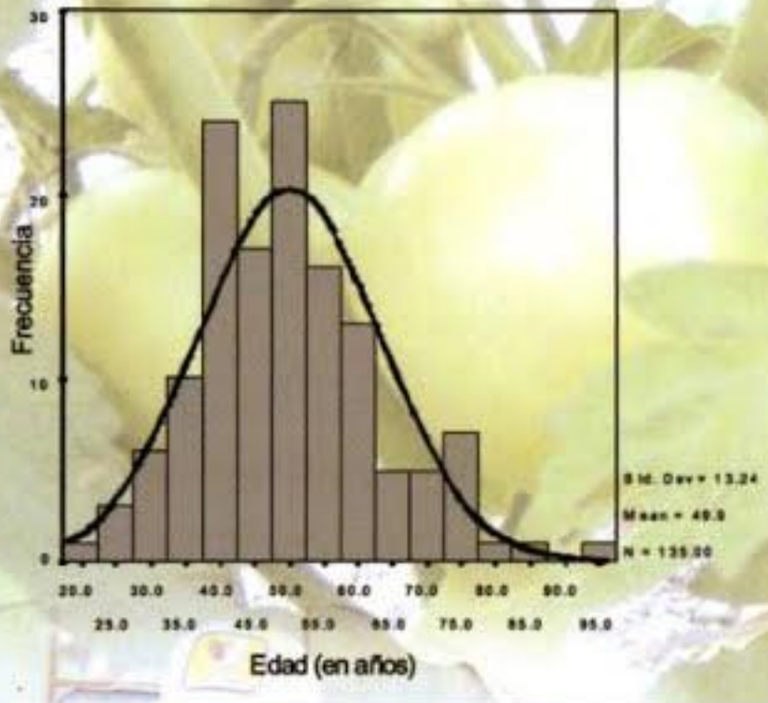
INTA

Instituto Nicaragüense de Tecnología Agropecuaria

IICA

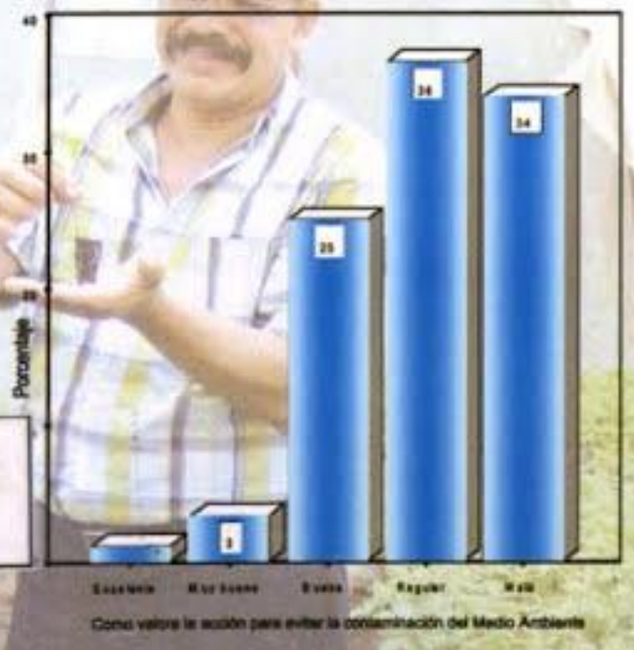
Sistema de Análisis Estadístico con SPSS

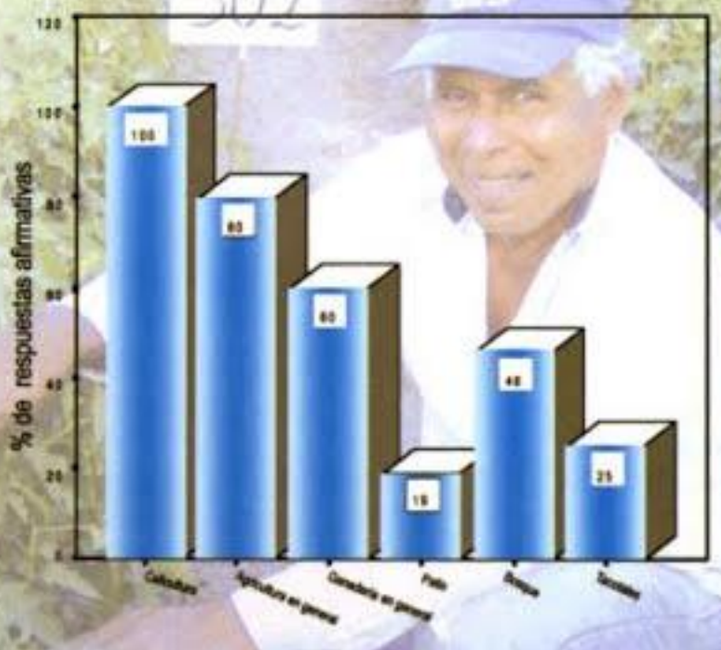


One-Sample Kolmogorov-Smirnov Test

		Edad (en años)
N		135
Normal Parame	Mean	49.92
	Std. Deviaton	13.24
Most Extreme	Absolute	.077
Differences	Positive	.077
	Negative	-.037
Kolmogorov-Smirnov Z		.896
Asymp. Sig. (2-tailed)		.399

a. Test distribución
b. Calculated from data





Dr. Henry Pedroza
Ing. MSc. Luis Dicovskyi

Managua, Nicaragua
Mayo, 2006

Copyrighted material

00002541

Copyrighted material

© Instituto Interamericano de Cooperación para la Agricultura (IICA).
Instituto Nicaragüense de Tecnología Agropecuaria (INTA), 2007.

El Instituto promueve el uso justo de este documento.
Se solicita que sea citado apropiadamente cuando corresponda

Esta publicación también está disponible en formato electrónico (PDF)
en el sitio Web institucional en www.iica.int.

Coordinación editorial: Esperanza Rodríguez
Corrección de estilo: Nestor Allan Alvarado Díaz
Diagramado: Hauny Mendieta
Diseño de portada: Ivana Alvarado
Impresión: LITONIC

Sistema de análisis estadístico con SPSS
Henry Pedroza, Luis Dicovskyi – Managua: IICA,
INTA, 2007.
167 p. ; 21.59 X 27.44 cm

ISBN13: 978-92-9039-790-8

1. Estadística. 2. SPSS I. IICA II. INTA III. Título

AGRIS E10
DEWEY 310

Copyrighted material

PROLOGO

El contenido de este libro, es muy original, ya que los autores Dr. Henry Pedroza Pacheco y MSc. Luis Elías Dicovskiy, realizan un minucioso estudio, señalando los múltiples aspectos que deben tomarse en cuenta en relación al análisis de variables cuantitativas o cualitativas, organizadas en un sistema de bases de datos, (DBMS). Los autores profundizan sobre las técnicas de análisis estadístico paramétricas y no paramétricas, apoyados tanto en información bibliográfica como en su propia experiencia. Todos los análisis presentados en este libro, son realizados con datos originales que corresponden a tesis de grado o maestría, o son datos de fuentes primarias, obtenidos a través de experimentos propios o consultarías a las cuales los autores estuvieron vinculados como asesores o como autores de las mismas.

Considero que este libro será de gran utilidad para diversas instituciones académicas, ONG´s, centros de investigación, tecnológicos o socio-económicos, las cuales realizan diversos estudios de investigación experimental y/o no experimental, tales como Líneas de Base, Diagnósticos, Evaluación de Impacto, Estudios Prospectivos, y toda la gama de estudios experimentales univariados o multivariados. También será de utilidad para los estudiantes de diversas universidades donde imparten carreras como Ingeniería Agronómica y Desarrollo Rural; Economía Aplicada, Administración de Empresas, Ingeniería de Sistemas de Producción, en Ingeniería Industrial, y otras.

Por otra parte, las técnicas de análisis estadísticos expuestas en este libro, serán de gran utilidad para los técnicos de los centros experimentales y extensionista del INTA, quienes encontrarán en este libro una guía tutorial para realizar con mayor efectividad el análisis estadístico de sus datos y mejorar así su desempeño profesional en el área de investigación, tan importante para el desarrollo y sostenibilidad económica de Nicaragua.

Felicito de manera muy especial a los autores de este nuevo libro, que en tres ocasiones, una en el año 2005, el 2006 y el 2007, han impartido exitosamente en el IICA sede Nicaragua, un curso con este mismo texto, demostrando su calidad profesional y la pertinencia de esta obra, por lo cual no dudo que será de mucha importancia tanto para la docencia universitaria como para los procesos de investigación e innovación tecnológica del país.

Dr. Gerardo Escudero
Director General IICA, Nicaragua

Copyrighted material

PROLOGO

En una realidad tan dinámica como la que actualmente se vive, los profesionales, técnicos y estudiantes que realizan investigación en Nicaragua han de sentirse altamente estimulados con la aparición de esta nueva obra del Dr. Henry Pedroza Pacheco y el MSc. Luis Elias Dicovskiy Riobóo, que versa sobre: "Sistema de Análisis Estadístico con SPSS".

Este libro será de gran utilidad para los estudiantes de la Universidad Nacional Agraria que cursan las asignaturas Experimentación Agrícola, en las carreras de Ingeniería Agronómica, Ingeniera Agrícola para el Desarrollo Sostenible e Ingeniería Forestal; Investigación Agrícola y Forestal, en la carrera de Ingeniería en Sistemas de Protección Agrícola y Forestal; Experimentación Pecuaria, en las carreras de Ingeniería en Zootecnia y Medicina Veterinaria.

En el libro "Sistema de Análisis Estadístico con SPSS", se realizan los análisis estadísticos de datos provenientes de investigación experimental y no experimental, con el software SPSS. En los primeros cuatro capítulos del libro, se enseña a manejar una amplia gama de comandos para realizar el procesamiento estadísticos de datos, tales como: estadística descriptiva, generación de gráficos univariados y multivariados, tablas de contingencia y medidas de asociación. En los capítulos del quinto al décimo, se realiza el análisis de varianza para datos provenientes de experimentos unifactoriales o multifactoriales; también se ejemplifica el uso de los modelos de regresión lineal simple y curvilínea. En los últimos cuatro capítulos del libro se profundiza sobre Técnicas de Análisis Multivariadas, tales como: el análisis de regresión múltiple, el análisis multivariante de la varianza, técnicas de análisis cluster, y el análisis discriminante.

Quiero expresar mis felicitaciones al Dr. Pedroza y al MSc. Luis Dicovskiy, profesionales de mucha experiencia tanto en docencia universitaria como en investigación, por lo que estoy seguro del gran éxito que tendrá esta nueva obra.

Ing. Agr. Néstor Allan Alvarado Díaz
Master en Sistemas Integrales de Producción en el Trópico
Jefe del Departamento de Ingeniería Agrícola,
Facultad de Agronomía, UNA.

Copyrighted material

INDICE GENERAL

Copyrighted material

Copyrighted material

Copyrighted material

INDICE DE CUADROS

Cuadro *Página*

Copyrighted material

Copyrighted material

Copyrighted material

INDICE DE FIGURAS

Copyrighted material

Copyrighted material

Capítulo 1. Sistema de Análisis Estadístico con SPSS

1.1 Introducción.

Un sistema de información ya sea documental, estadístico o geográfico, en general es una herramienta de apoyo al desarrollo institucional y los datos que contiene requieren que sean analizados para lograr respuestas concretas sobre el o los problemas que se desean resolver. Por lo tanto, los sistemas de información son instrumentos de trabajo para los analistas, técnicos, asesores de políticas y tomadores de decisiones, para apoyar los procesos de innovación y desarrollo a nivel regional, nacional y/o local. En la figura 1.1, se ilustran las interrelaciones necesarias a partir de los problemas identificados, hasta la selección de una alternativa de solución en el proceso de toma de decisiones, (Pedroza, H.P. 1995).

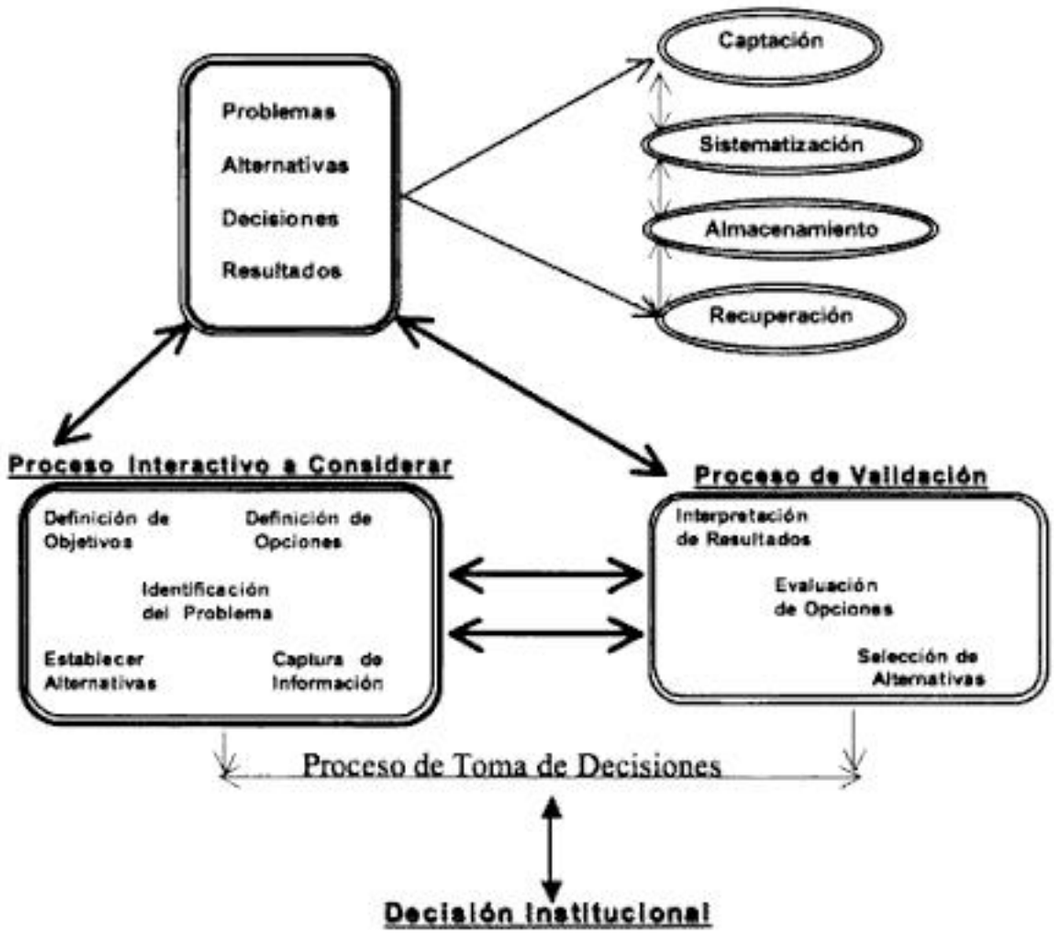


Figura 1.1. Los sistemas de información como instrumentos de apoyo institucional.

En términos estrictamente técnicos, los sistemas de información deben entenderse como medios importantes para capturar, almacenar, procesar y recuperar la información pertinente, por medio de reportes de salida, para sustentar las decisiones institucionales que tengan que implementarse. Hoy en día, las organizaciones confían en la técnica de investigación cuantitativa, ***la encuesta***, para conocer mejor y aprender lecciones en tiempo real sobre los problemas que deben resolver.

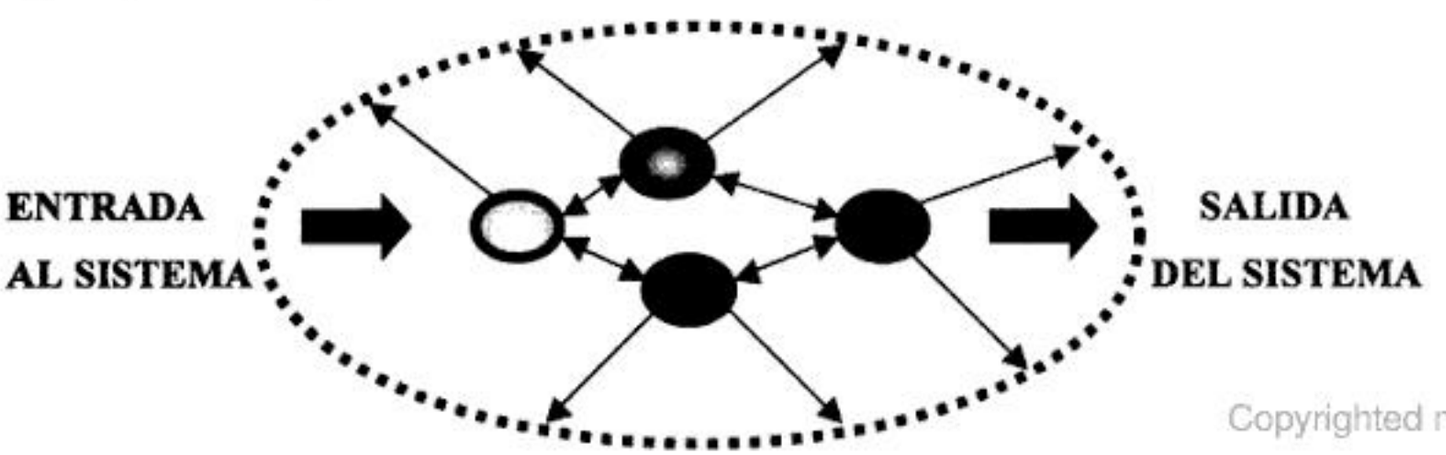
Copyrighted material

El problema principal a atender para muchas instituciones u organizaciones del sector agropecuario nacional, es que no cuentan con un sistema de información gerencial integrado en diversos aspectos, (producción, educación, población, salud, etc.), que facilite el proceso de identificación, standardización, captura y devolución de la información a usuarios, para mejorar el acceso/ disponibilidad de nuevos conocimientos y tecnologías, no solo de Nicaragua, sino de América Central, Latino América y el mundo, generadas por los diferentes actores del Sistema de Innovación Tecnológica de cada país, (Universidades, ONG`s, Centros de Investigación Públicos y Privados, Programas Internacionales de cooperación, etc).

Actualmente, en la "era de la información y el conocimiento", hay mucha información tecnológica dentro del país como tal, pero aún hace falta un serio impulso en cuanto a la gestión del conocimiento para evidenciar, sistematizar esa información, y ponerla en línea al servicio del usuario interesado: ***"traducirla en conocimientos y tecnologías disponibles, accesibles y útiles para sus usuarios"***, productores, técnicos, docentes, estudiantes, empresas de asistencia técnica privada y cofinanciada, etc. Por lo tanto, la información existente dispersa y **no** automatizada, dificulta los procesos de toma de decisión gerencial, soportados en un manejo efectivo de la información existente a nivel regional, nacional y local.

Toda esta información, ejemplo datos de producción, educación, salud, tecnología, población, etc., podría estandarizarse y hacerse disponible a los usuarios, mediante estadísticas útiles, si se constituyen en un Sistema de Análisis Estadístico con el software ***Statistical Packet for Social Science***, (**SPSS**). Implementando métodos de investigación cuantitativa, realizados de manera correcta, el SPSS proporcionará a las organizaciones mucha información valiosa, basado en datos confiables, para fundamentar las decisiones institucionales. El análisis de un sistema de análisis estadístico, con la flexibilidad del SPSS, le ayudan al usuario a responder las preguntas que se desean sobre diversos problemas a resolver.
Como todo sistema, el SPSS, funciona mediante un conjunto de elementos relacionados entre sí con un propósito determinado. Esto es lo que caracteriza al SPSS, ya que es un sistema diseñado para cumplir el propósito de aportar soluciones, mediante el análisis estadístico para el cual existe.

El SPSS funciona como un verdadero sistema, ya que maneja de forma integrada un sistema de base de datos (DBMS), con el que interactúan un conjunto ordenado de módulos y comandos, los cuales están estructurados y relacionados para efectuar los procedimientos estadísticos, sobre las entradas o variables y producir las salidas deseadas o reportes. Lo que el SPSS analiza como sistema son los elementos definidos por las variables de estudio y las relaciones entre ellas. Esta es la estructura imaginaria en la cual se basa el SPSS, para realizar los análisis estadísticos que transforman los insumos (input/datos) en productos (output/hojas de salidas).



Copyrighted material

La idea básica con que funciona el SPSS como sistema de análisis estadístico es, que el conjunto de elementos organizados, dado por las variables cualitativas y cuantitativas, se encuentran en una interacción o flujo continuo, buscando como cumplir una meta común. Para lograr tal meta, el SPSS actúa sobre los datos organizados en una estructura lógica, dentro de una base de datos, para producir la información de salida, como nuevos hallazgos o conocimientos, resultado del análisis estadístico realizado.

Como sistema, el SPSS establece diferentes relaciones entre sus elementos (variables), las que dentro del SPSS son relaciones sinérgicas fundamentalmente, ya que interactúa con c/u de los módulos (subsistemas) y entre los comandos (sub-subsistemas dentro de los módulos), es decir, se refuerzan entre sí para obtener los objetivos comunes; esto hace que el SPSS se comporte como un sistema ideal con relaciones optimizadas entre las variables.

Es importante destacar que, en el sistema de análisis estadístico con el SPSS, las interrelaciones entre las variables son orientadas al logro del resultado deseado, con mucha efectividad, de manera que sus procedimientos de análisis estadístico conduzcan al objetivo deseado, todo ello realizado en un ambiente gráfico, en el cual todo el sistema del SPSS está inmerso.

Tal como se expresa en el sitio web, http://www.spss.com/la/soluciones/analisis-encuestas.htm, el sistema de análisis estadístico con el SPSS, facilita construir bases de datos desde internet, incluir datos de cuestionarios, desde hojas electrónicas, desde otras bases de datos, o encuestas telefónicas y/o entrevistas personales, etc. Estas herramientas reducen en gran medida el tiempo empleado en preparar los datos extraídos de las encuestas para realizar el análisis de la información. El SPSS, ofrece a los usuarios una amplia gama de estadísticas y técnicas para el procesamiento de datos, siempre que se cuente con la preparación apropiada para realizar el análisis de los datos. Dependerá del interés del usuario elegir el procedimiento que considere más adecuado según los objetivos y el tipo de variables en estudio para obtener los reportes de salida que desee, (SPSS, 2004).

1.2 Operación de Variables en SPSS.

Para iniciar el reconociendo del sistema de análisis estadístico con el SPSS, se comienza por observar la barra de herramientas en la ventana de aplicación -dentro del programa SPSS-, para lo cual se cita la experiencia expuesta por el Profesor Luis Dicovskyi, en Pedroza y Dicovskiy, 2003.

Barra de Herramientas

Debajo del menú principal se encuentra la barra de herramientas, esta posee los siguientes comandos en forma de iconos.

Open File Abre documentos
Save File, Graba documentos
Print, imprime.
Dialog Recall, muestra un listado de operaciones recientes echas con SPSS. Con un clic se puede entrar en una de ellas.
Goto Chart , activa un gráfico fuera del archivo, es una ventana.

Copyrighted material

Goto case, permite buscar un caso determinado.
Variables, presenta información de las variables.
Find, permite buscar un dato dentro de una variable.
Insert case, permite insertar una fila.
Insert variable, permite insertar una columna.
Split File, permite partir la base de datos según los valores de una variable.
Wheight cases, permite analizar una variable según los valores de otra variable.
Select cases, permite seleccionar casos que cumplan unas condiciones dadas.
Value labels, muestra las etiquetas de las variables.
Use sets, **selecciona un conjunto de variables predefinidas, para usarlas en un análisis.**

1.3 Definición de Variables en SPSS.

Al definir una variable nueva se deben distinguir los siguientes aspectos que aparecen al hacer doble clic sobre ella.

Nombre de la variable, deben cumplir los siguientes requisitos:

- Máximo de 8 caracteres.
- Deben comenzar con una letra y no pueden terminar con un punto.
- No pueden tener espacios en blanco ni caracteres especiales (¡,?,*...).
- No puede haber dos nombres de variables repetidas.
- El programa no distingue entre mayúsculas y minúsculas
- No se pueden usar las letras: ALL, LT, AND, NE, AND, NE, BY, NOT, EQ, OR, GE, TO, GT, WIDT, LE.

Tipo, por defecto el SPSS asume que las variables son numéricas, puede cambiarse el tipo pulsando el botón TYPE... Los tipos de variables son:

- Numérica, admite valores numéricos, signos de + o -, decimales y notación exponencial. El ancho máximo es de 40 caracteres y el número de decimales 16.
- Coma, añade a lo anterior la posibilidad de una coma para la separación de miles.
- Dot, Funciona como el anterior, pero cambia comas por puntos.
- Fechas. Define variables con formato predefinido de fecha.
- Dóllar.
- Monedas.
- Atributos.
- String, para variables de texto.

Label (Etiquetas), permite dar nombre completo a las variables y asignar un nombre a cada valor de la variable, ***esto se hace con variables discretas de pocos valores***.

Copyrighted material

Missing values (valores perdidos), los que no serán tomados en cuenta al analizar, hay dos tipos de "missing values":

- Los del sistema, cualquier casilla en blanco dentro de la matriz de datos.
- Los del usuario, en este caso se debe entrar en ***"define missing values"*** y especificar el valor que se asigna como "missing". En las variables discretas de trabajo (un máximo de tres valores), en las variables continuas un intervalo o un intervalo y un valor fuera del mismo.

Formato de columna, se puede cambiar el ancho de la columna y la alineación del texto en el cuadro de diálogo de ***"Defin Column Format"***. El ancho definido no afecta los valores grabados en el archivo.

Plantilla (Template), **Si se abre** Data/Template, **se puede crear y guardar una plantilla para asignar formato a todo un grupo de variables. Se pueden hacer plantillas para variables como: días de la semana, meses del año, SI - NO, Varón - Mujer, etc. Hay que usar el comando** Define **para definirla, y** add **para que quede guardada en la memoria.**

1.4 Introducción de Datos al Sistema SPSS.

La celda activa aparece con un borde más grueso e identificado en la parte superior izquierda de la pantalla del Editor. **Una vez introducido un dato, dar Enter**. ***A manera de ejercicio inicial, introduzca los datos de* "Rendimiento Académico22"; *de esta forma, ya se esta creando una base de datos dentro del SPSS, con los datos del grupo de clase.***

Las variables a introducir son: 1) Número de boleta (variable numérica); 2) Nombre del estudiante (variable nominal); 3) Edad (variable numérica); 4) Sexo (variable dicotómica); 5) Procedencia (variable cuantitativa discreta: 1) Urbano, 2) Rural, 3) Periferia urbana); 6) Profesión (variable cuantitativa discreta); 7) Nombre del Municipio (variable cuantitativa discreta); 8) Estrato (variable cuantitativa continua); 9) Nota 1 (variable cuantitativa continua); 10) Nota 2 (variable cuantitativa continua); 11) Nota 3 (variable cuantitativa continua); 12) Dominio del tema (variable Likert); 13) Valoración de la Metodología (variable Likert); 14) Valoración del curso con el SPSS (variable Likert); 15) Pertinencia del curso (variable dicotómica Si/No).

1.5 Procedimiento Básico para realizar el Análisis Estadístico con el SPSS.

Se deben cumplir tres operaciones básicas: 1) Seleccionar una base de datos; 2) Seleccionar el procedimiento estadístico deseado; y 3) Seleccionar las variables a incluir en el análisis y otros parámetros adicionales. En el presente texto, se toma como guía el procedimiento básico para realizar el análisis estadístico con el SAS, realizado por Pedroza P.H. (1995), pero superando el sistema de manejo de base de datos (DBMS) que le falta al SAS. Este salto cualitativo, se logra mediante el uso del SPSS, el cual tiene integrado un DBMS, a diferencia del SAS, que integra los procedimientos de análisis estadístico con las variables a analizar dentro de un programa particular.

Copyrighted material

1.6 Control de Calidad de Datos.

Para iniciar a analizar los datos, la primera actividad sugerida es hacer un **"análisis exploratorio de los datos"** con el comando **Analyze/ Descriptive Statistics/ Explore.** Este comando ofrece una serie de opciones para representar gráficamente los datos, examinar visualmente las distribuciones de los valores, detectar valores anormales y realizar pruebas de normalidad con variables continuas.

El análisis exploratorio previo es útil para: a) Detectar errores en los datos, observando los valores anormales; b) Observar la distribución de los datos, permite conocer si hay valores extremos, variabilidad inesperada, rango de datos vacíos o un patrón extraño en el comportamiento de los datos; c) Preparar los datos para pruebas de hipótesis posteriores, esto en función de la distribución observada. Se puede incluso concluir que los datos deban sufrir transformaciones para prepararlos para un determinado análisis.

Copyrighted material

Capítulo 2. Estadísticas Descriptivas.

2.1 Análisis Descriptivo de una Variable Cualitativa en Escala Nominal.

Este análisis es aplicable en situaciones en que los valores de una variable son **no** numéricos, **con ausencia de orden entre ellos**. Se dice que **la variable** correspondiente **es de tipo cualitativo y** que la escala de medida de sus posibles valores **es nominal**. La presentación de datos cualitativos suele hacerse indicando los atributos considerados y su frecuencia de aparición. La tabla que recoge las frecuencias de las modalidades, se denomina ***Distribución de Frecuencias de la variable***, (Ferran, A. M., 1996).

Para ilustrar el análisis descriptivo de una variable cualitativa en escala nominal, se utilizará la variable "escolaridad", en la Base de Datos *"SURVEY33"*. En lo sucesivo el término Base de Datos será BDD. Para desarrollar el **análisis de Frecuencia**, se utiliza la rutina de comandos siguiente: **Analyze/Descriptives Statistics/Frequencies/** en la ventana de diálogo **Variable(s)**, debe incluirse la variable a analizar, en este caso se incluye la variable *Escolaridad*; se debe seleccionar **Display frequency tables**; luego en la opción **Charts**, se seleccionan ***Bar charts, Percentages*** o ***frequency*** y ***Continue***. En la opción **statistics**, se selecciona ***Mode***. El análisis de frecuencia facilita obtener tablas de Frecuencias, gráficos de barras e histogramas, cálculo de percentiles, índice de tendencia central e índices de dispersión. La salida solicitada al SPPS, es la siguiente.

Cuadro 2.1. Análisis de frecuencia de la variable cualitativa en escala nominal, "escolaridad".

Escolaridad de la persona

		Frequency	Percent	Valid	Cumulative
Valid	Primaria Incompleta	122	23.7	27.5	27.5
	Primaria Completa	54	10.5	12.2	39.7
	Secundar Incompleta	114	22.2	25.7	65.5
	Secundaria Completa	107	20.8	24.2	89.6
	Técnico medio	21	4.1	4.7	94.4
	Analfabeta	25	4.9	5.6	100.0
	Total	443	86.2	100.0	
Missing	0	71	13.8		
Total		514	100.0		

La tabla de frecuencias presenta la información en columnas con las descripciones siguientes:

- Valid: Que muestra la etiqueta definida para cada categoría.
- Frequency. Que muestra la frecuencia absoluta para cada categoría.
- Percent. Que muestra la frecuencia relativa, incluyendo valores perdidos.
- Valid Percent. Que muestra la frecuencia relativa, eliminando de la muestra los valores perdidos
- Cumulative Percent. Que muestra la frecuencia relativa acumulada, sin los valores perdidos.

Copyrighted material

El gráfico solicitado al SPSS, es el siguiente:

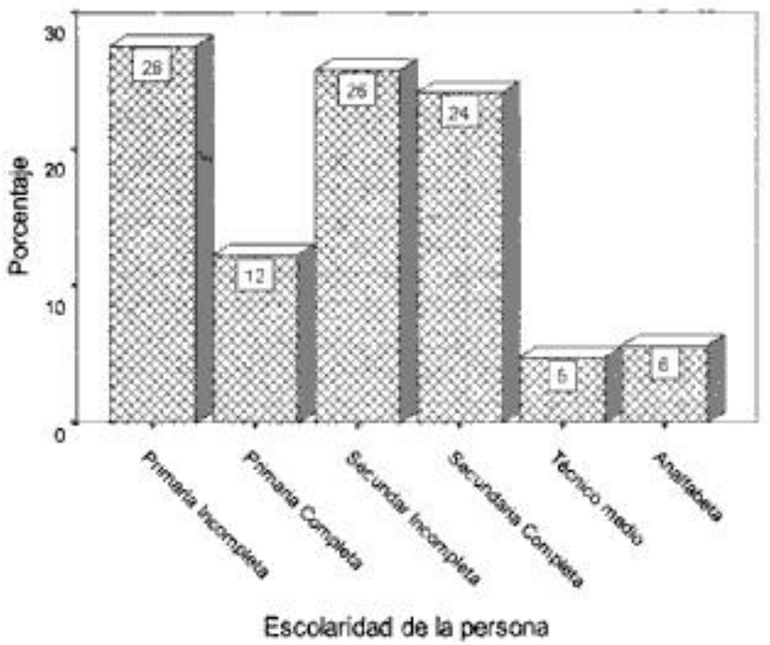


Figura 2.1. ***Porcentaje de escolaridad de las personas encuestadas.***

En este caso, el valor de la Moda solicitado es = 1. Esto indica que el valor que más se repite es la escolaridad "Primaria incompleta". El análisis del gráfico de barras, es el correcto para variables cualitativas en escala nominal, dado que son variables definidas en categorías con valores asignados de una variable discreta. Sin embargo, en el caso de variables continuas, el gráfico de histogramas -solicitando la frecuencia del mismo-, es el análisis gráfico correcto que se recomienda realizar.

2.2 Análisis Descriptivo de una Variable Cualitativa en Escala Ordinal.

Este análisis es aplicable en situaciones en que los valores de una variable son **no** numéricos, **con presencia de orden entre ellos**. Se dice que **la variable** correspondiente **es de tipo cualitativo y** que la escala de medida de sus posibles valores **es ordinal**. La tabla que recoge las frecuencias de las modalidades, se denomina ***Distribución de Frecuencias de la variable***. En el ejemplo anterior, por ser la variable medida en escala nominal, no tenia sentido analizar la acumulación de los porcentajes. No obstante, en el caso de una variable cualitativa en escala ordinal, si tiene sentido estudiar la suma de los porcentajes correspondientes a todos los valores inferiores o iguales a uno dado, (Ferran, A. M., 1996). Este análisis es apropiado para variables de tipo "Likert".

Para ilustrar el análisis descriptivo de una variable cualitativa en escala ordinal, se utilizará la Base de Datos *"SURVEY33"*, donde se encuentra la variable de tipo **"Likert", *"Como valora la accion para evitar la contaminacion del Medio Ambiente"***. Para el análisis de Frecuencia, se utiliza la rutina de comandos: **Analyze/Descriptives Statistics/Frequencies/** en la ventana de diálogo **Variable(s)**, debe incluirse la variable a analizar, en este caso se incluye la variable *Como valora la accion para evitar la contaminacion del Medio Ambiente*; se debe seleccionar **Display frequency tables**; luego en la opción **Charts**, se seleccionan ***Bar charts, Percentages*** y ***Continue***. En la opción **statistics**, se selecciona *Mode, Median, y Percentile 25, 50 y 75*. El análisis de frecuencia solicitado al SPPS, es el siguiente.

Copyrighted material

Cuadro 2.2. ***Análisis de frecuencia de la variable cualitativa en escala ordinal, "Como valora la accion para evitar la contaminacion del medio ambiente"***

Como valora la accion para evitar la contaminacion del Medio Ambiente

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Excelente	6	1.2	1.3	1.3
	Muy buena	16	3.1	3.4	4.6
	Buena	119	23.2	24.9	29.6
	Regular	174	33.9	36.5	66.0
	Mala	162	31.5	34.0	100.0
	Total	477	92.8	100.0	
Missing	0	37	7.2		
Total		514	100.0		

El análisis del gráfico de barras, es el correcto para variables cualitativas en escala ordinal por cuanto son variables definidas en categorías con valores asignados de una variable discreta.

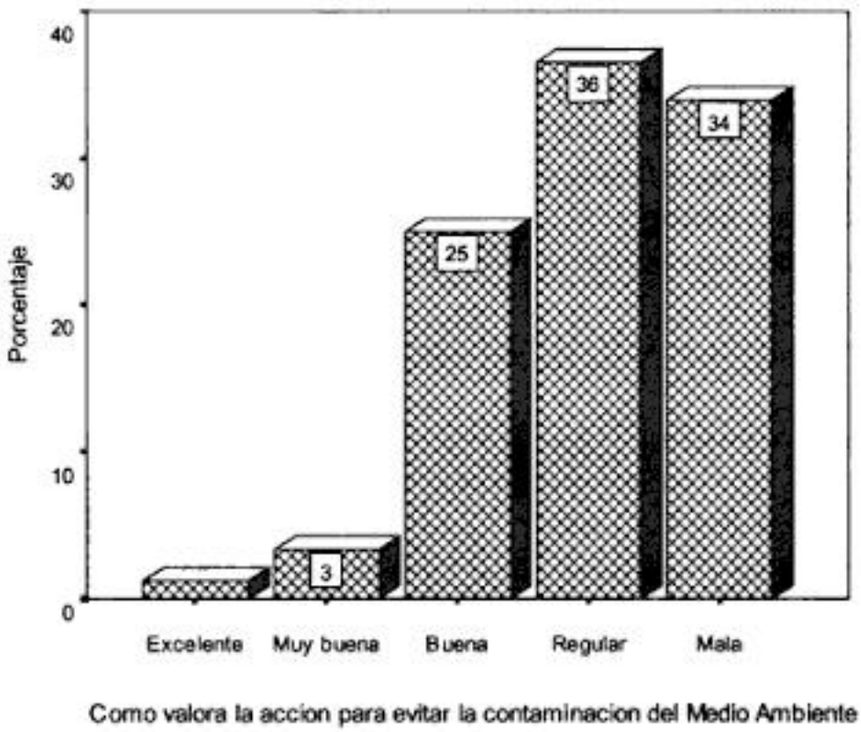


Figura 2.2. ***Porcentajes sobre valoración de accion para evitar contaminacion ambiental.***

El comando **statistics**, da un valor de la Moda = 4. Esto significa que el valor más frecuente es la categoría "Regular"; y la mediana es = 4, lo que indica que los datos ordenados según su magnitud, la categoría que esta en el centro es "Regular", esto significa que el 50 % de los datos es menor o igual que él y restante 50 % es mayor o igual que él.

Otros estadísticos para medir la posición de los datos son los **n-tiles**, los que representan los **n-1** valores que dividen la distribución de la variable en **n** partes, tales que todas ellas contiene el mismo porcentaje de observaciones.

Copyrighted material

Los **percentiles**, por ejemplo, dividen la distribución de la variable en 100 partes, tales que cada una contiene el 1% de las observaciones. Los **deciles**, dividen la distribución de la variable en 10 partes, tales que cada una contiene el 10% de las observaciones. Los **quartiles**, dividen la distribución de la variable en 4 partes, tales que cada una contiene el 25% de las observaciones. En particular, los **quartiles**, coinciden con los **percentiles 25, 50, y 75**, (Ferran, A. M., 1996).

En el caso de la variable "Como valora la accion para evitar la contaminacion del Medio Ambiente", en al menos el 75% de las personas encuestadas señalan que la acción es mala; en al menos el 50 % de las personas encuestadas señalan que la acción es regular; y en al menos el 25 % de las personas encuestadas señalan que la acción es buena., vease en la tabla de **statistics**, los valores de 5, 4, y 3 de los percentiles 75%, 50% y 25%, respectivamente.

2.3 Análisis Descriptivo de una Variable Cuantitativa en Escala de Intervalo o Razón.

Existen situaciones en que la técnica estadística que se utilizará, exige que las variables implicadas sigan una distribución Normal. En situaciones en que los valores de una variable **son numéricos**, pudiendo tomar cualquier valor en un intervalo determinado, se dice que la variable correspondiente es "***cuantitativa continua***". Si la variable unicamente pudiera tomar una cantidad finita de valores, se diría que es "***cuantitativa discreta***". Si además, tiene sentido hablar de la razón entre sus valores, se dirá que la variable está medida en "***escala de razón***". Si unicamente tuviera sentido hablar de la diferencia entre sus valores, careciendo de sentido numérico la razón entre ellos, se diría que "***la variable está medida en escala de intervalo***", (Ferran, A. M., 1996).

Para ilustrar el análisis de frecuencia de una variable cuantitativa continua o discreta, se utilizará la Base de Datos *"FARMERS22"*, donde se encuentra la variable "***edad***". El análisis de la normalidad de los datos, tiene un sentido de control de calidad de los datos, se utiliza la rutina de comandos: **Analyze/Descriptives Statistics/Frequencies/** en la ventana de diálogo **Variable(s)**, se incluye la variable a analizar, en este caso se incluye la variable *Edad*; se debe seleccionar **Display frequency tables**; luego en la opción **Charts**, se seleccionan ***Histograms, With Normal curve,*** *y dar* ***Continue***. En la opción **statistics**, se selecciona *Mean, Median, Mode, Skewness, Kurtosis y Quartiles, y dar* ***Continue***. El análisis solicitado al SPPS, se presenta en el cuadro siguiente.

Copyrighted material

Cuadro 2.3. Análisis de frecuencia de la variable edad.

Edad (en años)

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	22	1	.7	.7	.7
	24	2	1.5	1.5	2.2
	27	1	.7	.7	3.0
	30	2	1.5	1.5	4.4
	31	1	.7	.7	5.2
	32	3	2.2	2.2	7.4
	33	1	.7	.7	8.1
	34	3	2.2	2.2	10.4
	35	1	.7	.7	11.1
	36	4	3.0	3.0	14.1
	37	1	.7	.7	14.8
	38	4	3.0	3.0	17.8
	39	4	3.0	3.0	20.7
	40	8	5.9	5.9	26.7
	41	1	.7	.7	27.4
	42	7	5.2	5.2	32.6
	43	7	5.2	5.2	37.8
	44	3	2.2	2.2	40.0
	45	2	1.5	1.5	41.5
	46	3	2.2	2.2	43.7
	47	2	1.5	1.5	45.2
	48	2	1.5	1.5	46.7
	49	5	3.7	3.7	50.4
	50	6	4.4	4.4	54.8
	51	1	.7	.7	55.6
	52	11	8.1	8.1	63.7
	53	2	1.5	1.5	65.2
	54	3	2.2	2.2	67.4
	55	5	3.7	3.7	71.1
	56	3	2.2	2.2	73.3
	57	3	2.2	2.2	75.6
	58	4	3.0	3.0	78.5
	60	4	3.0	3.0	81.5
	61	2	1.5	1.5	83.0
	62	3	2.2	2.2	85.2
	63	2	1.5	1.5	86.7
	67	3	2.2	2.2	88.9
	68	2	1.5	1.5	90.4
	70	2	1.5	1.5	91.9
	72	1	.7	.7	92.6
	73	2	1.5	1.5	94.1
	74	1	.7	.7	94.8
	75	2	1.5	1.5	96.3
	76	1	.7	.7	97.0
	77	1	.7	.7	97.8
	82	1	.7	.7	98.5
	86	1	.7	.7	99.3
	94	1	.7	.7	100.0
	Total	135	100.0	100.0	

Copyrighted material

Cuadro 2.4. ***Análisis de normalidad de la variable, edad, mediante el uso de Frequencies.***

Statistics

N	Valid	135
	Missing	0
Mean		49.92
Median		49.00
Mode		52
Skewness		.585
Std. Error of Skewness		.209
Kurtosis		.431
Std. Error of Kurtosis		.414
Percentiles	25	40.00
	50	49.00
	75	57.00

En el cuadro 2.3., se muestra el análisis de frecuencia de la variable "edad". En el cuadro 2.4., la salida **statistics** proporciona los valores de la media, mediana y moda, así como los valores de **Kurtosis = 0.431, y Skewness (Asimetría) = 0.585**.

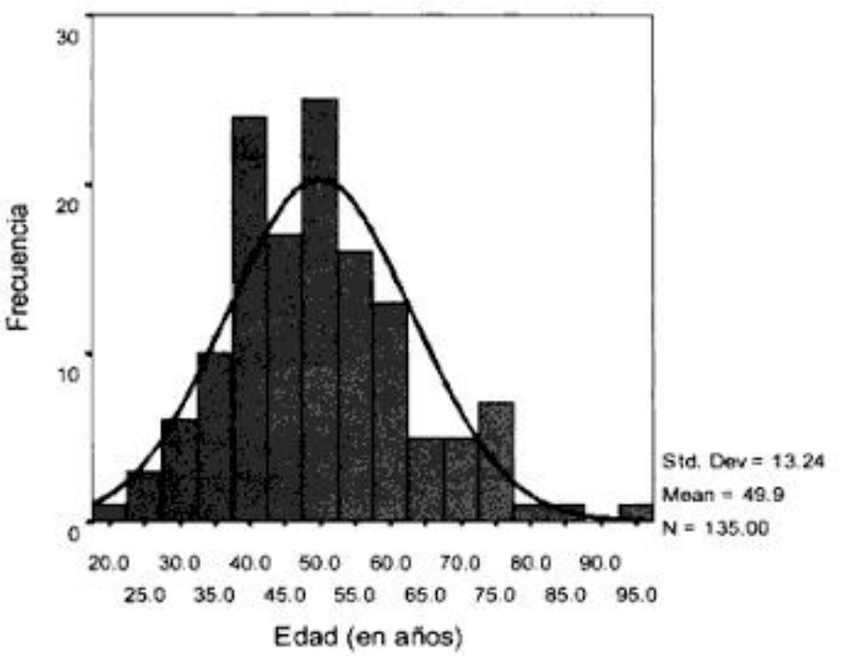


Figura 2.3. ***Ilustración de distribución Normal de la variable edad.***

Un histograma es un gráfico de la distribución de los valores de variables cuantitativas en la que los datos son agrupados en intervalos de la misma longitud, y cada uno de los posibles intervalos se representan mediante un rectángulo de área proporcional a la frecuencia de datos en el intervalo correspondiente. El punto medio de cada intervalo, denominado **marca de clase**, permitirá identificar el grupo valores en el intervalo dado, (Ferran, A. M., 1996).

El histograma con la curva normal, muestra la tendencia de normalidad de los datos. Una parte muy importante para el análisis de la normalidad de los datos es considerar los coeficientes de Kurtosis y Skewness (Asimetría), proporcionados por el comando statistics y presentados en el cuadro 2.4.

Copyrighted material

El coeficiente de Kurtosis, es una medida de la concentración de la distribución en torno a la media. Si la variable sigue una distribución Normal, su valor de Kurtosis será cero. Valores mayores a cero, indican que la distribución tiende a concentrase en torno a la media más que en una distribución Normal; mientras que valores menores que cero, indican que tiende a dispersarse más. **El coeficiente de Skewness (Asimetría),** como su propio nombre lo indica, es una medida de la asimetría de la distribución de los valores respecto a la media. Si la distribución de la variable es simétrica, su valor será igual a cero. Valores mayores que cero, indicarán que las desviaciones a la media son mayores para los valores superiores a la media que para los valores inferiores; mientras que valores menores que cero indicarán que las desviaciones a la media son mayores para los valores inferiores a la media que para los valores superiores. (Ferran, A. M., 1996).

En el análisis de la variable "edad", el valor de Kurtosis = 0.431, y de Skewness = 0.585, están en correspondencia con el histograma de frecuencias, que muestra la tendencia Normal de los datos. Esto se confirma con el analisis de Normalidad de los datos, por medio de **la prueba Kolmogorov-Smirnov. Para realizar la prueba de Kolmogorov-Smirnov (K-S), dentro del SPSS, se utiliza el módulo Analyse/ Nonparametric test/ Sample KS. Luego, en la ventana de diálogo que aparece, se declara la variable dependiente que se desea verificar la normalidad de los datos, (se marca en Test de Distribución la opción Normal).** El resultado es el siguiente.

Cuadro 2.5. ***Prueba de Kolmogorov-Smirnov para la variable edad.***

One-Sample Kolmogorov-Smirnov Test

		Edad (en años)
N		135
Normal Parameters[a,b]	Mean	49.92
	Std. Deviation	13.24
Most Extreme Differences	Absolute	.077
	Positive	.077
	Negative	-.037
Kolmogorov-Smirnov Z		.896
Asymp. Sig. (2-tailed)		.399

a. Test distribution is Normal.

b. Calculated from data.

El valor de Significancia obtenido de 0.399 > 0.05, implica que se acepta la hipótesis de normalidad para la variable "edad". En resumen la prueba de K-S reconoce como variable Normal las mediciones para variable "edad". Esto confirma la regla de que, en una distribución Normal, tanto el coeficiente de Kurtosis como el coeficiente de Skewness (Asimetría), deberían ser próximos a cero.

Otra rutina muy importante para analizar las características de una variable cuantitativa continua o discreta, es el uso del comando **"Descriptives"**, que facilita obtener directamente: Índices de Posición, Índices de Tendencia Central, Índices de Dispersión y Distribución. Realmente, la ruta del comando **"Descriptives"** del SPSS, representa una forma muy efectiva de obtener la misma información que proporciona el comando **"Frequencies"**, pero de otra manera.

Copyrighted material

Para ilustrar el uso del comando **"Descriptives"**, se carga la variable ***"edad"*** desde la BDD *"FARMERS22"*. La rutina de comandos es: **Analyze/Descriptives Statistics/ Descriptives** / en la ventana de diálogo **Variable(s)**, se incluye la variable *Edad*; luego en **Options**, se seleccionan *Mean, Sum, Minimum, Maximum, Std Deviation, Variance, Range, Skewness, Kurtosis, y dar* ***Continue***. Luego **OK**. El análisis descriptivo de los datos, se presenta en el cuadro 2.6.

Cuadro 2.6. Análisis descriptivo para la variable edad, mediante el comando "Descriptives".

Descriptive Statistics

		Edad (en años)	Valid N (listwise)
N	Statistic	135	135
Range	Statistic	72	
Minimum	Statistic	22	
Maximum	Statistic	94	
Sum	Statistic	6739	
Mean	Statistic	49.92	
Std.	Statistic	13.24	
Variance	Statistic	175.404	
Skewness	Statistic	.585	
	Std. Error	.209	
Kurtosis	Statistic	.431	
	Std. Error	.414	

Otra forma de obtener las estadísticas descriptivas que caracterizan una variable cuantitativa, es hacer uso del Comando **"Explore"**. La ruta del comando **"Explore"**, permite al usuario obtener de forma muy efectiva los índices siguientes: 1) Indices de posición son: Cuartiles q_1 25%, q_2 50% (mediana) y q_3 75% y centiles; 2) Indices de tendencia central: Media, Mediana, Moda, Suma de todos los valores; 3) Indices de dispersión: Desvío Estándar, Variancia, Rango, Error Estandar, Valor mayor y menor; 4) Índices de distribución: Coeficiente de Asimetría, y Kurtosis.

Para ilustrar el uso del comando **"Explore"**, se carga la variable ***"edad" en*** la Base de Datos *"FARMERS22"*. La rutina a seguir es: **Analyze/Descriptives Statistics/Explore** / en la ventana de diálogo **Dependent List,** se incluye la variable *Edad*; luego en **Statistics**, seleccionar **Descriptives**, *Outliers, y Percentiles, y dar* ***Continue***. En la opción **Plots**, seleccionar **Stem and Leaf**. Luego, dar **OK**. El análisis de los datos, se presenta en el cuadro siguiente.

Copyrighted material

Cuadro 2.7. Análisis descriptivo para la variable edad, mediante el comando "Explore".

Descriptives

			Statics	Std. Error
Edad (en años)	Mean		49.92	1.14
	95% Confidence Interval for Mean	Lower Bound	47.66	
		Upper Bound	52.17	
	5% Trimmed Mean		49.49	
	Median		49.00	
	Variance		175.404	
	Std. Deviation		13.24	
	Minimum		22	
	Maximum		94	
	Range		72	
	Interquartile Range		17.00	
	Skewness		.585	.209
	Kurtosis		.431	.414

En el Cuadro 2.7., el análisis descriptivo para la variable "edad", dado por el comando **"Explore",** el que permite obtener: 1) Media, mediana, desviación estándar, variancia, etc. ; 2) El 5% de Trimmed mean (media recortada) es la media aritmética eliminando el 5% de los datos con los valores más bajos y el 5% con los más altos; 3) El intervalo de confianza de la media al 95%; 4) Rango intercuartílico (puntuación del centil 75).

El análisis descriptivo dado por el comando **"Explore"**, presentado en el Cuadro 2.7., tiene un sentido de control de calidad de los datos y facilita conocer tanto la tendencia central como la dispersión de los valores de la variable "edad". En el ejemplo anterior, caso de una variable cualitativa en escala ordinal, se definió la mediana como una medida de tendencia central. Sin embargo, la medida de tendencia central más comúnmente utilizada para variables cuantitativas continuas, es la media aritmética de los valores observados, en este caso igual a 49.92.

Por otra parte, el rango como medida de la dispersión de los datos, presenta el inconveniente de que únicamente depende de los dos valores más extremos (mínimo y máximo), en este caso igual a 72. Una alternativa para medir la dispersión de los datos, que considera todos los valores observados, es la varianza, definida como el promedio de los cuadrados de las desviaciones de cada observación respecto a la media. Si los valores de los datos están muy concentrados, las desviaciones respecto a la media serán pequeñas y viceversa. En consecuencia, sus cuadrados también lo serán. En este caso la varianza y desviación estándar obtenidas son 175.40 y 13.24, respectivamente. Para saber que tan grande o pequeña es la desviación respecto a su media, lo que se hace es calcular el Coeficiente de Variación, definido como el cociente entre la desviación típica y la media, (Ferran, A. M., 1996).

Los valores extremos (**Outliers**), dados en la salida del comando **"Explore"**, presentan los cinco valores más altos (highest) y los cinco valores más bajos (lowest), lo que facilita encontrar aquellos "datos aberrantes" o "valores atípicos", en la distribución de la variable "edad".

Copyrighted material

Cuadro 2.8. ***Valores extremos (Outliers) para la variable edad.***

Extreme Values

			Case Number	Value
Edad (en años)	Highest	1	79	94
		2	65	86
		3	41	82
		4	118	77
		5	75	76
	Lowest	1	97	22
		2	3	24
		3	82	24
		4	7	27
		5	61	. [a]

a. Only a partial list of cases with the value 30 are shown in the table of lower extremes.

Por otra parte, los percentiles de la variable "edad" presentados en el Cuadro 2.9., son parte de la salida del comando **"Explore"**. Los percentiles 25 y 75, (igual a 40 y 57) denominados q_1 y q_3, ayudan a conocer la distribución de los datos, ya que contienen el 50 % de los datos más centrados, en el sentido de que el límite inferior q_1 deja por debajo de él al 25 % de los casos; y el límite superior q_3, deja por encima de él al 25% de los casos, (Ferran, A. M., 1996).

Cuadro 2.9. ***Percentiles para la variable edad, mediante el comando "Explore".***

Percentiles

Percentiles	Weighted Average(Definition 1)	Tukey's Hinges
	Edad (en años)	Edad (en años)
5	30.80	
10	34.00	
25	40.00	40.00
50	49.00	49.00
75	57.00	57.00
90	68.80	
95	75.00	

Copyrighted material

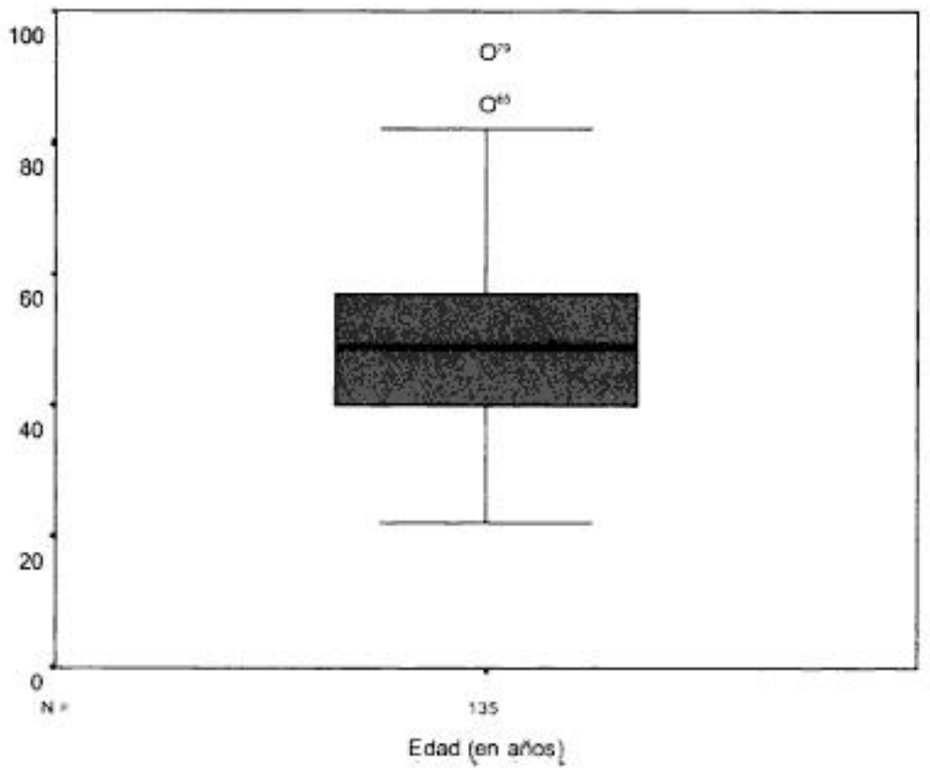


Figura 2.4. Ilustración del gráfico de Caja y Bigotes, (Box-Plot) para la variable edad.

La mediana es una medida de tendencia central basada en percentiles, funcionando como una medida alternativa a la media. **La mediana** se muestra en la gráfica como la línea horizontal más gruesa dentro de "la caja", **corresponde al segundo cuartil.** Por su parte, **el rango intercuartílico** definido como la diferencia entre el tercer y el primer cuartil, puede considerarse como una alternativa a la desviación típica para medir la dispersión de los datos. Los bigotes inferior y superior al mínimo y al máximo valor, tales que su distancia a los límites inferior y superior, respectivamente, de la caja es inferior a una vez el rango intercuartílico. En el caso de que un valor diste de los límites inferior o superior de la caja más de una vez el rango intercuartílico será considerado como un valor aislado o extremo, y se representará mediante los símbolos "**o**", si dista menos de una vez y medio, y "**x**" si dista más de una vez y media, (Ferran, A. M., 1996).

En la figura 2.4, se muestra el gráfico de Caja y Bigotes, (Box-Plot), que representa los **Cuartiles** en forma gráfica, la cual es parte de la salida del comando **"Explore"**, proporcionada de manera automática (by default). En esta gráfica, se observa "la caja" que representa **el rango intercuartílico** y contiene el 50 % de los datos. Los límites inferior y superior de "la caja" corresponden a los cuartiles primero y tercero respectivamente (q_1 y q_3); en consecuencia, la altura de "la caja" coincide con **el rango intercuartílico;** cada bigote representa un cuartil 25 y 75, que en el caso de la variable "edad" son igual a 40 y 57 años. Así mismo, **la mediana** ó segundo cuartil, en este caso coincide con la edad de 49 años, veáse el el gráfico de Caja y Bigotes.

Copyrighted material

En resumen, el gráfico de Caja y Bigotes, (Box-Plot), contiene el 25 % de los valores más pequeños entre el mínimo valor y el límite inferior de la caja; contiene el 25 % de los siguientes valores, entre el límite inferior y la barra dentro de la caja; contiene el 25 % de los siguientes valores, entre la barra y el límite superior de la caja; y contiene el 25 % de los valores restantes, por encima del límite superior de la caja, (Ferran, A. M., 1996).

Además del gráfico de Caja y Bigotes, (Box-Plot), otra forma gráfica alternativa para describir la distribución de una variable cuantitativa continua o discreta, es el gráfico de **"Tallo y Hoja"** (Stem and Leaf Plot). En la figura 2.5, se presenta la salida de este grafico solicitado al SPSS, para el caso de la variable "edad".

```
Edad (en años) Stem-and-Leaf Plot

Frequency     Stem &  Leaf

3.00          2 .  244
1.00          2 .  7
10.00         3 .  0012223444
14.00         3 .  56666788889999
26.00         4 .  00000000122222223333333444
14.00         4 .  55666778899999
23.00         5 .  00000012222222222233444
15.00         5 .  555556667778888
11.00         6 .  00001122233
5.00          6 .  77788
6.00          7 .  002334
4.00          7 .  5567
1.00          8 .  2
2.00 Extremes     (>=86)

Stem width:  10
Each leaf:        1 case(s)
```

Figura 2.5. Ilustración del gráfico de Tallo y Hoja, (Stem-and-Leaf Plot) para la variable edad.

El gráfico de Tallo y Hoja, igual que en el gráfico del histograma, proporciona información sobre la distribución de los datos presentándolos agrupados en intervalos de la misma longitud, cada uno de los cuales se presenta mediante una línea de dígitos, con longitud proporcional a la frecuencia de datos en el mismo, lo cual permite identificar los distintos valores en un mismo intervalo. En concreto, **cada línea de dígitos corresponde a la descomposición de los valores de la variable en dos partes: el Tallo y la Hoja**, (Ferran, A. M., 1996).

Copyrighted material

Por ejemplo, en la primera fila del gráfico de **Tallo y Hoja**, en el margen izquierdo, se indica que la frecuencia es igual a 3, coincidiendo con el número de hojas de la línea, y en la parte inferior del gráfico se indica que cada hoja corresponde a un caso (**"Each leaf: 1 case(s)"**); la amplitud del tallo es igual a 10 (**"Stem width: 10"**). En consecuencia, en la primera fila del gráfico de **Tallo y Hoja**, están representados tres casos y, teniendo en cuenta que cada valor se obtiene como un producto de AxT.H., donde **A** es la amplitud del tallo, **T** es el tallo y **H** la hoja, los tres valores correspondientes a la primera fila son: 22, 24, 24.

Otro ejemplo, tal como sigue: En la tercera fila, hay una frecuencia de 10 valores, estos son: 30, 31, 32, 33, y 34, con frecuencias de 2, 1, 3, 1, y 3 respectivamente.

Finalmente, se observa en la última línea del gráfico de **Tallo y Hoja**, el valor de 2 casos extremos mayores o igual que 86 años de edad. El criterio para considerar que un caso es extremo, es el mismo que en el gráfico de Caja y Bigotes.

Una de las ventajas del sistema de análisis estadístico con SPSS, es la gran flexibilidad que tiene para realizar una gama de análisis estadísticos. Una vez que ya se tiene la descripción de las variables que se desean analizar, se procede a aplicar el método de análisis pertinente según el caso. Esta va desde el más simple como una prueba de **"t"**, interactuando siempre con el conjunto variables disponibles en la base de datos, -cualitativas y/o cuantitativas- hasta realizar complejos análisis de varianza unifactoriales o multifactoriales, univariados o multivariados. Para ilustrar el uso de la prueba de **"t"** para muestras independientes, se carga la variable ***"edad"*** *en* la base de datos *"FARMERS22"*. La rutina a seguir es: **Analyze/Compare Means/Independent Samples – T Test** / en la ventana de diálogo **Test Variables (s)**, se incluye la variable ***Edad***; luego en **Grouping Variable(s)**, se debe incluir la variable que clasifica ambos grupos, en este caso incluir la variable ***Sexo***, luego en **Define Groups**, escribir en **Group 1: 1**; y en **Group 2: 2**; y dar ***Continue***. Luego, dar **OK**. El análisis de los datos, se presenta en el cuadro siguiente.

Group Statistics

	Sexo	N	Mean	Std. Deviation	Std. Error Mean
Edad (en años)	Varón	119	50.19	13.07	1.20
	Mujer	16	47.88	14.75	3.69

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	d	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Edad (en años)	Equal variances assumed	.521	.472	.656	133	.513	2.32	3.53	-4.67	9.31
	Equal variances not assumed			.598	18.309	.557	2.32	3.88	-5.82	10.45

Copyrighted material

Capítulo 3. El Módulo Operativo "Graphs" del SPSS.

3.1 El Sistema de Análisis Estadístico del SPSS.

El Sistema de Análisis Estadístico SPSS, es un sistema amplio y flexible de análisis estadístico y de gestión de base de datos en un entorno gráfico. En pocas palabras, SPSS en un software estadístico con grandes propiedades gráficas integradas dentro de un mismo sistema, que facilita tanto el análisis estadístico de los datos, como su ilustración gráfica. El SPSS, aunque se maneja mediante menús descriptivos y cuadros de diálogo, la comunicación con el sistema se realiza mediante instrucciones que se agrupan en módulos. **El módulo principal, llamado Base**, es indispensable para manejar cualquier otro módulo.

El módulo Base, permite manejar la programación en general, la definición y manejo de datos, manejo de archivos, etc., procedimientos estadisticos que van desde el análisis descriptivo, análisis gráfico, hasta realizar los **Modelos Paramétricos** de ANOVA, MANOVA, Regresión Lineal Simple, Regresión Múltiple, **Pruebas No Paramétricas**, etc., pasando por diversos métodos inferenciales y generación de gráficos de alta resolución, Ferran A., M. (1996).

3.2 El Análisis Gráfico con el Módulo Operativo Graphs.

Una de las propiedades más destacadas del SPSS, es su enorme capacidad y facilidad para generar gráficos de alta resolución. Dentro del SPSS, a su vez hay otros módulos operativos, tales como: **File, Edit, View, Data, Transform, Analyze, Graphs, Utilities, Windows y Help**. Dentro del módulo operativo **Graphs,** existen diversas opciones, que por su facilidad de aplicación, también conducen fácilmente a un mal uso o abuso de las propiedades gráficas del SPSS. Debe tenerse el extremo cuidado de solicitar al SPSS, el gráfico adecuado haciendo uso de las variables apropiadas. Hay que tener presente que cualquier variable que se desea analizar, éstas corresponden con uno de estos tres tipos: 1) Variable Cualitativa en Escala Nominal; 2) Variable Cualitativa en Escala Ordinal; 3) Variable Cuantitativa en Escala de Intervalo o Razón. Cada una de estas variables, tienen sus caracteristicas que las definen y diferencian de las demás; por tanto, según el tipo de variable y sus caracteristicas, así será definido el análisis gráfico que se le solicite realizar al SPSS.

3.3 El Comando "Bar" dentro del Módulo Operativo Graphs.

El comando **Bar**, tiene tres opciones de gráficos en forma de barras, entre las que se pueden seleccionar: 1) ***Simple, 2) Clustered (Agrupadas) y 3) Stacked (Apiladas)***. Cada una de estas tres opciones de barras, a su vez, se pueden combinar con tres alternativas más, tales son:

a) ***Summaries for groups of cases,*** *que es la opción para generar gráficos bivariados;*
b) ***Summaries of separate variables,*** *que es la opción para generar gráficos multivariados* y
c) ***Values of individual cases,*** *que es la opción para generar gráficos univariados.*

Para ilustrar el análisis gráfico ***Simple***, se utilizará las variables "edad" y "sexo", que se encuentran en la Base de Datos *"FARMERS22"*. Para desarrollar el gráfico ***Simple***, se utiliza la rutina de comandos siguiente: **Graphs/ Bar/ Simple/** ***Summaries for groups of cases / Define/*** en la ventana de diálogo,

Copyrighted material

primero se debe seleccionar ***Other Summary functions***; luego, en la ventana **Variable**, debe incluirse la variable a analizar, en este caso se incluye la variable *Edad*; después en la ventana ***Category Axis***, se debe incluir la variable que corresponderá al eje X, en este caso *Sexo*. Dar **Ok**. El gráfico solicitado al SPPS, es el siguiente.

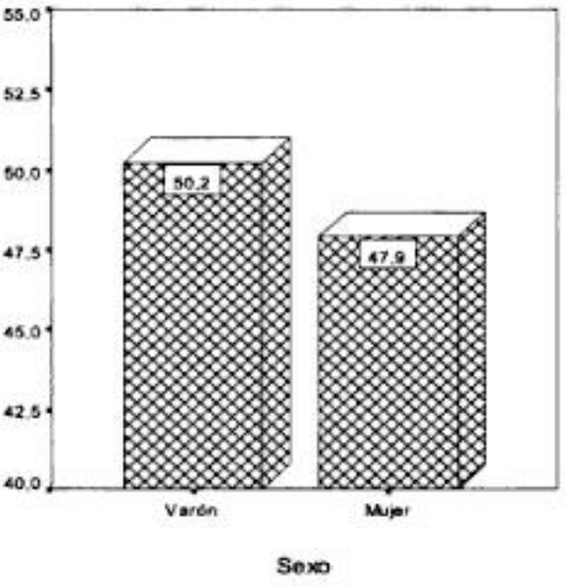


Figura 3.1. ***Gráfico Simple de las variables edad y sexo.***

Para ilustrar el análisis gráfico ***Clustered***, se utilizará las variables "edad" por "sexo", que se encuentran dentro de la Base de Datos *"FARMERS22"*. Para desarrollar el gráfico ***Clustered***, se utiliza la rutina de comandos siguiente: **Graphs/ Bar/ Clustered/** ***Summaries for groups of cases / Define*** / en la ventana de diálogo, primero se debe seleccionar ***Other Summary functions***; luego, en la ventana **Variable**, debe incluirse la variable a analizar, en este caso se incluye la variable *Edad*; después en la ventana ***Category Axis***, se debe incluir la variable que corresponderá al eje X, en este caso *Sexo*; después en la ventana ***Define Clusters by***, se debe incluir la variable que constituye el cluster, en este caso se incluye la variable ***Tipología de Productor(a)***. Dar **Ok**.

El gráfico solicitado al SPPS, es el siguiente.

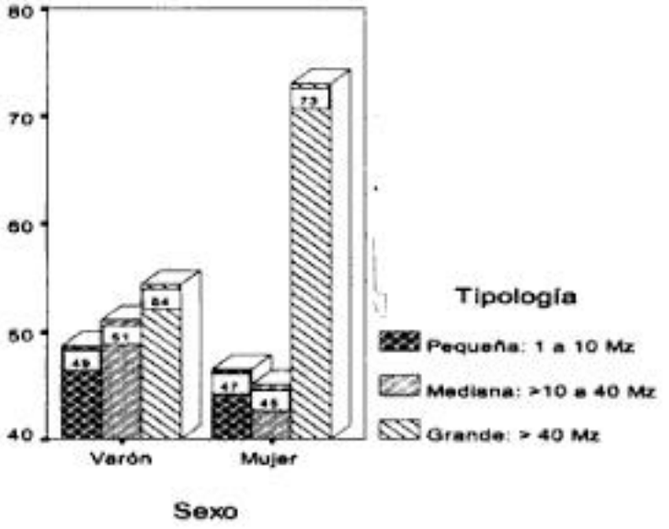


Figura 3.2. ***Gráfico Clustered - Bivariado de las variables edad, sexo y tipología de productor.***

Copyrighted material

Para ilustrar el análisis gráfico ***Stacked***, se utilizará las variables "edad" por "sexo", que se encuentran dentro de la Base de Datos *"FARMERS22"*. Para desarrollar el gráfico ***Stacked***, se utiliza la rutina de comandos siguiente: **Graphs/ Bar/ Stacked/** ***Summaries for groups of cases*** **/** ***Define*** /en la ventana de diálogo, primero se debe seleccionar ***Other Summary functions***; luego, en la ventana **Variable**, debe incluirse la variable a analizar, en este caso se incluye la variable *Edad*; después en la ventana, ***Category Axis***, se debe incluir la variable que corresponderá al eje X, en este caso *Sexo*; después en la ventana ***Define Stacks by***, se debe incluir la variable-criterio para apilar las barras, en este caso se incluye la variable ***Tipología de Productor(a)***. Dar **Ok**.

El gráfico solicitado al SPPS, es el siguiente.

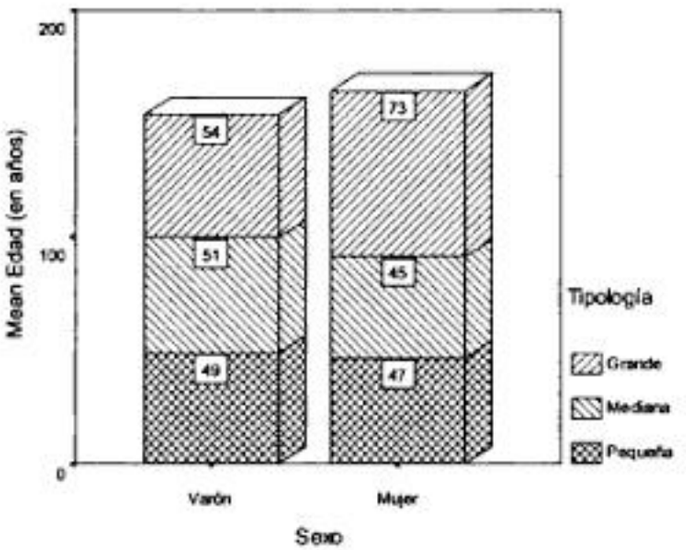


Figura 3.3. ***Gráfico Stacked - Bivariado de las variables edad, sexo y tipología de productor.***

3.4 El Comando "Bar" para Generar Gráficos Multivariados.

Una segunda ruta para generar gráficos en SPSS, es hacer uso de la opción ***"Summaries of separate variables"*****, que es la opción para generar gráficos multivariados**. Esta opción es muy útil para graficar **variables dicotómicas**, del tipo Si / No. Es una opción valiosa, sobre todo en aquellos casos que se tienen variables de selección múltiple, Si / No, tantas como sean necesarias. De este modo, el SPSS genera un gráfico multivariado, en el cual se muestra la información de muchas variables simultáneamente, en un mismo plano cartesiano.

De hecho, el SPSS computa todas las respuestas "Si", generando el gráfico multivariado en base al porcentaje de "respuestas afirmativas" de cada una de las variables, donde cada barra generada en el gráfico corresponde a una variable.

Para ilustrar el uso de la opción ***"Summaries of separate variables"***, se utilizarán las variables codificadas **sp1** hasta **sp6**, las que se describen en el cuadro 3.1, las mismas que se encuentran dentro de la BDD *"FARMERS22"*.

Copyrighted material

Cuadro 3.1. ***Actividad agropecuaria-forestal a la cual se dedican los productores es?:***

Marque con una X su *opción u opciones* seleccionada (s)	**Si**	**No**	**Escriba el Área en Mz**
(sp1). Caficultura			**(sp7).**
(sp2). Agricultura en general			**(sp8).**
(sp3). Ganadería en general			**(sp9).**
(sp4). Agricultura de Patio			**(sp10).**
(sp5). Explotación del Bosque			**(sp11).**
(sp6). Tacotales			**(sp12).**

Para desarrollar el gráfico multivariado, se utiliza la rutina de comandos siguiente: **Graphs/ Bar/ Simple/** ***Summaries of separate variables/ Define*** /en la ventana de diálogo **Bars Represent**, deben incluirse las variables a analizar, en este caso se incluyen las variables **sp1 hasta sp6**; luego las variables incluidas dentro de la ventana, **se marcan** para usar la opción **Change Summary**, después se selecciona la opción **percentage inside y marcar 1 en Low y 1 en High**. ***Dar Continue*** y **Ok**. El gráfico solicitado al SPPS, es el siguiente.

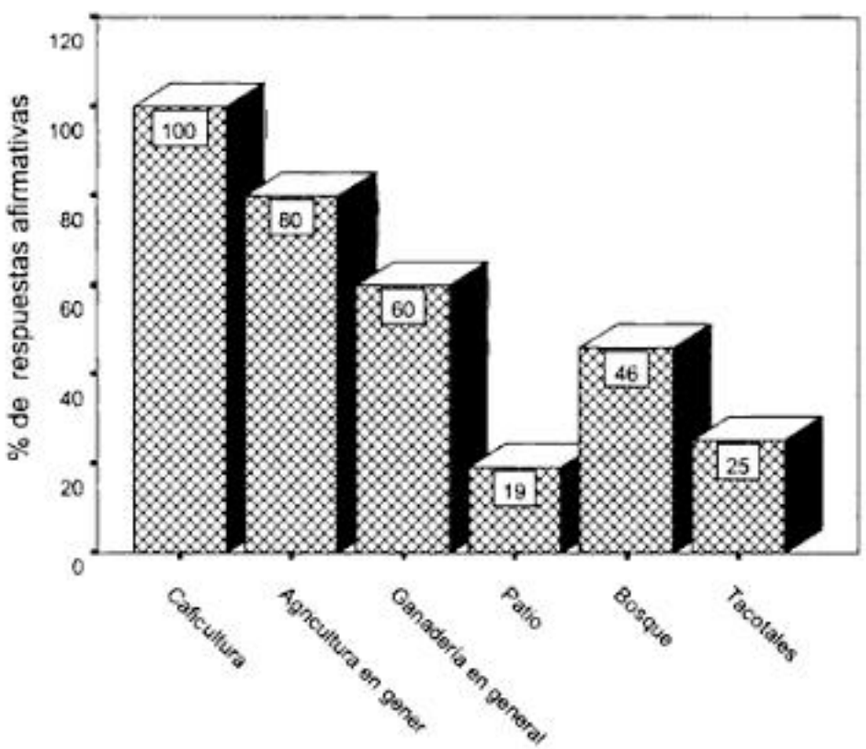


Figura 3.4. ***Gráfico Multivariado de las variables dicotómicas desde sp1 hasta sp6.***

Por otra parte, haciendo uso de la opción ***"Summaries of separate variables"***, también pueden generarse gráficos multivariados a partir de **variables cuantitativas continuas o discretas**. En este caso, las variables de selección múltiple (continuas o discretas), se marcan en función del estadístico que se desea, por ejemplo la media, la moda, la mediana, la varianza, la desviación estándar, etc. De este modo, el SPSS genera un gráfico multivariado, que muestra la información de muchas variables simultáneamente, pero en base al estadístico seleccionado.

Copyrighted material

Para ilustrar el uso de la opción ***"Summaries of separate variables"***, pero a partir de variables cuantitativas continuas o discretas, se utilizarán las variables **sp7** hasta **sp12**, que se describen en el cuadro 3.1, las mismas que se encuentran dentro de la BDD *"FARMERS22"*.

Para desarrollar el gráfico multivariado, se utiliza los comandos siguientes: **Graphs/ Bar/ Simple/ *Summaries of separate variables/ Define*/**en la ventana de diálogo **Bars Represent**, deben incluirse las variables a analizar, en este caso se incluyen las variables desde **sp7 hasta sp12**; luego las variables incluidas dentro de la ventana, **se marcan** para usar la opción **Change Summary**, después se selecciona la opción **Mean of values**. ***Dar Continue.*** Luego, ir a **Options** y seleccionar **exclude cases variable by variable**. Dar ***Continue. Luego darle*** **Ok**. El gráfico solicitado es el siguiente.

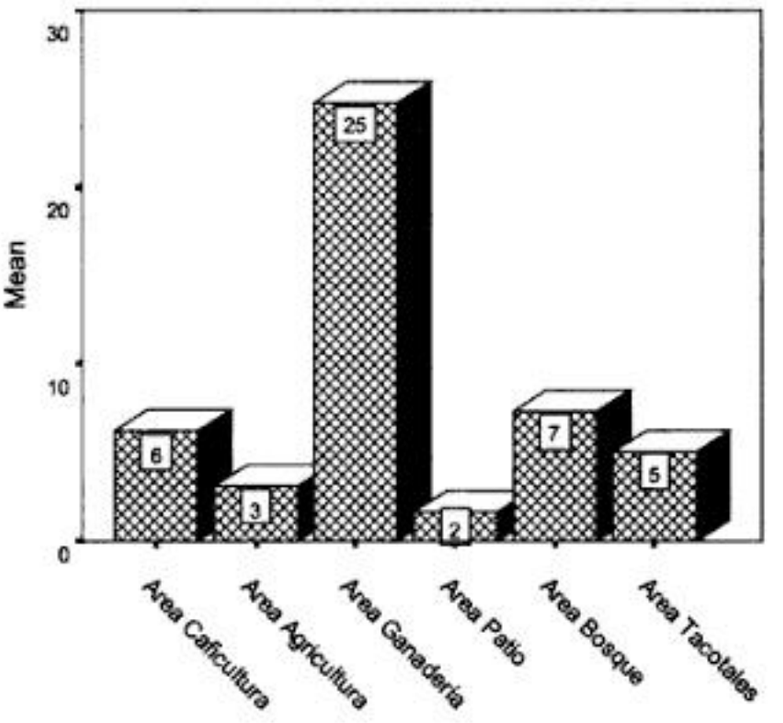


Figura 3.5. ***Gráfico Multivariado de las variables cuantitativas sp7 hasta sp12, (área en Mz).***

Una forma sencilla de verificar la información ofrecida en el gráfico multivariado para variables cuantitativas continuas o discretas, es solicitarle (por separado) al SPSS que presente las estadísticas de las variables incluidas en el gráfico multivariado; tal como se presenta en el cuadro siguiente.

Cuadro 3.2. Estadísticas de las variables cuantitativas continuas incluidas en el gráfico multivariado.

Statistics

		Area Caficultura	Area Agricultura	Area Ganaderia	Area Patio	Area Bosque	Area Tacotales
N	Valid	135	107	80	26	62	34
	Missing	0	28	55	109	73	101
Mean		6.2019	3.0958	24.7250	1.6154	7.3065	5.0735

Copyrighted material

Otra ruta muy valiosa para extraer la mayor riqueza posible de los datos disponibles, es solicitar al SPSS el gráfico multivariado, pero con un criterio de clasificación ex antes la información. Tal clasificación ex antes, se logra reorganizando la base de datos, dentro del modulo operativo **Data,** usando el comando **Split file**. Por ejemplo, para obtener el gráfico multivariado de las variables **sp7 hasta sp12**, pero por cada municipio involucrado en el estudio, se le puede solicitar al SPSS, que organice la salida por municipio. La rutina de comandos a usar es: **Data/ Split file/ Organize output by groups/** en la ventana de diálogo se debe incluir la variable de clasificación, en este caso **se incluye la variable Nombre del Municipio/ Dar OK.** Siguiendo esta ruta, la base de datos se reorganiza por municipio. Posteriormente, al realizar la rutina para solicitar el gráfico multivariado de las variables **sp7 hasta sp12**, la salida será ejecutada por municipio. La salida solicitada del gráfico multivariado por municipios, (por lo extensa que es), se presenta aquí solo para dos del total de municipios involucrados, tal como sigue.

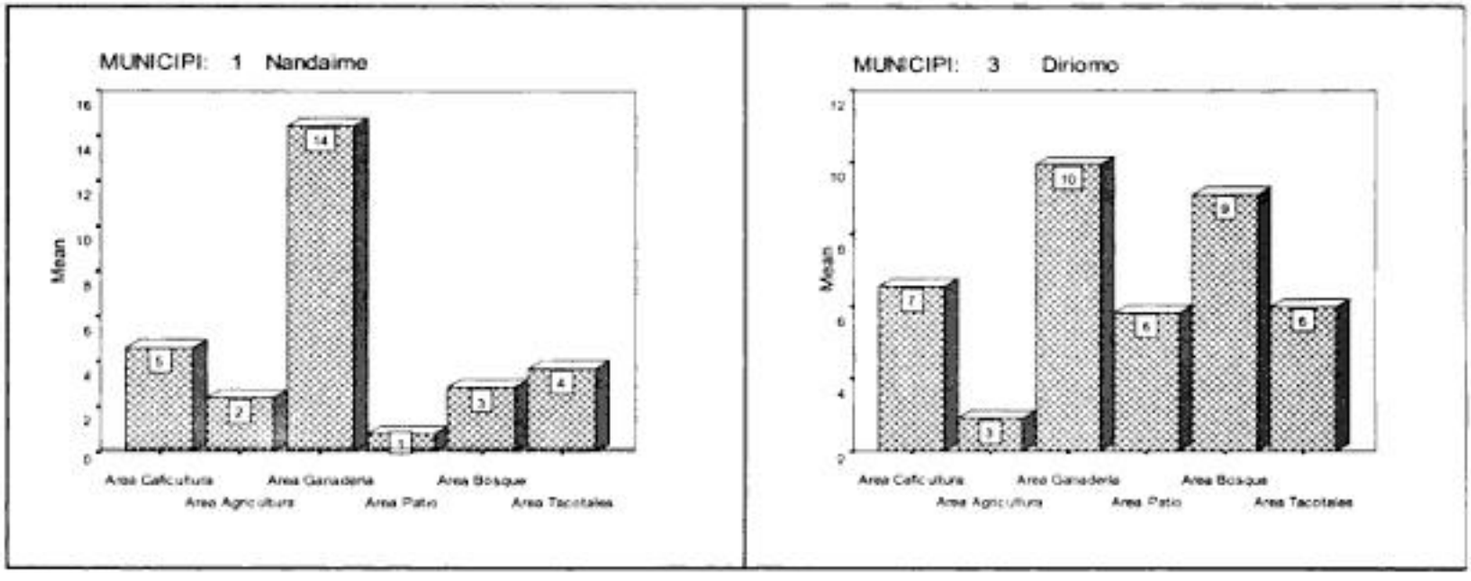


Figura 3.6. ***Gráfico Multivariado con un criterio de clasificación ex antes, por municipio, para las variables cuantitativas sp7 hasta sp12, (área en Mz).***

Después de usar el comando **Split file.**, se debe tener el sumo cuidado de regresar la base de datos a su estado original, es decir, dejar la base de datos **sin** la clasificación ex antes.

3.5 El Comando "Line" para Generar Gráficos.

Al igual que el comando **Bar**, el comando **Line** tiene tres opciones de gráficos en forma de líneas, entre las que se pueden seleccionar: 1) ***Simple, 2) Multiple y 3) Drop-Line (Línea de Gota).*** Cada una de estas tres opciones de líneas, a su vez, se pueden combinar con tres alternativas más, tales son: a) ***Summaries for groups of cases,*** b) ***Summaries of separate variables,*** y c) ***Values of individual cases***. En general, el gráfico de Líneas, se recomienda utilizar con variables cuantitativas continuas, las que se usan como variables independientes, siendo éstas unidas por medio de líneas.

Para ilustrar el análisis gráfico ***Line***, se utilizará la variable "edad", que se encuentran dentro de la Base de Datos *"FARMERS22"*. Se utiliza la rutina de comandos siguiente: **Graphs/ Line/ Simple/** ***Summaries for groups of cases/*** **Define/** en la ventana de diálogo, se debe seleccionar ***N of Cases***; luego, en la ventana ***Category Axis***, se debe incluir la variable que corresponderá al eje X, en este caso *Edad*; después dar **Ok**.

Copyrighted material

El gráfico solicitado al SPPS, es el siguiente.

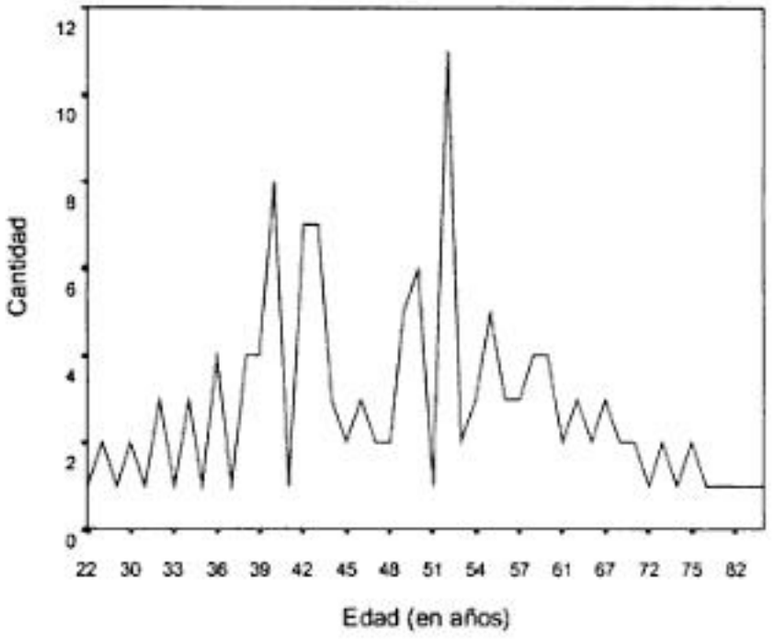


Figura 3.7. ***Gráfico de Línea con la opción Simple, para la variable edad.***

Otra alternativa, es la opción **Multiple.** Para ilustrar el análisis gráfico con la opción **Multiple**, se utilizarán las variables "edad" y "sexo", que se encuentran dentro de la Base de Datos *"FARMERS22"*. La rutina de comandos es: **Graphs/ Line/ Multiple/** ***Summaries for groups of cases/*** **Define/** en la ventana de diálogo, se debe seleccionar ***N of Cases***; luego, en la ventana ***Category Axis***, se debe incluir la variable que corresponderá al eje X, en este caso la variable *Edad*; luego en la ventana de diálogo ***Define Lines by,*** se debe incluir la variable que definirá las líneas en el gráfico, en este caso la variable *Sexo*; después dar **Ok**. El gráfico solicitado, es el siguiente.

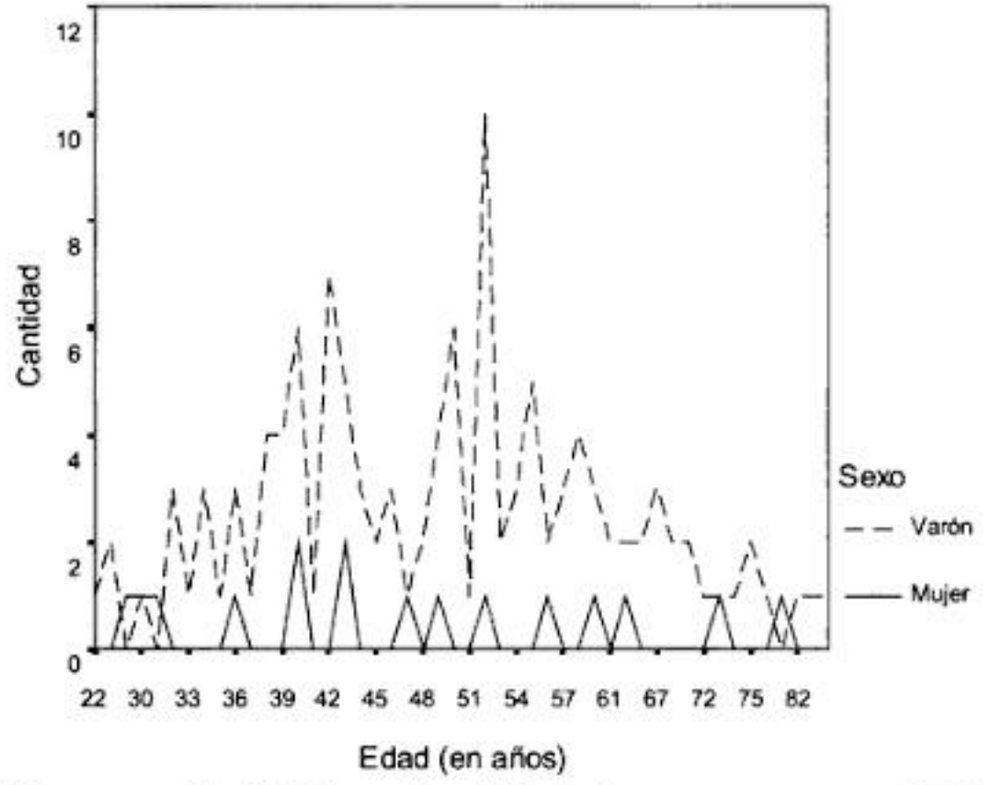


Figura 3.8. ***Gráfico de Linea, con opción Multiple para la variable edad y sexo.***

Copyrighted material

La siguiente alternativa, es la opción **Drop-Line (Línea de Gota).** Para ilustrar el análisis gráfico con ***Drop-Line***, se utilizarán las variables "edad" y "sexo", que se encuentran en la Base de Datos *"FARMERS22"*. La rutina de comandos es la siguiente: **Graphs/ Line/ Drop-Line/** ***Summaries for groups of cases/*** **Define/** en la ventana de diálogo, se debe seleccionar ***N of Cases***; luego, en la ventana ***Category Axis***, se debe incluir la variable que corresponderá al eje X, en este caso la variable *Edad*; luego en la ventana de diálogo ***Define Points by,*** se debe incluir la variable que definirá los puntos del gráfico, en este caso la variable *Sexo*; después dar **Ok**. El gráfico solicitado, es el siguiente.

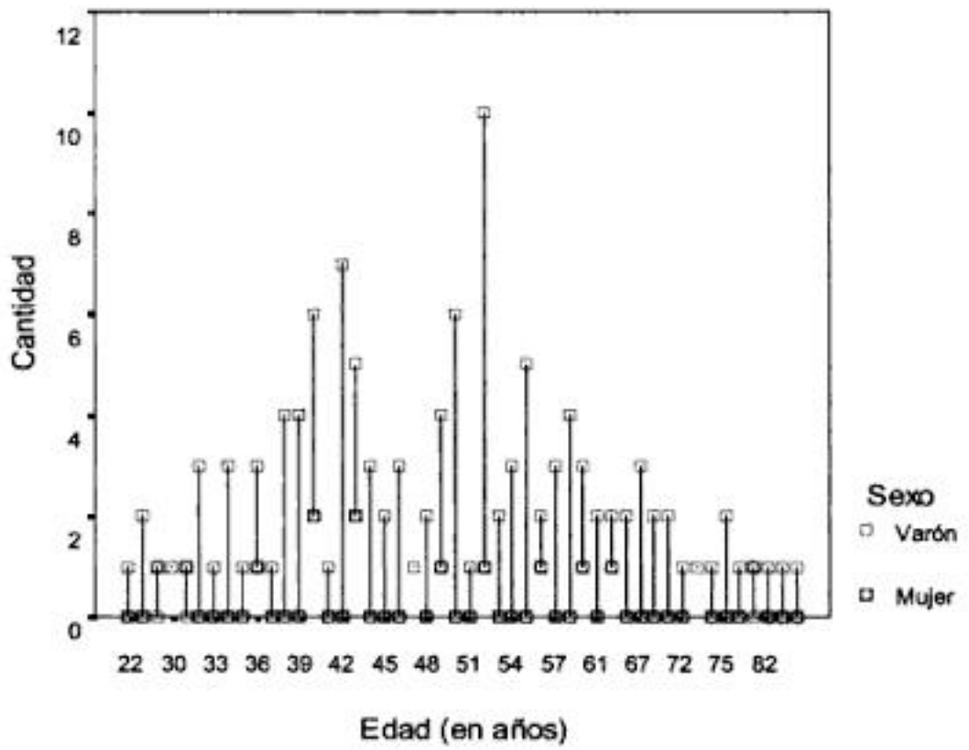


Figura 3.9. ***Gráfico de Gotas, con opción Drop-Line para la variable edad y sexo.***

3.6 El Comando "Pie" para Generar Gráficos.

Uno de los gráficos más atractivos generados por el SPSS, es el gráfico Pie (de Pastel). Se recomienda utilizar este tipo de gráfico en situaciones que los valores representan porcentajes, siendo que el pastel en su conjunto engloba el 100 % de las observaciones, igual a 1 en total. Al igual que el comando **Bar,** el comando **Pie** tiene tres opciones de gráficos en forma de pastel, tales son: a) ***Summaries for groups of cases,*** b) ***Summaries of separate variables,*** y c) ***Values of individual cases***. En general, para hacer el gráfico de Pastel, se recomienda utilizar la opción: ***Summaries for groups of cases,*** con variables cuantitativas discretas.

Para ilustrar el gráfico ***Pie (de Pastel)***, se utilizará la variable "Tipología del Productor (a)", que se encuentra en la Base de Datos *"FARMERS22"*. La rutina de comandos es la siguiente: **Graphs/ Pie/** ***Summaries for groups of cases/*** **Define/** en la ventana de diálogo, seleccionar ***N of Cases***; luego, en la ventana ***Define Slices by,*** se debe incluir la variable que definirá los pedazos del pastel, en este caso la variable "Tipología del Productor (a)"; después dar **Ok**.

Copyrighted material

El gráfico solicitado, es el siguiente.

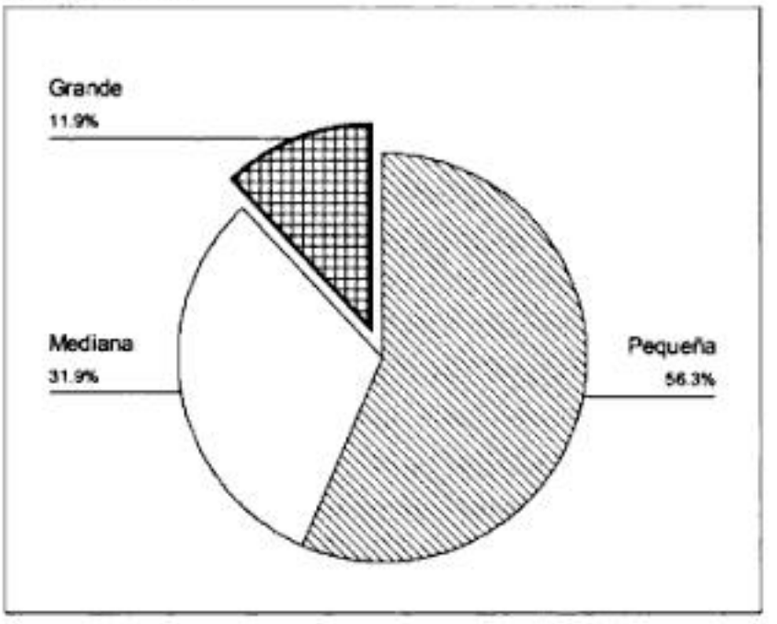


Figura 3.10. ***Gráfico de Pastel, para la variable tipología de productor (a).***

Dentro del gráfico de pastel, éste puede editarse, por ejemplo: abrir los slices, editar el texto, editar el formato, etc.

Otro ejemplo que parece muy aleccionador, es el gráfico de pastel del nivel de escolaridad de los productores. Se utilizará la variable "Escolaridad del Productor (a)", que se encuentra en la Base de Datos *"FARMERS22"*. Se destaca el hecho de que, el comando "Pie" permite realizar diversa opciones, tales como: editar el gráfico, variar el formato, abrir los slices, colocar las etiquetas en diferentes posiciones, etc.

La rutina de comandos es la siguiente: **Graphs/ Pie/** ***Summaries for groups of cases/*** **Define/** en la ventana de diálogo, seleccionar ***N of Cases***; luego, en la ventana ***Define Slices by,*** se debe incluir la variable que definirá los pedazos del pastel, en este caso la variable "Escolaridad del Productor (a)"; después dar **Ok**. El gráfico solicitado, es el siguiente.

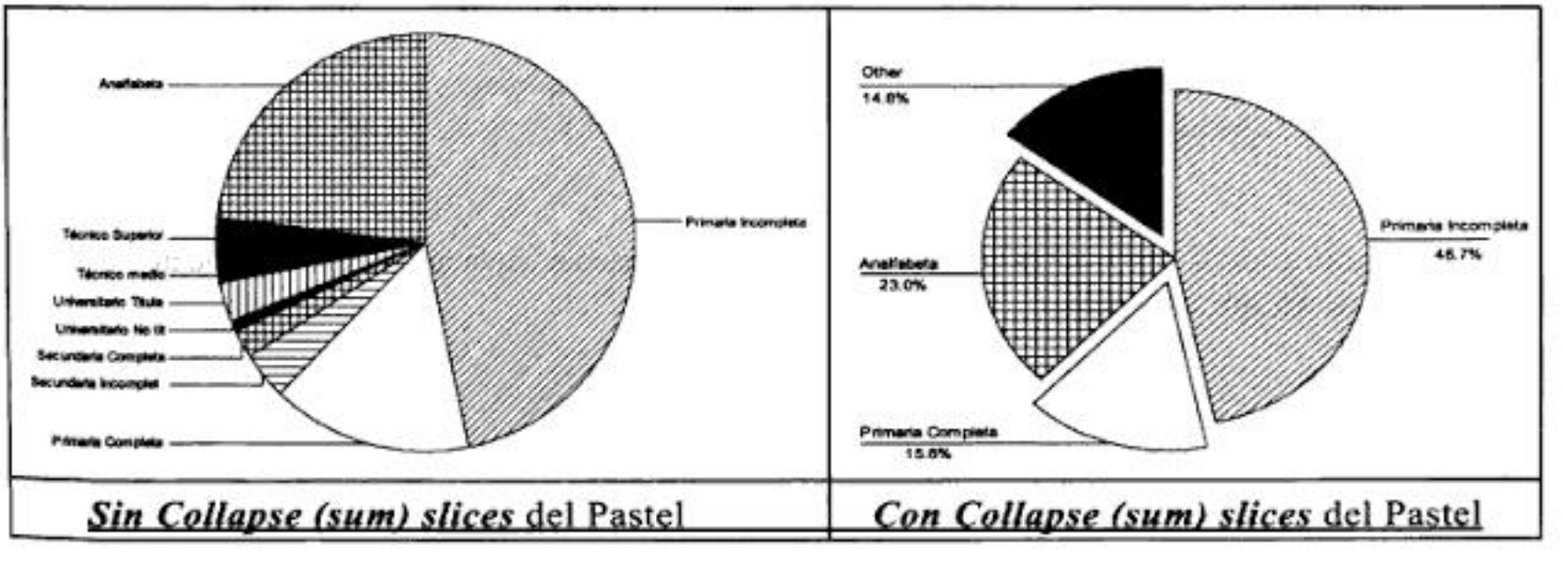


Figura 3.11. ***Gráfico de Pastel, para la variable Escolaridad del Productor (a).***

Copyrighted material

Dentro del archivo de salida del SPSS, al dar doble clic sobre una de las etiquetas del pastel, se puede editar el pastel, permitiendo ocultar algunos pedazos del pastel, usando la opción **Collapse (sum) slices less than: (5%)**, de modo que se puedan agrupar pedazos pequeños del pastel para dar una mejor presentación de la información, editar etiquetas, etc.

También otro ejemplo interesante, es el gráfico de pastel pero usando varias variables a la vez. Para este ejemplo, se utilizarán las variables: "Número total de mujeres", y "Número total de hombres", que se encuentra en la Base de Datos *"FARMERS22"*.

La rutina de comandos es la siguiente: **Graphs/ Pie/** ***Summaries of separate variables/*** **Define/** en la ventana de diálogo, ***Slices Represent,*** se deben incluir las variables que definirán los pedazos del pastel, en este caso son las dos variables "Número total de mujeres", "Número total de hombres"; después **se marcan esas dos variables**, y entrar en la tabla de diálogo **Change Summary**, ahí se debe seleccionar ***Sum of Values***; luego dar ***Continue***. Finalmente darle **Ok**. El gráfico solicitado, es el siguiente.

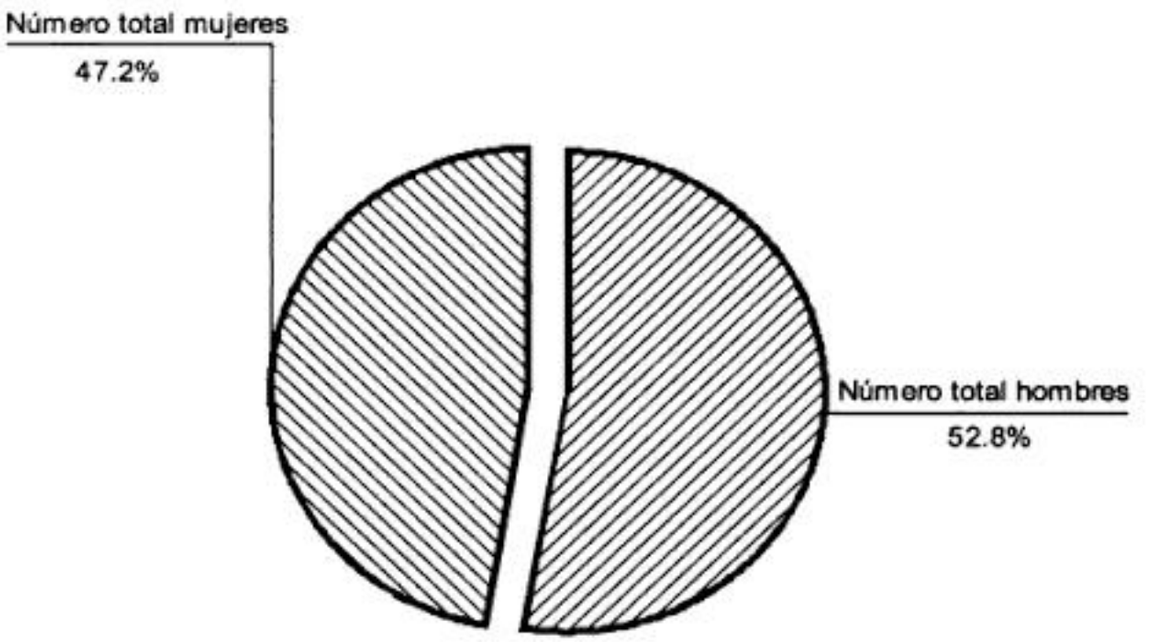


Figura 3.12. Gráfico de Pastel, para las variables número total de mujeres y número total de hombres.

Otras tipos de gráficos, que el SPSS también facilita al usuario, pueden explorarse dentro del módulo operativo de **Graph**, tales como:

Copyrighted material

Area: Es un gráfico parecido al gráfico de línea, pero dentro de él, el área definida por la línea se torna sombreada.
Boxplot: Esta opción genera el gráfico de Cajas y Bigotes al igual que el comando Explorer.
Error Bar: Es un gráfico especialmente útil cuando el gráfico deseado cuenta con una variable discreta que genera categorías o grupos, la cual se cruza con una variable continua. El gráfico de ***Error Bar,*** genera los promedios de cada categoría junto con sus intervalos de confianza al 95 %, los cuales se muestran en forma de dos bigotes, superior e inferior, con el promedio en el centro. En los capítulos siguientes cuatro y cinco, se ilustran los ejemplos del gráfico **Error Bar**.
Scatter: Es un gráfico que representa la "dispersión" de los valores observados de la variable analizada, presentando tales valores como una nube de puntos. Es importante para el ANARE.
Histogram: Esta opción genera el gráfico de histograma al igual que el comando **frequencies**.

3.7 Breves Sugerencia para Usar Mejor el Potencial del Sistema SPSS.

Existen muchas otras opciones y/o atributos que el sistema de análisis estadístico del SPSS le ofrece al usuario interesado en profundizar sus propiedades, esto es en los **Módulos File, Edit, Data, Transform, Analyze**, etc.; opciones sobre:

- recodificación de variables,
- sorteo de variables,
- cálculos de nuevos variables a partir de las variables ya existentes,
- fusión de bases de datos,
- división de bases de datos,
- modificaciones parciales o totales de variables dentro de una base de datos, etc, etc.

Todas estas capacidades del SPSS, pueden explorarse navegando dentro de los módulos antes citados. También existen muchas propiedades en la edición de gráficos y tablas, que el usuario debe practicar por su cuenta hasta adquirir las habilidades necesarias. De ahí que, se sugiere a los interesados explorar con paciencia todo el potencial que puede brindar el SPSS.

Capítulo 4. Tablas de Contingencia y Medidas de Asociación.

4.1 La prueba de Ji Cuadrado de Pearson en Tablas de Contingencia.

Una parte importante del análisis de datos provenientes de variables dicotómicas, variables en escala nominal, ordinal, o en escala de intervalo o de razón, se realizan con el SPSS por medio de tablas de contingencia, que facilitan la obtención de diversos estadísticos apropiados para realizar el análisis descriptivo e inferencial de la información social.

La prueba de *Ji* cuadrado de Pearson, se aplica en aquellos casos en que se disponga de una tabla de contingencia con ***"r"*** filas y ***"c"*** columnas correspondientes a la observación de muestras dos variables de ***X e Y***, con ***r*** y ***c*** categorías respectivamente. Se utiliza para contrastar la hipótesis nula:

Ho: Las variables X e Y son independientes.

Si el ***p***-valor asociado al estadístico de contraste es menor que α, se rechaza la Ho al nivel de significancia establecido, usualmente $\alpha = 0.05$, (Ferran, A. M., 1996).

Para ilustrar el uso de las tablas de contingencias y la prueba de *Ji* cuadrado de Pearson, se carga la BDD ***"SURVEY11"***. Para desarrollar las tablas de contingencia, la rutina de comandos a seguir es: **Analyze/ Descriptives Statistics/Croostabs/** en la ventana de diálogo **Row(s)**, debe incluirse la variable que se desea aparezca en la hilera o fila de la tabla, -es la variable independiente (X)-, en este caso se incluye la variable *Nombre del municipio*; y en la ventana de diálogo **Column(s)**, debe incluirse la variable que se desea aparezca en la columna de la tabla, -la variable de la columna es la variable dependiente (Y)-, en este caso se incluye la variable *Procedencia*. Luego, se selecciona la opción **Display clustered Bar charts**, para generar el gráfico bivariado correspondiente a estas variables. En la ventana **statistics**, seleccionar **Chi Square;** después en la ventana **Cells**, se selecciona **Observed, Expected**, a fin de obtener los valores observados y esperados; se selecciona **Row, Column y Total**, para obtener los porcentajes de la tabla por hilera, columna y total. Finalmente dar **OK**.

La prueba de *Ji* Cuadrado de Pearson en Tablas de Contingencia, se presenta a continuación.

Copyrighted material

Cuadro 4.1. Salida del SPSS para la prueba de Ji Cuadrado en Tablas de Contingencia.

Nombre del Municipio * Procedencia Crosstabulation

			Procedencia			
			Casco Urbano	Periferia Urbana	Area Rural	Total
Nombre del Municipio	Pueblo Nuevo	Count	60	46	63	169
		Expected Count	84.2	28.6	56.2	169.0
		% within Nombre del Municipio	35.5%	27.2%	37.3%	100.0%
		% within Procedencia	11.0%	24.7%	17.3%	15.4%
		% of Total	5.5%	4.2%	5.7%	15.4%
	Condega	Count	40	33	93	166
		Expected Count	82.7	28.1	55.2	166.0
		% within Nombre del Municipio	24.1%	19.9%	56.0%	100.0%
		% within Procedencia	7.3%	17.7%	25.5%	15.1%
		% of Total	3.6%	3.0%	8.5%	15.1%
	Municipio Jinotega	Count	84	37	32	153
		Expected Count	76.2	25.9	50.9	153.0
		% within Nombre del Municipio	54.9%	24.2%	20.9%	100.0%
		% within Procedencia	15.4%	19.9%	8.8%	13.9%
		% of Total	7.7%	3.4%	2.9%	13.9%
	El Sauce	Count	83	39	43	165
		Expected Count	82.2	28.0	54.8	165.0
		% within Nombre del Municipio	50.3%	23.6%	26.1%	100.0%
		% within Procedencia	15.2%	21.0%	11.8%	15.0%
		% of Total	7.6%	3.6%	3.9%	15.0%
	Municipio Matagalpa	Count	140	2	13	155
		Expected Count	77.2	26.3	51.5	155.0
		% within Nombre del Municipio	90.3%	1.3%	8.4%	100.0%
		% within Procedencia	25.6%	1.1%	3.6%	14.1%
		% of Total	12.8%	.2%	1.2%	14.1%
	Altagracia	Count	49	4	81	134
		Expected Count	66.8	22.7	44.5	134.0
		% within Nombre del Municipio	36.6%	3.0%	60.4%	100.0%
		% within Procedencia	9.0%	2.2%	22.2%	12.2%
		% of Total	4.5%	.4%	7.4%	12.2%
	Moyogalpa	Count	91	25	40	156
		Expected Count	77.7	26.4	51.9	156.0
		% within Nombre del Municipio	58.3%	16.0%	25.6%	100.0%
		% within Procedencia	16.6%	13.4%	11.0%	14.2%
		% of Total	8.3%	2.3%	3.6%	14.2%
Total		Count	547	186	365	1098
		Expected Count	547.0	186.0	365.0	1098.0
		% within Nombre del Municipio	49.8%	16.9%	33.2%	100.0%
		% within Procedencia	100.0%	100.0%	100.0%	100.0%
		% of Total	49.8%	16.9%	33.2%	100.0%

Copyrighted material

La hipótesis que se desea contrastar es que las variables Municipios por Procedencia son independientes. En el ejemplo dado en el cuadro 1, en el municipio Condega, para el casco urbano, el número observado de resultados favorables es igual a 40. Bajo la hipótesis de independencia, el número esperado (Expected count) es igual al producto de los valores marginales en la fila y en la columna correspondientes (Row total=166; y Column total= 547), todo ello partido por el total de observaciones, en este caso 1098 -> (166*547)/ 1098 = 82.7. Siguiendo ese procedimiento, se puede calcular c/u de los valores esperados, para realizar la prueba de Chi cuadrado.

El estadístico Chi cuadrado, se construye a partir de las diferencias entre las frecuencias observadas y esperadas bajo la hipótesis de independencia. Dado que se obtuvo un valor de significancia menor de 0.05, se rechaza la Ho de independencia entre las variables Municipio y Procedencia. Es decir, "la procedencia depende del municipio en cuestión", por ejemplo: En Pueblo Nuevo y Condega, predomina la procedencia Rural; en cambio, en Jinotega, Matagalpa, El Sauce, Altagracia y Moyogalpa, predomina la procedencia Urbana.

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	243.874 [a]	12	.000
Likelihood Ratio	264.983	12	.000
Linear-by-Linear Association	20.143	1	.000
N of Valid Cases	1098		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 22.70.

El gráfico bivariado solicitado, se presenta a continuación

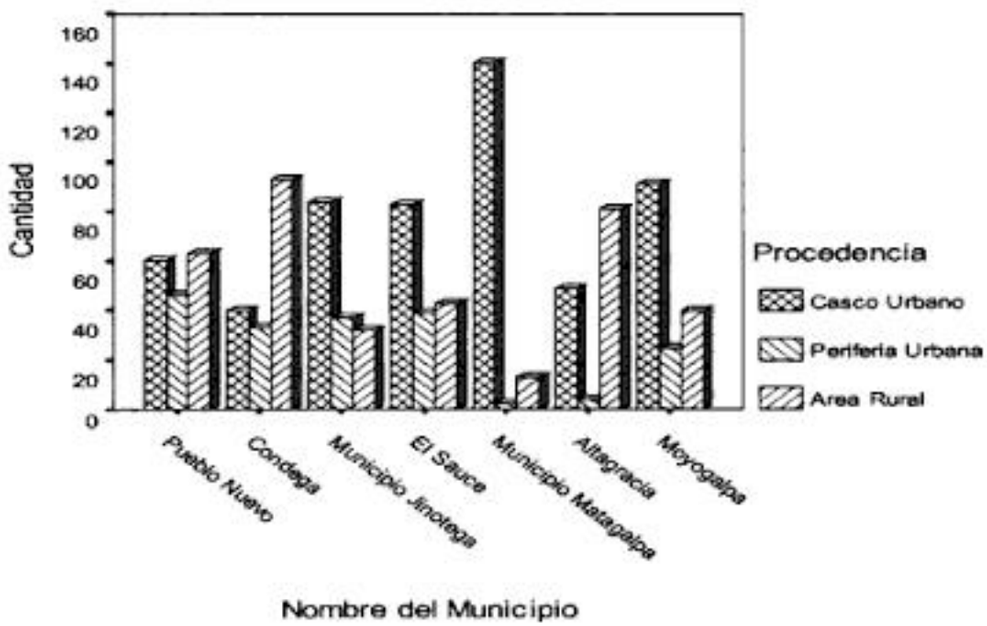


Figura 4.1. ***Relación bivariada de municipios por procedencia.***

Copyrighted material

4.2 Medidas de Asociación para dos Variables Dicotómicas en Tablas de Contingencia.

El coeficiente ***Phi***, es una medida del grado de asociación entre dos variables dicotómicas, basada en el estadístico *Ji-* Cuadrado de Pearson, que toma valores entre 0 y 1. Valores próximos a 0, indicarán **no** asociación entre las variables y valores próximos a 1, indicarán una fuerte asociación, (Ferran, A. M., 1996).

Para ilustrar el uso de la prueba de ***Phi***, se carga la BDD ***"SURVEY11"***. La rutina de comandos a seguir es: **Analyze/Descriptives Statistics/Croostabs/** en la ventana de diálogo **Row(s)**, debe incluirse la variable que se desea aparezca en la hilera o fila de la tabla, en este caso se incluye la variable *Sexo*; y en la ventana de diálogo **Column(s)**, debe incluirse la variable que se desea aparezca en la columna de la tabla, en este caso se incluye la variable *Visita ud. la Alcaldia*. Luego, se selecciona la opción **Display clustered Bar charts**, para generar el gráfico bivariado correspondiente a estas variables. En la ventana statistics, seleccionar ***Phi and Cramer's V***; en la ventana **Cells**, seleccionar **Observed, Expected**, para obtener los valores observados y esperados; seleccionar **Row, Column, Total**, para obtener los porcentajes de la tabla por hilera, columna y total. Finalmente dar **OK**. La prueba de ***Phi***, se presenta a continuación.

Sexo * Visita Ud. la alcaldia para exponer sus inquietudes o necesidades? Crosstabulation

			Visita Ud. la alcaldia para exponer sus inquietudes o necesidades?		
			No	Si	Total
Sexo	Varón	Count	368	257	625
		Expected Count	387.3	237.7	625.0
		% within Sexo	58.9%	41.1%	100.0%
		% within Visita Ud.la alcaldia	51.1%	58.1%	53.8%
		% of Total	31.7%	22.1%	53.8%
	Mujer	Count	352	185	537
		Expected Count	332.7	204.3	537.0
		% within Sexo	65.5%	34.5%	100.0%
		% within Visita Ud.la alcaldia	48.9%	41.9%	46.2%
		% of Total	30.3%	15.9%	46.2%
Total		Count	720	442	1162
		Expected Count	720.0	442.0	1162.0
		% within Sexo	62.0%	38.0%	100.0%
		% within Visita Ud.la alcaldia	100.0%	100.0%	100.0%
		% of Total	62.0%	38.0%	100.0%

Cuadro 4.2. Salida del SPSS para la prueba de Phi en Tablas de Contingencia.

El estadístico ***Phi***, al igual que la prueba de Chi cuadrada se construye a partir de las diferencias entre las frecuencias observadas y esperadas, solo que ***Phi*** toma valores entre 0 y 1. Valores de ***Phi*** próximos a 0, indicarán **no** asociación entre las variables y valores próximos a 1, indicarán una fuerte asociación.

En la tabla de salida dada por el SPSS, se obtuvo un valor de ***Phi*** con una significancia 0.020 que es menor de 0.05, por tanto se rechaza la Ho de independencia entre las variables Sexo y Visita a la Alcaldía, es decir "hay dependencia entre las variables estudiadas"; luego, la relación entre variables no es demasiado fuerte, al obtener un valor pequeño de ***Phi***=0.068.

Copyrighted material

Symmetric Measures

		Value	Approx. Sig.
Nominal by Nominal	Phi	-.068	.020
	Cramer's V	.068	.020
N of Valida Cases		1162	

a. Not assuming the null hypothesis.

b. ysing the asymptotic standard error assuming the null hypothesis.

El gráfico bivariado solicitado, se presenta a continuación.

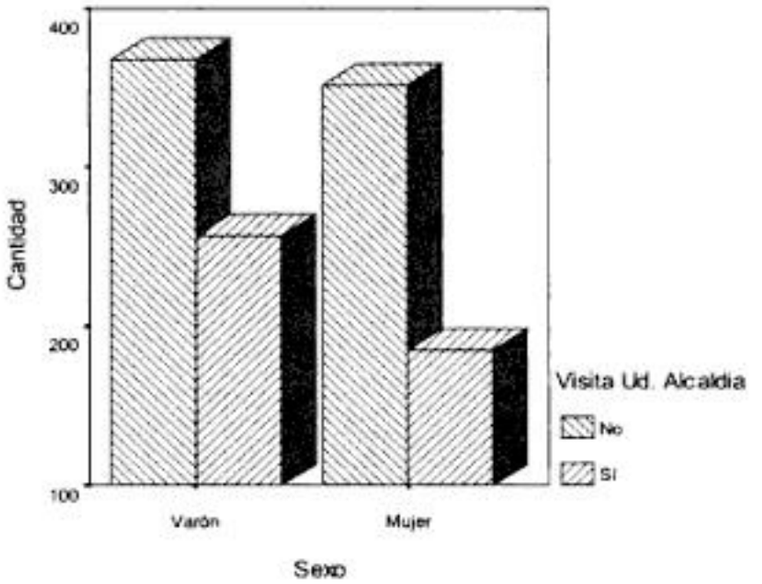


Figura 4.2. ***Relación bivariada de las variables sexo por ¿Visita Ud. la Alcaldía?***

4.3 Medidas de Asociación para dos Variables en Escala Nominal. - El Coeficiente de Contingencia y la V de Cramer -

El coeficiente de ***Phi***, únicamente toma valores entre 0 y 1, en el caso de tablas 2x2. En el caso de tablas mayores -"no simétricas"-, el estadístico ***Phi*** puede alcanzar valores superiores a 1. ***El coeficiente de Contingencia***, es una extensión del coeficiente de ***Phi***, ajustado al caso de que al menos una de las dos variables presente más de dos categorías. ***El coeficiente de Contingencia***, toma valores entre 0 y C_{max}, donde, si ***r*** y ***c*** son el número de categorías de cada una de las dos variables, entonces C_{max}, sería igual a:

$$C_{max} = \sqrt{\ } \textit{min (r - 1, c - 1)} / 1 + \textit{min (r - 1, c - 1)}$$

Valores del ***coeficiente de Contingencia*** próximos a 0, indicarán **no** asociación entre las variables y valores próximos a la cota C_{max}, indicarán una fuerte asociación; obsérvese que la cota C_{max}, será siempre inferior a 1, (Ferran, A. M., 1996).

Copyrighted material

Otra extensión del coeficiente de ***Phi***, ajustado a casos en que la tabla tiene al menos una de las dos variables con más de dos categorías, es ***la V de Cramer***, la cual a diferencia del coeficiente de contingencia, toma valores entre 0 y 1, **no dependiendo de una cota superior**; sin embargo, ***la V de Cramer***, tiende a subestimar el grado el grado de asociación entre las variables. Valores ***la V de Cramer***, próximos a 0, indicarán **no** asociación entre las variables y valores próximos a 1, indicarán una fuerte asociación, (Ferran, A. M., 1996).

Para ilustrar la prueba del ***coeficiente de Contingencia***, y ***la V de Cramer***, se carga la BDD ***"SURVEY33"***, diseñada para estudiar la asociación de escolaridad y sexo. La rutina de comandos a seguir es: **Analyze/ Descriptives Statistics/Croostabs/** en la ventana de diálogo **Row(s)**, debe incluirse la variable que se desea aparezca en la hilera o fila de la tabla, en este caso se incluye la variable *Sexo*; y en la ventana de diálogo **Column(s)**, debe incluirse la variable que se desea aparezca en la columna de la tabla, en este caso se incluye la variable *Escolaridad de la persona*. Luego, se selecciona la opción **Display clustered Bar charts**, para generar el gráfico bivariado correspondiente. En la ventana **statistics**, seleccionar ***Contingency coefficient*** y ***Phi and Cramer's V***; en la ventana **Cells**, seleccionar **Observed, Expected**; seleccionar **Row, Column, Total**, para obtener los porcentajes de la tabla por hilera, columna y total. Finalmente dar **OK**.

El ***coeficiente de Contingencia*** que se presenta a continuación del cuadro 4.3, es igual a 0.137, con una significancia de 0.131 que es mayor de 0.05, por tanto se acepta la Ho de asociación entre las variables Sexo y Escolaridad. El ***estadístico de la V de Cramer***, es igual a 0.139, con una significancia de 0.131 que es mayor de 0.05, por tanto se acepta la Ho de asociación entre variables.

Cuadro 4.3. ***Salida del SPSS para la prueba del coeficiente de Contingencia y la V de Cramer.***

Sexo * Escolaridad de la persona Crosstabulation

			Escolaridad de la persona						
			Primaria Incompleta	Primaria Completa	Secundar Incompleta	Secundaria Completa	Técnico medio	Analfabeta	Total
Sexo	Varón	Count	67	23	62	43	8	12	215
		% within Sexo	31.2%	10.7%	28.8%	20.0%	3.7%	5.6%	100.0%
		% within Escolaridad	55.4%	42.6%	54.4%	40.2%	38.1%	48.0%	48.6%
		% of Total	15.2%	5.2%	14.0%	9.7%	1.8%	2.7%	48.6%
	Mujer	Count	54	31	52	64	13	13	227
		% within Sexo	23.8%	13.7%	22.9%	28.2%	5.7%	5.7%	100.0%
		% within Escolaridad	44.6%	57.4%	45.6%	59.8%	61.9%	52.0%	51.4%
		% of Total	12.2%	7.0%	11.8%	14.5%	2.9%	2.9%	51.4%
Total		Count	121	54	114	107	21	25	442
		% within Sexo	27.4%	12.2%	25.8%	24.2%	4.8%	5.7%	100.0%
		% within Escolaridad	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
		% of Total	27.4%	12.2%	25.8%	24.2%	4.8%	5.7%	100.0%

Symmetric Measures

		Value	Approx. Sig.
Nominal by Nominal	Phi	.139	.131
	Cramer's V	.139	.131
	Contingency Coefficient	.137	.131
N of Valid Cases		442	

a. Not assuming the null hypothesis

b. Using the asymptotic standard error assuming the null hypothesis.

Copyrighted material

El gráfico bivariado solicitado, se presenta a continuación.

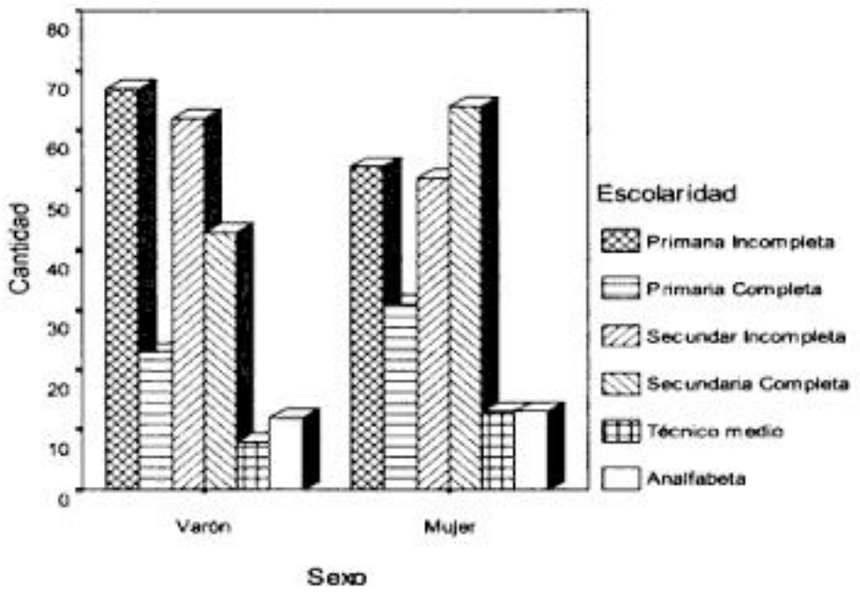


Figura 4.3. ***Relación bivariada de las variables sexo por escolaridad.***

4.4 Medidas de Asociación para Variables en Escala Ordinal.

Para analizar hasta el efecto de al menos una variable en escala ordinal, se consideran las pruebas de ***Gamma, Tau-b de Kendall, y Tau-c de Kendall.***

4.4.1 La Prueba de Gamma.

El estadístico de ***Gamma***, es una medida del grado y tipo de asociación, entre dos variables cualitativas en escala ordinal y toma valores entre -1 y +1. Valores próximos a 1, indican fuerte asociación positiva; a medida que aumentan los valores de una variable, aumentan los de la otra; por el contrario, valores próximos a -1, indican fuerte asociación negativa, es decir, a medida que aumenta una variable, disminuyen los de la otra. Valores próximos a 0, indican **no** asociación, lo que no significa que no pueda existir otro tipo de asociación, (Ferran, A. M., 1996).

Para ilustrar el uso de la prueba de ***Gamma***, se carga la BDD ***"SURVEY11"***. La rutina de comandos a seguir es: **Analyze/Descriptives Statistics/Croostabs/** en la ventana de diálogo **Row(s)**, debe incluirse la variable que se desea aparezca en la hilera o fila de la tabla, en este caso se incluye la variable *Sexo*; y en la ventana de diálogo **Column(s)**, debe incluirse la variable que se desea aparezca en la columna de la tabla, en este caso se incluye la variable *Como valora el servicio de recolección de Basura*. Esta prueba es importante para analizar variables de tipo "Likert".

Luego, se selecciona la opción **Display clustered Bar charts**, para generar el gráfico bivariado correspondiente a estas variables. En la ventana statistics, seleccionar ***Gamma***; en la ventana **Cells**, seleccionar **Observed, Expected**, para obtener los valores observados y esperados; seleccionar **Row, Column, Total**. Finalmente dar **OK**. La salida del SPSS, se presenta en el cuadro 4.4.

Copyrighted material

El estadístico ***Gamma***, toma valores entre -1 y 1. En la tabla de salida dada por el SPSS, debajo del cuadro 4.4, se obtuvo un valor de ***Gamma*** con una significancia de 0.581, que es mayor de 0.05, indica que se acepta la Ho de ausencia de asociación entre las variables *Sexo* y *Cómo valora el servicio de recolección de Basura*, es decir **no** hay asociación significativa entre las variables estudiadas".

Cuadro 4.4. ***Salida del SPSS para la prueba de Gamma.***

Sexo * Como valora el servicio de recoleccion de basura? Crosstabulation

			Como valora el servicio de recoleccion de basura?					
			Excelente	Muy buena	Buena	Regular	Mala	Total
Sexo	Varón	Count	18	36	202	190	60	506
		% within Sexo	3.6%	7.1%	39.9%	37.5%	11.9%	100.0%
		% within Como valora el servicio recoleccion de basura?	75.0%	51.4%	51.9%	56.4%	47.2%	53.4%
		% of Total	1.9%	3.8%	21.3%	20.1%	6.3%	53.4%
	Mujer	Count	6	34	187	147	67	441
		% within Sexo	1.4%	7.7%	42.4%	33.3%	15.2%	100.0%
		% within Como valora el servicio recoleccion de basura?	25.0%	48.6%	48.1%	43.6%	52.8%	46.6%
		% of Total	.6%	3.6%	19.7%	15.5%	7.1%	46.6%
Total		Count	24	70	389	337	127	947
		% within Sexo	2.5%	7.4%	41.1%	35.6%	13.4%	100.0%
		% within Como valora el servicio recoleccion de basura?	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
		% of Total	2.5%	7.4%	41.1%	35.6%	13.4%	100.0%

Symmetric Measures

		Value	Asymp. Std. Error [a]	Approx. T [b]	Approx. Sig.
Ordinal by Ordinal	Gamma	.029	.052	.552	.581
N of Valid Cases		947			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

El gráfico bivariado solicitado, se presenta a continuación

Copyrighted material

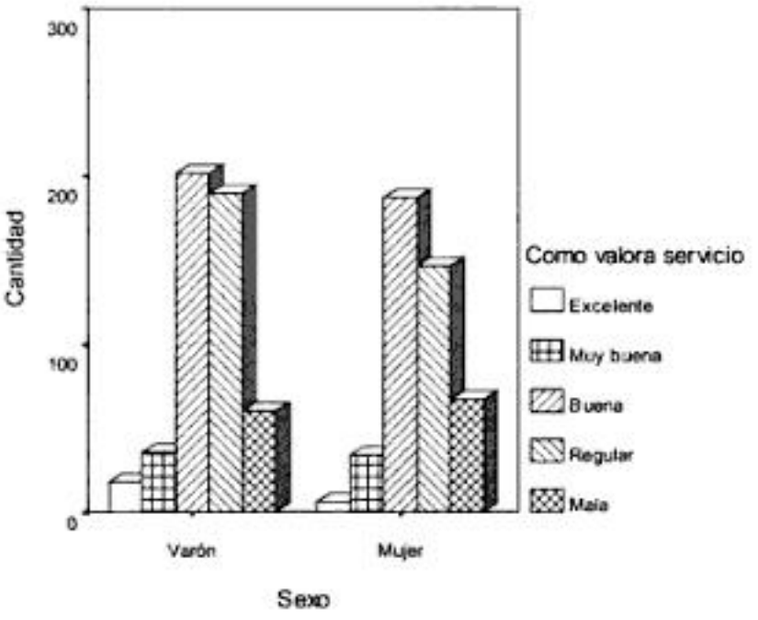


Figura 4.4. ***Relación bivariada de sexo por ¿Cómo valora el servicio de recolección de Basura?.***

4.4.2 Las pruebas de Tau-b de Kendall, y Tau-c de Kendall.

La medida de ***Tau-b de Kendall***, es una extensión de la ***Gamma***, en el sentido de que tanto la situación bajo la que puede ser aplicada como su interpretación, es la misma. Sin embargo, presenta el inconveniente de que dichos valores solo pueden ser alcanzados cuando la tabla de contingencia sea cuadrada, (2x2, 3x3, 4x4, etc.), (Ferran, A. M., 1996).

Por otra parte, la medida de ***Tau-c de Kendall***, es una corrección de la ***Tau-b de Kendall***, para el caso de tablas de contingencia que contienen variables con distinto número de categorias. Frente a ***Tau-b***, la prueba de ***Tau-c*** presenta la ventaja de poder alcanzar los valores de -1 y 1, cuando el número de categorías de las dos variables es distinto. Sin embargo, tiene la desventaja de subestimar el verdadero grado de asociación entre las variables, (Ferran, A. M., 1996). Esta prueba es importante para analizar variables de tipo "Likert".

Para ilustrar la prueba ***Tau-c de Kendall***, se carga la BDD ***"SURVEY22"***. La rutina de comandos a seguir es: **Analyze/Descriptives Statistics/ Croostabs/** en la ventana de diálogo **Row(s)**, debe incluirse la variable que se desea aparezca en la hilera o fila de la tabla, se incluye la variable *Municipio*; y en la ventana de diálogo **Column(s)**, debe incluirse la variable que se desea aparezca en la columna de la tabla, se incluye la variable *Como valora el servicio limpieza del mercado*.

Luego, se selecciona la opción **Display clustered Bar charts**. En **statistics**, seleccionar ***Tau-c de Kendall***; en la ventana **Cells**, seleccionar **Observed**; seleccionar **Row, Total**. Finalmente dar **OK**. La tabla de salida se presenta en el cuadro 4.5.

En la tabla de salida, debajo del cuadro 4.5, se presenta el estadístico ***Tau-c de Kendall*** para el que se obtuvo un valor de significancia igual a 0.000, que es menor de 0.05, esto indica que se rechaza la Ho de ausencia de asociación, entre las variables Municipio y Cómo valora el servicio de limpieza de mercado. es decir "hay una asociación significativa entre las variables estudiadas", el valor de asociación es negativo y bajo (-0.331).

Copyrighted material

Cuadro 4.5. ***Salida del SPSS para la prueba Tau-c de Kendall.***

Nombre del Municipio * Como valora el servicio limpieza de mercado? Crosstabulation

			Como valora el servicio limpieza de mercado?					
			Excelente	Muy buena	Buena	Regular	Mala	Total
Nombre del Municipio	Condega	Count			32	63	34	129
		% within Nombre del Municipio			24.8%	48.8%	26.4%	100.0%
		% of Total			7.4%	14.5%	7.8%	29.7%
	Municipio Jinotega	Count			7	39	95	141
		% within Nombre del Municipio			5.0%	27.7%	67.4%	100.0%
		% of Total			1.6%	9.0%	21.8%	32.4%
	El Sauce	Count	2	15	93	47	8	165
		% within Nombre del Municipio	1.2%	9.1%	56.4%	28.5%	4.8%	100.0%
		% of Total	.5%	3.4%	21.4%	10.8%	1.8%	37.9%
Total		Count	2	15	132	149	137	435
		% within Nombre del Municipio	.5%	3.4%	30.3%	34.3%	31.5%	100.0%
		% of Total	.5%	3.4%	30.3%	34.3%	31.5%	100.0%

Symmetric Measures

		Value	Asymp. Std. Error [a]	Approx. T [b]	Approx. Sig.
Ordinal by Ordinal	Kendall's tau-c	-.331	.038	-8.612	.000
N of Valid Cases		435			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

El gráfico bivariado solicitado, se presenta a continuación.

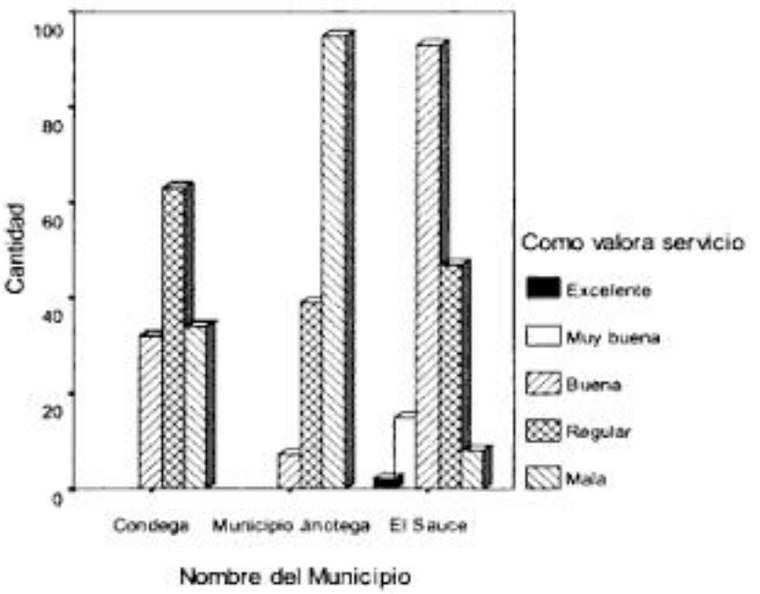


Figura 4.5. ***Relación bivariada de las variables municipio por ¿Cómo valora el servicio de limpieza de mercado.***

Copyrighted material

4.5 *Medidas de Asociación para Variables en Escala de Intervalo o Razón.*

4.5.1 *El coeficiente Eta.*

El coeficiente ***Eta***, es una medida similar al "**R**", la cual es apropiada para analizar los valores de ***una variable Y, en escala de intervalo o de razón***, en los distintos grupos o subpoblaciones establecidas por los valores de otra variable **X**, cualitativa. El coeficiente ***Eta***, toma valores entre 0 y 1; por tanto, valores de ***Eta*** próximos a 0 indicarán que el comportamiento de **Y** es independiente de los grupos (X): la media de Y es la misma en todos ellos. En cambio, valores de ***Eta*** próximos a 1, indicarán mucha dependencia: la media de Y es mayor o menor que la media global dependiendo del grupo. ***El cuadrado de Eta, puede interpretarse como la proporción de la variabilidad de la variable dependiente Y, explicada por los valores de la independiente, X***, (Ferran, A. M., 1996).

Para ilustrar el uso del **coeficiente *Eta***, se carga la BDD ***"FARMERS22"***. La rutina de comandos a seguir es: **Analyze/Descriptives Statistics/ Croostabs/** en la ventana de diálogo **Row(s)**, debe incluirse la variable que se desea aparezca en fila de la tabla, en este caso se incluye la variable *Tipologia del productor(a)*; y en la ventana de diálogo **Column(s)**, debe incluirse la variable que aparecerá en la columna de la tabla, en este caso se incluye la variable *Estratos*.
Seleccionar la opción **Display clustered Bar charts**. En **statistics**, seleccionar ***Eta***; en la ventana **Cells**, seleccionar **Observed, Expected**; seleccionar **Row, Column, Total**. Finalmente dar **OK**.

Cuadro 4.6. ***Salida del SPSS para la prueba Eta, en Tablas de Contingencia.***

Tipología del productor(a) * Estratos de la Microcuenca Crosstabulation

			Estratos de la Microcuenca			
			Estrato Bajo	Estrato Medio	Estrato Alto	Total
Tipología del productor(a)	Pequeña	Count	20	13	43	76
		Expected Count	19.1	15.2	41.7	76.0
		% within Tipología	26.3%	17.1%	56.6%	100.0%
		% within Estratos	58.8%	48.1%	58.1%	56.3%
		% of Total	14.8%	9.6%	31.9%	56.3%
	Mediana	Count	11	9	23	43
		Expected Count	10.8	8.6	23.6	43.0
		% within Tipología	25.6%	20.9%	53.5%	100.0%
		% within Estratos	32.4%	33.3%	31.1%	31.9%
		% of Total	8.1%	6.7%	17.0%	31.9%
	Grande	Count	3	5	8	16
		Expected Count	4.0	3.2	8.8	16.0
		% within Tipología	18.8%	31.3%	50.0%	100.0%
		% within Estratos	8.8%	18.5%	10.8%	11.9%
		% of Total	2.2%	3.7%	5.9%	11.9%
Total		Count	34	27	74	135
		Expected Count	34.0	27.0	74.0	135.0
		% within Tipología	25.2%	20.0%	54.8%	100.0%
		% within Estratos	100.0%	100.0%	100.0%	100.0%
		% of Total	25.2%	20.0%	54.8%	100.0%

Copyrighted material

El coeficiente ***Eta***, que se obtiene en el cuadro de salida es en dos sentidos: Un primer caso, considerando a *Tipología del productor(a)*, como la variable dependiente, para este caso se obtiene un ***Eta*** igual a 0.108. Un segundo caso, considerando a *Estratos de la microcuenca,* como la variable dependiente, para el cual se obtuvo un ***Eta*** igual a 0.014. Es notorio que en ambos casos, se obtienen valores de ***Eta*** próximos a 0, lo que indica que el comportamiento de *tipología del productor*, es independiente de los *estratos de la microcuenca*. En este caso, que el principal propósito del análisis está orientado a saber en que medida el *estrato de la microcuenca* determina una *tipología de productor(a)*, el valor dependiente que más interesaría sería ***Eta*** igual a 0.108.

Directional Measures

			Value
Nominal by Interval	Eta	Tipología del productor (a) Dependent	.108
		Estratos de la Microcuenca Dependent	.014

El gráfico bivariado solicitado, se presenta a continuación.

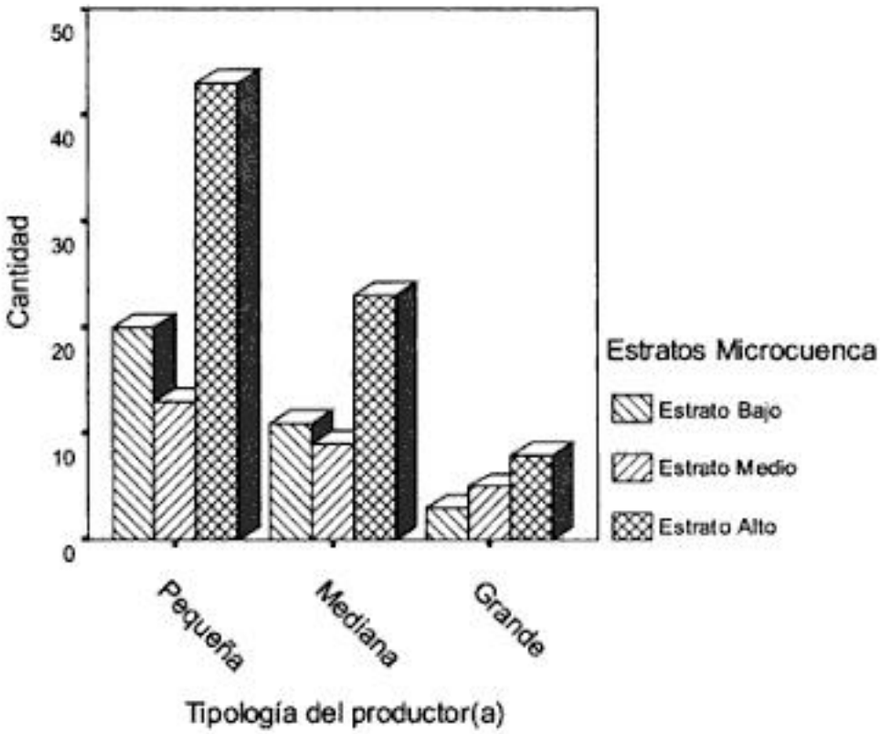


Figura 4.6. Relación bivariada de las variables tipología del productor(a) por estrato.

4.5.2 Los Coeficientes de Correlación de Pearson y Spearman.

Vinculado a las medidas de asociación en escala de intervalo o razón, se encuentran los **Coeficientes de Correlación de Pearson y Spearman.** El uso de ambos coeficientes, se ilustra como una opción del mismo ejemplo de la BDD ***"FARMERS22".*** La rutina de comandos a seguir, es igual al ejemplo inmediato anterior, pero en la ventana de diálogo **statistics**, se debe seleccionar ***Correlations.*** La hoja de salida que para ambos coeficientes solicitados se presenta a continuación.

Copyrighted material

Symmetric Measures

		Value	Asymp. Std. Error [a]	Approx. T [b]	Approx. Sig.
Interval by Interval	Pearson's R	.029	.83	-.032	.974 [c]
Ordinal by Ordinal	Spearman Correlation	947	.85	-.189	.850 [c]
N of Valid Cases					

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

c. Based on normal approximation.

El coeficiente de Correlación de Pearson **(R)**, mide el grado de asociación lineal entre dos variables medidas en escala de intervalo o de razón, tomando valores entre -1 y 1. Valores de **(R)** próximos a 1, indicarán una fuerte asociación lineal positiva; en cambio valores de **(R)** próximos a -1, indicarán una fuerte asociación lineal negativa; y valores de **(R)** próximos a 0 indicarán **no** asociación. ***Su cuadrado*** **(R^2)**, ***puede interpretarse como la proporción de la variabilidad de la variable Y, explicada en función de la variable X***, (Ferran, A. M., 1996).

En el ejemplo aquí realizado, el coeficiente de Correlación de Pearson **(R)** que se obtuvo es cercano a 0, es decir igual a -0.003, y tiene una significancia de 0.97 > 0.05, por lo tanto se acepta la Ho de ausencia de asociación entre las variables. Es decir, es **no** significativa la asociación entre *Tipología del productor(a)* y *estratos de la microcuenca.*

El coeficiente de correlación de Spearman, es una variante del coeficiente de Correlación de Pearson **(R)**, esta variante consiste en que, en lugar de medir el grado de asociación lineal a partir de los propios valores de las variables, se mide a partir de la asignación de rango de valores ordenados. ***En este sentido, el coeficiente de correlación de Spearman, es una medida también adecuada en el caso de variables en escala ordinal*** **(variables Likert).** ***Por lo demás, sus valores se interpretan exactamente igual al coeficiente de Correlación de Pearson*** **(R),** (Ferran, A. M., 1996).

En el ejemplo aquí expuesto, para *el coeficiente de correlación de Spearman* se obtuvo una significancia de 0.85 > 0.05, (con un valor bajo de -0.016), por lo que se acepta la Ho de ausencia de asociación. El *coeficiente de correlación de Spearman,* en este caso, tiene una interpretación igual a la del *coeficiente de Correlación de Pearson* **(R).**

Copyrighted material

Capítulo 5. Análisis de Varianza Univariado: Diseño Completo al Azar DCA (One Way ANOVA).

5.1 El Análisis de Varianza para un Diseño Completamente Aleatorizado.

El diseño completamente aleatorizado, D.C.A, es también conocido como One Way ANOVA. Es un diseño muy útil para condiciones en que las unidades experimentales presentan homogeneidad relativa, lo que permite colocar completamente al azar a los tratamientos en cada una de las unidades experimentales; es decir, este diseño no impone restricciones a las unidades experimentales. Este diseño es también útil para ensayos de campo en que las unidades experimentales **"no"** requieren de agrupamiento o bloqueo en particular, esto es cuando el efecto de los tratamientos en estudio **no estará determinado** por la heterogeneidad del suelo; tales como ensayos MIP, estudios de dietas alimenticias para aves en galerones, etc. El Modelo Aditivo Lineal (MAL) para un DCA, es el siguiente:

5.2 El Modelo Aditivo Lineal para un DCA.

$$Y_{ij} = \mu + \tau_i + \varepsilon_{ij} \quad \text{..... donde:}$$

i = 1, 2, 3, ...t... tratamientos.
j = 1, 2, 3, ...n...observaciones.

Y_{ij} = La j-ésima observación del i-ésimo tratamiento.

μ = Es la media poblacional a estimar a partir de los datos del experimento.

τ_i = Efecto del i-ésimo tratamiento a estimar a partir de los datos del experimento.

ε_{ij} = Efecto aleatorio de variación.

5.3 Procedimiento Estadístico para un Experimento establecido en D.C.A.

A continuación se presentan los datos de un experimento establecido en la Estación Experimental Raúl González del Valle de Sébaco, a partir del cual se realizó un muestreo completamente al azar para determinar el potencial agroindustrial de cinco variedades de tomate industrial. En el cuadro 5.1, se presentan los datos del peso de jugo obtenido. Como un estudio de caso, ver Pedroza, P.H., (1993), las páginas 73-81.

Se debe tener presente que para la correcta aplicación del análisis de varianza univariado (**ANOVA**), los datos obtenidos de las variables dependientes deben ser: a) muestras tomadas al azar de poblaciones normales, para lo que se realiza la prueba de **Normalidad de los datos ó Prueba de Kolmogorov-Smirnov**; y b) deben tener varianzas semejantes los diversos grupos en comparación, lo que se verifica mediante la prueba de **Homogeneidad de Varianzas ó Prueba de Levene**.

Copyrighted material

Cuadro 5.1. Peso del jugo (en gramos) obtenido para diferentes variedades de tomate industrial.

Variedades	OBSERVACIONES				$Y_{i.}$	$\bar{Y}_{i.}$
	1	2	3	4		
Martí	656.30	718.40	586.60	746.20	2707.50	676.87
Topacio	784.40	713.40	915.80	629.60	3043.20	760.80
Estela	924.50	822.80	824.20	978.50	3550.00	887.50
VF-134	534.40	685.10	567.20	655.50	2442.20	610.55
UC-82	640.70	658.80	532.70	614.40	2446.60	611.65

Con estos datos, se construye en SPSS la BDD llamada ***DCA en UNIFACTORIAL*** que contiene tres variables: 1ra) "Variedades", con valores de 1 a 5; 2da) "Observaciones" o repeticiones, con valores de 1 a 4; y 3ra) "Peso del Jugo en gr.", con los datos del peso de jugo obtenido para cada tratamiento.

Con el **SPSS**, para hacer el análisis estadísticos de un Diseño Completo al Azar, (DCA), se requiere de una variable dependiente continua; y de una variable independiente discreta que genere grupos **o Tratamientos**. El Diseño Completo al Azar se resuelve en **SPSS**, utilizando la siguiente rutina de comandos: **Analyze/ Compare Means/ One Way ANOVA/** en **Dependent List, se debe introducir la variable dependiente *–peso de jugo en gr-*; y en Factor se debe introducir la variable *"variedades"***. Luego, dentro del comando **Options,** se le solicita al programa que realice la prueba de Levene o de homogeneidad de varianza. Usando la opción **Post Hoc**, se le solicita realizar la prueba de separación de medias, para este ejemplo se solicito la prueba de **Duncan.** La hoja de salida para las pruebas solicitadas, se presentan a continuación.

Cuadro 5.2. Tabla de estadísticas descriptivas del DCA, One way ANOVA

Descriptives

Peso del Jugo en gr.

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
1	4	676.8750	70.9535	35.4767	563.9722	789.7778	586.60	746.20
2	4	760.8000	121.1640	60.5820	568.0010	953.5990	629.60	915.80
3	4	887.5000	77.1211	38.5605	764.7832	1010.2168	822.80	978.50
4	4	610.5500	71.3121	35.6561	497.0765	724.0235	534.40	685.10
5	4	611.6500	55.7007	27.8503	523.0178	700.2822	532.70	658.80
Total	20	709.4750	129.8857	29.0433	648.6866	770.2634	532.70	978.50

Copyrighted material

Cuadro 5.3. ***Prueba de homogeneidad de varianzas, o prueba de Levene.***

Test of Homogeneity of Variances

Peso del Jugo en gr.

Levene Statistic	df1	df2	Sig.
.956	4	15	.460

La prueba de homogeneidad de varianzas, basada en el estadístico de ***Levene***, obtuvo una Significancia de 0.46 > 0.05, por lo tanto se acepta la hipótesis nula de homogeneidad de varianzas. Esto indica que se puede proceder correctamente a realizar el ANOVA.

La prueba de normalidad de los datos o prueba de Kolmogorov-Smirnov, se solicita por separado, utilizando el comando **Analyze/ Nomparametric Tests / Simple K-S/ Test distribution-Normal.**

Cuadro 5.4. ***Prueba de normalidad de los datos o Prueba de Kolmogorov-Smirnov.***

One-Sample Kolmogorov-Smirnov Test

		Peso del Jugo en gr.
N		20
Normal Parameters[a,b]	Mean	709.4750
	Std. Deviation	129.8857
Most Extreme Differences	Absolute	.152
	Positive	.152
	Negative	-.094
Kolmogorov-Smirnov Z		.679
Asymp. Sig. (2-tailed)		.746

a. Test distribution is Normal.

b. Calculated from data.

La prueba de K-S obtuvo una Significancia de 0.746 > 0.05, por lo tanto se acepta la hipótesis nula de normalidad de los datos. Esto indica que se puede proceder correctamente a realizar el ANOVA.

Cuadro 5.5. Tabla de Análisis de Variancia, ANOVA.

ANOVA

Peso del Jugo en gr.

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	218983.2	4	54745.803	8.086	.001
Within Groups	101552.3	15	6770.151		
Total	320535.5	19			

Para analizar este cuadro, se debe observar el valor Sig. (la significación del valor "F") para la fuente de variación Entre Grupos o "Between Groups" que es el efecto para "Tratamientos". En este ejemplo:

Copyrighted material

La Sig. de los "Tratamientos" es 0.001 < 0.05, por tanto se rechaza la Ho de igualdad entre tratamientos, o bien se dice que existen diferencias significativas entre tratamientos, lo que indica que **"al menos uno de lo tratamientos tiene un promedio diferente"**.

El siguiente paso es determinar cuales son los tratamientos que difieren entre si, para esto se utiliza la Técnica de Separación de Medias. Como ejemplo en este caso, se utilizó la prueba de Rangos Múltiples de Duncan; la salida del SPSS se observa en el cuadro siguiente.

Cuadro 5.6. Salida del SPSS para la separación de medias por la prueba de Duncan.

Peso del Jugo en gr.

Duncan[a]

Variedades	N	Subset for alpha = .05		
		1	2	3
4	4	610.5500		
5	4	611.6500		
1	4	676.8750	676.8750	
2	4		760.8000	
3	4			887.5000
Sig.		.297	.170	1.000

Means for groups in homogeneous subsets are displayed.
a. Uses Harmonic Mean Sample Size = 4.000.

El cuadro de salida dado por la prueba de Duncan, se puede presentar de la siguiente manera:

Cuadro 5.7. Presentación de medias y su significación estadística dada por la prueba de Duncan.

Tratamientos	Peso del Jugo en gr.	Significancia Estadística
Estela	887.50	a
Topacio	760.80	b
Martí	676.87	bc
UC-82	611.65	c
VF-134	610.55	c

Nota: Letras iguales, indica promedios iguales, según prueba de Duncan al 5%.

Basados en la salida dada por la prueba de Duncan, se puede afirmar que los tratamientos se clasifican en cuatro categorías estadísticas: Categoría "a", determinada por la variedad Estela. La segunda categoría "b", está formada por la variedad Topacio. La tercera categoría "bc", está formada por la variedad Martí. La cuarta categoría "c", está formada por las variedades UC-82 y VF-134.

Copyrighted material

Estas mismas categorías pueden observarse en el gráfico de **"error bar"**, solicitada por aparte, dentro del Módulo de **Graphs**.

En el cuadro 5.2, se observan los promedios e intervalos de confianza para cada tratamiento, los cuales son ilustrados en la figura 5.1 de **"error bar",** demostrándose que la respuesta de tratamientos es significativa. La excepción a esta regla es el caso de la variedad Marti, que aparece en el gráfico como un subconjunto de la variedad Topacio, indicando claramente que ese caso es **NS**, y por tanto es la categoría **"bc"**. Así mismo, ocurre el caso de la variedad UC-82, que se observa en el gráfico 5.1., como un subconjunto de la variedad VF-134 lo cual indica un efecto **NS** de tratamientos, por lo tanto comparten la misma categoría **"c"**.

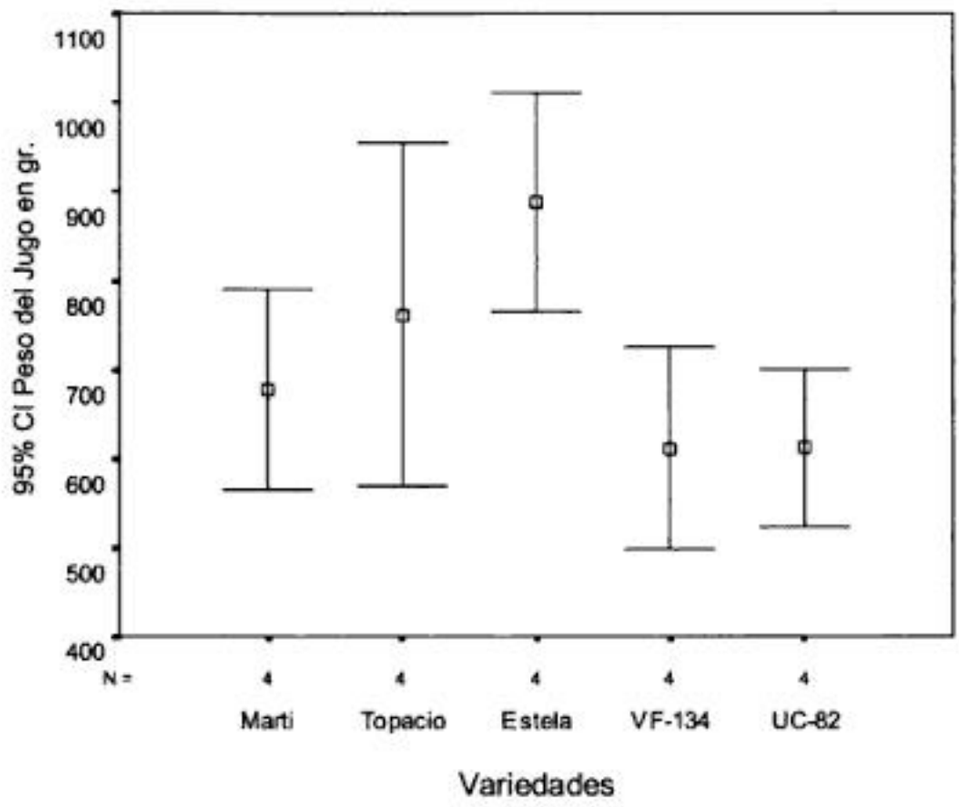


Figura 5.1. Gráfico de "error bar" para los tratamientos.

Copyrighted material

Capítulo 6. Análisis de Varianza Univariado: Diseño de Bloques Completos al Azar BCA - (Two Way ANOVA).

6.1 El Análisis de Varianza para un Diseño de Bloques Completos al Azar.

El diseño de bloques completos al azar, B.C.A, conocido también como Two Way ANOVA, es el diseño más usado en el campo agrícola cuando se hacen experimentos como por ejemplo: evaluación de variedades de un cultivo, épocas de siembra, distancias de siembra, prueba de niveles de nutrientes, etc.

Un diseño de bloques completos al azar (B.C.A) es aquel en que las U.E. se distribuyen en grupos, de manera tal que las U.E. dentro de un bloque o grupo son relativamente homogéneas, pero entre bloques son heterogéneas con relación al gradiente que se está bloqueando. En general este diseño se recomienda para experimentos con un número de tratamientos comprendido entre 3 y 15, y cuando es posible agrupar las unidades experimentales en bloques de igual tamaño.

Algunos criterios acerca de la disposición de los bloques en el campo, son los siguientes: 1) Cuando la gradiente de fertilidad del suelo es conocida, los bloques se colocan perpendicular a la gradiente; 2) Cuando la gradiente de fertilidad del suelo ocurra en dos direcciones aproximadamente perpendiculares entre si, un DCL debe ser usado. Sin embargo, si se utiliza un BCA, los bloques deben de ser cuadrados; 3) Cuando el gradiente de fertilidad del suelo no es conocido, o es errática, entonces los bloques deben se ser cuadrados, (Reyes, C., 1982). El Modelo Aditivo Lineal (MAL) para un BCA, es el siguiente:

6.2 El Modelo Aditivo Lineal para un BCA.

$$Y_{ij} = \mu + \tau_i + \beta_j + \varepsilon_{ij} \quad \ldots\ldots, \text{donde:}$$

$i = 1, 2, 3, \ldots t$ = tratamientos
$j = 1, 2, 3, \ldots r$ = repeticiones

Y_{ij} = La j-ésima observación del i-ésimo tratamiento.
μ = Es la media poblacional a estimar a partir de los datos del experimento.
τ_i = Efecto del i-ésimo tratamiento a estimar a partir de los datos del experimento.
β_j = Estimador del efecto debido al j-ésimo bloque.
ε_{ij} = Efecto aleatorio de variación.

6.3 Procedimiento estadístico para un experimento establecido en B.C.A.

Para ejemplificar el análisis de un BCA, se presentan los resultados de un experimento de campo realizado en la Estación Experimental Raúl González del Valle de Sébaco, para determinar el potencial agronómico de cinco variedades de tomate industrial. En el cuadro 6.1, se presentan los tratamientos en estudio y los datos obtenidos del ensayo. Como un estudio de caso, ver Pedroza, P.H., (1993), en páginas 82-98.

Copyrighted material

Cuadro 6.1. ***Datos del diámetro ecuatorial del fruto (en cm), obtenido para diferentes variedades de tomate industrial.***

Variedades	Bloques				Yi.	$\bar{Y}_{i.}$
	I	II	III	IV		
Martí	6.64	6.59	6.33	5.80	25.36	6.34
Topacio	7.37	6.21	6.19	6.39	26.16	6.54
Estela	6.87	7.03	6.53	6.66	27.09	6.77
VF-134	5.79	5.49	5.54	5.91	22.73	5.68
UC-82	5.19	5.48	5.42	5.46	21.55	5.38
Y.j	31.86	30.80	30.01	30.22	122.89	6.14

Con estos datos, se genera en SPSS la BDD llamada ***BCA en UNIFACTORIAL*** que contiene tres variables: 1ra) "Variedades ó Tratamientos", con valores de 1 a 5; 2da) "Bloques", con valores de 1 a 4; y 3ra) "Diámetro", con los datos del diámetro obtenido para cada tratamiento en cada bloque.

Para resolver en el **SPSS** el análisis estadísticos de un Diseño de Bloques Completo al Azar, se deben usar los comandos **Analyze/General Linear Model/ Univariate/** en **Dependent variable, se debe cargar la variable dependiente *–diámetro–*; y en Fixed Factor se deben cargar las variables *"tratamientos" y "bloques"***. Luego, dentro del comando **Model,** se deben definir los efectos principales del modelo, usando la opción **Custom (personalizado) / y se construyen los términos del modelo, incorporando una a la vez, c/u de las variables o factores fijos; recuerde NO debe pedirse interacción, *ya que el diseño de B.C.A., asume que NO existe interacción entre "tratamientos" y "bloques"***. Con **Post Hoc** se selecciona la prueba de separación de medias, en el ejemplo usaremos la prueba de **Duncan.**

Cuadro 6.2. ***Salida del ANOVA para un Diseño de Bloques Completos al Azar.***

Tests of Between-Subjects Effects

Dependent Variable: Diametro

Source	Type III. Sum of Squares	df	Mean	F	Sig.
Corrected Model	5.914 [a]	7	.845	7.818	.001
Intercept	755.098	1	755.098	6987.277	.000
BLOQUE	.412	3	.137	1.271	.328
TRAT	5.502	4	1.376	12.728	.000
Error	1.297	12	.108		
Total	762.308	20			
Corrected Total	7.211	19			

a. R Squared = .820 (Adjusted R Squared = .715)

Para analizar este cuadro, se debe observar el valor Sig. (la significacion del valor "F") para "Bloque" y para

Copyrighted material

"Tratamiento". En este ejemplo:

- La Significancia de "Bloque" es 0.328 > 0.05, por tanto se acepta la Ho de igualdad entre los bloques, o bien se dice que **no** hay diferencias significativas entre los bloques.
- La Significancia de los "Tratamientos" es 0.000 < 0.05, por tanto se rechaza la Ho de igualdad entre tratamientos, o bien se dice que existen diferencias significativas entre tratamientos, lo que indica que "**al menos uno de lo tratamientos tiene un promedio diferente**". El siguiente paso es determinar cuales son los tratamientos que difieren entre si, para esto se utiliza la Técnica de Separación de Medias. La prueba de Rangos Múltiples de Duncan, según la salida del SPSS se observa en el siguiente cuadro.

Cuadro 6.3. Salida del SPSS para la separación de medias dada por la prueba de Duncan.

Diametro del fruto

Duncan[a,b]

TRATAMIENTOS	N	Subset 1	Subset 2
UC-82	4	5.3875	
VF-134	4	5.6825	
Marti	4		6.3400
Topacio	4		6.5400
Estela	4		6.7725
Sig.		.228	.101

Means for groups in homogeneous subsets are displayed.
Based on Type III Sum of Squares
The error term is Mean Square(Error) = .108.

a. Uses Harmonic Mean Sample Size = 4.000.
b. Alpha = .05.

El cuadro de salida dado por la prueba de Duncan, se puede presentar de la siguiente manera:

Cuadro 6.4. Presentación de medias y su significación estadística dada por la prueba de Duncan.

Tratamientos	Diámetro del Fruto en cm	Significancia estadística
Estela	6.77	a
Topacio	6.54	a
Marti	6.34	a
VF-134	5.68	b
UC-82	5.38	b

Nota: Letras iguales, indica promedios iguales, según prueba de Duncan al 5%.

Basados en la salida dada por la prueba de Duncan, se puede afirmar que los tratamientos se clasifican en

Copyrighted material

dos categorías estadísticas:

La primera categoría "a", determinada por las variedades Estela, Topacio y Marti.
La segunda categoría "b", está formada por las variedades VF-134 y UC-82.

Un detalle importante que del SPSS, es que se le puede pedir los intervalos de confianza y el grafico de los tratamientos, aún dentro de la misma rutina del GLM Univariado, con lo cual se ilustra mucho mejor el efecto de los tratamientos.

El gráfico de **"error bar"** se solicita por aparte en SPSS, dentro del Módulo de **Graphs**.

TRATAMIENTO

Dependent Variable: Diametro

TRATAMIENTO	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
Marti	6.340	.164	5.982	6.698
Topacio	6.540	.164	6.182	6.898
Estela	6.773	.164	6.414	7.131
VF-134	5.683	.164	5.324	6.041
UC-82	5.388	.164	5.029	5.746

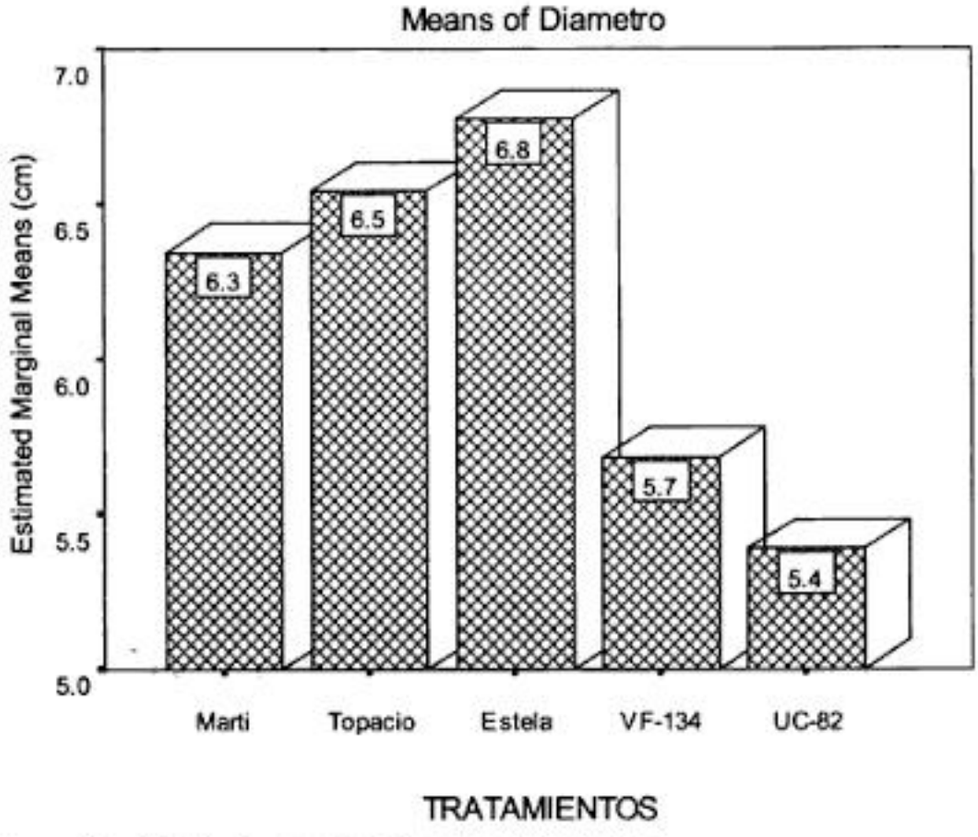


Figura 6.1. ***Promedios del diámetro ecuatorial para los tratamientos.***

Copyrighted material

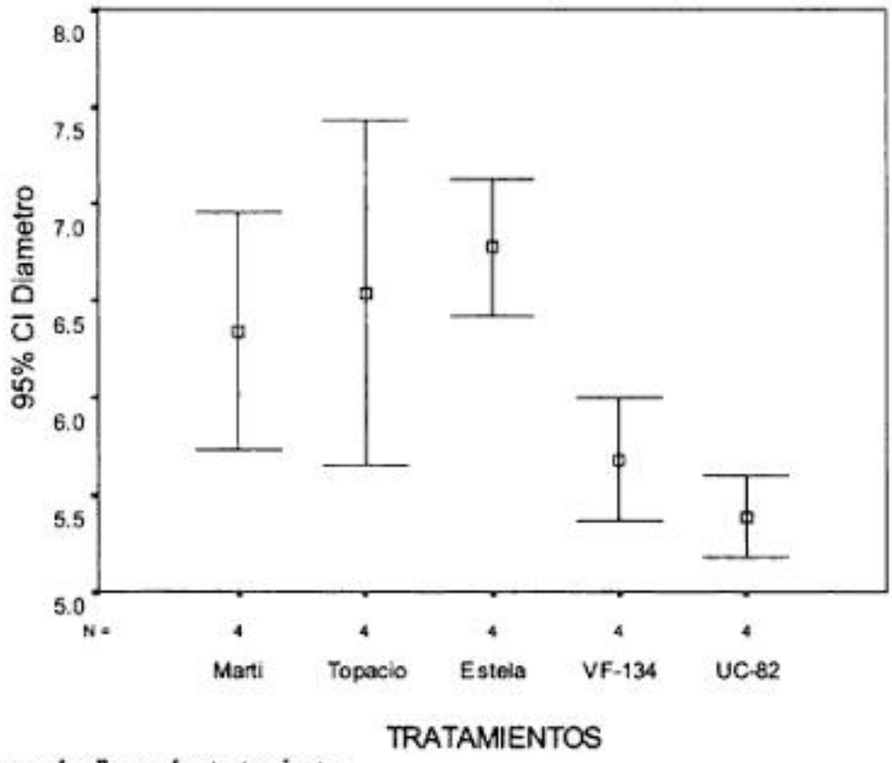


Figura 6.2. Gráfico de "error bar" para los tratamientos.

Con el gráfico del "error bar", en el que se grafica los intervalos de confianza para cada tratamiento, se evidencia la significancia de lo tratamientos definida por la separación de medias; por un lado hay tres intervalos de confianza que tiene grupos de medias similares con valores de 6.34 a 6.77, estos hacen la categoría estadística "a" (Marti, Topacio y Estela). Por otra parte, un grupo de dos intervalos de confianza con valores de media de 5.38 a 5.68, los que constituyen la categoría estadística "b" (VF-134 y UC-82).

Copyrighted material

Capítulo 7. Análisis de Varianza Univariado: Factoriales: Experimentos Bifactoriales establecidos en DCA.

7.1 El Análisis de Varianza para un Bifactorial en DCA.

El experimento factorial, es aquel en el cual los tratamientos son constituidos por la combinación de c/u de los niveles de un factor con todos y c/u de los niveles de los otros factores en el ensayo. Los experimentos factoriales, (dos o más factores en estudio), **no** son un diseño en sí, más bien son un arreglo de tratamientos que se distribuyen en los diseños comunes: D.C.A., B.C.A y D.C.L.

En los experimentos factoriales, dos o más factores son estudiados simultáneamente y cualquier factor puede proporcionar varios tratamientos. En los ensayos factoriales, se estudia por un lado ***los efectos principales,*** o acción independiente de los factores; por otro lado se estudia ***el efecto de interacción*** entre ellos. En la nomenclatura básica de los factoriales, cada factor en estudio se designa con letras mayúsculas (**A, B, C,** etc.); y los niveles o modalidades de cada factor, se designan con letras minúsculas y números subíndices (a_1, a_2; b_1, b_2, etc.). Normalmente, se construye una tabla de doble entrada para indicar los efectos principales y posibles efectos de interacción.

Cuadro 7.1. Cuadro de doble entrada para construir los tratamientos factoriales.

Densidades	**Variedades**		
	a_1	a_2	a_3
b_1	$a_1 b_1$	$a_2 b_1$	$A_3 b_1$
b_2	$a_1 b_2$	$a_2 b_2$	$a_3 b_2$

7.2 Los Efectos Simples, Principales y de Interacción.

Para definir los efectos individuales -simple y principal-, así como la interacción, supongamos un experimento bifactorial (A y B), con dos niveles cada factor (a_1, a_2 y b_1, b_2).

Cuadro 7.2. Efectos Simples, Principales y de Interacción entre factores.

	Factor A		**Efecto simple de A**	**Efecto principal de A**
Factor B	a_1	a_2	$(a_2 - a_1)$	$[(Y_{21.} - Y_{11.}) + (Y_{22.} - Y_{12.})]/2$
b_1	$Y_{11.}$	$Y_{21.}$	$Y_{21.} - Y_{11.}$	
b_2	$Y_{12.}$	$Y_{22.}$	$Y_{22.} - Y_{12.}$	
Efecto simple de B ($b_2 - b_1$)	$(Y_{12.} - Y_{11.})$	$(Y_{22.} - Y_{21.})$		
Efecto principal de B	$[(Y_{12.} - Y_{11.}) + (Y_{22.} - Y_{21.})]/2$			

Copyrighted material

Los efectos simples de un factor, son aquellos representados por las diferencias de los niveles de un factor, a un mismo nivel del otro factor. El efecto principal de un factor, es el promedio de los efectos simples para un mismo factor.

El efecto de Interacción:

Caso A: **Efecto de los factores es aditivo: No hay interacción entre los factores.**

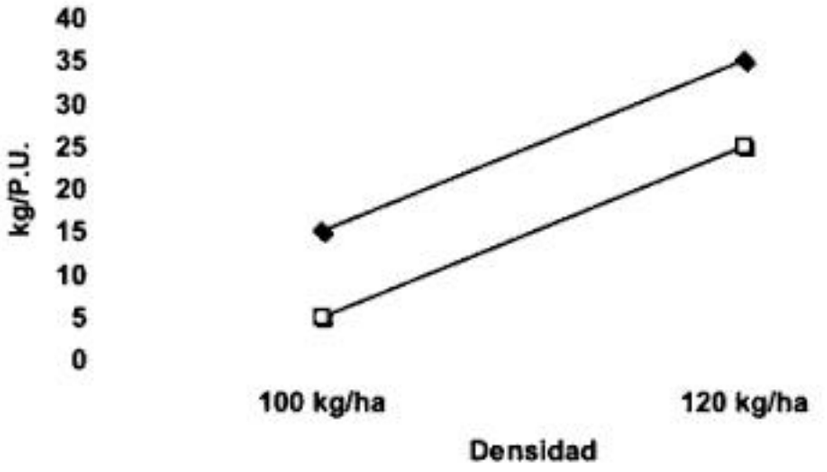


Figura 7.1. Ilustración de los efectos aditivos de dos factores, o los factores son independientes.

Cuando la diferencia de los efectos simple es cero (o puede estimar a cero), se dice que los efectos de los dos factores son aditivos o los factores son independientes; las líneas de tendencias son paralelas o tienden al paralelismo.

Caso B: **Efecto de los factores es interactivo.**

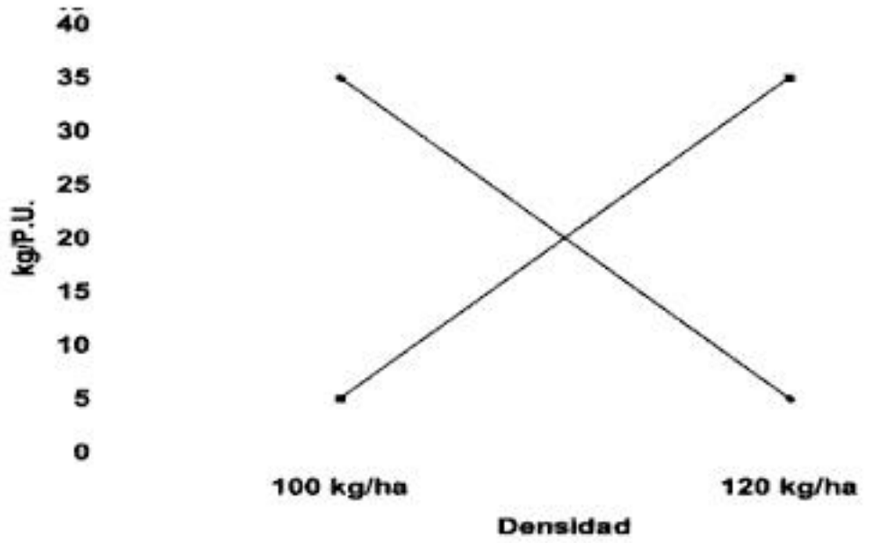


Figura 7.2. Ilustración de efectos interactivos de dos factores, o los factores no son independientes.

Cuando la diferencia de los efectos simples no es cero, se dice que el efecto de los factores es interactivo (o multiplicativo) y las respuestas de tendencias se cruzan o tienden a cruzarse.

Copyrighted material

Caso C: **Los datos sugieren efectos Interactivos.**

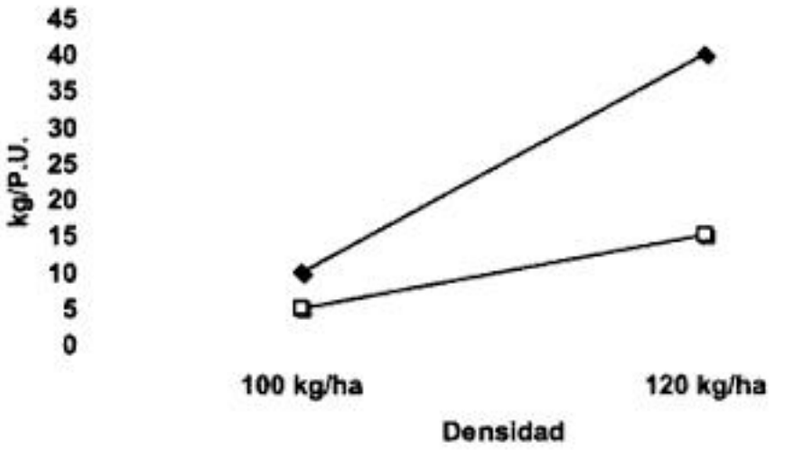


Figura 7.3. ***Ilustración de los efectos interactivos sugeridos por los datos.***

Cuando la diferencia de los efectos simple no es cero, se dice que los efectos de los dos factores son interactivos (o multiplicativos). En este caso, las líneas de tendencias no son paralelas, sino que tienden a cruzarse. (Reyes, C., 1982).

7.3 Proceso de Azarización de los Tratamientos.

Debido a que los experimentos factoriales propiamente dicho no son un diseño en sí, los tratamientos se asignan de acuerdo al proceso de azarización del diseño a establecer (D.C.A., B.C.A., D.C.L.)

7.4 El Modelo Aditivo Lineal para un bifactorial distribuido en D.C.A.

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$$

i = 1, 2, 3, ….a = niveles del factor A.
j = 1, 2, 3, …. b = niveles del factor B.
k= 1, 2, 3, ….n = observaciones.

Yij k = La k-ésima observación del i-ésimo tratamiento.
μ = Estima a la media poblacional.
∝i = Efecto del i-ésimo nivel del factor A.
βj = Efecto debido al j-ésimo nivel del factor B.
(∝ β) i j= Efecto de interacción entre los factores A y B.
Ɛi jk = Efecto aleatorio de variación.

Copyrighted material

7.5 Procedimiento estadístico para un experimento Bifactorial establecido en D.C.A.

Para ejemplificar el análisis correspondiente a un bifactorial establecido en DCA, se presentan los datos de un ensayo establecido con el objetivo de evaluar la fijación biológica del nitrógeno, inoculando tres variedades de frijol común, con tres diferentes cepas de ***Rhizobium***, usando N-15. El Experimento fue establecido en condiciones de invernadero y diseñado con el propósito de evaluar ambos factores con el mismo grado de precisión. En el cuadro 7.3., se presentan los tratamientos en estudio y los datos obtenidos del ensayo. Como un estudio de caso, ver Pedroza, P.H., (1993), en páginas 132-152.

Cuadro 7.3. Datos del Nitrógeno total (en mg) de la parte aérea de la planta.

Tratamientos	Observaciones			Totales
	1	2	3	$Y_{ij.}$
a_1b_1	85.25	98.49	90.37	274.11
a_1b_2	114.40	104.86	69.07	288.33
a_1b_3	73.90	70.91	65.12	209.93
a_1b_4	104.31	84.32	102.83	291.46
a_2b_1	85.06	82.08	101.96	269.10
a_2b_2	88.24	96.16	107.89	292.29
a_2b_3	97.87	71.25	92.19	261.31
a_2b_4	65.88	88.15	76.77	230.80
a_3b_1	152.20	197.06	175.82	525.08
a_3b_2	169.65	169.49	133.96	473.10
a_3b_3	124.34	178.43	150.14	452.91
a_3b_4	200.30	181.74	213.79	595.83

La descripción de los factores en estudio es la siguiente:

Factor A: Variedad

a_1: Rev-79
a_2: Rev-84
a_3: IMBAYO (de origen ecuatoriana)

Factor B: Cepas de Rhizobium

b_1: Cepa 1 (Ecuatoriana)
b_2: Cepa UMR - 1073
b_3: Cepa UMR - 1077
b_4: Cepa UMR - 1899

Con los datos presentados en el cuadro 7.3, se genera en SPSS la BDD llamada ***BIFACT en DCA*** que contiene cuatro variables: 1ra) "Variedad (Factor A)", con valores de 1 a 3; 2da) "Cepas (Factor B)", con valores de 1 a 4; 3ra) "Observaciones" o repeticiones estadísticas, con valores de 1 a 3; y 4ta) "Nittotal", con los datos del Nitrógeno Total (en mg) de la parte aérea de la planta, obtenido para cada tratamiento factorial en cada observación.

Copyrighted material

Para resolver en el **SPSS** el análisis estadísticos de un Bifactorial en DCA, se deben usar los comandos **Analyze/General Linear Model/ Univariate/** en **Dependent variable, se debe cargar la variable dependiente** – *Nitrógeno Total (en mg) de la parte aérea de la planta* -; y en **Fixed Factor(s)** se deben cargar las variables ***"Factor A-Variedad" y "Factor B-Cepas"***. Luego, dentro del comando **Model**, se deben definir los efectos principales y la interacción del modelo; usando la ventana de diálogo **Custom (personalizado),** se construyen los términos del modelo, incorporando una variable a la vez, se incluyen los factores A y B; luego para definir la interacción se toman simultáneamente el factor A y B, y se "jalan" con el botón de **"Build Term(s)"** hacia el cuadro derecho de la ventana de diálogo; la opción **"Type III" e "include intercept in model"** se dejan por defecto. Usando la ventana de diálogo **Options**, se le solicita al programa las tablas de medias para cada factor y la interacción; también puede solicitarse en esta ventana la prueba de Levene. En la ventana de diálogo **Plots** se le solicita el gráfico para la interacción. En la ventana de diálogo **Post Hoc** se selecciona la prueba de separación de medias para cada factor, en este ejemplo se utilizó la prueba de **S N K**.

Cuadro 7.4. Salida del ANOVA para un Bifactorial en DCA.

Tests of Between-Subjects Effects

Dependent Variable: Nitrógeno Total (en mg) de la parte aérea de la planta

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	60427.265[a]	11	5493.388	20.535	.000
Intercept	481693.8	1	481693.8	1800.658	.000
VARIEDAD	54262.665	2	27131.333	101.422	.000
CEPAS	2278.146	3	759.382	2.839	.059
VARIEDAD * CEPAS	3886.454	6	647.742	2.421	.057
Error	6420.238	24	267.510		
Total	548541.3	36			
Corrected Total	66847.503	35			

a. R Squared = .904 (Adjusted R Squared = .860)

Para analizar el cuadro de salida del bifactorial en DCA, se debe observar la significación del valor "F", para "Variedad", "Cepas" y la "Interacción". En este caso, la interpretación es la siguiente:

- La Significancia de "Variedad" es 0.000 < 0.05, por tanto se rechaza la Ho de igualdad entre variedades, esto indica que el efecto de las variedades tiene diferencias significativas entre ellas.

- La Significancia de "Cepas" es 0.05 < = 0.05, por tanto se rechaza la Ho de igualdad entre Cepas, esto indica que para el efecto de las cepas existen diferencias significativas entre si, lo que indica que "**al menos una de las cepas tiene un efecto promedio diferente**".

- La Significancia de interacción "Variedad*Cepas" es 0.05 < = 0.05, por tanto se rechaza la Ho de igualdad para la interacción, esto indica que existen diferencias significativas del efecto de interacción, por tanto "**al menos una de las combinaciones Variedad*Cepas tiene un efecto diferente**".

Copyrighted material

El siguiente paso es determinar cuales son los tratamientos que difieren entre si, para esto se utiliza la Técnica de Separación de Medias. La prueba solicitada de Rangos Múltiples de **SNK**, para el factor A y B, se presentan según la salida del SPSS en el siguiente cuadro.

Cuadro 7.5. Salida del SPSS para la separación de medias de SNK para el factor A.

Nitrógeno Total (en mg) de la parte aérea de la planta

Student-Newman-Keuls[a,b]

FACTOR A	N	Subset 1	Subset 2
2	12	87.7917	
1	12	88.6525	
3	12		170.5767
Sig.		.898	1.000

Means for groups in homogeneous subsets are displayed.
Based on Type III Sum of Squares
The error term is Mean Square(Error) = 267.510.
a. Uses Harmonic Mean Sample Size = 12.000.
b. Alpha = .05.

El cuadro de salida dado para la prueba de SNK del factor A, se presenta de la siguiente manera:

Cuadro 7.6. Presentación de medias del factor A y su significación estadística dada por la prueba de SNK.

Factor A	Promedios	Significancia estadística
a_3: IMBAYO	170.57	**a**
a_1: Rev-79	88.65	**b**
a_2: Rev-84	87.79	**b**

Nota: Letras iguales, indica promedios iguales, según prueba de SNK al 5%.

En el cuadro 7.6, se observan dos categorías estadísticas, a saber: la variedad Imbayo en primer lugar; seguida por las variedades Rev-79 y Rev-84 en segundo lugar.

Copyrighted material

Cuadro 7.7. ***Salida del SPSS para la separación de medias de SNK para el factor B.***

Nitrógeno Total (en mg) de la parte aérea de la planta

Student-Newman-Keuls[a,b]

FACTOR B	N	Subset 1	Subset 2
3	9	102.6833	
2	9	117.0800	117.0800
1	9	118.6989	118.6989
4	9		124.2322
Sig.		.116	.628

Means for groups in homogeneous subsets are displayed. Based on Type III Sum of Squares. The error term is Mean Square(Error) = 267.510.

a. Uses Harmonic Mean Sample Size = 9.000.

b. Alpha = .05.

El cuadro de salida dado para la prueba de SNK del factor B, se presenta de la siguiente manera:

Cuadro 7.8. ***Presentación de medias del factor B y su significación estadística dada por la prueba de SNK.***

Factor B	Promedios	Significancia estadística
b_4: Cepa UMR-1899	124.23	**a**
b_1: Cepa 1 (Ecuatoriana)	118.69	**a b**
b_2: Cepa UMR – 1073	117.08	**a b**
b_3: Cepa UMR – 1077	102.68	**b**

Nota: Letras iguales, indica promedios iguales, según prueba de SNK al 5%.

En base a la salida dada por la prueba de SNK, se puede afirmar que las cepas se clasifican en tres categorías estadísticas: Categoría **"a"**, determinada por la cepa **b_4: Cepa UMR-1899**. La segunda categoría **"ab"**, está formada por las cepas **b_1: Cepa 1 (Ecuatoriana)** y **b_2: Cepa UMR – 1073**. La tercera categoría **"b"**, la constituye la cepa **b_3: Cepa UMR – 1077.**

Por otra parte, se pueden observar las medias y los intervalos de confianza; así como el grafico de interacción solicitados al SPSS, con lo cual se ilustra mucho mejor el efecto de los tratamientos.

Copyrighted material

Cuadro 7.9. Presentación de medias e intervalos de confianza para la interacción.

3. FACTOR B * FACTOR A

FACTOR B	FACTOR A	Mean	Std. Error	95% Confidence Interval	
				Lower Bound	Upper Bound
1	1	91.370	9.443	71.881	110.859
	2	89.700	9.443	70.211	109.189
	3	175.027	9.443	155.537	194.516
2	1	96.110	9.443	76.621	115.599
	2	97.430	9.443	77.941	116.919
	3	157.700	9.443	138.211	177.189
3	1	69.977	9.443	50.487	89.466
	2	87.103	9.443	67.614	106.593
	3	150.970	9.443	131.481	170.459
4	1	97.153	9.443	77.664	116.643
	2	76.933	9.443	57.444	96.423
	3	198.610	9.443	179.121	218.099

En la figura 7.4., se evidencia de manera gráfica el efecto de interacción entre los factores.

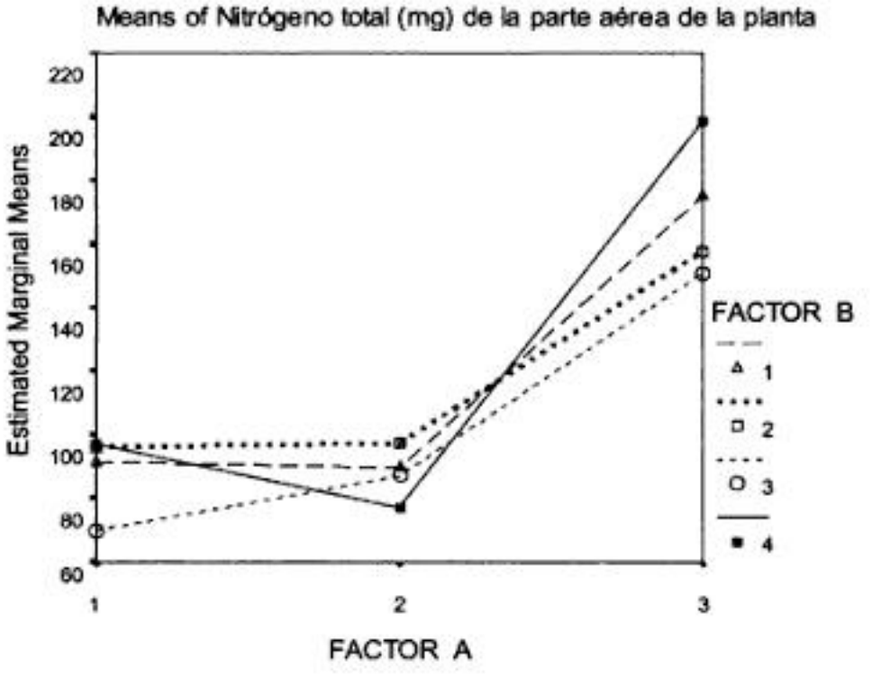


*Figura 7.4. Efecto de interacción entre Variedad*Cepas.*

El gráfico de **"error bar"**, se solicita aparte en SPSS, dentro del Módulo de **Graphs**.

Copyrighted material

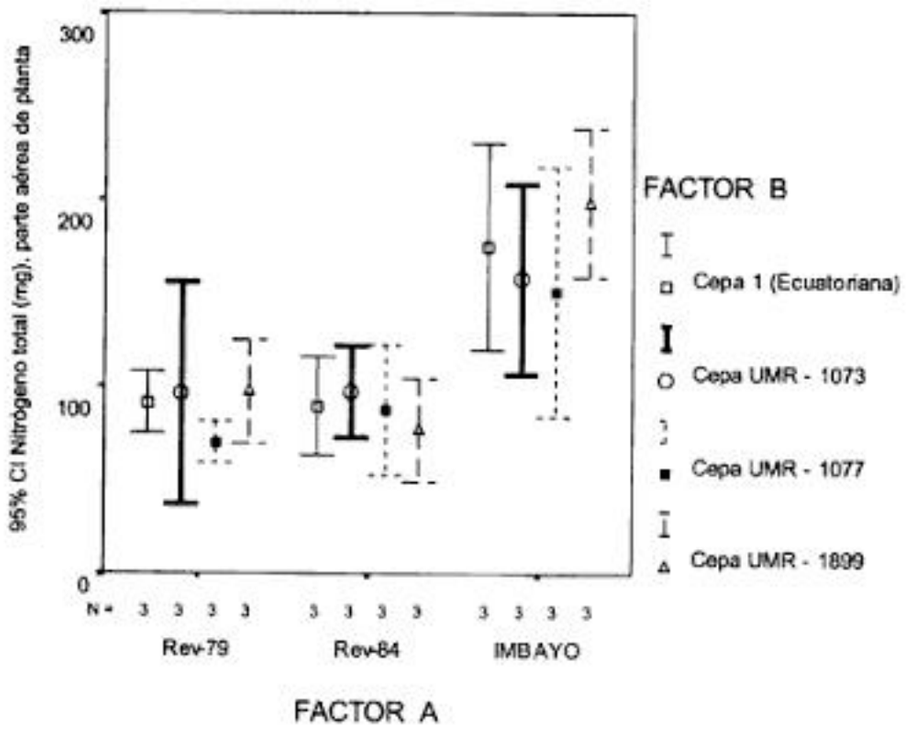


Figura 7.5. ***Gráfico de "error bar" para los tratamientos factoriales.***

Copyrighted material

Capítulo 8. Análisis de Varianza Univariado: Factoriales: Experimentos Bifactoriales establecidos en BCA.

8.1 El Análisis de Varianza para un Bifactorial en BCA.

Tal como se explicó en el capítulo anterior, los experimentos factoriales, **no** son un diseño en sí, sino un arreglo de tratamientos que se distribuyen en los diseños comunes: D.C.A., B.C.A y D.C.L. En este capítulo, abordaremos el caso de un experimento factorial establecido en BCA. De hecho, el bifactorial en BCA, es una extensión del MAL de un bifactorial en DCA, solamente que el modelo tiene un componente más que analizar, tal es el efecto de **"Bloques"**. Las implicaciones de campo de un bifactorial en BCA son muy marcadas en comparación con las del DCA, ya que en efecto, para el BCA debe garantizarse el agrupamiento de las unidades experimentales de forma tal, que el efecto del "bloqueo", sea efectivo en mejorar la precisión experimental de los datos. El bifactorial en BCA, igual que el DCA, permite estudiar por un lado ***los efectos principales,*** o acción independiente de los factores y por otro lado se estudia ***el efecto de interacción*** entre ellos.

8.2 El Modelo Aditivo Lineal para un Bifactorial distribuido en B.C.A.

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \rho_k + \varepsilon_{ijk}$$

i = 1, 2, 3, ...a = niveles del factor A.
j = 1, 2, 3, ... b = niveles del factor B.
k = 1, 2, 3, ...n = repeticiones ó bloques.

Yij k = La k-ésima observación del i-ésimo tratamiento.
μ = Estima a la media poblacional.
∝i = Efecto del i-ésimo nivel del factor A.
βj = Efecto debido al j-ésimo nivel del factor B.
(∝ β) i j= Efecto de interacción entre los factores A y B.
ρk = Efecto del k-ésimo bloque.
εi jk = Efecto aleatorio de variación.

8.3 Procedimiento estadístico para un experimento Bifactorial establecido en B.C.A.

Para ejemplificar el análisis de un bifactorial propiamente dicho establecido en BCA, se presentan los datos de un experimento de campo conducido para determinar el efecto de tres densidades de siembra y tres niveles de nitrógeno, sobre el rendimiento (en chilote), en el cultivo del maíz (Zea mays L). El experimento fue diseñado con el propósito de evaluar ambos factores con el mismo grado de precisión.

En el cuadro 8.1, se presentan los tratamientos en estudio y los datos obtenidos del ensayo. Como un estudio de caso, ver Pedroza, P.H., (1993), en páginas 153-159.

Copyrighted material

Cuadro 8.1. Datos del rendimiento total obtenido de Chilote (kg/P.U.).

Tratamientos	BLOQUES				Totales
	I	II	III	IV	$Y_{ij.}$
a_1b_1	4.15	7.90	5.50	3.50	21.05
a_1b_2	6.00	8.65	5.00	5.50	25.15
a_1b_3	8.25	8.95	8.60	8.40	34.20
a_2b_1	7.00	7.30	3.00	3.70	21.00
a_2b_2	7.35	7.70	4.70	5.10	24.85
a_2b_3	8.50	8.10	8.45	8.10	33.15
a_3b_1	5.70	8.90	11.10	5.50	31.20
a_3b_2	8.60	8.50	8.25	8.70	34.05
a_3b_3	9.85	9.30	8.80	8.40	36.35
$Y_{..k}$	65.40	75.30	63.40	56.90	261.00

La descripción de los factores en estudio es la siguiente:

Factor A: Densidad de Siembra

a_1: 136 000 plantas/ha
a_2: 90 750 plantas/ha
a_3: 68 600 plantas/ha

Factor B: Niveles de Nitrógeno

b_1: 50 kg/ha
b_2: 75 kg/ha
b_3: 100 kg/ha

Con los datos presentados en el cuadro 8.1., se genera en SPSS la BDD llamada ***BIFACT en BCA,*** que contiene cuatro variables: 1ra) "Densidad (Factor A)", con valores de 1 a 3; 2da) "Nivelden (Factor B)", con valores de 1 a 3; 3ra) "Bloques", con valores de 1 a 4; y 4ta) "Rendkg", con los datos de rendimiento total de Chilote, obtenido para cada tratamiento factorial en cada bloque.

Para resolver en el **SPSS** el análisis estadísticos de un Bifactorial en BCA, se deben usar los comandos **Analyze/ General Linear Model/ Univariate/** en **Dependent variable, se debe cargar la variable dependiente** – Rendimiento total de Chilote-; y en **Fixed Factor(s)** se deben cargar las variables ***"Densidad-Factor A", "Nivelden-Factor B" y "Bloques"***. Luego, dentro del comando **Model**, se deben definir los efectos principales y la interacción del modelo; usando la ventana de diálogo **Custom (personalizado),** se construyen los términos del modelo, incorporando una variable a la vez, se incluyen los factores A, B y Bloque; luego para definir la interacción se toman simultáneamente el factor A y B, y se "jalan" con el botón de **"Build Term(s)"** hacia el cuadro derecho de la ventana de diálogo; la opción **"Type III" e "include intercept in model",** se dejan por defecto. Usando la ventana de diálogo **Options**, se le solicita al programa las tablas de medias para cada factor y la interacción; también puede solicitarse en esta ventana la prueba de Levene. En la ventana de diálogo **Plots** se le solicita el gráfico para la interacción. En la ventana de diálogo **Post Hoc** se selecciona la prueba de separación de medias para cada factor, en el ejemplo se utilizó la prueba **S N K**.

Copyrighted material

Cuadro 8.2. ***Salida del ANOVA para un Bifactorial en BCA.***

Tests of Between-Subjects Effects

Dependent Variable: Rendimiento Total de Chilote (kg/P.U.)

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	91.321[a]	11	8.302	4.406	.001
Intercept	1892.250	1	1892.250	1004.312	.000
BLOQUES	19.352	3	6.451	3.424	.033
DENSIDAD	26.727	2	13.363	7.093	.004
NIVELDEN	39.721	2	19.861	10.541	.001
DENSIDAD * NIVELDEN	5.521	4	1.380	.733	.579
Error	45.219	24	1.884		
Total	2028.790	36			
Corrected Total	136.540	35			

a. R Squared = .669 (Adjusted R Squared = .517)

El análisis del cuadro de salida del bifactorial en BCA, debe referirse a la significación del valor "F", para "Bloques", "Densidad", "Nivel de Nitrógeno", y la Interacción. La interpretación es la siguiente:

- La Significancia de "Bloques" es 0.033 < 0.05, por tanto se rechaza la Ho de igualdad entre bloques, esto indica que el efecto de bloques ayudó a mejorar significativamente la precisión del experimento.

- La Significancia de "Densidad" es 0.004 < 0.05, por tanto se rechaza la Ho de igualdad entre densidades, esto indica que existen diferencias significativas entre las densidades.

- La Significancia de "Niveles de Nitrógeno" es 0.001 < 0.05, por tanto se rechaza la Ho de igualdad entre niveles de Nitrógeno, lo cual indica que existen diferencias significativas entre los diferentes niveles de Nitrógeno.

- La Significancia de interacción "Densidad*Niveles de Nitrógeno" es 0.579 > 0.05, por tanto se acepta la Ho de igualdad para la interacción, esto indica que **no** existen diferencias significativas del efecto de interacción.

El siguiente paso es determinar cuales son los tratamientos que difieren entre si, para esto se utiliza la Técnica de Separación de Medias. La prueba de Rangos Múltiples fue solicitada por medio de **SNK**, para el factor A y B. A continuación, se presentan las salidas del SPSS.

Copyrighted material

Cuadro 8.3. ***Salida del SPSS para la separación de medias de SNK para el factor A.***

Rendimiento Total de Chilote (kg/P.U.)

Student-Newman-Keuls[a,b]

Densidad de Siembra	N	Subset 1	Subset 2
a2: 90 750 plantas/ha	12	6.5833	
a1: 136 000 plantas/ha	12	6.7000	
a3: 68 600 plantas/ha	12		8.4667
Sig.		.837	1.000

Means for groups in homogeneous subsets are displayed.
Based on Type III Sum of Squares
The error term is Mean Square(Error) = 1.884.

a. Uses Harmonic Mean Sample Size = 12.000.

b. Alpha = .05.

El cuadro de salida dado para la prueba de SNK del factor A, se presenta de la siguiente manera:

Cuadro 8.4. Presentación de medias del factor A y su significación estadística dada por la prueba de SNK.

Factor A	Promedio	Significancia Estadística
a_3: 68600 plantas / ha	8.46	**a**
a_1: 136000 plantas / ha	6.70	**b**
a_2: 90750 plantas / ha	6.58	**b**

Nota: Letras iguales, indica promedios iguales, según prueba de SNK al 5%.

En el cuadro 8.4., se observan dos categorías estadísticas, a saber: la densidad de 68000 plantas/ha, en primer lugar; seguida por las densidades de siembra 136000 y 90750 plantas/ha, en segundo lugar.

Cuadro 8.5. ***Salida del SPSS para la separación de medias de SNK para el factor B.***

Rendimiento Total de Chilote (kg/P.U.)

Student-Newman-Keuls[a,b]

Niveles de Nitrogeno en Kg/Ha	N	Subset 1	Subset 2
b1: 50 kg/ha	12	6.1042	
b2: 75 kg/ha	12	7.0042	
b3: 100 kg/ha	12		8.6417
Sig.		.121	1.000

Means for groups in homogeneous subsets are displayed.
Based on Type III Sum of Squares
The error term is Mean Square(Error) = 1.884.

a. Uses Harmonic Mean Sample Size = 12.000.

b. Alpha = .05.

Copyrighted material

El cuadro de salida dado para la prueba de SNK del factor B, se presenta de la siguiente manera:

Cuadro 8.6. Presentación de medias del factor B y su significación estadística dada por la prueba de SNK.

Niveles de Nitrógeno	Promedio	Significancia estadística
b_3: 100 Kg/ha	8.64	**a**
b_2: 75 Kg/ha	7.00	**b**
b_1: 50 Kg/ha	6.10	**b**

Nota: Letras iguales, indica promedios iguales, según prueba de SNK al 5%.

En base a la salida dada por la prueba de SNK, se puede afirmar que los niveles de Nitrógeno se clasifican en dos categorías estadísticas: Categoría **"a"**, determinada por el nivel **b_3: 100 Kg/ha**. La segunda categoría **"b"**, está formada por los niveles **b_2: 75 Kg/ha** y **b_1: 50 Kg/ha**.

Por otra parte, se puede observar las medias y los intervalos de confianza; así como el grafico de interacción solicitados al SPSS, con lo cual se ilustra mucho mejor el efecto de los tratamientos.

Cuadro 8.7. Presentación de medias e intervalos de confianza para la interacción.

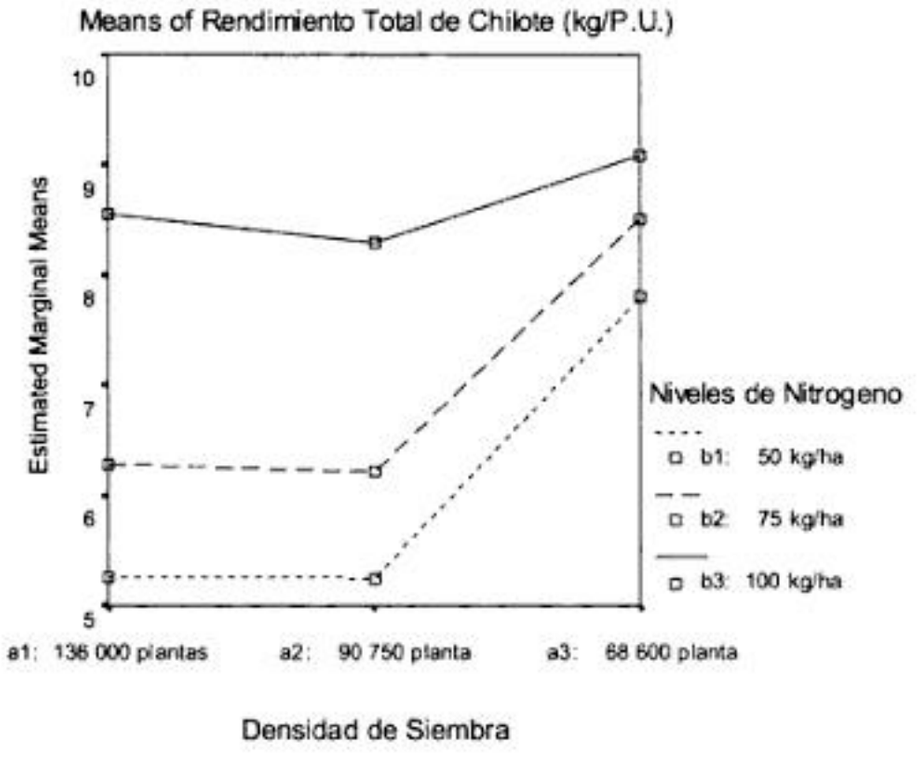
Densidad de Siembra * Niveles de Nitrogeno en Kg/Ha

Dependent Variable: Rendimiento Total de Chilote (kg/P.U.)

Densidad de Siembra	Niveles de Nitrogeno en Kg/Ha	Mean	Std. Error	95% Confidence Interval	
				Lower Bound	Upper Bound
a1: 136 000 plantas/ha	b1: 50 kg/ha	5.263	.686	3.846	6.679
	b2: 75 kg/ha	6.288	.686	4.871	7.704
	b3: 100 kg/ha	8.550	.686	7.134	9.966
a2: 90 750 plantas/ha	b1: 50 kg/ha	5.250	.686	3.834	6.666
	b2: 75 kg/ha	6.213	.686	4.796	7.629
	b3: 100 kg/ha	8.288	.686	6.871	9.704
a3: 68 600 plantas/ha	b1: 50 kg/ha	7.800	.686	6.384	9.216
	b2: 75 kg/ha	8.513	.686	7.096	9.929
	b3: 100 kg/ha	9.087	.686	7.671	10.504

Copyrighted material

Figura 8.1. ***Efecto Aditivo entre Densidad*Niveles de Nitrógeno***



En la figura 8.1., se evidencia el efecto no significativo de interacción o efecto aditivo entre los factores.

Con el gráfico de **"error bar"** se evidencia también el efecto aditivo entre factores. El gráfico de "error bar" es solicitado por aparte en SPSS, dentro del Módulo de **Graphs**.

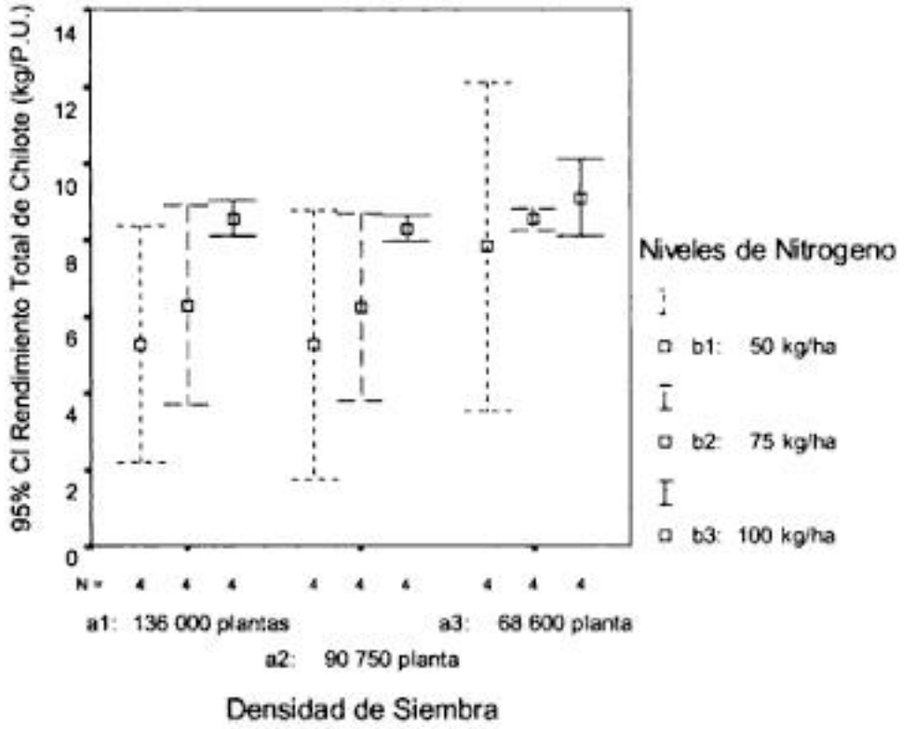


Figura 8.2. ***Gráfico de "error bar", de los tratamientos factoriales.***

Copyrighted material

8.4 El Análisis de Varianza para un Trifactorial en BCA.

Al agregar un factor más en estudio, se crea una extensión del modelo Bifactorial conocida como **Trifactorial en Bloques**, de hecho, el análisis estadístico varía muy poco. En el modelo de un trifactorial, se crean además de las interacciones de primer orden, (**AB, AC, BC**), también se crea la interacción de segundo orden en el modelo, (**ABC**).

La rutina dentro del SPSS, para el análisis estadístico de los datos de un experimento trifactorial, es prácticamente la misma que para un Bifactorial.

Copyrighted material

Capítulo 9. Análisis de Varianza Univariado: Factoriales: Diseño de Parcelas Divididas establecido en BCA.

9.1 El Análisis de Varianza para un Diseño de Parcelas Divididas en BCA.

El Diseño de Parcelas Divididas, es un diseño propiamente dicho basado en el principio básico de que las unidades experimentales, por razones de manejo de campo, tienen diferentes tamaños, de modo que a las parcelas grandes o parcelas principales, se les aplican mediante un primer proceso de azarización los niveles del primer factor (A). Luego, las parcelas grandes se sub-dividen en sub-parcelas o parcelas pequeñas, a las cuales se les aplican por medio de un segundo proceso de azarización, los niveles del segundo factor (B). De este modo, el Diseño de Parcelas Divididas estudia dos o más factores simultáneamente, pero uno de ellos (el factor B) se estudia con mayor precisión que el otro. El Parcelas Dividida es un diseño en si, porque los tratamientos tienen su propia azarización de manera muy diferente a los otros diseños; en él se estudia al igual que en otros factoriales, la acción independiente de los factores y el efecto de interacción entre ello.

Un análisis comparativo entre el Diseño de Parcelas Divididas y los bifactoriales propiamente dicho, indica que el Parcelas Divididas evalúa a los factores con diferente grado de precisión, es un diseño en si, utiliza dos diferentes tamaños de parcelas en el campo y en los procesos de azarización que utiliza se generan dos tipos de errores: E(a) y E(b). Por el contrario, los ensayos bifactoriales propiamente dicho, establecidos en DCA, BCA o DCL, evalúan los factores con el mismo grado de precisión, no constituye un diseño en si, utiliza en el campo un solo tamaño de parcela y su azarización genera un tipo de error.

Por otra parte, las situaciones prácticas en que normalmente se recomienda implementar un Parcelas Divididas, es por ejemplo cuando los niveles de un factor (A), requieran una mayor extensión o área para la unidad experimental que los niveles del otro factor (B). También es muy útil cuando se desea estudiar un factor con mayor precisión que el otro factor, siendo la norma que el factor establecido en la subparcela (factor B), es el que se estudia con mayor precisión, debido a que este factor se estudia con mayor número de repeticiones.

9.2 El Proceso de Azarización de Tratamientos en un Diseño de Parcelas Divididas.

Básicamente, se realiza en dos etapas: 1) Por cada bloque, se azarizan los niveles del factor A, lo que viene a constituir las parcelas grandes y este proceso genera el error del factor A, **(como una interacción entre el factor A y el bloque), con el cual se calcula el efecto del factor A y del bloque**. 2) Dentro de cada parcela grande, se azarizan los niveles del factor B, lo que viene a constituir las sub-parcelas y este proceso genera el error de B o residuo del modelo, con el cual se calcula el efecto del factor B y la interacción entre factores AB.

Copyrighted material

9.3 *El Modelo Aditivo Lineal para un Diseño de Parcelas Divididas.*

$$Yijk = \mu + \rho_k + \alpha_i + \varepsilon_{ik} + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$$

i = 1, 2, 3, ….a = niveles del factor A.
j = 1, 2, 3, …. b = niveles del factor B.
k = 1, 2, 3, ….n = repeticiones o bloques.

Yij k = La k-ésima observación del i-ésimo tratamiento.
μ = Estima a la media poblacional.
ρk = Efecto del k-ésimo bloque.
∝i = Efecto del i-ésimo nivel del factor A.
εi k = Error del Factor A, E(a).
βj = Efecto debido al j-ésimo nivel del factor B.
(∝β) i j = Efecto de interacción entre los factores A y B.
εi jk = Efecto aleatorio de variación o error del modelo E(b).

9.4 *Procedimiento estadístico para un Diseño de Parcelas Divididas en B.C.A.*

Para ejemplificar el análisis de un Diseño de Parcelas Divididas en BCA, se presentan los datos de un experimento de campo establecido con el objetivo de estudiar el efecto de sistemas de labranza (Factor A) y presencia o ausencia de malezas (Factor B), sobre la incidencia de la Chicharrita del Maíz (*Dalbulus maidis*). El experimento de campo fue establecido en el CNIA –Centro Nacional de Investigación Agropecuario-, utilizando la variedad de maíz NB-100. Se estableció un diseño de Parcelas Divididas con arreglos de parcelas grandes en B.C.A. En el cuadro 9.1., se presentan los tratamientos en estudio y los datos obtenidos del ensayo. Como un estudio de caso, ver Pedroza, P.H., (1993), páginas 160-178.

Cuadro 9.1. Datos del rendimiento de campo en kg/ha.

Factor A	Factor B	BLOQUES				
		I	II	III	IV	$Y_{ij.}$
a_1	b_1	1478.99	1268.23	1150.70	905.14	4803.06
	b_2	1304.30	1027.76	912.04	1013.78	4257.88
(Sub-Total) $Y_{i.k}$		2783.29	2295.99	2062.74	1918.92	9060.94
a_2	b_1	1877.33	1277.31	1534.20	1103.52	5792.36
	b_2	1891.54	1529.41	1545.54	1264.54	6231.03
(Sub-Total) $Y_{i.k}$		3768.87	2806.72	3079.74	2368.06	12023.39
(Total) $Y_{..k}$		6552.16	5102.71	5142.48	4286.98	21084.33

Copyrighted material

La descripción de los factores en estudio es la siguiente:

Factor A: Sistemas de labranzas

a_1: Labranza convencional
a_2: Labranza Cero

Factor B: Malezas

b_1: Sin malezas
b_2: Con malezas

Con los datos presentados en el cuadro 9.1, se genera en SPSS la BDD llamada ***PARCELA DIVIDIDA EN BCA,*** que contiene cuatro variables: 1ra) "Labranza (Factor A)", con valores de 1 a 2; 2da) "Maleza (Factor B)", con valores de 1 a 2; 3ra) "Bloques", con valores de 1 a 4; y 4ta) "Rendkg", con los datos de rendimiento de campo en Kg/ha, obtenido para cada tratamiento factorial.

Para resolver en el **SPSS** el análisis estadísticos de un Diseño de Parcelas Dividida en BCA, se deben usar los comandos **Analyze/General Linear Model/ Univariate/** en **Dependent variable, se debe cargar la variable dependiente** – Rendimiento de campo en Kg/ha-; y en **Fixed Factor(s)** se deben cargar las variables ***"Labranza-Factor A", "Maleza-Factor B" y "Bloques"***. Luego, dentro del comando **Model**, se deben definir los efectos principales y la interacción del modelo; usando la ventana de diálogo **Custom (personalizado),** se construyen los términos del modelo, incorporando **una variable a la vez en el orden que corresponde a las fuentes de variación de un Parcela Dividida**: se incluyen 1ro) Bloque; 2do) Factor A; 3ro) **Bloque*Factor A, esto define el E(a) en el modelo**. En 4to) orden se incluye el Factor B. Luego, para definir la interacción AB, en 5to) orden se toman simultáneamente el factor A y B, y se "jalan" con el botón de **"Build Term(s)"** hacia el cuadro derecho de la ventana de diálogo; la opción **"Type III" e "include intercept in model"**, se dejan por defecto.

El programa calcula por defecto el error del modelo ó **E(b)**. Usando la ventana de diálogo **Options**, se le solicita al programa las tablas de medias para cada factor y la interacción; también puede solicitarse en esta ventana la prueba de Levene; En la ventana de diálogo **Plots** se le solicita el gráfico para la interacción. En la ventana de diálogo **Post Hoc** se selecciona la prueba de separación de medias para cada factor, en este ejemplo se utilizó la prueba de **S N K**.

Cuadro 9.2. ***Salida del ANOVA dada por el SPSS, para un Diseño de Parcelas Dividida en BCA.***

Tests of Between-Subjects Effects

Dependent Variable: Rendimiento de campo en Kg/ha

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	1341890 [a]	9	149098.9	14.371	.002
Intercept	2.8E+07	1	2.8E+07	2677.944	.000
BLOQUES	663626.3	3	221208.8	21.321	.001
LABRANZA	548506.9	1	548506.9	52.867	.000
LABRANZA * BLOQUES	68550.327	3	22850.109	2.202	.189
MALEZAS	709.024	1	709.024	.068	.803
LABRANZA * MALEZAS	60497.551	1	60497.551	5.831	.052
Error	62251.441	6	10375.240		
Total	2.9E+07	16			
Corrected Total	1404142	15			

a. R Squared = .956 (Adjusted R Squared = .889)

Copyrighted material

La salida del SPSS para un Parcela Divididas proporciona los cuadrados medios (Mean Square) para cada uno de los términos del Modelo; sin embargo, el calculo de F para el Bloque y el Factor A, lo realiza contra el E(b). Por esta razón en particular, debe calcularse por aparte el valor de F para el Bloque y el Factor A, haciendo la relación de varianzas entre el cuadrado medio de bloque y Factor A, contra el E(a), definido en el cuadro 9.2 como LABRANZA*BLOQUE=22850.109. Luego se puede proceder a la correcta interpretación del efecto de Bloque y Factor A, tal como se presentan en el cuadro siguiente.

Cuadro 9.3. ***Tabla del ANOVA para un Diseño de Parcelas Dividida en BCA, con el valor de F_ para Bloque y el Factor A, calculados con el E(a).***

Tests of Between-Subjects Effects

Dependent Variable: Rendimiento de campo en Kg/ha

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	1341890.076(a)	9	149098.897	14.371	.002
Intercept	27784310.722	1	27784310.722	2677.944	.000
BLOQUES	663626.298	3	221208.766	9.68 *	.001
LABRANZA	548506.875	1	548506.875	24.00 *	.000
LABRANZA * BLOQUES	68550.327	3	22850.109		.189
MALEZAS	709.024	1	709.024	0.068 NS	.803
LABRANZA * MALEZAS	60497.551	1	60497.551	5.83 *	.052
Error	62251.441	6	10375.240		
Total	29188452.239	16			
Corrected Total	1404141.517	15			

a R Squared = .956 (Adjusted R Squared = .889)

El análisis del cuadro 9.3 para el Parcelas Divididas en BCA, debe referirse a la significación del valor "F", para "Bloques", "Labranza", "Maleza", y la Interacción. La interpretación es la siguiente:

- La significancia para "Bloques" es 0.001<0.05, esto indica que el efecto de bloques es significativo y por tanto, si ayudó a mejorar la precisión del experimento.

- La significancia para "Labranza" 0.000<0.05, esto indica que existen diferencias significativas entre los niveles de labranza.

- La significancia de "Malezas" es 0.803 > 0.05, por tanto se acepta la Ho de igualdad entre niveles de Malezas.

- La significancia de interacción "Labranza*Malezas" es 0.05 < = 0.05, por tanto se rechaza la Ho de igualdad para la interacción, esto indica que si existen diferencias significativas del efecto de interacción.

Copyrighted material

El siguiente paso es determinar cuales son los tratamientos que difieren entre si, para esto se utiliza la Técnica de Separación de Medias. La prueba de Rangos Múltiples fue solicitada por medio de **SNK**, para el factor A y B. No obstante, el SPSS envía un aviso señalando que hay menos de tres niveles para cada factor y que por tanto, no se puede realizar la prueba. Esto indica que se necesitan al menos tres niveles para cada factor para realizar la separación de medias. A continuación, se presentan las medias de factores e interacción, también solicitadas al SPSS.

Cuadro 9.4. Cuadro de medias para el factor labranza.

1. Labranza (Factor A)

Dependent Variable: Rendimiento de campo en Kg/ha

Labranza (Factor A)	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
a1: Labranza convencional	1132.617	36.013	1044.498	1220.737
a2: Labranza cero	1502.924	36.013	1414.804	1591.043

En el cuadro 9.4., se observa el incremento de medias para el nivel a_2, mayor que a_1.

Cuadro 9.5. Cuadro de de medias del factor malezas.

2. Malezas (Factor B)

Dependent Variable: Rendimiento de campo en Kg/ha

Malezas (Factor B)	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
b1: Sin maleza	1324.428	36.013	1236.308	1412.547
b2: Con maleza	1311.114	36.013	1222.994	1399.233

Hasta aquí dos detalles hay que destacar: El **R** ajustado = 0.889, presentado en el cuadro 9.2., indica que el ajuste del modelo de Parcelas Dividida en BCA ha sido muy adecuado para el análisis de los datos. Pero, el hecho de tener solo dos niveles para cada factor en estudio, induce a obtener una respuesta imperfecta **no** apropiada para cada uno de los factores, limitando la interpretación del efecto de los factores presentado en la figura 9.2. Por otra parte, se puede observar las medias y los intervalos de confianza; así como el grafico para cada factor y la interacción, solicitados al SPSS, con lo cual se ilustra mucho mejor el efecto de interacción entre los factores.

Copyrighted material

Cuadro 9.6. Presentación de medias e intervalos de confianza para la interacción.

3. Labranza (Factor A) * Malezas (Factor B)

Labranza (Factor A)	Malezas (Factor B)	Mean	Std. Error	95% Confidence Interval	
				Lower Bound	Upper Bound
a1: Labranza convencional	b1: Sin maleza	1200.765	50.929	1076.145	1325.385
	b2: Con maleza	1064.470	50.929	939.850	1189.090
a2: Labranza cero	b1: Sin maleza	1448.090	50.929	1323.470	1572.710
	b2: Con maleza	1557.758	50.929	1433.138	1682.377

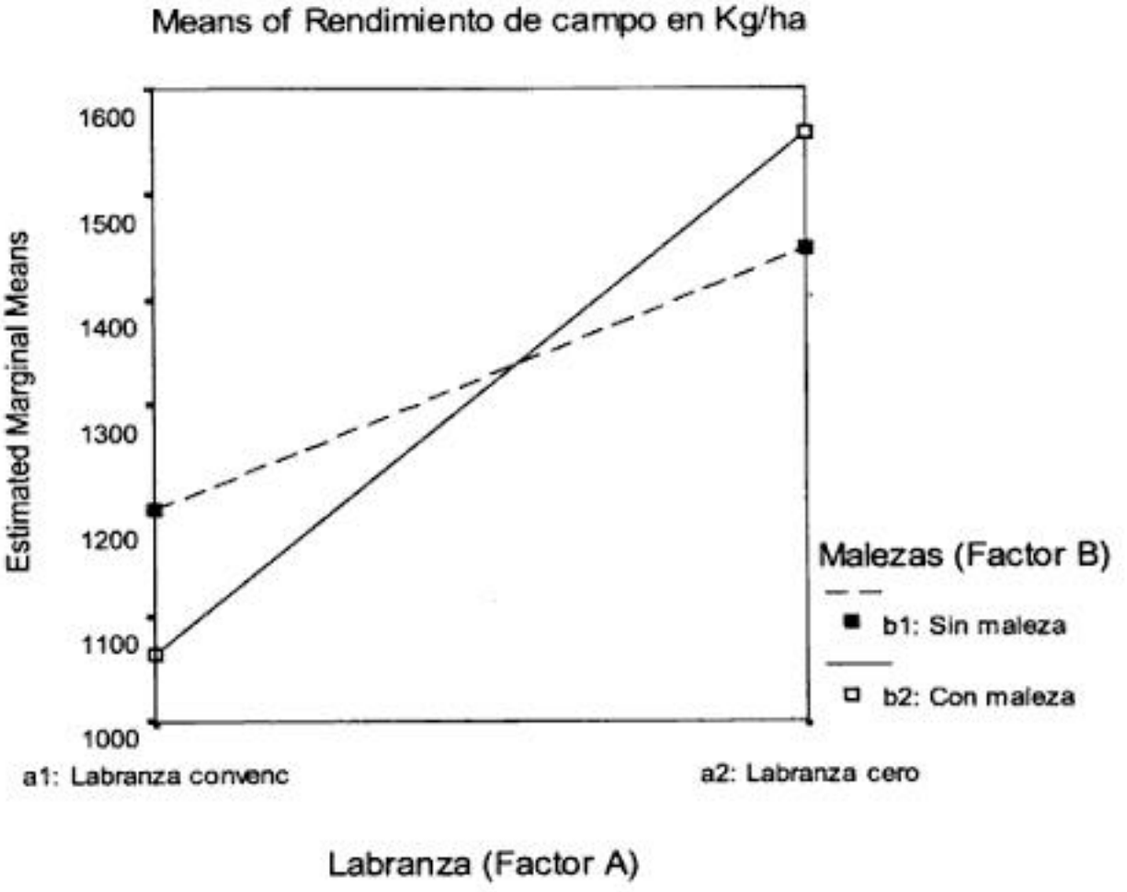


*Figura 9.1. Efecto de interacción Labranza*Malezas.*

Copyrighted material

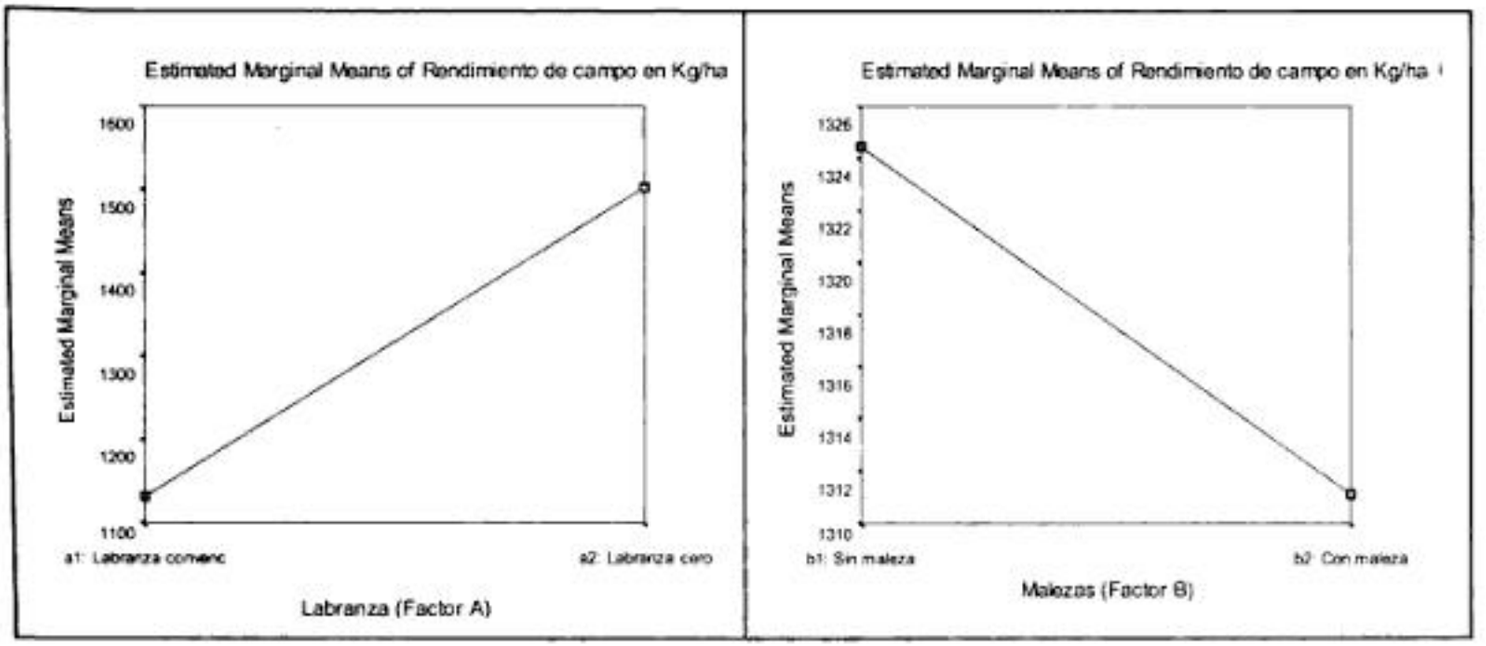


Figura 9.2. ***Efecto del factor Labranza y Malezas por separado.***

En la figura 9.1., se evidencia el efecto significativo de interacción Labranza*Malezas. Por otra parte, con el gráfico de **"error bar"** se evidencia también el efecto de interacción. El gráfico de **"error bar"** es solicitado aparte en SPSS, dentro del Módulo de **Graphs**.

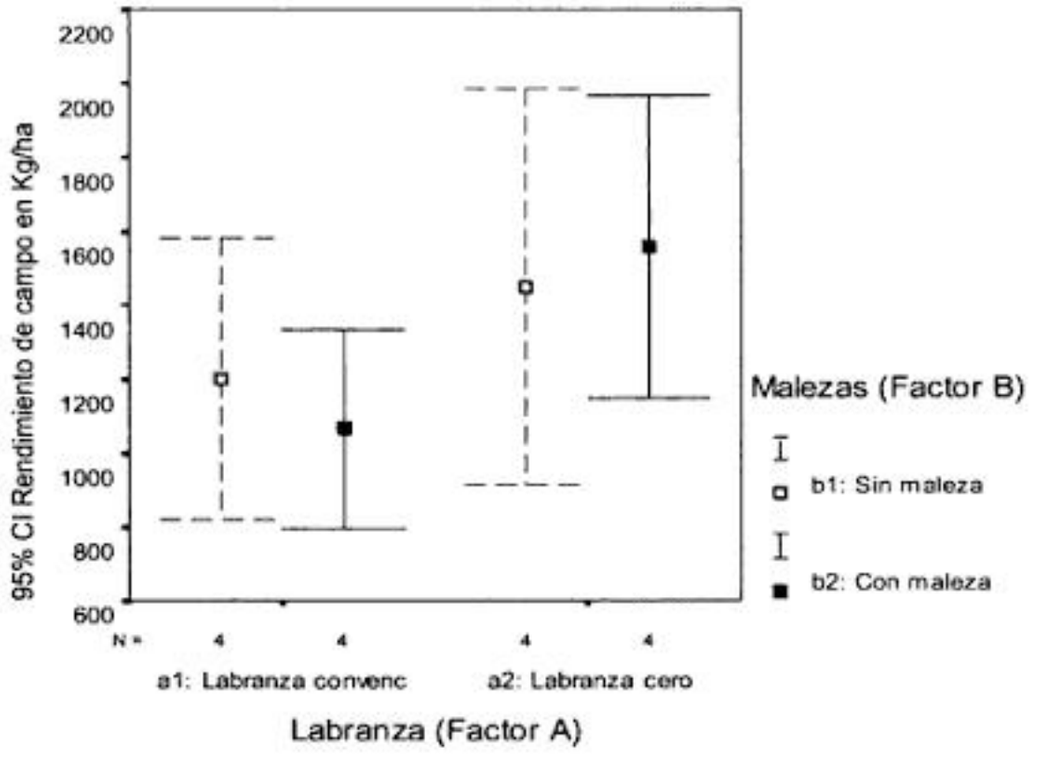


Figura 9.3. ***Gráfico de "error bar", de los tratamientos factoriales.***

Copyrighted material

Capítulo 10. Análisis de Regresión Lineal Simple.

10.1 El Análisis de Regresión Lineal Simple.

El concepto de regresión se refiere al "cuantum" o "cantidad de cambio" que experimenta una variable dependiente (Y), en relación al cambio de una unidad de una variable independiente (X). La regresión es un concepto estadístico estrechamente vinculado al concepto de correlación; mientras la regresión estudia la naturaleza de la relación entre dos variables dependientes, la correlación estudia la estrechez de la relación entre esas dos variables una dependiente de la otra, (Little y Hills, 1978). Al igual que con otros procedimientos estadísticos, (Dicovskyi L., 2002), destaca que en la regresión lineal se desea realizar una inferencia estadística partiendo de los valores muestrales obtenidos; por tanto, se deben cumplir ciertos requisitos, que en el caso de la regresión lineal son los siguientes:

1) **Normalidad y Homogeneidad de varianzas en la variable dependiente (Y) del modelo para los valores fijos de la variable independiente (X).**
2) **Independencia de las observaciones de Y**
3) **Linealidad en la relación entre las variables.**

El modelo de regresión simple es el siguiente $\mathbf{Y_i = B_0 + B_1 X_i + e_i}$ ..... donde:

$\mathbf{Y_i}$: es la variable dependiente.
$\mathbf{B_0}$: es la ordenada en el origen, o bien es el intercepto.
$\mathbf{B_1}$: es la pendiente de la recta de regresión.
$\mathbf{e_i}$: es el término del error, es decir la diferencia entre los valores predichos por la regresión y los valores reales.

Para desarrollar el tema del Análisis de Regresión Lineal, se toma como ejemplo parte de los datos del experimento de tomate, referido en el capítulo once. La variable dependiente es "Peso fresco de planta" (en gramos) y la variable independiente es "Desarrollo de la Planta", una variable cualitativa, codificada en escala cuantitativa discreta con cinco valores (1, 2, 3, 4 y 5), donde 1 es el peor estado, y 5 es muy bueno.

La pregunta lógica es: ¿ Si el peso fresco de planta está determinado o no por el estado de desarrollo de la planta, y si el efecto de regresión es significativo, entonces, en que medida el desarrollo de la planta induce a obtener un mayor peso fresco de la planta?. Este tipo de pregunta la puede responder un estudio de Regresión y Correlación. Para realizar el ANARE, es recomendable iniciar con el uso del comando **Descriptives**, a fin de conocer las características básicas de las variables en estudio.

Copyrighted material

Cuadro 10.1. ***Análisis descriptivo de las variables en estudio.***

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
PESOFRES	100	.50	7.30	3.1220	1.7544
DESARROL	100	1	5	3.63	1.12
Valid N (listwise)	100				

10.2 Rutina para el Análisis de Regresión Simple con el SPSS.

Primero, se carga la base de datos llamada **"EXPERIMENTO-SUSTRATOXFERTIRIEGO"**, luego la rutina para el ANARE se realiza en el SPSS, ejecutándo el comando **Analyze / Regression / Linear**. Siguiendo las indicaciones dadas en la ventana de diálogo, se introduce la variable "peso fresco de plántula" como dependiente, y luego se introduce la variable "desarrollo de la planta" como la variable independiente. Para desarrollar el análisis de regresión lineal simple, puede seleccionarse el método **Enter**, luego en la ventana de diálogo **statistics**.... se selecciona la prueba de ***Durbin-Watson***, y el ajuste del modelo (***model fit***). También se puede seleccionar en la venta de diálogo **plots**, el ***histograma y el gráfico de probabilidad normal***. En la ventana de **options**, se recomienda dejar por defecto la probabilidad de entrada de alfa 5%. Finalmente se ejecuta el procedimiento de la regresión lineal simple, dando **OK**. En el siguiente cuadro, se presenta la primera salida del ANARE.

Cuadro 10.2. ***Resumen de los coeficientes de Correlación de Pearson (R) y Determinación (R^2).***

Model Summary [b]

Model	R	R Square	Adjusted	Std. Error of the Estimate	Durbin-Watson
1	.749 [a]	.561	.557	1.1680	1.601

a. Predictors: (Constant), DESARROL

b. Dependent Variable: PESOFRES

En el cuadro 10.2., puede observarse el coeficiente de correlación de Pearson, (**R**), que mide el grado de asociación entre las variables X e Y. En ciencias biológicas, es aceptable un **R** cercano a un **80 %**. El valor obtenido para **R = 0.749**, es un valor de correlación alto y positivo, lo que indica una alta dependencia de la variable dependiente en función de la variable independiente "desarrollo de la planta". El coeficiente de determinación ($\mathbf{R^2}$) llamado en el cuadro 10.2 como **R Square**, es una medida -en porcentaje-, de la influencia en que la variable independiente (en este caso "desarrollo de la planta"), ***determina a la variable dependiente*** (en este caso "peso fresco de planta"). El valor obtenido de $\mathbf{R^2 = 0.561}$, indica que el 56 % de la variabilidad del "peso fresco de planta", se debe o se explica por la influencia de la variable "desarrollo de la planta".

El valor de $\mathbf{R^2}$ **ajustado**, tiene mayor relevancia y debe ser considerado en los casos de regresión lineal múltiple, ya que existe la tendencia a sobreestimar el valor de **R**, a medida que aumenta el número de variables independientes incorporadas en el modelo. No obstante, para el caso de la regresión lineal simple

puede ser que el ajuste sea insignificante, por cuanto solo existe una variable independiente incorporada en el modelo de regresión; de hecho, el valor obtenido de **R^2 y R^2 ajustado**, en este ejemplo es igual a 0.56, por lo que se puede utilizar cualquiera de ellos.

En el cuadro 10.2., se presenta la prueba de "*Independencia de los Residuos*", por el estadístico de Durbin-Watson = 1.601. El valor de Durbin-Watson aproximado a 2, indica que se cumple el principio de que los términos de los residuos **no** están correlacionados entre si. Por el contrario, si el estadístico Durbin-Watson se aproxima a 4, significa que los residuos estarán negativamente autocorrelacionados entre si. Finalmente, si el estadístico Durbin-Watson se aproxima 0, significa que los términos del error estarán positivamente autocorrelacionados, (Ferran A. M., 1996).

10.3 Construyendo el Modelo de Regresión Lineal Simple.

Para la regresión lineal simple, es fundamental determinar la significancia estadística del efecto de regresión en estudio, esto se determina en el cuadro del ANARE o ANOVA.

Cuadro 10.3. Análisis de Regresión de las variables en estudio.

ANOVA [b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	171.009	1	171.009	125.344	.000 [a]
	Residual	133.703	98	1.364		
	Total	304.712	99			

a. Predictors: (Constant), DESARROL

b. Dependent Variable: PESOFRES

En el cuadro 10.3., se observa la prueba de F para evaluar el efecto de regresión lineal. La Significancia observada = 0.000, es menor del 0.05, por tanto se rechaza la hipótesis nula de que el valor de Beta es igual a 0; es decir, se acepta que el efecto de regresión de la variable independiente "desarrollo de la planta" **es significativo** sobre la variable dependiente "peso fresco de planta". Queda establecido, el modelo de regresión lineal en función de una constante, más la influencia de la variable "desarrollo de la planta".

Otra salida importante para el ANARE lineal simple, se presenta en el cuadro 10.4.

Cuadro 10.4. Coeficientes Beta para construir el modelo de regresión.

Coefficients [a]

Model		Unstandardized		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	-1.153	.399		-2.887	.005
	DESARROL	1.178	.105	.749	11.196	.000

a. Dependent Variable: PESOFRES

Copyrighted material

En este ejemplo, el modelo queda definido por un **intercepto** de la función lineal, que es $\mathbf{B_0 = -1.153;}$ y el **coeficiente de regresión** es $\mathbf{B_1 = 1.178.}$ Se confirma la significancia de la regresión lineal por medio de la prueba de **"t"**, ya que el coeficiente estandarizado de Beta = 0.749, tiene un valor de Significancia 0.000 menor a 0.05. El modelo de regresión lineal simple queda dado por la ecuación....

$$\mathbf{Y_i = B_0 + B_1 X_{1i} + e_i}$$

Queda definido para este estudio, por los siguientes términos:

$$\mathbf{Y_i = -1.153 + 1.178 X_i + e_i}$$

El modelo de regresión lineal simple explica parcialmente comportamiento de los datos. Esto se observar mejor con un gráfico de dispersión, con el cual se pueden observar los puntos X-Y en forma de nube y la recta de regresión en ella. Para lograr este gráfico, se utiliza el comando **Analize/Regression/Curve Estimation/Linear**:

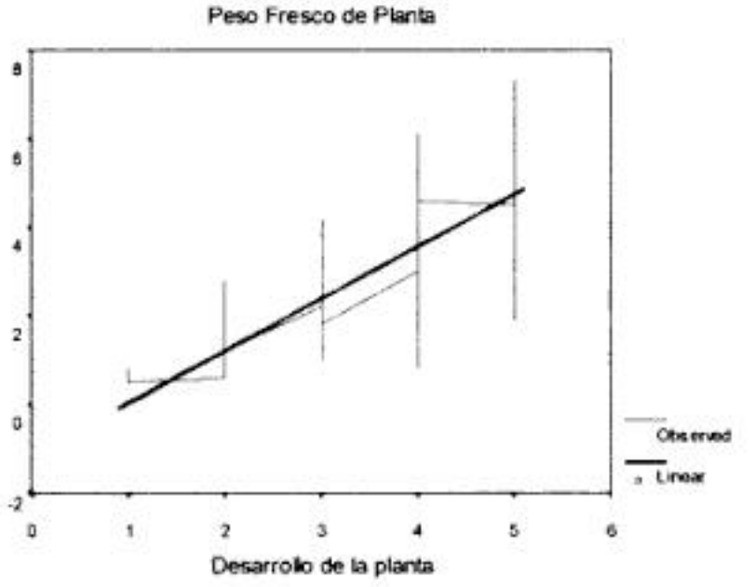


Figura 10.1. ***Gráfico de dispersión para la regresión lineal.***

10.4 Determinando el Modelo de Mejor Ajuste.

Este mismo comando **Analyze/Regression/Curve Estimation**, es importante utilizar en la simulación de modelos para determinar la curva de mejor ajuste, **ya que existen casos para los cuales el modelo lineal podría no ser el de mejor ajuste.** Para lograr la simulación, en la ventana de diálogo de este mismo comando, se solicitan diferentes modelos: ***Lineal, logarítmico, cuadrático, cúbico, exponencial,*** etc. Finalmente, basados en la prueba de F que proporciona el ANOVA y tomando muy en cuenta el $\mathbf{R^2}$, se procederá a seleccionar el modelo de mejor ajuste, que corresponderá al modelo que tenga mayor valor del $\mathbf{R^2}$. A continuación se presenta el gráfico y la hoja de salida solicitada:

Copyrighted material

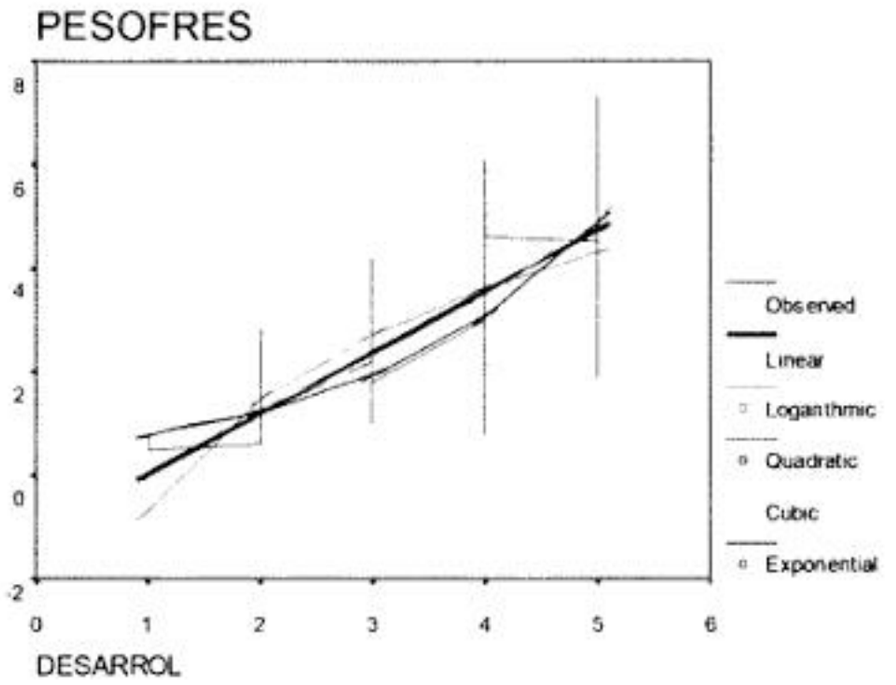


Figura 10.2. ***Gráfico de simulación de modelos para determinar la curva de mejor ajuste.***

```
MODEL:  MOD_1.
Dependent variable.. PESOFRES          Method.. LINEAR

Multiple R            .74914
R Square              .56122
Adjusted R Square     .55674
Standard Error       1.16804

            Analysis of Variance:
                 DF    Sum of Squares      Mean Square
Regression        1         171.00889        171.00889
Residuals        98         133.70271          1.36431
F =     125.34428       Signif F =  .0000
-------------------- Variables in the Equation --------------------
Variable                    B         SE B       Beta        T       Sig T
DESARROL             1.177634      .105186    .749143    11.196   .0000
(Constant)          -1.152810      .399291   -2.887               .0048
```

```
Dependent variable.. PESOFRES          Method.. LOGARITH

Multiple R            .68862
R Square              .47420
Adjusted R Square     .46884
Standard Error       1.27862
```

Copyrighted material

```
            Analysis of Variance:
                 DF    Sum of Squares      Mean Square
Regression        1         144.49467        144.49467
Residuals        98         160.21693          1.63487
F =       88.38315       Signif F =  .0000

-------------------- Variables in the Equation --------------------
Variable                    B         SE B       Beta        T     Sig T
DESARROL             3.088193     .328488     .688623     9.401   .0000
Constant)           -.665079     .422633                -1.574    .118
```

Dependent variable.. PESOFRES **Method.. QUADRATI**

```
Multiple R            .76349
R Square              .58292
Adjusted R Square     .57432
Standard Error       1.14463

            Analysis of Variance:
                 DF    Sum of Squares      Mean Square
Regression        2         177.62356         88.811778
Residuals        97         127.08804          1.310186
F =       67.78562       Signif F =  .0000

-------------------- Variables in the Equation --------------------
Variable                    B         SE B       Beta        T     Sig T
DESARROL            -.014591     .540524    -.009282     -.027   .9785
DESARROL**2          .177946     .079195     .772604     2.247   .0269
(Constant)           .610771     .877017                  .696   .4878
```

Copyrighted material

Dependent variable.. PESOFRES **Method.. CUBIC**

Multiple R	.76350
R Square	**.58293**
Adjusted R Square	.56990
Standard Error	1.15057

Analysis of Variance:

	DF	Sum of Squares	Mean Square
Regression	3	177.62602	59.208673
Residuals	96	127.08558	1.323808

F = 44.72602 Signif F = .0000

---------- Variables in the Equation ----------

Variable	B	SE B	Beta	T	Sig T
DESARROL	.074359	2.133134	.047303	.035	.9723
DESARROL**2	.146390	.736096	.635597	.199	.8428
DESARROL**3	.003340	.077445	.082222	.043	.9657
(Constant)	.540320	1.856453		.291	.7716

Dependent variable.. PESOFRES **Method.. EXPONENT**

Multiple R	.79718
R Square	**.63550**
Adjusted R Square	.63178
Standard Error	.39065

Analysis of Variance:

	DF	Sum of Squares	Mean Square
Regression	1	26.073999	26.073999
Residuals	98	14.955366	.152606

F = 170.85853 Signif F = .0000

---------- Variables in the Equation ----------

Variable	B	SE B	Beta	T	Sig T
DESARROL	.459838	.035179	.797180	13.071	.0000
(Constant)	.490165	.065458		7.488	.0000

Copyrighted material

10.5 El Análisis de Correlación.

Tal como se indicó al inicio de este capítulo, el análisis de correlación se realiza para medir el grado de asociación entre dos variables dependientes una de otra. La correlación es un indicador estadístico definido por el coeficiente de correlación –**R**- y es medido en una escala que varía entre –1 y +1. El valor de + 1, indica una correlación perfecta y directa; en cambio, el valor de -1, significa que existe una correlación perfecta e inversa. ***El valor de R = 0, significa ausencia de correlación entre las variables, lo cual es un indicador de que las variables son independientes entre si.*** El análisis de correlación puede aplicarse cuando se disponen de variables continuas o discretas de muchos valores donde se quiere saber si estas están asociadas o no.

Para ilustrar el análisis de correlación, se usarán las variables "Peso fresco de planta", "Desarrollo de la Planta", "Altura de planta" y "Diámetro del tallo". Primero, en SPSS se carga la BDD llamada **"EXPERIMENTO-SUSTRATOXFERTIRIEGO"**, luego, se ejecutan los comandos **Analyze/ Correlate/ Bivariate/,** luego en la ventana de diálogo, **"variables"**, ***se incluyen las variables que se desean analizar***. Se marcan la opción **Correlation Coefficients** y se solicita la prueba de significancia de dos colas en **Test of Significance two tailed**, y **OK**. En el cuadro 10.5, se presenta la matriz de correlación de Pearson, en la que se muestran los valores de **R**, de cada variable en relación a las otras.

Cuadro 10.5. Matriz de correlación de Pearson y sus niveles de Significación.

Correlations

		Peso Fresco de Planta	Desarrollo de la Planta	Altura de Planta	Diametro del Tallo
Peso Fresco de Planta	Pearson Correlation	1.000	.749**	.746**	.529**
	Sig. (2-tailed)	.	.000	.000	.000
	N	100	100	100	100
Desarrollo de la Planta	Pearson Correlation	.749**	1.000	.784**	.686**
	Sig. (2-tailed)	.000	.	.000	.000
	N	100	100	100	100
Altura de Planta	Pearson Correlation	.746**	.784**	1.000	.600**
	Sig. (2-tailed)	.000	.000	.	.000
	N	100	100	100	100
Diametro del Tallo	Pearson Correlation	.529**	.686**	.600**	1.000
	Sig. (2-tailed)	.000	.000	.000	.
	N	100	100	100	100

**. Correlation is significant at the 0.01 level (2-tailed).

Para interpretar esta hoja de salida, basta con observar el nivel de significancia de cada variable en relación a las demás, por ejemplo: El Peso Fresco de Planta tiene una correlación alta y positiva con las variables Desarrollo de la Planta y Altura de Planta; así mismo, tiene una correlación positiva y media con respecto al Diámetro del Tallo.

Copyrighted material

Capítulo 11. Análisis de Regresión Lineal Múltiple.

11.1 Regresión Lineal Múltiple.

El hecho que el modelo de Regresión Lineal Simple sea adecuado, no significa que no pueda ser mejorado a través de la información proporcionada por otras variables. Puede ser que al incorporar más variables al modelo, la proporción de la variabilidad explicada aumente significativamente, (Dicovskyi L., 2002). **La regresión lineal múltiple es una extensión del modelo simple al que se incorporan dos o más variables independientes. Este modelo puede ser expresado como:**

$$Y_i = B_0 + B_1 X_{1i} + B_2 X_{2i} + B_3 X_{3i} + ...+ B_p X_{pi} + e_i$$

Donde

- X_{pi}: es la puntuación de un sujeto i en la variable dependiente "***p***".
- **B:** son los parámetros estandarizados desconocidos.
- e_i: son los términos de residuos o errores, de media = 0 y variancia constante.

A continuación se presenta un ejemplo con datos de un experimento bifactorial con plántulas en semillero de tomate, cuyo objetivo era evaluar el efecto de 5 diferentes dósis de Raizal, (0, 1, 2, 3 y 4 gr/Lt de agua), en dos diferentes tipos de sustrato (Lombrihumus y Promix). Durante el crecimiento de las plántulas se evaluaron cuatro variables independientes, a saber: una variable cualitativa llamada "desarrollo de la planta", usando una escala de 1 a 5, donde el valor 1 representa la peor situación, el valor 2 es mal estado, el valor 3 es regular, el valor 4 representa un buen estado y el valor 5 es un estado muy bueno. Además, se evaluaron tres variables cuantitativas: "número de hojas verdaderas bien desarrolladas"; "altura de planta" (en cm) y "diámetro del tallo" (en mm). Estas variables independientes se relacionaron con la variable dependiente y final "peso fresco de plántula" (en gr). Los datos del experimento se presentan en el cuadro 11.1, y en la figura 11.1, se muestran los tratamientos utilizados en el experimento.

Si la regresión es real, significa que la evaluación visual a través de la variable cualitativa "desarrollo de la planta", es una buena forma de predecir una planta de mayor peso fresco y por lo tanto de mejor calidad para el transplante, 21 días después de germinadas.

Copyrighted material

Cuadro II.1. Datos del experimento bifactorial sustrato por fertiriego, en viveros de Tomate.

Tratamientos		Repeticiones	No. de Hojas	Altura de Planta (cm)	Diámetro del Tallo (mm)	Desarrollo de la Planta	Peso Fresco de Planta (gr)
Fertirriego	Sustrato						
Testigo	Lombrihumus	1	3	18	5	4	3
Testigo	Lombrihumus	2	4	16	5	3	2.2
Testigo	Lombrihumus	3	4	15	6.6	4	3.6
Testigo	Lombrihumus	4	3	17	5.5	4	3.4
Testigo	Lombrihumus	5	3	15.5	5	4	2.5
Testigo	Lombrihumus	6	3	16	5.25	4	3.7
Testigo	Lombrihumus	7	3	13	5	4	2.7
Testigo	Lombrihumus	8	3	21	5.3	4	3
Testigo	Lombrihumus	9	3	12	5.25	4	1.9
Testigo	Lombrihumus	10	3	14	3	4	2.5
Testigo	Promix	1	2	15	3	3	1.3
Testigo	Promix	2	2	12	3	3	1.7
Testigo	Promix	3	2	10	4	4	2.2
Testigo	Promix	4	2	8	5	3	1.5
Testigo	Promix	5	3	10	5	3	1.8
Testigo	Promix	6	3	13	5	3	2.5
Testigo	Promix	7	3	9.5	5	3	1.7
Testigo	Promix	8	2	8	5.5	3	1
Testigo	Promix	9	2	8.5	5.7	3	1.4
Testigo	Promix	10	2	9	5.25	3	1.2
1 gr por Lt de Agua	Lombrihumus	1	4	21	7.1	5	4.5
1 gr por Lt de Agua	Lombrihumus	2	4	20.5	7.9	5	6.4
1 gr por Lt de Agua	Lombrihumus	3	4	23	7.2	5	3.7
1 gr por Lt de Agua	Lombrihumus	4	4	19.5	6.6	5	4.4
1 gr por Lt de Agua	Lombrihumus	5	4	21	6	5	5.2
1 gr por Lt de Agua	Lombrihumus	6	3	22	6.5	5	5.7
1 gr por Lt de Agua	Lombrihumus	7	3	21	7.2	5	5
1 gr por Lt de Agua	Lombrihumus	8	4	18	6.8	5	4.5
1 gr por Lt de Agua	Lombrihumus	9	4	19.5	6.9	5	3.1
1 gr por Lt de Agua	Lombrihumus	10	4	18	7.4	5	5.6
1 gr por Lt de Agua	Promix	1	2	8	3	3	1.2
1 gr por Lt de Agua	Promix	2	3	8.5	3	3	1.9
1 gr por Lt de Agua	Promix	3	3	8	3.2	4	1.9
1 gr por Lt de Agua	Promix	4	3	7	3.7	3	1.7

Copyrighted material

Tratamientos Fertirriego	Sustrato	Repeticiones	No. de Hojas	Altura de Planta (cm)	Diámetro del Tallo (mm)	Desarrollo de la Planta	Peso Fresco de Planta (gr)
1 gr por Lt de Agua	Promix	5	2	7	3.8	3	3.8
1 gr por Lt de Agua	Promix	6	2	10.5	3.7	3	1.3
1 gr por Lt de Agua	Promix	7	2	5	3.2	3	2.8
1 gr por Lt de Agua	Promix	8	3	7.5	3.7	3	1.4
1 gr por Lt de Agua	Promix	9	2	7	3.8	3	1.8
1 gr por Lt de Agua	Promix	10	3	7.5	4.8	4	0.8
2 gr por Lt de Agua	Lombrihumus	1	3	19.5	5.1	5	5.6
2 gr por Lt de Agua	Lombrihumus	2	4	19.5	5.15	5	4.7
2 gr por Lt de Agua	Lombrihumus	3	4	23	5.9	5	3.4
2 gr por Lt de Agua	Lombrihumus	4	4	24.5	5.7	4	6.1
2 gr por Lt de Agua	Lombrihumus	5	3	23	5	4	5.4
2 gr por Lt de Agua	Lombrihumus	6	4	19.5	4.8	5	5.4
2 gr por Lt de Agua	Lombrihumus	7	4	21	4.7	5	6.8
2 gr por Lt de Agua	Lombrihumus	8	3	23	4.4	5	7.3
2 gr por Lt de Agua	Lombrihumus	9	3	25.5	4.6	5	3.1
2 gr por Lt de Agua	Lombrihumus	10	4	21	4.45	5	6.5
2 gr por Lt de Agua	Promix	1	3	6	3	3	2.4
2 gr por Lt de Agua	Promix	2	3	12	3.1	3	3.7
2 gr por Lt de Agua	Promix	3	3	13	2.5	4	3.5
2 gr por Lt de Agua	Promix	4	3	14	2.7	4	2.7
2 gr por Lt de Agua	Promix	5	3	16.5	2.3	4	1.6
2 gr por Lt de Agua	Promix	6	3	14.5	3.6	4	1.7
2 gr por Lt de Agua	Promix	7	3	15	3.6	3	1.9
2 gr por Lt de Agua	Promix	8	3	14	3.2	3	4.2
2 gr por Lt de Agua	Promix	9	3	13.5	3.2	3	4.2
2 gr por Lt de Agua	Promix	10	3	16	3.7	4	2.4
3 gr por Lt de Agua	Lombrihumus	1	4	24	5	5	6.2
3 gr por Lt de Agua	Lombrihumus	2	4	22	4.6	4	4.5
3 gr por Lt de Agua	Lombrihumus	3	3	23	4.65	5	6.1
3 gr por Lt de Agua	Lombrihumus	4	4	22.5	4.4	4	5.1
3 gr por Lt de Agua	Lombrihumus	5	4	20.5	4.55	4	4.4
3 gr por Lt de Agua	Lombrihumus	6	3	22	5.2	5	7.2
3 gr por Lt de Agua	Lombrihumus	7	3	21.5	4.75	4	3.1
3 gr por Lt de Agua	Lombrihumus	8	3	19	4.4	4	3.9
3 gr por Lt de Agua	Lombrihumus	9	4	21.5	4	5	4.9

Copyrighted material

Tratamientos		Repeticiones	No. de Hojas	Altura de Planta (cm)	Diámetro del Tallo (mm)	Desarrollo de la Planta	Peso Fresco de Planta (gr)
Fertirriego	Sustrato						
3 gr por Lt de Agua	Lombrihumus	10	4	20	4	4	5.7
3 gr por Lt de Agua	Promix	1	3	5	3	3	2
3 gr por Lt de Agua	Promix	2	3	7	2	3	2.4
3 gr por Lt de Agua	Promix	3	3	7	2	3	2.9
3 gr por Lt de Agua	Promix	4	3	13.5	2.7	3	1.1
3 gr por Lt de Agua	Promix	5	3	13.5	2.75	3	1.4
3 gr por Lt de Agua	Promix	6	3	12.5	3.55	3	2.4
3 gr por Lt de Agua	Promix	7	3	15	2.7	2	0.6
3 gr por Lt de Agua	Promix	8	2	16.5	2.4	2	1.6
3 gr por Lt de Agua	Promix	9	2	5.5	3.2	2	2.8
3 gr por Lt de Agua	Promix	10	2	5	3	2	1.7
4 gr por Lt de Agua	Lombrihumus	1	4	18	4	5	1.9
4 gr por Lt de Agua	Lombrihumus	2	3	15.5	4.7	5	4.3
4 gr por Lt de Agua	Lombrihumus	3	3	15	4.4	3	1.8
4 gr por Lt de Agua	Lombrihumus	4	4	20	4	5	3.7
4 gr por Lt de Agua	Lombrihumus	5	4	18.5	4.65	4	3.3
4 gr por Lt de Agua	Lombrihumus	6	3	20	4	4	5.5
4 gr por Lt de Agua	Lombrihumus	7	4	11	4	4	4.6
4 gr por Lt de Agua	Lombrihumus	8	4	18	3.8	4	4.2
4 gr por Lt de Agua	Lombrihumus	9	4	15	3.9	4	4.6
4 gr por Lt de Agua	Lombrihumus	10	4	11	4.5	5	4.1
4 gr por Lt de Agua	Promix	1	3	5.5	2	2	1.5
4 gr por Lt de Agua	Promix	2	3	5.5	1.8	2	1.3
4 gr por Lt de Agua	Promix	3	3	4.5	2.2	2	0.8
4 gr por Lt de Agua	Promix	4	3	4.5	2.4	2	1.9
4 gr por Lt de Agua	Promix	5	3	6	2.8	2	1.2
4 gr por Lt de Agua	Promix	6	3	5	2.9	1	0.8
4 gr por Lt de Agua	Promix	7	3	6	3	1	0.7
4 gr por Lt de Agua	Promix	8	3	5	2.3	1	0.7
4 gr por Lt de Agua	Promix	9	2	6.5	2	1	0.7
4 gr por Lt de Agua	Promix	10	2	8.5	2.3	1	0.5

Copyrighted material

11.2 Rutina para el Análisis de Regresión Múltiple con SPSS.

Primero, se carga la base de datos llamada "EXPERIMENTO-SUSTRATOXFERTIRIEGO", luego la rutina se realiza en el SPSS, ejecutándo en el módulo Analyze/Regression/Linear. En la ventana de diálogo, se introduce como dependiente la variable "peso fresco de plántula", y las restantes variables se introducen como variables independientes. Para desarrollar el análisis de regresión múltiple, se selecciona el método Forward, luego se dejan por defecto todo el resto de opciones. Finalmente se ejecuta el procedimiento de la regresión múltiple, dando el OK.

El programa selecciona como primera variable para entrar en el modelo aquella variable que en la matriz de correlaciones de Pearson, ***de entre todas las variables independientes del modelo***, tiene un coeficiente de correlación más significativo con la dependiente. El proceso se detiene cuando el grado de significación de "**t**" para las variables que quedan es menor que 0.05.

11.3 Análisis de los residuos.

11.3.1 La Normalidad de los Datos.

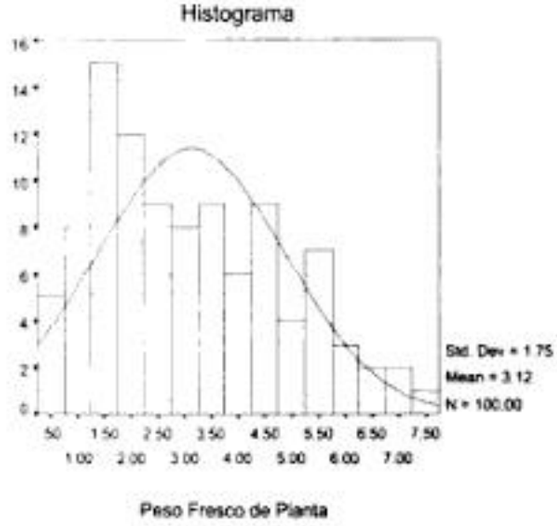
En la regresión lineal se supone que los verdaderos errores ε_i, son independientes con distribución N (0, σ^2). Respecto a la normalidad, la distribución de la variable dependiente formada por los residuos debe ser normal: los residuos observados y los esperados bajo hipótesis de distribución Normal deben ser parecidos, (Ferran A. M., 1996). Para verificar la normalidad de los datos de la variable dependiente Peso Fresco de Planta, se realiza la prueba de Kolmogorov-Smirnov, cuyo resultado se presenta en el cuadro 11.2.

*Cuadro 11.2. **Resultado de la prueba de Kolmogorov-Smirnov, para variable dependiente Peso Fresco de Planta.***

One-Sample Kolmogorov-Smirnov Test

		Peso Fresco de Planta
N		100
Normal Parameters a,b	Mean	3.1220
	Std Deviation	1.7544
Most Extreme Differences	Absolute	.127
	Positive	.127
	Negative	-.068
Kolmogorov-Smirnov Z		1.270
Asymp. Sig. (2-tailed)		0.080

a. Test distribution is Normal
b. Calculated from data



Copyrighted material

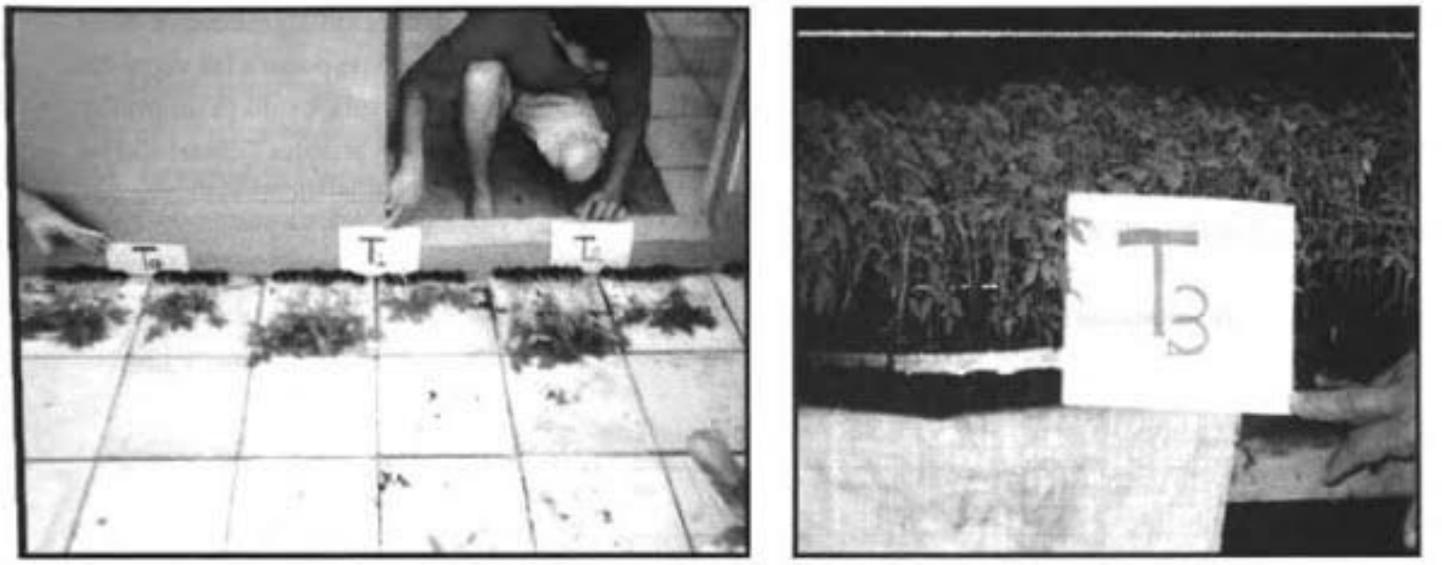


Figura 11.1. ***Tratamientos utilizados en el experimento bifactorial con plántulas de Tomate.***

En el cuadro 11.2., se verifica la normalidad de los datos para la variable "Peso Fresco de Planta", mediante la prueba de Kolmogorov-Smirnov, con una significancia de 0.08 > 0.05.

Una alternativa para comprobar la normalidad de los datos, es el gráfico de probabilidad normal, o el gráfico del histograma que permite comparar, gráficamente, la función de distribución observada en la muestra con la función de distribución Normal (0,1), (Ferran A. M., 1996).

Cuadro 11.3. ***Matriz de Correlación entre las cuatro variables independientes y su significación.***

Correlations

		NOHOJAS	Altura de Planta	Diametro del Tallo	Desarrollo de la Planta	Peso Fresco de Planta
NOHOJAS	Pearson Correlation	1.000	.595**	.440**	.600**	.578**
	Sig. (2-tailed)		.000	.000	.000	.000
	N	100	100	100	100	100
Altura de Planta	Pearson Correlation	.595**	1.000	.600**	.784**	.746**
	Sig. (2-tailed)	.000		.000	.000	.000
	N	100	100	100	100	100
Diametro del Tallo	Pearson Correlation	.440**	.600**	1.000	.686**	.529**
	Sig. (2-tailed)	.000	.000		.000	.000
	N	100	100	100	100	100
Desarrollo de la Planta	Pearson Correlation	.600**	.784**	.686**	1.000	.749**
	Sig. (2-tailed)	.000	.000	.000		.000
	N	100	100	100	100	100
Peso Fresco de Planta	Pearson Correlation	.578**	.746**	.529**	.749**	1.000
	Sig. (2-tailed)	.000	.000	.000	.000	
	N	100	100	100	100	100

** Correlation is significant at the 0.01 level (2-tailed).

La matriz de correlación de Pearson dada en el cuadro 11.3., indica una correlación significativa de la variable dependiente "peso fresco de la planta" (con significancia 0.00 < 0.05), con respecto a las variables "número de hojas", "altura de planta", "diámetro de planta" y "desarrollo de la planta"; esto es un primer indicador que estas variables **si** van a funcionar en la ecuación de regresión. Las variables "desarrollo de la planta" y "altura de planta", (con R=0.749 y R=0.746), serán las primeras variables independientes que entrarán en el modelo de regresión con el **método Forward.**

Cuadro 11.4. Incorporación de variable(s) al modelo de Regresión Lineal Múltiple.

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	DESARROL	.	Forward (Criterion: Probability-of-F-to-enter <= .050)
2	ALTURPTA	.	Forward (Criterion: Probability-of-F-to-enter <= .050)

a. Dependent Variable: PESOFRES

En el cuadro 11.4., se observa que las variables "desarrollo de la planta" y "altura de planta", son incorporadas al modelo de regresión múltiple, utilizando el criterio de incorporar aquellas variables que cumple la probabilidad de entrada F <= 0.05.

Cuadro 11.5. Correlación Múltiple (R), y Coeficiente de determinación, (R^2).

Model Summary[c]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.749[a]	.561	.557	1.1680	
2	.791[b]	.626	.619	1.0835	1.856

a. Predictors: (Constant), DESARROL

b. Predictors: (Constant), DESARROL, ALTURPTA

c. Dependent Variable: PESOFRES

En el cuadro 11.5., puede observarse el modelo 1, definido por la constante más el efecto de la variable "desarrollo de la planta", con el valor del coeficiente de correlación **R** = 0.749, y coeficiente de determinación ajustado $\mathbf{R^2}$ = 0.55, indica una alta y positiva asociación, en relación a la variable dependiente "peso fresco de plántula".

No obstante, para el modelo 2, se observa que la significancia de la regresión es mejorada al incorporar la información proporcionada por la variable "altura de la planta". En el modelo 2, es notoria la mejoría del coeficiente de correlación **R** = 0.791, y el coeficiente de determinación ajustado $\mathbf{R^2}$ = 0.619, indicando también una alta y positiva asociación, en relación a la variable dependiente "peso fresco de plántula". En este caso, el modelo 2 de regresión múltiple, aumenta la proporción de la variabilidad explicada por la variable independiente "desarrollo de la planta", al adicionarle la variable "altura de la planta"; es decir, que se explica mucho mejor la influencian de las variables independientes sobre la variable dependiente "peso fresco de la planta".

Copyrighted material

11.3.2 Independencia de los Residuos.

En el modelo 2, presentado en el cuadro 11.5, está declarada la prueba de **Independencia de los Residuos**, mediante el estadístico **Durbin-Watson=1.856**. ***Esta prueba mide el grado de autocorrelación entre los residuos. El valor de Durbin-Watson aproximado a 2, indica que se cumple correctamente el principio de que los términos de los residuos NO están correlacionados entre si.*** Por el contrario, si el valor del estadístico **Durbin-Watson** se aproxima a 4, significa que los términos del error estarán negativamente autocorrelacionados entre si. Finalmente, si el estadístico **Durbin-Watson** se aproxima 0, significa que los términos del error estarán positivamente autocorrelacionados, (Ferran A. M., 1996).

11.4. Construyendo el Modelo de Regresión Múltiple.

En el cuadro 11.6., se presenta el análisis de variancia (ANOVA), se observa que tanto para el modelo 1, como para el modelo 2, les corresponde un efecto significativo de la regresión; lo cual indica que hay un buen ajuste al modelo 1, léase efecto de regresión lineal simple entre la variable independiente "Desarrollo de la Planta" y la variable dependiente "Peso fresco de la planta".

Cuadro 11.6. ANOVA de los coeficientes "Beta" (β) de la Regresión Múltiple.

ANOVA[c]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	171.009	1	171.009	125.344	.000[a]
	Residual	133.703	98	1.364		
	Total	304.712	99			
2	Regression	190.836	2	95.418	81.277	.000[b]
	Residual	113.876	97	1.174		
	Total	304.712	99			

a. Predictors: (Constant), DESARROL

b. Predictors: (Constant), DESARROL, ALTURPTA

c. Dependent Variable: PESOFRES

Este efecto esta dado por el valor de F = 125.344 con Significancia = 0.000, que es mayor a 0.05, por tanto se rechaza la hipótesis nula de que el coeficiente de regresión **"Beta"** es igual a 0, esto implica una regresión lineal significativa. Igual interpretación corresponde al modelo 2, del cuadro 11.6, para el cual se obtuvo un valor de F = 81.277 con Significancia = 0.000, esto implica que la regresión múltiple también es significativa. Sin embargo, se destaca en este segundo caso, que la regresión está referida al efecto de regresión múltiple de las variables independientes "Desarrollo de la Planta" y "Altura de planta", sobre la variable dependiente "Peso fresco de la planta".

En los comentarios para el cuadro 11.5, ya se analizó que con el modelo de regresión múltiple, la proporción de la variabilidad explicada por las variables independientes aumenta significativamente su efecto sobre la variable dependiente "Peso fresco de la planta".

Material

Cuadro 11.7. Coeficientes Beta (β) de la ecuación de Regresión y su significación.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-1.153	.399		-2.887	.005
	DESARROL	1.178	.105	.749	11.196	.000
2	(Constant)	-1.009	.372		-2.711	.008
	DESARROL	.671	.157	.427	4.270	.000
	ALTURPTA	.119	.029	.411	4.110	.000

a. Dependent Variable: PESOFRES

En el cuadro 11.7., se presentan los valores de los coeficientes estandarizados **"Beta" (β) para el modelo 2, la prueba de significancia de la regresión múltiple**: el coeficiente **(β) es de 0.427**, para la variable "Desarrollo de la Planta", y para la variable, "Altura de planta", **(β) es de 0.411**, en ambos casos su valor de Significancia = 0.000 es < 0.05, por tanto la regresión múltiple es significativa.

Para construir la ecuación de regresión múltiple, se toma en el modelo 2, los coeficientes no estandarizados, **"B_1" que es 0.671,** para la variable "Desarrollo de la Planta", y para "Altura de planta", **"B_2" que es 0.119**. El modelo de regresión múltiple dado por la ecuación:

$$Y_i = B_0 + B_1 X_{1i} + B_2 X_{2i} + B_3 X_{3i} + ...+ B_p X_{pi} + e_i ;$$

La cual quedaría definida en este estudio, por los términos siguientes:

$$Y_i = -1.009 + 0.671 X_1 + 0.119 X_2 + e_i$$

Cuadro 11.8. Variables de exclusión del modelo.

Excluded Variables [c]

Model		Beta In	t	Sig.	Partial Correlation	Collinearity Statistics
						Tolerance
1	NOHOJAS	.201[a]	2.465	.015	.243	.641
	ALTURPTA	.411[a]	4.110	.000	.385	.385
	DIAMPTA	.028[a]	.306	.760	.031	.529
2	NOHOJAS	.129[b]	1.622	.108	.163	.600
	DIAMPTA	-.020[b]	-.232	.817	-.024	.519

a. Predictors in the Model: (Constant), DESARROL

b. Predictors in the Model: (Constant), DESARROL, ALTURPTA

c. Dependent Variable: PESOFRES

Copyrighted material

En el cuadro 11.8., se destaca para el modelo 2, que las variables "Número de hojas", y "Diámetro de planta", son **"no significativas"**, ya que tienen una Significancia = 0.108 y 0.817 respectivamente, en ambos casos son > 0.05. Por otra parte, se observa también para ambas variables que el valor de ***Tolerancia*** es alto, (0.600 y 0.519 respectivamente), lo que indica que esas variables deben ser excluidas del modelo, **por lo tanto no entran en la ecuación de regresión múltiple**.

En este ejemplo, el análisis de regresión múltiple solo incorporó dos variables independiente, tales son: "Desarrollo de la planta" y "Altura de planta" a la ecuación lineal, por lo tanto la regresión múltiple ha quedado definida para este estudio, en los términos:

$$\mathbf{Y_i = -1.009 + 0.671\ X_1 + 0.119\ X_2 + e_i}$$

Estas variables independientes son las que mejor pueden predecir la variable respuesta o dependiente, definida en este estudio como el "Peso fresco de planta". Desde el punto de vista práctico las variables "Desarrollo de la planta" y "Altura de planta", son las que se deberían usar para predecir una planta de tomate de buena calidad, lo cual indicará un mejor desarrollo vegetativo que la hace apta para el transplante a los 21 días después de germinadas en el vivero.

Copyrighted material

Capítulo 12. Análisis Multivariante de la Varianza.

12.1 Los Estudios Multivariados

Dentro de los métodos multivariados, una primera distinción está en los métodos descriptivos o exploratorios y en los métodos confirmatorios. Los métodos confirmatorios, se basan en un marco teórico que justifica y fundamenta una hipótesis que se intenta validar empíricamente, entre estos métodos se encuentra el ***Análisis Multivariado de la Varianza***, (Bizquerra, R. 1989, citado por Dicovskyi L., 2002). Por otra parte, entre los métodos multivariados descriptivos, el ***Análisis Cluster***, es uno de los más relevantes, tema del cual nos ocuparemos en el capítulo 13 de este texto.

Para el caso de una variable dependiente simple, son necesarias dos asunciones para la apropiada aplicación del análisis de varianza univariado (**ANOVA**). Los grupos deben ser: a) muestras tomadas al azar de poblaciones normales, (**Normalidad de los datos**); y b) tener varianzas semejantes, (**Homogeneidad de Varianzas**). Similares asunciones son necesarias para el análisis de varianza multivariado, (**MANOVA**).

La extensión de las asunciones del ANOVA al MANOVA, requiere que: 1) las variables dependientes tengan una **distribución normal multivariada,** 2) tengan matrices semejantes de varianza-covarianza entre cada grupo, llamada **Homocedasticidad multivariable,** y 3) tengan **Independencia Multivariable.** Es recomendable que las muestras sean grandes, n > 30, (Bizquerra, R. 1989, citado por Dicovskyi L., 2002).

Para realizar el MANOVA, se analizará solo una parte de los datos del experimento realizado con plántulas en invernadero de tomate, (micro túnel), los que se presentan en el cuadro 12.1. Los tratamientos se describen en el cuadro 12.2, y se muestran en la figura 12.1. El objetivo general de este experimento, fue evaluar el efecto de tratamientos factoriales definidos por 7 tipos de sustratos en tres tipos de bandejas (de 128, 98 y 72 nidos). Se registraron las variables: 1) "número de hojas verdaderas bien desarrolladas"; 2) "altura de planta" (en cm); 3) "diámetro del tallo" (en mm); 4) "peso fresco de plántula" (en gr); 5) "desarrollo de la planta", usando una escala de 1 a 5, donde el valor 1 representa la peor situación, el valor 2 es mal estado, el valor 3 es regular, el valor 4 representa un buen estado y el valor 5 es un estado muy bueno.

Copyrighted material

Cuadro 12.1. Datos del experimento sobre tipos de bandejas y tipos de sustratos, en vivero de tomate, establecido en estructura protegida de micro túnel.

Tratamientos		Repeti-ciones	No. de Hojas	Altura de Planta (cm)	Diámetro del Tallo (mm)	Peso Fresco de Planta (gr)	Desarrollo de la Planta
Tipo de Bandejas	Tipo de Sustratos						
De 72 Nidos	T1	1	4	13.0	4.00	6.00	4
De 72 Nidos	T1	2	4	14.0	4.00	4.40	4
De 72 Nidos	T1	3	3	13.0	4.00	4.40	4
De 72 Nidos	T1	4	3	13.0	3.00	4.60	3
De 72 Nidos	T1	5	4	14.0	4.00	4.30	4
De 72 Nidos	T1	6	4	14.0	4.00	5.90	4
De 72 Nidos	T1	7	4	19.0	4.00	5.20	4
De 72 Nidos	T1	8	4	17.0	4.00	4.50	4
De 72 Nidos	T1	9	4	14.5	4.00	4.00	4
De 72 Nidos	T1	10	4	15.0	4.00	4.50	4
De 72 Nidos	T2	1	5	19.0	4.00	7.30	4
De 72 Nidos	T2	2	5	25.0	5.00	8.30	5
De 72 Nidos	T2	3	5	24.0	5.00	8.30	5
De 72 Nidos	T2	4	5	24.0	5.00	7.60	5
De 72 Nidos	T2	5	5	22.0	5.00	6.90	5
De 72 Nidos	T2	6	5	28.0	5.00	4.50	5
De 72 Nidos	T2	7	5	27.0	4.00	8.50	4
De 72 Nidos	T2	8	5	27.0	4.00	7.60	4
De 72 Nidos	T2	9	5	24.0	5.00	8.60	5
De 72 Nidos	T2	10	5	25.0	5.00	7.80	5
De 72 Nidos	T3	1	5	24.0	5.00	8.60	5
De 72 Nidos	T3	2	5	27.0	4.00	5.60	4
De 72 Nidos	T3	3	5	26.0	4.00	8.70	4
De 72 Nidos	T3	4	5	20.0	5.00	9.10	5
De 72 Nidos	T3	5	5	24.0	5.00	10.30	5
De 72 Nidos	T3	6	5	25.0	5.00	10.90	5
De 72 Nidos	T3	7	5	25.0	5.00	9.70	5
De 72 Nidos	T3	8	5	22.0	5.00	6.80	5
De 72 Nidos	T3	9	5	24.0	5.00	9.40	5
De 72 Nidos	T3	10	5	25.0	5.00	9.50	5
De 72 Nidos	T4	1	4	20.0	5.00	5.60	5
De 72 Nidos	T4	2	4	23.0	4.00	8.70	4
De 72 Nidos	T4	3	5	22.0	4.00	9.10	4
De 72 Nidos	T4	4	5	22.0	4.00	6.20	4
De 72 Nidos	T4	5	4	23.0	4.00	7.30	4
De 72 Nidos	T4	6	5	22.0	4.00	6.90	4
De 72 Nidos	T4	7	5	22.0	5.00	8.00	5
De 72 Nidos	T4	8	5	23.0	5.00	7.10	5
De 72 Nidos	T4	9	5	22.0	4.00	5.40	4
De 72 Nidos	T4	10	4	22.0	4.00	8.00	4
De 72 Nidos	T5	1	4	19.0	4.00	7.80	4

Material

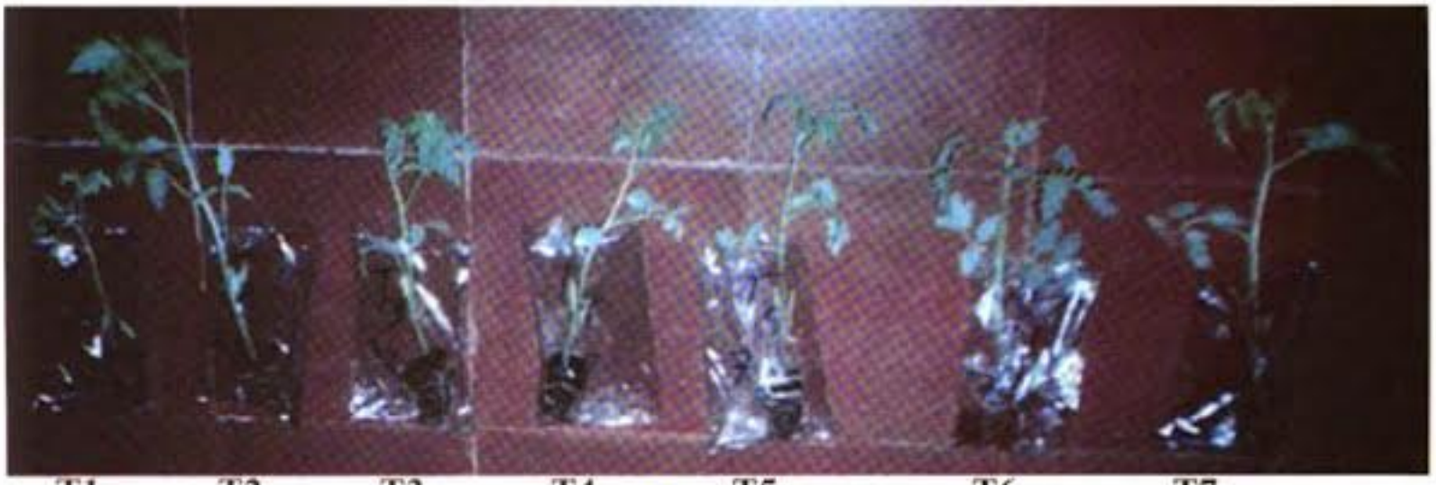
Tratamientos		Repeti-ciones	No. de Hojas	Altura de Planta (cm)	Diámetro del Tallo (mm)	Peso Fresco de Planta (gr)	Desarrollo de la Planta
Tipo de Bandejas	Tipo de Sustratos						
De 72 Nidos	T5	2	4	14.0	6.00	9.60	5
De 72 Nidos	T5	3	4	18.0	4.00	10.60	4
De 72 Nidos	T5	4	5	19.0	5.00	6.70	5
De 72 Nidos	T5	5	5	21.0	6.00	8.00	5
De 72 Nidos	T5	6	5	24.0	5.00	8.50	5
De 72 Nidos	T5	7	5	22.0	5.00	8.40	5
De 72 Nidos	T5	8	5	21.0	5.00	6.20	5
De 72 Nidos	T5	9	5	23.0	5.00	5.30	5
De 72 Nidos	T5	10	4	18.0	4.00	9.50	4
De 72 Nidos	T6	1	5	21.0	5.00	13.70	5
De 72 Nidos	T6	2	5	21.0	5.00	13.20	5
De 72 Nidos	T6	3	5	25.0	5.00	16.00	5
De 72 Nidos	T6	4	5	20.0	5.00	12.90	5
De 72 Nidos	T6	5	5	20.0	5.00	13.60	5
De 72 Nidos	T6	6	5	25.0	5.00	15.00	5
De 72 Nidos	T6	7	5	23.0	5.00	15.40	5
De 72 Nidos	T6	8	5	24.0	5.00	14.90	5
De 72 Nidos	T6	9	5	21.0	5.00	13.00	5
De 72 Nidos	T6	10	5	20.5	5.00	13.50	5
De 72 Nidos	T7	1	5	21.0	5.00	14.00	5
De 72 Nidos	T7	2	5	21.0	6.00	13.80	5
De 72 Nidos	T7	3	4	18.0	4.00	12.60	4
De 72 Nidos	T7	4	5	23.0	5.00	15.30	5
De 72 Nidos	T7	5	5	23.0	5.00	15.40	5
De 72 Nidos	T7	6	5	22.0	5.00	14.10	5
De 72 Nidos	T7	7	4	17.0	4.00	12.50	4
De 72 Nidos	T7	8	5	23.0	5.00	16.40	5
De 72 Nidos	T7	9	4	20.0	4.00	13.00	4
De 72 Nidos	T7	10	5	23.0	5.00	15.00	5

Cuadro 12.2. Descripción de los tratamientos del experimento sobre tipos de bandejas y tipos de sustratos, en vivero de tomate, establecido en estructura protegida de micro túnel.

Tratamientos		Descripción del Tipo de Sustrato
Tipo de Bandejas	Tipo de Sustratos	
De 72 Nidos	T1	Promix 100%
De 72 Nidos	T2	Lombrihumus 100%
De 72 Nidos	T3	Mitad inferior Promix y Mitad superior Lombrihumus
De 72 Nidos	T4	Mezcla uniforme de 50 % Promix y 50 % Lombrihumus
De 72 Nidos	T5	Mezcla uniforme de 50 % Promix y 50 % Abono Orgánico
De 72 Nidos	T6	Abono Orgánico MASINFA 100%
De 72 Nidos	T7	Mitad inferior Promix y Mitad superior Abono Orgánico

Copyrighted material

En el ejemplo particular que se desarrollará de aquí en adelante, por razones didácticas, se utilizarán solamente los datos correspondientes a los 7 sustratos en las bandejas de 72 nidos. Para realizar el MANOVA con estos datos, se carga la BDD **"EXPERIMENTO-TIPO DE SUSTRATO", dentro del SPSS versión 9,** y se realizan todas las pruebas estadísticas que a continuación se describen.



T1 T2 T3 T4 T5 T6 T7

T4 T3 T2 T1 T5 T6 T7

Figura 12.1 ***Tratamientos del experimento con plántulas en invernadero de tomate, (micro túnel).***

12.2 Normalidad Multivariable

Un requisito para que exista normalidad multivariable es que todas las variables dependientes sean normales. **Este requisito no implica necesariamente que todas las variables normales juntas sigan una distribución multivariable.** Una estrategia para verificar la normalidad multivariable es realizar **la prueba de normalidad (K-S) para cada una de las variables por separado. Una de las pruebas más utilizadas para comprobar la normalidad de cada una de las variables por separado, es la prueba de Kolmogorov-Smirnov (K-S), que compara la función de distribución teórica con la empírica**. La potencia de esta prueba está en función de que la muestra sea grande. (Bizquerra, R. 1989; Peña, D. 1987, citados por Dicovskyi L., 2002).

Para la prueba de **Kolmogorov-Smirnov (K-S)**, dentro del SPSS, se utiliza la rutina **Analyse/Nonparametric test/ Sample KS.** Luego, en la ventana de diálogo, se declaran el conjunto de variables dependientes para las cuales se desea verificar la normalidad de los datos, y se marca en Test de Distribución, **la opción Normal**). El resultado, se observa el cuadro 12.3.

Cuadro 12.3. Prueba de Kolmogorov-Smirnov para las variables dependientes en estudio.

One-Sample Kolmogorov-Smirnov Test

		Numero de Hojas	Altura de Planta (cm)	Diametro del Tallo (mm)	Peso Fresco de Planta (gr)	Desarrollo de la Planta
N		70	70	70	70	70
Normal Parameters[a,b]	Mean	4.67	21.2143	4.6286	9.2571	4.59
	Std. Deviation	.53	3.7208	.5940	3.5571	.52
Most Extreme Differences	Absolute	.432	.141	.334	.134	.385
	Positive	.268	.083	.255	.134	.268
	Negative	-.432	-.141	-.334	-.105	-.385
Kolmogorov-Smirnov Z		3.615	1.178	2.795	1.118	3.223
Asymp. Sig. (2-tailed)		.000	.125	.000	.164	.000

a. Test distribution is Normal.

b. Calculated from data.

El valor de Significancia > 0.05, implica que se acepta la hipótesis de normalidad para las variables "Altura de planta" y "Peso Fresco de Planta", con una significancia de 0.215 y 0.164 respectivamente. En resumen, la prueba de K-S no reconoce como variables normales las otras tres variables "Número de Hojas", "Diámetro de planta", y "Desarrollo de la Planta".

De estas tres variables, por su propia naturaleza de ser variables cuantitativa continua, es de esperar una respuesta de normalidad para la variable "Diámetro de planta", por lo que esta variable en particular será considerada para incluirla en el estudio multivariado aquí expuesto, fundamentados en el principio de normalidad dado por Little y Hills (1981), definido tal como sigue: "La normalidad significa que si se grafican todos los valores del error se obtendría una distribución aproximadamente normal. Es decir, se asume que los errores siguen una distribución normal. Las consecuencias de la no normalidad de los datos no son graves, si la desviación es moderada; solo distribuciones muy asimétricas afectan considerablemente los niveles de significancia" ... (fin de cita).

Copyrighted material

De hecho, en la figura 12.2., se observa la curva de normalidad sobre el histograma de frecuencias para la variable Diámetro de Tallo (mm). La curva de normalidad, muestra la tendencia de normalidad deseada para esta variable y no muestra una distribución muy asimétrica, por lo que no afectará los niveles de significancia al realizar el análisis multivariante de la varianza.

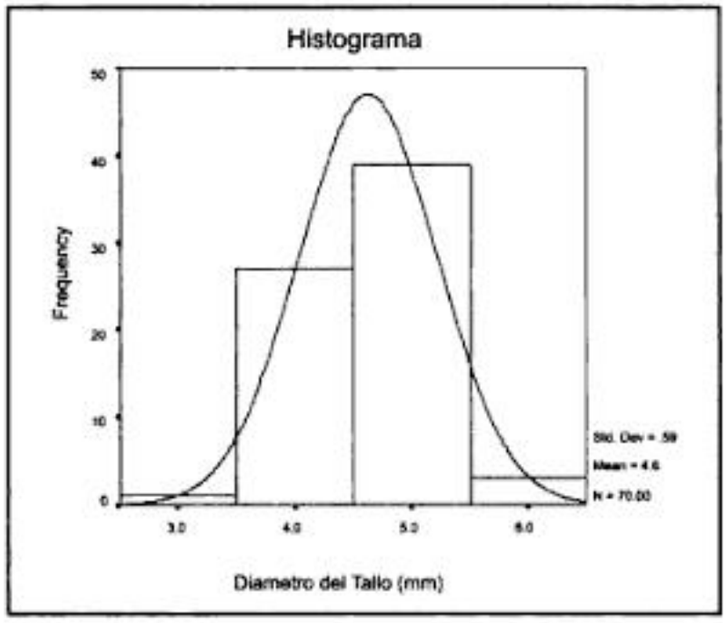


Figura 12.2. Histograma de frecuencia para la variable Diámetro de Tallo.

De tal manera que, las variables continuas a ser incluidas para el estudio multivariado serán las variables: 1) "Altura de planta", 2) "Peso fresco de planta", y 3) "Diámetro de planta".

12.3 Homocedasticidad Multivariable.

Una prueba para comprobar la homocedasticidad multivariable es la "M de Box". Esta prueba ***"sirve para comprobar la hipótesis de que las matrices de varianzas-covarianzas son iguales en todos los niveles del factor en estudio"***. Se basa en los determinantes de las matrices de varianzas-covarianzas. La M de Box se puede transformar en una "F" de Fisher o en una Chi-cuadrada, de esta forma se facilita su interpretación (Bizquerra, R. 1989, citado por Dicovskyi L., 2002).

Con las tres variables incluidas para el estudio multivariado, se realiza la prueba de **M de Box**. Esta prueba se ejecuta con los comandos **Analyze/General Linear Model/ Multivariate/** ***se declaran las variables dependientes/*** **en la ventana Fixed Factors** ***se declaran los tratamientos /*** **Option/ Homogeneity Test/ Continue/ OK. Agrupadas por los tratamientos, se obtuvieron los valores presentados en el cuadro 12.4.**

Copyrighted Material

Cuadro 12.4. Valores de la prueba M de Box.

Box's Test of Equality of Covariance Matrices[a]

Box's M	64.552
F	1.794
df1	30
df2	19632
Sig.	.005

Tests the null hypothesis that the observed covariance matrices of the dependent variables are equal across groups.

a. Design: Intrecept + SUSTRATO

El valor Significancia de 0.005 < 0.05, implica rechazar la hipótesis nula de igualdad de matrices de varianzas-covarianzas. Significa que estas matrices son diferentes para las tres variables incorporadas, según la agrupación de tratamientos. Debe recordarse que, para comprobar la homocedasticidad multivariable, es necesario alcanzar una respuesta NS en la prueba M de Box.

Una primera alternativa en la búsqueda de homocedasticidad multivariable, podría ser descartar alguna de las tres variables, mediante un proceso lógico de "tanteo" hasta llegar a definir al menos dos variables para el análisis multivariado, con las cuales si se lograra la homocedasticidad multivariable. Se realizaron estas "corridas de prueba", y el resultado fue que usando las variables "Peso Fresco de Planta" y "Diámetro del Tallo", **si** se logró la homocedasticidad multivariable.

Como un comentario aparte, este mismo procedimiento puede aplicarse para lograr la homocedasticidad multivariable, en el caso de considerar las tres variables dependientes, para tratamientos factoriales, aún estudiando el efecto de interacción. De hecho, la relación entre las variables es diferente, según si el modelo que se analiza es factorial e incluye el efecto de interacción o no. Al no contemplar el efecto de interacción, el modelo multivariado solamente analiza el efecto principal de los factores pero para cada una de las tres variables incluidas. La prueba **M de Box** obtenida para nuestro ejemplo se presenta en el cuadro 12.5.

Cuadro 12.5. Prueba M de Box, para comprobar la homocedasticidad multivariable.

Box's Test of Equality of Covariance Matrices[a]

Box's M	22.107
F	1.311
df1	15
df2	30866
Sig.	.185

Tests the null hypothesis that the observed covariance matrices of the dependent variables are equal across groups.

a. Design: Intrecept + SUSTRATO

Copyrighted material

Con el valor de Significancia de 0.185 > 0.05, si se acepta la hipótesis nula de igualdad de matrices de covarianzas. Es incluyendo las variable peso y diámetro, que se procede a realizar el MANOVA por ser normales y poseer homocedasticidad multivariable.
Pero antes debe evaluarse la independencia multivariable.

12.4 Independencia Multivariable.

Ya que no hay razón para usar el MANOVA, si las variables dependientes no están correlacionadas, una prueba para verificar la hipótesis de independencia multivariable es el **"Test de esfericidad de Bartlett"**. *Esta prueba somete a comprobación la hipótesis nula de que la matriz de correlaciones es una matriz identidad.* Esto significa que, las correlaciones entre las variables dependientes son cero. ***Si esto se confirma con la prueba de* Bartlett, *significa que las variables en estudio no están correlacionadas entre si,*** (Bizquerra, 1989, citado por Dicovskyi L., 2002).
En términos prácticos la hipótesis nula es:

Ho: ***La matriz de correlaciones es una matriz identidad -> las correlaciones entre las variables dependientes son cero,*** **-> las variables dependientes NO están correlacionadas entre si**

Ha: **Las variables dependientes SI están correlacionadas entre si.**

Se realizó el **"test de esfericidad de Bartlett"**, con las variables "Peso fresco de planta" y "Diámetro de planta". La rutina es **Analyze/General Linear Model/ Multivariate/** ***se declaran las variables dependientes/*** **en la ventana Fixed Factors** ***se declaran los tratamientos (sustratos)/*** **Option/ Residual SSCP matrix/ Continue/ OK**. En el cuadro 12.6, se presenta la prueba de independencia realizada.

Cuadro 12.6. Prueba de independencia multivariable por el "Test de esfericidad de Bartlett".

Bartlett's Test of Sphericity [a]

Likelihood Ratio	.000
Approx. Chi-Square	50.885
df	2
Sig.	.000

Tests the null hypothesis that the residual covariance matrix is proportional to an identity matrix.

a. Design: Intercept+SUSTRATO

La prueba de Bartlett, rechaza la hipótesis nula ya que el nivel de Significación es de 0.000 < 0.05, por lo tanto, ***esto indica*** que ***las variables dependientes SI están correlacionadas entre si. Esto nos faculta*** a que **si** se puede hacer uso del MANOVA con las dos variables, Peso y Diámetro.

Copyrighted material

12.5 Resolución del MANOVA.

El estadístico que más se usa para el análisis multivariante de la varianza, cuando el factor independiente en estudio tiene más de 2 tratamientos, es la distribución Lambda de Wilks que se puede aproximar a una distribución "F", (Mardia et al, 1979, citado por Dicovskyi L., 2002).

El análisis multivariante de la varianza con ***q*** factores, se basa en que la variabilidad total de la muestra puede descomponerse en la variabilidad debida a las diferencias entre grupos y a la debida a las diferencias dentro de los grupos: ***SCtotal = SCentre + SC dentro***. A partir de esta descomposición, para determinar qué parte de la variabilidad total es debida a cada uno de los dos términos, bastaría con calcular el cociente entre cada uno de ellos y la variabilidad total.

En este sentido, el estadístico **Lambda de Wilks,** compara las desviaciones dentro de cada grupo con las desviaciones totales sin distinguir grupos. Básicamente hay dos posibles respuestas o interpretaciones del estadístico de **Lambda de Wilks:**

Primera: Si el conjunto de variables dependientes ***YI, ..., Yp***, presentan un comportamiento, por un lado, muy distinto en los grupos y, por otro, muy homogéneo dentro de cada grupo, la variabilidad total será debida fundamentalmente a la variabilidad entre grupos. En consecuencia, la variabilidad dentro de los grupos será pequeña respecto a la total y el valor del estadístico Lambda de Wilks "será pequeño"; esto es lo que se interpreta como **efecto significativo multivariado**.

Segunda: Si el conjunto de variables dependientes presenta un comportamiento similar en los distintos grupos, la variabilidad entre grupos será pequeña. En consecuencia, la variabilidad total será debida fundamentalmente a la variabilidad dentro de los grupos y el valor del estadístico Lambda de Wilks "será grande"; esto es lo que se interpreta como efecto **no significativo multivariado**.

Cuanto menor se el valor del estadístico Lambda de Wilks, más diferenciados estarán los grupos y menos probable será la hipótesis nula multivareada. Tres estadísticos equivalentes a la **Lambda de Wilks,** para contrastar la hipótesis de igualdad de vectores de medias son **La Traza de Pillais, La Traza de Hotelling, y la Raíz máxima de Roy**, (Ferran, A., M. 1996).
El MANOVA verifica básicamente la hipótesis nula:

$$H_0 : \mu_1 = \mu_2 = ... = \mu_k$$

Es decir, que entre los vectores de medias de los grupos, no hay diferencias estadísticas.

En el programa SPSS, para realizar el MANOVA, es decir hacer la prueba Lambda de Wilks, primero, se carga la base de datos llamada **"EXPERIMENTO-TIPO DE SUSTRATO"**, luego la rutina se realiza con los comandos **Analyze/General Linear Model/ Multivariate/** ***se declaran las variables dependientes/*** **en la ventana Fixed Factors** ***se declara un factor fijo (sustrato)/*** **Option/ Homogeneity test/ Residual SSCP matrix/ Continue/ OK**.

Copyrighted material

En general, las variables dependientes deben ser cuantitativas. Los factores son categóricos y pueden tener valores numéricos o de cadenas de hasta ocho caracteres. Las covariables son variables cuantitativas que están relacionadas con la variable dependiente, (Estadísticas Avanzadas de SPSS 7.5, 1997). El cuadro 12.7, muestra la salida obtenida, para el análisis multivariado.

Cuadro 12.7. Test Multivariado.

Multivariate Tests [c]

		Value	F	Hypothesis df	Error df	Sig.
Intercept	Pillai's Trace	.993	4444.385 [a]	2.000	62.000	.000
	Wilks' Lambda	.007	4444.385 [a]	2.000	62.000	.000
	Hotelling's Trace	143.367	4444.385 [a]	2.000	62.000	.000
	Roy's Largest Root	143.367	4444.385 [a]	2.000	62.000	.000
SUSTRATO	Pillai's Trace	1.104	12.931	12.000	126.000	.000
	Wilks' Lambda	.092	23.753 [a]	12.000	124.000	.000
	Hotelling's Trace	7.753	39.410	12.000	122.000	.000
	Roy's Largest Root	7.468	78.411 [b]	6.000	63.000	.000

a. Exact statistic

b. The statistic is an upper bound on F that yields a lower bound on the significance level.

c. Design: Intercept+SUSTRATO

La prueba Lambda de Wilks, rechaza la hipótesis nula para el factor sustrato, dado por el valor de significancia obtenido $0.000 < 0.05$: por tanto, puede afirmarse que: Los vectores de medias de los diferentes tratamientos, no son iguales entre si, es decir las respuestas para las variables son diferentes por efecto de los tratamientos en estudio (los sustratos).

Por separado, el SPSS facilita la tabla del análisis univariado para cada uno de las variables dependientes en estudio, tal como se presenta en el cuadro 12.8.

Cuadro 12.8. Salida para el análisis univariado de las variables en estudio.

Tests of Between-Subjects Effects

		Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	Peso Fresco de Planta (gr)	768.717 [a]	6	128.120	77.362	.000
	Diametro del Tallo (mm)	9.143 [b]	6	1.524	6.316	.000
Intercept	Peso Fresco de Planta (gr)	5998.629	1	5998.629	3622.152	.000
	Diametro del Tallo (mm)	1499.657	1	1499.657	6215.684	.000
SUSTRATO	Peso Fresco de Planta (gr)	768.717	6	128.120	77.362	.000
	Diametro del Tallo (mm)	9.143	6	1.524	6.316	.000
Error	Peso Fresco de Planta (gr)	104.334	63	1.656		
	Diametro del Tallo (mm)	15.200	63	.241		
Total	Peso Fresco de Planta (gr)	6871.680	70			
	Diametro del Tallo (mm)	1524.000	70			
Corrected Total	Peso Fresco de Planta (gr)	873.051	69			
	Diametro del Tallo (mm)	24.343	69			

a. R Squared = .880 (Adjusted R Squared = .869)

b. R Squared = .376 (Adjusted R Squared = .316)

Copyrighted material

Pruebas de Separación de Medias o Comparaciones Múltiples post hoc.

Una vez que se ha determinado que existen diferencias entre las medias, las pruebas de rango **post hoc** y las comparaciones múltiples por parejas pueden determinar que medias específicamente difieren. Estas pruebas se utilizan solo para factores intersujetos; es decir, **las pruebas de comparaciones múltiples post hoc se realizan de forma separada para cada variable dependiente.** Por tanto, en el MANOVA para ver "cuales tratamientos son diferentes entre si", se procede ha realizar las pruebas de separación de medias para cada variable dependiente, ver opción de **General Lineal Model/GLM Multivariate/Post Hoc.**

Así mismo, en **Plots**, se puede solicitar al SPSS, los gráficos correspondientes, tanto para factores individuales como para tratamientos factoriales, etc.

Copyrighted material

Capítulo 13. Técnicas de Análisis Clusters

13.1 ¿Qué es el Análisis Cluster?

El análisis cluster es la denominación de un grupo de técnicas multivariantes cuyo principal propósito es agrupar objetos basándose en las características que poseen. El análisis cluster clasifica objetos, es decir, encuestados, productos u otras entidades, de tal forma que cada objeto es muy parecido a los que hay en el conglomerado con respecto a algún criterio de selección predeterminado. Los conglomerados resultantes, deberían mostrar un alto grado de homogeneidad interna dentro del conglomerado y un alto grado de heterogeneidad extrema (entre conglomerado). Por tanto, si la clasificación es acertada, los objetos dentro de los conglomerados estarán muy próximos cuando se representen gráficamente y los diferentes grupos estarán muy alejados, (Hair et al, citados por Bornemann G., 2004).

Cuando se tiene una muestra de individuos, de cada uno de los cuales se dispone de una serie de observaciones, el análisis cluster sirve para clasificarlos en grupos lo más homogéneos posibles en base a las variables observadas. La palabra ***<<cluster>>***, que define estas técnicas, se podría traducir como grupo, conglomerado, racimo, apiñarse. En general, en todos los paquetes estadísticos se conserva su nombre en inglés, aunque también se conoce como análisis de conglomerados, taxonomía numérica, análisis tipológico o clasificación automática. El cluster se puede utilizar de dos formas distintas: clasificación o representación de estructuras de datos. La clasificación, es la aplicación más común del análisis cluster, (Punj y Stewart, 1983, citados por Gómez S. M., 1998).

La formulación del problema en el análisis clusters, parte de la premisa que si "*n*" es el número de individuos en la muestra y "*p*" es el número de variables observadas, la matriz de datos que contiene las *n* x *p* observaciones tendrá "*n*" filas y "*p*" columnas. Cada fila puede ser considerada como un punto en un espacio de "*p*" dimensiones. Las coordenadas de cada punto se obtendrán a partir de de los valores en las "*p*" variables del individuo correspondiente. A partir de la representación de los "*n*" puntos-filas, teniendo en cuenta las distancias entre ellos, se trata de agruparlos en clusters o conglomerados de tal forma que, por un lado, las distancias dentro de un mismo conglomerado sean pequeñas y, por otro lado, que las distancias entre conglomerados sean grandes, (Ferran A. M., 1996).
El análisis cluster tiene como punto de partida una matriz de distancias o proximidades entre pares de sujetos (casos) o variables, la que permite cuantificar su grado de **similitud-semejanza** en el caso de proximidades -para variables-, o su grado de **disimilitud-desemejanza** en el caso de las distancias -para casos-, (Visauta, V., B. 1998).

Junto con los beneficios del análisis cluster, existen algunos inconvenientes. ***El análisis cluster puede caracterizarse como descriptivo y no inferencial.*** El análisis cluster no tiene bases estadísticas sobre las cuales deducir inferencias estadísticas para una población a partir de una nuestra, y ***se utiliza fundamentalmente como una técnica exploratoria***. Las soluciones no son únicas, en la medida en que la pertenencia al conglomerado para cualquier número de soluciones depende de muchos elementos del procedimiento y se pueden obtener muchas soluciones diferentes variando uno o más de estos elementos.

Material

El análisis clusters, se ha denominado como **análisis Q, construcción de tipología, análisis de clasificación y taxonomía numérica**. Esta variedad de nombres se debe en parte al uso de los métodos de agrupación en disciplinas tan diversas como psicología, biología, sociología, economía, ingenierías y negocios. Aunque los nombres difieren entre disciplinas, todos los métodos tienen una dimensión común: **clasificación de acuerdo a una relación natural**. Esta dimensión común representa la esencia de todas las aproximaciones del análisis cluster. Como tal, el valor fundamental del análisis cluster descansa en la clasificación de los datos, tal y como lo sugiere la agrupación "natural" de lo datos en si misma, (Hair et al, citados por Bornemann G., 2004).

13.2 Objetivo del Análisis Cluster

El objetivo del análisis cluster es definir la estructura de los datos colocando las observaciones mas parecida en grupos. La solución cluster es totalmente dependiente de las variables utilizadas como base para la medida de similitud, la adicción o asimilación de las variables relevantes pueden tener un impacto substancial sobre la solución resultante, por tanto, el investigador debe tener particular cuidado en evaluar el impacto de cada decisión implicada en el desarrollo de un análisis cluster

En el análisis cluster, el concepto de valor teórico es central, pero en una forma muy diferente del resto de las técnicas multivariantes. **El valor teórico del análisis cluster es el conjunto de variables que representan las características utilizadas para comparar objetos en el análisis cluster**. Dado que el valor teórico del análisis cluster incluye sólo las variables utilizadas para comparar objetos, **"determina el carácter de los objetos"**. El objetivo del análisis cluster es la comparación de objetos basándose en el valor teórico, **no** en la estimación del valor teórico en si misma. Esto hace crucial la definición que dé el investigador al valor teórico para el análisis cluster.

13.3 ¿Cómo Funciona el Análisis Cluster?

Para realizar correctamente el objetivo principal del análisis cluster, se deben tratar tres cuestiones básicas: En primer lugar, ¿Cómo se mide la similitud?; en segundo lugar, ¿Cómo se forman los conglomerados?; en tercer lugar ¿Cuántos grupos se forman?. Puede utilizarse cualquier número de "reglas", pero ***la tarea fundamental es evaluar la similitud "media" dentro de los conglomerados***. Se presenta a continuación una breve descripción de estos tres elementos, en base a los planeamientos dados por Hair et al, citados por Bornemann G., 2004.

13.3.1 Medición de la Similitud

Se ilustra un análisis cluster para siete observaciones (A-G), utilizando procedimientos sencillos para cada uno de ellos. La similitud será medida de acuerdo con **la distancia euclidea (en línea recta)** entre cada par de observaciones. En el cuadro 13.1 se presentan las medidas de proximidad entre cada uno de los siete encuestados.

Copyrighted material

Cuadro 13.1. Matriz de proximidad de distancias euclídeas entre observaciones.

Observación	Observación						
	A	B	C	D	E	F	G
A	-						
B	3.162	-					
C	5.099	2.000	-				
D	5.099	2.828	2.000	-			
E	5.000	2.236	2.236	4.123	-		
F	6.403	3.606	3.000	5.000	1.414	-	
G	3.606	2.236	3.606	5.000	2.000	3.162	-

Al utilizar la distancia como medida de similitud, se debe recordar que distancias mas pequeñas indican mayor similitud, de tal forma que las observaciones E-F son las mas parecidas con distancia 1,414; y A-F son las mas diferentes con distancia 6,403.

13.3.2 Formación de Conglomerados

Unas vez que se tiene la medida de similitud, se debe desarrollar el siguiente procedimiento para la formación de conglomerados. Se han propuesto muchos métodos, pero para propósito de este texto, se utilizará esta regla simple: **identificar las dos observaciones más parecidas (*cercanas*) que no están en el mismo conglomerado y combinar éstas**. Se aplica esta regla repetidas veces, comenzando con cada observación en su propio "conglomerado" y combinando dos conglomerados a un tiempo hasta que todas las observaciones estén en un único conglomerado. A esto se le denomina ***"un procedimiento jerárquico"*** dado que se opera paso a paso para formar un rango completo de soluciones cluster. Es también **un método aglomerativo**, dado que los conglomerados se forman para la combinación de los conglomerados existentes.

En el cuadro 13.2, se detallan los pasos del procedimiento jerárquico. En primer lugar representando el estado inicial con las sietes observaciones en conglomerados simples. A continuación, se unen los conglomerados en el proceso aglomerativo hasta que solo quede un conglomerado. El paso uno, identifica las dos observaciones mas cercanas (en este caso E y F) y las combina en un conglomerado, yendo de siete a seis conglomerados. A continuación, el paso dos, busca los pares de observaciones más cercanas. En este caso, tres pares tienen la misma distancia de 2.000 (E-G, C-D, y B-C). Se inicia con E-G. G es un miembro único de un conglomerado, pero E se combinó en el primer paso con F. Así que, el conglomerado formado a este nivel tiene tres miembros: **G, E, y F**. El paso tres combina los conglomerados de miembro único de C y D; y el paso cuatro combina B con el conglomerado de dos miembros C-D que se formo en el paso tres. Hasta este momento, se tienen 3 conglomerados: **Conglomerado 1 (A); Conglomerado 2 (B, C y D); y Conglomerado 3 (E, F y G).**

Copyrighted material

La siguiente distancia más pequeña es 2.236 para tres pares de observaciones (E-B; B-G y C–E). En este caso, se utiliza solo una de estas tres distancias; sin embargo, en la medida en que cada par de observaciones contiene un miembro de cada uno de los dos conglomerados existen (B, C y D, frente a E, F y G). Por tanto, el paso quinto, combina los dos conglomerados de tres miembros en un único conglomerado de seis miembros.

El paso final (paso seis), es combinar la observación A con el conglomerado restante (seis observaciones), en un único conglomerado a una distancia de 3.162. Es notorio que existen tres distancias iguales o menores a 3.162, pero que no se utilizan por que están entre los miembros del mismo conglomerado.

Cuadro 13.2. Proceso de cluster aglomerativo jerárquico.

Paso	Proceso de Aglomeración		Solución Cluster		
	Distancia Mínima entre Observaciones conjunta (distancias medias no aglomeradas)*	Par de Observaciones	Pertenencia al Conglomerado	Número de Conglomerados	Medida de Similitud número de conglomerados (dentro del conglomerado)
	Solución inicial		*(A) (B) (C) (D) (E) (F) (G)*	7	0
1	1.414	E-F	(A) (B) (C) (D) (E-F) (G)	6	1.414
2	2.000	E-G	(A) (B) (C) (D) (E-F-G)	5	2.192
3	2.000	C-D	(A) (B) (C-D) (E-F-G)	4	2.144
4	2.000	B-C	(A) (B-C-D) (E-F-G)	3	2.234
5	2.236	B-E	(A) (B-C-D-E-F-G)	2	2.896
6	3.162	A-B	(A-B-C-D-E-F-G)	1	3.420

*** Distancia euclidea entre observaciones.**

El proceso jerárquico de aglomeración puede representarse gráficamente de varias formas. En la figura 13.1, se ilustra dos de tales formas. En primer lugar, dado que el proceso es jerárquico, el proceso de aglomeración puede mostrarse como series de agrupaciones anidadas (ver figura 13.1 **a**). Este proceso, sin embargo, puede representar la proximidad de las observaciones para solo dos o tres variables de aglomeración del grafico o de dispersión. Una aproximación más aceptada convencionalmente, es **el dendrograma**, que representa el proceso de aglomeración en un gráfico con forma de árbol, (figura 13.1 **b**). **El eje horizontal representa *el coeficiente de aglomeración, en este caso la distancia utilizada en la unión de aglomerados***. Esta aproximación, es particularmente útil en la identificación de atípicos, como la observación A. También representa el tamaño relativo de los conglomerados que varían, aunque se hace difícil de manejar cuando aumenta el número de observaciones.

Copyrighted material

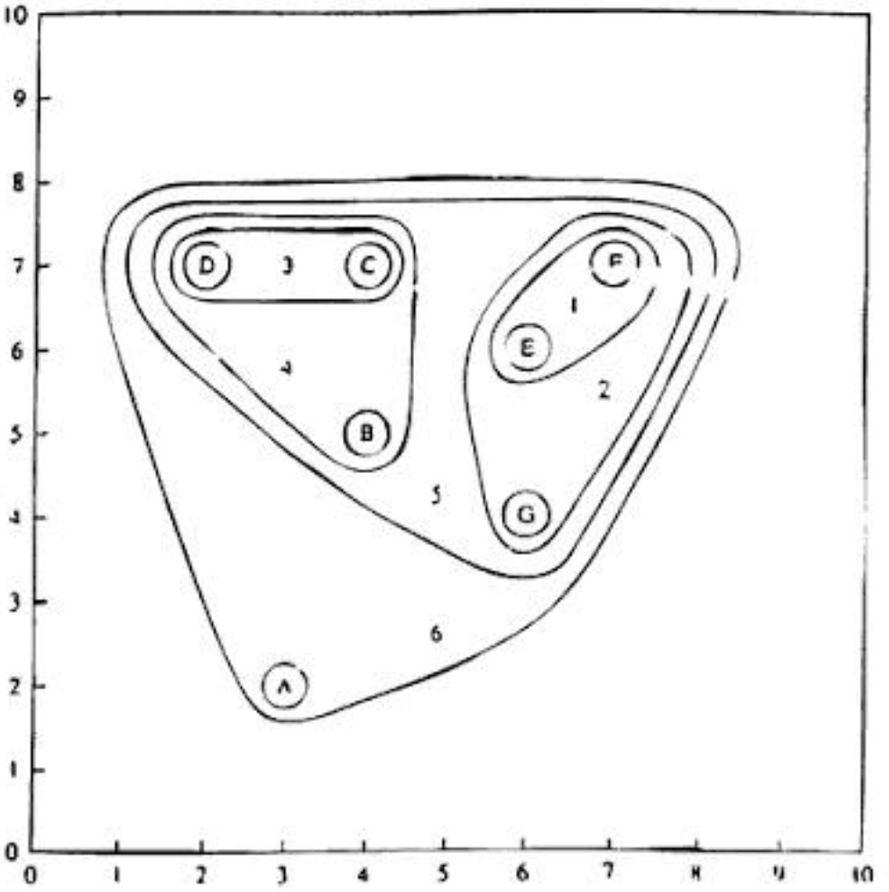


Figura 13.1 a. Representación gráfica del proceso de aglomeración en agrupaciones anidadas.

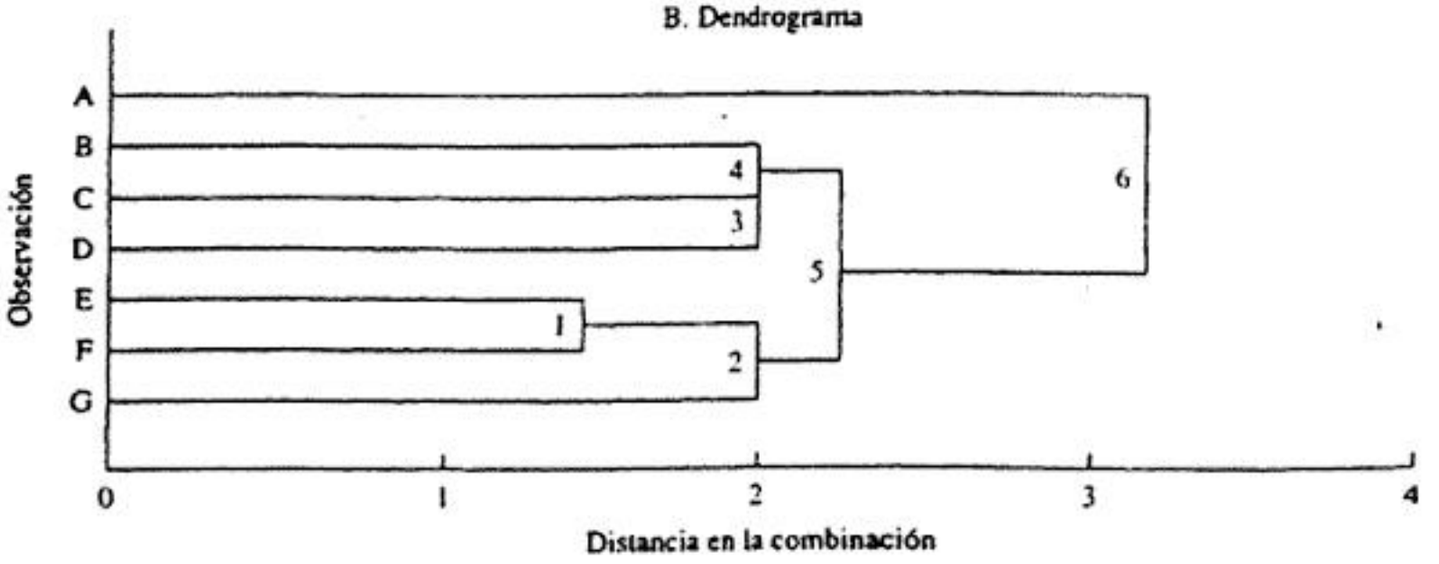


Figura 13.1 b. Representación gráfica del proceso de aglomeración en gráfico con forma de árbol, denominado como Dendrograma.

Copyrighted Material

13.3.3 Determinación del Número de Conglomerados en la Solución Final

Un método jerárquico produce un número de soluciones cluster –en el caso del inciso 13.3.2, van de una solución de un conglomerado a una solución de seis conglomerados. Pero ¿Cuál se debería elegir?. Se sabe que a medida que nos alegamos de los conglomerados de un único miembro, la homogeneidad disminuye. Así que, ¿por qué no quedarnos con los siete conglomerados, que son los más homogéneos posibles?. El problema es que no se definió ninguna estructura con siete conglomerados. De ahí que, el investigador debe ver cada solución cluster a partir de la descripción de su estructura compensada con la homogeneidad de los conglomerados. En este ejemplo, se utiliza una medida muy simple de homogeneidad: las distancias medias de todas las observaciones dentro de los conglomerados.

En la solución inicial con siete conglomerados, la medida de similitud conjunta es 0, --ninguna observación esta emparejada con otra. Para la solución de seis conglomerados, la similitud conjunta es la distancia entre las dos observaciones (1.414) unidas en el paso 1. El paso dos, forma un conglomerado de tres miembros (E, F y G), de tal forma que la medida de similitud total es la media de las distancias entre E y F (1.414); E y G (2.000); y F y G (3.162), para una media de 2.192. En el paso tres, se forma un nuevo conglomerado de dos miembros con una distancia de 2.000, que provoca que la media conjunta caiga ligeramente hasta 2.144. Se procede a formar nuevos conglomerados de esta forma hasta formar una solución de conglomerado único (paso 6), en que la media de todas las distancia de la matriz de distancias es 3.420.

Ahora bien, ¿como se utiliza esta medida conjunta de similitud para seleccionar una solución cluster?. Debe recordarse que se esta intentando coincidir la estructura más simple posible que represente agrupaciones homogéneas. Si se controla la medida de similitud conjunta a medida que disminuye el número de conglomerados, grandes aumentos en la medida conjunta indican que dos conglomerados no eran tan similares. En el ejemplo que aquí se aborda, la medida conjunta aumenta cuando en primer lugar juntamos dos observaciones (paso-1), y a continuación se hace de nuevo cuando se construye el primer conglomerado de tres miembros (paso-2). Pero en los dos pasos siguientes (3 y 4), la medida conjunta no cambia sustancialmente. Esto indica que se están formando otros conglomerados prácticamente con la misma homogeneidad de los conglomerados existentes.
Pero cuando se alcanza el paso 5, que combinada los dos conglomerados de tres miembros, se observa un gran aumento. Esto indica que al unir estos dos conglomerados se obtiene un único conglomerado marcadamente menos homogéneo. Considérese la solución cluster del paso cuatro, mucho mejor que la del paso 5.

Se puede ver también que en el paso 6, la medida conjunta de nuevo aumenta ligeramente indicando que, incluso aunque la última observación permanezca separada hasta el último paso, cuando se une cambia la homogeneidad del conglomerado. Sin embrago, dado el perfil bastante aislado de la observación A, comparada con el resto, puede ser mejor designar como miembro del **grupo de entropía,** aquellas observaciones que son atípicos e independiente de los conglomerados existentes. Por tanto, cuando se revisa el rango de las soluciones cluster, la solución de tres conglomerados del paso 4, parece ser la mas apropiada para una solución cluster definitiva, con dos conglomerados de igual tamaño y una observación atípica. De ahí que, ***la selección de la solución cluster definitiva se deja al juicio del investigador y es considerado por muchos como un proceso muy subjetivo.***

Copyrighted material

13.4 El Análisis de Conglomerados para Casos

Dada una muestra de observaciones en un conjunto grande de variables cuantitativas, el análisis cluster es una técnica para agrupar a los elementos de la muestra en grupos, denominados conglomerados, de tal forma que, respecto a la distribución de los valores de las variables, por un lado, cada conglomerado sea lo más homogéneo posible y, por otro lado, los conglomerados sean muy distintos entre si.

En los métodos jerárquicos conglomerativos ***para casos***, el análisis comienza con tantos conglomerados como individuos (cada individuo es un conglomerado inicial). A partir de estas unidades se van formando nuevos conglomerados de forma ascendente, agrupando en cada etapa "a los individuos" de los dos conglomerados más próximos. Al final del proceso, todos los individuos estarán agrupados en un único conglomerado. ***La diferencia entre los diversos métodos jerárquicos aglomerativos reside en la distancia considerada para medir la proximidad entre los conglomerados***, (Ferran A. M., 1996).

13.4.1 Medidas de Similitud

La similitud entre objetos es una medida de correspondencia, o parecido, entre objetos que van a ser agrupados. La similitud entre objetos pueden medirse de varias formas, pero tres métodos dominan las aplicaciones del análisis cluster, a saber: **1) medidas de correlación, 2) medidas de distancia, y 3) medidas de asociación.** ***Tanto las medidas de distancia, como la correlación, exigen datos paramétricos, mientras que las medidas de asociación son para datos no paramétricos***. Los conceptos sobre estas tres medidas de similitud, dados por Hair et al, citados por Bornemann G., 2004, se describen brevemente a continuación.

13.4.1.1 Medidas de Correlación

La medida de similitud entre objetos que probablemente se nos viene a la mente en primer lugar, es el coeficiente de correlación entre un par de objetos medido sobre varias variables. En efecto, en lugar de hacer la correlación entre dos conjuntos de variables, se invierte la matriz de las X variables de los objetos, de tal forma que las columnas representan los objetos; y las filas representan las variables. Por tanto, el coeficiente de correlación entre las dos columnas de números es la correlación (o similitud) entre los perfiles de los dos objetos. ***Elevadas correlaciones indican similitud, y bajas correlaciones indican falta de ella***. Por tanto, las correlaciones representan patrones para todas las variables más que las magnitudes. Las medidas de correlación, sin embargo, se utilizan rara vez por que el interés de la mayoría de las aplicaciones del análisis cluster, está en las magnitudes de los objetos, no en los patrones de valores.

13.4.1.2 Medidas de Distancia

Las medidas de similitud de distancia, que representan la similitud como la proximidad de las observaciones respecto a las otras, para las variables del valor teórico del análisis cluster, **son las medidas de similitud más utilizadas**. Las medidas de distancia, son en realidad medidas de diferencia, donde **los valores elevados indican una menor similitud**. **La distancia se convierte en medida de similitud utilizando una relación inversa**.

Copyrighted material

Hay diferentes tipos de medidas de distancia, entre ellas: **La distancia Euclidea** entre dos puntos, que es la longitud de la hipotenusa de un triangulo rectángulo, calculada por la formula que se presenta en la figura 13. 2. Este concepto es fácilmente generalizable para más de dos variables. La distancia euclidea se utiliza para calcular medidas específicas, tales como la **simple distancia Euclidea**, y la **distancia Euclidea al cuadrado o absoluta**, que es la suma de las diferencias al cuadrado sin tomar en cuenta la raiz cuadrada. La **distancia Euclidea al cuadrado** tiene la ventaja de **no** tener que tomar la raíz cuadrada, lo que acelera notablemente los cálculos, y es la medida más recomendada para los métodos de análisis cluster del ***Centroide y Ward***.

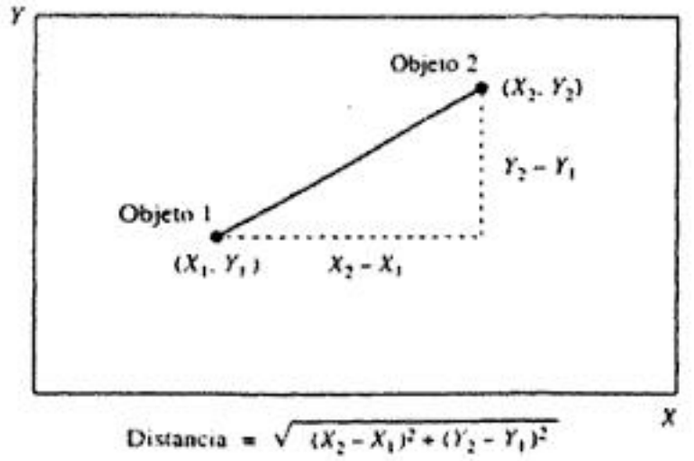


Figura 13. 2. Un ejemplo de distancia euclidea entre dos objetos medidos sobre dos variables X e Y.

13.4.1.3 Medidas de Asociación

Las medidas de asociación de similitud se utilizan para comparar objetos cuyas características se miden solo en términos **no paramétricos** (medida nominal y ordinal). Como ejemplo, véase el caso en que los encuestados responden si o no a cierto número de preguntas. Una medida de asociación podría evaluar el grado de acuerdo o de acercamiento entre cada par de encuestados. Existen diversas medidas de asociación para evaluar variables nominales de varias categorías o incluso medidas ordinales, léase el capítulo cuarto, de este libro.

13.4.2 Cómo Elegir las Variables que Participarán en la Formación de Conglomerados para Casos

Para ilustrar el análisis cluster para casos, se analizará solo una parte de los datos del estudio realizado en la micro cuenca "Pata de Gallina" por Mejía, Guzmán, Obregón y Palma, (2005). Esta micro cuenca, posee una gran diversidad de unidades de producción, y se desea clasificar los tipos de sistemas de producción agrícolas existente en la microcuenca, a partir de 14 variables relevantes que caracterizan estas unidades de producción. En este caso se incluyen las variables: 1) Edad del productor(a), 2) Salario quincenal del productor(a), 3) Área de siembra de Maíz en primera, 4) Área de siembra de Fríjol en primera, 5) Área de siembra de Sorgo en primera, 6) Área de siembra de Arroz en primera, 7) Área de siembra de Tomate en primera, 8) Área de siembra de Chiltoma en primera, 9) Área de siembra de Cebolla en primera, 10) Área de siembra de Pipián en primera, 11) Área de siembra de Sandia en primera, 12) Área de siembra de Yuca en primera, 13) Área de siembra de Camote en primera, y 14) Área de siembra de Quequisque en primera.

Copyrighted material

En este ejemplo, serán analizados solo los 17 casos correspondientes a la comunidad de Ochomogo. En la lógica del análisis cluster, cada caso puede ser considerado como un punto en un espacio de $p=14$ dimensiones (una dimensión es una variable). A partir de la representación de los $n = 17$ puntos, se trata de agruparlos –teniendo en cuenta las distancias entre ellos- en conglomerados de tal forma que respecto al resultado de clasificar los tipos de sistemas de producción agrícolas existente en la microcuenca, los casos pertenecientes a un mismo conglomerado sean semejantes entre si y diferentes a los pertenecientes a otro conglomerado.

Antes de establecer el criterio para la formación de los conglomerados, será necesario establecer una medida de la distancia entre individuos. De entre las distintas distancias disponibles, la más común utilizada es la **Distancia Euclidea** entre dos individuos, la cual se define como la raíz cuadrada de la suma de los p cuadrados de las diferencias entre los valores observados en las p variables para los individuos correspondientes. En consecuencia, será positiva cuando los dos individuos difieran en al menos un valor y nula cuando los dos individuos presenten exactamente los mismos valores en las p variables, (Ferran A. M., 1996).

Siguiendo con el ejemplo que aquí nos ocupa, se define la distancia euclidea entre dos casos cualquiera (i,j), los que pueden ser representados en el espacio de 14 dimensiones (uno por cada variable) como dos puntos de la forma:

$i = (\text{Edad}_i, \text{Salario}_i, \text{Areamaiz}_i, \text{Areafrijol}_i, \ldots\ldots\ldots\ldots \text{Areaquequisque}_i)$
$j = (\text{Edad}_j, \text{Salario}_j, \text{Areamaiz}_j, \text{Areafrijol}_j, \ldots\ldots\ldots\ldots \text{Areaquequisque}_j)$

... donde cada coordenada es el resultado observado en la medida correspondiente. Entonces, la distancia euclidea entre ellos se define como:

$$d(i,j) = \left((\text{Edad}_i - \text{Edad}_j)^2 + \ldots\ldots\ldots + (\text{Areaquequisque}_i - \text{Areaquequisque}_j)^2\right)^{1/2}$$

En este punto debe observarse que el número de variables implicadas en la distancia es grande y que, por su naturaleza, algunas de ellas podrían estar correlacionadas entre si, por tanto, contienen una información parecida. Aquí cabe destacar que, al calcular la distancia entre dos casos cualquiera, el componente debido a una variable tendrá el mismo peso que cada una de las restantes. Luego si, por ejemplo, tres variables contienen una misma información, dicha información tendrá un peso tres veces superior al de otra aportada por una única variable y, en consecuencia, en el proceso de formación de los conglomerados, la primera información será más determinante que la segunda. ***Para evitar este tipo de situaciones sesgadas, es conveniente reducir el conjunto original de variables consideradas, a un subconjunto de variables que estén incorreladas entre si, es decir que sean variables no correlacionadas entre si***, (Ferran A. M., 1996).

Copyrighted material

En el caso del ejemplo que nos ocupa, del conjunto de variables originalmente consideradas, pueden determinarse tres conjuntos, tales que, por un lado dentro de un mismo conjunto las variables están correlacionadas entre si, y por otro, cualquier par de variables en dos conjuntos diferentes están **no** correlacionadas entre si. Estos tres conjuntos son los siguientes:

(a) {Área de Maíz en primera, Área de Fríjol en primera, Área de Sorgo en primera, Área de Arroz en primera};
(b) {Área de Tomate en primera, Área de Chiltoma en primera, Área de Cebolla en primera, Área de Pipián en primera, Área de Sandia en primera};
(c) {Área de Yuca en primera, Área de Camote en primera, y Área de Quequisque en primera, Edad del productor(a), Salario quincenal del productor(a)}.

En nuestro caso, el subconjunto elegido estará formado por las variables: Área de Maíz, Área de Fríjol, Área de Sorgo, Área de Arroz, Área de Tomate, Área de Chiltoma, Área de Pipián, Área de Yuca, Área de Camote, Área de Quequisque, Edad del productor(a), Salario quincenal del productor(a). La matriz de correlaciones entre todas y c/u de las variables del subconjunto determinado, se presenta en el cuadro 13.3, dentro del cual puede observarse que la correlación muestral entre cada par de variables **es pequeña** y, para un tamaño muestral igual a 17, el ***p***-valor asociado al estadístico de prueba, (en este caso el coeficiente de correlación de Pearson), es mayor que 0.05, es decir se demuestra la **no** significancia evaluada en la mayoría de las variables. Luego entonces, al nivel de significancia del 0.05, se acepta la hipótesis nula de que tales variables, ***aquellas que no tiene asteriscos***, **no** están correlacionadas entre si.

La excepción a la afirmación del párrafo anterior, son los casos en que **si** hay significancia en la correlación evaluada, ***aquellas que tiene asteriscos***, para las variables siguientes: (1) Área Maíz primera y Área Fríjol primera; (2) Área Tomate primera y Área Fríjol primera; (3) Área Arroz primera y Área Chiltoma primera, (4) Salario quincenal y Área Pipián primera; las cuales **si** están correlacionadas entre si. Por tanto, de estos casos de variables correlacionadas entre si, se tomarán para realizar el análisis cluster, solamente una de ellas, en este caso ese tomarán las variables: Área Maíz primera, Área Tomate primera, Área Arroz primera, "Área Pipián primera". Por otra parte, quedarán excluidas para realizar el análisis cluster, las variables: Área Camote primera y Área Quequisque primera, debido a que estas dos variables se constituyeron con el valor constante de 0, por lo que no pueden ser computadas.

En base al análisis de correlación anterior, entre las variables preseleccionadas, las 7 variables que finalmente son seleccionadas y que se toman para realizar el análisis cluster son las siguientes:
1) Edad del productor(a), 2) Área de Maíz, 3) Área de Sorgo, 4) Área de Arroz, 5) Área de Tomate, 6) "Área Pipián primera", y 7) Área de Yuca.

Copyrighted material

Cuadro 13.3. Matriz de correlaciones entre las 12 variables del subconjunto seleccionado.

Correlations Coefficients

		Edad (años)	Salario quincenal	Area maíz primera	Area Frijol primera	Area Blanco primera	Area Arroz primera	Area Tomate primera	Area Chiltoma primera	Area Pipián primera	Area Yuca primera	Area Camote primera	Area Quiquisque Primera
Edad (años)	Pearson Correlation	1	-.191	.429	.196	.274	-.164	-.160	-.225	.057	.459	. [a]	. [a]
	Sig. (2-tailed)	.	.463	.085	.450	.287	.530	.539	.385	.827	.064	.	.
	N	17	17	17	17	17	17	17	17	17	17	17	17
Salario quincenal	Pearson Correlation	-.191	1	-.161	.007	.298	-.264	-.150	-.150	.639 [**]	-.014	. [a]	. [a]
	Sig. (2-tailed)	.463	.	.538	.977	.246	.305	.567	.567	.006	.957	.	.
	N	17	17	17	17	17	17	17	17	17	17	17	17
Area maíz primera	Pearson Correlation	.429	-.161	1	.698 [**]	.458	-.199	.220	-.279	.093	-.133	. [a]	. [a]
	Sig. (2-tailed)	.085	.538	.	.002	.065	.444	.396	.279	.723	.610	.	.
	N	17	17	17	17	17	17	17	17	17	17	17	17
Area Frijol primera	Pearson Correlation	.196	.007	.698 [**]	1	.121	-.182	.772 [**]	-.249	.074	-.074	. [a]	. [a]
	Sig. (2-tailed)	.450	.977	.002	.	.643	.485	.000	.336	.778	.778	.	.
	N	17	17	17	17	17	17	17	17	17	17	17	17
Area Sorgo primera	Pearson Correlation	.274	.298	.458	.121	1	-.174	-.339	.031	.373	.145	. [a]	. [a]
	Sig. (2-tailed)	.287	.246	.065	.643	.	.505	.183	.905	.140	.578	.	.
	N	17	17	17	17	17	17	17	17	17	17	17	17
Area Arroz primera	Pearson Correlation	-.164	-.264	-.199	-.182	-.174	1	-.110	.641 [**]	-.191	-.249	. [a]	. [a]
	Sig. (2-tailed)	.530	.305	.444	.485	.505	.	.673	.006	.464	.336	.	.
	N	17	17	17	17	17	17	17	17	17	17	17	17
Area Tomate primera	Pearson Correlation	-.160	-.150	.220	.772 [**]	-.339	-.110	1	-.063	-.108	-.141	. [a]	. [a]
	Sig. (2-tailed)	.539	.567	.396	.000	.183	.673	.	.812	.680	.590	.	.
	N	17	17	17	17	17	17	17	17	17	17	17	17
Area Chiltoma primera	Pearson Correlation	-.225	-.150	-.279	-.249	.031	.641 [**]	-.063	1	-.108	-.141	. [a]	. [a]
	Sig. (2-tailed)	.385	.567	.279	.336	.905	.006	.812	.	.680	.590	.	.
	N	17	17	17	17	17	17	17	17	17	17	17	17
Area Pipián primera	Pearson Correlation	.057	.639 [**]	.093	.074	.373	-.191	-.108	-.108	1	.046	. [a]	. [a]
	Sig. (2-tailed)	.827	.006	.723	.778	.140	.464	.680	.680	.	.861	.	.
	N	17	17	17	17	17	17	17	17	17	17	17	17
Area Yuca primera	Pearson Correlation	.459	-.014	-.133	-.074	.145	-.249	-.141	-.141	.046	1	. [a]	. [a]
	Sig. (2-tailed)	.064	.957	.610	.778	.578	.336	.590	.590	.861	.	.	.
	N	17	17	17	17	17	17	17	17	17	17	17	17
Area Camote primera	Pearson Correlation	. [a]	. [a]	. [a]	. [a]	. [a]	. [a]	. [a]	. [a]	. [a]	. [a]	. [a]	. [a]
	Sig. (2-tailed)	.	.	.	.	.	.	.	.	.	.	.	.
	N	17	17	17	17	17	17	17	17	17	17	17	17
Area Quiquisque Primera	Pearson Correlation	. [a]	. [a]	. [a]	. [a]	. [a]	. [a]	. [a]	. [a]	. [a]	. [a]	. [a]	. [a]
	Sig. (2-tailed)	.	.	.	.	.	.	.	.	.	.	.	.
	N	17	17	17	17	17	17	17	17	17	17	17	17

**. Correlation is significant at the 0.01 level (2-tailed).

a. Cannot be computed because at least one of the variables is constant.

Material

Finalmente, la distancia euclidea entre dos casos cualquiera ***i*** y ***j***, será calculada considerando únicamente la información del subconjunto formado por estas 7 variables, **no** correlacionadas entre si. La distancia euclidea estará dada por:

$$d(i,j) = ((\text{Edad}_i - \text{Edad}_j)^2 + (\text{Areamaiz}_i - \text{Areamaiz}_j)^2 + \ldots\ldots\ldots + (\text{AreaYuca}_i - \text{AreaYuca}_j)^2)^{1/2}$$

13.4.3 El Proceso de Tipificación de las Variables

El investigador debe resolver solo con una cuestión mas antes de proceder a realizar el análisis cluster: ¿Deberían tipificarse los datos antes de calcular las similitudes?. Esto debe ser así, dado que la medida de la distancia euclidea que se calculará, presenta el inconveniente de que su valor depende de las unidades de las variables. Es decir, tomando en cuenta que, en condiciones normales, los límites de los rangos de variación de las 7 variables seleccionadas para realizar el análisis cluster, son muy diferentes, tales como: años, Mz de Maíz, Mz de Sorgo, Mz de Arroz, hasta llegar a Mz de Tomate, Mz de Pipián, y Mz de Yuca.

Al utilizar diferentes unidades de medida, en las distintas variables, aquellas que se midan con grandes números solaparán a las variables que se miden con números pequeños, (Gómez, S. 1998). El problema surge, si dos casos cualquiera presentaran el mismo valor en dos de las variables y en la tercera variable difieren en una unidad, en ese caso, la distancia euclidea será igual a 1, independientemente de cuál sea la variable en la que difieren. Sin embargo, es muy claro que una diferencia de una unidad en el caso de la variable Edad, es una cantidad relativamente muy pequeña, en comparación con una unidad de la variable Área de Tomate. ***Para evitar este tipo de situaciones, es conveniente realizar el análisis sobre los valores tipificados*, proceso conocido como Estandarización de las variables**. El proceso de tipificación de las variables consideradas, consiste en restar a cada uno de sus valores la media aritmética de la variable y dividir la diferencia entre la desviación típica, (Ferran A. M., 1996).

En nuestro caso, la tipificación de las 7 variables consideradas para realizar el análisis cluster, se ejecuta mediante el comando **<Descriptives statistics/ Descriptives / Save standardized values as variables>**. Los resultados obtenidos del proceso de estandarización, se presentan en el cuadro 13.4, en el cual se observa la media ("Mean") y la desviación típica ("Std Dev") de cada una de las 7 variables. En la Base de datos correspondiente se generan las nuevas variables estandarizadas, tales como: 1) Zscore Edad del productor(a), 2) Zscore Área de Maíz, 3) Zscore Área de Sorgo, 4) Zscore Área de Arroz, 5) Zscore Área de Tomate, 6) Zscore Área de Pipiána, 7) Zscore Área de Yuca.

Copyrighted material

Cuadro 13.4. Tipificación de las 7 variables consideradas para realizar el análisis cluster.

Descriptive Statistics					
Variable	N	Minimum	Maximum	Mean	Std. Deviation
Edad (años)	17	33	80	50.94	15.971
Area maíz primera	17	.50	3.00	1.3382	.77531
Area Sorgo primera	17	.00	2.00	.9154	.69630
Area Arroz primera	17	.00	.50	.0735	.17150
Area Tomate primera	17	.00000	.25000	.0147059	.06063391
Area Pipián primera	17	.00000	.12500	.0147059	.03514348
Area Yuca primera	17	.00	.50	.0919	.16846
Valid N (listwise)	17				

Las nuevas variables generadas **"Zscore"**, a partir de la tipificación de los valores de la variable correspondiente, se realizan de tal como sigue:

1) **Zscore Edad del productor(a)**: Valor de la edad del productor - 50.94 / 15.971
2) **Zscore Área de Maíz**: Valor del Área de Maíz – 1.3382 / 0.77531
3) **Zscore Área de Sorgo**: Valor del Área de Sorgo – .9154 / 0.69630
4) **Zscore Área de Arroz**: Valor del Área de Arroz - .0735 / 0.17150
5) **Zscore Área de Tomate**: Valor del Área de Tomate - .0147059 / 0.06063391
6) **Zscore Área de Pipián**: Valor del Área de Pipián - .0147059 / 0.03514348
7) **Zscore Área de Yuca**: Valor del Área de Yuca - .0919 / 0.16846

En consecuencia, las 7 nuevas variables tipificadas tendrán **media = 0**; y **desviación típica = 1**.
En esta nueva situación, la distancia euclidea entre los casos ***i*** y ***j***, estará dada por:

$$d(i,j) = ((ZEdad_i - ZEdad_j)^2 + (ZAreamaiz_i - ZAreamaiz_j)^2 + \ldots\ldots\ldots\ldots + (ZAreaYuca_i - ZAreaYuca_j)^2)^{1/2}$$

Impactos de los valores de los datos no estandarizados: un problema al que se enfrentan todas las medidas de distancia es que el uso de datos no estandarizados implica inconsistencias entre las soluciones cluster cuando cambia la escala de las variables. El orden de las similitudes puede cambiar profundamente con un sólo un cambio en la escala de una de las variables. El investigador notara el tremendo impacto que la escala de las variables puede tener sobre la solución final. Por tanto, debería emplearse la estandarización de las variables de aglomeración, siempre que sea conceptualmente posible, para evitar casos como fuera de toda lógica.

Copyrighted material

Valores Atípicos: En la búsqueda de una estructura, el análisis cluster es muy sensible a la inclusión de variables irrevelantes. El análisis cluster es también sensible a los atípicos, es decir aquellos objetos que son muy diferentes del resto. Los atipicos representan tanto: (1) observaciones verdaderamente "aberrantes", **no** representativas de la población, o (2) una muestra reducida del grupo o grupos de la población que provoca una mala representación del grupo o grupos de la muestra. Siempre es necesaria una representación preliminar de los atípicos. Probablemente la forma más sencilla de llevar a cabo esta representación es preparar un diagrama de perfil grafico. *Los atipicos son aquellos objetos con perfiles muy diferentes, la mayoria caracterizados por valores extremos sobre una o más variables,* (Hair et al, citados por Bornemann G., 2004).

Una vez definidas: a) las variables a considerar para realizar el análisis cluster, b) la tipificación de tales variables, y c) la distancia entre las variables tipificadas, el siguiente paso será establecer el método que se utilizará para la conformación de los conglomerados. Para los propósitos del presente texto, se realizará el análisis cluster -en el siguiente acápite-, ***mediante los métodos Jerárquicos Aglomerativos,* tanto para casos como para variables**.

Los métodos Jerárquicos Aglomerativos, se caracterizan por ir agrupando o dividiendo los grupos sin tener que determinarse "a priori", el número de grupos final. En el procedimiento aglomerativo, se agrupan los casos en grupos hasta que se forma un único grupo. Si se utilizan métodos **no** Jerárquicos, se debe dar una solución "a priori" en cuanto al número de grupos que se forman, (Gómez S. M., 1998).

13.4.4 El Proceso de Formación de Conglomerados para Casos, por el Método Jerárquico Aglomerativo Promedio entre Grupos

En el método del promedio entre grupos se define la distancia entre dos conglomerados como el promedio de las distancias entre todos los pares de individuos, en los que cada componente del par pertenece a un conglomerado distinto. La ventaja de este método, radica en que el proceso de formación de conglomerados se puede seguir etapa por etapa. En consecuencia, el número de conglomerados, que se desea formar se puede elegir *a posteriori*, en función de la solución obtenida en cada etapa. Sin embargo, cuando el número de casos y de variables es elevado, requiere de un número de cálculos elevados. Para agilizar el proceso de cálculo, en lugar de la propia ***distancia euclidea***, se utiliza su cuadrado ***la distancia euclidea al cuadrado***, (Ferran A. M., 1996). Para ilustrar el análisis cluster **para casos**, tal como se definió en el inciso 13.4.2, se analizará una parte de los datos del estudio realizado en la micro cuenca "Pata de Gallina" por Mejía, Guzmán, Obregón y Palma, (2005). Para simplificar la clasificación de los sistemas de producción agrícola existentes en esta microcuenca, serán analizados solo 17 casos de la comunidad de Ochomogo.

La rutina para realizar el Análisis Cluster de Casos en SPSS, es la siguiente:
Primero, cargar la BDD OCHOMOGO, luego se procede a la estandarizaron de las variables que serán incluidas en el análisis cluster, para obtener sus valores **Zscores**, esto se logra mediante el comando **<Descriptives statistics/ Descriptives / Save standardized values as variables>**. En este caso se incluyen las 7 variables anteriormente consideradas para realizar el análisis cluster.

Copyrighted material

Segundo, se ejecuta el comando **<Classify/ Hierarchical Cluster/** en la ventana de diálogo se incluyen las 7 variables antes estandarizadas. También debe seleccionarse **en la ventana *Cluster la opción Cases*; y** en la ventana ***Display seleccionar las opciones Statistics y Plots.***

Tercero, especificar las opciones, tal como: **en la opción Method**, se selecciona el método de aglomeración dado por **"Between group linkage"**, también se selecciona el intervalo de **"Distancia Euclídea"** y **dar continue**. En la opción **Statistics**, se selecciona la pertenencia al conglomerado **marcando un chek en "Agglomeration schedule"** y después **seleccionar el "Range of solutions"** desde 2 hasta 17. En la **opción Plots**, se solicita el gráfico del **"Dendrograma"** correspondiente, y marcar **"All clusters"** y **dar continue**. Finalmente, **dar OK** para correr la rutina del análisis cluster.

El archivo de salida que proporciona el SPSS, brinda la información detallada de lo que sucede en cada una de las etapas, dada en el calendario de aglomeración presentado en el cuadro 13.5, "Aglomeration Schedule using Average Linkage (Between Groups).

Cuadro 13.5. Análisis cluster, usando el Método Jerárquico Aglomerativo Promedio entre Grupos.

Agglomeration Schedule

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	3	5	.328	0	0	6
2	7	17	.751	0	0	6
3	6	10	.965	0	0	12
4	4	9	1.118	0	0	14
5	1	12	1.521	0	0	9
6	3	7	1.602	1	2	7
7	3	14	1.970	6	0	10
8	2	11	2.094	0	0	11
9	1	8	2.208	5	0	13
10	3	15	2.254	7	0	11
11	2	3	2.924	8	10	12
12	2	6	3.307	11	3	13
13	1	2	3.693	9	12	14
14	1	4	4.018	13	4	15
15	1	13	4.170	14	0	16
16	1	16	5.140	15	0	0

La información detallada del cuadro 13.5, se representa de manera gráfica en el denominado ***dendograma***, (ver figura 13.3). A la izquierda del dendograma, (columna "seq") ***se observa que aparece un listado que identifica el número de casos, de cada uno de los casos observados***. La representación gráfica del proceso se realiza mediante líneas paralelas a la barra horizontal en la parte superior, correspondiente a la distancia entre los dos conglomerados que se combinan en cada etapa (transformada a enteros comprendidos entre 0 y 25). A la altura de la distancia cero saldrá una línea a la derecha cada caso (en este caso un total de diecisiete líneas). Las líneas consecutivas se irán cerrando mediante una línea vertical a medida que se vayan combinando los conglomerados. Después de un cierre vertical, cada línea horizontal que permanezca corresponderá a un conglomerado, el formado por todos aquellos casos que convergen en ella.

Copyrighted Material

Dendrogram using Average Linkage (Between Groups)

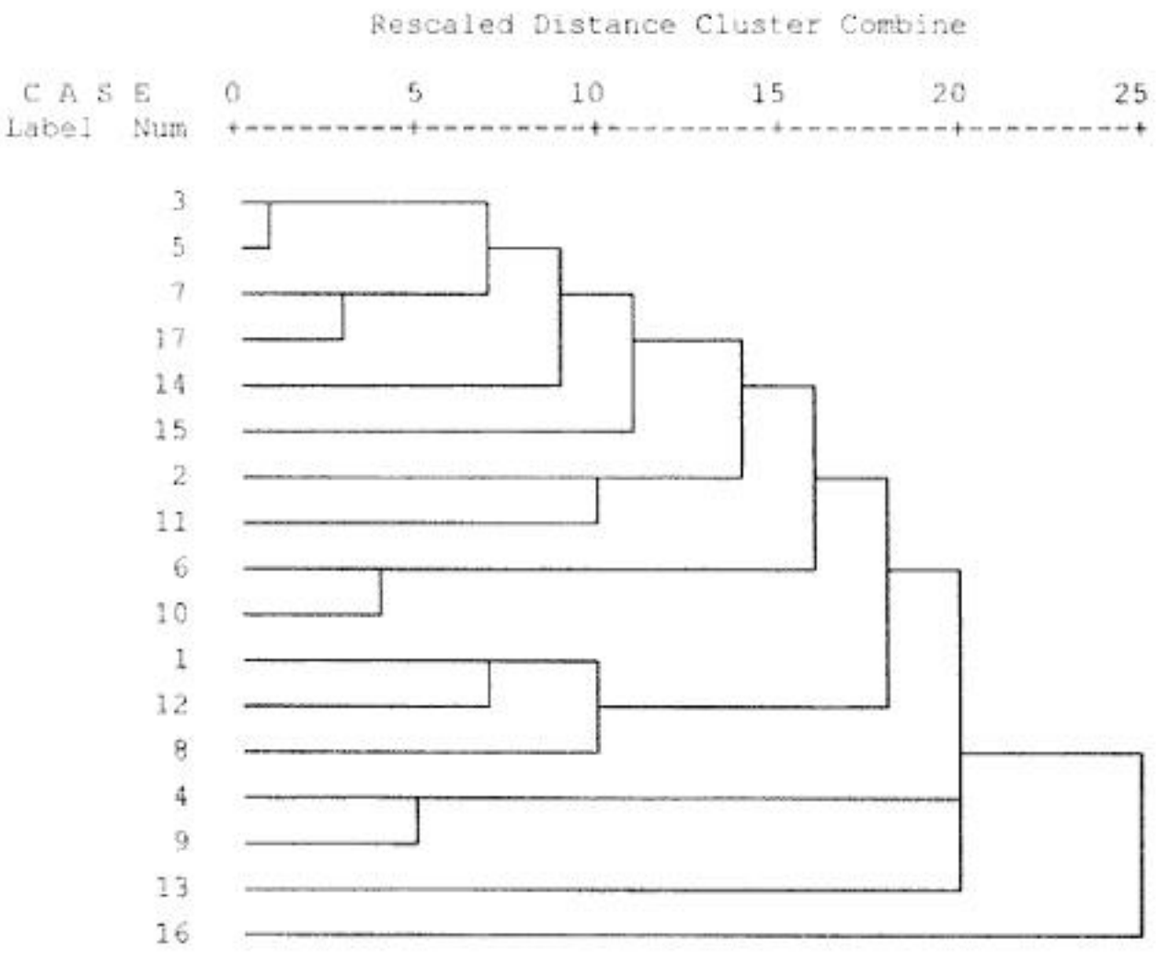


Figura 13.3. Formación Jerárquica de Conglomerados Aglomerativos de casos.

La primera columna del cuadro 13.5, ("Stage"), contiene el número de etapas del proceso. Dado que por un lado, en cada etapa se combinará el contenido de dos conglomerados y que, por otro lado, el número total de casos analizados es igual a 17, entonces el número total de etapas en el proceso será igual a 16. Después de la etapa decimosexta, todos los casos se encontrarán en un único conglomerado.

Para interpretar el calendario de aglomeración, junto con el dendograma, se inicia el proceso considerando que cada caso es un conglomerado, y cada uno de ellos adopta la denominación del número del caso correspondiente en el registro de la base datos. Inicialmente, los conglomerados son:

$$C_1 = \{1\},\ C_2 = \{2\},\ C_3 = \{3\},\ \ldots\ldots\ldots\ldots\ C_{16} = \{16\},\ C_{17} = \{17\},$$

Observe en el cuadro 13.5, que en la primera etapa ("Stage 1"), se combinan los casos 3 y 5, tales que la distancia euclidea entre ellos es la mínima de entre todas las posibles. Se combinan los casos "Cluster combined: Cluster 1 = 3; Cluster 2 = 5; o lo que es equivalente los conglomerados C_3 y C_5, y la distancia entre ellos *es la menor entre todas*, con un coeficiente igual a 0.328. En el dendograma se observan las líneas que corresponden a los casos 3 y 5, las cuales son las primeras que se cierran en una única línea.

En consecuencia, a partir de la altura del cierre, únicamente quedaran 16 líneas, correspondientes a los 16 conglomerados resultantes después de combinar los conglomerados iniciales C_3 y C_5, en un único conglomerado, ***el que adoptará el nombre del mínimo número de casos al que contenga, en este caso*** C_3. Esto indicará que, después de la primera etapa la solución obtenida es:

$$C_1 = \{1\},\ C_2 = \{2\},\ C_3 = \{3,5\},\ \ldots..\ C_7 = \{7\},\ \ldots\ldots\ldots\ldots\ C_{17} = \{17\},$$

La próxima vez que el conglomerado C_3 se combinará con algún otro conglomerado, ("Next Stage"), será en la etapa 6, en la cual se combinará con el conglomerado C_7 ("Cluster combined: Cluster 1 = 3; Cluster 2 = 7). De manera similar a lo que sucede con el conglomerado C_3, que contiene a los casos 3 y 5, podría suceder que el conglomerado C_7 contenga, además del caso 7, a cualquier otro caso o casos. Para resolver esta incógnita, bastaría con mirar en la columna "Stage Cluster 1st Appears: Cluster 1". Obsérvese que, si en la columna "Stage Cluster 1st Appears: Cluster 1" el valor que aparece es un 1, indica que el conglomerado C_7 procede de la etapa 1; en la columna "Stage Cluster 1st Appears: Cluster 2" el valor que aparece es un 2. Luego, el conglomerado C_7 se cierra en la etapa 2.

Efectivamente, en el dendograma se observa que la línea correspondiente al C_7 se cierra a partir de la línea que parte del caso 3. La distancia entre los conglomerados C_3 y C_7 será igual al promedio de las distancias euclídeas entre el caso 7 y cada uno de los casos 3 y 5. El conglomerado así obtenido se denominará según el nombre del mínimo número de casos al que contenga, en este caso se llamará C_3. Esto indicará que, después de la etapa 6, la solución obtenida es:

$$C_1 = \{1\},\ C_2 = \{2\},\ C_3 = \{3, 5, 7\},\ \ldots..\ C_8 = \{8\},\ \ldots\ldots\ldots\ldots\ C_{17} = \{17\},$$

La próxima vez que el conglomerado C_3 se combinará con otro conglomerado, ("Next Stage"), será en la etapa 7, en donde se combinará con el caso 14. Obsérvese en la columna "Cluster combined: Cluster 1 = 3; Cluster 2 = 14. El conglomerado así obtenido se denominará conglomerado C_3 y la solución obtenida, después de la etapa 7, es la siguiente:

$$C_1 = \{1\},\ C_2 = \{2\},\ C_3 = \{3, 5, 7, 14\}\ \ldots..\ C_8 = \{8\},\ \ldots\ldots\ C_{17} = \{17\},$$

La próxima vez que el conglomerado C_3 se combinará con otro conglomerado, ("Next Stage"), será en la etapa 10, en donde se combinará con el caso 15. Obsérvese en la columna "Cluster combined: Cluster 1 = 3; Cluster 2 = 15. El conglomerado así obtenido se denominará conglomerado C_3, y la solución obtenida, después de la etapa 10, es la siguiente:

$$C_1 = \{1\},\ C_2 = \{2\},\ C_3 = \{3, 5, 7, 14, 15\},\ \ldots..\ C_8 = \{8\},\ \ldots\ldots\ldots\ldots\ C_{17} = \{17\},$$

Material

La próxima vez que el conglomerado C_3 se combinará con otro conglomerado, ("Next Stage"), será en la etapa 11, en donde se combinará con el caso 2. Obsérvese en la columna "Cluster combined: Cluster 1 = 2; Cluster 2 = 3. El conglomerado así obtenido se denominará conglomerado C_2 y la solución obtenida, después de la etapa 11, es la siguiente:

$$C_1 = \{1\},\ C_2 = \{2, 3, 5, 7, 14, 15\},\ \ldots..\ C_8 = \{8\},\ \ldots\ldots\ldots\ldots\ C_{17} = \{17\},$$

La próxima vez que el conglomerado C_2 se combinará con otro conglomerado, ("Next Stage"), será en la etapa 12, en donde se combinará con el caso 6. Obsérvese en la columna "Cluster combined: Cluster 1 = 2; Cluster 2 = 6. El conglomerado así obtenido se denominará conglomerado C_2 y la solución obtenida, después de la etapa 12, es la siguiente:

$$C_1 = \{1\},\ C_2 = \{2, 3, 5, 6, 7, 14, 15\},\ \ldots..\ C_8 = \{8\},\ \ldots\ldots\ldots\ldots\ C_{17} = \{17\},$$

La próxima vez que el conglomerado C_2 se combinará con otro conglomerado, ("Next Stage"), será en la etapa 13, en donde se combinará con el caso 1. Obsérvese en la columna "Cluster combined: Cluster 1 = 1; Cluster 2 = 2. El conglomerado así obtenido se denominará conglomerado C_1 y la solución obtenida, después de la etapa 13, es la siguiente:

$$C_1 = \{1, 2, 3, 5, 6, 7, 14, 15\},\ \ldots..\ C_8 = \{8\},\ \ldots\ldots\ldots\ldots\ C_{17} = \{17\},$$

La próxima vez que el conglomerado C_1 se combinará con otro conglomerado, ("Next Stage"), será en la etapa 14, en donde se combinará con el caso 4. Obsérvese en la columna "Cluster combined: Cluster 1 = 1; Cluster 2 = 4. El conglomerado así obtenido se denominará conglomerado C_1 y la solución obtenida, después de la etapa 14, es la siguiente:

$$C_1 = \{1, 2, 3, 4, 5, 6, 7, 14, 15\},\ \ldots..\ C_8 = \{8\},\ \ldots\ldots\ldots\ldots\ C_{17} = \{17\},$$

La próxima vez que el conglomerado C_1 se combinará con otro conglomerado, ("Next Stage"), será en la etapa 15, en donde se combinará con el caso 13. Obsérvese en la columna "Cluster combined: Cluster 1 = 1; Cluster 2 = 13. El conglomerado así obtenido se denominará conglomerado C_1 y la solución obtenida, después de la etapa 15, es la siguiente:

$$C_1 = \{1, 2, 3, 4, 5, 6, 7, 13, 14, 15\},\ \ldots..\ C_8 = \{8\},\ \ldots\ldots\ldots\ldots\ C_{17} = \{17\},$$

Para finalizar con la ilustración del proceso de formación de conglomerados, se analiza lo que sucede en la última etapa. En la etapa 16, se combinan los conglomerados 1 y 16. En el extremo derecho del dendograma, cuando únicamente quedan dos líneas, la segunda de ellas arrastra al resto, al encadenarse con la etapa 15, cerrándose en la etapa 16, (obsérvese en la columna "Stage Cluster 1st Appears: Cluster 1" el valor que aparece es un 15, indica que el conglomerado C_{16} procede de la etapa 15, y en la columna "Stage Cluster 1st Appears: Cluster 2" el valor que aparece es un 0. Por tanto, el conglomerado C_{16} se cierra en la etapa 16 indicando que después de la etapa 16, al combinarse los dos anteriores, se conforman en un único conglomerado, la solución final será:

$$C_1 = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17\},$$

Copyrighted material

Luego de los 17 conglomerados iniciales, combinando paso a paso el contenido de dos de ellos, se logra alcanzar un único conglomerado formado por todos los casos. ***"Mediante el numero de líneas horizontales en el dendograma se puede conocer que casos forman la solución de cualquier número de conglomerados".***

Un aporte del análisis cluster es poder considerar el tipo de solución que se desea. Si lo que se desea es una solución en la que los conglomerados sean distantes entre si y, por otro lado, dentro de cada uno de ellos los elementos que lo forman estén próximos, una solución adecuada sería aquella tal que las líneas correspondientes tardaran en cerrarse. Si ***"a priori"***, se desea un número específico de conglomerados, la solución puede obtenerse directamente solicitando al comando **<Classify/ Hierarchical Cluster/ en la ventana de diálogo /...** En la opción **Statistics**, se selecciona **"Single solution"** ***y se escribe el número específico de clusters 4***. En la opción **Save**, se selecciona **Cluster Membership: "Single solution"** ***y se escribe el número específico de clusters 4***; dar **Continue**. Finalmente, **dar OK** para correr el análisis cluster, (Ferran A. M., 1996).

En el ejemplo aquí realizado, la generación de la variable para dicha solución fue solicitada ***"a posteriori"***, mediante el uso de la **opción Save, / Cluster Membership: "Single solution"/ ...** ***Así se genera la variable CLU4_1,*** *cuyos valores coinciden con el número de conglomerado al que ha sido asignado cada caso en la solución de cuatro conglomerados.* En el cuadro 13.6, se presentan los miembros de los 4 clusters, ***lo que facilita conocer cuales casos forman la solución de cada uno de los clusters, para establecer así la clasificación de los sistemas de producción agrícola existentes en la comunidad de Ochomogo.***

Cuadro 13.6. Membresía de cada uno de clusters. Solución con cuatro conglomerados.

Cluster Membership

Case	4 Clusters
1	1
2	1
3	1
4	2
5	1
6	1
7	1
8	1
9	2
10	1
11	1
12	1
13	3
14	1
15	1
16	4
17	1

Para presentar la clasificación obtenida, se sugiere al lector interesado, ver como un estudio de caso particular en Bornemann G., 2005, La aplicación de Modelos Multivariantes en Sistemas de Producción Agropecuarios del Municipio de Cardenas, Rivas, Nicaragua.

Copyrighted material

13.4.5 Validación de la Solución Cluster

Hasta aquí se ha desarrollado el ejemplo contemplando tres aspectos medulares del análisis cluster, tales como: la selección de variables, el método de agrupación, y la determinación del número de grupos. Sin embargo, todavía falta abordar, aunque sea brevemente, la validación de la solución cluster obtenida.

Hair et al, citados por Bornemann G., (2004), explican que dado la naturaleza de alguna forma subjetiva del análisis cluster, sobre la selección de una ***"solución cluster óptima"***, el investigador debería tener mucho cuidado en la validación de la solución cluster preliminar alcanzada y asegurarse de la pertinencia y relevancia práctica de la ***solución cluster definitiva***. La validación incluye los intentos del investigador por asegurarse que la solución cluster es representativa de la población en estudio. ***La aproximación más directa en este sentido es realizar el análisis cluster para muestras distintas***. El investigador también puede intentar establecer alguna forma de **criterio o validez predictiva.**, en este sentido dos técnicas son las más utilizadas, el Análisis de Varianza y la técnica del Análisis Discriminante.

En nuestro ejemplo, podemos considerar que los cluster preliminares obtenidos presentan muy poca diferenciación, y esto nos lleva a ***realizar el análisis cluster para una muestra distinta***. En este ejemplo en particular, la validación se realizará con una nueva muestra en la que se han removido (eliminado) los casos 13 y 16, por considerarlos como datos atípicos, creándose de esta manera, una nueva base de datos llamada "VALIDAR EL CLUSTER", la que contiene las mismas variables tipificadas. Se procedió de nuevo a correr el comando **<Classify/ Hierarchical Cluster/ en la ventana de diálogo /... En** la opción **Statistics**, seleccionar **"Single solution"** ***y se escribe el número específico de clusters 4***. En la opción **Save**, se selecciona **Cluster Membership: "Single solution"** ***y se escribe el número específico de clusters 4***, para generar la variable ***CLU4_1,*** dar **Continue**. Finalmente, **dar OK**.

Los resultados obtenidos de la validación de la solución cluster, se presentan en los cuadro 13.7, y 13.8, así como el dendrograma correspondiente se presenta en la figura 13.4.

Cuadro 13.7. Análisis cluster para la validación de la solución cluster preliminar.

Agglomeration Schedule

				Stage Cluster First Appears		
Stage	Cluster 1	Cluster 2	Coefficients	Cluster 1	Cluster 2	Next Stage
1	6	10	.707	0	0	3
2	3	5	1.031	0	0	4
3	6	14	1.233	1	0	9
4	3	13	1.254	2	0	9
5	2	8	2.018	0	0	13
6	9	11	2.475	0	0	12
7	1	4	2.739	0	0	12
8	7	15	3.373	0	0	10
9	3	6	4.070	4	3	11
10	7	12	5.056	8	0	11
11	3	7	8.805	9	10	13
12	1	9	9.103	7	6	14
13	2	3	15.914	5	11	14
14	1	2	33.448	12	13	0

Copyrighted material

Dendrogram using Average Linkage (Between Groups)

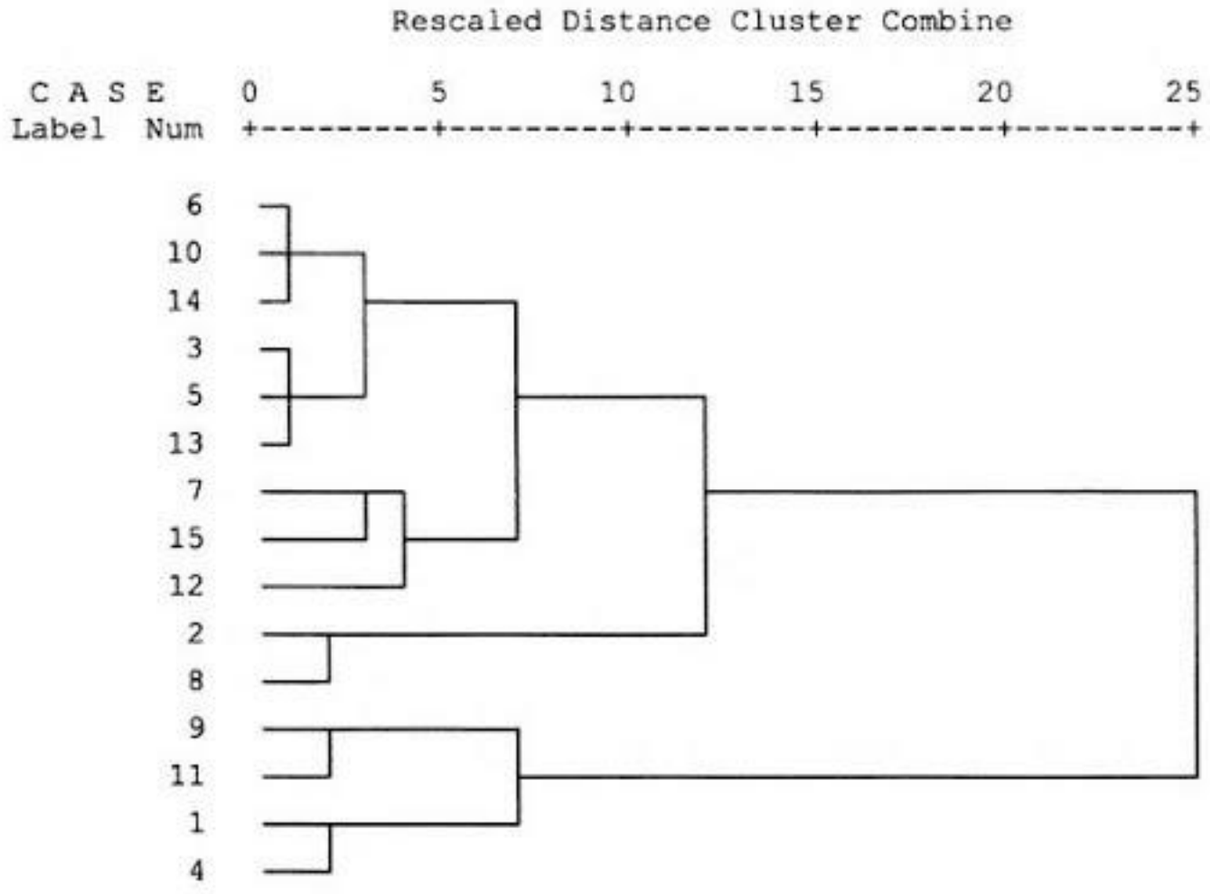


Figura 13.4. ***Formación Jerárquica de Conglomerados Aglomerativos, de casos, para la validación de la solución cluster preliminar.***

Cuadro 13.8. ***Membresía de cada uno de clusters. Solución con cuatro conglomerados.***

Cluster Membership

Case	4 Clusters
1	1
2	2
3	3
4	1
5	3
6	3
7	3
8	2
9	4
10	3
11	4
12	3
13	3
14	3
15	3

Los resultados obtenidos de la validación de la solución cluster, presentan ahora una mayor diferenciación de los grupos conformados. Esto se evidencia tanto en los casos de los cuatro clusters dados en el cuadro 13.8, como en el dendrograma presentado en la figura 13.4.

En base al cuadro 13.8, se presentan los miembros de los 4 clusters, ***para establecer así la clasificación de los sistemas de producción agrícola existentes en la comunidad de Ochomogo.***

Se debería determinar tanto la fiabilidad como la validez de las soluciones que se hayan alcanzado. En cuanto a la fiabilidad, se debe probar que los resultados son consistentes, utilizando diferentes métodos de agrupación. Si se ha demostrado la fiabilidad, aún queda por probar la validez, tanto externa (*la solución sobre la muestra es representativa de la población*), como interna (*si es útil para predecir resultados*). En cuanto a la validez externa, se debe acudir a una muestra parecida de la población que determine los mismos resultados. Respecto a la interna, se deben utilizar contrastes estadísticos que permitan establecer la consistencia de la solución, tales como el ANOVA y el análisis discrimante. ***Finalmente, cabe destacar que la mayoría de estos aspectos dependen del juicio del investigador, ya que ni los mismos expertos en la técnica cluster se ponen de acuerdo en cuanto a su utilización***, (Gómez S. M., 1998).

13.5 El Método Jerárquico de Conglomerados para Variables

El método jerárquico de conglomerados, se utiliza también para encontrar **grupos homogéneos de variables**. El criterio seguido en el proceso de aglomeración es exactamente el mismo que el utilizado para la agrupación de casos. Sin embargo, ***la medida de similitud entre los elementos del análisis cluster, es en general distinta. Cuando los elementos de análisis son las variables, una medida muy utilizada es el valor absoluto del coeficiente de correlación de Pearson*** (**"R"**), que tiene en cuenta el grado de asociación lineal entre cada par de variables, independientemente de la dirección de dicha asociación, es decir independientemente del signo que tenga **"R"**. En el análisis de conglomerados para variables, dos elementos del análisis estarán próximos cuando el valor de R sea próximo a 1, y estarán alejados entre si cuando este sea próximo a 0.

Cuando se desea realizar el análisis cluster de variables, se plantea el inconveniente de realizar ese análisis con variables que están correlacionadas entre si, y por tanto se recomienda determinar los subconjuntos del conjunto original de variables tales que, por un lado, dentro de un mismo subconjunto, las variables estuvieran correlacionadas entre si, y por otro lado, cualquier par de variables pertenecientes a dos o más subconjuntos diferentes estén incorreladas ***-no correlacionadas entre si-***, (Ferran A. M., 1996).

Para ilustrar el proceso de formación de conglomerados para variables, por el método promedio entre grupos, se analizará una parte de los datos del estudio realizado por Mejía, Guzmán, Obregón y Palma, (2005). Serán analizadas las variables correspondientes a la base de datos OCHOMOGO. Debe recordarse que, ***al inicio del análisis de conglomerados para casos***, las variables observadas sobre cada caso, eran 12, es decir, el subconjunto elegido estaba formado por las variables: Área de Maíz, Área de Frijol, Área de Sorgo, Área de Arroz, Área de Tomate, Área de Chiltoma, Área de Pipián, Área de Yuca, Área de Camote, Área de Quequisque, Edad del productor(a), Salario quincenal del productor(a).

Sin embargo, del subconjunto de 12 variables, se eliminaron las variables Área de Camote, Área de Quequisque, dado que presentan solo valores de 0. Así mismo, se planteó el inconveniente de realizar el análisis con aquellas variables que estuvieran correlacionadas entre si, y se comentó que era posible determinar tres subconjuntos del conjunto original de variables, tales que: por un lado, dentro de un mismo subconjunto, las variables estuvieran correlacionadas entre si, y por otro, cualquier par de variables pertenecientes a dos subconjuntos diferentes estuvieran **no** correlaciondas entre si.

Copyrighted material

La rutina para realizar el Análisis Cluster para Variables en SPSS, es la siguiente:

Primero, cargar la BDD OCHOMOGO.

Segundo, se ejecuta el comando **<Classify/ Hierarchical Cluster>** en la ventana de diálogo se incluyen las 10 variables del subconjunto elegido para el análisis cluster. En la ventana ***Cluster***, debe seleccionarse ***la opción Variables***; y en la ventana ***Display seleccionar las opciones Statistics y Plots***.

Tercero, especificar las opciones, tal como: en la opción **Method**, se selecciona el método de aglomeración dado por **"Between group linkage"**, ***también se selecciona el intervalo de medida "Pearson correlation"*** y en **Transform Measures**, dar **Absolute values**; luego **dar continue**. En la opción **Statistics**, se selecciona la pertenencia al conglomerado marcando con un check en **"Agglomeration schedule"**, también se debe marca con un check en **"Proximity Matrix"**, y dar **continue**. En la opción **Plots**, se solicita el gráfico del **"Dendrograma"** correspondiente y marcar **"All clusters"**, y dar **continue**. Finalmente, **dar OK** para correr la rutina del análisis cluster.

La formación de conglomerados de variables por el método promedio entre grupos, ***considerando como medida de similitud la correlación de Pearson***, se observará en el archivo de salida que proporciona el SPSS, el cual brinda la **"Matriz de correlación de Pearson en valor absoluto"**, presentada en el cuadro cuadro 13.9. Además, la salida del SPSS brinda la información detallada de lo que sucede en cada etapa, definido en el calendario de aglomeración presentado en el cuadro 13.10, "Aglomeration Schedule using Average Linkage (Between Groups), y el dendrograma.

Cuadro 13.9. Matriz de correlación de Pearson en valor absoluto.

Proximity Matrix

Case	Matrix File Input									
	Edad (años)	Salario quincenal	Area maiz primera	Area Frijol primera	Area Sorgo primera	Area Arroz primera	Area Tomate primera	Area Chiltoma primera	Area Pipián primera	Area Yuca primera
Edad (años)	.000	.191	.429	.196	.274	.164	.160	.225	.057	.459
Salario quincenal	.191	.000	.161	.007	.298	.264	.150	.150	.639	.014
Area maiz primera	.429	.161	.000	.698	.458	.199	.220	.279	.093	.133
Area Frijol primera	.196	.007	.698	.000	.121	.182	.772	.249	.074	.074
Area Sorgo primera	.274	.298	.458	.121	.000	.174	.339	.031	.373	.145
Area Arroz primera	.164	.264	.199	.182	.174	.000	.110	.641	.191	.249
Area Tomate primera	.160	.150	.220	.772	.339	.110	.000	.062	.108	.141
Area Chiltoma primera	.225	.150	.279	.249	.031	.641	.062	.000	.108	.141
Area Pipián primera	.057	.639	.093	.074	.373	.191	.108	.108	.000	.046
Area Yuca primera	.459	.014	.133	.074	.145	.249	.141	.141	.046	.000

Copyrighted Material

Observando los valores de coeficientes de Pearson, en valor absoluto, puede comprobarse que, por ejemplo, entre las variables Área Fríjol de primera y Área Tomate de primera, la asociación lineal es fuerte, *observe, que el valor de la correlación absoluta entre ellas es igual a 0.772*, mientras que la asociación de una de ellas con respecto a una tercera cualquiera, tiende a ser menor, y aún más débil hasta llegar a la ausencia de asociación entre ellas, con **R = 0.007**, tal como se observa para la correlación de la variable Área Frijol de primera versus Salario quincenal. Así mismo se observa una baja correlación de la variable Área Frijol de primera con relación a Área Pipián primera y Área Yuca primera. En consecuencia, se evidencia que pueden ser consideradas como un subconjunto de información separado del resto. En otros casos las relaciones de dependencia entre las variables no son tan fuertes y, por tanto, no es fácil determinar subconjuntos de variables similares entre sí. Para facilitar la interpretación, se procede a analizar el calendario de aglomeración y el dendograma.

Cuadro 13.10. Calendario de aglomeración, usando el Método Jerárquico Aglomerativo Promedio entre Grupos. Análisis cluster para Variables

Agglomeration Schedule

	Cluster Combined			Stage Cluster First Appears		
Stage	Cluster 1	Cluster 2	Coefficients	Cluster 1	Cluster 2	Next Stage
1	4	7	.772	0	0	5
2	6	8	.641	0	0	7
3	2	9	.639	0	0	6
4	1	10	.459	0	0	7
5	3	4	.459	0	1	8
6	2	5	.335	3	0	9
7	1	6	.194	4	2	8
8	1	3	.185	7	5	9
9	1	2	.150	8	6	0

Al inicio del proceso de aglomeración se considera que cada variable es un conglomerado, de ahí que, habrá tantos conglomerados como variables sean objeto de análisis. De manera abreviada, se presentan tales conglomerados a continuación:

$$C_1 = \{Var\ 1\},\ C_2 = \{Var\ 2\},\ C_3 = \{Var\ 3\},\ C_4 = \{Var\ 4\},\ \ldots\ldots\ldots\ C_{10} = \{Var\ 10\},$$

En el cuadro 13.10, se observa en la primera etapa del calendario de aglomeración, (Stage) que se combina el par de variables que tiene la máxima correlación en valor absoluto. Se combinan las variables *Var 4* y *Var 7*, ("Cluster combined: Cluster 1 = 4, Cluster 2 = 7), o lo que es equivalente los conglomerados C_4 y C_7, (las variables Área Fríjol de primera y Área Tomate de primera), y la similitud entre ellos ("Coefficients") es igual a 0.772.

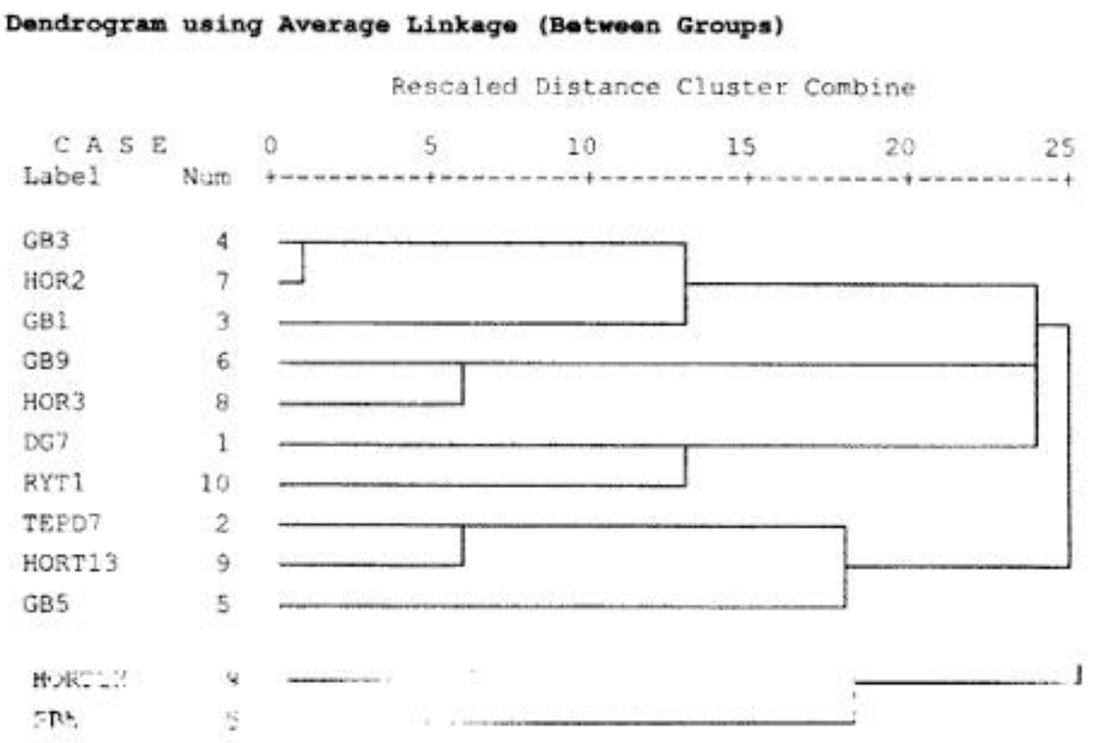


Figura 13.5. ***Formación Jerárquica de Conglomerados Aglomerativos, de variables.***

El dendograma, presentado en la figura 13.5, muestra las líneas correspondientes a las *Var*4 y *Var7*, siendo las dos primeras que se cierran en una única línea. En consecuencia, a partir de la altura del cierre, únicamente quedarán 9 líneas, correspondientes a los 9 conglomerados resultantes después de combinar los conglomerados iniciales C_4 y C_7 en un único conglomerado, ***el cual adoptará el nombre del mínimo número de variable a la que contenga, en este caso*** C_4.

Después de **la etapa 1**, la solución obtenida es:

C_1= *{Var 1}*, C_2= *{Var 2}*, C_3= *{Var 3}*, C_4= *{Var 4, Var 7}*, C_5= *{Var 5}*, ……..C_{10}= *{Var 10}*,

Resumiendo el proceso, los conglomerados obtenidos después de la segunda etapa y sucesivas etapas son los siguientes:

Etapa 2: C_1= *{Var 1}*, C_2= *{Var 2}*, C_3= *{Var 3}*, C_4= *{Var 4, Var 7}*, C_5= *{Var 5}*, C_6= *{Var 6, Var 8}*, …… C_{10}= *{Var 10}*,

Etapa 3: C_1= *{Var 1}*, C_2= *{Var 2, Var 9}*, C_3= *{Var 3}*, C_4= *{Var 4, Var 7}*, C_5= *{Var 5}*, C_6= *{Var 6, Var 8}*, …… C_{10}= *{Var 10}*,

Etapa 4: C_1= *{Var 1, Var 10}*, C_2= *{Var 2, Var 9}*, C_3= *{Var 3}*, C_4= *{Var 4, Var 7}*, C_5= *{Var 5}*, C_6= *{Var 6, Var 8}*,

Etapa 5: C_1 = *{Var 1, Var 10}*, C_2 = *{Var 2, Var 9}*, C_3 = *{Var 3, Var 4, Var 7}*, C_5 = *{Var 5}*, C_6 = *{Var 6, Var 8}*,

Etapa 6: C_1 = *{Var 1, Var 10}*, C_2 = *{Var 2, Var 9, Var 5}*, C_3 = *{Var 3, Var 4, Var 7}*, C_6 = *{Var 6, Var 8}*,

Etapa 7: C_1 = *{Var 1, Var 10, Var 6, Var 8}*, C_2 = *{Var 2, Var 9, Var 5}*, C_3 = *{Var 3, Var 4, Var 7}*,

Etapa 8: C_1 = *{Var 1, Var 10, Var 6, Var 8, Var 3, Var 4, Var 7}*, C_2 = *{Var 2, Var 9, Var 5}*,

Etapa 9: C_1 = *{Var 1, Var 10, Var 6, Var 8, Var 3, Var 4, Var 7, Var 2, Var 9, Var 5}*,

Para nuestro ejemplo, en el Cuadro 13.10, se observan las medidas de similitudes entre los conglomerados que se combinan en cada etapa, los que se presentan en la columna "Coefficient". Debe observarse que para las variables Área Fríjol de primera y Área Tomate de primera, mientras el primer valor de similitud es próximo a 1, (Coefficients = 0.772), a partir del segundo valor es inferior a 0.65. Luego, en nuestro ejemplo, a excepción de las variables Área Frijol de primera y Área Tomate de primera, cualquier solución, y en particular la de tres conglomerados, (ver etapas 6 y 7), se verifica que la similud de las variables incluidas en un conglomerado no es muy grande.

Copyrighted material

Capítulo 14. El Análisis Discriminante.

14.1 ¿Qué es el Análisis Discriminante?

El análisis discriminante, es un método multivariado el cual comúnmente se utiliza para discriminar un conjunto de datos. Por ejemplo, "los compradores" o "no compradores" de un producto determinado, se discriminan en base a una serie de características, tales como: sociodemográficas, forma de vida, etc., y en general este análisis se utiliza para discriminar diferentes grupos de individuos (plantas, animales, personas, productos, etc.), a partir de una serie de variables independientes, (Visauta V., 1998)

El análisis discriminante como método multivariante, permite:

- "Explicar" la pertenencia de un individuo a uno u otro grupo en función de variables independientes, cuantificando la importancia relativa de cada una de ellas.

- "Predecir" a que grupo pertenece un individuo que no forma parte de los datos analizados, y del cual conocemos el valor de las variables en ese individuo, pero no sabemos a que grupo pertenece, (González L., 1991).

El análisis discriminante asume ciertas asunciones, a saber:

a) Cada grupo o tratamiento de estudio debe ser una muestra de una población con una distribución normal multivariada.

b) La variable dependiente, la que hace grupos, debe ser discreta con más de dos grupos. En caso que la variable dependiente sea dicotómica, se puede usar otro tipo de análisis multivariado como la Regresión Logística, (Visauta V., 1998)

En situaciones con una mezcla de variables explicativas continuas y discretas, la función lineal discriminante no siempre es la óptima. En el caso de tener variables dicotómicas, la mayoría de las evidencias sugiere que la función lineal discriminante a menudo funciona razonablemente bien (SPSS/PC, 1988). Cuando el número de variables explicativas es grande, las variables discretas que tienen "n" categorías se deben transformar en "n-1" variables dicotómicas de valores 0 y 1 cada una.

14.2 Un Estudio de Caso realizado mediante el Análisis Discriminante

Para ilustrar el análisis discriminante, se analizará solo una parte de los datos del estudio realizado por Dicovskiy y Rizo, (1997). Se inicia con un proceso de selección de variables, agrupando aquellas variables que tengan altas correlaciones entre sí. Luego con cada grupo se hace un análisis discriminante, buscando agrupar aquellas variables que expliquen mejor la variable dependiente de la investigación. En nuestro ejemplo se evaluaron en total 100 plantas de tempate, (*Jatropha curcas L),* especie arbustiva de zona seca y con potencial energético. Lo que se trató de comprobar es si existía variabilidad genética entre diferentes poblaciones de cuatro localidades de la región norte de Nicaragua: 1) La Concordia- Yalí a 900-1050 msnm.; 2) Estelí, a 800-900 msnm.; 3) Condega a 600-700 msnm; y 4) Pueblo Nuevo a 650-750 msnm.

Copyrighted Material

En este trabajo, el análisis discriminante se utilizó para crear un modelo que permitiera predecir a que población pertenecía una planta que fuera tomada al azar en esta región. Si el modelo funciona, se puede deducir que las poblaciones son diferentes y se puede estimar en función de sus medidas, con que probabilidad una planta pertenece a un grupo dado. Los datos fueron recolectados sobre plantas adultas y cultivadas en una colección, con condiciones de manejo y suelos uniformes. En el cuadro 14.1 se presentan las variables en estudio.

Cuadro 14.1. Variables elegidas para realizar el análisis discriminante.

No	Nombre de la Variable	Tipo de Variable
1	Largo de la lámina de la hoja, en cm: (LARGO_H)	Cuantitativa continua
2	Ancho de hoja, en cm: (ANCHO_H)	Cuantitativa continua
3	Longitud del pecíolo de la hojas, en cm: (LONG_P)	Cuantitativa continua
4	Dos Lóbulos en las hojas, LOB_2	Dicotómica
5	Tres Lóbulos en las hojas, LOB_3	Dicotómica
6	Cuatro Lóbulos en las hojas, LOB_4	Dicotómica
7	Micropelos raros en las hojas MIC_RARO	Dicotómica
8	Micropelos comunes en las hojas MIC_com	Dicotómica
9	Hojas múltiples: (MUL_H)	Dicotómica
10	Porte de la planta : (PORTE)	Dicotómica
11	Distancia entre hojas, DIST_HO	Cuantitativa continua
12	Grozor del tallo en mm, GROSOR_T	Cuantitativa continua
13	Localidad, Local	Discreta

Primero se carga la BDD "DISCRIMINANTE-TEMPATE", y se realiza el análisis de correlación entre todas las variables. La rutina desarrollada ***para estudiar las correlaciones*** en SPSS fue la siguiente: **<Analize / Correlate / Bivariate>;** se declaran todas variables, excepto la variable localidad, que es la que hace los grupos de estudio.

A partir del análisis de correlación, se incorporaron al análisis discriminante las nueve variables que tenían algún grado de correlación. ***Luego, el estudio discriminante se realiza con los comandos*** **<Analize / Classify / Discriminant>,** como variable de agrupación se declara "Localidad", y se define el rango de la misma, en este caso de 1 a 4 localidades. Como variables independientes se toma el resto de las variables. En esta ventana de dialogo, **se toma la opción *"usar método de inclusión por pasos"***, para que durante el análisis se manifieste cuales variables no aportan a la mejora del modelo. ***Para que se realice el mapa territorial se debe buscar la ventana*** **<Discriminant /Classify/ Plots / Territorial map>.**

De las 12 variables iniciales las ***"funciones discriminantes finales"***, quedaron formadas por siete variables. El análisis discriminante removió de la función final a las variables "2 y 3 Lóbulos por hoja", "Hojas múltiples", "Porte de la plantas" y "Distancia entre Hojas"

Copyrighted material

14.2.1. Coeficientes no estandarizados de las Funciones Discriminantes

A continuación, en el cuadro 14.2, se presenta una parte de la hoja de salida del SPSS, en la cual se detallan los valores que tomaron las funciones discriminantes del grupo final. Como hay 4 poblaciones ó grupos, el número de funciones discriminantes calculadas es de k-1 o sea 4 - 1 = 3.

La primera función es el mayor cociente entre la variabilidad entre grupos y la variabilidad dentro de los grupos.

La segunda función está incorrelada (no correlacionada) con la primera y es el siguiente coeficiente mayor (González L., 1991y SPSS/PC, 1988). A continuación se detallan los valores que tomaron las funciones generadas.

Cuadro 14.2. Valores que tomaron las funciones discriminantes del grupo final.

Variable	Función 1	Función 2	Función 3
LOB_4	.6149168	1.536975	.2094860
LARGO_H	.2643860	-.4079299	.1317370
ANCH_H	.1847059	.4907421	-.2377801
LONG_PE	-.1122560	-.1596141	.1695111
GROSOR_T	-.1993729	.3550826	1.048689
MIC_RARO	-1.239799	.2728176	1.942288
MIC_COM	-.5010023	-1.154598	.9254884
(constante)	-5.539183	-.5076821	-5.242228

Los coeficientes **no** estandarizados son los multiplicadores de las variables, cuando estas están expresadas en su unidad original. Si se considera como ejemplo la planta número 1 de la localidad de la Concordia-Yalí, que tiene: sus hojas con cuatro lóbulos, 16.75 cm de largo de hoja, 19.5 cm de ancho de hoja, 17 cm de longitud de pecíolo, 2 cm de grosor de tallo, y tiene micropelos comunes en las hojas. Su valor Discriminante para la primera función es:

D = 1* .614 9168 + 16.75 * .2643860 + 19.5 * .1847059 - 17 * .1122560 - 2 * .1993729 - 0* 1.239799 - 1 * .5010023 - 5.539183

Este valor discriminante, **"D"**, permite ubicar la planta dentro de una población ó grupo dado y con una probabilidad de pertenencia asociada. Para esto, el modelo facilita una regla de clasificación basada en el teorema de Bayes (Visauta V., 1998) y la probabilidad de que una con una puntuación discriminante pertenezca a uno u otra población se estima a través de:

$$P(G_i / D) = \frac{P(D/G_i)P(G_i)}{\sum_{1}^{4} P(D/G_i)P(G_i)}$$

Material

Donde "$P(G_i)$" es la probabilidad previa, si no se tiene ninguna información previa sobre la misma, en este caso al tener 4 grupos es del 25 %. "$P(D/G_i)$" es la probabilidad condicional, que nos da una idea de cuán probable es una puntuación discriminante cualquiera para los miembros de uno u otro grupo. Y "$P((G_i/D)$" es la probabilidad posterior, que nos dice cuán probable es que un sujeto cualquiera de la muestra, con una puntuación discriminante determinada, pertenezca a uno u otro de los 4 grupos o poblaciones. Es así que el programa aporta una tabla de clasificación detallada por individuo con sus probabilidades de pertenecer a una u otra población.

14.2.2 Coeficientes Estandarizados de las Funciones Discriminantes

Como los coeficientes no estandarizados, no son un buen indicador de la importancia relativa de cada variable en la función discriminante, cuando estas difieren en la unidad de medida, es que se construye los "coeficientes estandarizados", con media 0 y desviación estándar de 1. Estos coeficientes estandarizados permiten una aproximación a la importancia relativa de cada variable en las funciones discriminantes y su interpretación es semejante a los de la regresión múltiple. Variables con mayor coeficiente, sin importar el signo, contribuyen con mayor peso en la función discriminante (Bizquerra, 1989 y SPSS/PC, 1988). En el cuadro 14.3, se ilustra los coeficientes estandarizados obtenidos.

Cuadro 14.3. Coeficientes estandarizados.

Variable	Función 1	Función 2	Función 3
LOB_4	.28785	.71948	..09806
LARGO_H	.72498	-1.11860	.36124
ANCH_H	.56602	1.50385	-.72866
LONG_PE	-.47435	-.67446	.71628
GROSOR_T	-.10939	.19482	.57538
MIC_RARO	-.53145	.11695	.83258
MIC_COM	-.23852	-.54970	.44062

14.2.3 Correlación Canónica y Variación porcentual

La correlación canónica de las funciones, es la raíz cuadrada del cociente de la suma de cuadrados entre los grupos para una función dada y la suma de cuadrados total. Es una proporción de la variación explicada por las diferencias entre los grupos (localidades) y la variación total.

"El Porcentaje de Variación", es la relación:

$$\frac{Suma\ de\ Cuadrados\ Función\ n_i}{\sum_{i=1}^{k-1} Suma.de.cuadrados.Función\ n_i}$$

Este porcentaje es una medida de los méritos de cada función en relación a las otras funciones, (SPSS/PC, 1988). En el cuadro 14.4, se presentan los valores obtenidos.

Copyrighted material

Cuadro 14.4. Porcentaje de Variación y Correlación Canónica

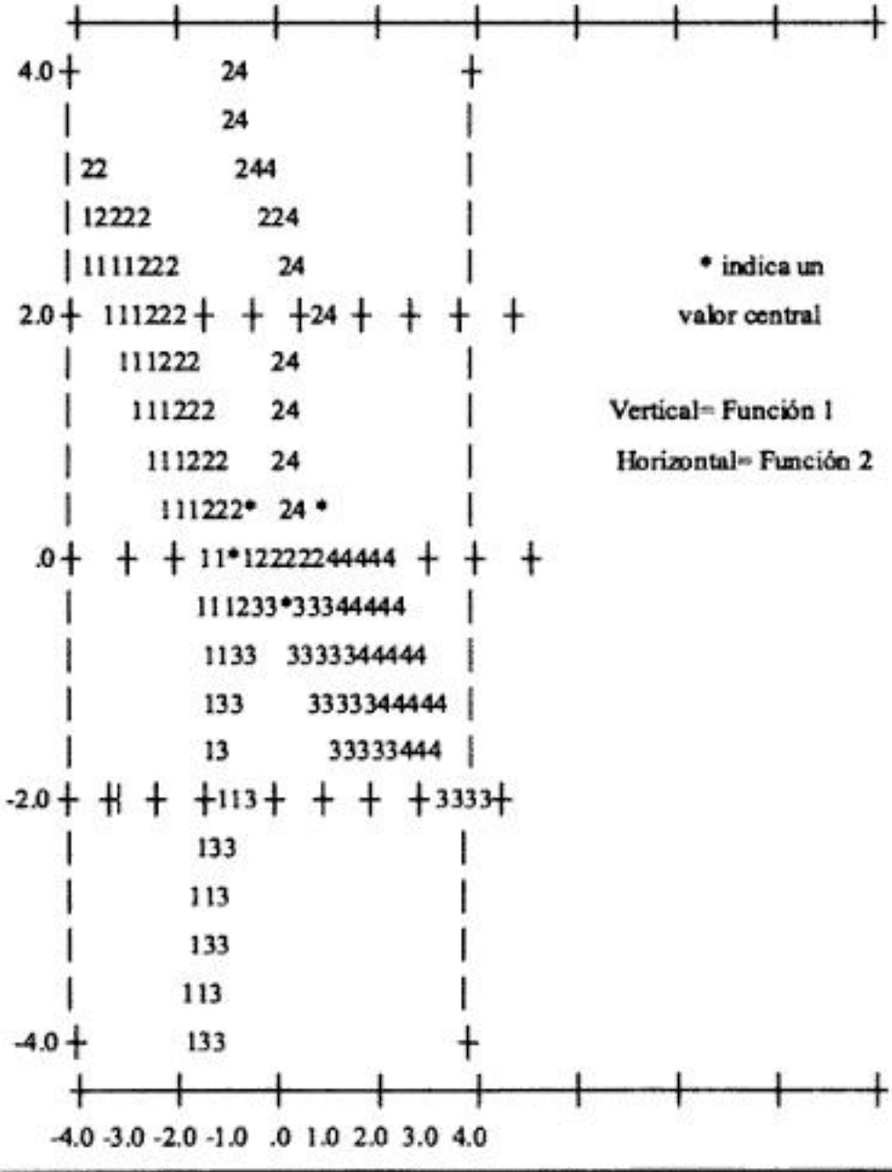
Función	Porcentaje de variación	Correlación canónica
1*	**51.04**	**0.6118**
2*	32.64	0.5260
3*	16.32	0.4007

Para la primera función, la de mejor ajuste, la correlación canónica tiene un valor de 0.6118 y contiene el 51.04 % de la variación total entre localidades.

14.2.4 Correlación Mapa Territorial

El SPSS diseña un mapa territorial donde se pueden ubicar las plantas según el valor de la primera y segunda función en un cuadrante dado, es decir ***"ubicar la planta en una localidad, con los valores de sus variables"***. En este ejemplo hay tantos cuadrantes como localidades.

Mapa Territorial de las Cuatro Poblaciones en estudio



14.3 Resultados de la Clasificación Final

El cuadro 14.5, muestra el número de sujetos correcta e incorrectamente clasificados sobre el total de la muestra utilizada en el análisis discriminante. Se puede ver en la tabla como el análisis discrimina correctamente 13 + 14 + 21 + 11 = 59 plantas, que sobre un total de 100 plantas, representa el **59 %** de los casos. De manera general, el número de casos correctamente clasificados lo encontramos en la diagonal del cuadro. Es así que el porcentaje de casos correctamente clasificados por el análisis discriminante quedó en un 59 %. Hay que considerar que teniendo 4 grupos (localidades), la probabilidad de clasificación correcta aleatoriamente, de una planta dentro de un grupo, es del 25%, por lo tanto el análisis discriminante, más que duplicó este valor, quedando así demostrado la utilidad del modelo en este ejemplo.

Cuadro 14.5. ***Tabla de clasificación: Número y Porcentaje de Miembros predecidos por grupo, según el análisis discriminante.***

Grupo	Casos / Grupos	1	2	3	4
1. Concordia-Yalí	20	13	5	1	1
Porcentaje de Clasificación		**65.0**	25.0	5.0	5.0
2. Estelí	25	4	14	6	1
Porcentaje de Clasificación		16.0	**56.0**	24.0	4.0
3. Condega	30	2	4	21	3
Porcentaje de Clasificación		6.7	13.3	**70.0**	10.0
4. Pueblo Nuevo	25	4	3	7	11
Porcentaje de Clasificación		16.0	12.0	28.0	**44.0**

Copyrighted material

Bibliografía Citada

1. Bornemann G. 2005. La aplicación de Modelos Multivariantes en Sistemas de Producción Agropecuarios del Municipio de Cardenas, Rivas, Nicaragua. Cuadernos de Investigación, colección Administración de Empresas # 18. Universidad Centroamericana, UCA. Managua, Nicaragua. 64 p.

2. Bornemann G. 2004. Enfoque Sistémico. Curso de Posgrado, en la Maestría de Desarrollo Rural. Universidad Centroamericana, UCA. Managua, Nicaragua. s.p.

3. Bizquerra Alsina, R. 1989. Introducción conceptual al Análisis Multivariable. Edit PPU. Barcelona. p 5, 29-31, 178, 260 y 295-296.

4. Dicovskiy L. 2002. Folletos del Curso"Estadística Aplicada para Análisis de Encuestas en SPSS para Windows". ADESO. Estelí, Nicaragua.

5. Dicovskiy, L y Rizo R. 1997. Análisis univariado y multivariado en plantas de tempate, (Jatropha curcas l.) de cuatro localidades de la zona norte central de nicaragua. Tesis de Maestría. Universidad de Valencia, España. 77p.

6. Ferran A., M. (1996). SPSS para Windows, Programación y Análisis Estadístico. Editorial McGraw-Hill. Mexico, D.F. 580 p.

7. Gómez S. M. 1998. El Análisis Cluster en Investigación de Marketing: Metodología y crítica. Universidad Autónoma de Madrid. Departamento Financiación e Investigación Comercial. Facultad de Ciencias Empresariales. UAM 28049. CANTOBLANCO. Madrid. pp 537-543.

8. González López, B. 1991 La Estadística Multivariante y la Investigación Sanitaria. España. p 175-177, 203-205.

9. Little, M.T & Hills, F. J. (1981). Métodos estadísticos para la investigación en la agricultura. Editorial Trillas, México, D.F. 268 p.

10. Mejía, I., Guzmán, M., Obregón S., y Palma, X. (2005). Estrategias de Desarrollo para la Micro Cuenca Pata de Gallina. Tesis de Maestría en Desarrollo Rural. Universidad Centroamericana, UCA. Managua, Nicaragua. 64 p. 152 p.

Copyrighted material

11. Munch Galindo, Lourdes. (1996). Métodos y Técnicas de Investigación. Editorial Trillas. Tercera Reimpresión. 165 p.

12. Piura, L. J. (1994). Introducción a la metodología de la investigación científica. Editorial el Amanecer, S.A. Managua, Nicaragua. 114 p.

13. Pedroza, P.H. (1993). Fundamentos de Experimentación Agrícola. Editora de Artes, S.A. Managua, Nicaragua. 226 p.

14. Pedroza P.H. (1995). Sistema de Análisis Estadístico aplicado a la Experimentación Agrícola. Curso de Post Grado, UNA- FAGRO. Managua, Nicaragua. 113 p.

15. Pedroza, H.P. 1995. Los Sistemas de Información: Instrumento vital para la sostenibilidad de los procesos de GTTA. Sociedad Agrícola. Año 2, No. 3. Managua, Nicaragua.

16. Pedroza, H.P., y Dicovskiy L. 2003. Manual del curso "Técnicas de Investigación Cuantitativa y Cualitativa". Maestría en Desarrollo Rural. UCA. Managua, Nicaragua.

17. Reyes, C. P. (1982). Diseño de experimentos aplicados. Segunda reimpresión, editorial Trillas. México, D.F. 343 p.

18. SPSS 7.5. (1997). Estadísticas Avanzadas de SPSS 7.5. SPSS Inc. Impreso en Irlanda. 107 p.

19. SPSS/PC V.3.0. Manual. 1988 Advanced Statistics. Edit SPSS Inc. USA. pp B-1, B-23/26, B-33, B-64, B-69/70.

20. Visauta, V. B. 1998. Análisis Estadístico con SPSS para windows, -Estadistica Multivariante-. Escuela Superior de Administración y Dirección de Empresa. (ESADE). Mc Graw Hill/ Interamericana de España, S.A.U. pp 167-212.

Copyrighted material

Copyrighted material

Henry Pedroza Pacheco, nació en Nandaime, Granada, el 4 de Octubre de 1958. En 1982, se graduó de Ingeniero Agrónomo Fitotecnista, en la Facultad de Ciencias Agropecuarias de la UNAN, hoy UNA. En 1991, obtuvo el grado científico de Doctor en Ciencias Agrícolas, en la Universidad de Agraria de Plovdiv, Bulgaria.

En los primeros once años de su vida profesional, (1982-1992), se desarrolló como docente investigador del ISCA, actual UNA, inicialmente en la cátedra de Economía Agrícola y luego en Diseños Experimentales, apoyando la formación básica de los investigadores agropecuarios del país.

En los últimos quince años, de 1992 a la fecha, se ha desempañado como consultor de sistemas de información tecnológica, que requieren Sistema de Manejo de Base de Datos (DBMS), en formulación de proyectos mediante el EML, en evaluación de programas/proyectos y como Biometrista para el análisis de datos tanto experimentales como no experimentales. De 1996 a 2001, se desempeñó como Director de Generación de Tecnología del INTA, cumpliendo con éxito funciones de planeación, diseño, monitoreo y evaluación de programas y proyectos de investigación. En el año 2000, realizó estudios de postgrado en la Universidad de California, Davis (UC Davis), obteniendo el Diploma Post graduate Certificate Program on Vegetable Crops. Se ha desempeñado como docente universitario en la UNA, UNI, UNN, UCA, e IICA, en diversos cursos de investigación. Ha asesorado varias tesis universitarias de pregrado y dos tesis de maestría.

Luis María Dicovskiy Riobóo, nació el 24 de Junio de 1956, en la ciudad de Rosario, Argentina, es nacionalizado nicaragüense y actualmente reside en Estelí. Se graduó de Ingeniero Agrónomo en la UNR, Universidad Nacional de Rosario, Argentina.

Obtuvo una especialidad en "Mejora Vegetal", en el Instituto de Altos Estudios del Mediterráneo, Zaragoza, España. Alcanzó el grado de Maestría en "Estadística e Investigación de Operaciones", otorgado por la Universidad de Valencia, España, y es egresado de la III Maestría en "Métodos de Investigación Social Cualitativa", de la UPOLI, Nicaragua.

Se ha destacado como Investigador y Docente Universitario por 24 años, tiene amplia experiencia en medición y análisis de datos socioeconómicos, y productivos; como Biometrista y analista de información experimental y no experimental por métodos estadísticos y cualitativos. Ha sido coordinador de varias investigaciones en el campo agropecuario y social en la zona norte de Nicaragua, y tutor de múltiples tesis universitarias.

Se ha desempeñado como: Jefe del departamento de Investigación y Post-grado en la EAGE de la Escuela de Agricultura y Ganadería de Estelí. Febrero de 1992-1999. Director Ejecutivo (EAGE,) de 1999-2001. Miembro de la Comisión de Directores Investigación y de la Comisión de Directores de Postgrado del Consejo Nacional de Universidades, CNU. 1998 a 2002. En el año 2005 se desempeñó como Coordinador de la Carrera de Agroindustria de la UNI, Sede Estelí. Desde 2006 a la

Copyright Material