



ANÁLISIS MULTIVARIANTE Y NO PARAMÉTRICO:

Casos de estudio aplicados a
experimentos agropecuarios



eBook



ANÁLISIS MULTIVARIANTE Y NO PARAMÉTRICO:

Casos de estudio aplicados a
experimentos agropecuarios

Julio Gabriel Ortega

Gema Burgos López

Alfredo Valverde Lucio

Agustín H. Álvarez Plúa

Autores Investigadores

ANÁLISIS MULTIVARIANTE Y NO PARAMÉTRICO:

Casos de estudio aplicados a experimentos agropecuarios

AUTORES

INVESTIGADORES

Julio Gabriel Ortega

Doctor dentro del Programa de Producción Agraria y
Aplicaciones Biotecnológicas;
Docente Investigador de la Facultad de
Ciencias Naturales y de la Agricultura,
Universidad Estatal del Sur de Manabí;
Jipijapa, Ecuador;

✉ julio.gabriel@unesum.edu.ec

ID <https://orcid.org/0000-0001-9776-9235>

Gema Burgos López

Magíster en Agropecuaria e
Ingeniera Agropecuaria de la
Universidad Estatal del Sur de Manabí;
Jipijapa, Ecuador;
Consultora independiente;

✉ burgos-gema4099@unesum.edu.ec

ID <https://orcid.org/0000-0002-0025-3679>

Alfredo Valverde Lucio

Doctor en Biociencias y Ciencias Alimentarias,
Docente Investigador de la Facultad de
Ciencia Naturales y de la Agricultura,
Universidad Estatal del Sur de Manabí;
Jipijapa, Ecuador;

 yhonny.valverde@unesum.edu.ec

 <https://orcid.org/0000-0002-9792-9400>

Agustín H. Álvarez Plúa

Magíster en Agroecología y Agricultura Sostenible;
Docente Investigador de la Facultad de
Ciencias Naturales y de la Agricultura,
Universidad Estatal del Sur de Manabí;
Jipijapa, Ecuador;

 agustin.alvarez@unesum.edu.ec

 <https://orcid.org/0000-0003-0176-1314>

ANÁLISIS MULTIVARIANTE Y NO PARAMÉTRICO:

Casos de estudio aplicados a
experimentos agropecuarios

REVISORES

ACADÉMICOS

Aldo José Loqui Sánchez

Magíster en Riego y Drenaje;
Ingeniero Agrónomo;
Facultad de Medicina Veterinaria y Zootecnia,
Universidad de Guayaquil;
Guayaquil, Ecuador;

✉ aldo_loqui@hotmail.com

ID <https://orcid.org/0000-0001-8953-5105>

Ricardo Paúl González Dávila

Ingeniero Agrícola de la Universidad Nacional de Loja UNL;
Maestro en Ciencias Sistemas Agropecuarios y Medio Ambiente de la
Universidad Autónoma de Tamaulipas UAT México;
Docente Universidad Laica Eloy Alfaro de Manabí,
El Carmen, Ecuador;

✉ ricardo.gonzalez@uleam.edu.ec

ID <https://orcid.org/0000-0001-7808-7642>

CATALOGACIÓN BIBLIOGRÁFICA

AUTORES: Julio Gabriel Ortega
Gema Burgos López
Alfredo Valverde Lucio
Agustín H. Álvarez Plúa

Título: Análisis multivariante y no paramétrico: Casos de estudio aplicados a experimentos agropecuarios

Descriptores: Ciencias Agrarias; Producción agrícola; Investigación agrícola; Tipos de café

Código UNESCO: 3103 Agronomía

Clasificación Decimal Dewey/Cutter: 633/P773

Área: Ciencias Agrarias

Edición: 1era

ISBN: 978-9942-654-70-0

Editorial: Mawil Publicaciones de Ecuador, 2025

Ciudad, País: Quito, Ecuador

Formato: 148 x 210 mm.

Páginas: 106

DOI: <https://doi.org/10.26820/978-9942-654-70-0>

URL: <https://mawil.us/repositorio/index.php/academico/catalog/book/176>

Texto para docentes y estudiantes universitarios

El proyecto didáctico: **Análisis multivariante y no paramétrico: Casos de estudio aplicados a experimentos agropecuario**, es una obra colectiva escrita por varios autores y publicada por SaberEC; publicación revisada bajo la modalidad de pares académicos y por el equipo profesional de la editorial siguiendo los lineamientos y estructuras establecidos por el departamento de publicaciones de SaberEC en la ciudad de Quito, Ecuador.

© Reservados todos los derechos. La reproducción parcial o total queda estrictamente prohibida, sin la autorización expresa de los autores, bajo sanciones establecidas en las leyes, por cualquier medio o procedimiento.



Usted es libre de:
Compartir — copiar y redistribuir el material en cualquier medio o formato.
Adaptar — remezclar, transformar y construir a partir del material para cualquier propósito, incluso comercialmente.

Director Académico: PhD. José María Lalama Aguirre

Dirección Central MAWIL: Office 18 Center Avenue Caldwell; New Jersey # 07006

Gerencia Editorial MAWIL-Ecuador: Mg. Vanessa Pamela Quishpe Morocho

Editor de Arte y Diseño: Leslie Letizia Plúa Proaño

Corrector de Estilo: Lic. Marcelo Acuña Cifuentes

ANÁLISIS MULTIVARIANTE Y NO PARAMÉTRICO:

Casos de estudio aplicados a
experimentos agropecuarios

Índice

Contenidos

Prólogo	12
Introducción	16

Capítulo I.

Técnicas de análisis multivariante	18
1.1 Introducción	19
1.2. Análisis factorial	20
1.2.1. Análisis de componentes principales (ACP)	20
1.2.1.1. Obtención de los componentes principales (García y Gómez, 2008)	20
1.2.2. Análisis factorial de correspondencias (AFC)	50
1.2.3. Análisis de conglomerados (clúster)	75

Capítulo II.

Nociones sobre estadística no paramétrica	81
2.1. Introducción	82
2.2. Clasificación de las principales pruebas no paramétricas	82
2.2.1. Problemas agropecuarios	86
2.2.1.1. Caso de una muestra	86
2.2.1.1.1. Prueba binomial	86
2.2.1.2. Caso de dos muestras independientes	87
2.2.1.2.1. Prueba de U de Mann-Whitney	87
2.2.1.3. Caso de dos muestras relacionadas	90
2.2.1.3.1. Prueba de Wilcoxon	90
2.2.1.4. Caso de k muestras independientes	91
2.2.1.4.1. Prueba de Kruskal-Wallis	91
2.2.1.5. Medidas de Asociación	94
2.2.1.5.1. Correlaciones no paramétricas (Spearman)	94
2.2.1.6. Aplicación de la distribución de Poisson	96
2.2.1.7. Aplicación de la distribución de chi-cuadrado	98

ANÁLISIS MULTIVARIANTE Y NO PARAMÉTRICO:

Casos de estudio aplicados a
experimentos agropecuarios

Índice

Figuras

Figura 1. Sedimentación para cinco variables cuantitativas agromorfológicas evaluadas en 30 híbridos de tomate (López, 2014; López <i>et al.</i> , 2015).	31
Figura 2. Representación tridimensional de los componentes principales.	35
Figura 3. Sedimentación para 17 variables cuantitativas del cerdo criollo.	38
Figura 4. Asociación entre estados de cada una de las variables cualitativas (Fuente: López 2014; López <i>et al.</i> 2015). Tipo de crecimiento de las plantas (Ind = Indeterminadas. Det = Determinadas); Forma predominante del fruto (Alre = Redondo alargado. Red = Redondo. Lach = Ligeramente achatado); Color exterior del fruto maduro (Roj = Rojo. Nar = Naranja. Ros = Rosada); color carne del pericarpio (Peroj = Rojo. Penar = Naranja. Peros = rosado); Corte transversal del fruto (Cored = Redondo, Corang = Angular); Número de lóculos (Lotres = tres lóculos. Locuatro = cuatro lóculos y Locinco = cinco lóculos) y los números son los 30 híbridos de tomate.	62

ANÁLISIS MULTIVARIANTE Y NO PARAMÉTRICO:

Casos de estudio aplicados a
experimentos agropecuarios

Índice

Tablas

Tabla 1. Métodos estadísticos de análisis multivariante	20
Tabla 2. Detalle de variables cualitativas y cuantitativas evaluadas en 30 híbridos de tomate	29
Tabla 3. Coeficientes de correlación de Pearson entre cinco variables cuantitativas evaluadas en 30 híbridos de tomate	29
Tabla 4. Matriz factorial correspondiente a variables relacionadas de los híbridos de tomate evaluados en invernadero	31
Tabla 5. KMO y prueba de Bartlett	32
Tabla 6. Análisis de correlación de Pearson para variables de respuesta evaluadas.	33
Tabla 7. Varianza total explicada	34
Tabla 8. Matriz de componentes(a) extraídos	34
Tabla 9. Prueba de KMO y prueba de Bartlett	36
Tabla 10. Análisis de correlaciones de variables cuantitativas	36
Tabla 11. Varianza total explicada	37
Tabla 12. Matriz de componentes(a) extraídos	39
Tabla 13. Métodos y niveles de interpretación de resultados	40
Tabla 13. Coeficientes de correlación de Spearman entre seis variables cualitativas evaluadas en 30 híbridos de tomate.	61
Tabla 14. Correlación de Spearman para variables cualitativas	64
Tabla 15. Medidas de discriminación	65
Tabla 16. Resumen de las principales pruebas estadísticas no paramétricas.	83
Tabla 17. Pruebas paramétricas y su alternativa no paramétrica	83
Tabla 18. Pruebas no paramétricas	84
Tabla 19. Distribución sexual de una muestra de 24 animales salvajes.	86

ANÁLISIS MULTIVARIANTE Y NO PARAMÉTRICO:

Casos de estudio aplicados a
experimentos agropecuarios

Prólogo

El análisis multivariante y no paramétrico puede contribuir a enriquecer el debate público sobre los fenómenos que son objeto de interés para los profesionales y los investigadores, gracias a la oportunidad que les ofrece para llevar a cabo un análisis complejo de los datos obtenidos en sus estudios. Al servicio de la investigación cuantitativa y cualitativa, y como extensión de las técnicas de análisis univariante y bivalente, el análisis multivariante y o paramétrico tiene como objetivo principal modelar las múltiples relaciones existentes entre diversas variables de manera simultánea.

La construcción de modelos multivariantes ejerce, pues, un papel importante en el desarrollo de las diferentes disciplinas basadas en el análisis de datos cuantitativos y cualitativos, y requiere, por lo tanto, una atención especial en la formación de futuros profesionales e investigadores. Conocer la lógica, las características específicas de las diferentes técnicas disponibles, los objetivos particulares que permiten lograr y las condiciones en que pueden ser utilizadas, son algunos de los retos importantes a los que nos enfrentaremos en este libro.

Para hacerlo, en este documento nos adentraremos en los aspectos básicos involucrados en el análisis multivariante de los datos, como el marco analítico general que se propone analizar e interpretar, las relaciones simultáneas entre diversas variables mediante la construcción de modelos estadísticos complejos que permiten distinguir la contribución independiente de cada una de ellas en el sistema de relaciones para, de este modo, describir, explicar o predecir los fenómenos que son objeto de interés.

La clave de este marco analítico general no se encuentra, por lo tanto, en el hecho de que los investigadores dispongan de múltiples variables, sino en la capacidad que las diferentes técnicas disponibles les ofrecen para estimar el peso específico o la importancia relativa de cada una de ellas en sus modelos. En este sentido, como veremos, el análisis multivariante puede proporcionar las evidencias necesarias que permitan establecer inferencias a partir de la observación de asociaciones entre las variables, de forma que sea posible extraer conclusiones no sesgadas que, además, sean generalizables más allá de los límites de los estudios particulares, siempre que sea posible.

No se pretende con esta obra cubrir o suplir la vasta bibliografía existente, sino simplemente servir de texto de consulta para docentes y estudiantes que generan conocimientos y tecnología.

Debemos mencionar que nos animó a escribir esta primera edición del libro *Análisis multivariante y no paramétrico: Casos de estudio aplicados a*

experimentos agropecuarios, el hecho de que mucha de la bibliografía existente está enfocada hacia complejas fórmulas y relaciones matemáticas, que no siempre son comprendidas por los estudiantes y docentes de las áreas biológicas y sociales.

Acerca de este libro

Esta obra se escribió teniendo presente y en mente a los estudiantes y docentes de pregrado y posgrado que han realizado tesis y proyectos de investigación, considerando variables cuantitativas y cualitativas. Al escribir esta obra se hizo un esfuerzo por proporcionar toda la información teórica y práctica para explicar el uso de las herramientas multivariantes y no paramétricas. Se documentaron ejemplos reales del análisis de experimentos multivariantes y no paramétricos, incluyendo el uso del software SPSS y otros; además de una bibliografía clásica amplia, a la cual los lectores pueden recurrir para más profundidad de los temas.

Organización flexible del material

Este libro de consulta fue dividido en dos partes: I. El análisis multivariante y II. El análisis no paramétrico. En cada tema se comparte ejemplos de casos agropecuarios reales, producto de tesis de investigación y de proyectos en el campo agrícola y pecuario. Para proporcionar una mayor flexibilidad de los temas se ha diseñado el contenido de las unidades de manera que se puedan seguir secuencialmente las instrucciones. Por lo que, en todos los casos, se identifica la pertinencia de cubrir los temas, la preparación de datos, la selección de las variables, la selección de la técnica adecuada, el análisis de los resultados y la interpretación de los mismos.

Agradecimientos

Escribir un libro sobre análisis multivariante y estadística no paramétrica fue una tarea desafiante y compleja, pero el cumplir este trabajo de manera exitosa fue en buena medida gracias a las personas que participaron e influyeron en su realización.

Agradezco a todos los que contribuyeron en la conclusión de esta obra, por el tiempo dedicado y el empeño puesto para que este libro sea comprensible al estudiante de pregrado y posgrado, donde se comparten los conocimientos en las diferentes ramas y aplicaciones de la estadística multivariante y no paramétrica.

A los estudiantes de los cursos de Metodologías de la Investigación, Estadística (Bioestadística en la nueva malla), Diseños Experimentales (Biometría

en la nueva malla), y ramas afines de pregrado y posgrado, quienes con sus preguntas, dudas y comentarios enriquecieron las diferentes unidades que se tratan en la presente obra.

Mucho de lo que aprendimos respecto a la manera de redactar los temas con mayor claridad fue impartiendo los cursos mencionados. Agradeceremos cumplidamente todos los comentarios y sugerencias provenientes de los lectores.

Queremos, asimismo, agradecer a los miembros del Comité Editorial, con cuyos comentarios y nuestras discusiones sobre las diferentes partes y capítulos estuvieron siempre solícitos a enriquecer el contenido con sus puntos de vista, a cualquier hora del día.

A la Universidad Estatal del Sur de Manabí (UNESUM), por financiar esta obra a través de los proyectos “Producción de hortalizas en campo e invernadero - Fase I” y “Desarrollo de estrategias de incremento productivo del café arábigo (*Coffe arábica L.*) - Fase II”.

ANÁLISIS MULTIVARIANTE Y NO PARAMÉTRICO:

Casos de estudio aplicados a
experimentos agropecuarios

Introducción

La presente compilación de investigaciones se adentra en el apasionante universo de la transformación digital y su impacto en diversos ámbitos de la sociedad ecuatoriana, con un enfoque particular en la gobernanza y la gestión pública. Los trabajos aquí reunidos exploran cómo las tecnologías de la información y la comunicación (TIC) están reconfigurando las relaciones entre el Estado, las empresas y los ciudadanos, generando nuevas oportunidades y desafíos.

A través de una mirada multidisciplinaria, los autores analizan cómo la digitalización está incidiendo en la eficiencia de la recaudación tributaria, fomentando la participación ciudadana en la toma de decisiones, promoviendo la transparencia en la gestión pública, y transformando las prácticas de gestión financiera y contable en diversos niveles de gobierno. Asimismo, se exploran los desafíos asociados a la cultura tributaria, la evasión de impuestos y la gestión del talento humano en el contexto de la transformación digital.

Los estudios de caso presentados en este libro ofrecen una visión detallada de las experiencias concretas en diferentes localidades del Ecuador, permitiendo identificar tanto los avances alcanzados como las brechas existentes. En este sentido, la compilación se convierte en una herramienta valiosa para académicos, tomadores de decisiones y profesionales interesados en comprender los desafíos y oportunidades que plantea la transformación digital en el país.

ANÁLISIS MULTIVARIANTE Y NO PARAMÉTRICO:

Casos de estudio aplicados a
experimentos agropecuarios

Capítulo 1

Técnicas de análisis multivariante

1.1 Introducción

La moderna investigación que se realiza demanda del uso de técnicas actualizadas de análisis estadísticos, que contribuyan y ayuden al investigador a llegar a interpretaciones cercanas a la realidad, en donde se involucran muchas variables. Esto no es posible hacer con las técnicas clásicas del diseño experimental, de ahí que muchos investigadores implementaron técnicas novedosas para estos análisis (Gabriel *et al.*, 2022). En esta obra se pretende compartir las más útiles con base en las experiencias logradas en más de 35 años de investigación. Se consideraron diversos y valiosos documentos de investigaciones realizadas en el ámbito agropecuario. Estas técnicas estadísticas son aplicadas a datos multidimensionales. Se pretende conseguir que estudiantes con distinto bagaje académico (ingenieros, biólogos, sociólogos, matemáticos, etc.) puedan realizar sus propios análisis estadísticos con rigor.

Franco e Hidalgo (2003) mencionan que el origen del análisis multivariado se remonta a los comienzos del siglo XX, con Pearson y Spearman, época en la cual se empezaron a introducir los conceptos de la estadística moderna (Infante Gil y Zárate de Lara, 1991). Las bases definitivas de este tipo de análisis se establecieron en la década de 1930-40 con Hotelling, Wilks, Fisher, Mahalanobis y Bartlett (Bramardi, 2002). En términos generales, el análisis multivariante se refiere a todos aquellos métodos estadísticos que analizan simultáneamente medidas múltiples (más de dos variables) de cada individuo. En sentido estricto, son una extensión de los análisis univariados (análisis de distribución) y bivariados (clasificaciones cruzadas, correlación, análisis de varianza y regresiones simples) que se consideran como tal si todas las variables son aleatorias y están interrelacionadas (Hair *et al.*, 1992).

Los métodos multivariantes, Franco e Hidalgo (2003) los clasifican en dos grandes grupos (Tabla 1): 1) los de dependencia, que son aquellos en los cuales una variable o conjunto de variables es identificado como dependiente de otro conjunto, conocidos como independiente o predictor; y 2) los de interdependencia, o aquellos en que ninguna variable o grupo de variables es definido como independiente o dependiente y, más bien, el procedimiento implica el análisis simultáneo de todo el conjunto de variables (Hair *et al.*, 1992).

Tabla 1.

Métodos estadísticos de análisis multivariante.

Métodos de dependencia (tipo de análisis)	Método de independencia (tipo de análisis)
Discriminante múltiple	Componentes principales
Correlación canónica	Factorial
Regresión múltiple	Conglomerados (clusters)
Multivariante de la varianza	Multidimensional
Conjunto	Correspondencia

Fuente: Hair *et al.* (1992); Franco e Hidalgo (2003).

En esta obra nos ocuparemos básicamente de los métodos de interdependencia, o aquellos en que ninguna variable o grupo de variables es definido como independiente o dependiente y, más bien, el procedimiento implica el análisis simultáneo de todo el conjunto de variables (Franco e Hidalgo, 2003). Haremos el esfuerzo de explicar de manera concisa y sencilla las técnicas más usadas y cómo analizarlas aplicando los comandos del software estadísticos SPSS (González, 2009).

1.2. Análisis factorial

1.2.1. Análisis de componentes principales (ACP)

Este método se basa en la transformación de un conjunto de variables cuantitativas originales en otro conjunto de variables independientes no correlacionadas, llamadas componentes principales (Gómez, 2017). Los componentes deben ser interpretados independientemente unos de otros, ya que contienen una parte de la varianza que no está expresada en otro componente principal (Uriel y Aldás, 2005).

En pocas palabras, se trata de disminuir la dimensión de los problemas con muy poca pérdida de información.

1.2.1.1. Obtención de los componentes principales (García y Gómez, 2008)

1. En definitiva: la obtención de componentes principales consiste en pasar la matriz $\sum_x$ simétrica a la matriz $\sum_p$ diagonal cuyos elementos van en orden decreciente. El paso se hace mediante $\sum_p = \mathbf{A}\sum_x\mathbf{A}'$ con $\mathbf{A}$ matriz cuyas filas son los vectores propios (ortogonales) de $\sum_x$ correspondientes a los valores propios ordenados de mayor a menor. Además $\mathbf{A}$ es ortogonal ($\mathbf{A}' = \mathbf{A}^{-1}$).

2. ¿Mantenemos la varianza? Si

$$\Sigma_1^p \text{Var}(X_1) = \Sigma_1^p \text{Var}(P_i)$$

$$\Sigma_1^p \text{Var}(P_i) = \text{tr}(\Sigma_p) = \text{tr}(A \Sigma_x A') = \text{tr}(A' A \Sigma_x) = \text{tr}(\Sigma_x) = \Sigma_1^p \text{Var} X_i$$

(pues $AA' = I$ por ser A ortogonal)

¿Cuál es la varianza total del modelo?

$$\Sigma \text{Var} X_i = \Sigma \text{Var} P_i = \Sigma \lambda_i \text{ con } \lambda_i \text{ valores propios de } \Sigma_x$$

¿Cuanta varianza retiene el primer componente?

$\text{Var}(P_1) = \lambda_1$; la promoción de varianza retenida por P_1 es:

$$\frac{\lambda_1}{\lambda_1 + \dots + \lambda_p}$$

¿Cuanta varianza retienen los dos primeros componentes?

$$\frac{\lambda_1 + \lambda_2}{\lambda_1 + \dots + \lambda_p}$$

3.- Si algún $\lambda_i = 0$ entonces los siguientes son 0 pues Σ_x es semidef. positiva. Si $\lambda_1 = 0 \rightarrow \text{Var} P_i = 0 \rightarrow P_i$ es constante.

X_1	X_2	$\dots$	X_p	P_1	P_2	P_i	$\dots$	P_p
$\vdots$	$\vdots$	$\vdots$	$\vdots$			C		
$\vdots$	$\vdots$	$\vdots$	$\vdots$			C		
$\vdots$	$\vdots$	$\vdots$	$\vdots$			C		
$\vdots$	$\vdots$	$\vdots$	$\vdots$			C		

$$a_{i1} + a_{i2}X_2 + \dots + a_{ip}X_p = C$$

Yo podría despejar cualquier variable X en función de las demás. $\Rightarrow$ una columna c.l. exacta de otras (en los datos originales) $\Rightarrow$ sobra una variable.

4 - Los componentes no son invariables frente a cambios de escala. El cambio en la unidad de medida produce cambios en los resultados. Esto se evita con variables tipificadas.

En lugar de diagonalizar Σ_x , diagonalizamos Σ_x que es la matriz de covarianza de las X tipificadas (o la matriz de correlaciones de las X originales).

$$Z_i = \frac{X_i - \bar{X}_1}{\sigma_i}$$

Hacemos todo exactamente igual pero sobre Σ_z en vez de sobre Σ_x con la ventaja de que:

$$\text{tr}(\Sigma_z) = p = \text{Var}(Z_1) + \text{Var}(Z_2) + \dots + \text{Var}(Z_p)$$

$\lambda_1/p \rightarrow$ proporción de información que retiene P_1

$\lambda_2/p \rightarrow$ proporción de información que retiene P_2

$\frac{\lambda_1 + \lambda_2}{p}$ proporción de información que retienen conjuntamente los ejes P_1 y P_2

5 – Si se ha decidido utilizar k componentes ¿Cómo queda representadas las variables X_i (o Z_i si hemos tipificado)?

Si os quedamos con algunos componentes perdemos cierta información, pero no la perdemos por igual en todas las variables.

$P = AX$ como A es ortogonal: $A' = A^{-1}$

$$X = A^{-1}P = A'P$$

$$\begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{pmatrix} = \begin{pmatrix} a_{11} & a_{21} & \cdots & a_{j1} & a_{p1} \\ a_{12} & a_{22} & \cdots & a_{j2} & a_{p2} \\ a_{1i} & a_{2i} & \cdots & a_{ji} & a_{pi} \\ a_{1p} & a_{2p} & \cdots & a_{jp} & a_{pp} \end{pmatrix} \cdot \begin{pmatrix} P_1 \\ P_2 \\ \vdots \\ P_p \end{pmatrix}$$

$$X_1 = a_{11}P_1 + a_{21}P_2 + \dots + a_{p1}P_p$$

$\vdots$

$$X_i = a_{1i}P_1 + a_{2i}P_2 + \dots + a_{pi}P_p$$

$\vdots$

$$X_p = a_{1p}P_1 + a_{2p}P_2 + \dots + a_{pp}P_p$$

$$\Rightarrow \text{Var}(X_1) = a_{11}^2\lambda_1 + a_{21}^2\lambda_2 + \dots + a_{p1}^2\lambda_p$$

$$\Rightarrow \text{Var}(X_i) = a_{1i}^2\lambda_1 + a_{2i}^2\lambda_2 + \dots + a_{pi}^2\lambda_p$$

Si yo decido emplear a sólo los 2 primeros componentes, del total de la varianza de X_1 :

$a_{11}^2 \lambda_1 + a_{21}^2 \lambda_2 + \dots + a_{p1}^2 \lambda_p$ me quedo con $a_{11}^2 \lambda_1 + a_{21}^2 \lambda_2$;

Comunalidad

$$\frac{a_{11}^2 \lambda_1 + a_{21}^2 \lambda_2}{a_{11}^2 \lambda_1 + a_{21}^2 \lambda_2 + \dots + a_{p1}^2 \lambda_p}$$

Si he decidido emplear k componentes, la comunidad X_1 será:

$$\frac{a_{11}^2 \lambda_1 + a_{21}^2 \lambda_2 + \dots + a_{k1}^2 \lambda_k}{a_{11}^2 \lambda_1 + a_{21}^2 \lambda_2 + \dots + a_{p1}^2 \lambda_p}$$

Así podemos conocer la comunidad de cada variable X_i al quedarnos con 1,2, k factores.

Si hemos trabajado con datos triplicados.

$$P = AZ;$$

$$Z = A'P;$$

$$Z_1 = a_{11}P_1 + a_{21}P_2 + \dots + a_{p1}P_p;$$

$$Var(Z_1) = 1 = a_{11}^2 \lambda_1 + a_{21}^2 \lambda_2 + \dots + a_{p1}^2 \lambda_p$$

Varianza retenida por las k primeras componentes.

$$a_{11}^2 \lambda_1 + a_{21}^2 \lambda_2 + \dots + a_{k1}^2 \lambda_k$$

Comunalidad

$$\text{para } Z_1 : a_{11}^2 \lambda_1 + a_{21}^2 \lambda_2 + \dots + a_{k1}^2 \lambda_k$$

$$\text{para } Z_i : a_{i1}^2 \lambda_1 + a_{i2}^2 \lambda_2 + \dots + a_{ik}^2 \lambda_k$$

¿Qué parte de la información de X_1 queda recogido en P_1 ?

$$\frac{a_{11}^2 \lambda_1}{a_{11}^2 \lambda_1 + a_{21}^2 \lambda_2 + \dots + a_{p1}^2 \lambda_p}$$

(Si hemos trabajado con variables triplicadas el denominador es 1)

6.- ¿Cuál es la correlación entre dos variables originales tipificadas?

Corr Z_i y Z_j .

$$\text{Corr}(Z_i, Z_j) = \text{Cov}(Z_i, Z_j) = E(a_{1i}P_1 + a_{2i}P_2 + \dots + a_{pi}P_p)(a_{1j}P_1 + a_{2j}P_2 + \dots + a_{pj}P_p) = \\ a_{1i}a_{1j}\text{Var}(P_1) + a_{2i}a_{2j}\text{Var}(P_2) + \dots + a_{pi}a_{pj}\text{Var}(P_p) = a_{1i}a_{1j}\lambda_1 + a_{2i}a_{2j}\lambda_2 + \dots + a_{pi}a_{pj}\lambda_p$$

Si sólo consideramos k variables contamos esa cadena de sumandos y tenemos la correlación reproducida.

P. ej. si sólo tomamos 2 factores P_1 y P_2 .

$$\text{Corr reproducida } (Z_i, Z_j) = a_{1i}a_{1j}\lambda_1 + a_{2i}a_{2j}\lambda_2$$

7.- Correlación entre variables originales y factores

$$\text{Corr}(Z_i, P_j) = \frac{\text{Cov}(Z_i, P_j)}{\sqrt{\text{Var}(P_j)}\sqrt{\text{Var}Z_i}} = \frac{\text{Cov}(Z_i, P_j)}{\sqrt{\lambda_j}}$$

$$\text{Cov}(Z_i, P_j) = E(Z_i, P_j) = E[(a_{1i}P_1 + a_{2i}P_2 + \dots + a_{pi}P_p)P_j] = a_{ji}\text{Var}(P_j) = a_{ji}\lambda_j$$

$$\text{Corr}(Z_i, P_j) = \frac{a_{ji}\lambda_j}{\sqrt{\lambda_j}} = a_{ji}\sqrt{\lambda_j}$$

En particular:

$$\text{Corr}(Z_i, P_1) = a_{1i}\sqrt{\lambda_1}$$

$$\text{Corr}(Z_i, P_2) = a_{2i}\sqrt{\lambda_2}$$

	P_1	P_2
Z_1	$a_{11}\sqrt{\lambda_1}$	$a_{21}\sqrt{\lambda_2}$
Z_2	$a_{12}\sqrt{\lambda_1}$	$a_{22}\sqrt{\lambda_2}$
Z_i	$a_{1i}\sqrt{\lambda_1}$	$a_{2i}\sqrt{\lambda_2}$
Z_p	$a_{1p}\sqrt{\lambda_1}$	$a_{2p}\sqrt{\lambda_2}$

TABLA que en **SPSS** se llama matriz de componentes.

Conociendo el contenido de esta tabla se puede obtener todo lo demás.

La suma por columnas de los elementos al cuadrado proporciona las λ_i .

La suma por filas de los elementos al cuadrado proporciona las comunales.

También se pueden obtener las correlaciones reproducidas.

8 – Gráfico de variables. Representación de las variables originales en los ejes formados por los componentes.

Normalmente representamos en P_1 y P_2 o P_1 y P_3 o P_2 y P_3

$$d = \sqrt{(a_{1i}\sqrt{\lambda_1})^2 + (a_{2i}\sqrt{\lambda_2})^2} = \sqrt{\text{comunalidad}(Z_i)}$$

Se representan las variables de acuerdo a las coordenadas obtenidas en la tabla anterior. (matriz de componentes.)

Cuanto más alejada está Z_i del origen, mayor comunalidad tiene, por lo tanto está mejor representada. Las variables cercanas al origen están mal representadas. Se necesita otro eje para representarlas bien (y por tanto otros gráficos).

La representación está limitada por el círculo unidad.

9.- Gráfico para observaciones.

Z_i tipificada.

Los factores P_j son tales que:

$$E(P_j) = 0$$

$$\text{Var}(P_j) = \lambda_j$$

Ahora hacemos componentes tipificadas.

$$Z_{pj} = \frac{P_j - 0}{\sqrt{\lambda_j}}$$

$$Z_{pj} = \frac{1}{\sqrt{\lambda_j}}(a_{j1}Z_1 + a_{j2}Z_2 + \dots + a_{jn}Z_n)$$

$$Z_{pj} = \frac{a_{j1}}{\sqrt{\lambda_j}}Z_1 + \frac{a_{j2}}{\sqrt{\lambda_j}}Z_2 + \dots + \frac{a_{jp}}{\sqrt{\lambda_j}}Z_p$$

Coeficientes para puntuaciones factoriales.

En el individuo (r)

$$Z_{pj}^{(r)} = \frac{a_{j1}}{\sqrt{\lambda_j}}Z_1^{(r)} + \frac{a_{j2}}{\sqrt{\lambda_j}}Z_2^{(r)} + \dots + \frac{a_{jp}}{\sqrt{\lambda_j}}Z_p^{(r)}$$

Puntuaciones tipificadas (o coordenada) del individuo r en el eje P_j = **Puntuaciones factoriales.**

10 – Criterios para decidir el número de ejes (componentes) a retener.

- Podemos fijar un valor de varianza que queremos retener.
- Podemos fijar un valor mínimo para las comunidades, o al menos para la comunalidad de alguna variable de especial interés.
- Podemos observar los valores propios (gráfico de sedimentación), ver cómo van disminuyendo y decidir con cuantos nos quedamos.
- Como las variables tipificadas tienen varianza mayor que 1, podemos seleccionar aquellos componentes que tengan varianza mayor a 1 (o sea expliquen más que las variables originales) el criterio en este caso es seleccionar los componentes asociados a valores propios > 1.

11.- Test de Barlett.

¿Tiene sentido aplicar CP?

¿Hay alguna estructura de correlación suficiente?

¿Hay suficiente dependencia entre las variables?

Hipótesis nula: No existe dependencia entre las Z_i .

$$\left(\Sigma_z \approx \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \right) \Rightarrow \left| \Sigma_z \right| = 1$$

Si se acepta la hipótesis nula no tiene sentido seguir adelante.

Estadístico de contraste

$$- \left| n - 1 - \frac{1}{6} (2p + 5) \right| \ln |R| \approx \chi^2_{(1/2).p(p-1)}$$

Estand. calculado < X^2 tabla $\rightarrow$ Acepto $\Rightarrow$ No aplicar CP. (sig > 0.05)

Estand. calculado > X^2 tabla $\rightarrow$ Rechazo $\Rightarrow$ Sí aplicar CP. (sig < 0.05)

12.- Rotación de los ejes factoriales.

Supongamos que nos hemos quedado con k ejes, en ocasiones una vez que se ha realizado el análisis y se ha cumplido el principal objetivo (reducir) es posible que se observe que una rotación de los ejes podría proporcionar una mejor interpretación de los resultados.

Existen varios tipos de rotación:

Rotación Varimax:

- Mantiene las comunalidades.
- Mantiene el Total de Varianza explicada.
- Cambia el reparto entre los ejes.

13 – Patrones con frecuencia se observan como componentes principales.

P_1 : Tamaño

P_2 : Forma

-Todas la variables correlacionando positivamente con P_1 “Cuanto más de cualquier variable mayor P_1 ”

-Algunas variables son específicamente correlacionadas (positiva o negativamente) con P_2 .

1.2.2. Órdenes en SPSS para análisis de componentes principales (ACP)

Las clasificaremos en tres tipos:

1. Las específicas del análisis.
2. Las de obtención de puntuaciones factoriales y representación gráfica de las observaciones.
3. Las de cambios (mejoras) en los gráficos.

1. Analizar ↴ Reducción de datos ↴ Análisis factorial ↴

Introducir las variables propias del análisis

En DESCRIPTIVOS: solución inicial, coeficientes, niveles de sig. determinante, KMO y prueba de esfericidad, reproducida.

En EXTRACCIÓN

–Método Componentes principales

–Analizar matriz de correlaciones

Mostrar: solución factorial sin rotar (solo en el caso SIN rotación) y gráfico de sedimentación.

Extraer: - Si se desea obtener los ejes con valor propio mayor que 1: autovalor mayor que 1 (la opción por defecto)

- Si se desea un n.º concreto de ejes: opción N.º de factores igual a

En ROTACIÓN: - Ninguno (sin rotación)

- VARIMAX (cuando se desea rotar)

Mostrar gráficos de saturaciones (siempre, se trata del gráfico de variables, construido con las correlaciones de la matriz de componentes).

2. En PUNTUACIONES:

- Guardar como variables (nuevas columnas en el archivo de datos FACij)
- Método regresión
- En OPCIONES: ordenar coeficientes por tamaño

Gráfico de observaciones: En el menú Gráfico

Dispersión Simple ↴ Definir ↴

- Al eje X el factor 1, al eje Y el factor 2 (u otros)
- Etiquetar por la variable que identifica las observaciones (si la hay)
- Opciones:** Mostrar etiquetas.

Si hay muchas observaciones este gráfico sale confuso. En ocasiones se puede segmentar el fichero antes de hacer el gráfico y entonces aparecen varios gráficos más claros.

3. Modificar un gráfico: Doble click en el gráfico (aparece una nueva ventana).

Cambiar de dimensión 3 a dimensión 2: edición propiedades variables (y se elimina una de las 3).

Colocar ejes en el gráfico: opciones línea de referencia etc.

Ejemplo 1. Evaluación de híbridos de tomate

Este ejemplo corresponde a un trabajo realizado por Gabriel *et al.* (2013) y López *et al.* (2015), en el que estos investigadores evaluaron 30 híbridos de tomate [*Solanum lycopersicum* L. (Mill.)] proporcionados por la Fundación PROINPA en Bolivia. Estos híbridos fueron obtenidos mediante el cruzamiento entre 21 líneas parcialmente puras obtenidas de cultivares comerciales de polinización abierta de tres ciclos de autofecundación (progenitores femeninos) y dos líneas silvestres obtenidas por el Dr. Mikel Stevens de San Diego State University-California, Estados Unidos: línea 70 con resistencia al tospovirus TSWV o peste negra y línea 71 susceptible a tospovirus (progenitores masculinos). Se evaluaron variables cuantitativas y cualitativas (Tabla 2), con las cuales se realizaron tres tipos de análisis: i) análisis de correlación, ii) análisis de componentes principales (ACP) y iii) análisis de correspondencia múltiple (ACM).

Tabla 2.

Detalle de variables cualitativas y cuantitativas evaluadas en 30 híbridos de tomate.

	Variables cuantitativas	Código	Variables cualitativas	Código
Variables morfológicas	1.Tamaño de fruto	TmF	1.Tipo de crecimiento	TCP
	2.Longitud de fruto	LnF	2.Forma predominante del fruto	FPF
	3.Ancho de fruto	AnF	3.Color exterior del fruto	CFM
	4.Rendimiento (Kg)	Y	4.Color de la carne del peicarpio	CCP
			5.Forma del corte transversal del fruto	CTF
			6.Número de lóculos	NOL
Variables agronómicas	1.Número de días a la madurez	NoDM		
	1.Pérdida en pH	PpH		
Variables agroindustriales	2.Pérdida en textura	pTxT		
	3.Ganancia en grados Brix	G°Bx		

Fuente: López (2014).

Análisis de correlación de Pearson

Este análisis permitió conocer el grado de asociación existente entre las diferentes variables sin afectar la relación presente entre unidades (cm, mm, g y días). Para este análisis se utilizó el coeficiente de Pearson aplicado para datos multiestratos cuantitativos, y para los datos cualitativos se utilizó el coeficiente de correlación de Spearman (ver el acápite 1.2.2).

La Tabla 3 muestra una matriz con las correlaciones de Pearson para cinco variables cuantitativas evaluadas.

Tabla 3.

Coefficientes de correlación de Pearson entre cinco variables cuantitativas evaluadas en 30 híbridos de tomate.

	NoDM	TmF	LNf	AnF	Y
NoDM		1			
TmF	0.26*		1		
LnF	0.33*	0.56**		1	
AnF	0.39*	0.55**	0.67**		1
Y	0.15	0.35**	0.33*	0.27*	

Fuente: López (2014).

NoM = Número de días hasta la madurez, TmF = Tamaño del fruto, LnF = Longitud del fruto, AnF = Ancho del fruto, Y = Rendimiento de fruto/planta. **: Altamente significativo al $p < 0.01$ de probabilidad, *: Significativo al $p < 0.05$ de probabilidad.

En la Tabla 3 se observa una correlación positiva altamente significativa entre las variables LnF con AnF ($r = 0.67$). Esto indicaría que a mayor longitud de fruto, más ancho es el fruto. Por otra parte, la correlación del TmF fue positivo y altamente significativo con LnF ($r = 0.56$) y AnF ($r = 0.55$). Por lo tanto, los híbridos con mayor TmF exhibieron mayor AnF y LnF. Hubo correlación positiva baja significativa entre las variables TmF y Y ($r = 0.35$).

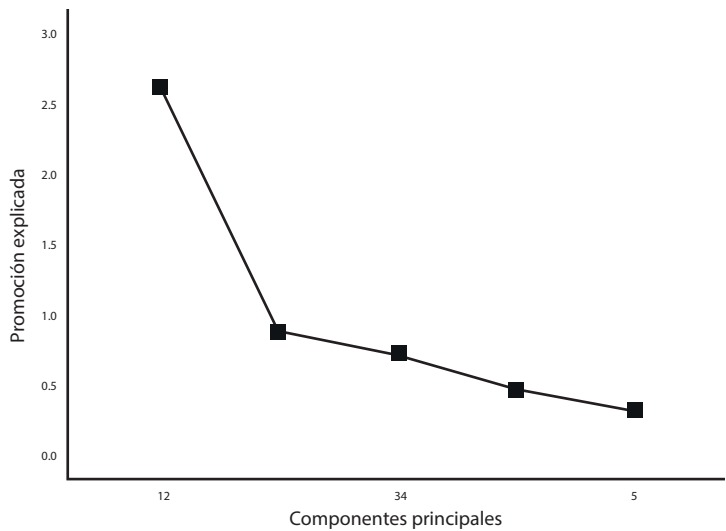
Análisis de los componentes principales (ACP)

El ACP consiste en transformar un conjunto de variables cuantitativas originales a un nuevo conjunto de variables independientes y no correlacionadas, donde los primeros componentes llevan la mayor información o variabilidad. Los componentes deben ser independientes unos de otros, ya que contienen una parte de la varianza que no está expresada en otros componentes principales y el número de componentes depende del número de variables incorporadas en el análisis y la contribución de estas a cada componente principal se expresa en valores y vectores propios. El valor propio representa la varianza asociada con el componente principal y decrece a medida que se generan estos componentes. En cambio, el vector propio contiene los coeficientes de las combinaciones lineales de las p variables originales.

El ACP generó componentes que representaron el número de componentes y su valor propio en su abscisa y el porcentaje de la varianza en la ordenada. Esto mostró el decrecimiento de los primeros componentes en relación con los demás y fueron seleccionados aquellos componentes más significativos.

Se consideró como componentes significativos aquellos valores anteriores al punto de inflexión (Figura 1). Se retuvo dos componentes cuyo valor propio fue ≥ 1 y que expresaron más del 69.88 % de la varianza total. Aunque el segundo componente tiene un valor de 0.9 es próximo a uno; por tanto, se le tomó en cuenta como componente con valor de uno.

Figura 1. Sedimentación para cinco variables cuantitativas agromorfológicas evaluadas en 30 híbridos de tomate (López, 2014; López *et al.*, 2015).



La Tabla 4 expresa la contribución de cada una de las variables que se asociaron al componente principal y la varianza total del componente. Mientras más altos fueron los coeficientes de la varianza, sin importar el signo, más eficaces serán los componentes en la discriminación de los híbridos.

Tabla 4.

Matriz factorial correspondiente a variables relacionadas de los híbridos de tomate evaluados en invernadero.

Componenetes	1º	2º
Valor propio	2.62	-0.63
Porcentaje de varianza	52.35	17.53
Variables	Coeficientes de correlación	
NoDM	0.55	-0.63
TmF	0,78	0.16
LnF	0.84	-0.02
AnF	0.83	-0.16
Y	0.53	0.66

Fuente: López (2014).

NoM = Número de días hasta la madurez, TmF = Tamaño del fruto, LnF = Longitud del fruto, AnF = Ancho del fruto, Y = Rendimiento de fruto/planta.

El primer componente contribuyó con el 52.35 % del total de la varianza (Tabla 4). Aportaron en forma positiva las variables TmF, LnF y AnF. Esto indicaría que este componente identificó a los híbridos con TmF, LnF y AnF. El segundo componente contribuyó con el 17.53 % de la varianza total, en el cual la variable Y aportó positivamente. Por otra parte, la variable NoDM, contribuyó en forma negativa. Este componente distinguió a los híbridos con mayor rendimiento y precocidad.

Ejemplo 2. Caracterización in situ de frutales no tradicionales

En este ejemplo, el objetivo de la investigación fue caracterizar in situ las especies de frutales no tradicionales presentes en las fincas cafetaleras de la parroquia La Unión-Zona Sur, Jipijapa, se analizó las especies de frutales no tradicionales de las fincas cafetaleras. El método de investigación fue descriptivo inferencial. Se realizó un análisis multivariante de componentes principales de nueve variables cuantitativas (Morán y Álvarez, 2023). Asimismo, se realizó un análisis de correspondencia para 16 variables cualitativas evaluadas.

Para iniciar el análisis de esta investigación se siguieron algunos pasos fundamentales, los cuales se describen a continuación.

Prueba de KMO y de Bartlett

La prueba de KMO mostró que no hubo diferencias significativas ($P<0.05$) de probabilidad en la adecuación muestral para las correlaciones parciales realizadas entre las variables evaluadas. Lo que indicó que el análisis factorial fue el más adecuado porque se acercó a la unidad (Tabla 4).

La prueba de esfericidad de Bartlett mostró diferencias altamente significativas ($P<0.01$), indicando esto que hubo correlaciones altas y significativas entre algunas de las variables evaluadas (Tabla 5).

Tabla 5.

KMO y prueba de Bartlett.

Medida de adecuación muestral de Kaiser-Meyer-Olkin		,697
Prueba de esfericidad de Bartlett	Chi-cuadrado aproximado	723,253
	gl	36
	Sig.	,000

Análisis de correlación de Pearson

El análisis de correlación de Pearson mostró que hubo correlaciones altas positivas y significativas entre las variables (Tabla 5): LDH vs. ADH ($r=0,88$), LDH vs. AF ($r=0,97$), ADH vs. AF ($r=0,90$), LDFRU vs. DDFRU ($r=0,88$), LDFRU vs. PDFRU ($r=0,96$) y DDFRU vs. PDFRU ($r=0,82$).

Tabla 6.

Análisis de correlación de Pearson para variables de respuesta evaluadas.

variables	ADP	DDP	LDH	ADH	AF	LDP	LDFRU	DDFRU	PDFRU
ADP	1	,77**	,58**	,62**	,57**	,010	,36**	,38**	,27*
DDP		1	,548**	,54**	,5**	,13	,27*	,33**	,17
LDH			1	,88**	,97**	,097	,482**	,52**	,32*
ADH				1	,902**	,34**	,52**	,65**	,32**
AF					1	,087	,47**	,49**	,27*
LDP						1	,14	,45**	,09
LDFRU							1	,88**	,96**
DDFRU								1	,82**
PDFRU									1

*: La diferencia de medias es significativa al nivel 5%, **: La diferencia de medias es altamente significativa al nivel 1%, ns: No significativo, *: Significativo al 5% de probabilidad, **: Altamente significativo al 1% de probabilidad, ADP: altura de planta, DDT: diámetro de tallo, LDH: largo de hoja, AF: área foliar, LDP: largo de peciolo, LDFRU: largo de fruto, DDFRU: diámetro de fruto y PDFRU: peso de fruto.

Análisis de los componentes principales (ACP)

Estos representan el número de componentes y su valor propio en su abscisa y el porcentaje de la varianza en la ordenada. Esto mostró el decremento de los primeros componentes en relación con los demás y fueron seleccionados aquellos componentes más significativos (Tabla 7). Se consideró como componentes significativos aquellos valores anteriores al punto de inflexión (Figura 2). Se retuvo tres componentes cuyo valor propio fue ≥ 1 y que expresaron más del 86 % de la varianza total.

Tabla 7.

Varianza total explicada.

Componente	Autovalores iniciales			Sumas de las saturaciones al cuadrado de la extracción		
	Total	% de la varianza	% acumulado	Total	% de la varianza	% acumulado
1	4,959	55,104	55,104	4,959	55,104	55,104
2	1,739	19,321	74,425	1,739	19,321	74,425
3	1,069	11,873	86,298	1,069	11,873	86,298
4	,809	8,984	95,282			
5	,231	2,565	97,847			
6	,097	1,080	98,928			
7	,057	,632	99,560			
8	,027	,303	99,863			
9	,012	,137	100,000			

La Tabla 8 muestra que el primer componente contribuyó con el 55 % del total de la varianza. Aportaron en forma positiva las variables altura de planta (ADP), diámetro de tallo (DDT), LDH: largo de hoja (LDH), AF: área foliar (AF), largo de peciolo (LDP), largo de fruto (LDFRU), diámetro de fruto (DDFRU) y peso de fruto (PDFRU). El segundo y tercer componentes contribuyeron con el 19 y 12 %, respectivamente, de la varianza total, en la cual las variables LDFRU, DDFRU y PDFRU aportaron positivamente (Figura 2).

Como consecuencia se generaron dos nuevas variables que explican más del 86 % de la varianza. El primer componente se refiere a las características morfológicas de las plantas, el segundo y tercer componentes agrupan a las características del fruto (Tabla 8).

Tabla 8.

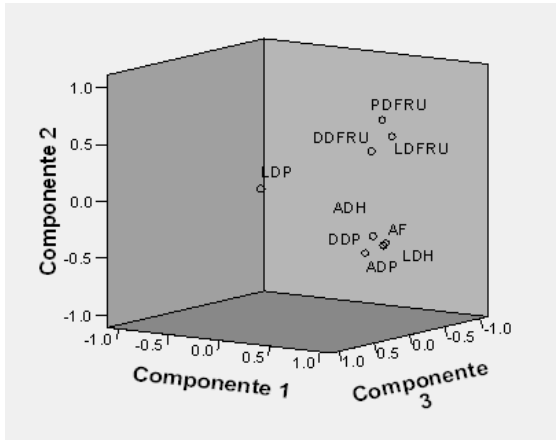
Matriz de componentes(a) extraídos.

Variables	Componente		
	1	2	3
ADP	.709	-.393	-.202
DDP	.646	-.443	-.034
LDH	.860	-.332	-.022
ADH	.896	-.234	.211
AF	.846	-.355	-.010
LDP	.276	.222	.912
LDFRU	.779	.568	-.227
DDFRU	.827	.496	.134
PDFRU	.643	.692	-.284

ns: No significativo, *: Significativo al 5 % de probabilidad, **: Altamente significativo al 1 % de probabilidad, ADP: altura de planta, DDT: diámetro de tallo, LDH: largo de hoja, AF: área foliar, LDP: largo de peciolo, LDFRU: largo de fruto, DDFRU: diámetro de fruto y PDFRU: peso de fruto.

Figura 2.

Representación tridimensional de los componentes principales.



Ejemplo 1.3. Caracterización del cerdo criollo

Esta investigación tuvo como objetivo caracterizar morfométrica y fanerópticamente el cerdo criollo (*Sus strofa domestica*) en Jipijapa-Manabí, para lo que se determinó los índices morfométricos y las características fanerópticas de los cerdos criollos en las comunidades La Cuesta, Santa Rosa, Sancán y Membrillal. Se hizo un análisis multivariante de componentes principales de 16 variables cuantitativas (Lino y Del Valle, 2023).

Prueba de KMO y de Bartlett

La prueba de KMO mostró que no hubo diferencias significativas al $P < 0.05$ de probabilidad en la adecuación muestral, para las correlaciones parciales realizadas entre las variables evaluadas. Lo que indicaría, que el análisis factorial que se realizó es el más adecuado porque se acerca a la unidad (Tabla 9).

En cambio, la prueba de esfericidad de Bartlett mostró diferencias altamente significativas al $P < 0.01$ de probabilidad, indicando esto que hubo correlaciones altas y significativas entre algunas de las variables evaluadas.

Tabla 9.

Prueba de KMO y prueba de Bartlett.

Medida de adecuación muestral de Kaiser-Meyer-Olkin.		0.832
Prueba de esfericidad de Bartlett	Chi-cuadrado aproximado	1002.58
	gl	136
	Sig.	0.000

Análisis de correlación de Pearson

Las correlaciones fueron altas y significativas al $P < 0.01$ de probabilidad para algunas variables analizadas (Tabla 10), como las siguientes: longitud de la cabeza (LCZ) vs. diámetro dorso esternal (DDE) (0.748), alzada de cruz (ALC) (0.735) vs. alzada de grupa (ALG) (0.726). Longitud de grupa (LGR) vs. anchura de la grupa (AGR) (0.771), diámetro bicotal (DBC) (0.804), diámetro dorso esternal (DDE) (0.767), alzada a la cruz (ALC) (0.818), alzada a la grupa (ALG) (0.798), alzada al nacimiento de la cola (ANC) (0.837) y perímetro torácico (PT) (0.864). AGR vs DBC (0.760) y PT (0.813). DBC vs. ALC (0.806), ALG (0.712) y PT (0.807). ALC vs. ALG (0.940), ANC (0.704) y PT (0.830). ALG vs. ANC (0.913), DL (0.704) y PT (0.780). ANC vs. PT (0.818).

Análisis de los componentes principales (ACP)

El ACP generó componentes que representaron el número de componentes y su valor propio en su abscisa y el porcentaje de la varianza en la ordenada. Esto mostró el decrecimiento de los primeros componentes en relación con los demás y fueron seleccionados aquellos componentes más significativos (Tabla 11). Se consideró como componentes significativos aquellos valores anteriores al punto de inflexión (Figura 3). Se retuvo tres componentes cuyo valor propio fue ≥ 1 , los que expresaron más del 76 % de la varianza total.

Tabla 11.

Varianza total explicada.

Componente	Autovalores iniciales			Sumas de las saturaciones al cuadrado de la extracción		
	Total	% de la varianza	% acumulado	Total	% de la varianza	% acumulado
1	9.467	55.688	55.688	9.467	55.688	55.688
2	2.200	12.940	68.628	2.200	12.940	68.628
3	1.282	7.540	76.168	1.282	7.540	76.168
4	.977	5.745	81.912			
5	.806	4.741	86.653			
6	.488	2.873	89.527			
7	.411	2.418	91.945			

Tabla 10.

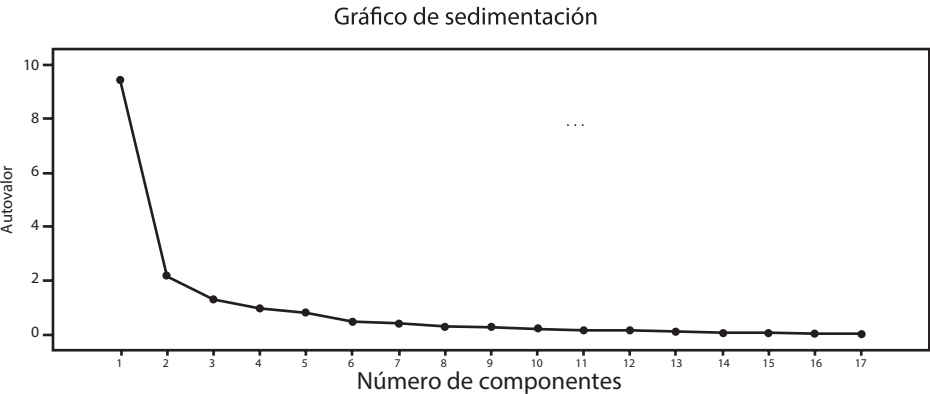
Análisis de correlaciones de variables cuantitativas

	PV	LCZ	ACZ	LH	AH	LGR	AGR	DBC	DDE	ALC	ALG	ANC	DL	PT	PCA	LO	AO
PV	1	.682*	ns	.322*	ns	.655**	.327*	.616**	.481**	.594**	.666**	.635**	.477**	.607**	.641**	.446**	.310*
LCZ		1	.284*	.638**	.428**	.791**	.679**	.740**	.748**	.735**	.726**	.686**	.478**	.640**	.617**	.293*	.381(**)
ACZ			1	.675**	.628**	.398**	.573**	ns	.437**	.312*	.301*	.422**	ns	.468**	Ns	ns	ns
AH				1													
1	.399**	.511**	ns	.655**	Ns	ns	.353**	ns	.408**	ns	ns	ns					
LGR						1	.771**	.804**	.767**	.818**	.798**	.837**	.512**	.864**	.468**	.441**	.426**
AGR							1	.760**	.666**	.683**	.610**	.677**	.272*	.813**	.305*	.308*	.314*
DBC								1	.666**	.806**	.712**	.664**	.391**	.807**	.451**	.410**	.338*
DDE									1	.573**	.540**	.563**	.289*	.660**	.450**	ns	.342*
ALC										1	.940**	.872**	.543**	.830**	.546**	.418**	ns
ALG											1	.913**	.704**	.780**	.680**	.473**	ns
ANC												1	.598**	.818**	.550**	.439**	ns
DL													1	.346*	.579**	.294*	ns
PT														1	.420**	.549**	.363**
PCA															1	.297*	ns
LO																1	ns
AO																	1

** : Correlación es altamente significativa (P<0.01). * : Correlación es significativa (P<0.05). Peso vivo (PV). Longitud de la cabeza (LCZ). Anchura de la cabeza (ACZ). Longitud del hocico (LH). Anchura del hocico (AH). Longitud de la grupa (LGR). Anchura de la grupa (AGR). Diámetro bi-costal (DBC). Diámetro dorso esternal (DDE). Alzada a la cruz (ALC). Alzada a la grupa (ALG). Alzada al nacimiento de la cola (ANC). Diámetro longitudinal (DL). Perímetro torácico (PT). Perímetro de la caña anterior y posterior (PCA). Longitud de la oreja (LO). Anchura de la oreja (AO).

8	.313	1.844	93.789
9	.273	1.606	95.395
10	.207	1.217	96.612
11	.168	.988	97.600
12	.156	.916	98.515
13	.096	.562	99.077
14	.060	.354	99.431
15	.051	.301	99.732
16	.030	.176	99.908
17	.016	.092	100.000

Figura 3.
Sedimentación para 17 variables cuantitativas del cerdo criollo.



La Tabla 12 mostró que el primer componente contribuyó con el 55.69 % del total de la varianza. Aportando en forma positiva las variables LGR, PT, ALC, ANC, ALG, LCZ, DBC, AGR, DDE, LH, PV, PCA y DL (Tabla 8). El segundo componente contribuyó con el 12.94 % de la varianza total, en el cual las variables ACZ y AH aportaron positivamente. El tercer componente contribuyó con 7.54 %, aportando las variables AO y LO positivamente.

Por lo que se generaron tres nuevas variables que explican más del 76 % de la varianza. El primer componente se refiere a las características morfológicas del cuerpo del animal; el segundo componente agrupa a las características de la cabeza, y el tercer componente agrupa a caracteres de la cola y oreja (Tabla 12).

Tabla 12.

Matriz de componentes(a) extraídos.

Variables	Componente		
	1	2	3
LGR	.938	.020	.023
PT	.909	.101	.063
ALC	.890	-.196	-.115
ANC	.890	-.140	-.124
ALG	.886	-.326	-.137
LCZ	.857	-.064	-.047
DBC	.848	-.073	.065
AGR	.824	.350	-.127
DDE	.780	.251	-.097
LH	.771	.470	-.048
PV	.704	-.409	.191
PCA	.597	-.531	-.159
DL	.553	-.524	-.170
ACZ	.462	.673	-.295
AH	.494	.574	-.027
AO	.420	.287	.752
LO	.514	-.143	.672

Ejemplo 1.4.

Se realizó una investigación físico-química del suelo, en cuatro cantones productores de café de la provincia de Manabí: Jipijapa, Paján, 24 de Mayo y Santa Ana (Valverde *et al.*, 2022). El sur de la provincia de Manabí es por excelencia cafetalero, sin embargo, no existen estudios que caractericen el contenido de macro y micronutrientes de sus suelos. Los estudios hasta ahora realizados se limitan a fincas o un sector en particular.

Toma de muestra

Se tomaron 181 de las parroquias rurales: La Unión, El América, El Anegado y de la cabecera cantonal en la comunidad El Matal; de Paján las muestras fueron tomadas de la parroquia Campozano; del cantón 24 de Mayo se consideró la parroquia Noboa, y del cantón Santa Ana las parroquias Ayacucho y Honorato Vázquez.

Para la toma de muestra se aplicó el método no probabilístico bola de nieve (Baltar, 2012), el cual consiste en preguntar al encuestado sobre otros finqueros cercanos y estos a la vez nos conducían a otro. Esta acción se tomó debido a que se desconocía quiénes disponían de cultivares de café. Una vez identificadas las fincas se procedía a tomar las muestras, se siguió el procedimiento en zigzag, y se tomaron 15 submuestras por finca. Se profundizó entre 20-30 cm, considerando que el 86 % de las raíces absorbentes se encuentran en este rango (López *et al.*, 2016). Las muestras se tomaron en la época seca y se empleó calicatas. La mezcla se realizó de manera manual empleando guantes, el peso por muestra fue de 1 kilo y se almacenaron en fundas de papel, previa entrega en el laboratorio.

Análisis de las muestras

Una vez registradas y codificadas, las muestras fueron enviadas a los laboratorios del Instituto Nacional de Investigación Agropecuaria, los cuales están autorizados para la realización de los análisis físico-químicos requeridos. La Tabla 13 señala los métodos implementados, así como los niveles de medición de cada parámetro.

Tabla 13.

Métodos y niveles de interpretación de resultados.

Parámetro	Unidad de medida	Niveles	Método aplicado	Cuantificación
pH		5 - 5,5 Ácido 5,5- 5,9 Medio ácido 6 - 6,5 Líder ácido 6,5 - 7,5 Neutro 7,5 - 9 Medio alcalino	Potenciométrico en pasta saturada	Potenciometría
NH 4	ppm	<21 Bajo 21- 40 Medio >40 Alto	Olsen modificado pH 8.5	Foto colorimétrico-Azul Indofenol en extracto
P	ppm	<8 Bajo 8- 14 Medio >14 Alto	Olsen modificado	Espectrofotometría de luz visible (UV)
K	meq/100 ml	<0,2 Bajo 0,2- 0,38 Medio >0,38 Alto	Olsen modificado	Espectrofotometría de absorción atómica (AA)
Ca	meq/100 ml	<5,1 Bajo 5,1- 8,9 Medio >8,9 Alto	Olsen modificado	Espectrofotometría de absorción atómica (AA)

Mg	meq/100 ml	<1,7 Bajo 1,7- 2,3 Medio >2,3 Alto	Olsen modificado	Espectrofotometría de absorción atómica (AA)
S	ppm	<4 Bajo 4- 19 Medio >19 Alto	Fosfato mono cálcico 0.008M	Turbidimétrica Ba Cl2
Zn	ppm	<3,1 Bajo 3,1- 7 Medio >7 Alto	Olsen modificado	Espectrofotometría de luz visible (UV)
Mn	ppm	<5,1 Bajo 5,1-15 Medio >15 Alto	Olsen modificado	Espectrofotometría de luz visible (UV)
Fe	ppm	<20 Bajo 20- 40 Medio >40 Alto	Olsen modificado	Espectrofotometría de luz visible (UV)
Cu	ppm	<1,1 Bajo 1,1- 4 Medio >4 Alto	Olsen modificado	Espectrofotometría de luz visible (UV)
B	ppm	<0,2 Bajo 0,2- 0,49 Medio >0,49 Alto	Ca - H4(PO4)2.2H2O	Espectrofotometría de luz visible (UV)
Materia orgánica	%	<3 Bajo 3- 5 Medio >5 Alto	WalkleyBlack	Volumetría
Textura	%		Bouyouocus	Hidrómetro

Fuente: INIAP

Se analizó en el laboratorio, el pH, materia orgánica (MO), los macronutrientes N, P, K, S, Ca, Mg; y entre los microelementos, Zn, Cu, Fe, Mn, B (Tabla 2). Como parte del análisis físico del suelo se consideró el contenido porcentual de arena, arcilla y limo, pautando a partir de esta información la clase textural, la capacidad de campo (CC), y el punto de marchitez permanente (PMP).

Base de datos en SPSS

Base de datos Café suabo.sav [ConjuntoDatos1] - IBM SPSS Statistics Editor de datos

Archivo Editar Ver Datos Transformar Analizar Gráficos Utilidades Ampliaciones Ventana Ayuda

9 P 167.00 Viable: 32 de 32 variables

	Productores	Café	Comunidad	Ph	NH4	P	K	Ca	Mg	S	Zn	Cu	Fe
1	M1 Marco Pincay Gutierrez	Jipijapa	1.00	6.40	11.00	21.00	.94	20.00	5.20	19.00	3.50	3.20	81.00
2	M2 Celso Quims	Jipijapa	1.00	7.30	18.00	29.00	1.07	21.00	4.70	14.00	5.20	5.00	39.00
3	M3 Alfonso Quims	Jipijapa	1.00	6.00	19.00	33.00	.96	20.00	6.00	12.00	4.10	3.50	88.00
4	M4 Douglas	Jipijapa	1.00	6.30	24.00	48.00	1.01	21.00	5.90	12.00	3.40	3.80	101.00
5	M5 Luisa Quims	Jipijapa	1.00	6.00	26.00	63.00	.97	21.00	3.80	10.00	4.70	3.50	95.00
6	M6 Tomalá Solís	Jipijapa	1.00	6.40	21.00	90.00	.64	20.00	.80	17.00	5.40	4.90	121.00
7	M7 Eduardo Velazquez	Jipijapa	1.00	5.50	13.00	266.00	.62	22.00	.60	12.00	4.10	2.40	119.00
8	M8 Santiago Villacreses	Jipijapa	1.00	6.20	11.00	14.00	.25	20.00	.60	10.00	1.10	1.90	108.00
9	M9 Mesin	Jipijapa	1.00	6.30	18.00	67.00	1.11	21.00	.80	13.00	3.40	4.10	78.00
10	M10 Gonzalo Menocal	Jipijapa	1.00	5.60	14.00	110.00	.71	19.00	1.10	11.00	3.20	4.20	111.00
11	M11 José Ponce	Jipijapa	1.00	6.00	16.00	285.00	.55	21.00	.50	13.00	2.20	2.70	82.00
12	M12 Serafin Bonilla Sánc	Jipijapa	1.00	7.40	9.00	85.00	3.20	21.00	.90	306.00	10.20	10.70	24.00
13	M13 Isaac Píiza Santana	Jipijapa	1.00	8.30	28.00	13.00	2.53	22.00	3.30	90.00	10.60	5.20	17.00
14	M14 Fernando Tapia Sarcón	Jipijapa	1.00	6.50	39.00	55.00	.88	22.00	2.80	9.00	4.00	4.60	116.00
15	M15 Candelario Lino Gúte	Jipijapa	1.00	6.10	13.00	90.00	.62	21.00	4.30	16.00	2.00	3.20	66.00
16	M16 Virgilio Lino Gúte	Jipijapa	2.00	6.00	8.00	71.00	1.51	20.00	5.30	9.00	2.80	4.10	113.00
17	M17 Domingo Niera Píiza	Jipijapa	2.00	6.60	30.00	40.00	1.39	22.00	4.70	21.00	10.00	5.20	102.00
18	M18 Julio Muñoz Sánchez	Jipijapa	2.00	5.80	34.00	83.00	1.79	22.00	3.70	25.00	5.60	4.00	123.00
19	M19 José Niera Píiza	Jipijapa	2.00	7.10	30.00	32.00	1.51	23.00	5.60	33.00	8.00	5.40	71.00
20	M20 Claudio Chepato Gata	Jipijapa	2.00	6.60	11.00	12.00	1.00	22.00	4.30	12.00	4.60	2.90	109.00
21	M21 Raúl Chóez Ray	Jipijapa	2.00	6.90	26.00	31.00	.99	22.00	2.80	19.00	9.80	8.10	74.00
22	M22 María Cháez Ray	Jipijapa	2.00	6.40	27.00	33.00	.95	20.00	2.70	18.00	3.70	3.40	83.00

Vista de datos Vista de variables

IBM SPSS Statistics Processor está listo Unicode ON

26°C 11/1/2025

Análisis de los datos

Base datos Café suabo.sav [ConjuntoDatos1] - IBM SPSS Statistics Editor de datos

ArchivoEditarVerDatosTransformarAnalizarGráficosUtilidadesAmpliacionesVentanaAyuda

</

En el segundo paso se deben incluir todas las variables cuantitativas que pueden ser discretas o continuas, como se observa en la figura a continuación.

Cantón	Comindad	Ph	NH4	P	K	Ca	Mg	
Jipijapa	1,00	6,40	11,00	21,00	,94	20,00	5,20	
Jipijapa	1,00	7,30	18,00	20,00	1,07	21,00	4,70	
Jipijapa							6,00	
Jipijapa							5,90	
Jipijapa							3,80	
Jipijapa							,80	
Jipijapa							,60	
Jipijapa							,60	
Jipijapa							,80	
Jipijapa							1,10	
Jipijapa							,50	
Jipijapa							,90	30
Jipijapa							3,30	9
Jipijapa							2,80	
Jipijapa							4,30	
Jipijapa	2,00	6,00	6,00	71,00	1,51	20,00	5,30	
Jipijapa	2,00	6,60	30,00	40,00	1,30	22,00	4,70	2
Jipijapa	2,00	5,80	24,00	62,00	4,70	22,00	2,70	2

Productores

Cantón [Cantón]

Comindad [Comind...

Ward Method

REGR factor score ...

REGR factor score ...

REGR factor score ...

Ph [Ph]

NH 4 [NH4]

P [P]

K [K]

Ca [Ca]

Mg [Mg]

S [S]

Variables:

Variable de selección:

Valor...

Descriptivos...

Extracción...

Rotación...

Puntuaciones...

Opciones...

Aceptar

Pegar

Restablecer

Cancelar

Ayuda

Posteriormente se abre la ventana **Análisis factorial - Descriptivos**, de la cual se abren varias opciones, como **Descriptivos univariados**, que nos muestra la media y la desviación típica; la **Solución inicial**, por su parte, permite obtener las comunidades iniciales y el porcentaje de varianza de cada valor.

La matriz de correlaciones abre varias opciones de análisis, destacando el **coeficiente** de correlación, los niveles de significancia que muestra la matriz de correlación. El **determinante**, por su lado, muestra la matriz de correlaciones, y se interpretan los resultados en función de la proximidad de los valores con el cero, lo que indicaría que existe relación entre las variables, con lo que se justifica la aplicación del estudio factorial.

Existen otras opciones como la **Inversa**, que muestra la inversa de la matriz de correlaciones. Encontramos también a la **Reproducida**, que muestra la matriz de correlaciones reproducidas con la matriz de correlaciones residuales, marcando las diferencias entre las correlaciones observadas y las correlaciones reproducidas. Si el número de residuos con valores elevados es mínimo, entonces el modelo es correcto.

También se cuenta con la opción **Anti-imagen**, que muestra la matriz de covarianzas **anti-imagen** y la matriz de correlaciones **anti-imagen**.

ANÁLISIS MULTIVARIANTE Y NO PARAMÉTRICO: CASOS DE ESTUDIO APLICADOS A EXPERIMENTOS AGROPECUARIOS

IBM SPSS Statistics Processor está listo. Unencode ON.

12:06 11/1/2025

La última opción es el KMO y prueba de esfericidad de Bartlett, contrasta si las correlaciones parciales entre las variables son suficientemente pequeñas. El estadístico KMO va desde 0 a 1, y la regla de decisión establece que valores menores a 0,5 determinan que no se justifica la aplicación del análisis factorial.

Se continua el análisis abriendo la ventana Análisis factorial extracción, de la que se despliegan varias opciones. Así tenemos: Primero el método, en el que ratificamos "Componentes principales"; luego tenemos la ventana Analizar, desde la cual se aprecian las opciones: la matriz de correlaciones y la matriz de covarianza, ambas estarían relacionadas con el ítem Anti-imagen, observada en la ventana de descriptivos.

ANÁLISIS MULTIVARIANTE Y NO PARAMÉTRICO: CASOS DE ESTUDIO APLICADOS A EXPERIMENTOS AGROPECUARIOS

IBM SPSS Statistics Editor de datos

Archivo Editar Ver Datos Transformar Analizar Gráficos Utilidades Ampliaciones Ventana Ayuda

9 - P 67.00

Análisis factorial: Extracción

Método: Componentes principales

Analizar:

- ☒ Matriz de correlaciones
- ☐ Matriz de covarianzas

Mostrar:

- ☒ Solución factorial sin rotar
- ☐ Gráfico de sedimentación

Editar:

- ☐ Basado en autovalor
- Autovalores mayores que: 1
- ☒ Número fijo de factores
- Factores que extraer: 2

Nº máximo de iteraciones para convergencia: 25

Continuar Cancelar Ayuda

	Producciones	Cantón	Comunidad	Ph	NH4	P	K	Ca	Mg	S
1	M1 Marco Pincay Gutierrez	Jipijapa							5.20	19.00
2	M2 Celso Guzmán	Jipijapa							4.70	14.00
3	M3 Alfonso Guzmán	Jipijapa							6.80	12.00
4	M4 Douglas	Jipijapa							5.90	12.00
5	M5 Luisa Guzmán	Jipijapa							3.80	19.00
6	M6 Tamala Solís	Jipijapa							.80	17.00
7	M7 Eduardo Velastegui	Jipijapa							.60	12.00
8	M8 Santiago Villacreses	Jipijapa							.60	19.00
9	M9 Morán	Jipijapa							.80	13.00
10	M10 Gonzalo Menocal	Jipijapa							1.10	11.00
11	M11 José Ponce	Jipijapa							.50	13.00
12	M12 Serafin Bonilla Sánc	Jipijapa							.90	396.00
13	M13 Isaac Píiza Santana	Jipijapa							3.30	90.00
14	M14 Fernando Tagle Sarcán	Jipijapa							2.80	9.00
15	M15 Candelario Lino Güite	Jipijapa							4.30	16.00
16	M16 Virgilio Lino Güite	Jipijapa							5.30	9.00
17	M17 Domingo Neira Píizas	Jipijapa							4.70	21.00
18	M18 Julia Muñoz Sánchez	Jipijapa	2.00	5.90	34.00	83.00	1.79	22.00	3.70	25.00
19	M19 José Neira Píizas	Jipijapa	2.00	7.10	30.00	32.00	1.51	23.00	5.60	33.00
20	M20 Claudio Chiquito Gola	Jipijapa	2.00	6.60	11.00	12.00	1.88	22.00	4.30	12.00

La ventana **“Mostrar”** nos muestra dos alternativas: **Solución factorial sin rotar y Gráfico de sedimentación**. Este último es importante porque de manera gráfica muestra el número de factores que deben estar en la solución.

IBM SPSS Statistics Editor de datos

Archivo Editar Ver Datos Transformar Analizar Gráficos Utilidades Ampliaciones Ventana Ayuda

9 - P 67.00

Análisis factorial: Rotación

Método: Ninguno

Analizar:

- ☒ Ninguno
- ☐ Qantimax
- ☐ Varimax
- ☐ Equimax
- ☐ Qúimán directo
- ☐ Promax

Mostrar:

- ☐ Solución rotada
- ☒ Gráfico de cargas

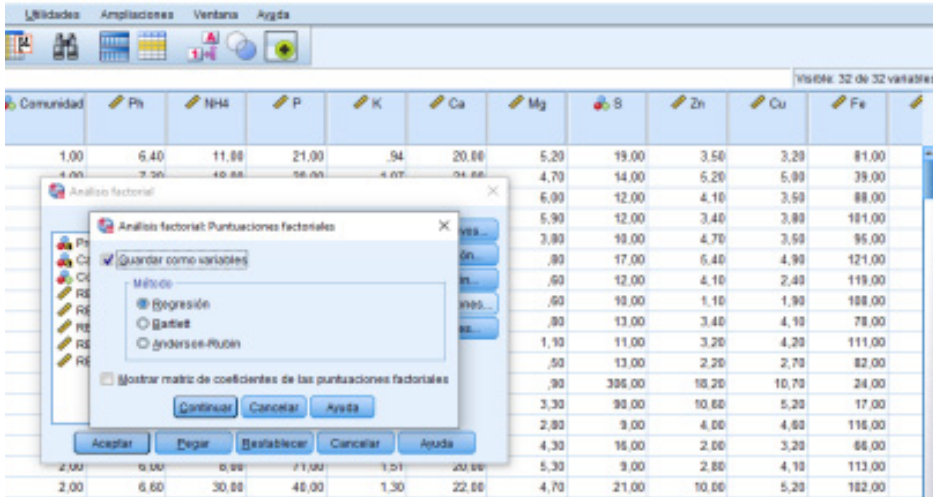
Nº máximo de iteraciones para convergencia: 25

Continuar Cancelar Ayuda

	Producciones	Cantón	Comunidad	Ph	NH4	P	K	Ca	Mg	S	Zn	Cu	Fe
1	M1 Marco Pincay Gutierrez	Jipijapa	1.00	6.40	11.00	21.00	.94	20.00	5.20	19.00	3.50	3.20	81.00
2	M2 Celso Guzmán	Jipijapa	1.00	7.50	18.00	35.00	1.07	19.00	4.70	14.00	5.20	5.00	39.00
3	M3 Alfonso Guzmán	Jipijapa	1.00	6.80	12.00	12.00	1.00	6.00	6.00	12.00	4.10	3.50	88.00
4	M4 Douglas	Jipijapa	1.00	5.90	12.00	3.40	3.80	6.00	12.00	3.40	3.80	101.00	
5	M5 Luisa Guzmán	Jipijapa	1.00	3.80	10.00	4.70	3.50	3.80	17.00	5.40	4.90	121.00	
6	M6 Tamala Solís	Jipijapa	1.00	.80	12.00	4.10	2.40	.80	12.00	4.10	2.40	119.00	
7	M7 Eduardo Velastegui	Jipijapa	1.00	.80	10.00	1.10	1.90	.80	10.00	1.10	1.90	108.00	
8	M8 Santiago Villacreses	Jipijapa	1.00	.80	13.00	3.40	4.10	.80	13.00	3.40	4.10	111.00	
9	M9 Morán	Jipijapa	1.00	1.10	11.00	3.20	4.20	1.10	11.00	3.20	4.20	111.00	
10	M10 Gonzalo Menocal	Jipijapa	1.00	.50	13.00	2.20	2.70	.50	13.00	2.20	2.70	82.00	
11	M11 José Ponce	Jipijapa	1.00	.90	306.00	18.20	10.70	24.00	.90	306.00	18.20	10.70	24.00
12	M12 Serafin Bonilla Sánc	Jipijapa	1.00	3.30	90.00	10.60	5.20	17.00	3.30	90.00	10.60	5.20	17.00
13	M13 Isaac Píiza Santana	Jipijapa	1.00	2.80	9.00	4.00	4.60	116.00	2.80	9.00	4.00	4.60	116.00
14	M14 Fernando Tagle Sarcán	Jipijapa	1.00	4.30	16.00	2.00	3.20	66.00	4.30	16.00	2.00	3.20	66.00
15	M15 Candelario Lino Güite	Jipijapa	1.00	5.30	9.00	2.80	4.10	113.00	5.30	9.00	2.80	4.10	113.00
16	M16 Virgilio Lino Güite	Jipijapa	1.00	4.70	21.00	10.00	5.20	102.00	4.70	21.00	10.00	5.20	102.00
17	M17 Domingo Neira Píizas	Jipijapa	1.00	2.00	5.80	34.00	83.00	1.79	2.00	5.80	34.00	83.00	1.79
18	M18 Julia Muñoz Sánchez	Jipijapa	1.00	2.00	7.10	30.00	32.00	1.51	2.00	7.10	30.00	32.00	1.51
19	M19 José Neira Píizas	Jipijapa	1.00	2.00	6.60	11.00	12.00	1.80	2.00	6.60	11.00	12.00	1.80
20	M20 Claudio Chiquito Gola	Jipijapa	1.00	2.00	6.90	26.00	31.00	.99	2.00	6.90	26.00	31.00	.99
21	M21 Raúl Chéz Ray	Jipijapa	1.00	2.00	6.90	26.00	31.00	.99	2.00	6.90	26.00	31.00	.99

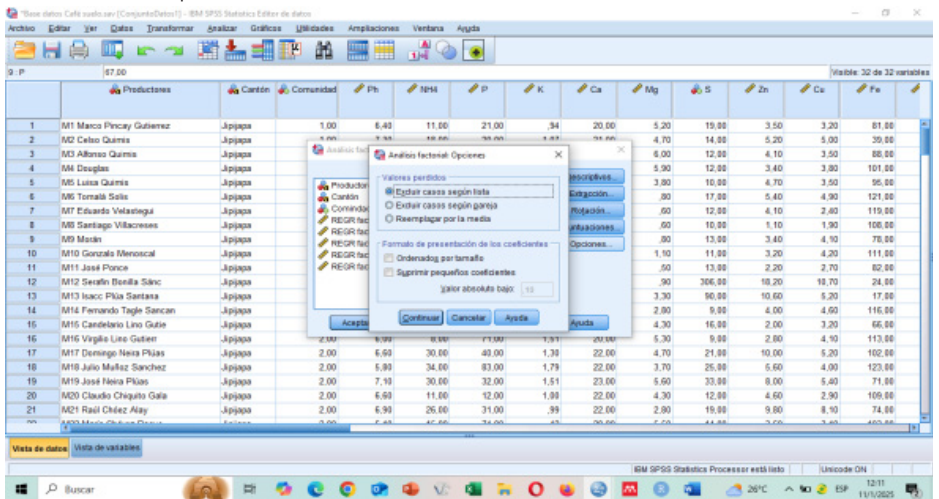
En la ventana Método se escoge Ninguno, dando la opción de que el análisis actúe por defecto. En la ventana Mostrar se selecciona el ítem Gráficos de carga. Este gráfico muestra la dispersión y ubicación de las variables, en tres factores asignados por defecto.

La siguiente ventana es el **Análisis factorial: Puntuaciones factoriales**, en el que encontraremos el ítem **“Guardar como variables”**, la cual consideraremos dentro del análisis, y que garantiza tener guardados los datos automáticamente en el editor, y para ese efecto crea una nueva columna, que pasaría a ser parte de futuros análisis.



La ventana métodos brinda tres opciones, la recomendada es Regresión, la cual da una estimación de las puntuaciones factoriales.

Desactivada por defecto se encuentra el ítem “Mostrar matriz de coeficientes de las puntuaciones factoriales”.



Finalmente, se presenta el “**Análisis factorial: opciones**”, que da lugar a la apertura de la ventana “**Valores perdidos**”, que a su vez presenta tres ítems: “**Excluir casos según lista**”, que es la opción por defecto. Se excluyen el análisis los sujetos que tengan valores perdidos en cualquiera de las variables trasladadas a la lista de variables; “**Excluir casos según pareja**”. Este método permite aprovechar más cantidad de información que la anterior, pero no todas las correlaciones se calculan sobre el mismo número de sujetos y podrían obtenerse matrices de correlaciones inconsistentes; y el ítem “**Reemplazar por la media**”, los valores perdidos de una variable se sustituyen por la media de esa variable.

Se encuentra, además, la ventana “**Formato de presentación de los coeficientes**”, en la que se encuentran dos ítems, “**Ordenados por tamaño**” y “**Suprimir pequeños coeficientes**”, en el ejercicio planteado no se las considera. En el análisis opciones, se deja la opción por defecto.

Una vez ingresada toda la información se da clic en aceptar en la ventana principal.

Presentación de resultados:

Matriz de correlacionesa														
Correlación	NH	4	P	K	Ca	Zn	Cu	Fe	M.O. %	Ca/Mg	Mg/K	Arena	Limo	Arcilla
Ph	1.00	0.22	0.13	0.51	0.02	0.40	0.39	-0.46	0.12	-0.12	-0.42	-0.23	0.15	0.21
NH 4	0.22	1.00	0.30	0.46	0.21	0.31	0.41	0.17	0.25	-0.08	-0.33	-0.34	0.18	0.31
P	0.13	0.30	1.00	0.39	0.17	0.19	0.27	0.24	0.28	0.50	-0.49	0.02	-0.02	-0.04
K	0.51	0.46	0.39	1.00	0.34	0.48	0.60	-0.06	0.14	0.01	-0.65	-0.38	0.42	0.17
Ca	0.02	0.21	0.17	0.34	1.00	-0.04	0.11	-0.24	-0.09	0.24	-0.29	-0.17	0.16	0.10
Zn	0.40	0.31	0.19	0.48	-0.04	1.00	0.58	0.10	0.44	0.09	-0.36	-0.18	0.19	0.05
Cu	0.39	0.41	0.27	0.60	0.11	0.58	1.00	0.07	0.13	-0.05	-0.43	-0.49	0.41	0.34
Fe	-0.46	0.17	0.24	-0.06	-0.24	0.10	0.07	1.00	0.28	0.06	-0.04	0.11	-0.01	-0.17
M.O. %	0.12	0.25	0.28	0.14	-0.09	0.44	0.13	0.28	1.00	0.28	-0.26	-0.02	-0.05	0.06
Ca/Mg	-0.12	-0.08	0.50	0.01	0.24	0.09	-0.05	0.06	0.28	1.00	-0.31	0.25	-0.13	-0.23
Mg/K	-0.42	-0.33	-0.49	-0.65	-0.29	-0.36	-0.43	-0.04	-0.26	-0.31	1.00	0.11	-0.16	0.01
Arena	-0.23	-0.34	0.02	-0.38	-0.17	-0.18	-0.49	0.11	-0.02	0.25	0.11	1.00	-0.70	-0.81
Limo	0.15	0.18	-0.02	0.42	0.16	0.19	0.41	-0.01	-0.05	-0.13	-0.16	-0.70	1.00	0.18
Arcilla	0.21	0.31	-0.04	0.17	0.10	0.05	0.34	-0.17	0.06	-0.23	0.01	-0.81	0.18	1.00
a. Determinante = .000														

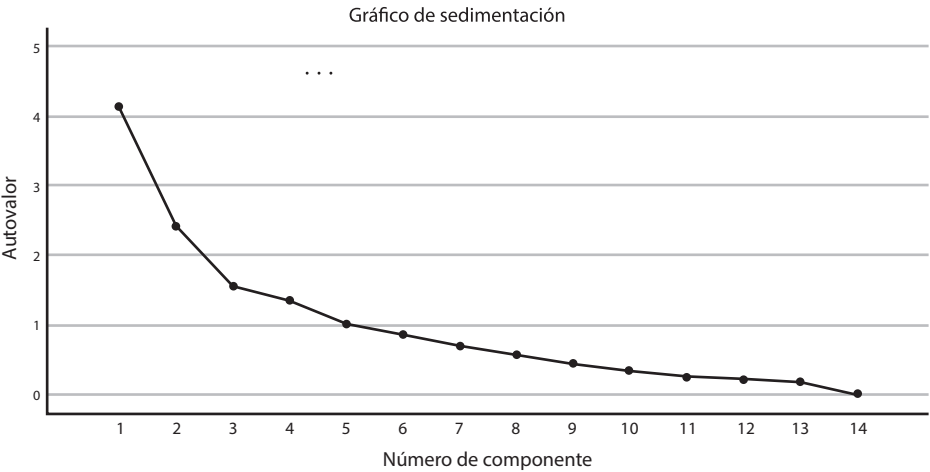
El determinante de la matriz de correlaciones establece que existe correlación entre variables, justificando de esta manera la aplicación del análisis de componentes principales.

Prueba de KMO y Bartlett		
Medida Kaiser-Meyer-Olkin de adecuación de muestreo		0.741
	Aprox. Chi-cuadrado	6197.123
Prueba de esfericidad de Bartlett	Gl	136
	Sig.	.000

La prueba de KMO y Bartlett al obtener menos de 1 confirma el resultado de la correlación, aprobando el desarrollo del análisis.

Varianza total explicada						
Componente	Autovalores iniciales			Sumas de cargas al cuadrado de la extracción		
	Total	% de varianza	% acumulado	Total	% de varianza	% acumulado
1	4.154	29.670	29.670	4.154	29.670	29.670
2	2.418	17.269	46.939	2.418	17.269	46.939
3	1.555	11.110	58.049	1.555	11.110	58.049
4	1.336	9.543	67.591			
5	1.009	7.210	74.802			
6	.856	6.112	80.914			
7	.680	4.860	85.774			
8	.557	3.976	89.750			
9	.437	3.123	92.872			
10	.340	2.425	95.298			
11	.255	1.821	97.118			
12	.216	1.544	98.662			
13	.177	1.262	99.924			
14	.011	.076	100.000			
Método de extracción: análisis de componentes principales.						

Este gráfico es una representación esquemática de la varianza y determina los componentes que mayormente la representan.



La aplicación del ACP determinó 3 componentes, que representan el 58.04 % de la varianza total.

El componente 1 se caracteriza por contar con suelos con altos contenido de NH₄ y K, así como Cu, Zn y Ph, representa el 29.67 % de la varianza.

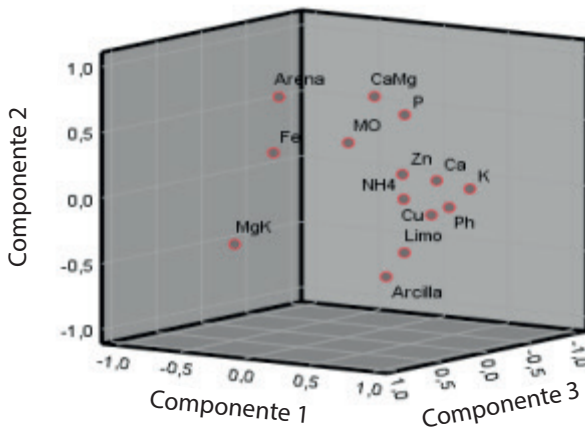
El componente 2, con el 17.26 % de varianza, está representado por suelos con alto contenido de P, MO, Ca/Mg y variables físicas como arena y arcilla.

El componente 3, conformado por suelos con alto contenido de Ca y Fe, representa el 11.11 % de la varianza total.

Matriz de componente ^a			
	Componente		
	1	2	3
Ph	.578	-.110	-.341
NH 4	.622	.039	.228
P	.437	.624	-.048
K	.841	.079	-.196
Ca	.313	.032	-.581
Zn	.626	.232	.247
Cu	.791	-.074	.161
Fe	-.029	.401	.745
M.O. %	.322	.466	.416
Ca/Mg	.056	.695	-.257
Mg/K	-.673	-.431	.259
Arena	-.630	.634	-.172
Limo	.521	-.402	.075
Arcilla	.437	-.581	.158
Método de extracción: análisis de componentes principales.			
a. 3 componentes extraídos.			

Representación tridimensional de los tres factores.

Gráfico de componente



1.2.2. Análisis factorial de correspondencias (AFC)

- Técnica de Interdependencia.
- Procedimiento factorial para variables cualitativas.
- Se basa en el análisis de las asociaciones entre las modalidades de dos o más variables cualitativas. Cuando el análisis se realiza para dos variables nos referimos a Correspondencias simples y si son más variables, Correspondencias múltiples.

Objetivo

La factorización (búsqueda o construcción de nuevos factores, nuevas variables a partir de iniciales).

Consiste en reducir un espacio de grandes dimensiones (número de modalidades o categoría) a otro de menor dimensión (generalmente dos) en el que se ubiquen o proyecten simultáneamente las modalidades de las variables. Además, podemos observar:

- Semejanzas entre modalidades de una misma variable.
- Asociaciones (y repulsiones) entre modalidades de distintas variables.
- Modalidades “raras”.

Presentación de datos
Correspondencias simples
(Dos variables, tabla contingencia)

	a ₁	a ₂	...	a _p
b ₁	Frecuencias n_{ij}			
b ₂				
b _n				

Correspondencias múltiples
(Varias variables, tabla disyuntiva completa)

Observación	a ₁ , a ₂ a _p	b ₁ , b ₂ ... b _r	c ₁ , c ₂ ... c _s
1	0,1,0,0,0	1,0,...,0	0,0,...,1
2	1,0,0,0,0	0,1,...,0	1,0,...,0
3			
⋮			

Extracción de factores

En ACP el criterio para extraer factores era que se mantuviera la mayor parte de la varianza.

En AFC el criterio para extraer factores es captar la mayor parte de ASOCIACIÓN entre modalidades. A esa asociación la llamamos inercia, es decir, extraemos los factores de modo que se produzca lo más fielmente posible las asociaciones entre modalidades.

Los datos se transforman de tal forma que las modalidades de una variable juegan el papel de variables y las de la otra de observaciones (y reciprocamente).

Obtendremos:

- Cuánta inercia retiene cada eje (en C.P. cuánta varianza explica cada factor).
- Contribución absoluta de las modalidades a la formación de cada eje.
- Contribución relativa con que cada eje representa a cada modalidad, es decir, cómo queda representada cada modalidad (en ACP comunalidad).
- Modalidades semejantes de la misma variable.
- Asociación (+ o -) de modalidades de distintas variables.
- Se puede proyectar en el mismo gráfico variables suplementarias que no intervienen en el proceso, pero que enriquecen la presentación.

Correspondencias simples

2 variables, A variable cualitativa con p modalidades (a1. a2. aj. ap)

B, variable cualitativa con n modalidades (b1. b2. bi. bn)

N, observaciones nij = número de observaciones en las modalidades aj y bi

		a_p	
b_1	$N_{.1} = \sum_{i=1}^n n_{i1}$	n_{1p}	$N_{1.} = \sum_{j=1}^p n_{1j}$
b_2	$N_{.2} = \sum_{i=1}^n n_{i2}$	n_{2p}	$N_{2.} = \sum_{j=1}^p n_{2j}$
b_i	$N_{.j} = \sum_{i=1}^n n_{ij}$	n_{ip}	$N_{i.} = \sum_{j=1}^p n_{ij}$
b_n	$N_{.p} = \sum_{i=1}^n n_{ip}$	n_{np}	$N_{n.} = \sum_{j=1}^p n_{nj}$
		$N_{.p}$	N
$N = \sum_{i=1}^n N_{i.} = \sum_{j=1}^p N_{.j}$			

erfiles de las filas.

Fila 1 = $b_1 = (\frac{n_{11}}{N_{1.}}, \frac{n_{12}}{N_{1.}}, \dots, \frac{n_{1j}}{N_{1.}}, \dots, \frac{n_{1p}}{N_{1.}})$

Fila i = $b_i = (\frac{n_{i1}}{N_{i.}}, \frac{n_{i2}}{N_{i.}}, \dots, \frac{n_{ij}}{N_{i.}}, \dots, \frac{n_{ip}}{N_{i.}})$

Perfiles promedio o MASA: (N.1 N . N.2 N N.j N N.p N)

Los perfiles-fila muestran cómo se distribuye la variable A en las modalidades de B. El perfil promedio o MASA muestra la distribución marginal de la variable A.

Si los perfiles de dos filas son idénticos significa que A se comporta igual en ellas, es decir, A se distribuye independientemente de las modalidades de B.

Por el contrario, perfiles muy distintos indican dependencia fuerte entre variables.

Análogamente se pueden calcular los perfiles-columna.

Podemos calcular distancias entre filas, distancia de una fila al promedio, distancia entre columnas, distancia de una columna al promedio:

$$d_{x^2}(b_i, b_{i'}) = \sqrt{\frac{(\frac{n_{i1}}{N_i} - \frac{n_{i'1}}{N_{i'}})^2}{\frac{N_1}{N}} + \frac{(\frac{n_{i2}}{N_i} - \frac{n_{i'2}}{N_{i'}})^2}{\frac{N_2}{N}} + \dots}$$

$$d_{x^2}(b_i, 0) = \sqrt{\frac{(\frac{n_{i1}}{N_i} - \frac{N_1}{N})^2}{\frac{N_1}{N}} + \frac{(\frac{n_{i2}}{N_i} - \frac{N_2}{N})^2}{\frac{N_2}{N}} + \dots}$$

Dependencia e independencia entre variables

Cuando para todo

$$n_{ij} = \frac{N_i \cdot N_j}{N}$$

Los perfiles resultan idénticos, ello implica que hay independencia entre las variables.

Por el contrario, cuando

$$\frac{N_i \cdot N_j}{N} = n'_{ij} \neq n_{ij}$$

decimos que hay asociación.

El cuantificador clásico de la asociación entre variables es:

$$X^2 = \sum \frac{(n'_{ij} - n_{ij})^2}{n'_{ij}}$$

Ho: Hay independencia entre valores.

Bajo Ho

$$X^2 \approx X^2_{(p-1)(n-1)}$$

Calculamos

$$X^2 = \sum \frac{(n'_{ij} - n_{ij})^2}{n_{ij}}$$

Si

$$X^2 > X^2_{(p-1)(n-1)}$$

Rechazamos la independencia (tiene sentido factorizar).

Si

$$\frac{X^2}{N} \leq X^2_{(p-1)(n-1)}$$

La inercia mide el grado de intensidad en la asociación entre variables.

Cuanto mayor es el valor X^2 calculado, mayor es la inercia que tienen los datos y más sentido tiene factorizar.

Extracción de factores

La inercia total se puede descomponer en:

i_1 (inercia en el eje 1) + i_2 (inercia en el eje 2) + i_3 (inercia en el eje 3) + etc.

$$i_1 > i_2 > i_3 > i_4 \dots$$

i_1/I = Porcentaje de inercia que retiene el primer eje.

i_2/I = Porcentaje de inercia que retiene el segundo eje.

$i_1 + i_2/I$ = Porcentaje que retienen conjuntamente los dos primeros ejes.

Los factores se extraen al diagonalizar la matriz $Qf = A.A'$ con

$$A = \left(\frac{P_{ij}}{\sqrt{P_i} \cdot \sqrt{P_j}} \right)_{n \times p}$$

y $P_{ij} = n_{ij}/N$ (frecuencias relativas)

$$P_i = \frac{N_i}{N}$$

$$P_j = \frac{N_j}{N}$$

La matriz Q siempre tiene "1" como valor propio (se omite).

Valores propios de $Q = 1, \lambda_1, \lambda_2, \lambda_{\min(p,n)-1}$

Número máximo de ejes: $\min(p,n)-1$

Con ellos retenemos el 100% de la inercia

La suma de los valores propios $\sum \lambda_i = 1$ es la inercia.

Para cada valor propio ($\lambda_i = 1$) obtendremos un vector propio que normalizaremos.

$$(\hat{v}_{11}, \hat{v}_{12}, \dots, \hat{v}_{1n})$$

Por ejemplo, para λ_1 obtenemos $(\hat{v}_{11}, \hat{v}_{12}, \dots, \hat{v}_{1n})$ que es el 1.º eje.

Para el segundo valor propio, λ_2 , obtendremos un vector propio normalizado.

El segundo eje es

$$\hat{v}_2 = (\hat{v}_{21}, \hat{v}_{22}, \dots, \hat{v}_{2n})$$

$$(\hat{v}_{21}^2 + \hat{v}_{22}^2 + \dots + \hat{v}_{2n}^2 = 1)$$

Si $\lambda_1 \neq \lambda_2$ entonces $\hat{v}_1$ y $\hat{v}_2$ son ortogonales.

Así sucesivamente.

Obtenemos

Coordenadas de cada modalidad fila en los nuevos ejes.

Sea λ_1 el mayor valor propio $\neq 0$ obtendremos $\hat{v}_1$, vector propio, $(\hat{v}_{11}, \hat{v}_{12}, \dots, \hat{v}_{1n})$.

$$B_i^{(1)} = \frac{\sqrt{\lambda_1}}{\sqrt{P_i}} v_{1i}$$

Coordenadas de las filas i en el eje 1.

Sea λ_2 el siguiente valor propio $\neq 0$, su vector propio $(\hat{v}_{21}, \hat{v}_{22}, \dots, \hat{v}_{2n})$.

$$B_i^{(2)} = \frac{\sqrt{\lambda_2}}{\sqrt{P_i}} v_{2i}$$

Coordenadas de las filas i en el eje 2.

$$\begin{array}{c|cc|cc}
 & \text{1er eje} & \text{2 eje} & & \\
 \hline
 \text{Coord. de } b_1 & \frac{\sqrt{\lambda_1}}{\sqrt{P_1}} v_{11} & \frac{\sqrt{\lambda_2}}{\sqrt{P_1}} v_{21} & B_1^{(1)} & B_1^{(2)} \\
 \text{Coord. de } b_2 & \frac{\sqrt{\lambda_1}}{\sqrt{P_2}} v_{12} & \frac{\sqrt{\lambda_2}}{\sqrt{P_2}} v_{22} & B_2^{(1)} & B_2^{(2)} \\
 \vdots & \vdots & \vdots & B_i^{(1)} & B_i^{(2)} \\
 \text{Coord. de } b_n & \frac{\sqrt{\lambda_1}}{\sqrt{P_n}} v_{1n} & \frac{\sqrt{\lambda_2}}{\sqrt{P_n}} v_{2n} & B_n^{(1)} & B_n^{(2)}
 \end{array} =$$

Estas coordenadas son las que se representan en la normalización principal. Contribución absoluta de cada modalidad fila a la formación del eje.

$$C_{b_1}^a = \frac{P_1}{\lambda_1} \left(B_1^{(1)} \right)^2$$

$$C_{b_2}^a = \frac{P_2}{\lambda_1} \left(B_2^{(1)} \right)^2$$

$$C_{b_n}^a = \frac{P_n}{\lambda_1} \left(B_n^{(1)} \right)^2$$

La suma de las contribuciones absolutas a un mismo eje es 1.

Contribución de $b_1, b_2, \dots, b_n$ a la formación del 1.er eje siendo $B_1^{(1)}, B_2^{(1)} \dots B_n^{(1)}$ las coordenadas de $b_1, b_2, \dots, b_n$ en ese primer eje.

Para la contribución de las filas a la formación del segundo eje, se cambiarán las coordenadas por $B_1^{(2)}, B_2^{(2)} \dots B_n^{(2)}$ etc.

$$C_{b_1}^a = v_{11}^2$$

$$C_{b_2}^a = v_{12}^2$$

$$C_{b_n}^a = v_{1n}^2$$

$$\left(\sqrt{C_{b_1}^a}, \sqrt{C_{b_2}^a}, \dots, \sqrt{C_{b_n}^a} \right) = (v_{11}, v_{12}, \dots, v_{1n})$$

Contribución relativa de los factores a las modalidades fila.

$$C_r^d = \frac{B_i^2}{d^2(0, i)}$$

Con B_i^2 coordenada al cuadrado de la fila considerada y $d^2(0, i)$ la distancia X^2 de la fila considerad a la media (MASA) de las filas.

Esto es para cada eje.

La suma de dos contribuciones relativas nos dice cómo ha quedado representada esa modalidad con esos dos ejes.

Mod Eje 1 Eje 2

$$b_1 \quad \frac{(B_1^{(1)})^2}{d^2(0,1)} \quad \frac{(B_1^{(2)})^2}{d^2(0,1)}$$

$$b_2 \quad \frac{(B_2^{(1)})^2}{d^2(0,2)} \quad \frac{(B_2^{(2)})^2}{d^2(0,2)}$$

$$b_i \quad \frac{(B_i^{(1)})^2}{d^2(0,i)} \quad \frac{(B_i^{(2)})^2}{d^2(0,i)}$$

Con $B_1^{(1)}, B_2^{(1)}, B_i^{(1)}$ coord. de b_1, b_2, b_i en el eje 1.

Con $B_1^{(2)}, B_2^{(2)}, B_i^{(2)}$ coord. de b_1, b_2, b_i en el eje 2.

Respecto de las columnas:

Coordenadas de las columnas:

$$A_j = \sum \frac{P_{ij}v_i}{\sqrt{P_i P_j}}$$

$$a_1 = (A_1^{(1)}, A_1^{(2)}) = \left(\sum \frac{P_{i1}v_i}{\sqrt{P_i P_{.1}}}, \sum \frac{P_{i2}v_i}{\sqrt{P_i P_{.1}}} \right)$$

$$a_2 = (A_2^{(1)}, A_2^{(2)}) = \left(\sum \frac{P_{i2}v_i}{\sqrt{P_i P_{.2}}}, \sum \frac{P_{i2}v_i}{\sqrt{P_i P_{.2}}} \right)$$

$$a_j = (A_j^{(1)}, A_j^{(2)}) = \left(\sum \frac{P_{ij}v_i}{\sqrt{P_i P_{.j}}}, \sum \frac{P_{ip}v_i}{\sqrt{P_i P_{.j}}} \right)$$

$$a_p = (A_p^{(1)}, A_p^{(2)}) = \left(\sum \frac{P_{ip}v_i}{\sqrt{P_i P_{.p}}}, \sum \frac{P_{ip}v_i}{\sqrt{P_i P_{.p}}} \right)$$

(Para una misma modalidad, las coordenadas en el 1.er y 2.o eje solo cambian los coeficientes $\hat{v}_{ij}$ del vector propio).

(Para distintas modalidades cambia la columna de referencia y su total de P._j). Contribución absoluta de cada modalidad columna a la formación del eje:

$$C_j^a = \frac{P_j}{\lambda} A_j^2$$

	eje 1	eje 2
C_{a1}^a	$\frac{P_1}{\lambda_1} (A_1^{(1)})^2$	$\frac{P_1}{\lambda_2} (A_1^{(2)})^2$
C_{a2}^a	$\frac{P_2}{\lambda_1} (A_2^{(1)})^2$	$\frac{P_2}{\lambda_2} (A_2^{(2)})^2$
$\vdots$	$\vdots$	$\vdots$
C_{aj}^a	$\frac{P_j}{\lambda_1} (A_j^{(1)})^2$	$\frac{P_j}{\lambda_2} (A_j^{(2)})^2$

(La suma de todas las contribuciones absolutas a la formación de un mismo eje es 1).

Contribución relativa de los ejes a las modalidades columna:

$$C_j^r = \frac{A_j^2}{d^2(0,j)}$$

	Eje 1	Eje 2
C_{a1}^r	$\frac{(A_1^{(1)})^2}{d^2(0,j)}$	$\frac{(A_1^{(2)})^2}{d^2(0,j)}$
C_{a2}^r	$\frac{(A_2^{(1)})^2}{d^2(0,j)}$	$\frac{(A_2^{(2)})^2}{d^2(0,j)}$
$\vdots$	$\vdots$	$\vdots$
C_{ar}^r	$\frac{(A_1^{(1)})^2}{d^2(0,j)}$	$\frac{(A_1^{(2)})^2}{d^2(0,j)}$

$d_2(0, j) = (\text{distancia } X^2)^2$ entre el perfil de la columna promedio (sin raíz) y la columna considerada.

(La suma por filas me da la calidad de la representación para la modalidad A_j considerada).

Correspondencias múltiples

Más de dos variables $\longrightarrow$ tabla disyuntiva completa.

Procedimiento de normalización y diagonalización diferente. Hay una inercia que solo depende del número de variables y modalidades.

$$I = \frac{N.\text{modalidades}}{N.\text{variables}} - 1$$

El grado de asociación viene dado por el reparto de esa inercia entre ejes.

Reparto equilibrado $\longrightarrow$ independencia

Reparto desequilibrado $\longrightarrow$ asociación

$N. \text{ max ejes} = N. \text{ de modalidades} - N. \text{ variables}.$

Órdenes en SPSS para análisis factorial de correspondencia (AFC)

A) Para introducir una tabla de contingencia

Abrir un archivo nuevo de datos y definir tres variables: en una se introducirán las modalidades-fila. en otra las modalidades-columna y en la tercera las frecuencias conjuntas. (Por ejemplo, si es una tabla 3x4 el archivo tendrá 12 filas y 3 columnas).

Datos $\longrightarrow$ Ponderar casos $\longrightarrow$ Ponderar mediante la variable frecuencia

(En la parte inferior derecha de la pantalla aparecerá la palabra “ponderado”)

B) Análisis factorial de correspondencias

Analizar $\longrightarrow$ Reducción de datos $\longrightarrow$ Análisis de correspondencias

Introducir en fila: la variable correspondiente a las modalidades-fila

Definir rango (mínimo 1 máximo: n.º de filas de la tabla) actualizar

Introducir en columna: la variable correspondiente a las modalidades-columna

Definir rango (mínimo 1 máximo: n.º de columnas de la tabla) actualizar

En **MODELO**: Cambiamos el método de normalización a Principal

(el resto como está)

En **ESTADÍSTICOS**: añadimos a las opciones marcadas (tabla, inspección de filas, inspección de columnas) los perfiles fila y perfiles columna.

Sin entrar en **GRÁFICOS**: Aceptar (Aparecerá toda la salida excepto gráficos).

Para obtener el gráfico conjunto con todas las modalidades fila y columna:

Repetimos el análisis señalando en **MODELO**: método de normalización simétrica (de forma automática en el apartado **GRÁFICOS** queda señalada la opción diagrama de dispersión biespacial) Aceptar (obtenemos la salida completa, donde únicamente habrán cambiado las coordenadas y aparece un gráfico que hemos de modificar).

Modificación del gráfico: doble clic en el gráfico

Galería → Dispersión Simple → Reemplazar.

Diseño → Línea de referencia ... (para obtener las líneas de los ejes)

Formato → Marcadores (si queremos que las modalidades-fila y las modalidades- columna tengan distinto símbolo, tamaño...)

C) Gráfico de barras

En el menú principal ir a GRÁFICOS → Barras → Agrupado → Definir

Las barras representan: número de casos

Eje de categorías: introducir la variable que deseemos en abscisas

Definir grupos por: la otra variable → Aceptar

Análisis de correlación de Spearman

Ejemplo 1.5

El análisis de correlación de Spearman y el de correspondencia múltiple (ACM), fueron realizados para el ejemplo 1, pero lo discutimos en esta sección para seguir un orden y que contribuya a una mejor comprensión del proceso por parte del lector.

La tabla 13 muestra una matriz con las correlaciones de Spearman entre variables cualitativas.

Tabla 13.

Coefficientes de correlación de Spearman entre seis variables cualitativas evaluadas en 30 híbridos de tomate.

	TCP	FPF	CFM	CCP	CTF	NOL
CTP	1					
PFP	-0.62	1				
CFM	0.23	-0.51	1			
CCP	0.14	0.23	0.09	1		
CTF	0.23	-0.26	0.32	-0.13	1	
NOL	0.08	-0.41	0.35	-0.07	0.33	1

Fuente: López (2014).

TCP = Tipo de crecimiento de la planta. FPF = Forma predominante del fruto. CFM = Color exterior del Fruto maduro. CCP = Color de la carne del pericarpio (interior). CFT = Forma del corte transversal del fruto. NOL = Número de lóculos- **: Altamente significativo al $p < 0.01$ de probabilidad. *: Significativo al $p < 0.05$ de probabilidad.

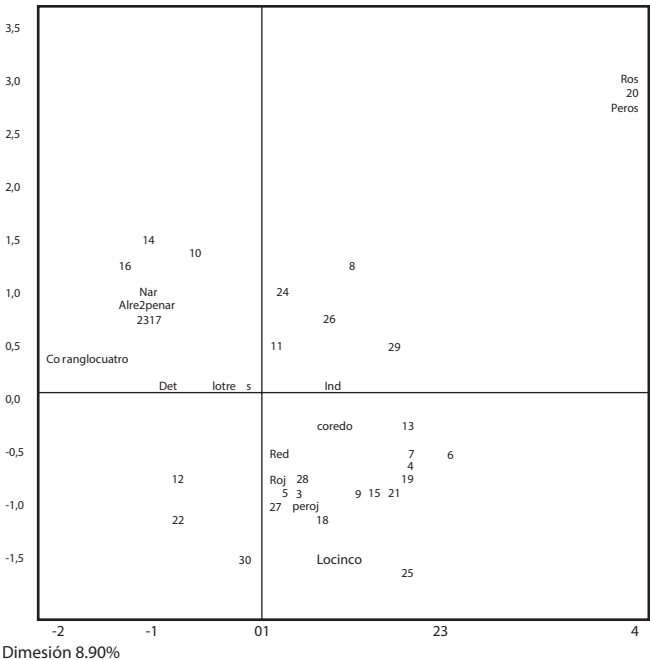
Hubo correlación negativa altamente significativa ($r = -0.62$) entre las variables TCP y FPF (Tabla 13). Esto indicaría que la FPF depende del TCP. En las plantas indeterminadas predominan los frutos redondos y redondos achatados (92 %) y solamente el 5 % fueron redondos alargados. En cambio, en las determinadas el porcentaje de frutos redondos alargados fue mayor (30 %) que en las indeterminadas y el porcentaje de frutos redondos y redondos achatados fue inferior que en las indeterminadas (65 %). Asimismo, se presentaron frutos cordiformes (tipo pera) en un 5 %. La variable FPF se correlacionó negativa altamente significativa con la variable CFM ($r = -0.51$). Esto mostraría que en los frutos alargados predomina el color naranja y en los frutos redondos los colores rojo y rosado. Asimismo, la variable FPF se correlacionó negativa y altamente significativa con la variable NOL ($r = -0.41$). Lo que denota que en los frutos alargados predominan tres lóculos por fruto y que en los frutos redondos predominan cuatro lóculos.

Análisis de correspondencia múltiples (ACP)

El ACM para variables cualitativas permitió explicar la mayor parte de la variabilidad genética que presenta la muestra original. Este método contribuye a explicar la relación existente entre las variables cualitativas heterogéneas conforme a los diferentes estados que presentaron los híbridos evaluados.

El ACM mostró variabilidad genética dentro de los 30 híbridos evaluados de tomate (Figura 4). Aunque la mayoría de los híbridos se agruparon por características similares. hubo casos aislados, como el híbrido 20 que no se agrupó con ninguno de los híbridos. Este híbrido tuvo frutos con color exterior rosado y carne del pericarpio del mismo color (características únicas de este híbrido). Por otro lado, los híbridos 2, 17 y 23 se agruparon al presentar frutos redondos alargados con un color exterior naranja y la carne del pericarpio del mismo color. Otros híbridos que se diferenciaron fueron los híbridos 3, 5, 18, 27 y 28 al presentar forma redonda y un color exterior rojo y la carne del pericarpio también del mismo color. En general, hubo una correlación entre la forma predominante del fruto, el color exterior y el color del pericarpio.

Figura 4. Asociación entre estados de cada una de las variables cualitativas (Fuente: López 2014; López *et al.* 2015). Tipo de crecimiento de las plantas (Ind = Indeterminadas. Det = Determinadas); Forma predominante del fruto (Alre = Redondo alargado. Red = Redondo. Lach = Ligeramente achatado); Color exterior del fruto maduro (Roj = Rojo. Nar = Naranja. Ros = Rosada); color carne del pericarpio (Peroj = Rojo. Penar = Naranja. Peros = rosado); Corte transversal del fruto (Coredo = Redondo, Corang = Angular); Número de lóculos (Lotres = tres lóculos. Locuatro = cuatro lóculos y Locinco = cinco lóculos) y los números son los 30 híbridos de tomate.



Ejemplo 1.6.

Análisis de correlación de Spearman

Este análisis se realizó en base al ejemplo 1.2, para las variables cualitativas, mismo que permitió discriminar en función de la correlación de Spearman (Tabla 14), qué variables están altamente asociadas, para luego interpretar la distribución de estas variables en un eje cartesiano con dos dimensiones.

En la tabla 14 se aprecia que no existen muchas variables altamente asociadas, las principales son las siguientes: la especie de árbol está alta y significativamente asociada a la textura del tallo ($r=0.69$) y al color de flor ($r=0.64$), la forma de hoja tiene asociación altamente significativa con el tipo de inflorescencia ($r=0.69$) y forma de inflorescencia ($r=0.69$), el tipo de inflorescencia tiene asociación altamente significativa con la forma de inflorescencia ($r=0.99$) y la forma del ápice del fruto ($r=0.69$), la forma de inflorescencia tiene asociación altamente significativa con la forma del ápice del fruto ($r=0.60$), la textura del tallo tiene asociación altamente significativa con el color de flor ($r=0.71$) y, finalmente, la forma de semilla del fruto tiene una asociación altamente significativa con el color de la semilla ($r=0.63$).

Correlación de Spearman para variables cualitativas

[illegible]

Análisis de correspondencias múltiples (ACP)

El grado de asociación puede ser apreciado en las dos dimensiones que se describen en la tabla 15, que confirma lo observado en las correlaciones de Pearson obtenidas. Como se observa en la Tabla 15, algunas variables contribuyen a las dos dimensiones (árbol de fruto. forma de hoja y color de flor, y otras solamente a una (componente 1: textura de hoja, forma de inflorescencia, textura de tallo, forma de fruto, forma del ápice del fruto. Componente 2: color interno del fruto y forma de semilla del fruto) y hay algunas que no contribuyen a ninguno.

Tabla 15.

Medidas de discriminación.

Variables	Ponderación de la variable	Dimensión		Media
		1	2	
árbol de fruto	9	.926	.929	.927
forma de árbol	5	.174	.149	.162
forma de hoja	9	.344	.636	.490
tipo de hojas	2	.160	.271	.216
color de hoja	3	.249	.252	.251
textura de hoja	5	.870	.044	.457
color de peciolo de hoja	3	.047	.113	.080
tipo de inflorescencia	2	.075	.502	.289
forma de inflorescencia	9	.133	.614	.373
textura del tallo	4	.862	.033	.447
color de fruto	8	.838	.938	.888
textura del fruto	2	.221	.000	.111
color interno del fruto	7	.884	.252	.568
forma de fruto	6	.390	.178	.284
color interno de fruto	7	.581	.072	.326
color de semilla	4	.836	.372	.604
Total activo		45.734	35.667	40.701

Las ponderaciones de las variables están incorporadas en los estadísticos de Total activo.

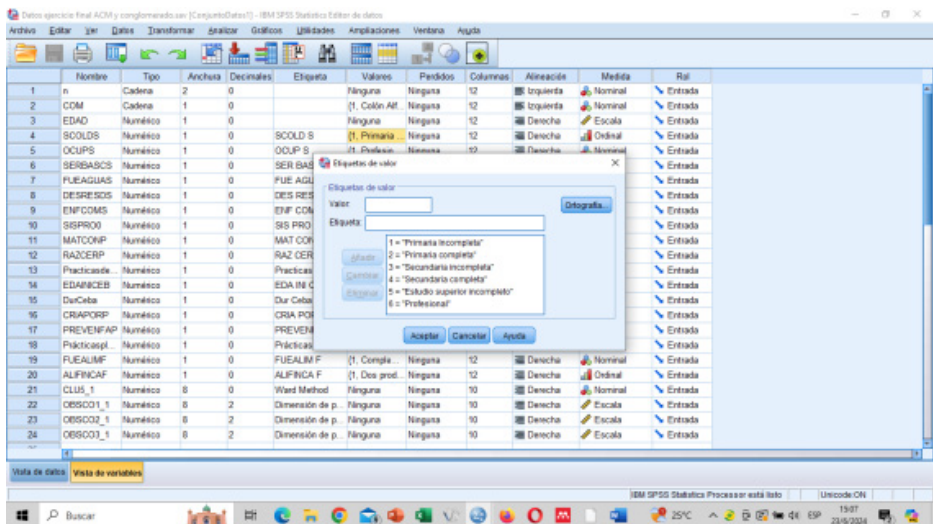
Ejemplo 1.7. Análisis de correspondencias múltiples en un estudio de caracterización socio productiva de cría de cerdos traspatio en comunidades rurales del cantón Jipijapa (Valverde *et al.*, 2021).

Base de datos

Se debe contar con una base de datos organizada y etiquetada, como se observa en la imagen, donde todas las variables deben ser categóricas, algunas nominales como:

Sistemas de producción:

- 1.Intensivos
- 2.Extensivos
- 3.Semi intensivos



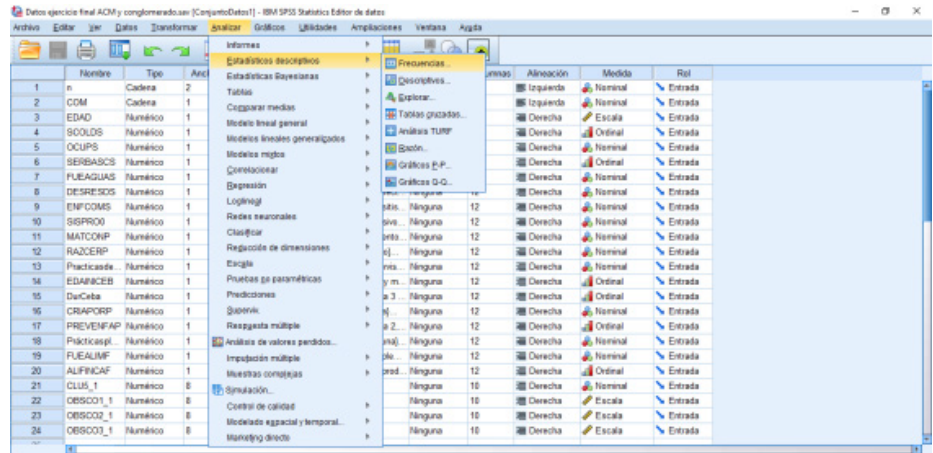
Un ejemplo de variable ordinal sería:

Nivel de estudio

1. Primaria
2. Secundaria
3. Preparación técnica
4. Profesional
5. Posgrado

Análisis de frecuencia

Es recomendado realizar el análisis de frecuencia de los datos, los cuales en todas las categorías mínimo se recomienda que debe haber 20 casos.



Fuente de alimento de la finca

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válido	Complemento con pienso	21	32.31	32.31	32.31
	Alimentos de la finca	20	30.77	30.77	63.08
	Compra alimentos de origen agrícola	24	36.92	36.92	100.00
	Total	65	100.00	100	

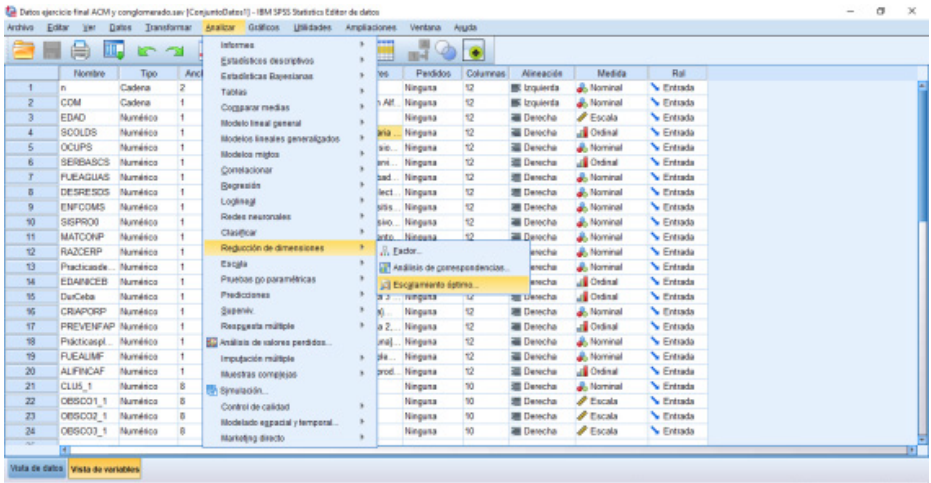
Análisis de correspondencia múltiple

El estudio presentado como ejemplo se realizó en la microrregión sur de la provincia de Manabí. Se tomó como referente, por su alto porcentaje de población rural (60 %), al cantón Jipijapa y la parroquia El Anegado. Ésta se encuentra ubicada geográficamente en 1°10' y 1°047' de latitud sur y 80° 25 y 80° de longitud oeste (Valverde Lucio *et al.*, 2021), con un clima estepa local considerado BSh, una variante del clima subtropical seco y del semi-árido cálido, la temperatura media anual es 23.7 °C, la precipitación promedio es de 537 mm. Se aplicó en 7 comunidades de Jipijapa, la mayoría de la parroquia El Anegado, considerando que por sus características climatológicas, relieve

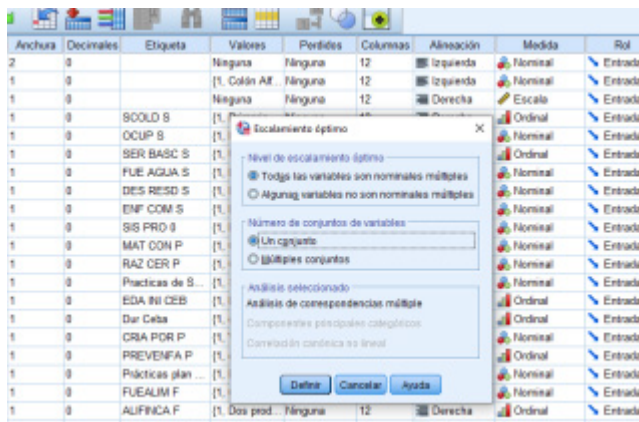
y producción agrícola diversa, provee de insumos para los animales, garantizando así la producción de cerdos durante todo el año.

La caracterización se efectuó bajo cuatro dimensiones: Social, que permitió identificar: edad, escolaridad, servicios básicos, tiempo de dedicación, afectaciones en salud; Organizativa, que determinó la cultura organizativa, el tipo de ayudas por parte de entes de desarrollo gubernamentales o no gubernamentales; Productiva, que dio lugar a caracterizar el sistema productivo, las razas empleadas, materiales de construcción de las porquerizas, número de animales, tiempo de cría, razón de crianza, peso, precio, registros productivos, planificación, autoconsumo, comercio, entre las destacadas; y los Recursos locales con los que cuenta el productor para la cría de sus animales, como insumos alimenticios locales, asistencia técnica, preparación o no de alimentos y equipos. Se diseñó para tal efecto una encuesta estática con preguntas cerradas sobre los ámbitos anteriormente referidos, aplicando la metodología operacionalización de las variables de estudio y respuesta.

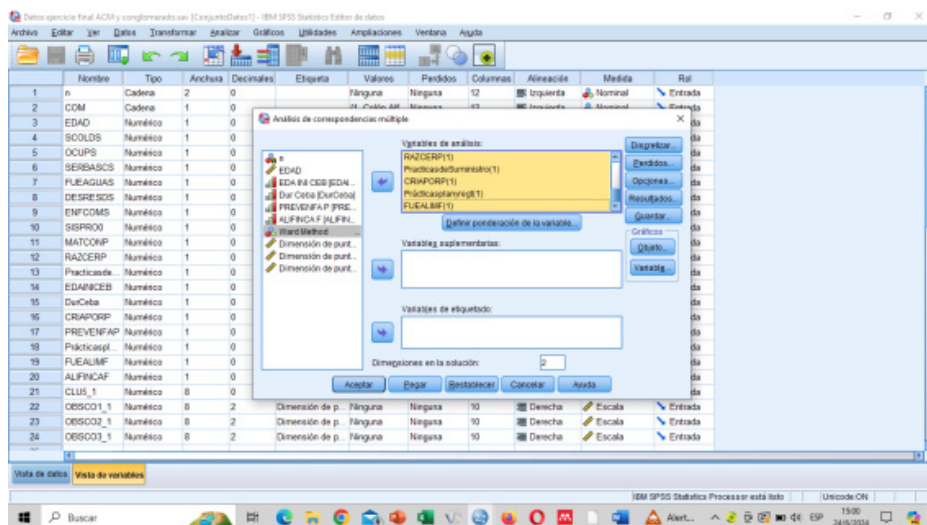
El procedimiento inicia con Analizar – Reducción de dimensiones – Escalamiento óptimo.



En la ventana nivel de Escalamiento óptimo, se selecciona “Todas las variables son nominales múltiples”; y en el número de conjuntos de variables se escoge “Un conjunto”. Con estos pasos queda seleccionado el análisis de correspondencia múltiple, y posteriormente se define.



El siguiente paso da lugar a la toma de decisiones con respecto al análisis propiamente dicho. Se inicia seleccionando las variables que van a ser parte del análisis y se las incorpora a la ventana “Variables de análisis”, se observará que a cada variable incorporada al final se le agrega, entre paréntesis (1), que significa que todas las variables cuentan con el mismo peso e importancia.

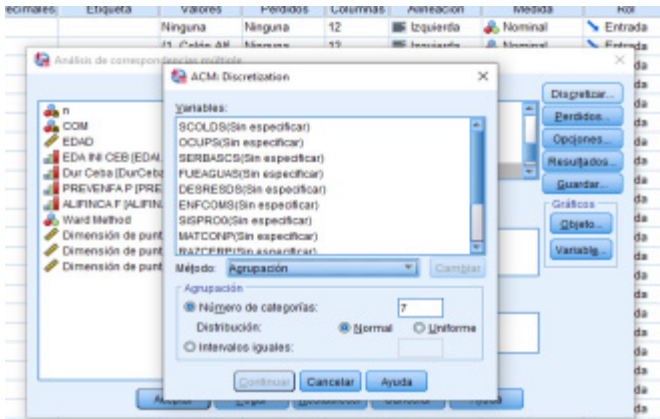


“Variables suplementarias”, en nuestro caso no ingresamos este tipo de variables, más, sin embargo, es una opción para quienes consideren agregar variables que, aunque no sean parte del modelo consideran oportuno analizarlas.

“**Variables de etiqueta**”; se aplica cuando se trabaja con pocos casos.

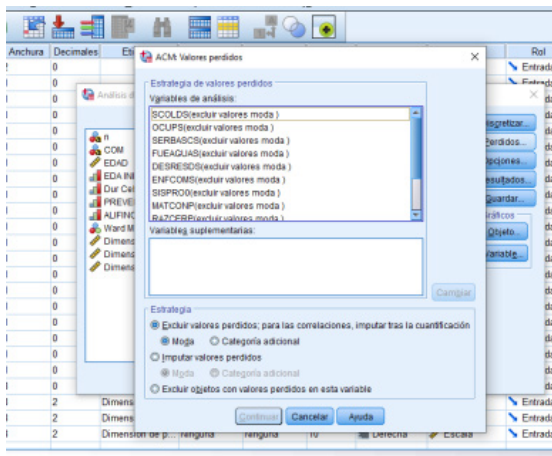
“**Dimensiones de la solución**”; se establece como valor predeterminado el 2, se recomienda no cambiarlo, salvo que sea necesario, porque al final del resultado sí se observa poca variabilidad.

El “**Discretizar**” permite transformar variables continuas en discretas, en nuestro caso no aplica.

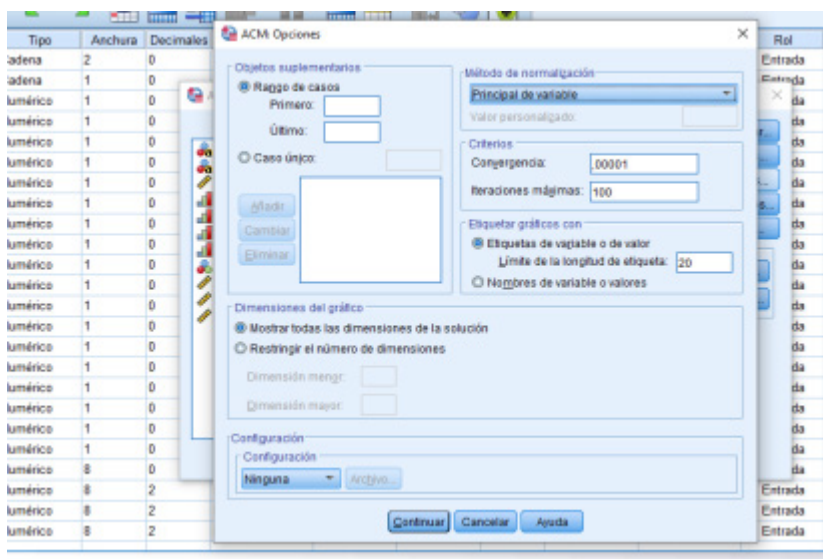


“**Valores perdidos**”; en este acápite se presentan tres opciones, excluir valores perdidos para las correlaciones, imputar tras la cuantificación; imputar valores perdidos; excluir objetos con valores perdidos en esta variable.

En este caso se deja la opción predeterminada, que es la primera “Excluir valores perdidos; para las correlaciones, imputar tras la cuantificación”.

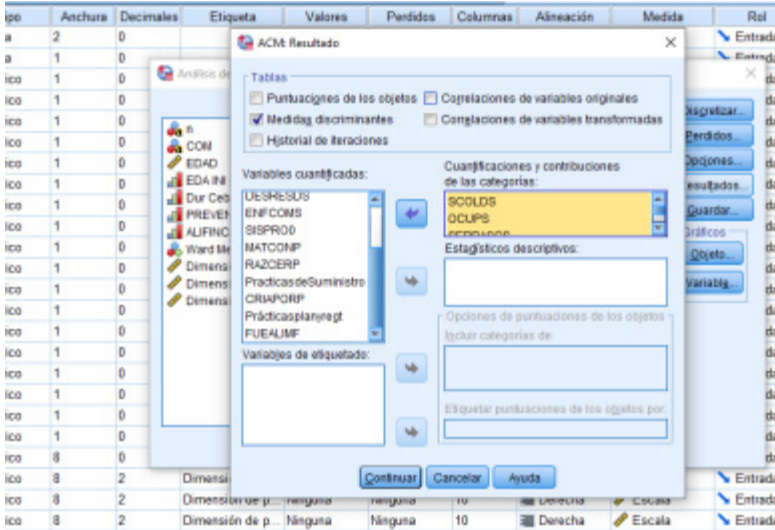


“Opciones”, en nuestro caso recomendamos lo predeterminado en cada uno de los aspectos citados, en Objetos suplementarios, dejar tal cual el rango de casos. En el método de normalización, se debe asumir la opción principal por variable, si bien es cierto se pueden aplicar otras opciones, en nuestro caso está bien la opción señalada. En criterios, al igual que en los casos anteriores, se asume lo predeterminado. En la opción Etiquetar gráficos con, se recomienda dejar la predeterminada, que da la opción máxima que es 20. En cuanto a las dimensiones de gráfico, se recomienda asumir la opción mostrar todas las dimensiones de la solución, se podría cambiar de tener muchas dimensiones, en este caso solo hemos seleccionado 2. En configuración se escoge ninguna.



En el acápite “Resultados”, en la opción tablas se selecciona las medidas discriminantes, que sería la opción más importante del estudio. En cuantificaciones y contribuciones de las categorías, se deben ingresar todas las variables en estudio.

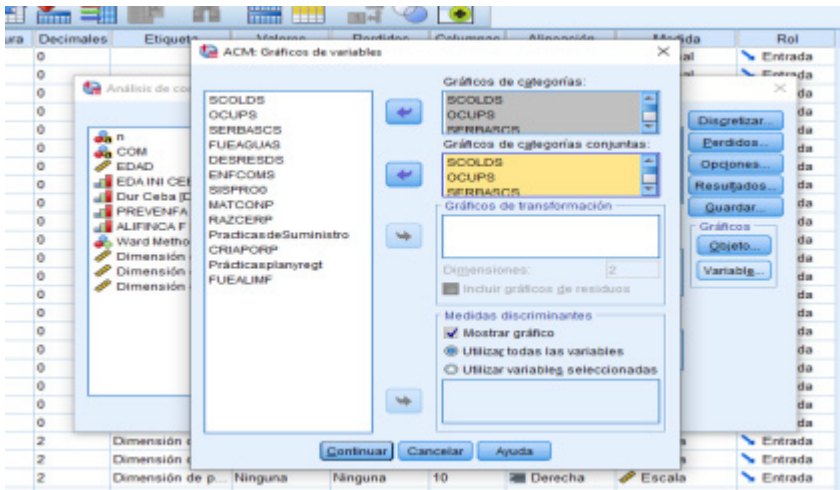
La opción “estadísticas descriptivas” no se aplica, pues se realizó al inicio como opción de justificación del planteamiento del análisis multivariante.



La opción “Guardar” se aplica solo cuando se desea guardar el ejercicio definitivo, y de esta manera se crea una nueva columna en la base de datos.

La opción final “Gráficos”, despliega dos opciones: Objetos y variables. La opción Objetos no se aplica en este caso.

Con respecto a la opción “Variable”, se ingresan todas las variables en las ventanas Gráficos de categorías y Gráficos de categorías conjuntas, este ejercicio ayuda en la interpretación de los gráficos, especialmente cuando se están analizando muchas variables.



Interpretación de resultados

Resumen de procesamiento de casos				
Casos activos válidos	55			
Casos activos con valores perdidos	0			
Casos complementarios	0			
Total	55			
Casos utilizados en análisis	55			
En el resumen se aprecia la cantidad de casos válidos.				
Historial de interacciones				
Número de iteración	Varianza contabilizada para Pérdida			
Número de iteración	Total	Aumentar		
27a	4.49	,000	8,51	
a. El proceso de iteración se ha detenido porque se ha alcanzado el valor de prueba de convergencia.				
Se observa que la iteración se paraliza en 27, señala el punto donde la varianza deja de ser significativa.				
Resumen del modelo				
Dimensión	Alfa de Cronbach	Varianza contabilizada para		
		Total (autovalor)	Inercia	% de varianza
1	0,868	5.041	0.388	38.780
2	0,808	3.940	0.303	30.310
Total		8.982	0.691	
Media	0,842a	4.491	0.345	34.545
a. La media de alfa de Cronbach se basa en la media de autovalor.				

En el resumen del modelo, se pueden apreciar las dos dimensiones seleccionadas en la aplicación del ejercicio, y se puede observar que la dimensión 1 con 0.388 tiene mayor inercia que la dimensión 2, por lo tanto, tiene una explicación mayor de la varianza. El alfa de Cronbach señala cuán correlacionadas están las variables observables que son parte de las dimensiones, el comportamiento del alfa de Cronbach con la inercia debe estar relacionado.

Cuantificaciones

Las tablas de cuantificaciones muestran la ubicación de las coordenadas de los ejes X e Y, pero, además, señalan la relación de la variable con cada dimensión; el puntaje más alto es el indicador de relación.

ESCOLARIDAD			
Puntos: Coordenadas			
Categoría	Frecuencia	Coordenadas del centroide	
		Dimensión	
		1	2
Primaria incompleta	10	-.251	-.398
Primaria completa	25	-.415	-.168
Secundaria incompleta	2	.373	-.429
Secundaria completa	13	.281	.469
Estudio superior incompleto	4	1.624	.420
Profesional	1	1.984	1.273

En el ejemplo se aprecia que “Primaria incompleta” está más relacionada a la dimensión dos, que primaria completa guarda mayor relación con la dimensión uno, y así sucesivamente.

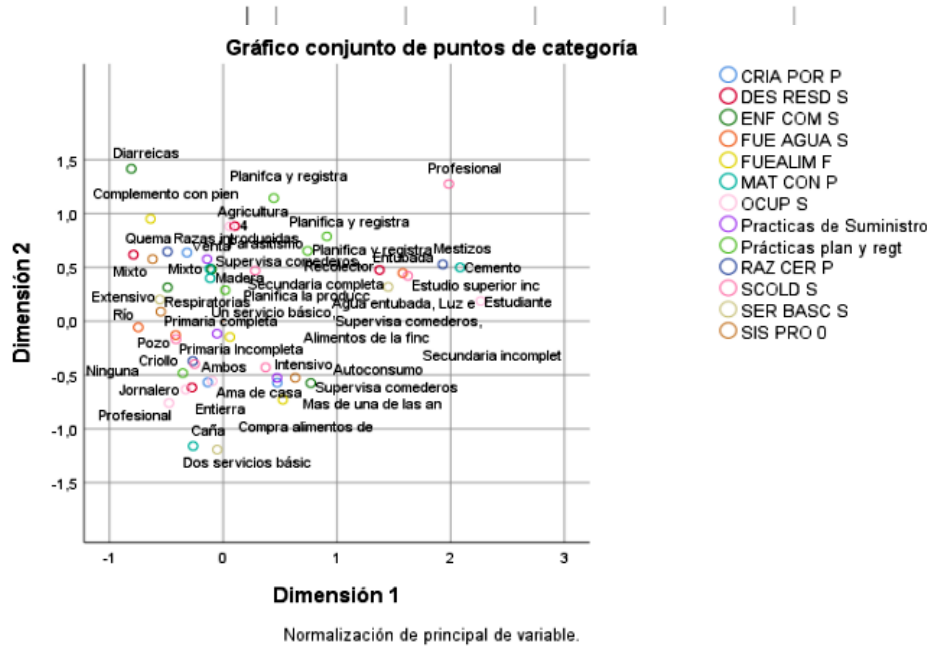
Medidas discriminantes

La tabla de medidas discriminantes determina cuando discrimina cada variable y cuán importante es en cada dimensión. Así tenemos la dimensión uno, que explica el 38,78 % de la varianza y que representa la importancia social de la producción porcina en los sectores rurales, las variables involucradas son: servicios básicos, fuentes de agua, manejo de desechos, razas de cerdo, sistemas de producción y enfermedades comunes.

La segunda dimensión explica el 30,31 % de la varianza, se refiere a la importancia del sistema de producción, destacan variables como: ocupación de la familia en la actividad de cría, materiales de construcción de las porquerizas, fuentes de alimentación, además las variables ocupaciones, razón de cría y prácticas de plan y registro, que también expresaron importancia.

	Dimensión		Media
	1	2	
Escolaridad	.377	.143	.260
Ocupación familiar	.326	.511	.419
Servicios básicos	.684	.281	.483
Fuentes de agua	.701	.056	.378
Destino de residuos	.586	.376	.481
Enfermedades comunes	.374	.227	.300
Sistema de producción	.387	.281	.334

Materiales construcción P	.344	.553	.449
Razas de cerdos	.643	.224	.433
Prácticas de suministro	.039	.133	.086
Razón de cría	.133	.363	.248
Prácticas plan y registro	.215	.296	.256
Fuentes de alimentación F	.234	.496	.365
Total, activo	5.041	3.940	4.491
% de varianza	38.780	30.310	34.545



El gráfico de conjuntos de puntos por categoría muestra la ubicación dentro del mapa de cada una de las variables. En este gráfico se puede apreciar la inercia que tienen las variables para con las dos dimensiones. Además, se evidencia el comportamiento de los ejes de Y así como de X y cómo están relacionadas con las dimensiones 1 y 2.

1.2.3. Análisis de conglomerados (clúster)

Es un método analítico que se puede aplicar para clasificar las accesiones de un germoplasma (o variables) en grupos relativamente homogéneos con base en alguna similitud existente entre ellas (Franco e Hidalgo, 2005).

El objetivo es reunir las observaciones en grupos (clúster) en función de su semejanza. Se trata de que las observaciones semejantes entre sí se hallen en un mismo grupo (homogeneidad interna) y los grupos difieran unos de otros (heterogeneidad entre grupos).

El método de conglomerados o análisis de clúster se puede aplicar sobre una matriz básica de datos $n \times p$ o sobre una matriz $n \times n$, o $p \times p$, donde n es el número de accesiones que se quieren agrupar y p son las variables.

Es importante aclarar que el análisis de conglomerados se aplica sobre una matriz de distancias y no sobre una de similitud. Para descriptores cualitativos, esta última debe ser transformada en una de distancia. Para datos cuantitativos, los programas de estadística actualmente disponibles calculan directamente los valores de distancia según el método que se aplique.

Básicamente, los métodos de agrupamiento más usados en el análisis de conglomerados son: 1) jerárquico, que forma grupos a varios niveles; y 2) no jerárquico o de partición, que también forma grupos a través de criterios predefinidos (Franco e Hidalgo, 2005).

Relación con otras técnicas

- comparte con las técnicas factoriales la idea de reducción, solo que ahora se trata de agrupar observaciones (filas) y no variables (columnas)
- comparte con el análisis discriminante la idea de clasificar, solo que ahora los grupos no existen a priori (no hay una variable que indique la pertenencia de las observaciones a grupos), sino que los grupos se van a construir a partir de la semejanza entre observaciones respecto a las variables disponibles.

Procedimientos

Existen distintos métodos de agrupación que se clasifican en: técnicas jerárquicas y técnicas de optimización.

En las técnicas jerárquicas los grupos se van formando en sucesivas etapas por unión (caso ascendente) o por división (caso descendente) de grupos ya existentes, de modo que si dos observaciones se han unido (caso ascendente) o se han separado (caso descendente), permanecerán en ese mismo estado (unidas o separadas) hasta el final del proceso.

En las técnicas de optimización se fija una función a optimizar (por ejemplo el cociente entre la variabilidad externa y la interna) y generalmente mediante procedimientos iterativos se van ubicando y reubicando las observaciones

en distintos grupos. En cada etapa se mide la “mejoría” experimentada por la función a optimizar y, cuando la mejoría no supera cierto umbral prefijado, el proceso se detiene.

Tanto en unas como en otras existen distintas variantes que pueden producir resultados diferentes.

Aspectos teórico-prácticos a considerar

1. La influencia de las unidades de medida de las variables en los resultados. Se aconseja, particularmente cuando las variables se miden en unidades diferentes unas de otras, realizar el proceso con variables tipificadas.
2. La estabilidad de la solución. Valorar los cambios que se producen en el resultado final (es decir, qué observaciones pertenecen a cada grupo) al introducir cambios (distintas variantes) en el proceso. Los grupos son más consistentes (podríamos decir que están geométricamente más diferenciados) cuando, pese a utilizar distintas variantes en el procedimiento de agrupación, el resultado final es similar.
3. Comparación entre diferentes soluciones (que posean el mismo número de grupos):
 - a. a)utilizando solamente las variables del análisis: la mejor es aquella en la que el cociente entre la variabilidad externa y la interna sea máximo (L de Wilks mínima).
 - b. b)utilizando otras variables auxiliares distintas a las del análisis: puede considerarse mejor aquella solución donde la variable auxiliar se comporte de forma más diferenciada en los grupos.
4. Número de grupos en la solución final: no hay reglas fijas. Puede depender del criterio del analista. Si nos basamos únicamente en las variables del análisis, se pueden utilizar criterios basados en la minimización de la variabilidad interna: siempre que aumentemos el número de grupos ocurre que la variabilidad interna disminuye, con lo cual merecerá la pena considerar $k+1$ grupos en lugar de k si la disminución que se consigue es superior a un cierto umbral. También podemos guiarnos por la evolución del proceso y detenerlo cuando se produzca un “salto” importante (se ve en la salida de ordenador).
5. Caracterización de los grupos. Tan importante como identificar los grupos es caracterizarlos, observar cómo son. Para ello se pueden

calcular las medias de las variables utilizadas en el análisis en cada grupo y representarlas gráficamente para ver su perfil.

6. En ocasiones, y para evitar que tengan excesiva influencia en el resultado varias variables muy correlacionadas entre sí, puede ser conveniente aplicar componentes principales y, con los factores resultantes, realizar el clúster.

Descripción del proceso en el caso jerárquico ascendente

En este caso se considera que cada una de las N observaciones es un clúster y el proceso consiste en ir uniendo dichas observaciones en grupos, hasta llegar a configurar un número de grupos fijado de antemano por el analista (por ejemplo, 2, 3 o 4 grupos) o bien observar el proceso hasta el final (cuando todas las observaciones configuran un único grupo) para después decidir el número de grupos a considerar como mejor solución.

Dado que vamos buscando observaciones semejantes (en las variables consideradas) hay una primera decisión a tomar: ¿qué medida de semejanza entre observaciones se va a elegir?

Cuando las variables son cuantitativas. las más usadas son:

- Distancia euclídea
- Distancia euclídea al cuadrado
- Distancia de bloques

Con la medida elegida se construye la matriz de distancias entre observaciones (matriz $N \times N$, simétrica y con ceros en la diagonal principal).

Observando esta matriz, encontraremos la pareja de observaciones “más parecidas”. Son las que dan lugar a la primera unión, al primer clúster.

En ese momento disponemos de $N-2$ observaciones y un clúster con dos observaciones. Para continuar el proceso, debemos tomar otra decisión ¿qué medida de distancia entre clúster (vinculación) se va a utilizar?

Dentro de las más utilizadas están:

Vinculación vecino más próximo (simple)

$D_{AB} = \min d_{ij}$ (la menor de las distancias entre pares de observaciones).

Donde i representa cualquier observación del clúster A; j representa cualquier observación del clúster B

Vinculación vecino más lejano (completa)

$D_{AB} = \max d_{ij}$ (la mayor de las distancias entre pares de observaciones).

Donde i representa cualquier observación del clúster A; j representa cualquier observación del clúster B.

Vinculación inter-grupos (promedio)

$D_{AB} = S d_{ij} / n_A n_B$ (la media aritmética de las distancias entre pares de observaciones).

Donde i representa cualquier observación del clúster A; j representa cualquier observación del clúster B; n_A y n_B representan los tamaños ($n.^\circ$ de observaciones) de los clústeres A y B, respectivamente.

Una vez elegida la vinculación, procedemos a construir una nueva matriz de distancias, en este caso $N-1 \times N-1$ (que será simétrica, con ceros en la diagonal principal), cuyos elementos habrán sido calculados de acuerdo a la vinculación elegida. De esta matriz se localiza el menor valor, que da lugar a la segunda unión.

Y se reitera el proceso. La vinculación elegida se ha de mantener durante todo el proceso y las d_{ij} se toman siempre de la primera matriz de distancias.

Ejemplo

Se dispone de 6 observaciones a, b... de las que se han medido dos variables X_1 y X_2 . Realizar el proceso de agrupamiento, utilizando la distancia euclídea al cuadrado y la vinculación promedio (en este ejemplo se realizará sin tipificar).

Datos

Observaciones	X_1	X_2
A	1	6
B	2	6.5
C	2	7
D	2.5	3
E	6	4
F	9	6

Cálculo de distancias (euclídea al cuadrado):

$d_{ab} = (1-2)^2 + (6-6.5)^2 = 1.25$, lo mismo las demás.

Primera matriz de distancias

	a	b	c	d	e	f
a	0	1.25	2	11.25	29	64
b		0	0.25	12.5	22.25	49.25
c			0	16.25	25	50
d				0	13.25	51.25
e					0	13
f						0

El primer clúster lo forman b y c a la distancia de 0.25.

Segunda matriz de distancias

	a	b	c	d	e	f
a	0	1.625	2	11.25	29	64
bc		0	0.25	14.375	23.625	49.625
d				0	13.25	51.25
e					0	13
f						0

Con la vinculación promedio inter-grupos (solo cambiarán respecto a la primera matriz las distancias respecto al grupo bc, en cursiva):

$d_{bc,a} = (d_{ab} + d_{ac})/2 = (1.25+2)/2 = 1.625$, las demás igual. Por lo tanto, en esta 2.^a etapa se une “a” al grupo bc a la distancia de 1.625.

Tercera matriz de distancias

	abc	d	e	f
abc	0	13.333	13.333	54.417
d		0	0	51.25
e			13.333	13
f				0

Donde $13.333 = (11.25 + 12.5 + 16.25)/3$, aquí se han promediado tres distancias. En esta etapa se unen e y f a la distancia 13.

Cuarta matriz de distancias

	abc	d	ef
abc	0	13.333	¿?
d		0	¿?
ef			0

Continuar el ejercicio hasta unir las seis observaciones.

ANÁLISIS MULTIVARIANTE Y NO PARAMÉTRICO:

Casos de estudio aplicados a
experimentos agropecuarios

Capítulo 2

Nociones sobre estadística
no paramétrica

2.1 Introducción

Esta unidad nueva, fue escrita considerando la necesidad de los investigadores de utilizar técnicas de análisis no paramétrico, esto debido a que en la realidad, en muchas investigaciones, los datos de las variables de estudio no siempre cumplen los requisitos claves de una distribución normal, homocedasticidad, independencia y aleatoriedad. Es en este sentido que se quiso contribuir con la aplicación de algunas técnicas de estadística no paramétrica para análisis de este tipo de datos. De ninguna manera se pretende sustituir a ninguna obra existente, sino al contrario, dar a los lectores herramientas de análisis acordes a las investigaciones que realizan.

Berlanga y Rubio (2012) mencionan que en el ámbito de las ciencias sociales es habitual el uso de pruebas no paramétricas, puesto que existen muchas variables que no siguen las condiciones de parametricidad. Pero no solo en las ciencias sociales ocurre esto, sino también en las ciencias naturales y otras. Las condiciones de parametricidad se refieren al uso de variables cuantitativas continuas, distribución normal de las muestras, varianzas similares (homogéneas u homocedasticidad), independencia, aleatorización y tamaño de las muestras, mayores a 30 casos. Estos criterios se recogen ampliamente en Rubio y Berlanga (2012). En caso de que no se cumplan estos requisitos, y sobre todo cuando la normalidad de las distribuciones de la variable en estudio esté en duda y el tamaño de la muestra sea menor a 30 casos, el empleo de las pruebas no paramétricas o de distribución libre sería lo más indicado.

Las pruebas no paramétricas reúnen las siguientes características: 1) son más fáciles de aplicar; 2) son aplicables a los datos jerarquizados; 3) se pueden usar cuando dos series de observaciones provienen de distintas poblaciones; 4) son la única alternativa cuando el tamaño de muestra es pequeño, y 5) son útiles a un nivel de significancia previamente especificado.

2.2. Clasificación de las principales pruebas no paramétricas

Ferrán (2002) las agrupa en contrastes para una muestra y el resto en no paramétricas. Por su parte Visauta (2007) engloba todas las pruebas en no paramétricas, mientras que otros autores las clasifican por tipo de muestra. También se aprecia como cada autor recoge un número diferente de pruebas no paramétricas, así como el uso de terminología también diferente para nombrarlas. Se ha tratado de recoger las aportaciones de cada uno de estos autores para ofrecer una clasificación completa de las pruebas no paramétricas (17 en total). En su clasificación se han asumido dimensiones relacionadas con el número de muestras y la relación o independencia entre esas muestras

tal y como muestran las tablas 16 y 17. No se han incluido las correlaciones no paramétricas a pesar de que determinados autores las incluyen dentro de pruebas no paramétricas (Pérez Juste *et al.*, 2009), con el objetivo de recopilar solamente los contrastes no paramétricos.

Tabla 16.

Resumen de las principales pruebas estadísticas no paramétricas.

Variable de- pendiente	Una muestra (Bondad de ajuste)	Muestras relacionadas		Muestras independientes	
		2 muestras	>2 muestras	2 muestras	>2muestras
Nominal	Binomial Chi-Cuadra- do Rachas	McNemar	Cochram	Rachas de Wald-Wol- fowitz U de Mann-Whit- ney	Mediana Kruskal-Wallis Jonckhee- re-Tepstra
Ordinal/inter- valo	Kolmoro- gov-Smirnov	Signos Wil- coxon	Friedman Kendall	Moses Kolmoro- gov-Smirnov	

Fuente: Berlanga Silvente y Rubio Hurtado (2012).

Se menciona que las pruebas no paramétricas son alternativas a las pa-ramétricas, por ello se consideró apropiado establecer la equivalencia entre ambas, tal y como se muestra en la Tabla 17 (Berlanga y Rubio, 2012).

Tabla 17.

Pruebas paramétricas y su alternativa no paramétrica.

Muestra	Prueba paramétrica	Prueba no paramétrica
Muestras relacionadas		
2 muestras	t-Student	Wilcoxon
> 2 muestra	ANOVA	Friedman
Muestras independientes		
2 muestras	t-Student	U de Mann-Whitney
> 2 muestra	ANOVA	Kruskal-Wallis

Fuente: Berlanga y Rubio (2012).

En la Tabla 18 se recopilan las características principales de 17 métodos de las 17 pruebas no paramétricas halladas (Berlanga y Rubio, 2012).

Tabla 18.

Pruebas no paramétricas.

Pruebas	Una muestra	Variables
Prueba de Chi-cuadrado de Pearson	Es una prueba de bondad de ajuste, que permite averiguar si la distribución empírica de una variable categórica se ajusta o no (se parece o no) a una determinada distribución teórica (uniforme, binomial, multinomial, etc.).	VD: Nominal
Prueba Binomial	Es una prueba de bondad de ajuste, que permite averiguar si una variable dicotómica sigue o no un determinado modelo de probabilidad. Permite contrastar la hipótesis de que la proporción observada de aciertos se ajusta a la proporción teórica de una distribución binomial (lo cual se traduce en la posibilidad de contrastar hipótesis sobre proporciones y sobre cuartiles).	VD: Nominal
Prueba de Rachas	Es una prueba de independencia o de aleatoriedad, que permite determinar si el número de rachas (R) observado en una determinada muestra de tamaño n es lo suficientemente grande o lo suficientemente pequeño para poder rechazar la hipótesis de independencia (o aleatoriedad) entre las observaciones. Una racha es una secuencia de observaciones de un mismo atributo o cualidad. Una serie de datos en los que hay muchas o pocas rachas permite concluir que estas no han ocurrido por azar.	VD: Nominal
Prueba de Kolmogorov-Smirnov (K-S)	Es una prueba de bondad de ajuste, que sirve para contrastar la hipótesis nula de que la distribución de una variable se ajusta a una determinada distribución teórica de probabilidad que puede ser con tendencia a la normal, a la de Poisson o exponencial.	VD: Ordinal/ Intervalo
Pruebas	Dos muestras relacionadas	Variables
Prueba de McNemar	Sirve para contrastar hipótesis sobre igualdad de proporciones. Se usa cuando hay una situación en la que las medidas de cada sujeto se repiten, por lo que la respuesta de cada uno de ellos se obtiene dos veces: una vez antes y otra después de que ocurra un evento específico.	VI: Dicotómica VD: Nominal

Prueba de los signos	Permite contrastar la hipótesis de igualdad entre dos medianas poblacionales. Puede ser usada para saber si una variable tiende a ser mayor que otra.	VI: Dicotómica VD: Ordinal/ Intervalo
	También es útil para probar la tendencia que sigue una serie de variables ordinales positivas, o para una valoración rápida de un estudio exploratorio.	
Prueba de Wilcoxon	Permite contrastar la hipótesis de igualdad entre dos medianas poblacionales. Paralela a la prueba paramétrica de contraste t para muestras relacionadas.	VI: Dicotómica VD: Ordinal/ Intervalo
Pruebas	K-muestras relacionadas	Variables
Prueba de Friedman	Es una extensión de la prueba de Wilcoxon para incluir datos registrados en más de dos periodos de tiempo o grupos de tres o más sujetos pareados, con un sujeto de cada grupo que ha sido asignado aleatoriamente a una de las tres o más condiciones.	VI: Politómica VD: Ordinal/ Intervalo
Prueba de Cochran	Esta prueba es idéntica a la prueba de Friedman, pero se aplica cuando todas las respuestas son binarias. La Q de Cochran prueba la hipótesis de que varias variables dicotómicas que están relacionados entre sí, tienen el mismo promedio. En observaciones múltiples las variables son medidas en el mismo individuo o en individuos pareados. Tiene la ventaja de examinar cambios en las variables categóricas.	VI: Dicotómica VD: Ordinal/ Intervalo
Coeficiente de concordancia de W de Kendall	Tiene las mismas indicaciones que la prueba de Friedman, aunque se usó en investigación, ha sido, principalmente, para conocer la concordancia entre rangos, más que para probar que existe una diferencia entre las medianas.	VI: Dicotómica VD: Ordinal/ Intervalo
Pruebas	K-muestras independientes	Variables
Prueba de la Mediana	Contrasta diferencias entre dos o más grupos en relación con su mediana, bien porque no cumplen las condiciones de normalidad para usar el promedio como medida de tendencia central, bien porque la variable es cuantitativa discreta. Es similar a la prueba Chi-cuadrado.	VI: Politómica VD: Ordinal/ Intervalo

Prueba de Jonckheere-Terpstra	Es más potente que sus homónimas, la prueba de Kruskal-Wallis y la de la mediana, cuando existe una ordenación a priori (ascendente o descendente) de las K poblaciones de las que se extraen las muestras.	VI: Politémica VD: Ordinal/Intervalo
Prueba H de Kruskal- Wallis	Es una extensión de la de U de Mann-Whitney y representa una excelente alternativa al ANOVA de un factor completamente aleatorizado.	VI: Politémica VD: Ordinal/Intervalo

Fuente: Berlanga y Rubio (2012)

2.2.1. Problemas agropecuarios

A continuación, se resolverán problemas agropecuarios con las principales pruebas no paramétricas, para que los lectores tengan una referencia de cómo hacer estos análisis prácticos e interpretarlos (<https://www.studocu.com/ec/document/universidad-de-las-americas-ecuador/bioestadistica-ii/ejercicios-de-pruebas-no-parametricas-resuelto/5654317>).

2.2.1.1. Caso de una muestra

2.2.1.1.1. Prueba binomial

Se desea conocer si en una población dada de animales salvajes, la relación entre la proporción de machos y hembras es de 1:1, es decir, si $P = Q$. Se asume la distribución binomial de los sexos (Wiedenhofer, 2013). Se ha tomado una muestra aleatoria de 24 animales ($n = 24$). En la tabla siguiente se muestran los resultados

Tabla 19.

Distribución sexual de una muestra de 24 animales salvajes.

Sexo	Machos	Hembras	Total
Frecuencia	15	9	24

Fuente: Wiedenhofer (2013)

Hipótesis

H0: $p=q$

H1: $p \neq q$

Estadístico de prueba: se utiliza la Prueba Binomial, ya que los datos están en dos clases o categorías (machos y hembras).

Se asume que las observaciones son independientes (el sexo de un animal no depende del sexo del otro) y que las probabilidades son constantes, es decir, que la probabilidad de que un animal sea macho o hembra es constante para todos los animales (antes de nacer, por supuesto). a) Nivel de significación: se utilizará un nivel de significación de 5 % ($\alpha = 0.05$). b) Región crítica o área de rechazo: se rechaza la H_0 si el valor de x o $(n - x)$, (el que sea el mayor) es mayor que el tabulado en la binomial, para dos colas, $\alpha = 0.05$ y $n = 24$. Se rechaza la H_0 si x es mayor que 18

Cálculo. Se tomará como clase de interés la de los machos, cuya frecuencia (x) es de 15, mayor que la de las hembras ($n - x$) que es 9. Se tiene que $n = 24$ y $x = 15$

Decisión: como $x = 15$ no es mayor que 18, el valor crítico, no se rechaza H_0 .

Conclusión: la proporción de machos y hembras en la población dada de animales salvajes son iguales (no son distintas)

2.2.1.2. Caso de dos muestras independientes

2.2.1.2.1. Prueba de U de Mann-Whitney

Se realiza un estudio para investigar el efecto de la presencia de una gran planta industrial sobre la población de invertebrados en un río que atraviesa la planta. Se tomaron muestras de siete especies de invertebrados, donde se contó el número de individuos en dos zonas del río: antes de la planta "Aguas arriba" y después de la planta "Aguas abajo". Los datos se muestran en la siguiente tabla:

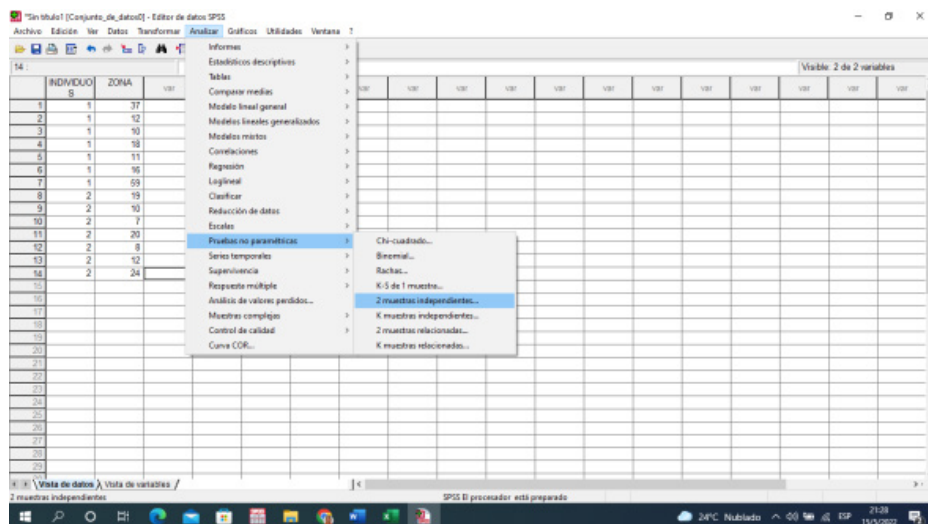
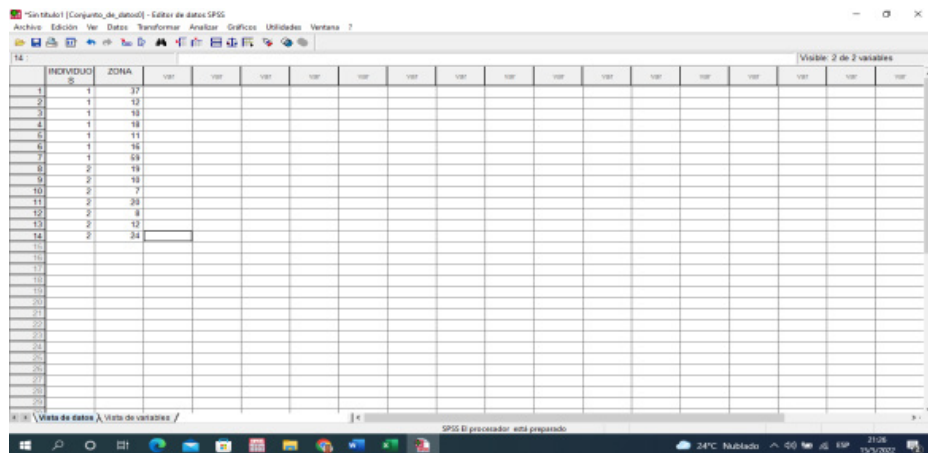
Especie							
Zonas	1	2	3	4	5	6	7
1. Aguas arriba	37	12	10	18	11	16	59
2. Aguas abajo	19	10	7	20	8	12	24

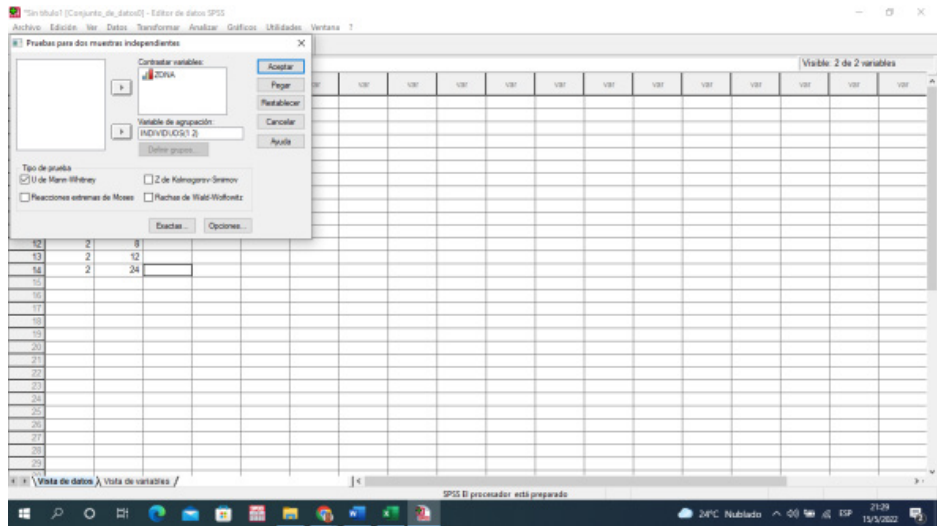
Formular la hipótesis

H₀: No hay efecto de la planta industrial en las especies entre aguas arriba y aguas abajo.

H₁: Sí hay efecto de la planta industrial en las especies entre aguas arriba y aguas abajo.

Procedimiento SPSS





Prueba de Mann-Whitney

Rangos

	Individuos	N	Rango Promedio	Suma de rangos
Zona	1	7	8.43	59.00
	2	7	6.57	46.00
Total		14		

Estadísticas de contraste^b

	Zona
U de Mann-Whitney	18.000
W de Wilcoxon	46.000
Z	-.832
Sig. asintót. (bilateral)	.405
Sig. exacta [2*(Sig. unilateral)]	.456*

a. No corregidos para los empates.

b. Variable de agrupación: INDIVIDUOS

Aceptación y rechazo de hipótesis nula y alternativa: Se acepta la hipótesis nula y se rechaza la hipótesis alternativa.

Conclusión: No existe evidencia suficiente para rechazar la hipótesis nula, por tanto, las medianas de las muestras son iguales y no hay diferencia entre los contenidos de azufre en los tres distintos yacimientos.

2.2.1.3. Caso de dos muestras relacionadas

2.2.1.3.1. Prueba de Wilcoxon

En una zona muy seca se pretende comprobar si determinado tratamiento, aplicado durante un mes, ayuda a reducir las agallas provocadas por *Agrobacterium tumefaciens*. Para ello, se toman aleatoriamente 10 plantas y en sus raíces se cuenta el número de agallas antes y después del tratamiento. Los resultados se presentan a continuación:

Antes	11	10	13	8	13	15	11	11	5	19
Después	11	8	9	3	9	11	11	9	6	11

¿Existen diferencias en el número de agallas antes y después del tratamiento?

Formular la hipótesis

Ho: Tratamiento no reduce las agallas de *A. tumefaciens*

H1: Tratamiento sí reduce las agallas de *A. tumefaciens*

Wilcoxon Signed Ranks Test

Ranks				
		N	Mean Rank	Sum of Ranks
Después - Antes	Negative Ranks	7a	5,00	35,00
	Positive Ranks	1b	1,00	1,00
	Ties	2c		
	Total	10		

- a. DESPUÉS < ANTES
- b. DESPUÉS > ANTES
- c. DESPUÉS = ANTES

	DESPUÉS - ANTES
Z	-2,395 ^b
Asymp. Sig.(2-tailed)	,017

- a. Wilcoxon signed ranks tests
- b. Based on positive ranks

Aceptación y rechazo de hipótesis nula y altesrnativa: Rechazo la hipótesis nula y acepto la hipótesis alternativa.

Conclusión: Existe evidencia suficiente para rechazar la hipótesis nula, por tanto, la mediana de las diferencias del número de agallas antes y después del tratamiento no es igual a cero, de manera que existen diferencias entre el número de agallas antes y después del tratamiento.

2.2.1.4. Caso de *k* muestras independientes

2.2.1.4.1. Prueba de *Kruskal-Wallis*

Se realiza un estudio del contenido de azufre en tres yacimientos de carbón. Se toman muestras aleatoriamente de cada uno de los yacimientos y se analizan. Los datos del porcentaje de azufre por muestra se indican en la siguiente tabla:

Yacimientos	Porcentaje de azufre								
1	51	92	8	4	14	76	17		
2	69	64	90	41	1	28	59	56	
3	22	32	39	33	54	4	25	49	30

¿Existe diferencia en los contenidos de azufre en los distintos yacimientos?

Formular la hipótesis

Ho: No hay diferencia en los contenidos de S en los distintos yacimientos.

H1: Sí hay diferencia en los contenidos de S en los distintos yacimientos.

Procedimiento SPSS

SPSS Editor de datos SPSS

Archivo Edición Ver Datos Transformar Analizar Gráficos Utilidades Ventana ?

1: Yacimiento

Yacimiento	Págufo
1	51
2	92
3	8
4	4
5	14
6	76
7	17
8	69
9	64
10	90
11	41
12	1
13	28
14	59
15	56
16	22
17	32
18	39
19	33
20	54
21	4
22	26
23	49
24	30
25	
26	
27	
28	
29	
30	

SPSS El procesador está preparado

24°C Lluvia ligera 15/5/2022

SPSS Editor de datos SPSS

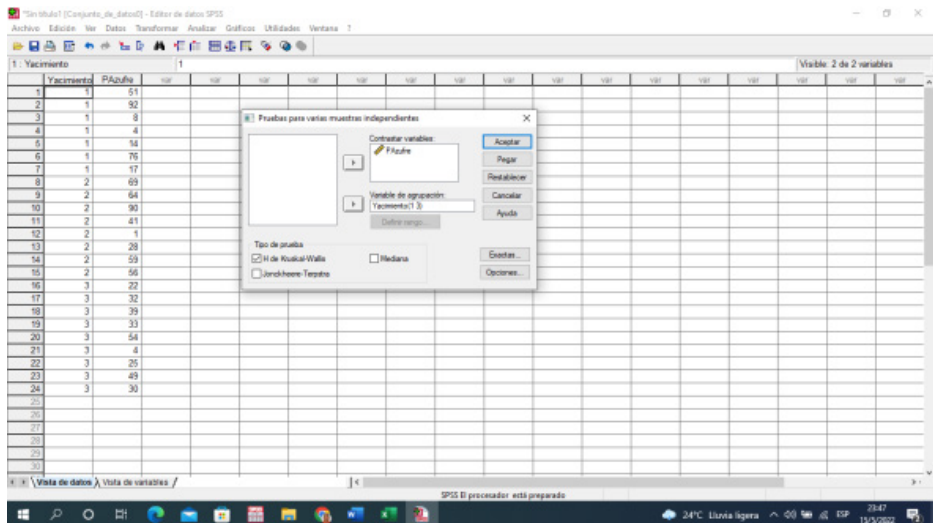
Archivo Edición Ver Datos Transformar **Análizar** Gráficos Utilidades Ventana ?

1: Yacimiento

Yacimiento	Págufo
1	51
2	92
3	8
4	4
5	14
6	76
7	17
8	69
9	64
10	90
11	41
12	1
13	28
14	59
15	56
16	22
17	32
18	39
19	33
20	54
21	4
22	26
23	49
24	30
25	
26	
27	
28	
29	
30	

SPSS El procesador está preparado

24°C Lluvia ligera 15/5/2022



Prueba de Kruskal-Wallis

	Yacimiento	N	Rango Promedio
PAzufre	1	7	11,36
	2	8	15,63
	3	9	10,61
Total		24	

	PAzufre
Chi-cuadrado	2,389
gl	2
Sig. asintót.	,303

- Prueba de Kruskal-Wallis
- Variable de agrupación: Yacimiento

Aceptación y rechazo de hipótesis nula y alternativa: Acepto hipótesis nula y rechazo la hipótesis alternativa.

Conclusión. No existe evidencia suficiente para rechazar la hipótesis nula, por tanto, las medianas de las muestras son iguales y no hay diferencia entre los contenidos de azufre en los tres distintos yacimientos.

2.2.1.5. Medidas de Asociación

2.2.1.5.1. Correlaciones no paramétricas (Spearman)

Se realiza un estudio para investigar la relación entre el nivel de humedad del suelo y la presencia de lombrices de tierra. Se obtuvieron los siguientes datos:

HS	0	0	0	0.32	0.32	0.32	0.63	0.63	0.63	0.94	0.94	0.94
NLT	10	8	10	4	6	6	0	2	0	2	4	2

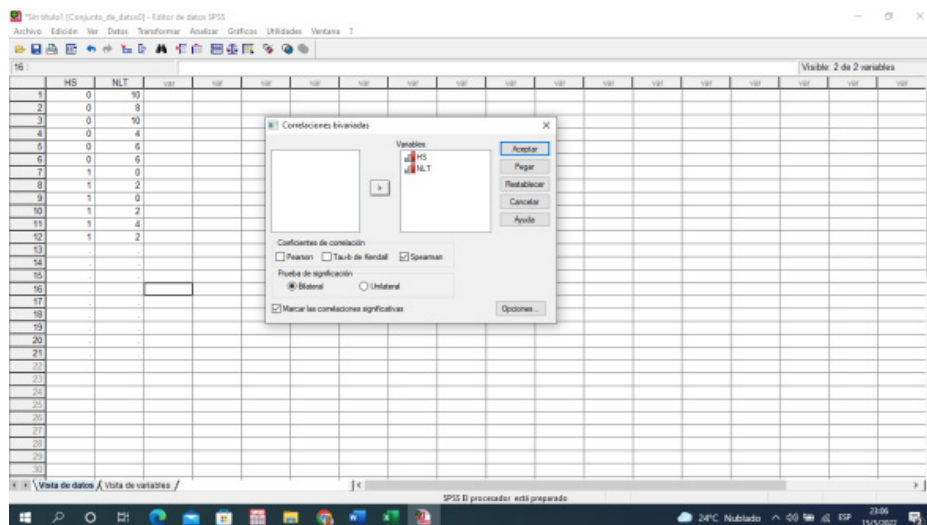
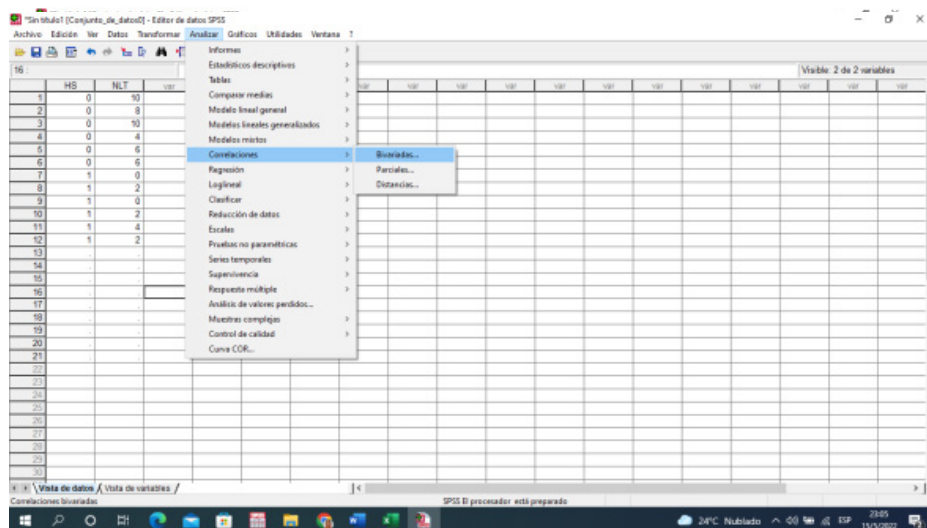
HS: Humedad del suelo (mmH2O por cm2 de suelo). **NLT:** Número de lombrices de tierra

¿Existe correlación entre ambas variables?

Formular la hipótesis

- Ho: El número de lombrices no está asociado con la humedad.
- H1: El número de lombrices sí está asociado con la humedad.

Procedimiento SPSS



Correlaciones no paramétricas

[conjunto de datos]

		HS	NTL
Rho de Spearman	HS	Coeficiente de correlación	1,000
		Sig. (bilateral)	-,788**
		N	,002
	NTL	Coeficiente de correlación	12
		Sig. (bilateral)	12
		N	-,788**

** La correlación es significativa al nivel 0.01 (bilateral).

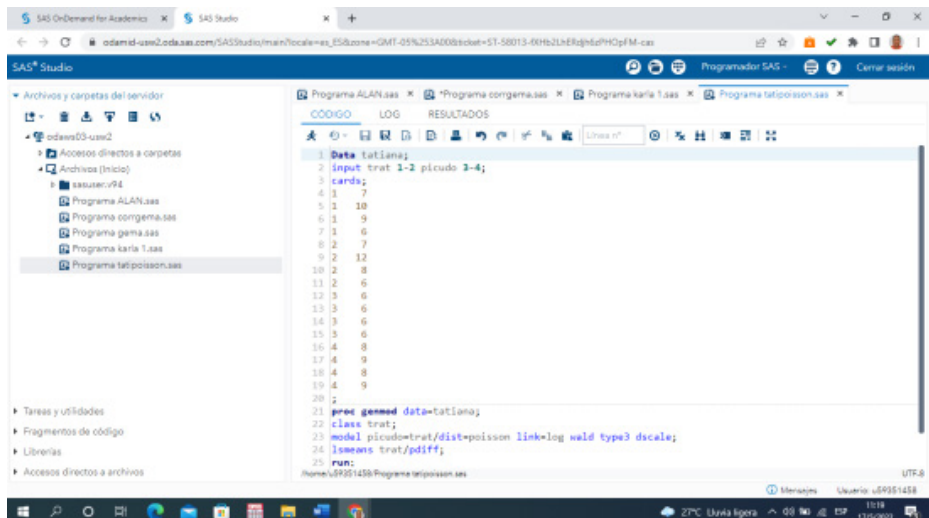
2.2.1.6. Aplicación de la distribución de Poisson

En un experimento realizado en la Granja Andil se evaluó la eficiencia de captura de tres tipos de atractantes y un testigo (solo agua) para capturar el picudo (*Cosmopolites sordidus Germar*) del plátano (*Musa paradisiaca*). El experimento fue implementado en un diseño experimental completamente aleatorio con cuatro trampas por tratamiento. Los tratamientos fueron: T1: Trampa con melaza (tarrinas de color amarillo llena hasta el 50 % con melaza). T2: Trampa con vinagre de guineo (tarrina de color amarillo llena hasta el 50 % con vinagre de guineo). T3: Trampa con pseudotallo de plátano (se utilizará la trampa tipo sándwich para la captura de insectos adultos del picudo). T4: Trampa con agua (testigo) (tarrina de color amarillo llena hasta el 50 % con agua). Se evaluó el número de insectos caídos en cada trampa cada ocho días, durante seis semanas (para cada semana se cambió la trampa, para que sean experimentos independientes) (Chompol y Castro, 2022).

Está claro que los datos en este tipo de experimentos no son de distribución normal y tampoco cumplen la homocedasticidad u homogeneidad de varianzas, debido a que es una variable cuantitativa discreta, y a que se realizan contajes del número de individuos capturados en las trampas durante las seis semanas (experimentos independientes).

En este caso se aconseja utilizar un análisis estadístico utilizando la distribución de tipo Poisson, que es muy útil para muestras de tamaño pequeño (menor a 30).

Para resolver este tipo de problemas se sugiere utilizar el software SAS University (SAS, 2020). que ofrece una alternativa factible para este tipo de distribuciones. La forma de introducir los datos sería la siguiente:



Salida

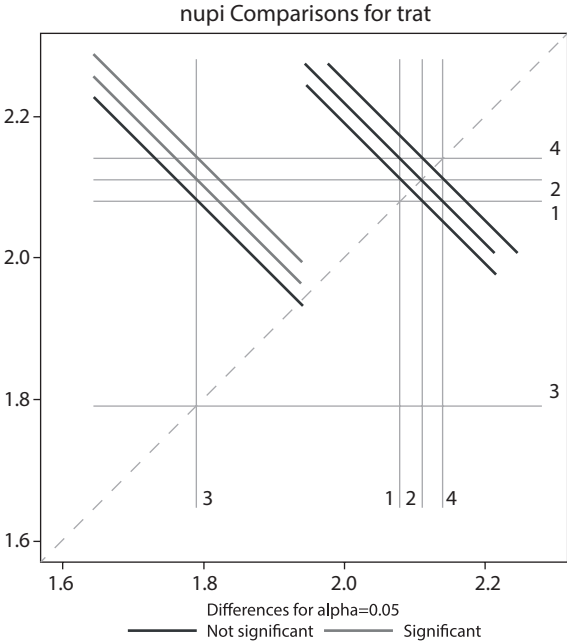
Análisis de estimadores de parámetro de verosimilitud máxima						
Parámetro	DF	Estimación	Error estándar	Límites de confianza de Wald al 95%	Chi-cuadrado de Wald	Pr > ChiSq
Intercept	1	2.1401	0.0980	1.9520	2.3262	497.27 <.0001
trat	1 1	-0.0606	0.1378	-0.3308	0.2095	0.19 0.6600
trat	2 1	-0.0299	0.1367	-0.2979	0.2362	0.05 0.8272
trat	3 1	-0.3483	0.1492	-0.6407	-0.0559	5.45 0.0196
trat	4 0	0.0000	0.0000	0.0000	0.0000	. .
Escala	0	0.5596	0.0000	0.5596	0.5596	

Note: The scale parameter was estimated by the square root of DEVIANCE/DOF.

Estadísticos de Wald para análisis de tipo 3			
Origen	DF	Chi-cuadrado	Pr > ChiSq
trat	3	8.46	0.0912

trat Medias de mínimos cuadrados					
trat	Estimación	Error estándar	Valor Z	Pr > z	
1	2.0794	0.09692	21.02	<.0001	
2	2.1102	0.09741	21.86	<.0001	
3	1.7918	0.1142	15.69	<.0001	
4	2.1401	0.09597	22.30	<.0001	

Diferencias de trat medias de mínimos cuadrados					
trat	_trat	Estimación	Error estándar	Valor Z	Pr > z
1	2	-0.03077	0.1388	-0.22	0.8246
1	3	0.2877	0.1511	1.90	0.0599
1	4	-0.06062	0.1378	-0.44	0.6600
2	3	0.3185	0.1501	2.12	0.0339
2	4	-0.02985	0.1367	-0.22	0.8272
3	4	-0.3483	0.1492	-2.33	0.0196



Interpretación de los resultados

La comparación de las medias de tratamiento (ver tablas y figuras) mostró que la comparación de los tratamientos T2 (trampa con vinagre de guineo y T3 (trampa con pseudotallo de plátano) mostraron diferencias significativas al $P < 0.05$ de probabilidad, asimismo, ocurrió entre los tratamientos T3 (trampa con pseudotallo de plátano) y T4 (testigo, con agua). Esto muestra que los mejores tratamientos fueron el T2 y T3.

2.2.1.7. Aplicación de la distribución de chi-cuadrado

En un estudio sobre la situación actual de los plaguicidas en las comunidades de Puerto La Boca y Cantagallo, de la parroquia Puerto Cayo de Manabí, Ecuador (Gabriel *et al.*, 2023). se aplicó una encuesta con 45 preguntas a 31 agricultores. La encuesta fue dividida en tres grandes grupos: 1. Treinta y tres preguntas fueron acerca de la compra, almacenamiento, aplicación y manejo de los plaguicidas. 2. Nueve preguntas fueron sobre los conocimientos en la aplicación de plaguicidas, y 3. Tres preguntas sobre daños causados en la salud por los plaguicidas. Para el análisis estadístico se utilizó la distribución de chi-cuadrado o de bondad de ajuste para las variables categóricas, para encontrar diferencias significativas. utilizando el software SPSS.

.....

100%

Category	Item 1	Item 2	Item 3
Row 1			
Row 2			

The screenshot displays the SPSS (Statistical Package for the Social Sciences) interface. The top window shows a data view with columns for various variables including 'GRUPO', 'SEXO', 'EDAD', 'TIPO', 'EST', 'GRUPO', 'CAP', 'LSP', 'FSP', 'APS', 'DOP', and 'CDA'. A dialog box for 'Pruebas no paramétricas' (Non-parametric Tests) is open, showing a list of variables to be tested. The bottom window shows the results of the non-parametric tests, including the 'Prueba de chi-cuadrado' (Chi-square test) and 'Frecuencias' (Frequencies).

Pruebas no paramétricas

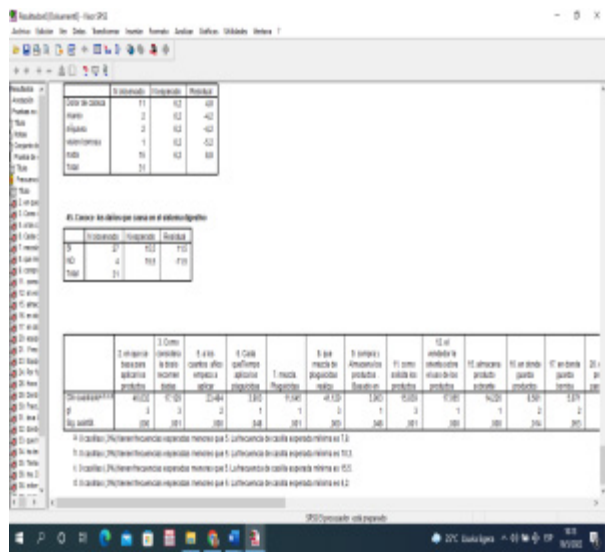
Chi-square test results:

Variable	Chi-square	df	Asymptotic Sig.
GRUPO	11.111	3	.003
SEXO	1.111	1	.291
EDAD	1.111	1	.291
TIPO	1.111	1	.291
EST	1.111	1	.291
GRUPO	11.111	3	.003
CAP	1.111	1	.291
LSP	1.111	1	.291
FSP	1.111	1	.291
APS	1.111	1	.291
DOP	1.111	1	.291
CDA	1.111	1	.291

Frecuencias

Frequency table for 'GRUPO':

GRUPO	N	%
1	11	55.0
2	7	35.0
3	2	10.0
Total	20	100.0



Interpretación de los resultados. El análisis de Chi-cuadrado al $P < 0.01$ de probabilidad para 30 preguntas, mostró diferencias altamente significativas entre lo observado y esperado, con excepción de las preguntas 4, 7, 11, 12, 13, 15, 19, 20 y 27 que fueron significativos al $P < 0.05$ de probabilidad, denotando en todos los casos que al menos un criterio fue diferente estadísticamente.

Referencias bibliográficas

- Berlanga, V. & Rubio, M.J. (2012). Clasificación de pruebas no paramétricas. Cómo aplicarlas en SPSS. REIRE, 5(2), 101-113. DOI:10.1344/reire2012.5.2528
- Bramardi, S. J. (2002). Análisis multivariado. Su aplicación en la caracterización de recursos genéticos. 60.
- Chompol, T., & Castro, C. (2022). Control etológico del picudo (*Cosmopolites sordidus* Germar) aplicando tres tipos de fermentos en el cultivo del plátano (*Mussa paradisiaca*) en el cantón Jipijapa (Tesis. Universidad Estatal del Sur de Manabí). Jipijapa. Ecuador. 46 p.
- Ferrán Aranaz, M. (2002). Curso de SPSS para Windows. Madrid: McGraw-Hill.
- Franco, T.L. & Hidalgo, R. (eds.). (2003) Análisis estadístico de datos de caracterización morfológica de recursos fitogenéticos. Boletín técnico n.º 8. Instituto Internacional de Recursos Fitogenéticos (IPGRI). Cali. Colombia. 89 p. <https://www.bioversityinternational.org/fileadmin>.
- Gabriel, J., López E., Magne, J., Angulo, A., Luján, R., La Torre, J., & Crespo, M. (2013) Genetic basis of inheritance for morphological. agronomic and agro-industries characteristics in hybrid tomato *Solanum lycopersicum* L. (Mill) (en línea). J Selva Andina Biosph. 1(1):45-54. http://www.scielo.org.bo/pdf/jsab/v1n1/v1n1_a05.pdf
- Gabriel, J., Valverde, A., Indacochea, B., Castro, C., Vera, M., Alcívar, J., & Vera, R. (2022). Diseños experimentales: teoría y práctica para experimentos agropecuarios. Mawil Publicaciones, Quito, Ecuador. <https://doi.org/10.26820/978-9942-602-26-8>
- Gabriel, J., Ávila, J., Ayón, F., Morán, J., Álvarez, A. & Flores, H. (2023). Utilización de plaguicidas por agricultores en Puerto La Boca, Manabí. Una reflexión sobre sus posibles consecuencias. Journal of the Selva Andina Biosphere, 11(1), 47-65. Epub 01 de mayo de 2023. <https://doi.org/10.36610/j.jsab.2023.110100044>
- García, C. & Gómez, S. (2008). Notas de curso sobre análisis multivariante. Universidad Pública de Navarra, Pamplona España.
- Gómez, S. (2017). Uso de software para análisis de datos cuantitativos. <http://softwareanalisisdedatoscuantitativos.blogspot.com/2014/04/tipos-de-paquetes-estadisticos.html>.

- González, J. (2009). Manual básico SPSS. Universidad de Talca. 70 p. https://www.fibao.es/media/uploads/manual_basico_spss_universidad_de_talca.pdf
- Hair, J.F., Anderson, R.E., Tatham, R.L., & Black, W.C. (1999): Análisis multivariante. Prentice Hall Iberia. Madrid.
- <https://www.studocu.com/ec/document/universidad-de-las-americas-ecuador/bioestadistica-ii/ejercicios-de-pruebas-no-parametricas-resuelto/5654317>
- Infante Gil, S. & Zárate de Lara, G.P. (1991). Métodos estadísticos, un enfoque interdisciplinario. Trillas, México D.F., México. 637 p.
- Lino, A.F. & Del Valle, W. (2023). Caracterización morfométrica y faneróptica del cerdo criollo (*Sus strofa domesticus*) en Jipijapa-Manabí (Tesis de maestría, Universidad Estatal del Sur de Manabí). Jipijapa, Ecuador.
- López, E. (2014). Estimación de la herencia y varianzas genéticas para caracteres morfológicos. agronómicos y agroindustriales en híbridos de tomate [*Solanum lycopersicum* L. (Mill.)]. (Tesis Ing. Agrónomo. Fac. de Ciencias Agrícolas y Pecuarias "Martín Cárdenas". Universidad Mayor de San Simón). Cochabamba. Bolivia. 69 p.
- López, E., Gabriel, J., Angulo, A., Magne, J., La Torre J., & Crespo, M. (2015). Herencia y relación genética asociados al rendimiento, madurez en híbridos de tomate [*Solanum lycopersicum* L. (MILL.)] [en línea]. Agro-nomía Costarricense, 39(1): 107-119. <https://www.scielo.sa.cr/pdf/ac/v39n1/a08v39n1.pdf>
- Morán, S.B. & Álvarez, A.H. (2023). Caracterización in situ de especies de frutales no tradicionales presentes en las fincas cafetaleras de la parroquia La Unión, zona sur, Jipijapa (Tesis pregrado. Universidad Estatal del Sur de Manabí). Jipijapa, Ecuador.
- Pérez Juste, R., García Llamas, J.L., Gil Pascual, J.A., & Galán González, A. (2009). Estadística aplicada a la educación. Madrid: UNED-Pearson.
- Rubio, M. J. & Berlanga, V. (2012). Cómo aplicar las pruebas paramétricas bivariadas t de Student y ANOVA en SPSS. Caso práctico. Revista d'Innovació i Recerca en Educació, 5(2). 83-100. <http://www.ub.edu/ice/reire.htm>
- SAS University. (2020). An Introduction to SAS University Edition. <https://www.oreilly.com/library/view/an-introduction-to/9781629600079/>

- Uriel, E. & Aldás, J. (2005). Análisis multivariante aplicado; aplicaciones al marketing, investigación de mercados, economía, dirección de empresas y turismo. Paraninfo Cengage Learning. Madrid. España. 531 p.
- Valverde Lucio, A., González-Martínez, A., Alcívar Cobeña, J.L., & Rodero Serrano, E. (2021). Characterization and Typology of Backyard Small Pig Farms in Jipijapa, Ecuador. *Animals*, 11(6), 1728. <https://doi.org/10.3390/ani11061728>
- Valverde Lucio, A., Ayón Villao, F., García Cabrera, J., Gabriel Ortega, J., & Vera Velázquez (2022). Análisis computacional de las caracterizaciones físico químicas de los suelos cafetaleros del sur de Manabí-Ecuador. *Serie Científica de la Universidad de las Ciencias Informáticas*, 15(1), 37-55. <https://dialnet.unirioja.es/servlet/articulo?codigo=8955475>
- Visauta, B. (2007). Análisis estadístico con SPSS 14: Estadística básica (3.a ed.). Madrid: McGraw-Hill Interamericana.
- Wiedenhofer, H. (2013). Pruebas no paramétricas para las ciencias agropecuarias. Muestras pequeñas. 2.a ed. rev y amp. Maracay. VE. Instituto Nacional de Investigaciones Agrícolas. 261 p. <https://docplayer.es/14084308-Pruebas-no-parametricas-para-las-ciencias-agropecuarias.html>

ANÁLISIS MULTIVARIANTE Y NO PARAMÉTRICO:

Casos de estudio aplicados a experimentos agropecuarios



Publicado en Ecuador
Junio 2025

Edición realizada desde el mes de Enero de 2025 hasta
Junio del año 2025, en los talleres Editoriales de MAWIL
publicaciones impresas y digitales de la ciudad de Quito.

Quito – Ecuador

Tiraje 30, Ejemplares, A5, 4 colores; Offset MBO
Tipografía: Helvetica LT Std; Bebas Neue; Times New Roman.
Portada: Collage de figuras representadas y citadas en el libro.