o ancestros, evaluación de muerte así como la determinación del sexo.

Según (Rosing, *et al*., 2007) la existencia de dimorfismo sexual del esqueleto y su evaluación representa una presunción en la cual los científicos basan los métodos actuales para la determinación del sexo en esqueletos humanos. Se puede entender que el dimorfismo sexual es resultado en parte de las diferencias reproductivas y en parte a la fuerte presión a la que los humanos han sido sometidos por la evolución.

Un acercamiento cuantitativo al estudio del dimorfismo sexual está representado por una clasificación convencional de técnicas estadísticas univariadas y multivariadas. Estas aproximaciones utilizan ángulos, distancias o distancias de radios. La mayoría de los científicos que trabajan en antropología utilizan la metodología de la morfometría geométrica, ya que esta incluye métodos basados principalmente en coordenadas 3D de *landmarks* homologas que describen al objeto de estudio. De esta manera las coordenadas representan la información geométrica completa relacionada al objeto de estudio. Cuando se realiza un análisis de las formas de los objetos biológicos, la morfometría geométrica permite una diferenciación de la variabilidad de acuerdo al tamaño y la forma. La cuantificación de la forma y tamaño usando procedimientos estadísticos de la morfometría geométrica específica y agrupa resultados más precisos que aquellos que han sido obtenidos con otros métodos.

La morfometría geométrica representa una nueva técnica de aproximación en la evaluación de la variabilidad, no sólo en las disciplinas biomédicas, sino también en otras áreas como bioarqueología, evolución y ecología.

(Bigonio, *et al*., 2010) Realizan un estudio sobre el dimorfismo sexual cráneo-encefálico en personas de sexo conocido en Europa, con el propósito de conocer las regiones del cráneo donde el dimorfismo sexual es más pronunciado así como para investigar la efectividad del método para determinar el sexo a partir de la forma del cráneo. La muestra consistió en 139 cráneos (73 hombres y 66 mujeres), los cuales vivieron durante la primera mitad del siglo XX en Bohemia. Los resultados de este trabajo demuestran que es mejor analizar partes específicas del cráneo en vez del cráneo completo. Las diferencias sexuales significativas (la significancia se determinó usando análisis multivariados de varianza) fueron notados en la curvatura de la bóveda craneal, parte superior del rostro, región de la nariz, orbitales y placas; no existiendo diferencias en la forma del cráneo en total o en la base de las regiones del neurocráneo.

(Gómez-Valdés, *et al*., 2007) Analizaron 6 cráneos de la época prehispánica del área de Mesoamérica, para cada cráneo colectaron un total de 78 puntos craneométricos, para la posición frontal diseñaron un protocolo con una colección de 28 puntos craneométricos que definen los componentes anatómicos: orbito-nasal y zigo-maxilar. En la posición lateral se emplearon 29 puntos craneométricos para definir dos componentes de la región facial: orbito-nasal y zigo-maxilar, glenoidea y mastoido-occipital, para la posición basal el protocolo incluyo 21 puntos craneométricos que definen la región zigo-maxilar, glenoidea, esfero-basilar y mastoido-occipital. Los puntos anatómicos son convertidos en puntos en el plano bidimensional. La matriz de coordenadas que representaba las configuraciones geométricas de los puntos craneométricos se procesó usando el análisis generalizado de Procrustes. Los resultados permitieron comparar y enriquecer el conocimiento que se tiene sobre la deformación craneal en la época prehispánica.

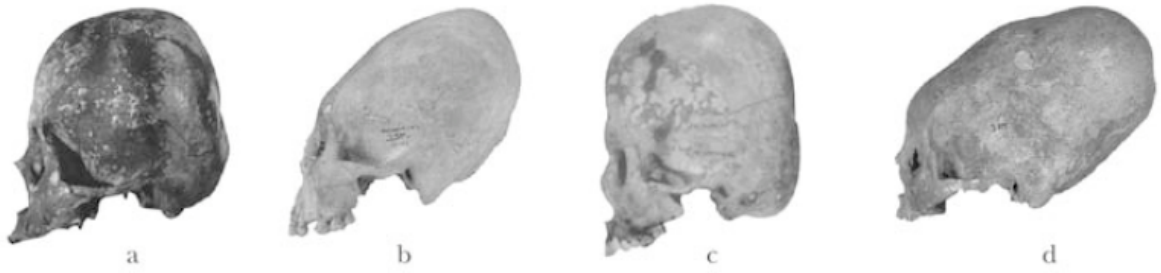


Figura 5.- Detalle de los cráneos utilizados por Gómez-Valdez, *et al.,* (2007) para el estudio de morfometría en la época prehispánica de México.

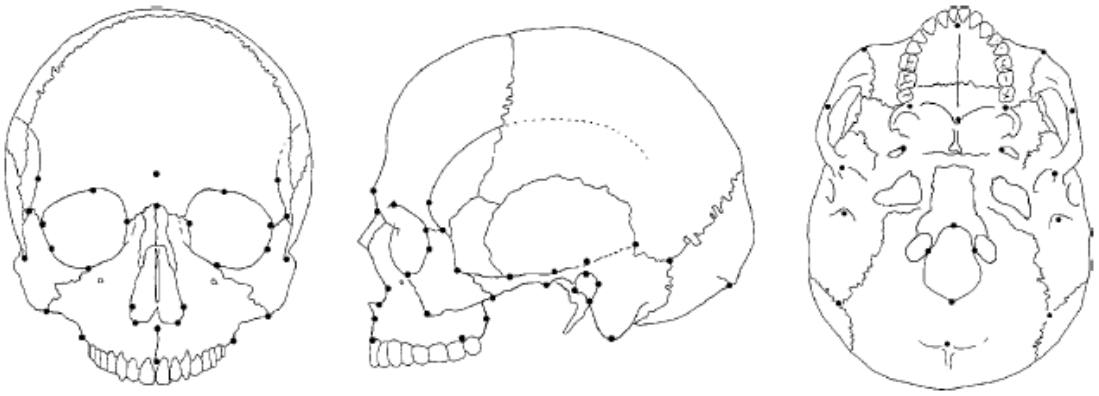

Figura 6.- Detalle de los puntos homólogos colocados para analizar los cráneos. Tomado de Gómez-Valdez, *et al.,* (2007).

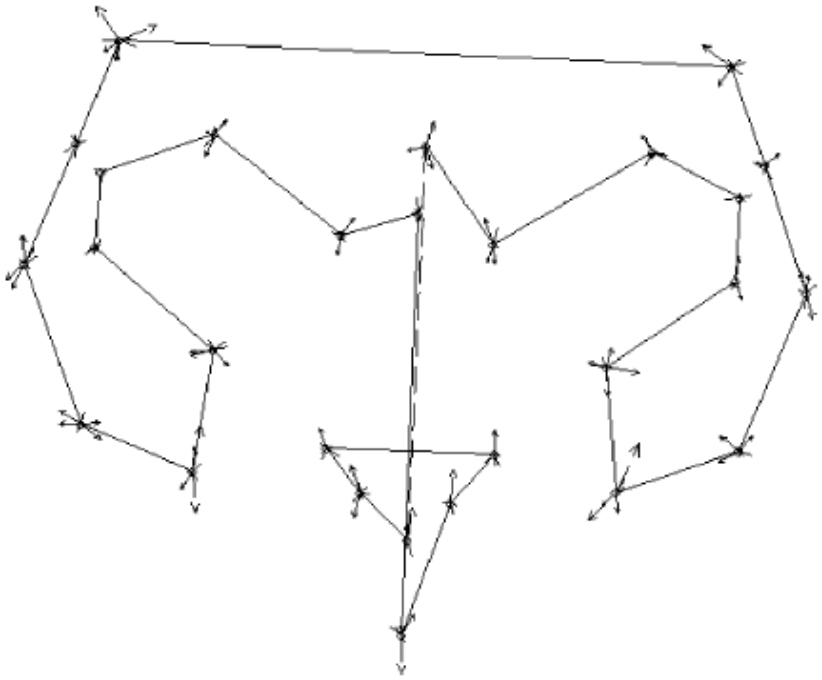

Figura 7.- Variación de los puntos homólogos posicionados en los cráneos. Tomado de Gómez-Valdez, *et al.,* (2007).

## 7. Conclusiones

La necesidad del hombre por entender el mundo natural exige el avance de la ciencia, para de esta manera poder realizar clasificaciones naturales cada vez más precisas. La Morfometría Geométrica, representa una ciencia con un gran potencial, no solamente por los avances logrados sino por su unión con la estadística y la biología cuantitativa.

La Morfometría Geométrica tiene una gran cantidad de campos para su desarrollo, principalmente aquellos relacionados con la forma, por ejemplo el evaluar el crecimiento de ciertas estructuras con fines clínicos y de la salud, no restringiéndose de esta manera solamente al entendimiento estricto de las formas antropológicas – biológicas. Todas las estructuras que se encuentren en desarrollo pueden ser analizadas mediante morfometría geométrica, incluyéndose dimorfismo sexual, afinidad de linajes y especies, evaluación forense, entre otras. El hecho de que la Morfometría Geométrica trabaje con datos naturales y con aplicaciones de la estadística multivariada la convierte en una poderosa herramienta para la evaluación de manera objetiva de las variaciones de la forma.

En este trabajo ha quedado claro la relación estrecha que existe entre las matemáticas (estadística) y la biología, unidas por la morfometría geométrica. Aunque se considere una ciencia relativamente reciente ha alcanzado un nivel de madurez como para formar un morfológo.

Un morfólogo por lo tanto debe de poseer una formación matemática o biológica, pero con sólidos conocimientos de la estadística y la variación de la forma.

## 8. Bibliografía

Adams, D., Rohlf, J. & Slice, D., 2004. Geometric Morphometrics: Ten years of progress following the "revolution". *Italian Journal of Zoology,* 71(1).

Álvarez, C.R., 2007. Estadística Aplicada a las Ciencias de la Salud. Editorial Díaz de Santo, España.

Anzelmo, M., Sardi, M. L., Barbeito-Andrés, J. & Pucciarelli, H. M., 2012. Alometrías ontogénicas y dimorfismo sexual facial en dos poblaciones humanas modernas. *Revista Argentina de Antropología Biológica,* 14(1), pp. 89-100.

Baltanás, A. & Danielopol, D., 2011. Geometric Morphometrics and its use in ostracod research: a short guide. *Joannea Geol. Palaont.,* Volumen 11, pp. 235-272.

Berns, A., 2014. *A geometric morphometric analysis of wing shape variation in monarch butterflies Danaus plexippus.* s.l.:Tesis Profesional .

Bigoni, L., Velemínská, J. & Bruzek, J., 2010. Three-dimensional geometric morphometric analysis of cranio-facial sexual dimorphism in a Central European sample of known sex. *HOMO - Journal Comparative Human Biology,* Volumen 61, pp. 16-32.

Bigonio, L., Veleminska, J. & Bruzek, J., 2010. Three-dimensional geometric morphometric analysis of cranio-facial sexual dimorphism in a central European sample of known sex. *Homo,* Volumen 61, pp. 16-32.

Brookstein, F., 1982. Foundations of morphometrics. *Annual Review of Ecology and Systematics,* Volumen 13, pp. 451-470.

Brookstein, F., 1986. Size and shape spaces for landmark data in two dimensions. *Statistical Science,* Volumen 1, pp. 181-242.

Brookstein, F., 1991. *Morphometric tools for landmark data: Geometry and Biology.* s.l.:Cambridge Press University.

Brookstein, F., 1996. Landmark methods for forms without landmarks: morphometrics of group differences in outline shape. *Medical Image Analysis,* 1(3), pp. 225-243.

Calle, D. A., Quiñones, M. L., Erazo, H. F. & Jaramillo, N., 2008. Discriminación por morfometría geométrica de once especies de Anopheles (Nyssorhynchus) presentes en Colombia. *Biomédica,* Volumen 28, pp. 371-385.

Clifford, B., Taylor, R., 2008. Bioestadística. Editorial Pearson / Educación. México.

Cuadras, C. M., 2014. *Nuevos métodos de análisis multivariante.* Barcelona: CMC Editions.

Douglas, M. E., Douglas, M. R., Lynch, J. M. & McElroy, D. M., 2001. Use of Geometric Morphometrics to Differentiate Gila (Cyprinidae) within the Upper Colorado River Basin. *Copeia,* Issue 2, pp. 389-400.

García, F. J. & Sánchez-González, E., 2013. Morfometría Geométrica craneal en tres especies de roedores arborícolas neotropicales (Rodentia: Cricetidae: Rhipidomys) en Venezuela. *Therya,* 4(1), pp. 157-178.

Gómez-Valdés, J. A., Bautista Martínez, J. & Romano Pacheco, A., 2007. Morfometría Geométrica aplicada al estudio de la deformación cefálica intencional. *Estudios de Antropología Biológica,* Volumen XIII, pp. 117-134.

Hurley, J. & Cattell, R., 1962. Producing direct rotation to test a hypothesized factor structure. *Behavioral Science,* 7(2), pp. 258-262.

Klingenberg, C. & Monteiro, L., 2005. Distances and directions in multidimensional shape spaces: implications for morphometric applications. *Systematic Biology,* Volumen 54, pp. 678-688.

Lawing, A. M. & Polly, P. D., 2010. Geometric morphometrics: recent applications to the study of evolution and development. *Journal of Zoology,* Volumen 280, pp. 1-7.

López, G. A., 2015. Morfometria geométrica: el estudio de la forma y su aplicación en biología. *Temas de Ciencia y Tecnología*, 19(55), pp. 53-59.

Rohlf, F., 1990. Morphometrics. *Annual Review of Ecology, Evolution, and Systematics,* Volumen 21, pp. 299-316.

Rohlf, F. & Marcus, L., 1993. A revolution in Morphometrics. *Trends in Ecology and Evolution,* 8(4).

Rosing, F. W. y otros, 2007. Recommendations for the forensic diagnosis of sex and age from skeletons. *Homo,* Volumen 58, pp. 75-89.

Torcida, S. & Pérez, I., 2012. Análisis de Procrustes y el estudio de la variación morfológica. *Revista Argentina de Antropología biológica,* 14(1), pp. 131-141.

Toro Ibacache, M. V., Soto Manríquez, G. & Suazo Galdamas, I., 2010. Morfometría Geométrica y el Estudio de las formas biológicas: de la Morfología Descriptiva a la Morfología Cuantitativa. *Int. J. Morphol.,* 28(4), pp. 977-990.

Wayne, W.D., 2002. Bioestadística: Base para el Análisis de las Ciencias de la Salud. UTEHA, Noriega Editores, México.

Webster, M., 2010. A practical introduction to landmark-based Geometric morphometrics. *Quantitative methods in Paleobiology,* Volumen 16, pp. 163-188.

Zelditch, M., Swiderski, D. & Sheets, H., 2004. *Geometric morphometrics for biologists: A primer.* San Diego, California, USA: Elsevier Academic Press.

Zuliani, R. P., 2012. *Evaluación comparativa de las Técnicas Multivariadas Análisis factorial Múltiple y Análisis de Procrustes Generalizado para el tratamiento de datos de tres Modos.* Córdoba.: Universidad Nacional de Córdoba. Tesis Profesional.