



Universidad Agraria de La Habana, Cuba
"Fructuoso Rodríguez Pérez"

MAESTRÍA EN BIOMATEMÁTICA

" Metodologías para el procesamiento de datos en Investigaciones Científicas. Uso de Software



Dra.C Lucía Fernández Chuairey
Profesor Titular

lucia@unah.edu.cu, chuairey@yahoo.com

2023

CONTENIDO DEL CURSO:

Papel de la Estadística vinculada a investigaciones Matemáticas, Agrarias y Educacional. La Estadística como Método Científico. Escalas de medidas. Fuentes de información y necesidad de su evaluación. Procesamiento y presentación de la información: tablas, diagramas, esquemas, y gráficos. Consideraciones sobre el manejo y resumen de datos. Estadística Descriptiva. Análisis de datos e interpretación de resultados. Probabilidad e Inferencia Estadística. Estimación y Docima para pruebas de hipótesis de media. Métodos paramétricos y no paramétricos. Diseño experimental y no experimental (introducción al trabajo con encuestas). Aplicaciones. Uso de Software estadísticos. **Elementos generales de análisis multivariado. Método de componentes principales.**

ANTECEDENTES DE ESTUDIO

- Estadística
- Metodología de la investigación



OBJETIVOS

OBJETIVOS GENERAL

Emplear Métodos Estadísticos con el correspondiente procesamiento, presentación e interpretación de los resultados en función de las características propias de su objeto de estudio, que permitan arribar a conclusiones certeras, con vistas a lograr una mayor eficiencia de su uso en la actividad docente y científico-investigativa.

EVALUACIÓN FINAL

•Trabajo final: Entrega y discusión de un trabajo relacionado con una problemática de la especialidad que permita medir el usos de las herramientas estadísticas - matemáticas abordadas en el curso, así como los software empleados para su realización.

Gran parte de los procesos de investigación generan un conjunto de **datos** de diferentes características, muchos de los cuales necesitaran de un **análisis estadístico**. De ahí la necesidad la necesidad de comprender la naturaleza del proceso estadístico íntimamente relacionado con el de investigación (**objetivo**) y de conocer algunos métodos mas **tradicionales** y **actuales** usados en el en el análisis de datos, diseño de experimentos, encuesta entre otros (Pereyra y col, 2004)



ASPECTOS GENERALES DEL TEMA ANTERIOR

¿ Estadística? ¿El Proceso Estadístico?

Estadística, es la rama de la **MATEMÁTICA** que se ocupa de reunir, organizar y analizar datos numéricos y que ayuda a resolver problemas como la inferencia, la toma de decisiones, predicciones, entre otros

CIENCIA (que se ocupa en general de los fenómenos observables); **Método de Investigación**, tecnología del **MÉTODO CIENTÍFICO**, herramientas sumamente útiles para la investigación. (Bernard Ostle)



¿Qué instrumento utilizar? ¿Cómo usarlo?
¿Cómo interpretar el resultado?

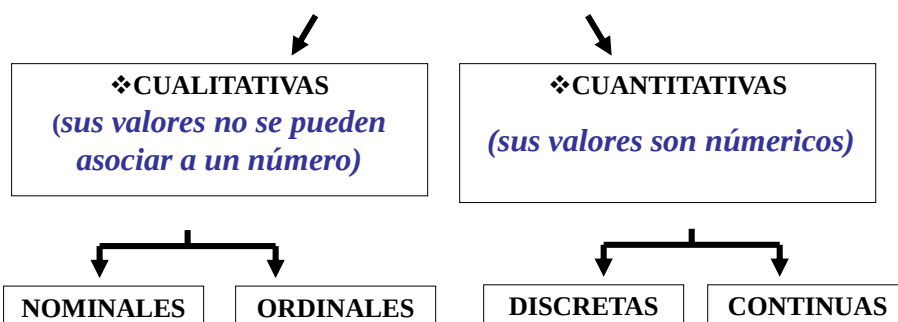
TRABAJO CON VARIABLE

¿VARIABLE? Una variable es una propiedad que puede variar y cuya variación es susceptible de medirse

Estadística Descriptiva

¿Tipo de variable? ¿Qué dice el dato? ¿Para que me sirve?

TIPO DE VARIABLES



¿Herramientas Estadísticas generalmente empleadas en la descripción de **VARIABLES CUALITATIVAS**?

DIAGRAMA DE BARRAS Y TABLAS DE FRECUENCIA

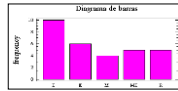


Tabla de Frecuencia para las evaluaciones

Clase	Valor	Frecuencia	Relative Frequency	Cumulative Frequency	Cum. Rel. Frequency
1	B	10	0.3333	10	0.3333
2	B	4	0.2000	14	0.5333
3	B	4	0.1333	18	0.6667
4	MB	4	0.1667	22	0.8333
5	B	4	0.1667	26	1.0000

¿¿¿¿Herramientas Estadísticas generalmente empleadas en la descripción de **VARIABLES CUANTITATIVAS**????

•Medidas de tendencia central {
Media ($\bar{x}$, μ)
Moda (M_o)
Mediana (M_e)

•Medidas de dispersión {
Varianza (S^2 , σ^2)
Desviación Estándar (S , σ)
Coefficiente de Variación $CV = \frac{S}{\bar{X}} * 100$

ESTADÍSTICA
PUBLICACIONES
CIENTÍFICAS, TESIS, ETC

MATERIALES Y MÉTODOS

RESULTADOS Y DISCUSIÓN

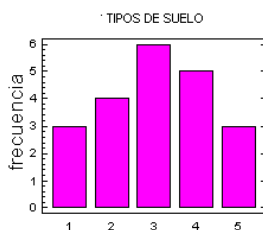
MATERIALES Y MÉTODOS

- bajo que condiciones se realizó la investigación (condiciones climáticas, procedimiento empleado para tomar las muestras, etc)
- Variables aleatorias observadas
- periodo (tiempo)
- tamaño de muestra,
- herramienta estadística utilizada
- softwear

SE PARTE REVISIÓN
BIBLIOGRÁFICA Ó RESEÑA

En una Investigación Agraria, relacionada con la caracterización de tipos y subtipos de suelos en una región de la Provincia Mayabeque, se analizaron un total de 21 subregiones. Para la caracterización de esta variable se utilizó tablas de frecuencia y..... En el procesamiento de los datos se empleo el software

RESULTADOS Y DISCUSIÓN



Interpretación de la salida
(destacar extremos)

Se pudo mostrar (tabla I.1) ó (figura II.1) que en la región el 28.57% de los suelo son de tipo ferralíticos rojos y **DISCUTIR** (coinciden superan.....están por debajo.....)

EXPLICACIÓN BIOLÓGICA, FISICA.....

Uso de Software

• **Statgraphics (versión 6.1)**

• **SAS (versión 6.21, 1989-1996)**

• **CSS (año 1990)**

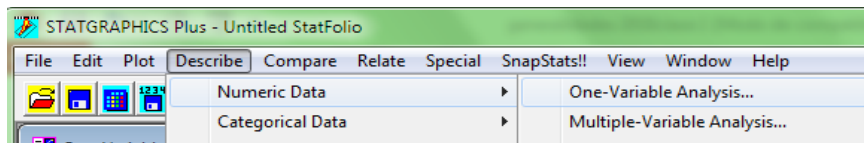
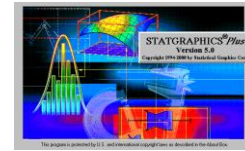
• **SPSS (o la SPSS- PC)**

• **Statística (versión 4.0, año 1993)**

• **Systat (versión 5.0)**

• **Sigma Plot (versión 5. 1999)**

• **Software libre R (aparece 1993)**

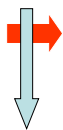


Estadística Inferencial

Parámetros: Indicador constante que caracteriza la población

Estadígrafos: Es una función de n valores muestrales

Muestra



Población

Probabilidad

**Estimación
puntual**

$$\left[\bar{X} \rightarrow \hat{\mu} \quad s \rightarrow \hat{\sigma} \quad \tilde{p} \rightarrow \hat{p} \quad \dots \right]$$

**Estimación
intervalo de
confianza**

$$P\{\mu \in I\} = \left\{ \bar{x} \pm \frac{s}{\sqrt{n}} t_{1-\frac{\alpha}{2}}^{n-1} \right\} = 1 - \alpha$$

Nivel de confianza

$1 - \alpha$
cercanos a 1
0.90
0.95
0.99

SE ESTIMA CON UN 95% DE CONFIANZA DE QUE....

PROBLEMAS RELACIONADOS CON LA TOMA DE DECISIONES

DÓCIMAS DE HIPÓTESIS O PRUEBAS DE HIPOTESIS

Comparar dos medias??

Comparar dos o mas medias??

-  **Pasos para docimar**
- **Plantear la hipótesis H_0 y H_1**
 - **Fijar el nivel de significación α**
 - **Calcular el estadígrafo**
 - **Determinar la región crítica**
 - **Tomar la decisión de aceptación o rechazo**

Se utiliza:
estadígrafo **t** – **student**
(para comparar DOS
poblaciones
estadígrafo **F** para
comparar dos o mas
poblaciones (ANOVA)



MATERIALES Y MÉTODOS

RESULTADOS Y DISCUSIÓN

Hacer el análisis a partir del valor de probabilidad

$p \{ \text{rechazar } H_0 / H_0 \text{ cierta} \} < \alpha$ entonces rechazo a H_0

ESTADISTICA MULTIVARIADA



ELEMENTOS GENERALES DE ANÁLISIS MULTIVARIADO.
MÉTODO DE COMPONENTES PRINCIPALES



DEFINICIÓN

HAIR y col. (1999) “El análisis multivariado en un sentido amplio, se refiere a todos los métodos estadísticos que analizan simultáneamente medidas múltiples de cada individuo u objeto sometido a investigación. Cualquier análisis simultáneo de más de dos variables puede ser considerado aproximadamente como un análisis multivariante.



... agrupa un conjunto de técnicas estadísticas que se encargan del análisis de datos correspondientes a mediciones de *p* variables observadas en *n* individuos; permitiendo el estudio de las interrelaciones

Métodos de Análisis Multivariado o Análisis Multidimensional (Permite la descripción simultanea de varias variables, así como relaciones entre las variables, reduce la dimensión del problema, describen, deciden entre otros aspectos)

En un experimento se estudió el comportamiento del rendimiento de 10 variedades de calabaza sometidas a 8 ambientes o condiciones diferentes de estrés de temperatura y humedad. Los resultados se presenta en la siguiente tabla:

ESTADÍSTICA MULTIVARIADA

Ambientes/ Variedades	E1	E2	E3	E4	E5	E6	E7	E8
V1	8.44	8.01	2.50	3.29	1.33	0.80	1.84	1.19
V2	9.57	5.63	3.60	6.27	0.37	1.35	1.76	1.54
V3	10.93	5.42	3.70	6.10	0.90	0.99	1.30	0.14
V4	8.60	8.87	5.10	4.90	0.90	0.91	1.22	1.57
V5	9.60	3.37	1.45	2.80	0.35	0.56	0.97	0.62
V6	6.58	2.53	2.83	2.97	0.57	0.13	3.10	0.13
V7	4.66	3.64	2.80	1.52	0.35	0.32	2.74	0.37
V8	5.70	4.65	2.74	1.83	0.92	0.20	1.46	0.37
V9	5.33	3.03	1.92	2.91	1.09	0.27	1.71	0.54
V10	3.40	3.68	2.38	1.62	0.41	1.65	0.72	0.52

FUENTE: Varela, M. Análisis Multivariado de Datos. Aplicación a las Ciencias Agrícolas. Monografía. ISBN:959-7023-04-0. Ediciones INCA. 1998.

Se presentan los datos correspondientes a siete variables socioeconómicas para 14 países (Batista Forquet y Martines Áreas 1989)

Estas son : Densidad de población, porcentaje de personas empleadas en la agricultura, renta (ingreso nacionales percápita), Inversiones de rendimiento de capital en maquinaria y otros , . tasa de mortalidad Infantil, consumo de energía por 100 habitantes, aparatos de TV por 100 habitantes

ESTADÍSTICA MULTIVARIADA

	pais	densidad	empleo	renta	inversiones	mortalidad	energia	TV
1	Austalia	2	6	8,4	10,1	12	5,2	36
2	Francia	97	9	10,7	9,2	10	3,7	28
3	Alemania	247	6	12,4	9,1	15	4,6	33
4	Gracia	72	31	4,1	8,1	19	1,7	12
5	Islandia	2	13	11	6,6	11	5,8	25
6	Italia	189	15	5,7	7,9	15	2,5	22
7	Japon	311	11	8,7	10,9	8	3,3	24
8	NZelandia	12	10	6,8	8	14	3,4	26
9	Portugal	107	31	2,1	5,5	39	1,1	9
10	España	74	19	5,3	6,9	15	2	21
11	Suecia	18	6	12,8	7,2	7	6,3	37
12	Turquia	56	61	1,6	8,8	153	0,7	5
13	Reino Unido	229	3	7,2	9,3	13	3,9	39
14	EU	24	4	10,6	7,3	13	8,7	62

Manual de Estadística para proyectos de Investigación Autor Pereyra A.M; Abbiati,N ,N; Fernández, E.N Universidad Nacional de Lomas de Zamora ISBN 987-9455-33-9, Argentina, año 2004) Pág. 124

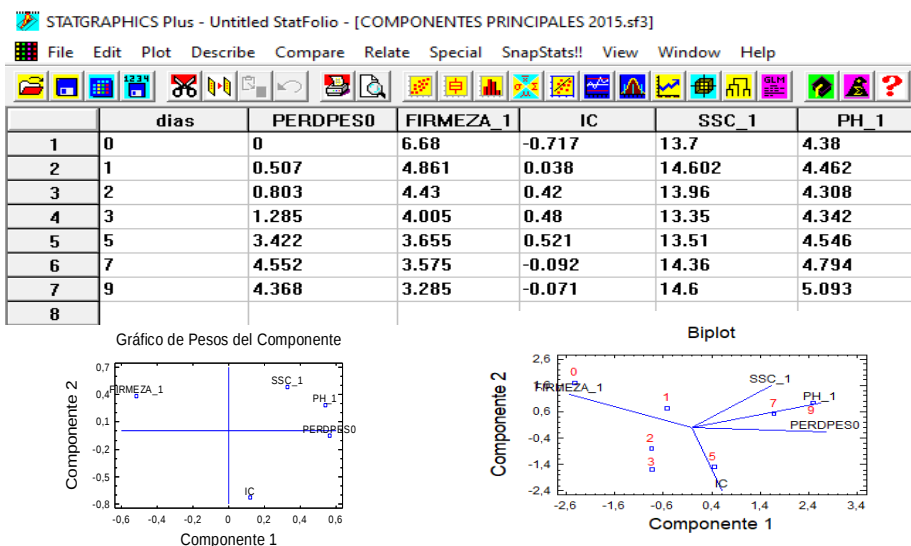
Se estudió el comportamiento de 18 variedades de boniato, a las que se le evaluaron un total de 10 variables

Var.	Peso (g)	Largo (cm)	Ancho (cm)	.	.	.	Azuc. Total.	Ceniza (%)	Fibra (%)
1	191.2	11.0	4.5	.	.	.	5.95	1.46	1.28
2	150.4	13.5	5.0	.	.	.	4.39	1.53	1.51
3	155.7	15.0	6.5	.	.	.	4.43	1.25	1.26
.				.	.	.			.
.	.	.	.	.	.	.	.	.	.
.				.	.	.			.
17	237.8	23.5	6.0	.	.	.	3.98	0.83	1.11
18	176.8	15.5	4.0	.	.	.	3.71	0.82	1.04

FUENTE: Varela, M. Análisis Multivariado de Datos. Aplicación a las Ciencias Agrícolas. Monografía. ISBN:959-7023-04-0. Ediciones INCA. 1998.

Propiedades físico de los frutos, en el transcurso de los días, post cosecha

ESTADÍSTICA MULTIVARIADA



BIBLIOGRAFIA

Varela, M. Análisis Multivariado de Datos. Aplicación a las Ciencias Agrícolas. Monografía. ISBN:959-7023-04-0. Ediciones INCA. 1998.

Pereyra A.M; Abbiati,N ,N; Fernández, E.N. Manual de Estadística para proyectos de Investigación Autor Universidad Nacional de Lomas de Zamora ISBN 987-9455-33-9, Argentina, año 2004.

Bouxa,C.N; Sistachs Vivian. Estadística. Teoría Básica y ejercicios. Editorial Félix Varela, 2006

Linares, Gladys. Análisis de Datos. Universidad de La Habana, 1990

¿Técnicas Multivariadas?

Las técnicas multivariadas se encarga del análisis de datos correspondientes a mediciones múltiples

Primeras ideas de técnicas multivariadas surgen a Principios del siglo XX



¿Por qué se plantea que es una técnica reciente?
¿TEORIA DE LAS TECNICAS MULTIVARIADAS ?



IMPACTO DE LA REVOLUCIÓN INFORMÁTICA

Solo cuando los ordenadores tuvieron capacidad de cálculo y memorias necesarias para llevar a cabo el análisis multivariante, es que se empezó a generalizarse el uso de estas técnicas (Hair y col, 1999)

Uso de Software Estadístico

- SAS (versión 6.21, 1989-1996)
- SPSS (o la SPSS- PC)
- Statgraphics (versión 5.1, 1999)
- Statistica (versión 4.0, año 1993)
- Systat (versión 5.0)



Pereyra y col. (2004) Métodos Multivariados se pueden distinguir:

Métodos Factoriales (se fundamentan en el Algebra lineal y la Geometría Euclideana) Consiste en proyectar nube de puntos de individuos y las variables sobre planos denominados factoriales. Se clasifican según el número y tipos de variable

•Análisis de Componentes Principales (ACP)

•Análisis Factorial de Correspondencia Simple (AFCS)

•Análisis Factorial de Correspondencia Multiple (AFCM)

Métodos de Clasificación (Se basa en medidas de distancia entre objetos (individuos o variable) y permite obtener tipologías (particiones estable e interpretables), permite descubrir agrupamientos naturales de observaciones o variables, o busca grupos similares. (Análisis Discriminante, Cluster)



	dias	PERDPESO	FI			
1	0	0	6.			
2	1	0.507	4.			
3	2	0.803	4.			
4	3	1.285	4.005	0.48	13.3	
5	5	3.422	3.655	0.521	13.5	
6	7	4.552	3.575	-0.092	14.3	
7	9	4.368	3.285	-0.071	14.6	5.093

Matriz de datos X_{ij} :

i	j	1	...	j	...	P
1						
.						
.						
.						
i				X_{ij}		
.						
.						
.						
N						P

P: Número de variables
N: Número de individuos

Se quiere realizar el estudio simultaneo de estas variables y se trata de determinar la relación entre estas variables

STATGRAPHICS Plus - StatFolio sin Nombre - [DOS EJEMPLOS DE COMP PRINC.sf3]										
Archivo Edición Gráficos Descripción Comparación Dependencia Avanzado SnapStats! Ver Ventana Ayu										
	varied	e1	e2	e3	e4	e5	e6	e7	e8	Col. 10
1	v1	8,44	8,01	2,5	3,29	1,33	0,8	1,84	1,19	
2	v2	9,57	5,63	3,6	6,27	0,37	1,35	1,76	1,54	
3	v3	10,93	5,42	3,7	6,1	0,9	0,99	1,3	0,14	
4	v4	8,6	8,87	5,1	4,9	0,9	0,91	1,22	1,57	
5	v5	9,6	3,37	1,45	2,8	0,35	0,56	0,97	0,62	
6	v6	6,58	2,53	2,83	2,97	0,57	0,13	3,1	0,13	
7	v7	4,66	3,64	2,8	1,52	0,35	0,32	2,74	0,37	
8	v8	5,7	4,65	2,74	1,83	0,92	0,2	1,46	0,37	
9	v9	5,33	3,03	1,92	2,91	1,09	0,27	1,71	0,54	
10	v10	3,4	3,68	2,38	1,62	0,41	1,65	0,72	0,52	
11										

¿¿¿Variables????

e1, e2, e3, e4, e5, e6, e7, e8

8 ambientes diferentes

¿¿¿Individuos???

v1 - v10

10 variedades de calabazas diferentes

Análisis de Componentes Principales

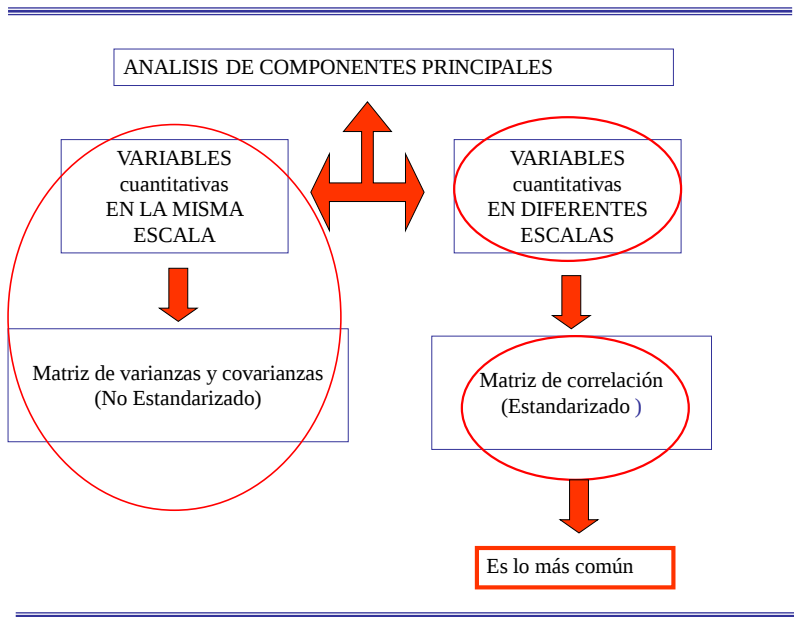
El método consiste en construir un conjunto de p nuevas variables o componentes ($Y_i, i=1..p$), como combinación lineal de las variables originales (X_i). Estas nuevas variables o componentes tienen la característica de concentrar la mayor parte de la variabilidad de los datos en las primeras componentes, o sea, cada componente extrae mayor varianza que las siguientes. Además son variables incorrelacionadas.

$$(X_1 \ X_2 \ \dots \ X_p) \rightarrow (Y_1 \ Y_2 \ \dots \ Y_p)$$

$$\begin{pmatrix} X_{11} & X_{12} & \dots & X_{1p} \\ X_{21} & X_{22} & \dots & X_{2p} \\ \dots & \dots & \dots & \dots \\ X_{i1} & X_{i2} & \dots & X_{ip} \\ \dots & \dots & \dots & \dots \\ X_{n1} & X_{n2} & \dots & X_{np} \end{pmatrix} \rightarrow \begin{pmatrix} Y_{11} & Y_{12} & \dots & Y_{1p} \\ Y_{21} & Y_{22} & \dots & Y_{2p} \\ \dots & \dots & \dots & \dots \\ Y_{i1} & Y_{i2} & \dots & Y_{ip} \\ \dots & \dots & \dots & \dots \\ Y_{n1} & Y_{n2} & \dots & Y_{np} \end{pmatrix} \quad \begin{aligned} Y_{i1} &= \alpha_{11}X_{i1} + \alpha_{12}X_{i2} + \alpha_{13}X_{i3} + \dots + \alpha_{1p}X_{ip} \\ Y_{i2} &= \alpha_{21}X_{i1} + \alpha_{22}X_{i2} + \alpha_{23}X_{i3} + \dots + \alpha_{2p}X_{ip} \\ Y_{i3} &= \alpha_{31}X_{i1} + \alpha_{32}X_{i2} + \alpha_{33}X_{i3} + \dots + \alpha_{3p}X_{ip} \\ &\dots \\ Y_{ip} &= \alpha_{p1}X_{i1} + \alpha_{p2}X_{i2} + \alpha_{p3}X_{i3} + \dots + \alpha_{pp}X_{ip} \end{aligned}$$

$$V(X_1) + V(X_2) + \dots + V(X_p) = \text{Varianza Total} = V(Y_1) + V(Y_2) + \dots + V(Y_p)$$

$$V(Y_1) \geq V(Y_2) \geq V(Y_3) \geq \dots \geq V(Y_p) \quad \text{cor}(Y_i, Y_j) = 0 \quad i \neq j$$



En un experimento se estudió el comportamiento del rendimiento de 10 variedades de calabaza sometidas a 8 ambientes o condiciones diferentes de estrés de temperatura y humedad.

Variables

Ambientes/ Variedades	E1	E2	E3	E4	E5	E6	E7	E8
V1	8.44	8.01	2.50	3.29	1.33	0.80	1.84	1.19
V2	9.57	5.63	3.60	6.27	0.37	1.35	1.76	1.54
V3	10.93	5.42	3.70	6.10	0.90	0.99	1.30	0.14
V4	8.60	8.87	5.10	4.90	0.90	0.91	1.22	1.57
V5	9.60	3.37	1.45	2.80	0.35	0.56	0.97	0.62
V6	6.58	2.53	2.83	2.97	0.57	0.13	3.10	0.13
V7	4.66	3.64	2.80	1.52	0.35	0.32	2.74	0.37
V8	5.70	4.65	2.74	1.83	0.92	0.20	1.46	0.37
V9	5.33	3.03	1.92	2.91	1.09	0.27	1.71	0.54
V10	3.40	3.68	2.38	1.62	0.41	1.65	0.72	0.52

Individuos



¿Por qué aplicamos en esta problemática el Método de Componentes Principales?

¿¿Comportamiento del rendimiento en estas variedades de calabazas en los diferentes ambientes?

¿Cuándo es efectivo este método?

¿Por qué aplicamos en esta problemática el Método de Componentes Principales?

MÉTODO DE COMPONENTES PRINCIPALES



El objetivo es condensar la información contenida en el número de variables originales en un conjunto mas pequeños de variables (factores) con una perdida mínima de información

....construir un conjunto de nuevas variables o componentes Y_i (incorrelacionadas), combinación lineal de las variables originales X_{ij} , de forma tal que recoge toda la variabilidad existente

$\text{Var}(Y_1) + \text{Var}(Y_2) + \dots + \text{Var}(Y_p) = \text{varianza total} = \text{Var}(X_1) + \text{Var}(X_2) + \dots + \text{Var}(X_p)$

tal que: $\text{Var}(Y_1) > \text{Var}(Y_2) > \dots > \text{Var}(Y_p)$

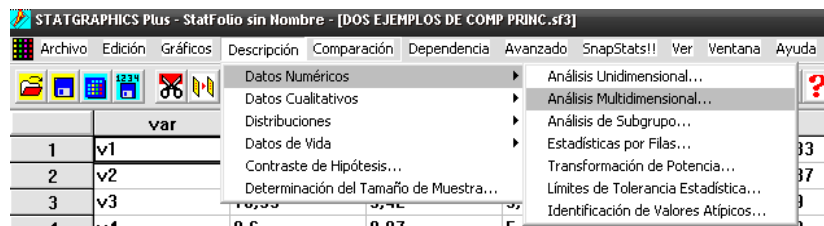


¿Cuándo es efectivo este método?



¿Existe grado de asociación entre las variables???

•La correlación mide **el grado de asociación** entre dos variables $(-1 \leq r \leq 1)$.



STATGRAPHICS Plus - StatFolio sin Nombre - [DOS EJEMPLOS DE COMP PRINC.sf3]

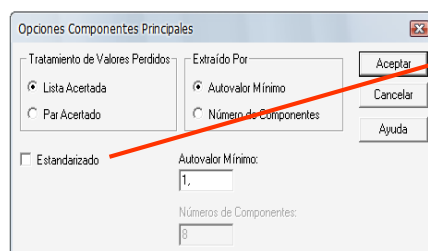
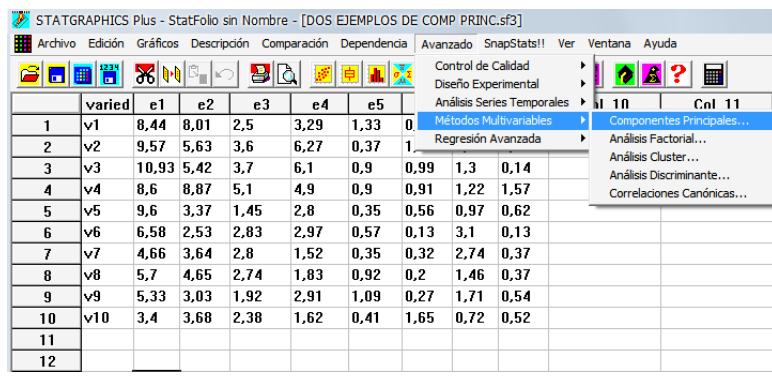
	e1	e2	e3	e4	e5	e6	e7	e8
1		0,4738 (10) 0,1666	0,3515 (10) 0,3192	0,8254 (10) 0,0833	0,1490 (10) 0,6812	0,1694 (10) 0,6480	-0,1814 (13) 0,6159	0,3478 (10) 0,3247
2	0,4738 (10) 0,1666		0,6820 (10) 0,0298	0,5023 (10) 0,1390	0,5117 (10) 0,1305	0,3436 (10) 0,0817	-0,2737 (13) 0,4441	0,7527 (10) 0,0120
3	0,3515 (10) 0,3192	0,6820 (10) 0,0298		0,6318 (10) 0,0501	0,1226 (10) 0,7358	0,2935 (10) 0,4105	0,0037 (13) 0,9984	0,4707 (10) 0,1698
4	0,8254 (10) 0,0833	0,5023 (10) 0,1390	0,6318 (10) 0,0501		0,1293 (10) 0,7219	0,4055 (10) 0,2451	-0,1394 (13) 0,7089	0,4666 (10) 0,1740
5	0,1490 (10) 0,6812	0,5117 (10) 0,1305	0,1226 (10) 0,7358	0,1293 (10) 0,7219		-0,2146 (10) 0,5516	-0,0671 (10) 0,8539	0,1328 (10) 0,7146
6	0,1694 (10) 0,6480	0,3636 (10) 0,3017	0,2935 (10) 0,4105	0,4055 (10) 0,2451	-0,2146 (10) 0,5516		-0,5985 (13) 0,0676	0,4558 (10) 0,1855
7	-0,1814 (10) 0,6159	-0,2737 (10) 0,4441	0,0037 (10) 0,9984	-0,1394 (10) 0,7089	-0,0671 (10) 0,8539	-0,5985 (10) 0,0676		-0,2494 (10) 0,4871
8	0,3478 (10) 0,3247	0,7527 (10) 0,0120	0,4707 (10) 0,1698	0,4666 (10) 0,1740	0,1328 (10) 0,7146	0,4558 (10) 0,1855	-0,2494 (13) 0,4871	

E1-E4, E2-E3, E3-E4, E2-E8

En un experimento se estudió el comportamiento del rendimiento de 10 variedades de calabaza sometidas a 8 ambientes o condiciones diferentes de estrés de temperatura y humedad.

Ambientes Variedades	E1	E2	E3	E4	E5	E6	E7	E8
1	8.44	8.01	2.50	3.29	1.33	0.80	1.84	1.19
2	9.57	5.63	3.60	6.27	0.37	1.35	1.76	1.54
3	10.93	5.42	3.70	6.10	0.90	0.99	1.30	0.14
4	8.60	8.87	5.10	4.90	0.90	0.91	1.22	1.57
5	9.60	3.37	1.45	2.80	0.35	0.56	0.97	0.62
6	6.58	2.53	2.83	2.97	0.57	0.13	3.10	0.13
7	4.66	3.64	2.80	1.52	0.35	0.32	2.74	0.37
8	5.70	4.65	2.74	1.83	0.92	0.20	1.46	0.37
9	5.33	3.03	1.92	2.91	1.09	0.27	1.71	0.54
10	3.40	3.68	2.38	1.62	0.41	1.65	0.72	0.52

Aplicar Análisis de Componentes Principales



Desmarcar el Estandarizado, ya que en este caso se trabaja con la misma escala, trabaja con la Matriz de Varianza y Covarianza

Resumen del Análisis

Datos/Variables:

e1
e2
e3
e4
e5
e6
e7
e8

Entrada de datos: observaciones
Número de casos completos: 10
Tratamiento de valor perdido: lista considerada
Estandarizado: no

Número de componentes extraídos: 3

¿Número de componentes Principales a considerar ?

Criterio de Porcentaje (Linares, G)
Incluir el criterio suficiente de componentes principales que den un porcentaje de la varianza aceptable. Usualmente se toma alrededor del 85%

Criterio de Kaiser (Linares, G)
Si se trabaja con R, tomar sólo aquellos auto valores que sean mayores o iguales a 1

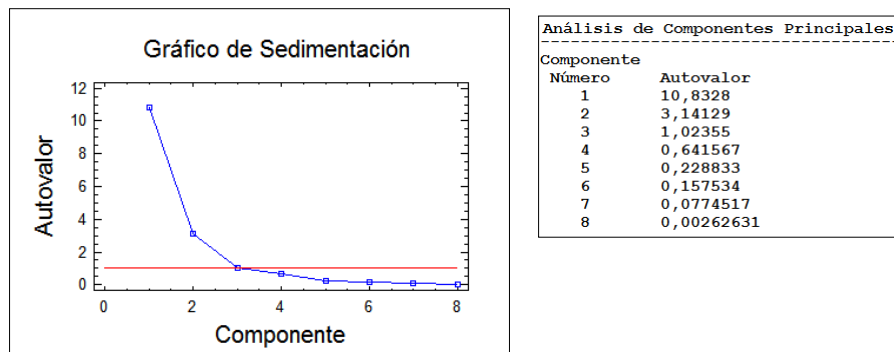
Análisis de Componentes Principales

Componente Número	Autovalor	Porcentaje de Varianza	Acumulado Porcentaje
1	10,8328	67,261	67,261
2	3,14129	19,504	86,765
3	1,02355	6,355	93,120
4	0,641567	3,983	97,104
5	0,228833	1,421	98,525
6	0,157534	0,978	99,503
7	0,0774517	0,481	99,984
8	0,00262631	0,016	100,000

Las dos primeras componentes explican el 86.76% de la variabilidad total

Esto indica que en lugar de trabajar con 8 variables iniciales, se puede describir los datos a partir de las dos primeras componentes

Bibliografía a Consultar: Linares, Gladys. Análisis de Datos. Universidad de La Habana. Impreso en Andrés Voisin 1990.



Criterio de las Pendientes Se hace un gráfico componente y porcentaje de varianza explicada por cada componente. Se unen los puntos mediante segmentos de rectas. Se determina a partir del punto en que comienza la caída. Ahí se detiene y se toma el número de componentes.

LAS EXPERIENCIAS PRÁCTICAS INDICAN TRABAJAR EN EL SENTIDO DE UN COMPROMISO ENTRE LOS CRITERIOS

Matriz de varianzas y covarianzas:

	e1	e2	e3	e4	e5	e6	e7	e8
e1	6,19	2,51	0,9	3,61	0,13	0,22	-0,34	0,47
e2		4,54	1,49	1,88	0,39	0,4	-0,43	0,87
e3			1,06	1,14	0,04	0,16	0	0,26
e4				3,08	0,08	0,37	-0,18	0,44
e5					0,13	-0,04	-0,02	0,03
e6						0,26	-0,23	0,13
e7							0,56	-0,1
e8								0,29

VARIANZA
TOTAL=16.10

Número de componentes
Principales a considerar

Análisis de Componentes Principales			
Componente Número	Autovalor	Porcentaje de Varianza	Acumulado Porcentaje
1	10,8328	67,261	67,261
2	3,14129	19,504	86,765
3	1,02355	6,355	93,120
4	0,641567	3,983	97,104
5	0,228833	1,421	98,525
6	0,157534	0,978	99,503
7	0,0774517	0,481	99,984
8	0,00262631	0,016	100,000

¿ Análisis de las variables ? Relación o peso de variables en cada componente ???

Correlaciones
E1-E4....0.82
E2-E3.....0.68
E3-E4..... 0.63
E2-E8..... 0.75

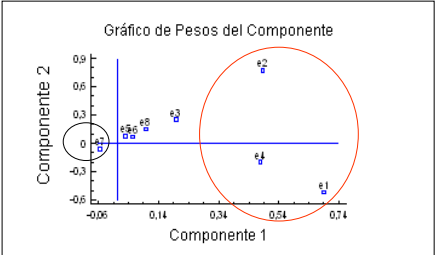
Opciones Tabulares

☐ Resumen del Análisis
☒ Ponderación del Componente
☐ Hoja de Datos

Aceptar Cancelar Todos Ayuda

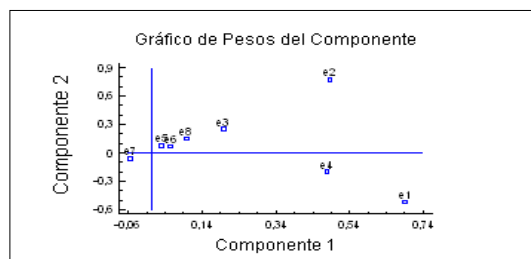
Se resaltan en la primera
componente los ambientes e1,
e2, y e4

Análisis de Componentes Principales			
Tabla de Pesos de los Componentes			
	Componentes 1	Componentes 2	Componentes 3
e1	0,691384	-0,514619	-0,393233
e2	0,487665	0,77339	-0,261338
e3	0,197409	0,249324	0,555622
e4	0,478552	-0,196941	0,627063
e5	0,0305941	0,0757494	-0,0951572
e6	0,0542925	0,0671327	0,099627
e7	-0,0539742	-0,0613133	0,23699
e8	0,0969041	0,148601	-0,00157475



¿Sentido biológico de las componentes a partir de su relación con las variables iniciales?

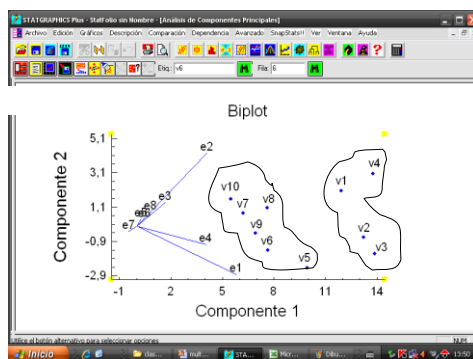
Las variedades de calabaza con mejores rendimiento condiciones de Estrés (e1, e2, e4), serán menos productivas bajo las condiciones de estrés 7



¿Cuáles son las variedades mas favorecidas en estos ambientes?

¿Cuáles son las variedades mas favorecidas en estos ambientes?

Fila	Etiqueta	Componentes 1	Componentes 2	Componentes 3
1	v1	11,9096	2,04532	-1,57279
2	v2	13,2122	-0,668381	1,21131
3	v3	13,8742	-1,63612	0,487246
4	v4	13,7864	3,02855	0,498077
5	v5	9,95575	-2,42717	-1,84293
6	v6	7,63285	-1,42769	0,879301
7	v7	6,19315	0,750749	0,372484
8	v8	7,62124	1,03426	-0,50892
9	v9	6,94236	-0,417853	0,331339
10	v10	5,50406	1,54567	0,334688



CP 1 (e1, e2, e3 y e4)

¿Cuáles son las variedades más favorecidas en estos ambientes?

Existen dos grandes grupos

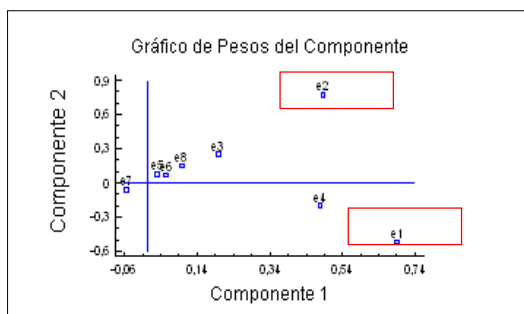
Grupo 1

formado por las variedades 1, 2, 3 y 4

Grupo 2

Formado por el resto de las variedades

¿ Relación o peso de variables en la Segunda Componente ???



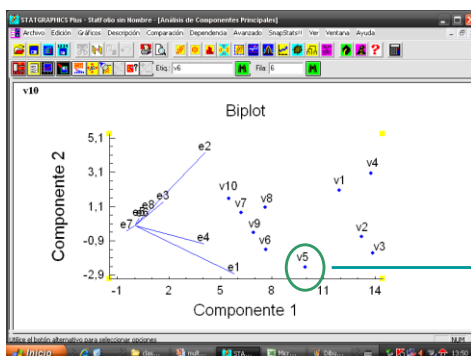
¿Cuáles son las variedades mas favorecidas en estos ambientes?

Análisis de Componentes Principales
Tabla de Pesos de Los Componentes

	Componentes 1	Componentes 2	Componentes 3
e1	0,691384	-0,514619	-0,393233
e2	0,487665	0,77339	-0,261338
e3	0,197409	0,249324	0,555622
e4	0,478552	-0,196941	0,627063
e5	0,0305941	0,0757494	-0,0951572
e6	0,0542925	0,0671327	0,099627
e7	-0,0539742	-0,0613138	0,23699
e8	0,0969041	0,148601	-0,00157475

¿Cuáles son las variedades mas favorecidas en estos ambientes (e2)?

Fila	Etiqueta	Componentes 1	Componentes 2	Componentes 3
1	v1	11,9096	2,04532	-1,57279
2	v2	13,2122	-0,668381	1,21131
3	v3	13,8742	-1,63612	0,487246
4	v4	13,7864	3,02855	0,498077
5	v5	9,95575	-2,42717	-1,84293
6	v6	7,63285	-1,42769	0,879301
7	v7	6,19315	0,750749	0,372484
8	v8	7,62124	1,03426	-0,50892
9	v9	6,94236	-0,417853	0,331339
10	v10	5,50406	1,54567	0,334688

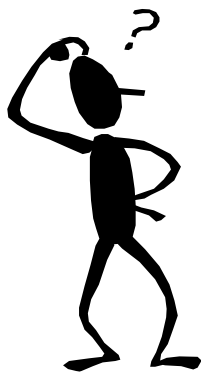


CP 2 (e2) y -e1
¿Cuáles son las variedades más favorecidas en estos ambientes?
v4, v1, v10 y v8
Menos favorecida en el ambiente 2 es la variedad 5

Análisis de Componentes Principales

Algunas consideraciones:

- Resulta más efectivo en la medida que exista una estructura de correlación marcada entre las variables.
 - Puede ser usada para formar grupos de individuos.
 - Si los dos primeros ejes no explican un porcentaje alto de la variación total, cualquier agrupamiento formado carece de sentido.
 - Para asegurar la calidad del agrupamiento, se sugiere combinar esta técnica con un Análisis Factorial Discriminante; o simplemente un Análisis de Clúster.
-



Nueva Problemática



Ejercicio 1



Manual de Estadística para proyectos de Investigación Autor Pereyra A.M; Abbiati, N, N; Fernández, E.N Universidad Nacional de Lomas de Zamora ISBN 987-9455-33-9, Argentina, año 2004) Pág. 124

EJERCICIO 1. Se presentan los datos correspondientes a siete variables socioeconómicas para 14 países (Batista Forguet y Martines Áreas 1989) :
 Densidad de población, porcentaje de personas empleadas en la agricultura, renta (ingreso nacionales per cápita), Inversiones de rendimiento de capital en maquinaria y otros , tasa de mortalidad Infantil, consumo de energía por 100 habitantes, aparatos de TV por 100 habitantes

	pais	densidad	empleo	renta	inversiones	mortalidad	energia	TV
1	Australia	2	6	8,4	10,1	12	5,2	36
2	Francia	97	9	10,7	9,2	10	3,7	28
3	Alemania	247	6	12,4	9,1	15	4,6	33
4	Gracia	72	31	4,1	8,1	19	1,7	12
5	Islandia	2	13	11	6,6	11	5,8	25
6	Italia	189	15	5,7	7,9	15	2,5	22
7	Japon	311	11	8,7	10,9	8	3,3	24
8	NZelandia	12	10	6,8	8	14	3,4	26
9	Portugal	107	31	2,1	5,5	39	1,1	9
10	España	74	19	5,3	6,9	15	2	21
11	Suecia	18	6	12,8	7,2	7	6,3	37
12	Turquia	56	61	1,6	8,8	153	0,7	5
13	Reino Unido	229	3	7,2	9,3	13	3,9	39
14	EU	24	4	10,6	7,3	13	8,7	62

¿Variables?

7 variables diferentes

densidad	empleo	renta
inversiones	mortalidad	energia
TV		

¿Individuos?

países

14 países diferentes

ESTADÍSTICA DESCRIPTIVA. ESTADÍSTICOS UNIVARIADOS				
Etq: Fila:				
Resumen Estadístico				
	densidad	empleo	renta	inversion
Frecuencia	14	14	14	14
Media	102,857	16,0714	7,67143	8,20714
Desviación típica	101,331	15,7307	3,61863	1,46154
Coef. de variación	98,5163%	97,8801%	47,1703%	17,8081%
	mortalidad	energia	TV	
Frecuencia	14	14	14	
Media	24,5714	3,77857	27,0714	
Desviación típica	37,7496	2,21053	14,398	
Coef. de variación	153,692%	58,5017%	53,1852%	

Se observa una alta variabilidad , heterogeneidad de varianza

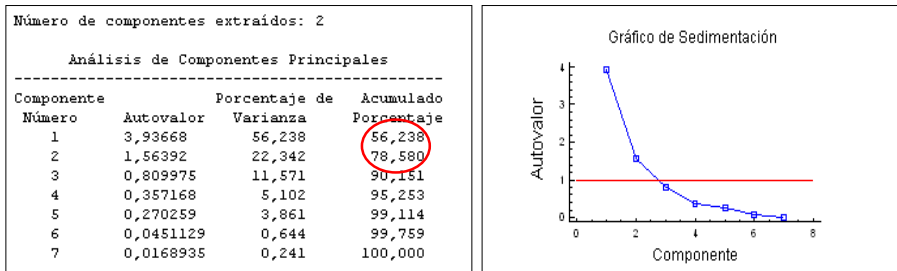
Matriz de correlación

	densidad	empleo	renta	inversiones	mortalidad	energia
densidad		-0,1502 (14) 0,6084	0,0194 (14) 0,9475	0,4902 (14) 0,0751	-0,1315 (14) 0,6542	-0,2548 (14) 0,3794
empleo	-0,1502 (14) 0,6084		-0,7858 (14) 0,0009	-0,1827 (14) 0,5319	0,8897 (14) 0,0000	-0,7147 (14) 0,0041
renta	0,0194 (14) 0,9475	-0,7858 (14) 0,0009		0,1955 (14) 0,5029	-0,6018 (14) 0,0228	0,8300 (14) 0,0002
inversiones	0,4902 (14) 0,0751	-0,1827 (14) 0,5319	0,1955 (14) 0,5029		0,0017 (14) 0,9953	0,0091 (14) 0,9754
mortalidad	-0,1315 (14) 0,6542	0,8897 (14) 0,0000	-0,6018 (14) 0,0228	0,0017 (14) 0,9953		-0,4940 (14) 0,0726
energia	-0,2548 (14) 0,3794	-0,7147 (14) 0,0041	0,8300 (14) 0,0002	0,0091 (14) 0,9754	-0,4940 (14) 0,0726	
TV	-0,0693 (14) 0,8140	-0,7825 (14) 0,0009	0,7216 (14) 0,0036	0,1341 (14) 0,6476	-0,5256 (14) 0,0536	0,9148 (14) 0,0000

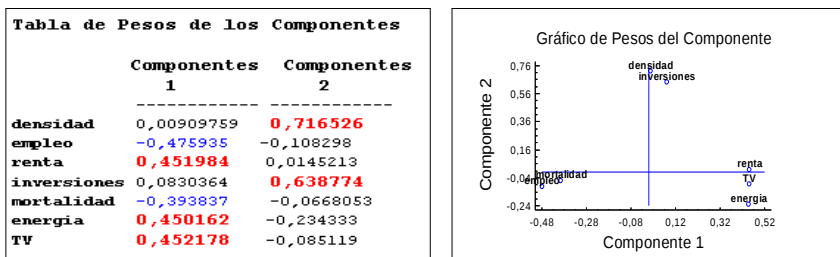
empleo y renta
empleo y mortalidad
empleo y energia
empleo y TV
renta y mortalidad
renta y energia
renta y TV
energia y TV

Densidad de población, porcentaje de personas empleadas en la agricultura, renta (ingreso nacional per cápita), Inversiones de rendimiento de capital en maquinaria y otros, tasa de mortalidad Infantil, consumo de energía por 100 habitantes, aparatos de TV por 100 habitantes

¿Número de componentes Principales a considerar ?

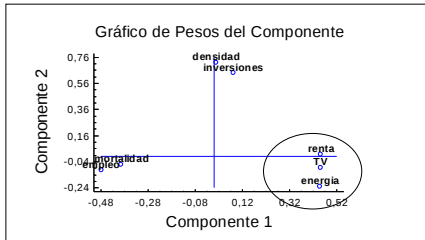


¿ Análisis de las variables ? ¿ Relación o peso de variables en cada componente ???



CP 1 (renta , TV y energía) (-empleo , mortalidad)

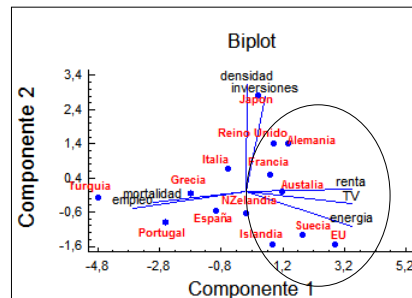
¿Cuáles son los países más favorecidos en estos indicadores?



Los países que reportan mayores rentas (ingresos nacionales per cápita), mejores consumo de energía por cada 100 habitantes y mayor cantidad de aparatos de TV cada 100 habitantes, serán los que tendrán menos tasa de mortalidad infantil y menos porcentaje de personas empleadas en la agricultura.

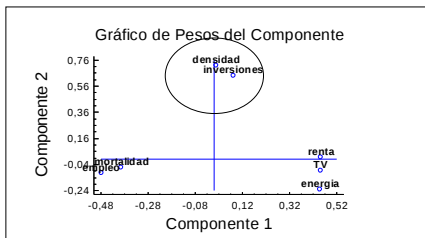
Se puede verificar estos elementos en la matriz de correlación

Fila	Etiqueta	Componentes 1	Componentes 2
1	Australia	1,19523	0,00514973
2	Francia	0,813296	0,48198
3	Alemania	1,41233	1,39261
4	Gracia	-1,7451	-0,0628254
5	Islandia	0,896505	-1,55911
6	Italia	-0,543329	0,656813
7	Japon	0,432559	2,78599
8	NZelandia	0,0544676	-0,629521
9	Portugal	-2,56455	-0,913759
10	España	-0,914678	-0,563651
11	Suecia	1,88907	-1,24519
12	Turquia	-4,74817	-0,17634
13	Reino Unido	0,930085	1,3948
14	EU	2,8923	-1,56695



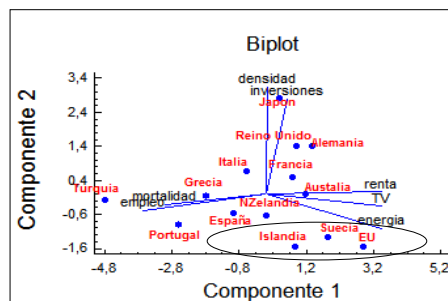
CP 2 (aporta un 22,34% de la varianza) Densidad de población e Inversiones

¿Cuáles son los países más favorecidos en estos indicadores?



La segunda componente se caracteriza por agrupar a Japón, Reino Unido y Alemania con elevados valores en densidad de población e inversión de capital en maquinaria a los que se le opone el grupo de Islandia, Suecia y Estados Unidos que alcanzan bajos valores en esas dos variables

Fila	Etiqueta	Componentes 1	Componentes 2
1	Australia	1,19523	0,00514973
2	Francia	0,813296	0,48198
3	Alemania	1,41233	1,39261
4	Gracia	-1,7451	-0,0628254
5	Islandia	0,896505	-1,55911
6	Italia	-0,543329	0,656813
7	Japon	0,432559	2,78599
8	NZelandia	0,0544676	-0,629521
9	Portugal	-2,56455	-0,913759
10	España	-0,914678	-0,563651
11	Suecia	1,88907	-1,24519
12	Turquia	-4,74817	-0,17634
13	Reino Unido	0,930085	1,3948
14	EU	2,8923	-1,56695



EJERCICIO 3

Se realizó un estudio pos cosecha de la Piña (*variedad Cayena Lisas*), en áreas de la empresa de cultivos varios ubicados dentro de la Llanura Habana-Matanzas, donde se observó durante los días (1, 2, 3, 5, 6, 8 y 10) las variables como: Pérdida de Peso (PP), firmeza, índice de color (IC), contenido de sólidos solubles (SSC) y PH. Se desea analizar el comportamiento de estas variables en los días pos cosecha.

(5 variables)

días (6 individuos).	días	PERDPESO	FIRMEZA	IC	SSC_1	PH_1
	1	0	6.68	-0.717	13.7	4.38
	2	0.507	4.861	0.038	14.602	4.462
	3	0.803	4.43	0.42	13.96	4.308
	5	1.285	4.005	0.48	13.35	4.342
	6	3.422	3.655	0.521	13.51	4.546
	8	4.552	3.575	-0.092	14.36	4.794
	10	4.368	3.285	-0.071	14.6	5.093

Number of components extracted: 2

Principal Components Analysis			
Component Number	Eigenvalue	Percent of Variance	Cumulative Percentage
1	2.77171	55.434	55.434
2	1.64655	32.931	88.365
3	0.519198	10.384	98.749
4	0.0505872	1.012	99.761
5	0.0119526	0.239	100.000

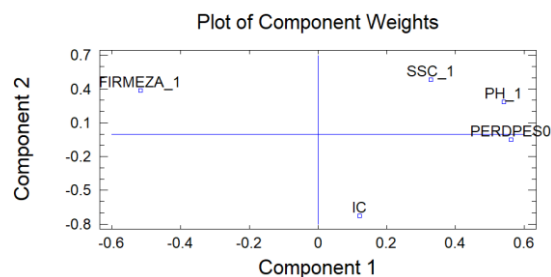
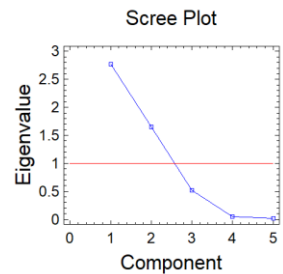
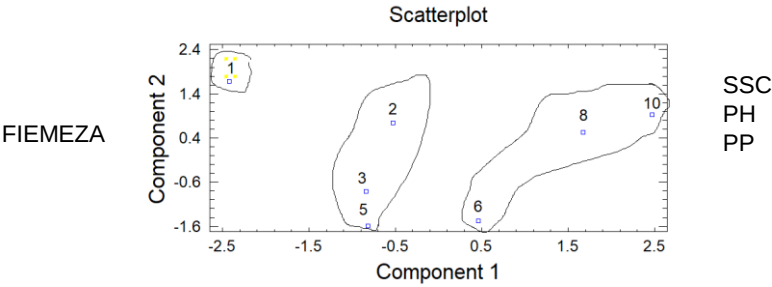
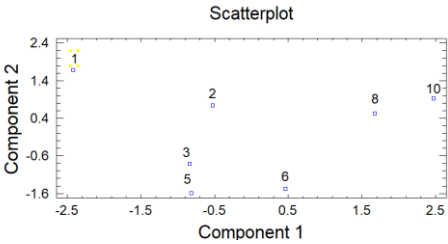


Table of Component Weights

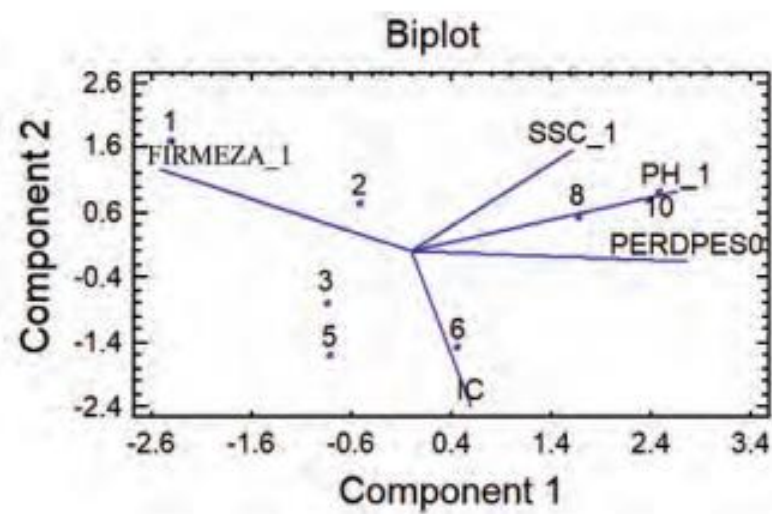
	Component 1	Component 2
PERDPESO	0.562604	-0.048823
FIRMEZA_1	-0.516914	0.386097
IC	0.120952	-0.728645
SSC_1	0.329301	0.484943
PH_1	0.541487	0.287145

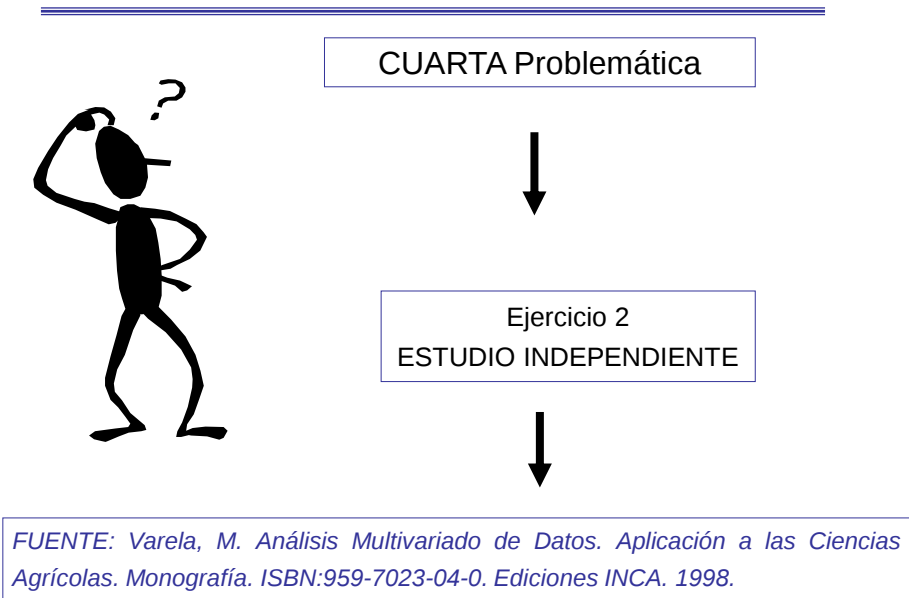
Pérdida de Peso (PP), firmeza, índice de color (IC), contenido de sólidos solubles (SSC) y PH

Row	Label	Component 1	Component 2
1	1	-2.42274	1.68195
2	2	-0.524755	0.738421
3	3	-0.840422	-0.802432
4	5	-0.817106	-1.59375
5	6	0.463305	-1.4782
6	8	1.67103	0.529105
7	10	2.47068	0.926015



Pérdida de Peso (PP), firmeza, índice de color (IC), contenido de sólidos solubles (SSC) y PH





Se estudió el comportamiento de 18 variedades de boniato, a las que se le evaluaron un total de 10 variables

varied	Peso	Largo	Ancho	Protein	SolidosTot	SolidosSol	Amin	AzucToles	Cenizas	Fibra
v1	191,2	11	4,5	1,99	47,16	9,2	37,69	5,95	1,46	1,28
v2	150,4	13,5	5	1,75	37,14	9	28,14	4,39	1,53	1,51
v3	155,7	15	6,5	1,91	35,65	9,5	29,35	4,43	1,25	1,26
v4	175,8	16	4,5	1,89	33,73	8,8	25,22	3,62	0,94	1,42
v5	160	14,5	3,5	2,03	31,81	10	25,45	4,12	1,31	1,02
v6	146,7	15	4	2,59	38,76	10,4	27,65	5,09	1,09	1,73
v7	231,3	20,5	5,5	1,94	33,76	10	26,22	3,92	0,79	1,19
v8	172,5	12,5	5	1,21	31,17	10,8	23,39	3,65	0,75	1,36
v9	240,2	24	7	2,14	32,73	12	24,63	4,56	0,82	1,01
v10	151,7	15	3,5	1,9	30,91	8,2	25,64	2,83	1,07	1,21
v11	220,1	20,5	3,5	1,64	31,25	8,3	24,61	3,02	0,7	1,19
v12	167,2	15	6,5	1,29	31,55	8,7	24,63	3,29	1,31	1,31
v13	202,2	20,5	4	1,85	33,38	11,2	27,62	4,86	1,13	1,1
v14	204,1	17	4,5	1,64	36,11	10,6	28,73	4,83	1,21	1,38
v15	217,8	17,5	4,5	1,22	31,11	9,4	19,9	3,38	1,01	1,25
v16	241,7	16	3,5	2,19	40,44	10,3	31,02	4,55	0,97	1,66
v17	237,8	23,5	6	1,65	31,41	10,2	24,19	3,98	0,87	1,11
v18	176,8	15,5	4	1,73	30,21	10	23,61	3,71	0,82	1,04
